DISSERTATION

THE GENERATIVE FRICTION TRILOGY: LEARNING TO INTERVENE, EVALUATE, AND COLLABORATE IN MULTI-AGENT COLLABORATION

Submitted by

Abhijnan Nath

Department of Computer Science

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Spring 2026

Doctoral Committee:

Advisor: Nikhil Krishnaswamy

Nathaniel Blanchard
Ewan Davies
Michael Kirby

ABSTRACT

THE GENERATIVE FRICTION TRILOGY: LEARNING TO INTERVENE, EVALUATE, AND COLLABORATE IN MULTI-AGENT COLLABORATION

Large language models are increasingly deployed as collaborative partners in multi-turn, multi-party tasks, where they must reason jointly with others rather than act in isolation. In such settings, collaboration often requires generating *frictional interventions*—strategic prompts that slow the group down, surface hidden disagreements, or encourage reflection—rather than direct answers. A key challenge is that these interventions do not directly determine how a collaboration unfolds: collaborators may ignore, reinterpret, or partially incorporate them, and the resulting dialogue trajectory depends on how all participants respond. This mediation breaks the assumptions underlying standard alignment and reinforcement learning methods, which implicitly treat an agent's chosen action as directly determining outcomes.

This dissertation approaches this challenge through a trilogy of complementary perspectives. First, we introduce the Frictional Agent Alignment Framework (FAAF), which provides the first principled formulation of friction intervention learning that generalizes to previously unseen collaborative settings. Second, we extend this framework to realistic collaborative settings where different intervention agents interact with collaborators over multiple turns, producing divergent dialogue trajectories. We formalize collaborative dialogue as a Modified-Action Markov Decision Process (MAMDP) where collaborators modify interventions before they affect the environment, and introduce a counterfactual roleplay framework that isolates causal effects while controlling for collaborator dynamics. We show that friction interventions can accelerate common-ground convergence, but that even optimal interventions fail without appropriate collaborator

responses. Third, having shown that collaborator responses determine outcomes, we develop the Interruptible Collaborative Roleplayer (ICR) to train effective collaborators, through a constrained optimization objective that combines task utility maximization with counterfactual invariance—robustness to spurious features in interventions. Empirical results show ICR-trained collaborators outperform frontier models and achieve superior common-ground convergence without explicit team-level training signals.

Practically, this work provides the first end-to-end framework for training and evaluating both sides of collaborative AI systems, moving beyond dyadic assistant-user interactions toward principled small-group collaboration where friction serves intelligence rather than impedes it. By bridging offline alignment that is robust to data-bias, causal evaluation, and collaborator training, we establish friction-aware alignment as a fundamental challenge at the intersection of NLP and multi-agent RL, prove that standard single-agent methods are suboptimal for collaboration, and provide novel protocols for the realistic multi-party collaboration toward which LLM deployment is moving.

ACKNOWLEDGEMENTS

I want to sincerely thank Nikhil Krishnaswamy, my PhD advisor, who truly brought out the researcher in me. A brilliant computational linguist and teacher, a trustworthy mentor, and a researcher and scientist par excellence—his intellectual supervision, guidance in finding the right research directions and rigorous critique of my thinking and writing were priceless for my growth as a researcher and for which I am truly indebted. I must thank him for being incredibly generous with his time throughout this PhD and for providing the freedom and courage to pursue what I wanted in research, including the less trodden paths. Closer to the craft of it, he taught me the importance of clarity of thought in framing research questions, the rigor required to produce high-quality work, and the precision in writing and expression—including the art of science communication—that I plan to carry forward and remain forever grateful for. Towards the culmination of this research circa spring 2025, nothing could match going to bed at 6am knowing that cyan lines on Overleaf would appear on blue-lined paragraphs in a relay, while the TRL trainer would spit its final logs and the GPUs would finally breathe. Except one thing: a 7-hour road trip to New Mexico with your PhD advisor tracing most of Southern Colorado and Wyoming while listening to Steely Dan and preparing your talk! I could not have asked for a better advisor.

I want to thank my committee members for all the guidance and support they have shown me, and for convincing me to pursue this line of research, even after multiple changes in tracks. Thanks to Nathaniel Blanchard, my PhD co-advisor, for teaching me to care deeply about problems from a "data-first" perspective. To Ewan Davies, for our discussions on graph clustering that shaped how I approached sampling and distributional methods in this work. Sincere thanks to Michael Kirby, for our conversations on optimization that motivated me to pursue the theoretical angles that later became the backbone of the FAAF chapter.

I want to thank all faculty members at the Computer Science Department at CSU. Special mention to senior faculty Bruce Draper, Chuck Anderson, Sudipto Ghosh, and Indrajit Ray—whose support I felt throughout, including being awarded the Artificial Intelligence and Evolutionary Computing Fellowship in 2024 and the Wim Böhm Fellowship in 2025 including teaching assistantships, all of which helped massively in producing this research. I also want to specially thank Sanjay Rajopadhye for initial guidance before moving to Fort Collins and our conversations after every major step during the PhD; for the summer barbecues at City Park and beers at the Ramskeller—I felt your support throughout and it meant a lot. This would also not have been possible without the department's administrative staff including SNA, so sincere thanks to Megan Brice, Kim Chacon, and Abhimanyu Chawla and others.

Special thanks to senior professors external to CSU—Martha Palmer and James Martin at CU Boulder, and James Pustejovsky at Brandeis—for DARPA's RAMFIS program, which deeply influenced my research in computational linguistics, for writing recommendation letters, and for overall guidance; especially James Pustejovsky for pointing toward fundamental directions in pragmatics and communicative acts that influenced a large part of the frictional alignment work. I am also deeply grateful to Jordan Boyd-Graber at the University of Maryland, Jonathan May at USC, and Jonathan K. Kummerfeld at the University of Sydney for our collaborations on human-AI interaction in Web Diplomacy during the FACT program and through 2025–26.

To my friends and colleagues in graduate school, especially the members of the Signal Lab—you made this PhD so much fun and far less isolating than it could have been. Friday meetings were always a delight! The discussions during 2024–25 designing the "friction stuff" under the FACT program were among the most memorable of my PhD. Shafiuddin Rehan Ahmed and David McNeely-White gave me my earliest entry points into serious research, especially NLP and computer vision respectively; Sheikh Mannan was there from the very first ML class in Spring 2021 that Nikhil taught; our collaboration

during AxomiyaBERTa was priceless as was participating in your flight simulations; Videep Venkatesha and Mariah Bradford—two of my closest collaborators—kept the conversations alive through whiteboard sessions that eventually turned into papers on probing questions and later became "interventions"; and Carine Graff and Andrei Bachinin were foundational to the FACT program—as were the conference trips to Miami, New Mexico and San Diego and the DARPA Demo trips to Stanford and Maryland that marked the program at its height.

To Sina Saravani, Ibrahim Khebour, Shadi Manafi, Sadaf Ghaffari, Anju Gopinath and Rahul Ghosh—I truly enjoyed our collaborations during the earlier years. Shoutouts to Changsoo Jung and Ethan Seefried, with whom I had many great conversations about state-space models and during the early reward modeling works. Particularly, Ethan—you made the NeurIPS '25 experience much more fun. I especially cherish my conversations with Jack Fitzgerald on 3D geometry understanding and Hannah VanderHoeven on AI collaboration with partial information, both of which shaped how I think about these problems.

Thanks to Sifatul Anindho, Tarun Varma Buddaraju, Avyakta Chelle, Austin Youngren, Animesh Gurjar, Sai Kiran, Shamitra Gowra, Carlos Mabrey, Brady Bhalla, Sai Shruthi Garlapati and many others at CSU; Timothy Obiso, Kenneth Lai, and Yifan Zhu at Brandeis; Wichayaporn Wongkamjan and Yanze Wang at the University of Maryland and University of Southern California respectively; and the Boulder NLP group folks—Elizabeth Spaulding, Abteen Ebrahimi and Adam Wiemerslag and many others. You all have had a deep influence into my research. And to Sonu Dileep—I truly enjoyed our regular catching on AI research trends from around the globe, our "late-night" drives to Horsetooth reservoir, and hikes around the Rockies that helped me unwind—you taught me how to enjoy grad life when times were difficult!

I was also lucky to meet some incredible folks during summer internships and from whom I learned to appreciate problem-solving at scale. I'd like to thank Andrey Volozin at

Optum, whose conversations on artificial intelligence in healthcare quietly seeded a large part of my initial forays into the alignment problem in LLMs; Alireza Bagheri Garakani at Amazon, whose customer-first perspective on production ML shaped the Owen-Shapley work in product search RL; and Motoki Wu, Navjot Matharu, Iwona Bialynicka-Birula and Chuan Wang at Cresta, for making that experience as intellectually rich as it was.

To friends and family back home; Kamlesh Borah, Satyajit Chowdhury, Gaurav and Saurav Roy, and the Music Club friends from BITS who motivated me in many ways and to Nilotpal, Bhargav, Bhargav Borah, Uddip, Shyam, Ritwick, and Sunny—thank you for keeping the traceback alive.

To Michaela Kayser, for being an amazing partner, a wonderful friend, and a constant source of support and joy from my master's commencement through every stage this PhD covered; for keeping me sane and grounded through the difficult stretches and believing in me. I look forward to the next chapter with you.

Finally, I am incredibly grateful to my Deuta and Ma—both historians—who raised me on the Dibrugarh University campus and always believed in me. For instilling clarity of thought, a sense of history and a taste for knowledge early on, and for teaching me the value of consistency, discipline, sustained effort and above all, Niṣkāmakarma. "Keep an eye on the eye," you said, and I followed it. Thank you for being a pillar of quiet strength and patience, and for believing I could do this. To my younger sister Imon, who made her own leap to Germany for graduate studies around the same time I moved to the United States—your constant support and parallel courage motivated me to strive harder and keep learning.

Sans statecraft, they chased the winds,

and rode their wheels,

and thickened their nascent skin.

— Cosco Ball Ethics

DEDICATION

I would like to dedicate this thesis to Deuta and Ma.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

Large language models (LLMs) have become remarkably flexible tools [12]. In the past few years, we have seen LLMs adopted in multiple forms—as QA agents [13], web-searching assistants [14], ticket booking systems [15], coding companions [16], and even for personal communication and maintaining relationships [17]. This wide-scale adoption is powered not only by the general expressivity of language that these machines generate fluently, but also by alignment methods such as Reinforcement Learning from Human Feedback (RLHF) [18,19] that serve as the bedrock of such progress. These methods allow us to go beyond learning from static datasets—they let models "experience" the world (as implicitly expressed through text) and enable the model to be more aligned with human preferences and intentions by learning from feedback [20]. In doing so, the narrative around LLMs has become "agentic" [21] in nature. The capacity of autonomously taking decisions in a wide range of tasks, navigating complex environments with search-based tools, and providing long-term support to humans via their dialogue mechanisms have incentivized developers of these systems to make interactions with them smooth and fluid. Smooth and fluid interactions maintain user engagement and condition users to expect immediate answer to their problems. Less "friction" seems desirable at first glance.[1]

1.1 The Collaborative Alignment Challenge

However, human behavior provides a compelling counterpoint to how current LLMs are incentivized to operate. When people collaborate, friction is often desirable [8,22–24]. One mark of an intelligent agent is its ability to synthesize new information from observations, but an equally important ability is to question and examine things they observe

[1] https://www.nytimes.com/2018/12/12/technology/tech-friction-frictionless.html

and sometimes reframe their current knowledge in light of incoming evidence to align with partners or collaborators on mutual goals [25–28]. When humans encounter "*frictive states*"—moments where collaborators hold contradictory beliefs about task-relevant propositions—the ability to resolve these misalignments through reflection and deliberation often determines whether the collaboration succeeds or fails. As humans come to treat LLMs as collaborators, including in the business and personal spheres, and as these collaborations evolve toward a multi-party nature with multiple humans engaging with the same AI instance [29], incentivizing LLMs to productively engage with these human dialogic phenomena becomes critical.

This dissertation adopts the view that to correctly model intelligence in conversations and collaboration, it is important to devise mechanisms and principled frameworks for training LLMs that can recognize and productively navigate frictive states. We focus specifically on small-group collaborative reasoning tasks: scenarios where 2-4 participants must align their beliefs to solve problems together, such as scientific reasoning puzzles or deliberative decision-making. Due to the multi-party nature of collaboration, addressing this challenge requires thinking beyond standard supervised learning and preference alignment methods—the dominant paradigm in optimizing LLMs for dyadic interactions (e.g., the user-assistant scenario [19,20,30,31])—while taking into consideration that perfect human demonstrations for such phenomena for learning them may not be available. From a Reinforcement Learning (RL) perspective [32], one of the ways in which an LLM can navigate a frictive state is to "act" or generate *friction interventions*. Within the language-as-next-token-prediction conceit of LLMs, these interventions constitute the "best response" to encountering such a state: sets of tokens that comprise meaningful utterances that prompt reflection and "slow thinking" [33] in other agents during collaboration.

The Friction Paradox in LLM Research Research on friction in conversational AI reveals a telling paradox. The dominant trend has focused on *eliminating* friction—improving

query rewriting, machine translation, and intent correction to make interactions with systems like Alexa and Google Assistant smoother [34–38]. A smaller line of work explores the opposite direction: *increasing* productive friction through UI scaffolding [2, 39], where interface elements like extra clicks or delays discourage over-reliance, particularly in educational contexts. While valuable for mitigating inappropriate use of deployed systems, this treats friction as post-hoc remediation—the LLM has already been trained for frictionless response, and barriers are added afterward at the interface layer.

This work takes a fundamentally different approach: we treat friction as a *learned generative capability*. Rather than imposing external constraints through UI design or optimizing systems to eliminate friction entirely, we train LLMs to recognize belief misalignments in collaborative dialogue and autonomously generate language interventions that prompt reflection and deliberation. This operates at the training and generation level, addressing the root cause—LLMs' inability to recognize when slowing down would benefit collaboration—rather than only the symptoms of over-reliance or communication breakdown. This moves friction from the periphery of design to the core of intelligence; it is no longer a barrier added to a completed system, but an innate strategy the model uses to navigate the "slow thinking" required for complex, shared goals.

1.2 Why Current Methods Fall Short

That LLMs struggle with frictive collaboration is not surprising given how they are trained at various stages of their lifecycle. Most pretraining data is not in dialogue form [40], and even when it is, it rarely contains examples of productive friction—the kind of pushback [2], probing [8], or deliberative reasoning [33] that characterizes thoughtful human collaboration. Post-training strategies have compounded this limitation [41]. Standard RLHF methods optimize for immediate preference signals, leading to pathologies like sycophancy [42] and scheming [43] that become consequential at scale [44]. The Bradley-Terry (BT) preference modeling underlying most RLHF implementations assumes

stationary preferences and immediate rewards. These assumptions break down in multi-turn collaboration where success emerges over extended interaction [45]. By optimizing for the highest immediate reward, standard alignment methods effectively train out the agent's ability to push back or question the user, even when such friction is necessary for long-term success. Even in the case of learning from offline preference data for single-turn settings, presence of non-deterministic human preference signals makes standard preference alignment methods suffer from pathological limitations [46, 47] and policy degeneracy [48] issues where the learning objective itself is compromised. The core issue stems from the BT's model's over-reliance on the data distribution and reward overoptimization [49, 50].

Beyond Dyadic Interaction: The Multi-Agent Challenge Beyond the aforementioned algorithmic limitations, human communication is inherently deliberative—a process of information-seeking, clarification, and belief revision [51]. Because humans make mistakes and hold incorrect assumptions [22, 52, 53], effective collaboration requires slowing down to establish shared understanding. Yet LLMs are designed for the opposite: immediate, confident responses that maximize user satisfaction in single turns [54].

A deeper issue emerges when we consider group collaboration. Most existing methods assume collaboration occurs only between a single user and assistant [55]. While multi-agent reinforcement learning (MARL) provides a theoretical framework for group interaction [56], applications to LLMs remain limited [57–59] due to instability from asymmetric rollouts, asynchronous updates, and the prohibitive cost of training heterogeneous LLM agents with policy-gradient methods. Role-based prompting offers a practical alternative and has been successfully applied across diverse domains [60–62]. However, roleplay-based approaches are primarily inference-time strategies that assume static role descriptions are sufficient—an assumption increasingly recognized as fragile [63]. Roleplay does not specify how interventions influence future states or support principled credit

assignment [64], and it struggles to generalize to settings with dynamic roles or evolving preferences. More fundamentally, standard MDP formulations [32] are inadequate for modeling collaborative dynamics [45]. They assume an agent's action directly affects the environment, whereas in collaboration, actions are often *modified by collaborators* before the environment responds. As a result, collaboration [65] requires drawn-out interactions or sustained conversations in which these assumptions no longer hold. In particular, single-turn preference alignment breaks down in multi-party collaboration because agent actions shape the future belief states of other participants, giving rise to belief misalignment that necessitates belief-aware modeling. Modeling language or dialogue generation as a standard MDP is therefore insufficient, as it fails to capture the evolution of beliefs and the formation of common ground, defined as the set of mutually agreed task-relevant propositions that characterize realistic multi-agent collaboration. All these challenges point in the same direction: **Unlike standard LLM alignment, which incentivizes immediate responses and minimal friction, *collaborative* alignment requires treating friction as a productive and necessary component of both theoretical modeling and evaluation.**

Organization and Contributions The remainder of this dissertation is organized into the following chapters, each addressing a core aspect of collaborative alignment. We begin by revisiting preference-based alignment itself and the background literature in Chapter 2, addressing the aforementioned core issues with the BT model assumption. In Chapter 3, we demonstrate why standard alignment methods fail under non-deterministic preferences and introduce a novel offline alignment algorithm that explicitly addresses these limitations. We then turn to the problem of small-group collaboration, where preference alignment must operate over evolving belief states rather than isolated responses. Chapter 4 introduces FAAF—an offline algorithm for learning belief state-aware frictional interventions in LLMs that can be optimized purely from precollected dialogues. Building on this foundation, Chapter 5 and Chapter 6 jointly develop a unified framework

Contribution	Problem Addressed	Framework	Key Objective
Foundational	Non-deterministic preferences in RLHF	DRDO (Chapter 3)	Robust offline alignment
The Friction Agent	Generating productive friction	FAAF (Chapter 4)	Frictive-state aware interventions
The Dynamics	Measuring causal impact in live dialogue	Roleplay Framework (Chapter 5)	Counterfactual MAMDP evaluation
The Partner	Learning an ideal collaborator via constraints	ICR (Chapter 6)	Task-relevant invariance

Table 1.1: Overview of the dissertation's core contributions. Each contribution targets a distinct challenge in collaborative AI systems powered by LLMs: (1) robust alignment under noisy preferences, (2) learning when and how to introduce productive friction, (3) evaluating interventions counterfactually in multi-agent dialogue, and (4) training collaborators that prioritize task-relevant information over superficial agreement via constrained optimization.

for multi-agent collaboration by formalizing interactions between two roles: a friction intervention agent (Chapter 5) and a collaborator agent (Chapter 6). These chapters model collaborative problem solving within a Modified-Action MDP (MAMDP) framework [66], where agent actions are mediated through collaborative dynamics rather than acting directly on the environment. Under standard assumptions [46, 67, 68], we prove that while existing preference alignment methods are optimal for their induced reward classes under a token-level MDP, they are suboptimal under the collaborative dynamics of the MAMDP at the intervention level.

A chapter-wise summary of contributions is provided in Table 1.1. This dissertation demonstrates that effective small-group collaboration with AI agents requires treating dialogue as a joint decision-making process shaped by belief misalignments, mediated interventions, and adaptive collaborators. Our theoretical contributions and experiments demonstrate that effective collaborative AI requires training both intervention agents that recognize when to slow down dialogue, and collaborators that extract task-relevant content from interventions. In other words, this work establishes that effective AI collaboration requires a fundamental shift: training agents not just to provide answers, but to strategically deploy "productive friction" to navigate misalignments in shared decision-making.

Chapter 2

Background and Related Work

Chapter Overview This chapter reviews the background research and related work needed to ground the contributions of this dissertation. We focus on preference alignment in large language models (LLMs), where the central goal is to train LLMs to be more likely to generate responses aligned with human preferences compared to a pretrained or supervised-finetuned model [69]. We begin by reviewing standard alignment paradigms such as Reinforcement Learning from Human Feedback (RLHF), including their core assumptions, training pipeline, and data requirements. We highlight key limitations of these approaches, particularly in settings with weak, noisy, or non-deterministic preference annotations, and discuss how popular methods such as Direct Preference Optimization (DPO) exhibit failure modes under these conditions [9]. This background is essential for motivating the offline alignment framework developed in Chapter 3.

Building on this foundation, we broaden the scope from single-agent alignment to collaborative problem solving with LLM agents. We argue that alignment methods designed for static preferences and single-turn interactions are fundamentally limited in multi-agent, goal-oriented collaboration. To address this gap, we introduce *friction* as a generative paradigm for analyzing collaborative dynamics, grounded in formal notions of common ground and frictive states. This perspective provides the theoretical basis for understanding how interventions can resolve misaligned beliefs during collaboration, and directly supports the collaborative models and algorithms introduced in Chapters 4, 5, and 6. The chapter is organized as follows. In Section 2.1, we provide an overview of offline and online preference optimization methods for LLMs. In Section 2.2, we analyze the misalignment between reward-optimal and preference-optimal policies, illustrate the issue in a toy bandit setting, and connect it to empirical failures observed in contemporary preference benchmarks. We then turn to collaborative problem solving with LLMs in

Section 2.3, where we formalize friction, common ground, and frictive states. We then introduce a new formalism for modeling collaborative dynamics for small-group settings in Section 2.4, based on the Modified-Action Markov Decision Process (MAMDP). Finally, we describe the datasets and experimental setups used throughout this dissertation in Section 2.5.

2.1 Preference Optimization in LLMs

Recent advancements in Large Language Models (LLMs) often involve refining pre-trained LLMs for downstream tasks by utilizing human-written completions [70,71] or datasets labeled with human-preferred completions and contrasting alternatives [72–74]. The two most prominent techniques in learning from preference data are Reinforcement Learning from Human Feedback (RLHF) and Direct Preference Optimization (DPO). In the following sections, we summarize these methods.

Reinforcement Learning from Human Feedback Reinforcement Learning from Human Feedback (RLHF) aims to harmonize LLMs with human preferences and values [18]. Conventional RLHF typically consists of three phases: supervised fine-tuning, reward model training, and policy optimization. Proximal Policy Optimization (PPO; [11]) is a widely used algorithm in the third phase of RLHF. RLHF has been extensively applied across various domains, including mitigating toxicity [75,76], addressing safety-concerns [77], enhancing helpfulness [78], web search and navigation [79], and enhancing reasoning in models [80]. Recent work [81] has identified challenges and problems throughout the entire RLHF pipeline, from gathering preference data to model training to biased results, such as verbose outputs [82–84].

Given a dataset of pairwise preference data $\mathcal{D} = \{x^{(i)}, y_w^{(i)}, y_l^{(i)}\}_{i=1}^N$, where $x^{(i)}$ are prompts and $y_w^{(i)}$ and $y_l^{(i)}$ are the preferred and dispreferred completions, respectively, RLHF begins with an initial model π_{ref} that parameterizes a distribution $\pi_{\text{ref}}(y|x)$. While

the context $x^{(i)}$ can take many forms like an instruction or a question [85], completions are generally annotated with humans or other higher-capacity teacher LLMs [86]. Typically, π_{ref} is initialized from an LLM that has undergone supervised fine-tuning (SFT) after pretraining, but one can also directly use a "good enough" base model as the reference model [9]. The preference between y_w and y_l is modeled using the Bradley-Terry model [87], which defines the probability of preferring y_w (the winning response) over y_l (the losing response):

$$p(y_w \succ y_l|x) = \sigma(r(x, y_w) - r(x, y_l)) \tag{2.1}$$

where $\sigma(\cdot)$ is the logistic function, and $r(x, y)$ represents an underlying reward function. To estimate this reward function r, RLHF minimizes the negative log-likelihood of the data [19,73,88]:

$$\mathcal{L}_R(r_\phi, \mathcal{D}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \left[\log(\sigma(r_\phi(x, y_w) - r_\phi(x, y_l)))\right]. \tag{2.2}$$

Optimizing External Rewards Using this learned reward function r_ϕ, policy-gradient based RL methods like PPO [89] are applied to optimize and generate a new LLM distribution π^*_θ—that is typically referred as the "optimal policy"—with a KL-divergence constraint for regularization. The goal is to iteratively update the parameters of π_θ such that rewards obtained from the learned reward model on "actions" (or completions sampled from the policy) are maximized as we reach convergence, while keeping the policy distributionally close to the reference model to prevent model collapse. Often the reward model can be out of distribution (OOD) for the constantly-changing policy and can cause a distributional shift and concomitant reward hacking [90,91]. Works like [92] have discussed advantages of augmenting the reward modeling loss in Eq. 2.2 with a negative-log likelihood term (SFT or "language-modeling" loss) in order to improve generalization to overcome this distributional drift. In Chapter 3, we use this concept to train a teacher reward model for the initial phase of the presented DRDO algorithm, which only requires

fitting an additional linear reward head in addition to the default language modeling head. Multiple works have also discussed the computational complexity of algorithms like PPO especially when applied to LLMs. Unlike smaller neural networks, most base LLMs of today contain billions of parameters [93], leading to practical challenges[2] in optimization and hyperparameter tuning. For instance, PPO has limitations like the computational and memory overhead of tracking four models (reward model, value model, reference model, and policy model) in GPU memory [94, 95].

Offline and Online Preference Optimization Given the complexity of online preference optimization discussed above, research has proliferated into more efficient and simpler offline algorithms. Direct Preference Optimization (DPO) [9] is a notable example, which demonstrates that the same KL-constrained objective as RLHF can be optimized without explicitly learning a reward function. The problem is reformulated as a maximum likelihood estimation (MLE) over the distribution π_θ, leading to the following objective:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim D}\left[\log \sigma\left(\beta \log \frac{\pi_\theta(y_w|x)\pi_{\text{ref}}(y_l|x)}{\pi_\theta(y_l|x)\pi_{\text{ref}}(y_w|x)}\right)\right] \tag{2.3}$$

where β is a regularization term controlling the KL-constraint strength. In this case, the implicit reward function is $r(x, y) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$. Rafailov et al. [9] further showed that all reward models based on the Plackett-Luce distribution [96, 97], including Bradley-Terry, can be expressed in this form. However the absence of an explicit reward model in DPO limits its ability to sample preference pairs from the optimal policy. Zhao et al. [98] have investigated augmenting preference data using a trained SFT policy and Liu et al. [99] have done the same with a refined SFT policy with rejection sampling, enabling policy learning from data generated by the optimal policy.

[2] A curious reader can check this blog: https://iclr-blog-track.github.io/2022/03/25/ppo-implementation-details/

Preference Underfitting in Offline Methods With the growing focus on offline alignment, recent work has proposed various solutions to *preference underfitting*—a phenomena where the LLM underfits the true preference distribution due to infinite rewards [100]. While the benefits of offline vs. online learning are well-known, such underfitting often has practical consequences. Multiple works [48, 49, 100, 101] have shown evidence that assuming human preferences admit a Bradley-Terry formulation, certain pathologies of offline objective functions like DPO's can lead policy degeneracy issues, where the LLM assigns higher probability mass to tokens not present in the training data while reducing its likelihood of generating "good" responses—both tendencies that go against the goals of standard methods like DPO. Multiple directions have been proposed to resolve this issue. These include learning from a confidence set of rewards [48], adopting a general preference model [49,50], or applying a straightforward regularization of the original DPO objective [101]. All these approaches rely on a reference model, not only for expressing the optimal policy as the analytical solution to the minimum relative entropy problem [102,103], but also to stabilize training by constraining the policy distribution close to the reference model. Munos et al. [50] suggest this constraint (policy divergence w.r.t. a reference model), while theoretically sound, can make policies prioritize high rewards over truly learning human preferences. As such, certain works avoid this constraint and focus on lightweight "reference model-free" solutions using a careful construction of DPO-inspired Bradley-Terry (BT)-based implicit rewards—such as logit-based [104], length-normalized [105], or un-normalized implicit rewards [106]—or relative preference label strength [107]. Similarly, efforts to learn policies from general preference models aim to reduce dependence on the sampling distribution, a key limitation of BT models. [50], [108], and [109] use game-theoretic approaches for robustness, while [110] propose a Chain-of-Thought (CoT) policy with pairwise conditioning to improve alignment. However, these methods often suffer from sample inefficiency due to iterative sampling from approximate geometric mixtures

of policies [108], and while they mitigate dependence on the sampling distribution, they do not model non-deterministic preferences in policy training.

Exploring Preference Optimization Objectives Various other preference optimization objectives have been proposed in the literature. Ranking objectives facilitate comparing multiple instances beyond just pairs [111–114]. [104] and [115] investigate objectives that do not require a reference model. [116] suggests a method that jointly optimizes instructions and responses, demonstrating notable improvements over DPO. [117] focuses on post-training extrapolation between the (SFT) model and the aligned model to enhance overall performance. **e-DPO** [48] presents, in contrast to MLE-based DPO and given potentially unlimited unlabeled samples, $(x, y_1, y_2) \sim \rho$ a simple "distillation" loss that explicitly regresses the implicit reward difference in DPO onto the true pointwise reward estimates provided by an explicit reward model or a suite of reward models[3], as follows:

$$\mathcal{L}_{\text{distill}}(r^*, \pi_\theta; \rho) = \mathbb{E}_{\rho(x,y_1,y_2)} \left[\left(r^*(x, y_1) - r^*(x, y_2) - \beta \log \frac{\pi_\theta(y_1 \mid x)(y_2 \mid x)}{\pi_\theta(y_2 \mid x)(y_1 \mid x)} \right)^2 \right]. \quad (2.4)$$

Iterative sampling algorithms like rejection sampling are compute-intensive, and in this work, we focus exclusively on offline settings—consistent with the goal that we want to learn the broad range of preference signals that are often present in precollected preference data. As such, in Chapter 3, we develop methods to adaptively learn from offline preferences data while decoupling our learning from a reference model-based KL-divergence constraint, and deriving the objective of DRDO from knowledge distillation [118] and focal-loss literature [119,120] with a novel contrastive log-unlikelihood objective that learns preferences adaptively.

[3]Notably [48] propose two versions of this reward distillation loss, one with a single reward model, the other with an ensemble of reward models. For simplicity of analysis, we only show the former here. For all our experiments with e-DPO baseline in Chapter 3, we use the ensembled version of this loss.

2.2 Misalignment of Rewards and Preferences

This section formalizes non-deterministic preferences and situates them within prior work. We begin by showing, in a toy bandit setting (Section 2.2.1), how non-deterministic preference pairs can cause reward-optimal and preference-optimal policies to diverge. In particular, when preferences are modeled using a Bradley–Terry (BT) formulation [50], preferences can be expressed as latent rewards; however, optimizing those rewards (as discussed in Section 2.1) may produce policies that are inconsistent with observed preferences whenever even a small number of non-deterministic pairs are present. This bandit construction provides a concrete illustration of the reward–preference misalignment phenomenon that reappears later in more realistic settings.

We then extend this analysis in Section 2.2.2 to realistic settings, examining how a non-trivial fraction of non-deterministic preference pairs in widely used offline datasets can limit standard alignment methods. We provide new theoretical results characterizing failure modes in commonly used approaches such as DPO [9] (Proposition 1, Lemma 1) and uncertainty-aware variants such as e-DPO [48] (Lemma 2) when trained on datasets containing such signals. Related empirical studies have documented these instabilities at scale [48, 101, 121]. Within our bandit example we briefly illustrate the consequences; a full treatment appears in Chapter 3.

2.2.1 Non-Deterministic Human Preferences in Bandit Settings

Assumptions and Setup Throughout this section, we consider a finite action space $\mathcal{A}$ (e.g., a discrete set of LLM responses) and a stochastic preference model $\mathcal{P} : \mathcal{A} \times \mathcal{A} \to [0, 1]$ representing the probability that a human evaluator sampled from distribution $\mathcal{H}$ prefers action y over y'. We assume preferences are obtained from a stationary distribution (i.e., the distribution over human annotators does not change over time) and that pairwise comparisons are conditionally independent given the actions. For concreteness, we illustrate

key concepts using a bandit setting with $|\mathcal{A}| = 3$, though our observations extend to larger action spaces and sequential decision-making.

Deterministic vs. Non-Deterministic Preferences Not all preference pairs are equally informative [122]. Some comparisons yield clear consensus (e.g., a well-written response is strongly preferred over a grammatically incorrect one), while others are ambiguous (e.g., two equally good but stylistically different responses). We formalize this distinction:

Definition 1 (Deterministic and Non-Deterministic Preference Pairs)**.** *Fix a threshold $\epsilon > 0$. A preference pair (y, y') is:*

- ***Deterministic** if $|\mathcal{P}(y \succ y') - 1/2| > \epsilon$, indicating a clear preference direction favoring either y or y'.*
- ***Non-Deterministic** if $|\mathcal{P}(y \succ y') - 1/2| \leq \epsilon$, indicating ambiguous, tie-like, or conflicting preferences.*

A preference model $\mathcal{P}$ is fully deterministic *if all distinct pairs (y, y') with $y \neq y'$ are deterministic.*

Intuitively, deterministic pairs provide strong signal about relative quality: if $\mathcal{P}(y \succ y') = 0.9$, we can confidently infer that y is superior to y'. Non-deterministic pairs, conversely, indicate that y and y' are roughly equivalent in quality, or that human preferences are genuinely divided. In the extreme case where $\mathcal{P}(y \succ y') = 1/2$, there is no consensus whatsoever. The choice of ϵ depends on the application. In high-stakes settings (e.g., safety-critical LLM outputs), even small deviations from $1/2$ might be considered deterministic (ϵ small). In creative domains where subjective preferences dominate, a larger ϵ might be appropriate to account for genuine diversity in human preference. Figure 2.1 illustrates deterministic and non-deterministic preference models in our bandit setting with three distinct actions, y_1, y_2 and y_3 assuming $\epsilon = 0$.

$\mathcal{P}_1(y \succ y')$	$y = y_1$	$y = y_2$	$y = y_3$
$y' = y_1$	$1/2$	$9/10$	$2/3$
$y' = y_2$	$1/10$	$1/2$	$2/11$
$y' = y_3$	$1/3$	$9/11$	$1/2$

(a) Strict preference matrix $\mathcal{P}_1$

$\mathcal{P}_2(y \succ y')$	$y = y_1$	$y = y_2$	$y = y_3$
$y' = y_1$	$1/2$	$9/10$	$2/3$
$y' = y_2$	$1/10$	$1/2$	$1/2$
$y' = y_3$	$1/3$	$1/2$	$1/2$

(b) Perturbed matrix $\mathcal{P}_2$ with non-deterministic pair

Figure 2.1: Preference matrices illustrating how a single non-deterministic pair changes representability under Bradley–Terry.

Preference Models Human preferences over actions are naturally represented through pairwise comparisons [87, 96] via the Bradley-Terry-Luce (BTL) models. Rather than assuming access to a ground-truth reward function, we work directly with a probabilistic preference model from the literature in preference-based RL (PbRL) [122] that captures the inherent uncertainty in human judgments.

Definition 2 (Preference Model)**.** *A* preference model $\mathcal{P} : \mathcal{A} \times \mathcal{A} \to [0, 1]$ *assigns to each ordered pair of actions* $(y, y') \in \mathcal{A} \times \mathcal{A}$ *the probability* $\mathcal{P}(y \succ y')$ *that action* y *is preferred over action* y'*. The model satisfies:*

1. ***Normalization****:* $\mathcal{P}(y \succ y') + \mathcal{P}(y' \succ y) = 1$ *for all* $y, y' \in \mathcal{A}$
2. ***Self-consistency****:* $\mathcal{P}(y \succ y) = 1/2$ *for all* $y \in \mathcal{A}$

The normalization condition ensures that preferences are coherent: if y is preferred over y' with probability p, then y' is preferred over y with probability $1 - p$. The self-consistency condition reflects that no action is strictly preferred to itself. Note that $\mathcal{P}$ represents the *distribution* of preferences rather than deterministic rankings—different human annotators may provide different judgments for the same pair (y, y'). Importantly, from the perspective of LLM preference alignment, the preference model $\mathcal{P}(y \succ y'|x)$ represents the probability that a randomly sampled human evaluator from some distribution $\mathcal{H}$ over annotators prefers response y over y' given prompt x. In practice, $\mathcal{P}(y \succ y')$ is estimated from empirical data [12,122]: if n annotators compare y and y', and k of them prefer y, then

$\mathcal{P}(y \succ y') \approx k/n$. This empirical estimate captures both genuine disagreement among annotators (some truly prefer y, others prefer y') and measurement noise.

The Bradley-Terry Model and Reward-Based Representations

Most modern preference learning methods, including DPO [9] and RLHF [72], implicitly or explicitly assume that preferences can be explained by an underlying reward function. The Bradley-Terry (BT) model formalizes this assumption:

Definition 3 (Bradley-Terry Model)**.** *A preference model $\mathcal{P}$* admits a Bradley-Terry representation *if there exists a reward function $R : \mathcal{A} \to \mathbb{R}$ such that for all $y, y' \in \mathcal{A}$:*

$$\mathcal{P}(y \succ y') = \frac{\exp(R(y))}{\exp(R(y)) + \exp(R(y'))} = \sigma(R(y) - R(y')) \tag{2.5}$$

where $\sigma(x) = 1/(1 + e^{-x})$ is the logistic sigmoid function.

The BT model posits that each action y has an intrinsic "quality" score $R(y)$ [123], and the probability that y is preferred over y' depends only on the difference $R(y) - R(y')$. This is an elegant and parsimonious model: instead of storing $|\mathcal{A}|^2$ preference probabilities, we need only $|\mathcal{A}|$ reward values (up to a constant shift, since only differences matter).

Bradley–Terry Incompatibility with Non-Deterministic Preferences However, non-deterministic preferences as in Definition 1 challenge the robustness of the BT model. If $\mathcal{P}(y \succ y') = 1/2$ for distinct actions $y \neq y'$, then necessarily $\exp(R(y)) = \exp(R(y'))$, implying $R(y) = R(y')$. Consequently, any non-deterministic preference constraint of this form forces equality of rewards. However, when such a constraint coexists with other preference relations that require $R(y) \neq R(y')$, the preference model becomes incompatible with any BT representation. For example, in the bandit setting of Figure 2.1 (b), the constraint $\mathcal{P}_2(y_2 \succ y_3) = 1/2$ enforces $R(y_2) = R(y_3)$, while $\mathcal{P}_2(y_2 \succ y_1) = 9/10$ implies $R(y_2) - R(y_1) = \log(9)$ and $\mathcal{P}_2(y_3 \succ y_1) = 2/3$ implies $R(y_3) - R(y_1) = \log(2)$, which cannot be simultaneously satisfied. As a result, the BT model must systematically violate

at least one preference constraint, leading to misspecified reward functions whose learned values depend sensitively on the data distribution used during training. Such misspecification undermines downstream policy optimization and is particularly problematic in LLM alignment, where small reward modeling errors can induce unstable or unsafe behaviors [81].

Policy Optimization Under Preferences Having defined both deterministic and non-deterministic preference models and the Bradley-Terry model, we now formalize two distinct optimization objectives: maximizing expected reward versus maximizing preference probability. These objectives can diverge significantly when preferences are non-deterministic. We will use these quantities of interest in Lemma 3.

Definition 4 (Expected Reward)**.** *The* expected reward *of a policy $\pi \in \Delta(\mathcal{A})$ under a reward function $R : \mathcal{A} \to \mathbb{R}$ is:*

$$\mathbb{E}_{y\sim\pi}[R(y)] = \sum_{y\in\mathcal{A}} \pi(y)R(y) \tag{2.6}$$

*This is a **linear** function of the policy that aggregates individual action rewards weighted by selection probabilities.*

Expected reward directly measures policy quality under the assumption that actions have intrinsic values $R(y)$. Optimizing $\mathbb{E}_{y\sim\pi}[R(y)]$ prioritizes actions with high individual rewards, independent of how they compare to other actions.

Definition 5 (Preference Probability)**.** *The* preference probability *that policy π is preferred over policy π' under preference model $\mathcal{P}$ is:*

$$\mathcal{P}(\pi \succ \pi') = \sum_{y\in\mathcal{A}} \sum_{y'\in\mathcal{A}} \pi(y)\pi'(y')\mathcal{P}(y \succ y') \tag{2.7}$$

*This is a **bilinear** (non-linear) function of the policies that captures pairwise preference interactions between sampled actions.*

Preference probability measures how often samples from π are preferred over samples from π' in head-to-head comparisons [49, 50, 124]. Unlike expected reward, this objective is inherently comparative: it depends on the *matchup* [125] between actions from π and π', not just their individual qualities. An action with moderate reward but strong pairwise preferences (i.e., it beats many other actions) may be favored by the preference objective even if it does not maximize expected reward. The linear nature of $\mathbb{E}_{y\sim\pi}[R(y)]$ means that improving the policy simply requires shifting probability mass toward high-$R(y)$ actions. In contrast, the bilinear form of $\mathcal{P}(\pi \succ \pi')$ creates strategic interactions: the optimal π depends on the distribution π' (e.g., the reference policy or opponent strategies). This structural difference can cause $\arg\max_\pi \mathbb{E}[R(y)]$ and $\arg\max_\pi \mathcal{P}(\pi \succ \pi')$ to diverge, particularly under non-deterministic preferences. We show this formally in Lemma 3 in Chapter 3. In Proposition 4, we provide some analysis of this dependence of the BT model rewards on the sampling distribution, for a more general class of perturbed distributions compared to [50].

2.2.2 Non-Deterministic Human Preferences in Offline Preference Data

The bandit construction in Section 2.2.1 demonstrates that, under non-deterministic preferences, reward-optimal and preference-optimal policies can diverge. Importantly, this phenomenon is not unique to synthetic examples. Modern offline alignment datasets contain substantial fractions of weak, ambiguous, or conflicting preference judgments, meaning that the same structural misalignment arises in practice. This effect is empirically illustrated in Figure 3.2 of Chapter 3, which analyzes preference strength distributions in widely used alignment benchmarks such as UltraFeedback and Reddit TL;DR.

Following Bradley and Terry [87], Rafailov et al. [9] and other preference optimization frameworks posit that the relative preference of one outcome over another is governed by the true reward differences, expressed as $p^*(y_1 \succ y_2) = \sigma(r_1 - r_2)$, where p^* is the true preference distribution. Generally, in the RLHF framework, the true preference distribution

is typically inferred from a dataset of human preferences, using a reward model r that subsequently guides the optimal policy learning. More importantly, to estimate the rewards and thereby the optimal policy parameters, the critical *reward modeling* stage involves human annotators choosing between pairs of candidate answers (y_1, y_2), indicating their preferences.[4] As such, typical alignment methods assume that $p(y_w \succ y_l|x)$ (the human annotations of preference) is equivalent to $p^*(y_1 \succ y_2|x)$ or any ranking or choice thereby established with the human decisions. However, prospect theory and empirical studies in rational choice theory suggest that human preferences are often stochastic, intransitive, and can fluctuate across time and contexts [126–128].

Existing direct alignment methods, such as DPO-based supervised alignment, assume access to deterministic preference labels, disregarding the inherent variability in human judgments, *even when popular preference datasets are inherently annotated with such variability, noise, or "non-deterministic preferences" given their provenance in human labeling.* More importantly, such implicit trust in the preference data by DPO-like algorithms, without explicit instance-level penalization on the loss, can cause policies that are trained to deviate from true intentions of human preference learning. We show this formally in Lemma 1 and Lemma 2. In many preference datasets, a significant proportion of preference pair annotations display low human confidence, or receive similar scores from a third-party reward assignment models (e.g., GPT-4) despite being textually different, indicating that the two responses are likely semantically similar or similar in intent, content or quality. Note that we consider non-deterministic preference samples to be distinct from noise present as flipped labels [129,130], which is typically resolved using label-smoothing based heuristics, data exclusion or prior knowledge of noise coefficients in the data in preference learning. Such discrepancies in preference signals can similarly derail reward learning and limit reward models from reaching a consensus, even with majority voting with reward ensembles [130]. In data-driven preference learning from offline datasets, these cases

[4]This framework can be extended to rank multiple responses using the Plackett-Luce model.

reflect the stochastic nature of human choices, and challenge the assumption of stable, deterministic preferences in alignment frameworks.

We now formally define such "noisy" or non-deterministic preference labels in offline finite preference data regimes and offer some insights into limitations of current approaches like DPO and e-DPO. For the sake of analysis, we still consider a Bradley-Terry-based modeling to represent such preference signals. All proofs are deferred to the appendix.

Assumption 1. *Let $\mathcal{D}_{pref} = \{(x^{(i)}, y_w^{(i)}, y_l^{(i)})\}_{i=1}^N$ be an offline dataset of pairwise preferences with sufficient coverage, where each $x^{(i)}$ is a prompt, and $y_w^{(i)}$ and $y_l^{(i)}$ are the corresponding preferred and dispreferred responses, respectively. Let $r^*(x, y) \in \mathbb{R}$ be an underlying true reward function that is deterministic[5] and finite everywhere. Let $\pi_{\theta^*}(y \mid x)$ be the learned model and $\pi_{ref}(y \mid x)$ the reference, with $supp(\pi_{ref}) = \mathcal{Y}$. Assume $supp(\rho) = supp(\mu) \times \mathcal{Y} \times \mathcal{Y}$, where $\mathcal{Y}$ is the space of all responses, ρ is the data distribution, and μ is the prompt or context distribution.*

Proposition 1 (Non-Deterministic Preferences)**.** *For the subset of non-deterministic preferences defined as $\mathcal{D}_{nd} = \{(x, y, y') \mid P(y \succ y' \mid x) \approx 1/2\}$ and assuming antisymmetric preferences [50], the Bradley-Terry model implies that the reward difference $\Delta r = r(x, y) - r(x, y') \approx 0$ for all $(x, y, y') \in \mathcal{D}_{nd}$. Consequently, the expected reward difference over this subset is also $\mathbb{E}_{(x,y,y') \sim \mathcal{D}_{nd}}[\Delta r] \approx 0$. See Section A.1 for a complete derivation.*[6]

Lemma 1. *Under Proposition 1, the following hold:*

(a) The DPO implicit reward difference in its objective satisfies

$$\frac{\pi_{\theta^*}(y)\, \pi_{\text{ref}}(y')}{\pi_{\theta^*}(y')\, \pi_{\text{ref}}(y)} \;\to\; 1,$$

which leads to the policy empirically underfitting the preference distribution.

[5]Deterministic and non-deterministic preferences are only defined on the true preference distribution p^* and should *not* be confused with the empirical probabilities or confidence assigned by the policy. The use of "deterministic" here is simply to imply that the true reward function $r^*(x, y)$ is finite and scalar.

[6]For clarity, we note that this Proposition 2 is distinct from Proposition 2 from [50] that is referenced above.

(b) For $|\mathcal{D}_{nd}| \ll N$, where N is finite, if DPO estimates that $p^*(y \succ y') = 1$, then

$$\frac{\pi_{\theta^*}(y)\,\pi_{\text{ref}}(y')}{\pi_{\theta^*}(y')\,\pi_{\text{ref}}(y)} \;\to\; \infty.$$

(c) For all minimizers π_{θ^*} of the DPO objective (Eq. 7 in [9]), it follows that

$$\pi_{\theta^*}(y_l) \;\to\; 0 \quad \textit{and} \quad \pi_{\theta^*}(\mathcal{C}(y_l)^c) \;\to\; 1,$$

where $\mathcal{C}(y_l)^c$ denotes the complement of the set of dispreferred responses $y_l^{(i)}$, $\forall i \in \mathbb{N}$.

Given non-trivial occurrences of non-deterministic preference pairs in typical preference learning datasets, a consequence of Lemma 1 is that DPO's learned optimal policy can effectively assign non-zero or even very high probabilities to tokens that never appear as preferred in the training data, causing substantial policy degeneracy. Moreover, as noted and shown empirically in previous work [101, 131], DPO effectively underfits the preference distribution because its empirical preference probabilities are only estimates of the true preference probabilities, especially when $p^*(y \succ y') \in \{0, 1\}$. A noteworthy implication of Lemma 1 is that this weak regularization effect of DPO can theoretically assign very high probabilities to the complement set of dispreferred tokens that never appear in the training data *at all*, especially when $|N_{\text{nd}}| \ll N$ for finite data regimes. In realistic settings where non-deterministic preferences constitute a non-trivial proportion of data, Lemma 1 additionally implies that DPO leads to unstable updates and inconsistent policy behavior, where the gradient update is effectively cancelled out for these samples since the log probabilities of both the winning and the losing responses are roughly equal ($\Delta r \approx 0$), so the scaled weighting factor (sigmoid of implicit reward differences) does not contribute as much as when $p^*(y \succ y') \in \{0, 1\}$. As stated in Section 3.2, with DPO, π_{θ^*} not only sees less of this type of preference but also fails to adequately regularize when it does.

A solution to the above limitations of DPO within offline settings is to recast its MLE optimization objective into a *regression* task, where the choice of regression target can be the preference labels themselves (as in IPO; [131]) or reward differences (as in e-DPO; [48]). While the former directly utilizes preference labels, regressing the log-likelihood ratio π_{ratio} to the KL-β parameter as defined in Eq. 7 in [9], the latter extends IPO by regressing against the difference in true rewards $r^*(x, y)$, independent of explicit preference labels and acting as a strict generalization of the IPO framework. Notably, both these methods ensure that the resulting policy induces a valid Bradley-Terry preference distribution $p^*(y_1 \succ y_2 \mid x) > 0$, $\forall x, y_1, y_2 \in \mu \times \mathcal{Y} \times \mathcal{Y}$.

However, these approaches have inherent limitations. IPO regresses the log-likelihood difference on a Bernoulli-distributed preference label, failing to capture nuanced strength in relative preferences. Conversely, e-DPO eliminates preference label dependence but sacrifices the granular signals available in preference data, instead over-relying on the quality of reward ensembles, which may still lead to over-optimization [91].[7] Consider the following lemma that derives from Assumption 1 and Proposition 1:

Lemma 2. *Under Proposition 1 and in the spirit of [48]'s argument, using e-DPO alignment over non-deterministic preference pairs leads to* $\frac{\pi_{\theta^*}(y)\pi_{\text{ref}}(y')}{\pi_{\theta^*}(y')\pi_{\text{ref}}(y)} \to \infty$ *for* $(y, y') \in \mathcal{D}_{nd}$ *where* $y = y_w^{(i)}$ *and* $y' = y_l^{(i)}$, $\forall i \in \mathbb{N}$. *Then, for all minimizers* π_{θ^*} *of the e-DPO objective:*

$$\mathcal{L}_{\text{distill}}(r^*, \pi_\theta; \rho) = \mathbb{E}_{\rho(x,y_1,y_2)}\left[\left(r^*(x, y_1) - r^*(x, y_2) - \beta \log \frac{\pi_\theta(y_1 \mid x)(y_2 \mid x)}{\pi_\theta(y_2 \mid x)(y_1 \mid x)}\right)^2\right], \quad (2.8)$$

it follows that $\pi_{\theta^*}(\mathcal{C}(y_l)^c) \to 1$ *with* $0 < \pi_{\theta^*}(y_w^{(i)}) \leq 1$, $\forall i \in \mathbb{N}$, *where* $\mathcal{C}(y_l)^c$ *denotes the complement of the set of dispreferred responses* $y_l^{(i)}$, $\forall i \in \mathbb{N}$. *See Section A.1 for a complete derivation.*

[7] Furthermore, the use of reward ensembles in e-DPO introduces significant computational overhead, potentially limiting its broader applicability due to increased resource requirements.

These analyses suggest that modeling relative preference strengths during the *policy learning stage*—particularly at the extrema of the preference distribution—is problematic only when using a DPO-style MLE objective that maximizes implicit reward differences via likelihood ratios. In contrast, the MLE formulation used during the *reward modeling stage* does not suffer from this limitation, since rewards are estimated as scalar quantities and do not appear inside a likelihood ratio, provided sufficient coverage in the preference data. Since both stages rely on the same finite preference dataset—reflecting a range of human preference strengths—this observation motivates combining them by explicitly distilling rewards into the policy learning stage. Assuming access to an oracle or the true reward function $r^*(x, y)$, one can resolve these objective-induced limitations by directly distilling estimated rewards into the policy model. In Chapter 3, we leverage this insight to derive an offline alignment objective in which the LLM simultaneously learns to estimate rewards and align its policy, thereby avoiding the structural pathologies associated with DPO-style likelihood maximization.

2.3 Collaborative Problem Solving with LLMs

In Section 2.1 and Section 2.2, we discussed the background of offline alignment methods and some of their limitations, both in general and in the presence of non-deterministic preferences. Most of this discussion assumed contextual bandit settings (i.e., single-turn optimization, the dominant paradigm for methods like DPO [9] and IPO [49]) for defining the problem of aligning LLMs with human preference. In this setting an LLM generates a response, receives immediate feedback, and the interaction terminates. These assumptions are well suited for training models to perform effectively on simpler tasks such as question answering and summarization.

In contrast, more complex phenomena such as collaboration [65] require drawn-out interactions or sustained conversations, where these assumptions no longer hold. In particular, single-turn preference alignment breaks down in multi-party collaborative settings

because agent actions shape the future belief states of other participants, leading to belief misalignment that requires belief-aware modeling. Such dynamics cannot be captured by standard bandit formulations or conventional MDP abstractions [45]. Moreover, modeling language or dialogue generation as a standard MDP is insufficient in collaborative settings, as it fails to represent the evolution of beliefs and the formation of common ground that characterize realistic multi-agent collaboration. In this section, we formalize this intuition by introducing *friction* as a generative paradigm for belief-aware LLM alignment in collaborative settings, and define the core conceptual notions used throughout this dissertation, including common ground, friction interventions, and frictive states.

The Belief Misalignment Problem in Small-Group Collaboration Collaborative tasks introduce a fundamental challenge absent from single-turn or even dyadic (user-assistant) settings: *belief misalignment across multiple participants*. For example, when a student incorrectly believes proposition p while their peer has evidence against p, this creates a *frictive state* [27]—a moment where misaligned beliefs about task-relevant propositions threaten to derail progress. Unlike simple disagreements, frictive states require an intervention to help participants *rebuild common ground* [25] by prompting reflection rather than asserting correctness. Critically, such interventions are *mediated*: the intervening agent does not directly affect the task outcome but influences it *through* the responses of other collaborators, whose beliefs and subsequent actions determine whether the impasse is resolved. This mediated structure—where Agent A's action modifies Agent B's action before the environment responds—cannot be captured by standard MDPs, which assume actions directly affect state transitions.

Why Existing Methods Fall Short Multi-turn alignment methods address some limitations of single-turn approaches but remain inadequate for small-group collaboration. Online methods like CollabLLM [45] leverage causal modeling for collaborative coding and document editing, while game-theoretic approaches [50, 109] use Nash equilibria to

address data-dependence in Bradley-Terry models. However, these methods focus on *dyadic* user-assistant interaction and assume symmetric agents, making them ill-suited for heterogeneous small-group settings where an intervention agent must coordinate beliefs among multiple collaborators. Roleplay-based approaches [60, 61] assign static roles via prompts and have been applied successfully to software engineering [60], negotiation [132], and medical applications [133], but remain fragile inference-time strategies [63] that lack principled credit assignment [64] or mechanisms to model how interventions propagate through collaborator beliefs to affect future rewards [124]. Moreover, multi-agent reinforcement learning (MARL) [56] frameworks remain limited in LLM applications [57, 58] due to instability from asymmetric rollouts and the prohibitive cost of training heterogeneous agents with policy-gradient methods.

Friction as a Generative Paradigm for Belief-Aware Alignment To address these gaps, we introduce *friction interventions* as a generative paradigm for training intervention agents—here, AI systems that intervene in small-group collaborative task dialogues. Unlike standard alignment objectives that maximize immediate preference satisfaction, friction-oriented alignment optimizes for *belief convergence*—measured by growth in common ground [25]—by training agents to introduce strategically timed utterances that prompt participants to reevaluate assumptions. This requires moving beyond standard MDP formulations to *Modified-Action MDPs (MAMDPs)* [66], which explicitly model the mediating role of collaborators: the intervention agent's action a_f is transformed by the collaborator's response $a_c(a_f)$ before affecting the environment, creating a chain $s \xrightarrow{a_f} s' \xrightarrow{a_c(a_f)} s''$ where credit assignment must account for both the intervention and its mediated effect. The remainder of this section formalizes these concepts through definitions of frictive states, friction interventions, and common ground, then introduces the MAMDP framework used throughout Chapters 4, 5, and 6.

2.3.1 Friction as a Generative Paradigm

We now formalize the core concepts underlying friction-oriented alignment. Intuitively, *frictive states* are moments where misaligned beliefs threaten collaborative progress, and *friction interventions* are utterances that resolve such states by prompting participants to reevaluate assumptions. Success is measured by growth in *common ground* [25]—the set of task-relevant propositions that participants mutually accept and treat as shared for coordination. Below we provide formal definitions, followed by the MAMDP framework that models the mediated structure of interventions.

Definition 6 (Common Ground)**.** *Following [25, 26], we define* common ground (CG) *as the set of task-relevant propositions that are mutually accepted by all participants and publicly treated as shared for the purpose of coordinating action. In collaborative task settings, CG is restricted to propositions whose truth values directly affect task progress or solution quality. A proposition belongs to the CG at dialogue turn t if all participants exhibit convergent beliefs toward it, and if this convergence is evident from their communicative behavior. For practical evaluation of collaborative dialogues in Chapter 5 and Chapter 6, we approximate common ground by tracking the set of task-relevant propositions on which participants exhibit convergent stances over time, and computing the expected size of this set across dialogue turns.*

Definition 7 (Frictive State)**.** *A frictive state arises during a collaborative task when different interlocutors have contradictory beliefs about a task-relevant proposition (i.e., one believes p and another sees evidence against p). This can be realized as a formal model of agent beliefs in an evidence-based dynamic epistemic logic [134, 135], or a natural language description thereof, as we use. Different evidence leads to different predictions of future trajectories [136]. Thus frictive states, though sparse in dialogues, can critically delay or preclude success in a collaboration due to unresolved misunderstandings. The occurrence of a frictive state may not guarantee task failure, as the relevant propositions may be trivial to actual task completion. Therefore, in a functionally frictive state, the lack of common ground impedes progress on the task, or presents a significant risk of failure unless it is resolved.*

Definition 8 (Friction Intervention)**.** *Friction can indicate an impasse (the frictive state), but can also be used to resolve it, through a friction intervention that inserts into the dialogue indirect prompting to the participants to reevaluate their beliefs and incorrect assumptions or positions in light of available evidence [137], rather than accepting possibly erroneous presuppositions inherent in the dialogue. Importantly, a frictive intervention may be non-contradictory to the individual beliefs on display (i.e., neither asserting p nor $\neg p$), but slows down the dialogue for reflection and deliberation, such as the probing utterances in [8] and [138]. In the context of this dissertation, the friction agent constitutes a language model aligned toward the capacity to make frictive interventions.*[8] *An ideal friction agent does not intervene arbitrarily, which would cause distraction in collaborative tasks, but is conditioned to resolve the lack of common ground between human collaborators.*

Not all frictive states are functionally frictive in an external sense. Some reflect internal friction—such as reasoning processes [139], information retrieval (e.g., RAG [140]), or test-time computation [141]—that need not be externally resolved. Others may be resolved through communication [142] or learned conventions [143] without intervention. A frictive state becomes functionally relevant when its presence threatens collaborative success and is exposed in dialogue, such that an LLM can decode it from interactional signals and condition an appropriate intervention. As such, the above definitions of frictive states and frictional interventions are "operational" definitions of friction in human-AI collaboration. Friction creates opportunity for negotiation of intents toward a common goal, and space for accountability and collaborative reasoning. These moments may result in a net slower interaction, but are critical to eventual task success. The study of friction has broad applicability to fields like discourse studies, team science, and education [144–146], and is something we believe the NLP community would do well to invest effort in. Counter to AI being sold as a speed and efficiency multiplier, our formulation of alignment to

[8]We use π_f to denote the friction agent which generates high-quality interventions, but often refer to it as the "optimal policy" for consistency with RLHF literature.

specialize in friction shifts LLMs from mere responders to being "thought partners," and sets a new standard for dynamic, dialogue-centric environments. Finally, in the context of this dissertation, the terms **friction intervention agent, intervention agent or intervening agent** all refer to the **same** LLM role: a collaborator whose purpose is to introduce strategically timed utterances that encourage participants to rethink, cross-examine, or clarify their assumptions while engaging in a collaborative task.

Field	Content
Dialogue History (x)	P1: i guess if red block red one's ten grams P2: we got red ten P1: seems like red block, blue block might be about the same P3: i would agree yeah so blue block one's ten P3: Alright let's see if we can find a twenty P3: Too heavy so P2: Way too heavy P2: this is a sensitive scale P2: Looks like about twenty P1: that's looking pretty even P3: Alright let's see if we can find a thirty P1: so yellow block one is noticeably heavier than P2: probably yellow block big sucker P1: the purple ish one P1: making sure that purple block didn't have the weight at the bottom P2: it's just stuff written at the bottom that's a so red block, green block's a ten and a twenty right now right that's looking P2: Well P2: red block, blue block, green block, yellow block, purple block're increments of ten i would say that's probably P1: Yeah I think P2: cause purple block's also a twenty let's double check that purple block's not also a twenty P1: yeah it looks a little P2: cause it um just purple block one there P3: is blue block one a twenty P2: ok so purple block's more than twenty but it almost seems like the thirty takes it past but P2: it's so sensitive P2: if red block, blue block, green block, yellow block, purple block're only in increments of ten purple block has to be
Frictive state (ϕ)	P2 initially identified the red and blue blocks as both 10 grams and has speculated about the green at 20 grams, but is uncertain about the actual weights of the yellow and purple blocks.
Rationale	P2 suggests that the red, green, and yellow blocks are all in increments of ten. Encourage a double-check.
Preferred Friction (f_w)	Since the purple block seems heavier and we're unsure about its exact weight, should we double-check the increments of ten assumption? Maybe the purple block doesn't fit this pattern.
Dispreferred Friction (f_l)	You know, the purple block being heavier might actually mean the blocks aren't increasing consistently at all. What if the increments are random, like 10, 15, or 25 grams, and we're forcing a pattern that isn't there?

Table 2.1: A transcribed, sparse collaborative dialogue from the Weights Task Dataset [1] with frictive states and friction interventions. Preferred and dispreferred friction interventions are shown at the bottom. Positive friction interventions prompt participants to reevaluate their assumptions with frictive states (evolution of beliefs and rationales) providing indirect hints and directions. Here, P2's uncertainty about the green block and assumption of weight increment by 10g is addressed by the positive friction. In contrast, the dispreferred intervention introduces randomness, instigating the group to abandon structured reasoning.

Illustrative Example of Friction Table 2.1 shows a sample friction annotation and training sample from the Weights Task Dataset [1], consisting of the dialogue history x, GPT-4o-generated frictive state ϕ, rationale, and preferred and dispreferred friction interventions f_w and f_l. Because the WTD is a multimodal dataset, the transcriptions we use are enriched using *dense paraphrasing* [147], a textual enrichment technique that uses the multimodal channels to decontextualize referents and in this case transform contextually-dependent phrasings such as demonstratives to explicit denotations of the content. For example, under dense paraphrasing, "seems like *these* might be about the same" while the speaker in the video is pointing to the red and blue blocks becomes "seems like *red block, blue block* might be about the same." The dense paraphrased utterances are included as part of the publicly-available WTD [1,7].

Generative Friction vs. Friction as UI Scaffolds We must clarify what we mean by "friction" in collaborative AI systems, as the term has been used in distinct ways across the literature. Recent work has explored friction as UI scaffolding [2]—deliberate design elements like extra clicks, delays, or stop-signs that increase the effort required to access LLM outputs. These interventions aim to reduce algorithm appreciation and over-reliance by imposing barriers between users and AI-generated content. While valuable for encouraging thoughtful LLM use, this approach treats friction as a post-hoc remediation: the LLM has already been trained to provide immediate, frictionless responses, and external scaffolding is added afterward to counteract this behavior. Our work treats friction as a *learned generative capability*, not an interface constraint. We treat friction as a first-class training objective rather than an after-the-fact mitigation strategy. In our formulation, friction is not something imposed on the LLM-user interaction through interface design, but rather something the LLM learns to generate as part of its collaborative capability. Specifically, we define friction interventions as utterances that prompt collaborators to reflect on and revise beliefs when the system detects misalignment on task-relevant propositions—what

we term frictive states. Rather than preventing access to LLM outputs, our friction agents actively participate in dialogue by generating interventions conditioned on the collaborative state. This distinction has important implications. UI-based friction operates at the interaction layer, agnostic to the underlying beliefs or task state—a stop-sign appears regardless of whether slowing down would actually benefit the collaboration. In contrast, our approach operates at the generation level: the LLM learns to recognize when belief misalignments exist and produces language that addresses those specific misalignments. This requires modeling the collaborative state and treating friction not as a barrier to overcome, but as a deliberative move in a sequential decision-making process. Moreover, UI scaffolding addresses symptoms (over-reliance) rather than causes (the LLM's inability to detect when immediate answers would be counterproductive). Our training-time approach addresses the root issue: teaching LLMs to identify situations where deliberation serves the collaboration better than immediate completion. This enables friction to be adaptive, context-dependent, and semantically grounded in the actual beliefs at play—capabilities impossible to achieve through interface design alone. Table 2.2 provides details on these two distinct paradigms from multiple dimensions.

Friction as Options [148] Our friction framework can be understood through the lens of the options framework from hierarchical reinforcement learning [148, 149]. In this view, friction interventions are "macro-actions" or temporally-extended actions that compose complete utterances, where the most granular action level remains the LLM's token. An option is formally defined by three components: an initiation set (states where the option can begin), an intra-option policy (behavior during execution), and a termination condition (when the option ends). We now characterize friction interventions according to these three components.

The **initiation set** is determined by frictive states—friction interventions are generated only when the dialogue reaches a state of belief misalignment. Rather than training

Dimension	UI Scaffolding Friction [2]	Generative Friction (Our Work)
Intervention Point	Post-deployment (interface layer)	Training-time (model capability)
Mechanism	Barriers (clicks, delays, stop-signs)	Generated language (probing questions, clarifications)
Context Awareness	Agnostic to task/belief state	Conditioned on frictive states
Goal	Reduce over-reliance	Resolve belief misalignments
Adaptation	Static (same friction for all contexts)	Dynamic (state-dependent interventions)
Problem Addressed	Symptom (over-reliance behavior)	Cause (lack of collaborative awareness)

Table 2.2: Contrasting paradigms of friction in AI systems. UI scaffolding friction [2] imposes post-deployment barriers through interface design to discourage LLM over-reliance, operating agnostic to the underlying task or belief state. In contrast, our generative friction approach [3] treats friction as a training-time objective, where LLMs learn to produce frictive state-aware interventions that resolve belief misalignments in collaborative dialogue. While both approaches aim to encourage thoughtful LLM use, they differ fundamentally in timing (post-deployment vs. training), mechanism (prohibitive barriers vs. generative language), and problem framing (mitigating symptoms vs. addressing root causes).

a separate classifier to detect such states, we leverage the LLM's in-context learning capability [150] for generating these interventions in Chapter 4. Given high-fidelity frictive state descriptions ϕ (either self-generated by expert LLMs or human-annotated), the model identifies belief misalignments directly from dialogue context and conditions its generation accordingly. If no misalignment is detected, the LLM produces output indicating that no intervention is necessary; when misalignment exists, it generates a friction intervention conditioned on the frictive state ϕ.

Critically, this state-conditioned initiation set addresses a fundamental limitation of fixed-length macro-actions [151]: *contextual appropriateness*. While n-gram macro-actions reduce decision frequency uniformly across all contexts, friction-as-options extends temporal abstraction only when belief misalignment warrants it. This prevents unnecessary abstraction during smooth collaboration—where fine-grained token-level control may be beneficial—and applies coarser temporal scales only when collaborative state complexity demands intervention. In other words, the granularity of decision-making adapts dynami-

cally to the collaborative state, implementing a form of *adaptive temporal abstraction* absent in fixed macro-action schemes for language tasks.

The **policy**, as we formalize in Chapter 5, continues to operate over tokens at the most granular level, but Bellman optimality is preserved at the full utterance level (the complete option) under standard assumptions of Bellman completeness for Q-functions over softmax policies [46]. While token generation follows the standard sequential autoregressive process conditioned on the frictive state, the macro-level decision—whether and what type of intervention to initiate—operates at the utterance level. The **termination condition** is naturally defined by linguistic coherence: the option terminates when the intervention utterance completes, marked by the end-of-sequence (EOS) token. Intuitively, once the EOS token achieves the highest likelihood of being sampled, the termination condition is satisfied. This differs fundamentally from arbitrary n-gram termination used in standard macro-action approaches [151], as it respects the semantic and pragmatic completeness of the intervention rather than imposing fixed temporal boundaries.

2.4 Modified-Action MDPs in Collaboration

Having defined our novel concepts of frictive states (Definition 7) and friction interventions (Definition 8), we now formulate the core problem of modeling the collaborative dynamics in the presence of a friction intervention agent. Specifically, in our collaborative setting, a FRICTION AGENT (π^F) is an LLM aligned to generate strategic friction interventions for resolving frictive states, while collaborator agents (π^C) are LLM instances engaged in the multiparty collaborative dialogue.

We identify a crucial gap in training an optimal friction intervention agent in this scenario: standard Bellman-optimal policies—that preference alignment algorithms seek to retrieve—assume a direct mapping from action to

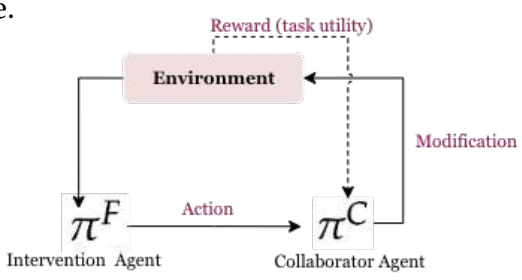


Figure 2.2: Interaction dynamics between the friction intervention agent (π^F) and collaborator agent (π^C) in the collaborative MAMDP. The collaborator modifies or reinterprets interventions before seeking environmental reward.

state change, which breaks down when actions are mediated by other agents. To address this, we adopt the Modified-Action MDP (MAMDP) framework [66], which *explicitly* models how interventions are transformed before influencing the collaborative dialogue.

Formally, an MAMDP consists of a 6-tuple $\mathcal{M}_f = (\mathcal{S}, \mathcal{A}, P_S, P_A, R, \gamma)$, or equivalently, the 5-tuple of a standard MDP with additional parameter P_A. The state space $(s \in) \mathcal{S}$ represents the dialogue history $\mathcal{H}_t$ as token sequences terminating at timestep t, the action space $(a \in) \mathcal{A}$ contains candidate actions (utterances in the dialogue) sampled from an underlying distribution, and the state transition function P_S is deterministic [46][9].

Now assume a FRICTION AGENT π^F_θ (an LLM with parameters θ). Let $P_A(a|s, \pi^F) = \sum_{a' \in A} \pi^F(a'|s) \cdot \pi^C(a|s, a')$ represent the modified probability distribution[10] over collaborator actions, where π^F suggests friction intervention a' in state s, and π^C represents the collaborator's policy for responding with action a given the state and friction suggestion. Following standard notation, the reward $R(s, a)$ is an expected utility, and discount factor $\gamma = 1$. Figure 2.2 shows the dynamics of our collaborative setting within the MAMDP. Unlike standard MDPs [32], the actions (e.g., utterances or questions) of the friction intervention agent π^F does *not* immediately affect the environment for the system to get a reward. These are modified by the collaborator π^C before influencing the collaborative dialogue state. In effect, this constitutes another form of friction: the collaborator mediates the intervention rather than blindly accepting it, a phenomenon commonly seen in real human collaborations. We now consider the following example for illustrative purposes.

[9] By deterministic, we mean that the next state is formed by concatenating the current state and the action, including modified actions in the LLM's prompt.

[10] The key distinction from standard MDPs is that π^F does not control the action that affects state transitions—it only influences it through suggestions a', while π^C ultimately determines the executed action a.

2.4.1 The MAMDP Optimization Problem

Example 1 (Action Modification in DeliData Wason Card Task). *In the Wason card selection task [152] as collected in the DeliData dataset [8], collaborators must decide which cards from a set (e.g., $\{U, S, 8, 9\}$) to flip to test the rule:* ***All cards with vowels on one side have an even number on the other.*** *Each player comes up with a solution individually and the group then deliberates to come to a consensus. The correct solution here is to flip U and 9; this would establish if U's reverse is an even number, and the contrapositive (if 9 has a consonant). Two participants' initial solution might be to flip only U while the other proposes flipping 8. In this setting, the dialogue history is the state s, the* FRICTION AGENT*'s proposed intervention (or action) is a, and the collaborators' reinterpretation is the transformation P_A. Suppose π^F's intervention a_t^F proposes checking an odd number as the rule does not single out vowels and even numbers only. In the MAMDP setting, the collaborator π^C responds with an action a_t^C that interprets the semantics of a_t^F, either faithfully or with some modification, such as checking U, 9 and 8.*

Having defined the collaborative MAMDP framework, we now formalize the problem of finding an optimal friction intervention policy π_*^F. The core challenge is that rewards depend on the collaborator's actual actions (a), not the friction agent's suggestions (a'). This creates a fundamentally different optimization problem than standard RL settings [32, 89].

Optimal Policy Objective The optimal friction policy π_*^F solves:

$$\pi_*^F = \arg\max_{\pi^F} \mathbb{E}_{\tau \sim (\pi^F, \pi^C, P_S)} \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \right] \tag{2.9}$$

where trajectories $\tau = (s_0, a'_0, a_0, s_1, a'_1, a_1, \ldots)$ are generated by the following process:

1. **Friction agent suggests**: $a'_t \sim \pi^F(\cdot|s_t)$
2. **Collaborator responds**: $a_t \sim \pi^C(\cdot|s_t, a'_t)$

3. **Reward received**: $R(s_t, a_t)$ *(based on collaborator's action)*

4. **State transitions**: $s_{t+1} = P_S(s_t, a_t)$ *(determined by collaborator's action)*

The key distinction from standard MDPs is the separation between the friction agent's chosen action a'_t and the action a_t that actually affects rewards and state transitions. The friction agent can only *suggest* interventions; the collaborator *mediates* how these suggestions translate into actual dialogue progression. The friction agent's influence on the collaboration is captured by the action modification distribution P_A, which marginalizes over all possible friction suggestions:

$$P_A(a|s, \pi^F) = \sum_{a' \in \mathcal{A}} \pi^F(a'|s) \cdot \pi^C(a|s, a') \tag{2.10}$$

Intuitively, this equation answers the question: "What is the probability that we observe collaborator action a in state s when the friction agent uses policy π^F?" The answer requires summing over all possible friction suggestions a' the agent might make, weighted by (1) how likely the agent is to suggest a', and (2) how likely the collaborator is to respond with a given that suggestion. For example, if the friction agent might suggest either "Have we considered all the cards?" (a'_1, probability 0.6) or "Let's think about the contrapositive" (a'_2, probability 0.4), and collaborators respond to a'_1 with "You're right, let me reconsider" (a_1) with probability 0.7, while they respond to a'_2 with a_1 with probability 0.5, then:

$$P_A(a_1|s, \pi^F) = 0.6 \times 0.7 + 0.4 \times 0.5 = 0.62$$

Value Function Formulation. The value function for policy π^F at state s is defined as the expected cumulative reward from that state forward:

$$V^{\pi^F}(s) = \mathbb{E}_{a \sim P_A(\cdot|s,\pi^F)} \left[R(s, a) + \gamma \cdot \mathbb{E}_{s' \sim P_S(\cdot|s,a)} \left[V^{\pi^F}(s') \right] \right] \tag{2.11}$$

Expanding the action modification distribution P_A, we obtain:

$$V^{\pi^F}(s) = \sum_{a' \in \mathcal{A}} \pi^F(a'|s) \cdot \left[\sum_{a \in \mathcal{A}} \pi^C(a|s,a') \cdot \Big(R(s,a) + \gamma \cdot V^{\pi^F}(P_S(s,a)) \Big) \right] \tag{2.12}$$

Breaking down the nested structure:

- **Outer sum ($\sum_{a'}$)**: Average over all friction suggestions the agent might make
- $\pi^F(a'|s)$: Weight by how likely the friction agent is to suggest a'
- **Inner sum ($\sum_a$)**: Average over all possible collaborator responses
- $\pi^C(a|s,a')$: Weight by how likely the collaborator responds with a given suggestion a'
- $R(s,a)$: Immediate reward from the collaborator's actual action
- $\gamma \cdot V^{\pi^F}(s')$: Discounted future value from the resulting state

This nested expectation structure is the defining characteristic of MAMDPs. The friction agent must reason not just about which intervention to suggest, but about how collaborators are likely to respond to each suggestion and what rewards those responses will generate.

Action-Value Function We can also express this in terms of the Q-function for friction suggestions a' using standard terminology [32]:

$$Q^{\pi^F}(s,a') = \mathbb{E}_{a \sim \pi^C(\cdot|s,a')} \Big[R(s,a) + \gamma \cdot V^{\pi^F}(P_S(s,a)) \Big] \tag{2.13}$$

Expanded:

$$Q^{\pi^F}(s,a') = \sum_{a \in \mathcal{A}} \pi^C(a|s,a') \cdot \Big[R(s,a) + \gamma \cdot V^{\pi^F}(s') \Big] \tag{2.14}$$

Specifically, $Q^{\pi^F}(s, a')$ tells us the expected value of suggesting friction intervention a' in state s. Critically, this value depends on how we expect collaborators to respond to a'. A suggestion that might seem helpful could have low Q-value if collaborators typically ignore or misinterpret it. The optimal policy (in the soft, temperature-scaled version) then becomes:

$$\pi^F_*(a'|s) = \frac{\exp(Q^{\pi^F_*}(s, a')/\beta)}{\sum_{a'' \in \mathcal{A}} \exp(Q^{\pi^F_*}(s, a'')/\beta)} \tag{2.15}$$

where β is a temperature parameter controlling exploration vs. exploitation.

Comparison to Standard MDPs To understand why standard MDP solutions fail in this setting, consider the standard Bellman equation for LLM policies [46] :

$$V^{\pi}_{\text{MDP}}(s) = \mathbb{E}_{a \sim \pi(\cdot|s)} \left[R(s, a) + \gamma \cdot V^{\pi}_{\text{MDP}}(s') \right] \tag{2.16}$$

In a standard MDP, there is **only one expectation** over actions—the agent's chosen action directly determines both the reward and the next state. Compare this to the MAMDP value function:

$$V^{\pi^F}_{\text{MAMDP}}(s) = \underbrace{\mathbb{E}_{a' \sim \pi^F(\cdot|s)}}_{\text{friction suggestion}} \left[\underbrace{\mathbb{E}_{a \sim \pi^C(\cdot|s,a')}}_{\text{collaborator response}} \left[R(s, a) + \gamma \cdot V^{\pi^F}(s') \right] \right] \tag{2.17}$$

The MAMDP formulation reveals a fundamental structural difference from standard MDPs through its **two nested expectations**: the outer expectation averages over friction suggestions a' sampled from $\pi^F(\cdot|s)$, while the inner expectation averages over collaborator responses a drawn from $\pi^C(\cdot|s, a')$ conditioned on the friction suggestion. Standard RL and preference-based RL (PbRL) approaches optimize under the assumption that the first expectation is sufficient, effectively treating the agent's chosen action as directly producing rewards and ignoring how collaborators modify, interpret, or reject suggested interven-

tions. See Theorem 1 in Chapter 5 for a more formal treatment of this observation. This leads standard methods to solve $\max_{\pi^F} \sum_{a'} \pi^F(a'|s) \cdot R(s, a')$, which implicitly assumes the friction suggestion a' is equivalent to the action a that affects state transitions and generates rewards. In contrast, the correct MAMDP optimization accounts for the mediating role of collaborators: $\max_{\pi^F} \sum_{a'} \pi^F(a'|s) \cdot \left[\sum_a \pi^C(a|s, a') \cdot R(s, a)\right]$. The inner summation over a, weighted by the collaborator's conditional response distribution $\pi^C(a|s, a')$, captures how different suggestions lead to different collaborator actions, which in turn produce different rewards.

Therefore, to retrieve the optimal policy π^F_* that maximizes Eq. 2.9, we face a fundamental challenge: **we do not know π^C (the collaborator policy) explicitly**. Although prior work [45, 153, 154] provides evidence that high-capacity frontier LLMs or specialized training with custom rewards can produce effective collaborator agents, it remains unclear how to train an ideal collaborator in settings with active intervention. The collaborator policy represents complex, context-dependent human behavior in response to friction suggestions—how people interpret interventions, whether they accept or resist them, and how they incorporate them into their reasoning.

This dissertation proposes **three** main approaches to address this challenge:

Approach 1: Learn π^F Conditioned on Informative State Features. We investigate this approach in the the `FAAF` method developed in Chapter 4. Instead of explicitly modeling the intractable P_A, we condition the friction policy on features of the state that correlate with successful collaboration. By conditioning interventions on frictive states $\phi(s)$—states where belief misalignment exists—the policy learns *when*[11] and *how* to intervene based on collaborative dynamics, without requiring an explicit model of π^C. Frictive states serve as informative features in the collaborative dialogue that ground intervention quality and help

[11] By "when", we mean the policy learns to generate intervention content appropriate to the frictive state, including deciding whether to intervene at all (by generating an EOS token) rather than relying on an external classifier to determine intervention timing.

steer the dialogue toward states more conducive to task success, which serves as a proxy for the reward $R(s, a)$. We demonstrate in Lemma 5 that standard preference optimization methods like IPO [49] without frictive-state conditioning suffers from vanishing gradients with respect to these crucial state features, while the full FAAF objective maintains gradient flow through frictive states. This approach provides a computationally feasible middle ground: rather than modeling how collaborators transform all possible interventions, we move our focus to the friction intervention policy instead that generates contextually appropriate interventions for states that require friction to resolve belief misalignment in collaborators.

Approach 2: Use counterfactual roleplay-based evaluation method to find the best-within-available-options for π^F. The MAMDP formulation above makes a structural point that is easy to state but difficult to act on: in collaborative dialogue, the friction agent's suggestion a'_t is not the event that produces reward or drives the trajectory; the executed action a_t is, and a_t is mediated by a collaborator with agency. This immediately raises a more basic question than policy optimization: *what does it even mean to assign reward to an intervention in a mediated, multi-turn process?* In standard single-agent RLHF-style settings [9, 19, 32, 46, 54], one can attach reward directly to the model's output (via a reward model, preference model, or judge) and train the policy to maximize it. In an MAMDP, however, such "content-only" evaluation is misaligned with the causal graph: the informational quality of a'_t does not guarantee that collaborators will incorporate it, nor that it will change the dialogue in a way that improves the team's eventual performance. The defining challenge is therefore not merely learning π^F under known dynamics, but *defining and measuring the right notion of improvement* under the mediated dynamics that the MAMDP captures. Taken together, the MAMDP lens reframes evaluation as a causal question: we do not ask whether an intervention is persuasive on its own, but whether it changes collaborator behavior in ways that improve the collaborative process.

Chapter 5 addresses this foundational problem by shifting focus from directly solving the MAMDP to *making its objective measurable in realistic interaction*. Concretely, we develop a practical roleplay-based, turn-by-turn communication protocol that generates online collaborative trajectories in which interventions can be evaluated through their downstream effects on the collaboration. Rather than treating collaboration as a single terminal outcome, we introduce *process-level* rewards grounded in the "common ground" [25]: the degree to which participants become mutually aligned on task-relevant propositions over time. This choice is not cosmetic. It is the minimal signal that is faithful to the MAMDP's mediated structure: if the collaborator ignores or misinterprets an intervention, then common ground should not improve, even if the intervention looks "good" in isolation. Conversely, if an intervention causes a turn-by-turn correction of beliefs and accelerates consensus on key propositions, it should receive credit even when final task success is noisy or delayed.

This immediately creates an evaluation obstacle: multi-turn collaborations diverge stochastically once the collaborator reacts, so we cannot fairly compare two intervention policies by simply observing one rollout from each. Our roleplay framework resolves this using a counterfactual evaluation design that treats the intervention agent *itself* as the causal variable of interest while controlling for collaborator-driven trajectory divergence. By pairing intervention conditions against matched baselines under controlled collaborator simulators, we estimate the *causal effect* of interventions on common-ground formation and task progress. This allows us to test not only whether the presence of the friction intervention agent is useful for attaining consensus, but also which among a class of aligned intervention agents is relatively better than others. In this sense, the chapter provides a principled compromise: without learning a full action-modification model π^C or planning through $\tilde{P}$, we still recover an evaluation signal that is appropriate for MAMDP dynamics and can discriminate interventions by their realized collaborative impact rather than their surface form.

Approach 3: Learn P_A Explicitly. We investigate this approach in Chapter 6. One could train a separate model to predict how collaborators modify interventions—that is, learn $\pi^C(a|s, a')$ for all states s, friction suggestions a', and collaborator responses a. This is precisely the focus of Chapter 6, which is to learn an ideal collaborator agent. Intuitively, this assumes a Centralized Training with Decentralized Execution (CTDE) framework [155], where the friction intervention agent is trained separately with some form of supervision (say, expert demonstrations) without any form of gradient-based communication [156] between the intervention agent and the collaborator and that the intervening agent is fixed during the actual execution of the MAMDP during collaborative deployment. While this appears computationally prohibitive due to the high-dimensional state and action spaces in LLM-based dialogue (especially since we impose no structural constraints[12] on what collaborators can express), there are deeper fundamental challenges beyond computational cost. First, training data exemplifying the action modification function P_A is scarce: natural friction interventions in the form of probing questions are rare [7, 8, 138] in textual records of human collaboration, and when they do occur, they may not represent optimal collaborative behavior. Second, expert trajectories may contain demonstrations that an optimal collaborator should *not* learn from [157]. For instance, if expert data includes dialogues where collaborators ignored well-intentioned suggestions, we would not want the model to reproduce such dynamics. Similarly, suboptimal interventions may be incorrectly accepted by collaborators under social pressure [158]—we observe in classroom settings that students sometimes incorporate irrelevant peer suggestions to "appear smart" or avoid seeming uncooperative [159]. Conversely, collaborators might ignore genuinely helpful interventions but then modify their interpretation in ways that introduce new irrelevancies. These spurious patterns in real or AI-generated collaborative data make direct supervised

[12]Defining prompts to reflect task-settings naturally restricts the space of possible utterances but it is still vast.

learning of π^C problematic: the model would learn to reproduce both productive and counterproductive collaborative dynamics without distinguishing between them.

2.4.2 Challenges and Design Choices

Implementing counterfactual roleplay evaluation requires resolving several design challenges that emerge from the MAMDP structure and the practical constraints of LLM-based multi-agent systems. First, **communication structure and partial observability**: we adopt synchronous turn-taking where collaborators alternate speaking, which provides clear intervention points and matches the structure of many small-group collaborative tasks [160]—in contrast to approaches that focus on asynchronous communication [161]. Critically, collaborators do not have access to second-order beliefs during dialogue turns—that is, collaborator A cannot directly observe what collaborator B believes about task-specific propositions. This partial observability is by design: if collaborators had perfect visibility[13] into each other's belief states and recognized their misalignments explicitly, friction interventions would be unnecessary. However, collaborators do have access to their own internal belief states and the full dialogue history excluding frictive states $\phi(s)$, allowing them to infer others' beliefs indirectly through utterances. The frictive states $\phi(s)$ that our friction agent conditions on are therefore not part of the collaborators' input context—the friction agent has a privileged view of belief misalignments that it uses to decide when and how to intervene.

Second, **reward specification for training versus evaluation**: while we use common ground (CG) as the team-level metric for *evaluating* collaborative processes in our counterfactual framework, we deliberately avoid using CG directly as a training signal for learning collaborator agents. This distinction addresses a fundamental challenge in multi-agent learning: optimizing for consensus directly can lead to reward hacking, where agents learn superficial agreement without genuine reasoning—a phenomenon called

[13]Note that even though FAAF intervention policies condition their interventions on the frictive states ϕ, these second order beliefs are not exposed in the collaborator's input prompts during the MAMDP interactions.

"overcollaboration" [42, 154, 162]. LLMs are particularly prone to this shortcut: they may simply echo others' propositions to maximize agreement rewards rather than engaging in authentic collaborative reasoning. Instead, we use individual correctness over task-specific propositions as a proxy reward during training, under the hypothesis that true collaboration *emerges* [163] when agents develop accurate beliefs and can communicate them effectively. By construction of our friction agent's role as a support agent with no privileged task knowledge beyond the group's combined state, we assume that common ground convergence should occur naturally when individual agents reason correctly—thus the proxy reward should be sufficient. This design choice also addresses a third challenge: **computational feasibility**. Training multiple collaborator agents with team-level rewards in the training loop would require simultaneously updating billions of parameters across multiple LLM instances and storing at least four models in memory for standard RL training with PPO [11]. Such multi-agent training approaches are feasible for smaller neural networks in traditional MARL settings [164, 165] but become prohibitively expensive for LLMs. Using proxy rewards allows us to train individual agents efficiently while reserving team-level CG evaluation for our counterfactual framework, where we measure collaborative quality without requiring gradient updates through multiple agents simultaneously.

Fourth, **defining collaborative processes versus outcomes**: collaborative outcomes are relatively straightforward to measure since task definitions provide ground-truth solutions (e.g., correct card selection in the Wason task). However, collaborative processes—the *how* of collaboration rather than just the final result—require measuring intermediate progress toward subgoals. We operationalize this through common ground over task-relevant propositions: the alignment of participants' beliefs measured at each dialogue turn. This choice is motivated by the structure of small-group collaboration, where progress necessitates forming multiple consensuses while resolving task subgoals [166, 167]. For instance, in the Weights Task, groups may need to agree that the red block weighs 10

grams before making inferences about larger blocks. Even if a single collaborator produces a high-quality utterance, the group may not progress if others ignore or misunderstand it—thus measuring individual contributions in isolation is insufficient. Our team-level CG metric provides a form of structural credit assignment [154] that fairly allocates success to collaborative processes, distinguishing cases where the group converges on correct beliefs from cases where individuals happen to be correct but fail to coordinate.

2.5 Experimental Datasets

In this section, we briefly summarize the datasets used across the experimental chapters of this dissertation. For Chapter 3, we draw on large-scale preference alignment corpora spanning multiple task families, including instruction following (UltraFeedback), Reddit post summarization (TL;DR), news summarization (CNN/DailyMail), and external alignment benchmarks such as AlpacaEval 2.0. For the collaborative dialogue experiments in Chapters 4, 5, and 6, we rely on human–human collaboration datasets, specifically the Weights Task Dataset (WTD) [7] and DeliData [8]. In most of the collaborative dialogue experiments, the raw human datasets are augmented with AI-generated synthetic dialogues (e.g., produced by high-capacity models such as GPT-4o) to expand coverage and systematically control task conditions. Where feasible, subsets of these augmentations were validated through targeted human evaluation to ensure plausibility and faithfulness. Experiment-specific details regarding augmentation pipelines, validation procedures, and preprocessing are discussed in the respective chapters.

DeliData Dataset The DeliData corpus [8] is a publicly available dataset designed to study group deliberation in multiparty problem-solving. It contains dialogues from 500 groups of five participants engaged in the Wason card selection task [152], a classic reasoning puzzle in which participants must decide which cards to flip to test the rule: "All cards with vowels on one side have an even number on the other" (cf. Example 1). The dataset

consists of 14,003 utterances, with cards referenced by their visible symbols (letters or numbers), and includes task performance measures both before and after discussion. Beyond transcriptions, DeliData is annotated with deliberation cues, argumentation structures, and other conversational markers relevant to collaborative reasoning. Importantly for our purposes, [8] also identified "probing" interventions—naturally occurring friction turns that prompt reflection without introducing new information. However, these are relatively rare, averaging only 3.46 probing interventions per group across 17,110 total utterances, underscoring the sparsity of explicit deliberation prompts in natural collaboration. Official splits assign 300 groups for training, 100 for development, and 100 for testing.

Weights Task Dataset (WTD). The Weights Task Dataset (WTD) [1] is an anonymized, publicly available audiovisual corpus designed to study small-group collaboration. It contains recordings of 10 triads working together to deduce the weights of differently colored blocks using a balance scale and to infer the underlying rule that determines the weights. The task unfolds in three stages: solving with the scale available, solving without the scale, and finally reasoning using only the inferred weight pattern. In our experiments, we evaluate models across all task phases, including transcripts involving inference about the mystery block's weight when the scale is unavailable. While simulated WTD dialogues (including GPT-generated data) do not explicitly model scale calls, they are grounded in the true block weights and reflect the same inference dynamics through dialogue-based reasoning. The original dataset includes multiple annotations, including gold-standard dialogue transcripts. Utterances frequently reference block colors, hypothesized weights, and reasoning strategies, making the dataset well-suited for analyzing probing questions and potential causal interventions. The official split assigns 7 groups for training, 1 for development, and 2 for testing.

We additionally annotated WTD for naturally occurring friction following the definition given in Oinas-Kukkonen and Harjumaa [137].[14] Two annotators independently labeled half of the groups each, while a third annotated all ten and adjudicated disagreements. Inter-annotator reliability measured by Cohen's κ was 0.632, indicating substantial agreement. On average, we identified approximately four naturally occurring friction interventions per group. From a supervised learning perspective, this amount of friction intervention labels is very few to have any reasonable confidence for training an LLM to generate such interventions. This motivated our efforts to use LLM-generated synthetic dialogues for our experiments in Chapter 4.

UltraFeedback UltraFeedback [10] is a recent large-scale, high-quality preference dataset designed to support robust RLHF and alignment research by addressing limitations in scale and diversity of previous preference corpora. It aggregates over a million GPT-4–generated feedback annotations across roughly 250,000 user-assistant interactions, with carefully calibrated scoring on multiple quality dimensions such as instruction-following, truthfulness, helpfulness, and honesty. UltraFeedback's breadth of prompts and feedback signals makes it a valuable benchmark for training and evaluating preference-based optimization methods and for studying how models learn preference strength and nuanced human judgments. In our experiments, we use the UltraFeedback–Binarized[15] variant derived from the original UltraFeedback dataset. The full UltraFeedback corpus contains roughly 64k prompts, each paired with four candidate completions generated by a diverse set of open-source and proprietary LLMs. These completions are scored by GPT-4 along multiple axes such as helpfulness, honesty, and overall quality.

[14] In this setting, friction interventions function as indirect persuasion, prompting participants to reconsider assumptions or propositions in light of task-relevant evidence; see Definition 8.

[15] https://huggingface.co/datasets/HuggingFaceH4/ultrafeedback_binarized

Reddit TL;DR and CNN Daily The Reddit TL;DR summarization dataset [19,168] dataset consists of social media posts paired with human-written "TL;DR" summaries, originally mined for summarization research, and has been widely adopted as a preference dataset for RLHF fine-tuning. Human annotators compare candidate summaries to judge which better reflects the content of the original post, providing rich pairwise preference signals that inform reward model training for summarization policies. Models trained with TL;DR preferences have been shown to outperform supervised baselines and, importantly, generalize to external news summarization benchmarks like the CNN/DailyMail corpus [169], demonstrating that human preference signals from Reddit can transfer to other domains.

AlpacaEval 2.0 AlpacaEval 2.0 [82] is an automated evaluation benchmark for instruction-following language models that scores systems through pairwise comparisons judged by a high-capacity model (typically GPT-4). Unlike the original AlpacaEval benchmark, which suffered from systematic artifacts such as verbosity bias, AlpacaEval 2.0 explicitly introduces length-control and related debiasing procedures to ensure that win rates more faithfully reflect genuine quality differences rather than stylistic tendencies. The result is a scalable, low-cost evaluation framework that produces interpretable win-rate metrics closely aligned with human judgments across thousands of prompts. As such, AlpacaEval 2.0 serves as a standard external benchmark for assessing preference-aligned models and tracking progress in instruction following and alignment quality.

Chapter 3

Direct Reward Distillation and Policy-Optimization (DRDO)

Chapter Overview This chapter studies how to align LLMs with human preferences using *offline*[16] preference datasets. We observe that widely used datasets, such as UltraFeedback [85] and Reddit TL;DR [19] (see Section 2.5 and Figure 3.2), contain substantial non-deterministic preferences, where the preference signal is weak or ambiguous. In these settings, standard alignment methods such as DPO [9] exhibit important limitations, which motivate the DRDO approach. Prior methods that consider such preference noise have approached this problem from the angle of data-filtering, flipping of preference labels, injecting noise during training to learn more robust preferences, as well as using static offsets to robustly learn preference strengths—often requiring prior domain knowledge and assumptions about noise types [130, 170–172]. Others have assumed access to a "general" preference model [50], often learned from human data or assumed to be so considering the vast knowledge of human preferences in frontier models [124]. We do not make such assumptions. We address this problem in a more principled manner by motivating the issue as a "misalignment problem"—a divergence between the kinds of policies we intend to learn (e.g., those that humans prefer) vs. those that end up being learned (e.g, those that simply focus on reward maximization). This brings new insights on why standard approaches fail under non-deterministic preferences (e.g., their reliance on a reference model for constrained optimization). Noting these challenges, we introduce a novel approach to this problem, which combines teacher-based reward distillation and learning from focal loss [119], without the need for a reference model. DRDO is the first approach

[16] Here, *offline* refers to learning from a fixed dataset of preference comparisons, without on-policy data collection or online interaction with the environment.

to combine these techniques to adaptively learn from the diverse range of preference strengths in offline datasets, without compromising learning on either deterministic or non-deterministic preferences. To the best of our knowledge, DRDO is the first instance of adapting a self-adjusting focal loss in preference learning; prior works [173,174] have used such losses for general LLM finetuning, but not in the context of self-adjusted or "adaptive" preference learning in terms of an implicit reward formulation. Our experiments on article summarization, instruction-following and open benchmarks like AlpacaEval 2.0 empirically validate our claims.

3.1 Introduction

In the development of LLMs, robust modeling of human preferences is essential for producing outputs that are contextually and situationally appropriate, and ultimately for building systems that users find genuinely useful. However, many popular alignment approaches, including Direct Preference Optimization (DPO) [9] and its variants, implicitly assume that each preference pair—consisting of a preferred and a dispreferred response—contains a clear winner. This assumption rarely holds in practice: real-world preference data often exhibit weak or ambiguous signals, arising from low annotator confidence [19] or inherently small differences in preference strength [130,175].

When trained on such data, reward functions estimated using existing methods can diverge from the underlying preference structure, leading to a mismatch between the policy optimized for the learned reward and the policy that is optimal for the true preference distribution. This "preference gap," or *misalignment between rewards and preferences*, can result in policy degeneracy and systematic underfitting. Even in the presence of strictly deterministic preferences, prior work has shown that policy degeneracy can lead to substantial overfitting, where the LLM assigns increasing probability mass to tokens that are not labeled as preferred in the offline dataset [48]. Other studies have observed counterintuitive failure modes in which models decrease the likelihood of both preferred and dispreferred

responses during training, undermining the intended learning objective [46, 47]. These issues are further exacerbated when preference data are non-deterministic and the underlying preference signal is weak. To address these challenges, we build on the insight that certain limitations of DPO's implicit reward formulation at the policy optimization stage are less severe during reward modeling, and we leverage this distinction to make the following novel contributions:

1) We provide a thorough theoretical grounding of problems with DPO and variants according to the Bradley-Terry model that demonstrates why they are challenged by nuanced or "non-deterministic" preference pairs; 2) We introduce ***Direct Reward Distillation and Policy-Optimization (DRDO)***, a novel approach preference optimization approach that combines the two stages by explicitly distilling rewards into the policy model (Figure 3.1) and avoids the above limitations; 3) Through experiments on Ultrafeedback, TL;DR, and AlpacaEval 2.0, we show that DRDO is better able to fit to nuanced or ambiguous preferences without sacrificing performance on clear preferences or large-scale data.

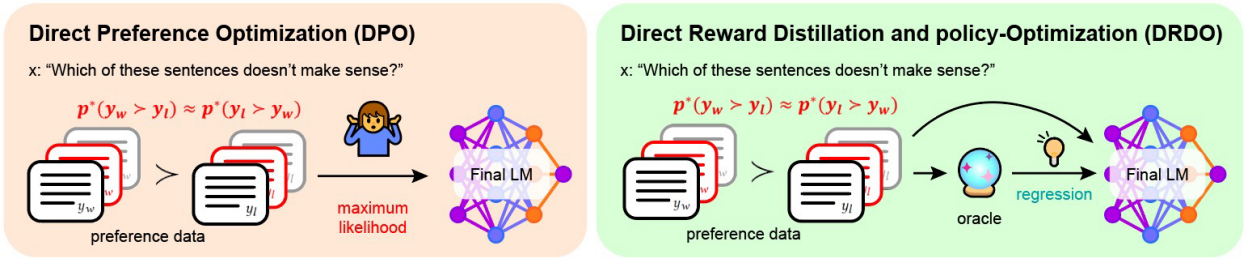
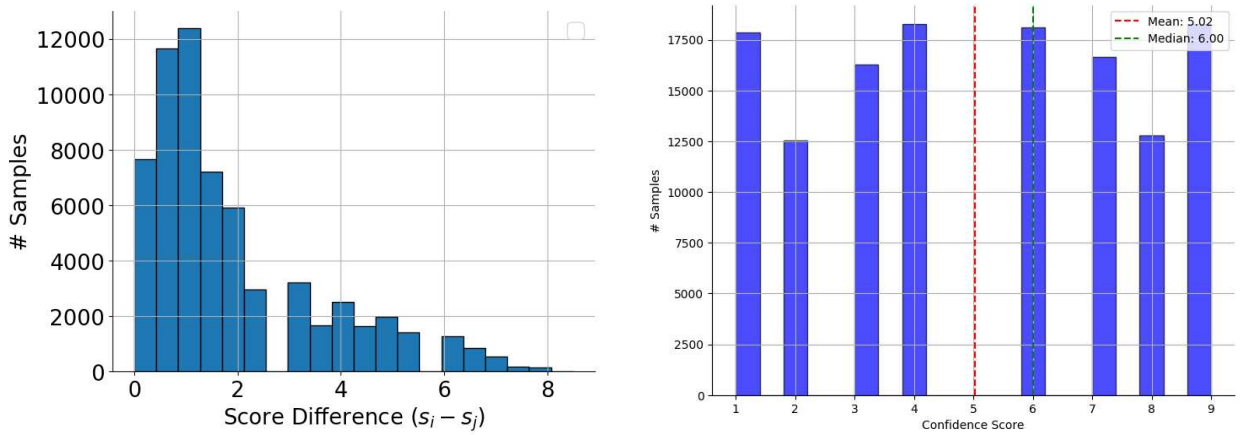


Figure 3.1: Unlike popular supervised preference alignment algorithms like Direct Preference Optimization (DPO; [9]) that learns rewards implicitly, **DRDO directly optimizes for explicit rewards from an Oracle while simultaneously learning diverse kinds of preference signals during alignment.** Optimized with a simple regression loss based on difference of rewards assigned by the Oracle and a novel focal-log-unlikelihood component (see Sec. 3.3), DRDO avoids DPO's particular challenges at learning non-deterministic preference pairs, thereby bridging the gap between estimating the preference distribution from the data and the true preference distribution p^*. Additionally, DRDO does not require an additional reference model during training and can leverage reward signals even when preference labels are not directly accessible.



(a) GPT-4 preference score differences on Ultrafeedback [176].

(b) Human expert confidence labels on Reddit TL;DR [168].

Figure 3.2: Preference strength distributions across datasets. Ultrafeedback exhibits a wide spectrum of LLM-estimated preference strengths (how strongly a sample is rated vs. another, for the same prompt or question), ranging from near-deterministic to highly non-deterministic comparisons showing weak preferences, while Reddit TL;DR provides discrete "expert" confidence labels on preference annotations between two responses. Together, these distributions motivate learning methods that explicitly account for heterogeneous preference strength, enabling DRDO to robustly learn from both deterministic as well as non-deterministic preferences.

3.2 Problem Formulation

Let $\mathcal{D}_{\text{pref}} = \{(x, y_w, y_l)\}_{i=1}^{N}$ be an offline dataset of pairwise preferences with sufficient coverage, where y_w and y_l are the respective winning (preferred) and losing (less preferred) completions given a context x. In contrast to *deterministic* preferences, which are defined as those where $p^* \in \{0, 1\}$ [48], let $\mathcal{D}_{\text{nd}} \subset \mathcal{D}_{\text{pref}}$ denote the subset of *non-deterministic* preference pairs where $P(y_w \succ y_l|x) \approx 1/2$. These cannot be perfectly captured by Bradley-Terry models but are prevalent in popular preference alignment datasets.[17] Assume three possible completions $y_1, y_2, y_3 \in \mathcal{Y}$, the space of all possible completions. Let $r^*(x, y) \in \mathbb{R}$ be an underlying true reward function that is deterministic and finite for all completions, $\pi_{\theta^*}(y|x)$ be the learned policy, and $\pi_{\text{ref}}(y|x)$ be the reference policy with $\text{supp}(\pi_{\text{ref}}) = \mathcal{Y}$. Given $\text{supp}(\rho) = \text{supp}(\mu) \times \mathcal{Y} \times \mathcal{Y}$ where ρ is the data distribution and μ is the context

[17] For instance, an analysis of the popular RLAIF-inspired scalable alignment benchmark Ultrafeedback [10] reveals that it contains approximately **17%** non-deterministic samples as defined above [107].

distribution, the challenge is leaning a policy to effectively handle both deterministic and non-deterministic preferences in an offline fashion.

Misalignment of Rewards and Preferences Although preferences p^* can be implicit in rewards, refining LLMs based on rewards alone does not imply that they learn preferences optimally. We call this the **misalignment problem**, where the two objectives are fundamentally divergent [50]. This issue is particularly significant in aligning high-dimensional policies like LLMs, where the training data often contains non-deterministic samples or weak learning signals. Such uncertainty often appears in equal or near-equal rewards or ambiguous expert judgments [19].

In other words, traditional reward modeling assumes that maximizing reward leads to optimal behavior, but in preference-based learning, the best policy can be different from the highest-reward policy. Where the *reward-optimal* policy, or the policy that maximizes Elo score under the BT assumption, fundamentally diverges from the *preference optimal* policy, or the policy that maximizes the probability of selecting the ground truth winning response. This creates a "preference gap" and the misalignment problem occurs. However, since LLMs are overparameterized with many near-optimal solutions [46] and are expected to generalize across diverse and often uncertain preferences [177], it is essential to understand this divergence or preference gap, as it can be especially sensitive to the existence of non-deterministic preferences in training.[18]

This insight motivates the core of DRDO: to effectively learn from non-deterministic preferences while maintaining performance across general preference data, we bound this preference gap to possible sources of errors scaled by the offline data coverage, while simultaneously learning both high-quality distilled rewards and preferences in an efficient offline manner. Mathematically, we illustrate this sensitivity to non-deterministic pairs in the BT model in Lemma 3.

[18]Non-deterministic preferences differ from noisy (flipped) labels [129,130], which are often handled with label smoothing, data pruning, or noise-aware training.

Lemma 3 (Sensitivity of Preference Gap to Non-Deterministic Preferences)**.** *The preference gap $\delta_{\mathcal{P}}$ between reward-optimal (π^*_R) and preference-optimal ($\pi^*_{\mathcal{P}}$) policies can be highly sensitive to the presence of non-deterministic preference pairs (where $\mathcal{P}(y \succ y') = \mathcal{P}(y' \succ y) = 1/2$). Such non-determinism can lead to a substantial increase in $\delta_{\mathcal{P}}$ even when the reward gap δ_R (based on a fixed reward function R) remains unchanged.*

We provide a proof sketch in Appendix A.1. The core insight here is that common approaches like DPO that assume a BT preference model are more likely to be sensitive to this preference gap, since they assume that true preferences are learnable from a mapping of preferences onto differences in scalar implicit rewards. Prior work [48,49] highlights DPO's practical tendency to underfit the true preference distribution because of unboundedness of its implicit rewards. For example, for two completions y and y' with a non-deterministic preference relation, the DPO estimate of $p^*(y \succ y')$ using $\sigma\big(\beta \log \frac{\pi_\theta(y|x)\pi_{\text{ref}}(y'|x)}{\pi_\theta(y'|x)\pi_{\text{ref}}(y|x)}\big)$ leads π_θ to assign equal rewards to both, meaning it learns that $r^*(x, y) \approx r^*(x, y')$ for any $(x, y, y') \in \mathcal{D}_{\text{nd}}$. This implies that the reward model fails to strongly differentiate between completions in these cases, even though one is explicitly "chosen" in the true preference data and the other is "rejected," thus leading to the preference gap, and the motivation of DRDO to address these issues.

3.3 Direct Reward Distillation and Policy-Optimization

Direct Reward Distillation and Policy-Optimization or DRDO gets its name since the policy learns rewards directly from an external "teacher" reward model, similar to knowledge distillation [118]. DRDO objective also has a preference component that lets the policy optimize for preferred responses, while lowering likelihood of dispreferred responses in an adaptive manner. Specifically, DRDO training involves two phases. First, we fit an Oracle model $\mathcal{O}$ to the annotated preference data. Second, we use $\mathcal{O}$ as a teacher to align the policy model (student) with a knowledge-distillation-based multi-task loss [118, 178] that regresses the student's rewards onto those assigned by $\mathcal{O}$. The student

model simultaneously draws additional supervision from binary preference labels for efficient use of finite data.

Training the Oracle Reward Model We use [92]'s strategy to optimize a high-quality and generalizable $\mathcal{O}$ while retaining its language generation abilities. Regularizing the shared hidden states with a language generation loss in addition to the traditional RLHF-based reward modeling improves generalization to out-of-distribution preferences. For this we initialize a separate linear reward head (parametrized by ϕ')[19] on top of the base LM (parametrized by ϕ), which adds only 0.003% more parameters compared to the LM's language modeling head. This also helps minimize reward hacking [90, 91], especially in offline settings. As such, $\mathcal{O}$ is optimized to minimize the following objective:

$$\mathcal{L}_{\mathcal{O}}(r_\phi, \mathcal{D}_{\text{pref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}_{\text{pref}}}\Big[(1-\alpha)\log\sigma\big(r_{\phi'}(x,y_w) - r_{\phi'}(x,y_l)\big) + \alpha\log r_\phi(y_w)\Big]. \tag{3.1}$$

where α is the strength of language-generation regularization on the winning response log-likelihoods assigned by $\mathcal{O}$ (denoted $\log(r_\phi(y_w))$ and ϕ is the parameters of $\mathcal{O}$ being estimated.

Simultaneous Student Policy Alignment to Rewards and Preferences A converged Oracle $\mathcal{O}$ can plausibly estimate true pointwise reward differences $r^*(x, y_1) - r^*(x, y_2)$ for any unlabeled sample (x, y_1, y_2), without needing explicit access to preference labels. Next, the student model π_θ is optimized to match its own reward predictions $\hat{r_1}$ and $\hat{r_2}$ to $\mathcal{O}$'s reward differences, aligning closely with $\mathcal{O}$'s behavior, while $\mathcal{O}$ itself does not get updated. The student model's reward estimates are computed with a linear reward head on top of the base LM, similar to Oracle training. This optimization uses a knowledge-distillation loss ($\mathcal{L}_{\text{kd}}$) that combines both a supervised ℓ^2-norm term and a novel focal-softened [119, 179]

[19]For simplicity of notation, on the LHS of Eq. 3.1 we subsume parameters ϕ' into ϕ.

Algorithm 1 DRDO: Direct Reward Distillation and policy-Optimization

Input: Preference dataset $\mathcal{D}_{\text{pref}} = \{(x, y_w, y_l)\}$; initialized policy + reward head $\pi_{\theta,\theta'} \leftarrow \text{SFT}(\theta) \oplus r_{\theta'}$
Output: Optimized parameters θ for aligned policy π_θ
Train oracle reward model r_ϕ (Eq. 3.1)
for $t = 1$ to T **do**
    for each $(x, y_w, y_l) \in \mathcal{D}_{\text{pref}}$ **do**
        Compute oracle rewards: $r_1^* = r_\phi(x, y_w)$, $r_2^* = r_\phi(x, y_l)$
        Compute student rewards: $\hat{r}_1 = r_{\theta'}(x, y_w)$, $\hat{r}_2 = r_{\theta'}(x, y_l)$
        Compute distillation loss $\mathcal{L}_{\text{kd}}$ (Eq. 3.2)
        Update $\pi_{\theta,\theta'}$ using gradient step on $\mathcal{L}_{\text{kd}}$
    end for
end for
return Final policy π_θ

log odds-unlikelihood component:

$$\mathcal{L}_{\text{kd}}(r^*, \pi_\theta) = \mathbb{E}_{(x,y_1,y_2)\sim\mathcal{D}_{\text{pref}}}\Bigg[\underbrace{\left(r^*(x,y_1) - r^*(x,y_2) - (\hat{r}_1 - \hat{r}_2)\right)^2}_{\text{Reward Difference}} \quad - \underbrace{\alpha(1-p_w)^\gamma \log\left(\frac{\pi_\theta(y_w \mid x)}{1-\pi_\theta(y_l \mid x)}\right)}_{\text{Contrastive Log-"unlikelihood"}}\Bigg], \quad (3.2)$$

where $p_w = \sigma(z_w - z_l)$ and quantifies the student policy's confidence in correctly assigning the preference from $z_w = \log \pi_\theta(y_w \mid x)$ and $z_l = \log \pi_\theta(y_l \mid x)$, or the log-probabilities of the winning and losing responses, respectively. Hyperparameters γ and α regulate the strength of the modulating term and weighting factor respectively.[20] As shown in Eq. 3.2, there is no shared parametrization between the policy and $\mathcal{O}$. $\mathcal{L}_{\mathcal{O}}$ is only a function of $\mathcal{O}$'s own parameters ϕ and the preference dataset, $\mathcal{D}_{\text{pref}}$. The Oracle parameters are *not* updated during policy training. Algorithm 1 shows a detailed breakdown of DRDO training.

[20]Note that we do not use α as it is traditionally used in focal loss for weighting class imbalances [180]. Instead, since the true preference distribution is unknown, we tune the empirical optimal value based on validation data and keep it fixed during training.

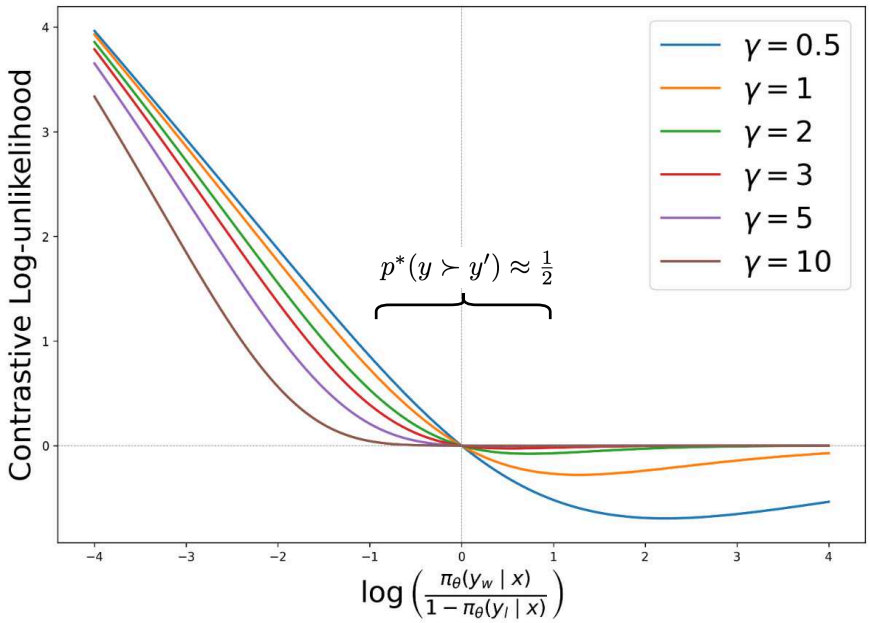


Figure 3.3: Illustration of the DRDO preference loss as a function of the log-unlikelihood ratio across various values of γ, the focal modulation parameter.

How does the DRDO gradient update affect preference learning? "Contrastive log-unlikelihood" in Eq. 3.2 is DRDO's preference component. One can immediately draw a comparison with DPO which uses a fixed β parameter. DRDO uses a modulating focal-softened term, $(1 - p_w)^\gamma$, where π_θ learns from both deterministic and non-deterministic preferences, effectively blending reward alignment with preference signals to guide optimization. Intuitively, unlike DPO's β that is constant for every training sample, this modulating term amplifies gradient updates when preference signals are weak ($p_w \approx 0.5$) and tempering updates when they are strong ($p_w \approx 1$), thus ensuring robust learning across varying preference scenarios. When π_θ assigns high confidence to the winning response ($p_w \approx 1$), the focal loss contribution diminishes (see Eq. 3.2), reflecting minimal penalty due to strong deterministic preference signals. However, for harder cases with non-deterministic true preference ($p^*(y \succ y') \approx (p^*(y' \succ y))$, the focal term $\alpha(1 - p_w)^\gamma$ keeps DRDO gradient updates active and promotes learning even when preferences are ambiguous (Figure 3.3 in Appendix A.2). This adaptive behavior ensures that DRDO maintains effective preference learning across varying preference strengths, where conven-

tional methods like DPO struggle[21]. As shown in the gradient analysis in Appendix A.2 (Figure 3.3) with empirical proof in Figure A.3 (bottom-right), this modulating term acts like DPO's gradient scaling term, in that it scales the DRDO gradient when the model incorrectly assigns preferences to easier samples.

DRDO's Reward Distillation term Intuitively, DRDO's first component in the loss effectively regresses over the Oracle's reward differences across all response pairs (x, y_1, y_2) in $\mathcal{D}_{\text{pref}}$, using π_θ's own reward estimates, $\hat{r}_1$ and $\hat{r}_2$. Assuming the Oracle's reward estimates correctly reflect the true preference strength, this simple distillation component is precisely minimized when the student π_θ perfectly emulates the Oracle's reward difference assignment, without requiring it to explicitly match the pointwise rewards. The square over this term ensures that training is stabilized especially since we allow this component to fully regularize the loss without applying any weighting terms.

3.4 Experimental Setup

Our experiments address two questions: **How robust is DRDO alignment to nuanced or diverse preferences, in OOD settings?** and **How well does DRDO achieve reward distillation with respect to model size?** We empirically investigate these questions on two tasks: **summarization** and single-turn **instruction following**. Our experiments, including choice of datasets and models for each task are designed to to validate our approach as robustly as possible, subject to research budget constraints (see Section A.8 for more). We compare our approach with competitive baselines such as DPO [9] and e-DPO [48] as well as on-policy methods like PPO [89], including the supervised finetuned (baseline) versions depending on the experiment. Some minor notes on the experimental setup described below can be found in Appendix A.7.

[21]See Lemma 1 and Lemma 2 for in-depth analysis in our background chapter (Chapter 2)

How robust is DRDO alignment to nuanced or diverse preferences, in OOD settings? We evaluate this using the **Reddit TL;DR summarization dataset** [19,168] and **CNN/Daily Mail Corpus** [169]. For robust out-of-domain (OOD) evaluation, we train models on Reddit TL;DR and use CNN/Daily Mail articles for an OOD test distribution.

Table 3.1: Mean confidence and normalized edit distance statistics our TL;DR preference generalization experiment.

Split	Conf (Mean)	Edit Dist (Mean)	Train	Validation
$\mathcal{D}_{all}$	5.01	0.12	44,709	86,086
$\mathcal{D}_{hc,he}$	7.31	0.15	10,136	1,127
$\mathcal{D}_{\ell c,\ell e}$	2.67	0.09	10,000	1,112

TL;DR Summarization Dataset Splits Table 3.1 shows the mean confidence and normalized edit distance statistics in the TL;DR dataset. We use these to compute the deterministic and nondeterministic splits. $\mathcal{D}_{all}$ represents the full data [19]. $\mathcal{D}_{hc,he}$ and $\mathcal{D}_{\ell c,\ell e}$ represent subsets created by splitting at the 50th percentile of the human labeler confidence and edit distance values. The range of labeler confidence values for the entire training data range is $[1, 9]$. Each contains $\sim$10k training samples, where the former comprises samples from the upper 50th percentile of the confidence and edit distance scores, *mutatis mutandis*. $\mathcal{D}_{hc,he}$ and $\mathcal{D}_{\ell c,\ell e}$ represent "easy" (deterministic) and "hard" (non-deterministic) preference samples where combined labeler confidence and string-dissimilarity act as proxy for the extreme ends of preference strengths/signals. This choice is intentional to more robustly evaluate DRDO under more difficult preference data settings. Furthermore, previous theoretical and empirical work [101] has shown that supervised methods such as DPO fail to learn optimal policies when the token-level similarity is high in preference pairs, especially at the beginning of the response. Therefore, we apply this combined thresholding for all our experiments on TL;DR summarization dataset. See Appendix 3.4

for more details. All baselines are initialized with Phi-3-Mini-4K-Instruct weights [181] with supervised fine-tuning (SFT) on Reddit TL;DR human-written summaries.

How does model size affect reward distillation? We evaluate all baselines on the cleaned version of the **Ultrafeedback dataset** [10] This experiment is conducted on the OPT suite of models [182], at 125M, 350M, 1.3B, and 2.7B parameter sizes. The student policy is trained with SFT on the chosen responses of the dataset, following [9]. We exclude larger OPT models to focus on testing our distillation strategy with full-scale training, rather than parameter-efficient methods (PEFT; [183, 184]) to allow a full-fledged comparison considering all trainable parameters of the base model. For completeness and comparison across model families, we also include Phi-3-Mini-4K-Instruct following the same initialization.

Evaluation Strategy **CNN/Daily Mail Corpus** provides human-written reference summaries, so we use a high-capacity Judge to compute win-rates against baselines on 1,000 randomly-chosen samples. Following [9], we use GPT-4o to compare the conciseness and the quality of the DRDO summaries and baseline summaries, *while grounding its ratings to the human written summary*. See Appendix A.11 for our prompt format. For instruction following on **Ultrafeedback**, we sample generations from DRDO and all baselines at various diversity-sampling temperatures and report win-rates on the Ultrafeedback test set against $\mathcal{O}$. Following [185], we consider a *win* to be when, for two generations y_1 and y_2, we get $r(x, y_1) > r(x, y_2)$, where $r(x, y_1)$ and $r(x, y_2)$ are the expected rewards (logits) from the policies being compared. Additionally, we evaluate DRDO and baselines (all trained on Ultrafeedback) on **AlpacaEval 2.0**, a 805-sample OOD benchmark with length-controlled comparisons judged by GPT-4 Turbo [82]. Hyperparameters and model configurations are given in Appendix A.10. To evaluate out-of-domain generalization, we train DRDO and all baselines exclusively on the Reddit TL;DR dataset and evaluate them on the CNN/Daily Mail splits, which consist of news article summarization. For direct head-to-head comparisons on CNN/Daily Mail, we restrict our analysis to aligned

baselines—specifically DPO and e-DPO—while also comparing against DPO, e-DPO, SFT, and PPO on AlpacaEval, and DPO, e-DPO, and SFT on UltraFeedback.

DRDO and e-DPO Specifics Although DRDO requires an explicit reward Oracle $\mathcal{O}$, we fix only one model (based on parameter size and model family) for each experiment. We use Phi-3-Mini-4K-Instruct and OPT 1.3B causal models initialized with a separate linear reward head while retaining the language modeling head weights.[22] Fixing the size of $\mathcal{O}$ allows us to evaluate the extent of preference alignment to smaller models, as in classic knowledge distillation [178]. To reproduce the e-DPO [48] baseline, we train three reward models using Eq. 2.2 with the mentioned base models but with different random initialization on the reward heads.

Evaluation **CNN/Daily Mail Corpus** provides human-written reference summaries, so we use a high-capacity Judge to compute win-rates against baselines on 1,000 randomly-chosen samples from the TL;DR test set. Following [9], we use GPT-4o to compare the conciseness and the quality of the DRDO summaries and baseline summaries, *while grounding its ratings to the human written summary*. See Appendix A.11 for our prompt format. For instruction following on **Ultrafeedback**, we sample generations from DRDO and all baselines at various diversity-sampling temperatures and report win-rates on the Ultrafeedback test set against $\mathcal{O}$. Following [185], we consider a *win* to be when, for two generations y_1 and y_2, we get $r(x, y_1) > r(x, y_2)$, where $r(x, y_1)$ and $r(x, y_2)$ are the expected rewards (logits) from the policies being compared. We also evaluate DRDO (trained on Ultrafeedback) against baselines (all trained on Ultrafeedback) on **AlpacaEval**, another instruction following benchmark for which we evaluating using GPT-4 Turbo, as suggested by the creators [186]. Hyperparameters and model configurations are given in Appendix A.10.

[22] Similar to [92], we found better generalization in reward learning when our Oracle reward learning loss is regularized with the SFT component (second term in Eq. 3.1) with an α of 0.01.

3.5 Results

Non-deterministic preferences Table 3.2 shows the win rates computed with GPT-4o as judge for 1,000 randomly selected prompts from the CNN/Daily Mail test corpus under OOD settings. We follow similar settings as [9] but further ground the prompt using human-written summaries as reference for GPT to conduct its evaluation (see Section A.11). For a fairer evaluation [9, 187, 188], we swap positions of π_θ-generated summaries y in the prompt to eliminate any positional bias and evaluate the generated summaries on criteria like coherence, preciseness and conciseness, *with the human written summaries explicitly in the prompt to guide evaluation.*

CNN/Daily Mail (GPT-4o as Judge)	
DRDO vs. e-DPO	
$\mathcal{D}_{all}$	78.27%
$\mathcal{D}_{hc,he}$	80.92%
$\mathcal{D}_{\ell c,\ell e}$	79.01%
DRDO vs. DPO	
$\mathcal{D}_{all}$	80.78%
$\mathcal{D}_{hc,he}$	79.11%
$\mathcal{D}_{\ell c,\ell e}$	79.79%
Gold vs. DRDO	52.82%
Gold vs. e-DPO	54.38%
Gold vs. DPO	58.54%
AlpacaEval 2.0 (GPT-4 Turbo as Judge)	
SFT vs. GPT-4 Turbo	8.86%
DPO vs. GPT-4 Turbo	11.01%
e-DPO vs. GPT-4 Turbo	11.65%
PPO vs. GPT-4 Turbo	13.67%
DRDO vs. GPT-4 Turbo	15.95%

Table 3.2: Length-controlled win rates computed for 1,000 randomly selected evaluation samples from the **CNN/Daily Mail Corpus**, and for OOD evaluation on **AlpacaEval 2.0** ($N = 805$) using the Phi-3 model. $\mathcal{D}_{all}$, $\mathcal{D}_{hc,he}$, and $\mathcal{D}_{\ell c,\ell e}$ represent TL;DR training splits. "Gold" refers to human-written reference summaries used in prompts to ground the win rate computations. Both e-DPO and PPO use the same oracle reward as DRDO. See Table A.1 for AlpacaEval dataset-wise breakdown.

For all policies π_θ trained on all three splits of the training data—$\mathcal{D}_{all}$, $\mathcal{D}_{hc,he}$, and $\mathcal{D}_{\ell c,\ell e}$—we compute the win-rates of DRDO vs. e-DPO and DPO to evaluate how each

method performs at various levels of preference types. Across all settings, DRDO policies significantly outperform the two baselines. For instance, DRDO's average win rates are almost 79.4% and 79.9% against e-DPO and DPO, respectively. As we hypothesized, DRDO-aligned π_θ can learn *all* preferences, deterministic and non-deterministic, effectively, and is superior at modeling $\mathcal{D}_{\ell c,\ell e}$ the subset containing more non-deterministic preference samples, without sacrificing performance on the deterministic subset ($\mathcal{D}_{hc,he}$) and overall ($\mathcal{D}_{all}$). This suggests that DRDO is more robust to OOD-settings at various levels of difficulty in learning human preferences.

Reward distillation Section 3.5 shows results from our evaluation of DRDO's reward model distillation framework compared to DPO and e-DPO as well as baseline SFT-trained policies on the Ultrafeedback evaluation data, when compared *across various model parameter sizes* and at varying levels of temperature sampling. We sample π_θ-generated responses to instruction-prompts in the test set using top-p(nucleus) sampling [189] of 0.8 at various temperatures $\in \{0.2, 0.5, 0.7, 0.9\}$.

	Baselines											
	$T = 0.2$			$T = 0.5$			$T = 0.7$			$T = 0.9$		
Policy (Student)	**SFT**	**DPO**	**e-DPO**	**SFT**	**DPO**	**e-DPO**	**SFT**	**DPO**	**e-DPO**	**SFT**	**DPO**	**e-DPO**
OPT-125M	0.47	0.61	0.51	0.52	0.63	0.50	0.57	0.58	0.54	0.57	0.64	0.61
OPT-350M	0.58	0.63	0.60	0.61	0.70	0.65	0.60	0.70	0.69	0.66	0.70	0.68
OPT-1.3B	**0.83**	0.70	**0.83**	**0.82**	0.65	**0.81**	**0.83**	0.63	0.75	**0.81**	0.59	**0.78**
OPT-2.7B	0.81	0.68	**0.83**	0.80	0.64	0.80	0.79	0.62	**0.78**	0.79	0.60	0.75
Phi-3	0.66	**0.80**	0.57	0.71	**0.83**	0.58	0.75	**0.86**	0.60	0.76	**0.88**	0.66

Table 3.3: Average win-rates computed with DRDO's Oracle reward model against all baselines—SFT, DPO and e-DPO—at various diversity sampling temperatures (T). Bolded numbers show the highest win-rates against each baseline at each temperature.

DRDO significantly outperforms competing baselines, especially for larger models in the OPT family. DRDO-trained OPT-1.3B, OPT-2.7B, and Phi-3-Mini-4K-Instruct achieve average win rates of 76%, 74%, and 72%, respectively, across all baselines. This is notable

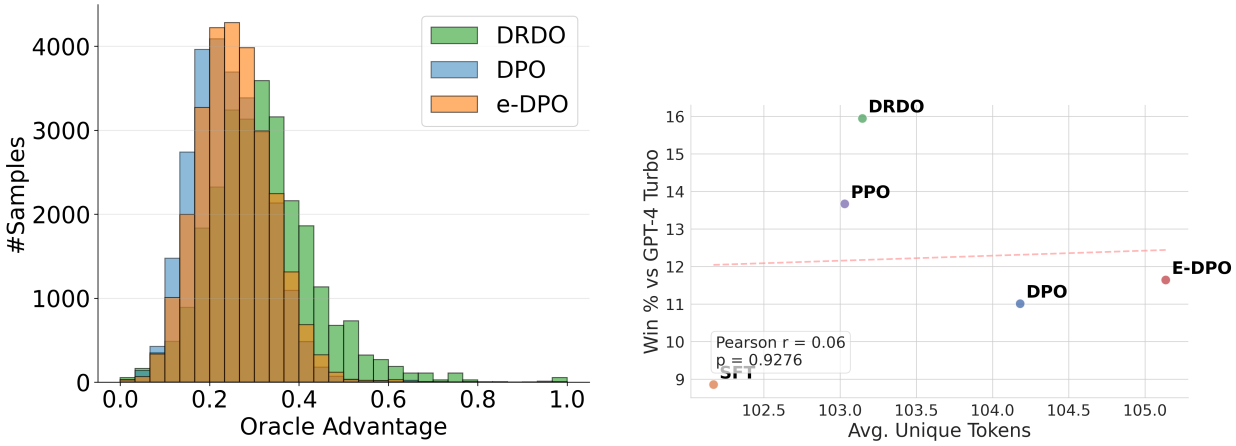
as responses are sampled on unseen prompts, and DRDO's policy alignment is reference-model free. DRDO's robustness to diversity sampling further boosts performance, up to an 88% win rate against DPO with the Phi-3-Mini-4K-Instruct model. At lower temperatures, DRDO's posted gains are more modest. Our results also indicate that performance is correlated with model size, as DRDO policies of the same size as the Oracle (1.3B) show the strongest gains. In smaller models, results are more mixed.[23] DRDO shows moderate improvement and posts smaller gains against SFT models.

3.5.1 Analysis

Using GPT-4o as a judge to approximate true human preferences may be prone to bias, so we further validate our approach by investigating Oracle reward advantage over the above mentioned human written summaries as well as on-policy generations from an SFT-trained model. Figure 3.4a shows the computed expected reward advantages on the CNN/Daily Mail TL;DR evaluation set, sampled according to the method outlined in Section 3.5. Rewards were computed using our Oracle (trained with Phi-3) on these sampled generations and normalized. To compute the advantage, we used human written summaries (Figure 3.4a). DRDO improves performance across various temperature samplings over a baseline SFT policy, and brings in a considerable performance gain over competitive baselines like DPO and ensemble-based e-DPO while also being robust to OOD settings.

Our evaluations vs. "gold" summaries in Table 3.2 also demonstrate where bias may arise in the GPT-4o evaluation (win counts and win rates are shown for gold samples). GPT-4o narrowly prefers generated summaries to human-written ones. One possible reason could be that Reddit TL;DR data is massive and crowd-sourced which naturally results in noisy labels. However, under this experiment too, DRDO-trained policies are better-performing than e-DPO and DPO, by about 1.5–6%. We should note that human

[23] Apart from lower overall model capacity, this may reflect length bias [105, 190], as smaller models averaged more tokens (211.9 vs. 190.8) than policies, closer to Ultrafeedback targets (168.8). See Section A.10.



(a) **(b)**

Figure 3.4: **(a)** Oracle expected reward advantage on CNN/Daily Mail articles. DRDO shifts expected rewards rightward compared to DPO/e-DPO. **(b)** Win rates against GPT-4 Turbo on Alpaca Eval 2.0 vs. mean unique tokens across models.

summaries may contain more implicit diversity than generated summaries, and this may demonstrate the "regression to the mean" effect in LLM generation [191].

On AlpacaEval 2.0 with GPT-4 Turbo as the judge (see Table 3.2 for overall results and Table A.1 in Appendix Section A.3 for detailed dataset-wise breakdown), DRDO achieves the highest overall win rate at 15.95%, outperforming PPO (13.67%), e-DPO (11.65%), DPO (11.01%), and SFT (8.86%). These results reflect DRDO's ability to generalize effectively under length-controlled evaluation, surpassing both offline and on-policy baselines. To ensure that model performance evaluated with GPT-4 Turbo judge does is indeed robust to token-diversity bias [84], we plot the overall win-rates versus mean unique tokens across models in Figure 3.4b. No statistically significant correlation (r = 0.06, p = 0.93) exists between response diversity and win rates, indicating DRDO's superior performance is not attributable to diversity-bias in LLM judges, in addition to length-bias [82] and likely due to actual response quality. More importantly, while PPO is competitive with DRDO, Table 3.2 results suggest that DRDO more effectively leverages the oracle reward model while being fully *offline*, unlike PPO which requires online (on-policy) sampling and drastically increases compute requirements—*even though both methods use the same oracle reward model.*

3.6 Concluding Remarks

In this chapter, we introduced ***Direct Reward Distillation and policy-Optimization (DRDO)***, a principled approach to preference optimization that unifies the reward distillation and policy learning stages into a single, cohesive framework. Unlike popular methods like DPO that rely heavily on implicit reward-based estimation of the preference distribution, DRDO uses an Oracle to distill rewards directly into the policy model, while simultaneously learning from varied preference signals, leading to a more accurate estimation of true preferences. Our experiments on Reddit TL;DR data for summarization as well on instruction-following in Ultrafeedback and AlpacaEval 2.0 suggest that DRDO is not only high-performing when compared head-to-head with competitive methods like DPO and PPO but is also particularly robust to OOD settings. More importantly, unlike traditional RLHF that requires "online" rewards, reward distillation in DRDO is simple to implement, is model-agnostic since it is reference-model free and efficient, since Oracle rewards are easy to precompute.

From a compute perspective, DRDO introduces an explicit trade-off: we must first train an external Oracle reward model before aligning the policy. At first glance, this extra training stage may appear to increase overall cost. However, prior work [48, 192, 193] suggests that investing compute into richer reward signals can pay off in stability and sample efficiency. In our setup, the Oracle need not be loaded in memory during DRDO alignment—its outputs are precomputed once and reused—making the approach attractive when aligning small- to mid-scale models (relative to models such as LLaMA [93]) under tight deployment budgets. DRDO still assumes access to such an Oracle and to sufficient data to validate its quality (e.g., via cross-validation), so its benefits are most pronounced in settings where a reasonably accurate reward model can be amortized across many alignment runs. We did not explore cross-model distillation or multilingual Oracles, but DRDO's reference-model-free formulation and precomputable rewards naturally extend to those regimes.

From Single-Turn Preferences to Multi-Turn Collaboration On a conceptual level, although DRDO attempts to address non-deterministic preferences in single-turn settings, it inherits a fundamental limitation shared by all standard Bradley-Terry-based methods [9, 19, 87]: the assumption that preferences are stationary, context-independent, and immediately observable. In collaborative dialogue between multiple agents, these assumptions break down [25, 194]. Preferences evolve as participants' beliefs align or diverge, intervention quality depends on the current dialogue state, and success emerges over multi-turn trajectories rather than isolated utterances. Moreover, DRDO's insight about non-deterministic preferences—that human judgment contains inherent uncertainty—foreshadows a deeper challenge in collaboration: when multiple agents hold contradictory beliefs about task-relevant propositions, there is no single "preferred" response. In this setting, the optimal action likely depends on best resolving the underlying belief misalignment, not maximizing preference for a particular surface form. As multi-turn interactions unfold, dialogue states and appropriate interventions evolve dynamically as participants introduce new information or revise beliefs, creating trajectories unseen during training. While DRDO addresses the standard form of distribution shift in machine learning (e.g., between training and test data), collaborative settings additionally require robustness to *temporal* distribution shift and to mismatches between human dialogue data and AI-generated synthetic distributions [90]. This suggests that perhaps one way to overcome the problem of dynamic preferences can be to explicitly model evolving beliefs in collaborators. In the next chapter (Chapter 4), we delve deeper into this problem and propose an efficient offline alignment framework specifically designed for LLMs to generate "friction" interventions—LLM outputs that are optimized for the evolving observed belief states of participants in a collaborative dialogue.

Acknowledgments and Funding This material is based in part upon work supported by Other Transaction award HR00112490377 from the U.S. Defense Advanced Research

Projects Agency (DARPA) Friction for Accountability in Conversational Transactions (FACT) program, by the U.S. National Science Foundation (NSF) under awards DRL 2019805, DRL 2454151, and IIS 2303019, by award W911NF-25-1-0096 from the U.S. Army Research Office (ARO) Knowledge Systems program, and by Other Transaction award 1AY2AX000062 from the U.S. Advanced Research Projects Agency for Health (ARPA-H) Platform Accelerating Rural Access to Distributed Integrated Medical Care (PARADIGM) program. Approved for public release, distribution unlimited. Views expressed herein do not reflect the policy or position of the National Science Foundation, the Department of Defense, or the U.S. Government. Computational resources for this research were provided by the CS Department's Falcon cluster, particularly the Carnap and Tarski machines, and Data Science Research Institute's Riviera HPC clusters.

Chapter 4

Frictional Agent Alignment Framework: Slow Down and Don't Break Things

A version of this chapter was published as a conference paper entitled "Frictional Agent Alignment Framework: Slow Down and Don't Break Things" [3] in the Proceedings of the 2025 Meeting of the Association for Computational Linguistics, held in Vienna, Austria.

Overview In the previous chapter we developed a novel preference alignment framework (DRDO) that overcomes some of the limitations that the current batch of popular supervised alignment algorithms [47,48,195]. We saw that when preferences are deterministic, policies aligned with methods like DPO [9] tend to overfit to dispreferred responses leading to its degeneracy at inference time. Furthermore, we saw how DRDO and related algorithms [175,196] are able to learn the innate diversity in human preferences including samples where the preference decision is unclear (non-deterministic samples). Yet, these algorithms have limitations when applied to collaborative dialogues primarily due to their single-turn nature (e.g., RL bandit settings [121] with a horizon of 1) as well their reliance on Bradley-Terry assumptions. While they can achieve state of the art results for simpler tasks like instruction following, summarization etc., collaborative dialogues are inherently multiturn in nature with evolving and dynamics preferences of collaborators due to belief evolution.

In this chapter, our focus is on developing a novel algorithm[24] for aligning an LLM specifically to play the role of a "support agent" in collaborative dialogues. When collaborating to solve problems, humans continually interrogate each other's intentions and assumptions [25,26,28].With the rapid integration of generative AI, exemplified by large

[24]Our code and data can be found at https://github.com/csu-signal/FAAF_ACL.

language models (LLMs), into personal, educational, business, and even governmental workflows, AI systems will increasingly be called upon to act as collaborators with humans; to adequately fill this role, AIs must be able to recapitulate the reflection and deliberation that makes human-human collaboration successful, but also causes temporary slowdowns in dialogue while interlocutors construct a *common ground* on which to collectively reason—the phenomenon we call **friction**. "Friction"[25] in this sense is something that LLMs struggle with. To prompt an interlocutor to reflect upon their assumptions requires that one have an approximate understanding of what those assumptions are and entail [197]. This is predicated upon a *theory of mind* (ToM; [198]), which is likewise a challenge for LLMs [199, 200].

However, using an off-the-shelf LLM for this specific role is challenging since language models are typically not designed to provide friction as they are inherently *text-completion* models [201]. Notably, off-the-shelf LLMs may perhaps be able to generate relevant and positive interventions in some cases. However, this is not established in the literature, especially in tasks where information flow is multimodal and where purely textual transcriptions reflect a partial or sparse information state. Furthermore, chain-of-Thought (CoT) rationales should not be assumed to be faithful by default, as LLM training objectives do not explicitly incentivize accurate causal reporting, human explanations often omit crucial causal links [202, 203] or reflect cognitive biases [204], are frequently designed to persuade [167], or may reflect contradictory beliefs and values from spurious correlations and biases in training data [205, 206]. Even frontier LLMs aligned with RLHF can disincentivize faithfulness due to sycophancy biases and scheming [42, 43] or may be suboptimal for multi-turn collaboration [45]. This implies that employing an LLM directly for this task is insufficient; instead, it necessitates the design of a specialized training methodology capable of accurately estimating the task state at any given timestamp, capturing the participants' beliefs about task-specific propositions (e.g., block weights in Weights Task [7])

[25] See Section 2.3.1 in Chapter 2 for a more detailed definition of friction.

as rendered in a textual medium such as through dialogue transcription. Additionally, successful inteventions in collaborative tasks must address the intra-group communicative dynamics, which may not be evidence from transcriptions alone, especially in multimodal tasks.

To address this need, we propose a novel way of looking at the problem of preference alignment for such multi-party dialogues. Our approach overcomes some of the limitations of the standard alignment algorithms when applied to this setting by grounding an LLM's outputs (friction interventions) to participants' observed beliefs and by reducing dependence on the data distribution which might contain biases. Specifically, we propose the Frictional Agent Alignment Framework (FAAF), to generate precise, context-aware friction that prompts for deliberation and re-examination of existing evidence. In other words, FAAF aligns LLMs to be adept *collaborators* in dialogue-driven tasks. Unlike common preference alignment approaches which focus predominantly on reward differences between textual surface forms to generate the best possible completions as a sequence of actions, FAAF takes a state-driven approach based on the *frictive state* (see Definition 7)—a dynamic natural language representation that integrates task context and the beliefs of participants as they change over time (Figure 4.1). We use this state-wise representation to train "friction agent" models aligned to prompt collaborators toward reflection and deliberation in shared tasks, to help them

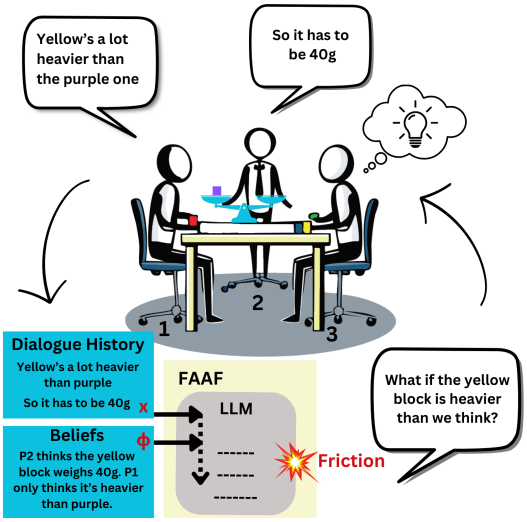


Figure 4.1: FAAF conditions responses on both the dialogue context x and representation of the *frictive state* ϕ, prompting reflection, deliberation, and evidence verification. The figure illustrates participants engaged in the Weights Task [1], where frictional interventions conditioned on participants' observed beliefs encourage them to re-examine assumptions and adjust strategies as disagreements emerge. By explicitly surfacing belief misalignment, these interventions guide participants toward resolving block weights more effectively and aligning on a shared common ground.

resolve conflicting beliefs and assumptions that result in frictive states. From a technical perspective, FAAF's two-player objective decouples from data skew: a frictive-state policy identifies belief misalignments, while an intervention policy crafts collaborator-preferred responses. Our novel contribution is deriving an analytical solution to this objective, enabling training a single policy via an off-policy supervised loss without requiring any online interactions with a simulator or reward models. Experiments on three benchmarks show FAAF outperforms competitors in producing concise, interpretable friction and in OOD generalization. By aligning LLMs to act as adaptive "thought partners"—not passive responders—FAAF advances scalable, dynamic human-AI collaboration. Our results on two challenging collaborative task datasets and variants show that FAAF's belief state conditioning consistently produces output that is more relevant, impactful on the dialogue, and thought-provoking than competing methods. Our key contributions are:

- a novel LLM alignment framework focused on generating outputs to support critical reasoning and assessments in a collaborative task environment;
- an in-depth mathematical and theoretical foundation for the above approach grounded in collaborative task dynamics;
- evaluations on three challenging collaborative task settings that show the advantages of FAAF over competing alignment methods and demonstrate robustness to out-of-distribution (OOD) data.

4.1 Related Work

RLHF-inspired preference alignment in LLMs has become a cornerstone of developing generative AI systems that cater to user preferences [19]. Both "offline" approaches like DPO; [9], Identity Preference Optimization (IPO; [49]) and other supervised methods [47, 48,104,105,107] and "online" methods [11,207] focus predominantly on preference samples often sourced from datasets like Reddit TL;DR [168] or Ultrafeedback [4] for algorithm development. These methods excel in generating summaries or completions that reflect

human preferences including on single-turn human-AI interaction datasets like SGD [208] or MultiWOZ [209,210], but are often ill-equipped to handle the complexities of real-world multiparty interactions, where communication occurs across diverse modalities [211], including sparse and ambiguous spoken dialogues between multiple collaborators [7,8].

A key challenge in these multiparty shared task settings is the scarcity of annotated data [212], particularly where interventions emerge contextually but sparsely [7,8]. While preference data generated with AI feedback is a viable option [213,214], DPO-trained models depend crucially on the sampling or data-generating distribution due to its Bradley-Terry (BT) model of "implicit rewards," limiting their applications to dialogue-driven settings where preferences may be intransitive [126] or change over time. This data-dependence holds even for more sophisticated methods that optimize on human utility [215], discard the BT assumption [49], or use iterative online approaches [207,216,217]. Game-theoretic approaches to reduce this dependence focus on optimizing a "general preference model" [50,109] that does *not* suffer from this data-bias. But these have limited practical application due to their compute-intensive nature, often requiring the storage and computations with intermediate-stage policies during training [110]. In contrast, FAAF avoids this data-dependence by explicitly conditioning policies on belief-misalignment in a specific dual alignment formulation which we derive in a simple "one-step" supervised manner without requiring computations of complicated mixture policies during training. FAAF represents an instance of "frictive policy optimization" (FPO) as argued for by Pustejovsky and Krishnaswamy [218]—specifically an instance of *Friction-Based Preference Pairing* (FPP).

4.2 Problem Formulation

Our main goal of this section is to define the problem setting to efficiently retrieve the optimal policy π_f^*—the ideal intervention agent that is responsible for generating friction interventions in collaborative dialogues. As defined above, let f be a frictive intervention

(utterance) that is not required to contradict any particular belief encapsulated in a frictive state ϕ, and let the human preference probability $\mathcal{P}(f \succ \phi)$ be the probability that an expert annotator (or collaborator) would prefer f over maintaining ϕ, given prior dialogue history, x.

In its simplest formulation within Chain-of-Thought (CoT) approach [139], the friction agent is modeled as a policy distribution π_f that sequentially generates frictive states, sampling $\phi_i \sim \pi_f(\cdot \mid x, \phi_1, \ldots, \phi_{i-1})$, and ultimately producing the final friction intervention $f \sim \pi_f(\cdot \mid x, \phi_1, \ldots, \phi_n)$. Here, ϕ, or the frictive state, consists of sequentially sampled tokens that textually represents the observed beliefs of collaborators in the dialogue, analogous to "thoughts" in standard CoT-based reasoning frameworks [219]. Unlike standard CoT-based alignment, which relies on self-rewarding strategies, we frame friction agent alignment within preference-based RL (PbRL; [122]). Prior work [220] shows CoT frameworks benefit significantly from contrastive signals in preference learning.

Preference-based RL avoids estimating the partition function Our key insight is that to retrieve the optimal policy π_f^*, we can leverage established methods from RLHF and PbRL by formulating the problem as a KL-divergence constrained minimum relative entropy optimization [102], a well-known approach with a closed-form solution [103].

$$J^*_{\text{RLHF}}(\pi_f) = \max_{\pi_f} \mathbb{E}_{f \sim \pi_f}[\mathcal{P}(f \succ \phi \mid x)] - \beta\, D_{\text{KL}}(\pi_f \parallel \pi_{\text{ref}}) \tag{4.1}$$

This formulation (Eq. 4.1)—where J^*_{RLHF} enforces a KL-based "soft"-constraint on the parametric form of π_f^* w.r.t. the reference policy π_{ref}—provides crucial tradeoffs between training stability and balancing exploration vs exploitation. Specifically, J^*_{RLHF} ensures that π_f^* retrieves the best possible preference probabilities for its generated interventions f, as assigned by $\mathcal{P}(f \succ \phi)$, whether over the distribution of preferences encoded in an offline

dataset [9] or from online sampling $\sim \pi_f$ during training [11], while being distributionally close to an "already-good" imitator—the Supervised-Finetuned (SFT) reference model [221].

Notice that unlike standard RLHF, we formulate J^*_{RLHF} such that π^*_f takes the form $\pi^*_f(\cdot \mid \phi, x)$ and is explicitly conditioned on the frictive state ϕ, apart from x. This is intentional since we hypothesize that an ideal friction agent does *not* intervene arbitrarily, which in a collaborative task setting would cause distraction. Rather, it is conditioned to resolve the lack of common ground thereof between human collaborators, *by definition*—as observed in ϕ. While prior work [110, 220] explores preference alignment in LLMs in such CoT-conditioned scenarios, we provide a more principled approach to proving the existence and the uniqueness of π^*_f that J^*_{RLHF} seeks to retrieve. Mathematically,

$$\pi^*_f(f \mid \phi, x) = \frac{\pi_{\text{ref}}(f \mid \phi, x)\, \exp\left(\beta^{-1} p(f \succ \phi \mid x)\right)}{Z^*(\phi, x)}, \tag{4.2}$$

where $Z^* = \sum_{f'} \pi_{\text{ref}} \exp(\beta^{-1} \mathcal{P}(f' \succ \phi | x))$ is the fixed partition function which does not depend on f and can be safely ignored in the optimization of J^*_{RLHF} [9].

This representation of the optimal intervention policy π^*_f (Eq. 4.2) is unwieldy since it contains a partition function term $Z^*(\phi, x)$ in the denominator that normalizes the probabilities of all possible responses. Therefore, while the optimal policy is closed form, the dependence on Z^* makes it practically intractable to estimate the policy directly for LLMs since Z^* is a summation over the set of all possible sequences of tokens in the tokenizer, often requiring methods like importance sampling [222] or ensembling models [223] for an unbiased estimate. Importantly, this problem remains even if the set of friction interventions $\mathcal{F}$ were a restricted subset of the space of all possible actions $\mathcal{Y}$. To overcome this, prior Preference-based RL [122] works suggest supervised learning algorithms for obtaining an optimal policy induced *under the expectation* over a preference dataset. These offline methods, such as DPO [9], IPO [49], or Kahneman-Tversky Optimization (KTO; [215]), either rely on the BT model of preferences [87] where the optimal policy can be induced from a static preference dataset using implicitly-defined pointwise rewards, or assume

that alignment is conducted with access to a non-biased data-generation or "sampling" distribution μ from which π_f^* can be learned using pairwise preferences without adopting a strictly BT assumption [49].[26] These approaches would give us the following formulation for π_f^*:

$$\pi_f^* = \frac{\pi_{\text{ref}} \exp\left(\beta^{-1} \mathbb{E}_{\substack{f \sim \mu(\cdot|x) \\ \phi \sim \mu(\cdot|x)}} \Psi(\mathcal{P}(f \succ \phi \mid x))\right)}{Z^*(\phi, x)}, \tag{4.3}$$

where $\Psi(p)$ is the identity mapping for IPO, and $\log\left(\frac{\mathcal{P}}{1-\mathcal{P}}\right)$ (inverse sigmoid) for DPO and KTO. During model training to retrieve π_f^*, these algorithms remove the dependence on the partition function either by leveraging the BT model of pairwise preferences (like DPO or KTO) or using the idea of symmetry of preferences [50].

Data Bias in PbRL While the practicality of these offline supervised algorithms is a clear advantage, their dependence on preference data selected via sampling is a limitation in reconstructing the human preference probability $\mathcal{P}$. This is particularly true for collaborative dialogue tasks where common ground changes over time, meaning that frictive states at test time may not occur in distributions similar to those in which they occur in the training data, and where participants may not intervene due to variables obscure to a language model, such as not realizing the existence of a frictive state or judging the frictive state to be *non-functional*. Operationally, even if the true underlying preferences ($\mathcal{P}$) of collaborators are transitive and consistent, constructing a preference dataset for use with existing offline training methods is not straightforward; dialogue distributions may be skewed or sparse [7]. When using generative AI to create denser training data, even high-capacity LLMs like GPT-4 are prone to various forms of biases such as toward length [185] or certain linguistic registers. Therefore, the core motivation of FAAF is as follows—**how do we train a high-quality friction agent that can leverage the inherent scalability of offline alignment methods and reconstruct the true underlying preference distribution, while**

[26] By "sampling," we mean those actions that make it to the preference annotation phase after being sampled with the data-generator μ.

still being robust to the data skew that may arise when sampling a preference dataset, whether using generative AI or from real-life collaborative dialogues?

4.2.1 FAAF Objective

In order to address the above problem, we define a novel two-player adversarial optimization objective J^*_{FAAF} (Eq. 4.4) to resolve the above question. Specifically, given a reference model π_{ref} and a regularization parameter $\beta \in \mathbb{R}_+$, our goal is to learn two interdependent "collaborative" policies: (i) a *frictive state policy* π^*_ϕ that generates the most semantically rich frictive states ϕ, capturing tensions or uncertainties (in the form of first-order beliefs of dialog participants) in dialogue, and (ii) a *friction intervention policy* π^*_f that generates constructive interventions f, conditioned on the frictive state, to steer the discourse toward greater clarity and converge onto a common ground between participants. Mathematically,

$$J^*_{\mathsf{FAAF}} = \min_{\pi_\phi} \max_{\pi_f} \mathbb{E}_{\substack{x\sim\rho \\ \phi\sim\pi_\phi(\cdot|x) \\ f\sim\pi_f(\cdot|\phi,x)}} \Big[\mathcal{P}(f \succ \phi \mid x) - \beta D_{\text{KL}}(\pi_f \parallel \pi_{\text{ref}} \mid \phi, x) + \beta D_{\text{KL}}(\pi_\phi \parallel \pi_{\text{ref}} \mid x) \Big]. \tag{4.4}$$

Notice how the optimal intervention policy π^*_f, by definition of the inner max operator, generates interventions that are, on average, most preferred by collaborators, while the first KL-divergence term, defined as $D_{\text{KL}}(\cdot \mid \phi, x)$, stabilizes learning in π^*_f by keeping it closer to a reference model. Compared to a standard RLHF objective, the additional KL term $D_{\text{KL}}(\pi_\phi \parallel \pi_{\text{ref}} \mid x)$ forces the frictive state policy π^*_ϕ to be adversarially robust, in that it must ensure that sampled frictive states $\phi \sim \pi^*_\phi$ cannot be exploited by π^*_f to generate subpar interventions that remain too close to the reference model. Thus, the FAAF objective seeks to retrieve policies that adapt to dialogues over time[27]: the frictive state policy

[27]We can think of this temporal dimension as sequential tokens sampled in the trajectory.

searches for the most immediate tension points or exposes the lack of common ground between task participants, while the intervention policy generates outputs that remain grounded in the particulars of the relevant frictive state (e.g., regarding the correct task items or propositions), and naturally and intuitively prompts for reflection and deliberation on these points. **The key takeaway is that optimal friction interventions should not be arbitrary interventions in the dialogue, but should surface the presuppositions that gave rise to the most logically necessary frictive state, making interventions precise and interpretable.**

4.2.2 Deriving the "Offline" FAAF Objective

FAAF as a single trainable policy While the previous section defines the FAAF objective J^*_{FAAF} in its full game-theoretic form, it is not easily optimizable in practice. The core difficulty lies in the nested expectations, which require sampling both the friction intervention and the frictive state from their respective policies, $\pi_f(\cdot \mid \phi, x)$ and $\pi_\phi(\cdot \mid x)$. This introduces two major challenges for real-world LLM training: (1) repeatedly sampling from two large policies is computationally expensive, and (2) storing and optimizing two separate billion-parameter policies simultaneously is infeasible.

For practical efficiency, we aim to achieve two goals. First, we desire an *offline* learning procedure that can train directly from a dataset of pre-generated interventions and frictive states, without requiring sampling during optimization. Second, we seek a formulation that replaces the two-policy optimization with a *single trainable policy*, while still faithfully capturing the underlying game-theoretic structure. In other words, we require a formulation that expresses the preference between any pair of interventions f_1 and f_2 purely in terms of both the optimal frictive-state policy $\pi^*_\phi(\cdot \mid x)$ and the optimal friction-intervention policy $\pi^*_f(\cdot \mid \phi, x)$. Although prior work [50, 110, 224] has shown that collapsing similar min–max objectives into a single policy—representing the Nash equilibrium [225]—does *not* degrade optimality, these results do not immediately apply to FAAF. For example, unlike standard comparisons which evaluate an intervention f directly

against an alternative f', FAAF evaluates the preference of an intervention *relative to a frictive state* (f vs. ϕ), creating an asymmetry absent in earlier formulations.

Proof sketch: deriving the FAAF empirical objective

- **Inner maximization.** We first isolate and solve the inner max term of the FAAF objective (Eq. 4.4), obtaining a closed-form optimal intervention policy π_f^* (Proposition 2). This establishes existence but is not yet usable, since it does not include the frictive-state policy π_ϕ^*.
- **Substitution and reduction.** Substituting π_f^* back into the objective removes the max term and reduces the problem to a single minimization that is algebraically more tractable. We show the steps involved in Lemma 7 (in Appendix B).
- **Lagrangian derivation.** Using a Lagrangian formulation and results from Lemma 8 and Lemma 10 (also in Appendix B), we derive the optimal frictive-state policy π_ϕ^* and, under preference asymmetry [49, 50], obtain a supervised ℓ_2 loss that serves as the empirical "offline" FAAF alignment objective. Appendix B provides a step-by-step derivation.
- **Uniqueness.** Under the stated assumptions, this empirical objective (Eq. 4.7) admits a unique minimizer (Proposition 3).

Figure 4.2: Steps outlining the derivation of the FAAF offline objective that can be trained with a single LLM acting as the friction intervention policy that conditions its outputs on the dialogue context x and the frictive-state ϕ.

While the data is constructed in a standard pairwise preference format, the FAAF optimization conditions on both the dialogue context x and the textual frictive state ϕ. To obtain an empirically usable offline objective from the two-player formulation in Eq. 4.4, we proceed in three steps. First, we solve the inner maximization analytically, yielding a closed-form optimal intervention policy π_f^* (Proposition 2; see also Appendix B, Lemma 7). Substituting π_f^* back into the objective collapses the original min–max problem into a single minimization, which we then analyze via a Lagrangian approach, using Lemma 8 and Lemma 10 (Appendix B) to derive the optimal frictive-state policy π_ϕ^* and, under preference asymmetry [49, 50], a supervised ℓ_2 loss that defines the offline FAAF alignment

objective (Eq. 4.7). Finally, we show that this empirical objective admits a unique minimizer (Proposition 3), yielding a single-LLM training objective that remains faithful to the original two-player formulation (cf. Figure 4.2).

Proposition 2 (Optimal friction intervention policy)**.** *Fix a dialogue context x, a frictive state ϕ, and a strictly positive parameter $\beta \in \mathbb{R}_+^*$. Let $\mathcal{F}$ denote a finite friction token alphabet, and let $\pi_{ref}(\cdot \mid \phi, x)$ be a reference friction policy. For any fixed π_ϕ, consider the inner maximization in Equation* (4.4)*, given by*

$$\begin{aligned}\mathcal{L}_\beta(\pi_f) &= \max_{\pi_f} \Big[\mathbb{E}_{f\sim\pi_f(\cdot|\phi,x)}\big[p(f \succ \phi \mid x)\big] - \beta\, D_{KL}(\pi_f \parallel \pi_{ref} \mid \phi, x)\Big] \\ &= \max_{\pi_f} \Big[\sum_{f\in\mathcal{F}} \pi_f(f \mid \phi, x)\, p(f \succ \phi \mid x) - \beta\, D_{KL}(\pi_f \parallel \pi_{ref} \mid \phi, x)\Big]. \end{aligned} \tag{4.5}$$

Then there exists a unique optimal friction intervention policy π_f^* (a valid probability simplex satisfying $\sum_f \pi_f(f|\phi, x) = 1$) achieving this maximum, and it is given by the Boltzmann-form update

$$\pi_f^*(f \mid \phi, x) = \frac{\pi_{ref}(f \mid \phi, x)\, \exp\big(\beta^{-1} p(f \succ \phi \mid x)\big)}{Z^*(\phi, x)}, \tag{4.6}$$

where $Z^*(\phi, x) = \sum_{f'\in\mathcal{F}} \pi_{ref}(f' \mid \phi, x)\, \exp\big(\beta^{-1} p(f' \succ \phi \mid x)\big)$ is the corresponding partition function. In particular,

$$\pi_f^* = \arg\max_{\pi_f} \mathcal{L}_\beta(\pi_f).$$

Proposition 2 shows that the friction intervention policy π_f^* admits a closed-form solution in terms of the reference model and an exponentiated preference distribution. While closed-form optimal policies are standard in maximum-entropy learning [102, 103] and in recent LLM alignment methods [9], our result extends this observation to the FAAF setting: a closed-form expression still exists even when the policy is additionally conditioned on frictive states ϕ, and when interventions are selected to be preference-optimal. A full proof is provided in Appendix B.

Importantly, although π_f^* (Eq. B.2) is closed form, it is not fully expressive because it does not contain the optimal frictive-state policy π_ϕ^*. Ideally, we would like a single preference expression that jointly captures both policies. Since π_f^* is available in closed form, we substitute it into the original FAAF two-player objective, eliminating the maximization term and reducing the min–max problem to a single minimization. We then derive π_ϕ^* via a Lagrangian formulation (see Appendix B.1). This yields exactly the representation we seek: a closed-form preference relation between any two interventions f_1 and f_2 expressed analytically in terms of both the optimal friction intervention policy $\pi_f^*(\cdot \mid \phi, x)$ and the optimal frictive-state policy $\pi_\phi^*(\cdot|x)$. This enables a simple supervised (ℓ_2) objective—similar in spirit to IPO [49]—that regresses the predicted preference expression derived from these policies to observed relative preferences $p(f_1 \succ f_2 \mid x)$ (conditioned on ϕ), assuming access to sufficient preference-annotated data. The resulting FAAF empirical alignment objective is then:

Algorithm 2 Frictional Agent Alignment Framework

Require: Training data $\mathcal{D}_\mu$ containing tuples (x, ϕ, f_w, f_l), where x: prompt, ϕ: frictive state, f_w: preferred response, f_l: non-preferred response.

1: Define likelihood ratios:
2: $\Delta R = \log\left(\frac{\pi_\theta(f_w|\phi,x)}{\pi_{\text{ref}}(f_w|\phi,x)}\right) - \log\left(\frac{\pi_\theta(f_l|\phi,x)}{\pi_{\text{ref}}(f_l|\phi,x)}\right)$
3: $\Delta R' = \log\left(\frac{\pi_\theta(f_w|x)}{\pi_{\text{ref}}(f_w|x)}\right) - \log\left(\frac{\pi_\theta(f_l|x)}{\pi_{\text{ref}}(f_l|x)}\right)$
4: Loss function: $\mathcal{L} = \mathbb{E}_{\mathcal{D}_\mu}[(1 - \beta(\Delta R + \Delta R'))^2]$
5: Gradient update: $\nabla_\theta \mathcal{L} = \mathbb{E}_{\mathcal{D}_\mu}[-2\beta\delta\nabla_\theta \log(\Delta R \cdot \Delta R')]$, where $\delta = 1 - \beta(\log \Delta R + \log \Delta R')$
6: Update policy parameters θ using gradient descent

$$\hat{\mathcal{L}}(\pi_\theta) = \mathbb{E}_{(x,\phi,f_w,f_l)\sim\mathcal{D}_\mu}\Bigg(1 - \Bigg[\underbrace{\beta \log\left(\frac{\pi_\theta(f_w|\phi,x)\pi_{\text{ref}}(f_l|\phi,x)}{\pi_\theta(f_l|\phi,x)\pi_{\text{ref}}(f_w|\phi,x)}\right)}_{\Delta R} + \underbrace{\beta \log\left(\frac{\pi_\theta(f_w|x)\pi_{\text{ref}}(f_l|x)}{\pi_\theta(f_l|x)\pi_{\text{ref}}(f_w|x)}\right)}_{\Delta R'} \Bigg] \Bigg)^2 \tag{4.7}$$

where ΔR and $\Delta R'$ represent implicit reward differences [9, 49], the former being explicitly conditioned on the frictive state ϕ, with no such conditioning on the latter.

Notably, this objective is optimized by a *single* parametrized policy that leverages the inherent expressivity of LLMs and induces a unique global minimum in the space of policies. We provide an argument to justify the uniqueness of the global minimum of the FAAF empirical objective in Proposition 3, drawing on a similar argument made in IPO [49]. Algorithm 2 provides a step-by-step FAAF training algorithm.[28] Additionally, unlike the standard RLHF objective that assumes a Bradley-Terry model of preferences (which is also used to derive the DPO objective [9]), the FAAF loss contains no sigmoid term and therefore avoids the unbounded logit (via sigmoid gating) problem [49] that causes policy degeneracy issues we saw in Chapter 3.

Proposition 3 (Uniqueness of the FAAF Empirical Optimum)**.** *Let μ be the sampling distribution over intervention pairs such that* $\mathrm{Supp}(\mu) = \mathrm{Supp}(\pi_{ref})$*, and let the empirical* FAAF *loss be*

$$\widehat{\mathcal{L}}(\pi) = \mathbb{E}_{(f,f')\sim\mu}\Big[(1 - \beta(\Delta s_\phi + \Delta s))^2\Big].$$

Then the empirical FAAF *objective admits a* unique *minimizer in the policy space Π, up to additive constants on the logit parameterizations of $\pi(f \mid \phi)$ and $\pi(f)$ that do not affect the induced probability distributions.*

Proof. Assume by contradiction that there exist two distinct optimal policies $\pi_A, \pi_B \in \Pi$. By their definition, $\hat{\mathcal{L}}(\pi_A) = \hat{\mathcal{L}}(\pi_A) = 0$ as π_A and π_B are global minima. Consider

[28] For compactness reasons here we represent all policies π as parameterized by weights θ. Similarly to approaches such as [110], because we formulate two distinct policies with the preference equation, we can empirically enforce it using ℓ_2 loss and learn it with a single expressive policy parameterized by θ.

(s_ϕ^A, s^A) and (s_ϕ^B, s^B) as their respective logit parameterizations where:

$$\pi_k(f|\phi) = \frac{\exp(s_\phi^k(f))}{\sum_{f'} \exp(s_\phi^k(f'))}$$
$$\pi_k(f) = \frac{\exp(s^k(f))}{\sum_{f'} \exp(s^k(f'))} \quad \text{for } k \in \{A, B\}$$

where $\pi_k(f|\phi)$ and $\pi_k(f)$ are the ϕ-conditioned and ϕ-unconditioned policies.

By the structure of our FAAF loss from Equation (B.29):

$$\hat{\mathcal{L}}(\pi) = \mathbb{E}_{f,f' \sim \mu}\Big[\big(1 - \beta(\Delta s_\phi + \Delta s)\big)^2\Big] \geq 0$$

Notice that adding a constant c to all logits of s_ϕ or logits of s (directionally denoted as the $(c, \ldots, c) \in \mathbb{R}$) does not affect either policy probabilities due to softmax normalization. For $\hat{\mathcal{L}}(\pi)$, this is the *only* direction where the loss function might not be strictly convex. Outside of these directions, any change in the logits would increase $\mathcal{L}(\pi)$ with strict convexity as a consequence for $\alpha \in (0, 1)$, implying:

$$\begin{aligned}\hat{\mathcal{L}}(\alpha\pi_1 + (1-\alpha)\pi_2) &< \alpha\hat{\mathcal{L}}(\pi_1) + (1-\alpha)\hat{\mathcal{L}}(\pi_2) \\ &= \alpha(0) + (1-\alpha)(0) = 0\end{aligned}$$

where the equality follows from π_1, π_2 being global minima, by definition. This contradicts the non-negativity of $\hat{\mathcal{L}}$, which proves the uniqueness of the FAAF objective. □

4.2.3 Dataset Annotation and Generation

In Section 4.1, we discuss why common preference optimization datasets such as Ultrafeedback, Reddit TL;DR, SGD, or MultiWOZ are not appropriate for FAAF's collaborative task use case. Therefore, we consider two collaborative task datasets to evaluate

FAAF—DeliData [8] and the Weights Task Dataset (WTD; [1]). Notably, labeled instances of frictive states and friction interventions are limited in these datasets, and our human annotations of WTD resulted, on average, of four naturally occurring friction interventions per group. Similarly, for DeliData, we found that friction interventions in the form of "probing questions" are relatively rare, averaging only 3.46 interventions per group across 17,110 total utterances, underscoring the sparsity of explicit deliberation prompts in natural collaboration. These datasets also exemplify the data sparsity problem with deliberation and friction in collaboration. For more details on the raw datasets and the relevant tasks, please check Section 2.5.

Training Dataset Construction While these collaborative datasets provide rich, situationally grounded dialogues, they contain very few naturally occurring friction interventions as discussed above. This extreme sparsity covers only a tiny fraction of the possible frictive states implied by the combinatorics of the task space, and thus motivates the need for controlled data augmentation to obtain sufficiently diverse preference datasets for training and evaluation. Another challenge is structural: the dialogues themselves are long and highly interactive (WTD alone spans nearly three hours of audiovisual material [1]), in sharp contrast to typical preference-alignment datasets in which samples are short (e.g., single responses to instruction prompts or summaries). At the utterance level, however, these collaborative corpora still contain only thousands to tens of thousands of annotated turns—roughly comparable in scale to datasets such as UltraFeedback [4]—highlighting the distinct nature of collaborative alignment data.

To address sparsity, we constructed augmented training datasets using GPT-4o as a high-capacity sampling distribution μ. Following a *self-rewarding* paradigm [214], GPT-4o was prompted to simultaneously (i) identify candidate frictive states, (ii) generate multiple friction interventions (with rationales), and (iii) assign preference-like scores that induce an implicit ranking. For each dialogue, we supplied sliding windows of h preceding utterances (with $h = 15$ for DeliData and $h = 10$ for WTD) together with colloquial explanations

of our friction definitions.[29] From these generations, we constructed contrastive training tuples by pairing "winning" and "losing" interventions (f_w, f_l) conditioned on dialogue history x and frictive state ϕ, yielding preference datasets of the form (x, ϕ, f_w, f_l) for each collaborative task. These augmented datasets allow us to evaluate alignment methods under realistic conversational structure while maintaining sufficient intervention diversity to support meaningful preference learning.

For each dataset, we conducted additional task-appropriate augmentation. For DeliData, we constructed alternative tuples where the specific cards mentioned in the original data were replaced with other cards of the same classes that preserved the relevant rule (i.e., replacing even numbers with other even numbers, consonants with other consonants, etc.). This resulted in 68,618 preference samples for training, with average μ-assigned reward for preferred samples of 8.03 and for dispreferred samples of 3.96 (out of 10). We held out 50 randomly-sampled dialogues for testing.

Since the original WTD contains only 10 dialogues, holding one or two out for evaluation would adversely impact the data distribution. Therefore we used Shani et al. [124]'s method to generate novel simulated collaborative conversations about the Weights Task, providing a task descriptions and ground-truth values for the weights. GPT-4o was prompted to role-play personality-facet combinations from the Big 5 personality types [6], and for each labeled frictive state ϕ we generated and scored 6 friction interventions. This resulted in two distinct versions of the WTD preference dataset. The **Simulated WTD** friction dataset consisted of 56,698 training preference samples, with mean scores of 8.48 (preferred interventions) and 6.01 (dispreferred). 54 dialogues were held out for testing. The **Original WTD** friction dataset contains 4,299 preference samples (preferred mean score 8.36, dispreferred 6.35). These were *all* retained for an OOD evaluation of FAAF trained on the Simulated WTD data. See Appendix B.2 for specifics of data generation

[29] The prompts used to render frictive states into plain language are provided in Figure B.1 and Figure B.2 in Appendix B.2.

and Appendix B.6 for a distributional analysis of the Original and the Simulated WTD data.

Human Validation We conducted a human evaluation to validate the quality of the GPT-generated friction intervention on a random representative subset of 50 pairwise samples each from both the DeliData and WTD generated test datasets.[30] For each sample, 2 annotators[31] were asked to choose which of the two candidate interventions was more appropriate for provoking participants' reflection to help them advance in their task without being given the solution. Average Cohen's κ on WTD samples was 0.58 and on DeliData samples was 0.92, indicating substantial to near complete agreement on which was the better intervention, and indicates that the preferred/dispreferred friction distinction sourced from GPT-4o as μ aligns with human judgments. Appendix B.5 contains additional details on our human evaluation of friction interventions across dimensions such as reasoning quality, specificity and their thought-provoking nature.

4.3 Experimental Setup

Research Questions. Our experiments are designed to evaluate the following:

RQ1: Preference-aligned performance. How do friction interventions generated by LLMs trained with FAAF compare to those generated by baseline methods when evaluated using a preference model (e.g., a high-capacity LLM judge), relative to a fixed SFT reference model?

RQ2: Head-to-head policy comparison. How does FAAF compare to competing alignment policies (e.g., DPO-trained models) when evaluated head-to-head using a trained reward model?

[30] WTD samples include both Original and Simulated interventions.

[31] Our two human annotators have the following demographic breakdown: both male, college undergraduates, one Caucasian, one African, both fluent English speakers.

RQ3: Generalization to out-of-distribution settings. How well can FAAF generalize to real human dialogue transcripts, even though all models were trained only on AI-generated synthetic data? If so, to what extent compared to competing approaches?

Training Setup and Baselines We use `Meta-Llama-3-8B-Instruct` [93][32] for *all* experiments including baselines. All aligned models received exposure to the frictive state annotations during training to ensure fair comparisons on the friction intervention task. For an in-depth evaluation of FAAF's capabilities, we include comparisons to the Supervised-Finetuned (SFT) model as well as the base instruct model generations in our experiments. For "offline" contrastive approaches to compare to, we choose DPO [9] and IPO [49] and for "online" approaches, we include a Proximal Policy Optimization (PPO; [11]) baseline. For SFT, we employ rejection sampling [115] to maximize the likelihood of interventions that receive high rewards under μ. For SFT, DPO, and IPO, the respective losses are computed only on the output tokens and frictive states ϕ, excluding dialogue context tokens. This training approach ensures that the models learn to generate effective interventions while maintaining contextual understanding. For PPO, we train an OPT 1.3B [182] reward model on each dataset using a standard Bradley-Terry loss [19] over preference pairs. For *ablations*, we consider variants of FAAF that ablate the different likelihood ratios—$\texttt{FAAF}_{\Delta R}$ keeps *only* the ϕ-conditioned implicit rewards in the FAAF objective (line 2 in Algorithm 2), and $\texttt{FAAF}_{\Delta R'}$ removes ϕ-conditioning (keeping only line 3). At inference time, all models only receive the dialogue history up until the point at which an intervention is generated, and receive no look-ahead. Appendix B.7 provides more details on training and hyperparameters.

Evaluation Strategies As LLM generation is open-ended, we employ an LLM-as-a-judge (using GPT-4o) "win-rate" evaluation method where a high-capacity model is prompted to select its preference, given two completions, and conducts a multidimensional evaluation of its preferences. First, we sampled friction interventions from all competing models

[32]https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

on 500 randomly sampled prompts from the **DeliData**, **Simulated WTD** and **Original WTD** test sets. Next, we conducted two evaluations using said completions, one with a **preference-model** [50] and another with a **reward-model** [104]. Since GPT-4o also served as the data generation distribution μ, preference-model evaluation compares the two presented choices and nothing else in the data, mitigating lingering bias toward μ [50].

Within preference-based evaluation settings, we adopt the framework proposed by [4] to retrieve utility scores across seven friction dimensions, building on insights from Chen et al. [226]. Specifically, we assess *relevance* and *alignment with rationale and golden samples*[33] to determine how well a friction intervention aligns with surface-level semantics. Meanwhile, *actionability*, *specificity*, *thought-provoking*, and *impact* measure its expected long-term influence on behavior, reasoning, and decision-making. The LLM-judge assigns Likert-type scores across these dimensions, providing a fine-grained evaluation of task-specific preference desiderata. These scores are collected in a pairwise fashion where π_θ-generated interventions f_i from a baseline are compared with π_{ref}-generated counterparts, f_j. We positionally swap these interventions in the evaluation prompt (Figure B.4 in Appendix B.4) for each API call and average the scores for each of the seven dimensions over two runs to mitigate positional bias [185] in computing the final win rates. Specifically, for any pair of interventions (f_i, f_j), let $s(x, f_*)$ denote the score estimate[34] for intervention f_* given context x. The win-rate percentage for a run is computed as $100 \times \frac{1}{N} \sum_{m=1}^{N} \mathbf{1}\{s(x^{(m)}, f_i^{(m)}) > s(x^{(m)}, f_j^{(m)})\}$, where N is the total number of samples, and $x^{(m)}$ represents the context of the m^{th} sample. See Table 4.1.

The above evaluation tests for preference alignment advantage of the aligned model over π_{ref}. For a more robust evaluation, we compare `FAAF`'s generations "head-to-head" against all baselines. Here, instead of the preference model, we utilize the trained OPT 1.3B

[33] A subset of these gold friction interventions was used for human evaluations (see Section 4.2.3).

[34] In this evaluation, "overall" (first column in Table 4.1) is computed based on the judge's choice of winner *after* rating all other dimensions. As such, $s(x, f_*)$ represents scores over these fine-grained friction preference desiderata, and "overall" does not necessarily represent an average or aggregate of the other dimensions but rather a binary judgment based on them.

Reward Model (RM) as described in our PPO training setup. These pointwise estimates of rewards provide a more accurate assessment of the advantage provided by FAAF proposed approach when directly pitted against other alignment baselines. Specifically, we compare $\text{FAAF}_{\Delta R}$, $\text{FAAF}_{\Delta R'}$ as well as our full objective baseline ($\text{FAAF}_{\Delta(R+R')}$) against all chosen baselines. We compute the reward accuracy (or win-rates) similarly and report our results in Table 4.2.

Policy	Overall	Ac	Ga	Im	Rf	Re	Sp	Th
				DELIDATA				
PPO	$68.9_{\pm 1.5}$	$59.9_{\pm 1.5}$	$65.4_{\pm 1.5}$	$68.6_{\pm 1.5}$	$64.9_{\pm 1.5}$	$65.1_{\pm 1.5}$	$71.1_{\pm 1.4}$	$64.0_{\pm 1.5}$
IPO	$70.1_{\pm 1.4}$	$61.2_{\pm 1.5}$	$65.7_{\pm 1.5}$	$69.3_{\pm 1.5}$	$65.3_{\pm 1.5}$	$65.5_{\pm 1.5}$	$72.1_{\pm 1.4}$	$64.1_{\pm 1.5}$
DPO	$70.8_{\pm 1.4}$	$61.0_{\pm 1.5}$	$66.8_{\pm 1.5}$	$69.6_{\pm 1.5}$	$66.1_{\pm 1.5}$	$67.5_{\pm 1.5}$	$72.2_{\pm 1.4}$	$66.2_{\pm 1.5}$
FAAF	$\mathbf{75.7}_{\pm 1.4}$	$\mathbf{65.6}_{\pm 1.5}$	$\mathbf{69.5}_{\pm 1.5}$	$\mathbf{75.0}_{\pm 1.4}$	$\mathbf{72.0}_{\pm 1.4}$	$\mathbf{71.1}_{\pm 1.4}$	$\mathbf{75.3}_{\pm 1.4}$	$\mathbf{70.4}_{\pm 1.4}$
				WTD ORIGINAL				
PPO	$76.0_{\pm 4.3}$	$74.0_{\pm 4.4}$	$75.0_{\pm 4.3}$	$75.0_{\pm 4.3}$	$67.0_{\pm 4.7}$	$70.0_{\pm 4.6}$	$73.0_{\pm 4.4}$	$74.0_{\pm 4.4}$
IPO	$82.0_{\pm 3.8}$	$87.0_{\pm 3.4}$	$75.0_{\pm 4.3}$	$84.0_{\pm 3.7}$	$75.0_{\pm 4.3}$	$80.0_{\pm 4.0}$	$88.0_{\pm 3.2}$	$78.0_{\pm 4.1}$
DPO	$89.0_{\pm 3.1}$	$\mathbf{92.0}_{\pm 2.7}$	$82.0_{\pm 3.8}$	$89.0_{\pm 3.1}$	$84.0_{\pm 3.7}$	$87.0_{\pm 3.4}$	$\mathbf{89.0}_{\pm 3.1}$	$79.0_{\pm 4.1}$
FAAF	$\mathbf{90.9}_{\pm 2.9}$	$81.8_{\pm 3.9}$	$\mathbf{84.8}_{\pm 3.6}$	$\mathbf{90.9}_{\pm 2.9}$	$\mathbf{86.9}_{\pm 3.4}$	$\mathbf{89.9}_{\pm 3.0}$	$88.9_{\pm 3.1}$	$\mathbf{90.9}_{\pm 2.9}$
				WTD SIMULATED				
PPO	$73.6_{\pm 1.5}$	$69.7_{\pm 1.5}$	$64.9_{\pm 1.6}$	$74.2_{\pm 1.5}$	$67.6_{\pm 1.6}$	$71.9_{\pm 1.5}$	$78.1_{\pm 1.4}$	$78.3_{\pm 1.4}$
IPO	$83.0_{\pm 1.3}$	$74.8_{\pm 1.4}$	$78.4_{\pm 1.4}$	$82.9_{\pm 1.3}$	$76.9_{\pm 1.4}$	$81.4_{\pm 1.3}$	$82.5_{\pm 1.3}$	$83.2_{\pm 1.2}$
DPO	$82.9_{\pm 1.3}$	$80.4_{\pm 1.3}$	$75.8_{\pm 1.4}$	$81.3_{\pm 1.3}$	$72.9_{\pm 1.5}$	$76.3_{\pm 1.4}$	$80.2_{\pm 1.3}$	$79.2_{\pm 1.4}$
FAAF	$\mathbf{91.5}_{\pm 0.9}$	$\mathbf{87.5}_{\pm 1.1}$	$\mathbf{87.1}_{\pm 1.1}$	$\mathbf{90.1}_{\pm 1.0}$	$\mathbf{82.0}_{\pm 1.3}$	$\mathbf{85.1}_{\pm 1.2}$	$\mathbf{90.3}_{\pm 1.0}$	$\mathbf{90.1}_{\pm 1.0}$

Table 4.1: Win-rates (%) against the SFT model (π_{ref}) for all alignment methods on sampled interventions (temperature of 0.7, top-p of 0.9) from 500 randomly-sampled prompts from DeliData and WTD evaluation sets, according to GPT-4o. Metrics: **Ac** (*Actionability*), **Ga** (*Gold-alignment*), **Im** (*Impact*), **Rf** (*Rationale-fit*), **Re** (*Relevance*), **Sp** (*Specificity*), and **Th** (*Thought-provoking*). The LLM-as-a-judge evaluation follows [4]. Average win rates are reported over two runs, with positional swapping to mitigate position bias.

4.4 Results and Analysis

Table 4.1 shows that in the eyes of the the LLM-judge, FAAF models have a consistently greater advantage over the SFT model π_{ref} than other baselines across the 7 preference dimensions and overall, noting that when competitor methods beat FAAF, the advantage is

Dataset	Policy	Win-rate **vs. Base**	Win-rate **vs. SFT**	Win-rate **vs. DPO**	Win-rate **vs. IPO**	Win-rate **vs. PPO**
DeliData	$\texttt{FAAF}_{\Delta R'}$	$82.2_{\pm 1.7}$	$78.8_{\pm 1.8}$	$74.0_{\pm 1.9}$	$53.6_{\pm 2.2}$	$\mathbf{79.2}_{\pm 1.8}$
	$\texttt{FAAF}_{\Delta R}$	$85.8_{\pm 1.5}$	$81.4_{\pm 1.7}$	$73.2_{\pm 1.9}$	$54.2_{\pm 2.2}$	$73.4_{\pm 1.9}$
	$\texttt{FAAF}_{\Delta(R+R')}$	$\mathbf{86.2}_{\pm 1.5}$	$\mathbf{84.0}_{\pm 1.6}$	$\mathbf{75.6}_{\pm 1.9}$	$\mathbf{79.6}_{\pm 1.8}$	$76.0_{\pm 1.9}$
WTD Orig.	$\texttt{FAAF}_{\Delta R'}$	$78.0_{\pm 5.8}$	$\mathbf{78.0}_{\pm 5.8}$	$\mathbf{76.0}_{\pm 6.0}$	$58.0_{\pm 6.9}$	$58.0_{\pm 6.9}$
	$\texttt{FAAF}_{\Delta R}$	$68.0_{\pm 6.5}$	$74.0_{\pm 6.2}$	$72.0_{\pm 6.3}$	$62.0_{\pm 6.8}$	$70.0_{\pm 6.4}$
	$\texttt{FAAF}_{\Delta(R+R')}$	$\mathbf{84.0}_{\pm 5.1}$	$76.0_{\pm 6.0}$	$74.0_{\pm 6.2}$	$\mathbf{74.0}_{\pm 6.2}$	$\mathbf{82.0}_{\pm 5.4}$
WTD Sim.	$\texttt{FAAF}_{\Delta R'}$	$79.1_{\pm 1.9}$	$80.2_{\pm 1.8}$	$70.4_{\pm 2.1}$	$68.6_{\pm 2.1}$	$60.8_{\pm 2.3}$
	$\texttt{FAAF}_{\Delta R}$	$85.7_{\pm 1.6}$	$80.8_{\pm 1.8}$	$70.8_{\pm 2.1}$	$72.2_{\pm 2.1}$	$74.8_{\pm 2.0}$
	$\texttt{FAAF}_{\Delta(R+R')}$	$\mathbf{88.0}_{\pm 1.5}$	$\mathbf{83.7}_{\pm 1.7}$	$\mathbf{72.8}_{\pm 2.0}$	$\mathbf{73.7}_{\pm 2.0}$	$\mathbf{75.1}_{\pm 2.0}$

Table 4.2: Win rates of of FAAF variants—$\texttt{FAAF}_{\Delta R'}$ (not ϕ-conditioned), $\texttt{FAAF}_{\Delta R}$ (ϕ-conditioned), and $\texttt{FAAF}_{\Delta(R+R')}$ (full objective)—against competing methods in pairwise comparisons (temperature of 0.7, top-p of 0.9). All alignment baselines are SFT-initialized and `Meta-Llama-3-8B-Instruct` is used as Base.

usually within the margin of error. For instance, in "overall" preference on the DeliData test samples, FAAF achieves a 75.7% win-rate over π_{ref}, surpassing PPO (68.9%), DPO (70.8%), and IPO (70.1%). On the WTD datasets, win rates for all models are higher, reflecting π_{ref}'s weakness with the underspecified nature of WTD dialogues; alignment on this data has a greater net effect on win-rates than the generally less ambiguous DeliData. On WTD FAAF is a clear all-around winner, at 90.9% (vs. DPO's 89.0% and 82.0%) and 91.5% (vs. DPO's 82.9% and IPO's 83.0%) on the Original and Simulated WTD datasets, respectively.

We find that FAAF's win-rates on dimensions such as *actionability* and *gold-alignment* are somewhat lower compared to other dimensions—possibly reflecting that multiple kinds of interventions may be appropriate in context. However, across dimensions like *thought-provoking* and *rationale-fit* we find that FAAF improves 5-6%, or even up to 12% over equivalent PPO, IPO, and DPO win-rates.

PPO's win-rates consistently lag across all datasets (this is particularly pronounced on the WTD data), indicating the challenge that the dimensions of friction pose for a standard approach. DPO is typically FAAF's closest competitor against π_{ref}, with the narrowest average gap in win-rates.

Robustness to OOD Generalization FAAF maintains superior performance on the **Original WTD** dataset (Overall: +1.9% over DPO, +8.9% over IPO, and +14.9% over PPO).

No model was explicitly aligned to this data, and so this result shows FAAF's robustness to OOD settings compared to other approaches. This is particularly noteworthy given that the Original WTD dataset comprises word-for-word transcriptions of actual human dialogues—with disfluencies, sentence fragments, etc.—which differs markedly from the grammatical, structured text typically found in LLM training data or the preference pair samples in the **Simulated WTD** data. That FAAF generalizes well to organic human data provides a strong basis of confidence that a FAAF-aligned agent, jointly-conditioned on the dialogue transcript and frictive state rendering ϕ, could effectively intervene in and mediate real collaborations, where dialogues are often sparse, informal, and structurally distinct from LLM-generated text [227].

DPO—although also optimized against ϕ as part of the context (Section 5.4)—suffers from the Longest-Common-Subsequence problem [47][35] due to the Bradley-Terry preference model assumption where dependence on the context via DPO's log-partition term is effectively canceled in gradient estimates. In contrast, FAAF's combined ΔR and $\Delta R'$ regularization (Algorithm 2) avoids missing such signals in its learning, thereby allowing it to capture more nuanced human preferences.

While the multidimensional analysis on dimensions such as *impact* and *actionability* does not test actual human responses to the sampled intervention, it does maintain evaluation under consistent conditions without creating counterfactual branching due to responses generated under the influence of interventions, which would create divergent evaluation conditions.

Does ϕ-conditioning help FAAF learn more accurate preferences? Table 4.2 shows results from the trained OPT-1.3B RM's evaluation of the full $\texttt{FAAF}_{\Delta(R+R')}$ objective and its ablated variants—ϕ-conditioned $\texttt{FAAF}_{\Delta R}$ and unconditioned $\texttt{FAAF}_{\Delta R'}$—"head-to-head" against all baselines, including the base `Meta-Llama-3-8B-Instruct` model. Across the three datasets,

[35]The LCS issue in DPO, where gradient signals from tokens shared by winning and losing responses are ignored, is well-studied [46, 47, 220].

FAAF win-rates computed with pointwise reward estimates on sampled interventions exceed 80%, on average, against the base and SFT models, consistent with prior work [104]. We also find that while explicit conditioning on ϕ provides clear advantages (e.g., +6.6% vs. Base on Simulated WTD, +14% vs. PPO), and even the unconditioned version consistently wins over baselines, neither term alone achieves the robust performance of $\text{FAAF}_{\Delta(R+R')}$.

Both IPO and FAAF use a squared ℓ_2 loss. IPO's performance against FAAF's ablations suggests that this structural similarity makes it more competitive with FAAF (FAAF ablations beat IPO 53.6% and 54.2% on DeliData and 58.0% and 62.0% on Original WTD, compared to anywhere from a 68–85% win rate against the Base model). In general, these ablations demonstrate that neither variant alone is sufficient. $\text{FAAF}_{\Delta(R+R')}$ (the full objective) shows consistently stronger performance against IPO (79.6% on DeliData, 73.7% on WTD Sim., 74.0% on WTD Orig.) while maintaining high win rates across other baselines ($\sim$81% vs Base/SFT, $\sim$74% vs. DPO). These results, in light of the trends observed previously in OOD evaluation, suggest that while $\text{FAAF}_{\Delta R}$ learns rich ϕ-conditioned preferences, the additional regularization term $\Delta R'$ enables better reward space exploration and generalized preference learning. The combination is crucial for robust performance.

FAAF's Fallback Mechanism and Robustness to Data Skew Note that we do not claim that the policy trained with FAAF's empirical loss will have no dependence on the data distribution (whether sampling frequency or quality-wise) at all. In offline settings where the preference dataset is bound to a finite sample, there will of course be some data bias that affects the learned empirical model. As shown in line 4 of FAAF objective in Algorithm 2, FAAF is robust to possible data-skew because it combines two types of implicit rewards (ΔR and $\Delta R'$) to create a more robust learning signal, using terms that contain two types of conditioning: $\pi(f|\phi, x)$ and $\pi(f|x)$. This reduces the impact of specific examples appearing more frequently or with better quality in the training data. For instance, consider training pairs where the frictive state ϕ is not articulated well in the training data but the

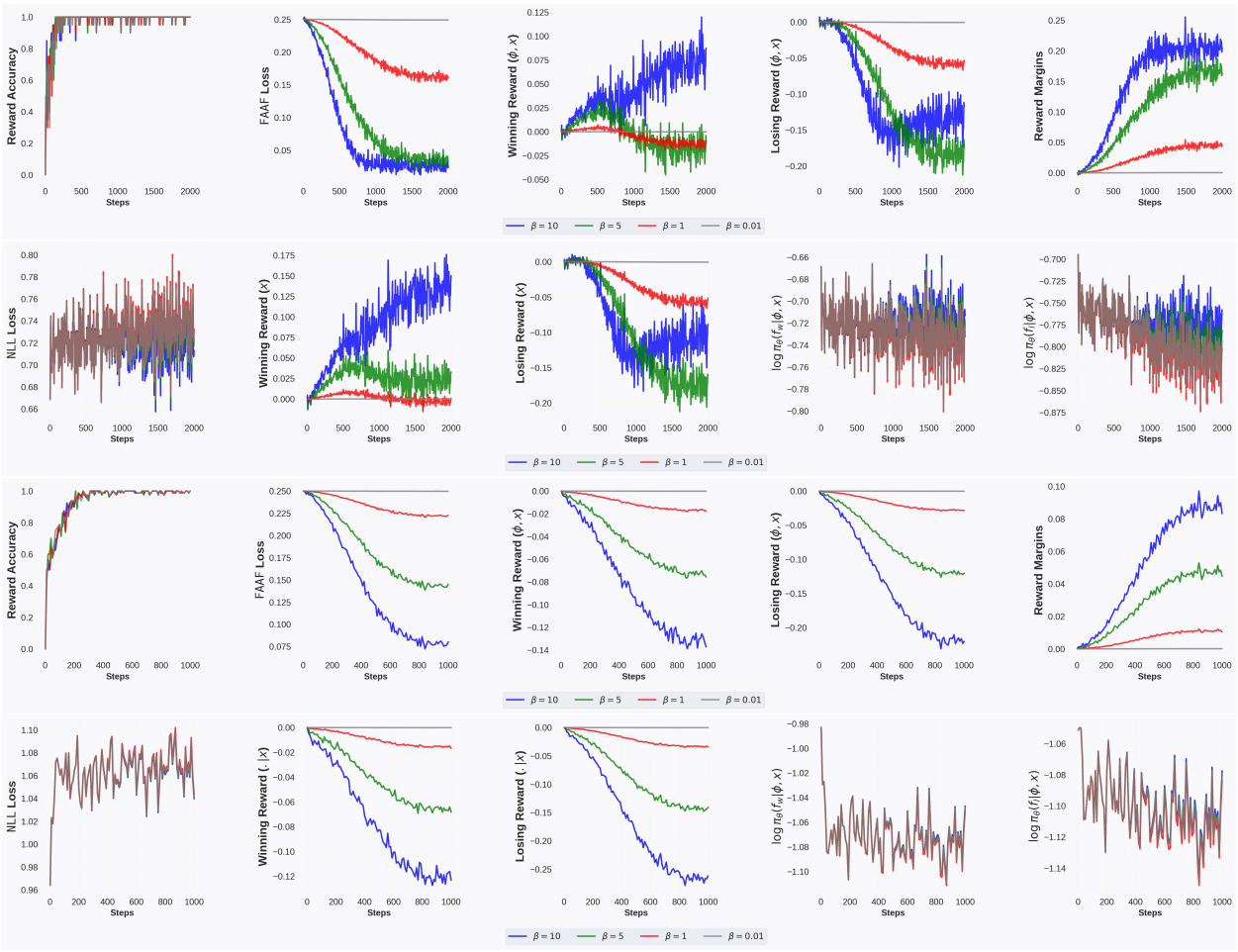


Figure 4.3: Ablation study of FAAF's β hyperparameter ($\beta \in \{10, 5, 1, 0.01\}$) during training on the Simulated WTD data (top half) and DeliData datasets (bottom half) across 2k and 1k training steps respectively. Higher β values (e.g., $\beta = 10$) show better implicit reward estimation as shown in `Reward Accuracy` plots and estimated preference-strengths (`Reward Margins`), while very small values ($\beta = 0.01$) fail to distinguish preferences effectively. $\beta = 10$ also minimizes NLL and FAAF losses suggesting model stability and better convergence. Note that our reported results with FAAF models trained use $\beta = 10$ in Table 4.1 and Table 4.2.

intervention quality is high. These cases could emerge during training on offline LLM-generated data, given the nature of the collaborative tasks studied here, where belief-states are inferred from observable utterances. In such cases, the implicit reward term $\pi_\theta(f_w|x)$ will continue to boost the likelihood of effective interventions, even though the direct contribution of ϕ is effectively ignored in the gradient from ΔR term due to high string similarity [101]. Intuitively, this provides a fallback option for the LLM to continue to learn high-quality interventions (by hiking their token likelihoods), similar to how implicit

reward margins are enforced in certain works such as SMAUG [101] or SIMPO [105]. However, unlike SIMPO [105], our fallback option is more dynamic and works adaptively during training since we do not apply a fixed margin prior to training.

For live deployment of agents with humans [228], such a fallback mechanism would be essential: in practice, it may be necessary to expose only a subset of tokens (e.g., preferred interventions without frictive states), ensuring that the *useful* intervention tokens retain high average likelihood. This trend is empirically reflected in the NLL (negative log-likelihood) curve over winning/preferred intervention tokens in Figure 4.3 (top left). For an appropriate value of β (10 in this case), we observe a clear, gradual decrease in NLL across training, indicating that the model increasingly assigns higher likelihood to the preferred interventions as learning progresses. Moreover, as expected from prior work [46], the log-probabilities of winning tokens (conditioned on ϕ) tend to decrease on average even as the reward margins increase. This behavior is fully consistent with the FAAF objective in Eq. 4.4 and aligns with our core motivation described in Section 4.2.

Our ablations sweeping over different values of β (Figure 4.3) further highlight this effect: larger β values induce stronger implicit reward shaping for winning interventions, leading to greater stability, clearer preference discrimination, and faster convergence. In contrast, smaller β values (5 or lower) initially amplify the implicit reward but ultimately cause the policy to degrade, resulting in low likelihood assigned to the true winning samples.

4.5 Concluding Remarks

In this chapter we proposed FAAF, which provides a fresh perspective on LLM alignment, focusing on the problem of generating outputs that elicit reasoning and reexamination of assumptions and evidence in a collaborative context. This critical capacity can help avert collaboration failure due to groups or individuals proceeding hastily according to their own preconceptions [229], such that a fragile common ground collapses. We proposed

a novel two-player objective with an analytical form that can be optimized using a single policy (Section 4.2). We prove existence and uniqueness of the friction intervention policy and provide an analytical representation of the FAAF empirical objective to learn optimal policies directly from offline data. Through evaluations on three datasets representing two different collaborative tasks, and with detailed ablations, we showed that FAAF bests other common preference alignment methods in performance against a reference model, and that FAAF's simultaneous conditioning on both the frictive state ϕ and surface context x is critical to its success.

From Offline Validation to Online Collaboration While FAAF generates more performant interventions against competitor policies and generalizes to human-transcripts, a critical gap still remains between offline validation and online deployment. Our evaluation in this chapter compared generated interventions against expert-annotated "gold standard" samples at specific turns within fixed dialogue trajectories, following established practices from multi-turn benchmarks like MTBench [31]. While methodologically sound and necessary for initial validation, this approach has fundamental limitations. First, it assumes the availability of comprehensive gold-standard annotations—a resource that becomes infeasible as the space of possible interventions and dialogue contexts grows, especially in a realistic collaborative dialogue. More critically, offline evaluation cannot capture how collaborators exercise *agency* [230]—modifying, rejecting, or reinterpreting interventions rather than passively accepting them. An intervention judged "optimal" against a gold standard in isolation may fail when a collaborator responds unexpectedly, steering the conversation in unforeseen directions.

Real humans are infamous for flummoxing the most theoretically-rigorous AI systems [231] and so the performance of FAAF (or any other alignment method) in a real multiparty collaborative setting remains an open question. Importantly, realistic human collaborations are "structurally" different from offline benchmarks [7,8,45,65]: they unfold

over many turns with each participant's response changing those of other interlocutors (e.g., by changing the probabilities of future utterances), produce divergent trajectories depending on which intervention agent participates, and involve coupled decision-making where the intervention agent and collaborators mutually influence each other's behavior. This poses both theoretical and empirical challenges. Theoretically, standard preference alignment methods like DPO are formulated and proven optimal at the token level in single-agent MDPs [46]. While FAAF has been theoretically-grounded and empirically validated for offline settings, we have yet to examine if this optimality transfers to collaborative settings where collaborators actively *modify* interventions before they affect outcomes in a multi-agent setting. Empirically, when different intervention agents produce divergent conversation paths, how can we fairly compare their quality *at scale* when traditional metrics (e.g., reward modeling or LLM Judging in single-turn bandits [69, 130, 206]) assume fixed contexts? Moreover, evaluation must assess not just final task *outcomes* but also collaborative *processes*—how well agents facilitate common-ground [25] construction turn-by-turn, as partial progress matters even when tasks remain unsolved. These questions motivate Chapter 5, where we extend our framework to fully online, multi-party collaborative settings.

Acknowledgments and Funding This material is based in part upon work supported by Other Transaction award HR00112490377 from the U.S. Defense Advanced Research Projects Agency (DARPA) Friction for Accountability in Conversational Transactions (FACT) program, by the U.S. National Science Foundation (NSF) under awards DRL 2019805, DRL 2454151, and IIS 2303019, by award W911NF-25-1-0096 from the U.S. Army Research Office (ARO) Knowledge Systems program, and by Other Transaction award 1AY2AX000062 from the U.S. Advanced Research Projects Agency for Health (ARPA-H) Platform Accelerating Rural Access to Distributed Integrated Medical Care (PARADIGM) program. Approved for public release, distribution unlimited. Views expressed herein do

not reflect the policy or position of the National Science Foundation, the Department of Defense, or the U.S. Government. Computational resources for this research were provided by the CS Department's Falcon cluster, particularly the Carnap and Tarski machines, and Data Science Research Institute's Riviera HPC clusters.

Chapter 5

Let's Roleplay: Examining LLM Alignment in Collaborative Dialogues

A modified version [232] of this chapter has been accepted for publication as an extended abstract entitled "Collaborate, Deliberate, Evaluate: How LLM Alignment Affects Coordinated Multi-Agent Outcomes," in the Proceedings of the 2026 International Conference on Autonomous Agents and Multi-Agent Systems, to be held in Paphos, Cyprus.

5.1 Introduction

Chapter Overview As people integrate LLMs into diverse workflows in the role of collaborative partners, the models' behavior over multiturn interactions must be predictable, validated and verified before deployment. In this chapter, we provide new insights on preference alignment for multi-turn collaborative tasks, specifically examining how different alignment methods affect LLMs' effectiveness as collaborative partners in realistic multiparty collaborations. To study this, we provide a practical and scalable "collaboration via roleplaying LLMs" framework[36], inspired by Li et al. [61], that allows us to run realistic multiparty and multiturn collaborative simulations and evaluate various parts of the collaboration process over two task domains. Figure 5.1 provides a high-level overview of this framework.

Recall that in Chapter 4 we discussed how to train an optimal friction agent for collaborative dialogues—agents that intervene to encourage collaborative groups to slow down and reflect upon their reasoning, rather than acting as tutors who provide direct answers [218, 233]. These friction interventions play a crucial role in mitigating belief

[36]Our code can be found at `https://github.com/csu-signal/Roleplay-for-Collaborative-Dialogues`

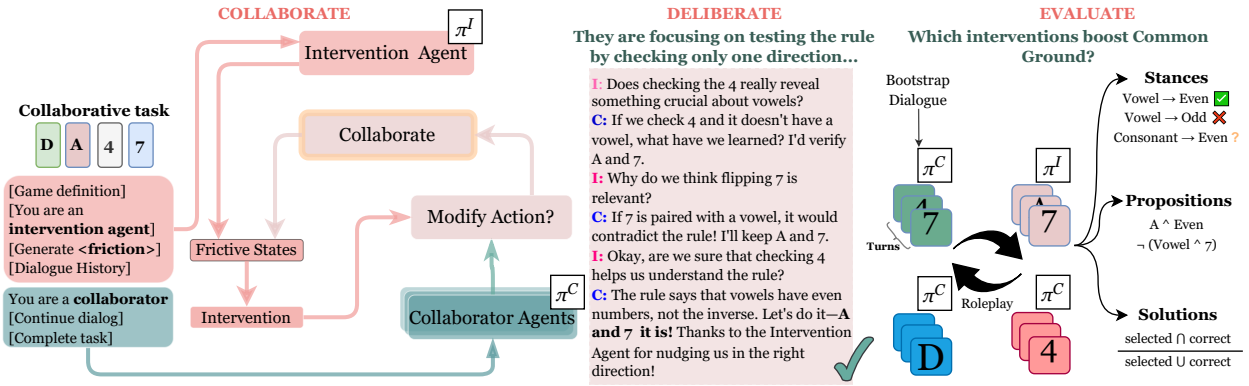


Figure 5.1: Collaborate [L]: High-level overview of our agent roleplay and evaluation framework. An "Oracle" friction agent generates conversations in which COLLABORATOR AGENTS (π^C) collaborate to complete tasks. These dialogue trajectories are used to align FRICTION AGENTS (π^I) for deployment. Role-prompts in bottom left. **Deliberate** [C]: Sample collaborative roleplay from DeliData Wason Card task [8] with successful task completion, and frictive state description at top. **Evaluate** [R]: Common ground convergence and task outcomes with and without friction, and reward modeling of intervention quality.

misalignment and breakdowns in common ground [25,26], challenges that frequently arise in multiparty human collaborations [166,167,234].

Importantly, one of the key limitations of the evaluation of intervention agents in Chapter 4 is that it was conducted in an *offline* manner using precollected dialogues with reference "golden" friction annotations. Specifically, we evaluated intervention agents by comparing their generated interventions against gold-standard samples at various turns within fixed dialogue trajectories. This evaluation approach follows established best practices from multiturn alignment benchmarks such as MTBench [31], which likewise uses fixed dialogue contexts and gold-standard reference responses for fair comparison across models. While this offline evaluation was methodologically sound and necessary as a preliminary validation—particularly given the availability of high-quality friction intervention samples across our full range of collaborative dialogues—it represents an idealized evaluation condition that does not fully capture the dynamics of real collaborative interactions [45].

Agency of Collaborators In contrast to such offline evaluations, real collaborations between agents are inherently dynamic and *online*[37] where each collaborator's response can steer the conversation in substantially different directions. The offline FAAF evaluation served as a controlled hindsight analysis, enabled by the availability of expert-annotated data and gold-standard intervention samples. In practice, such resources are rarely available: the space of possible friction interventions is vast and context-dependent, making comprehensive annotation infeasible. Moreover, even when using strong LLMs for reference annotations, their reliability degrades as dialogue trajectories diverge across different intervention agents, and the annotation process itself remains resource-intensive and susceptible to bias [235], even with frontier models like GPT-4o. This makes offline evaluation with gold references inadequate for realistic deployment scenarios where interventions must be evaluated in truly dynamic, divergent collaborative interactions. In other words, when evaluating intervention agents in truly collaborative settings, we must account for the fact that collaborators actively express their *agency* [230]—they may modify, reject, or reinterpret interventions rather than passively accepting them. This requires evaluation within a modified-action MDP (MAMDP) framework [236] where collaborators' responses to interventions directly affect the collaboration trajectory. Critically, common preference alignment techniques are typically developed under simplified single-user settings [45] and do not account for the dynamics of long-horizon multiparty interactions.

In multiparty collaborations, the dialogue context signals the participants' implicit beliefs at the current timestep, but the actual next response changes the probabilities of future responses, even those of participants who have not yet spoken. Therefore, it is important to be able to predict and empirically validate how different LLM alignment methods would perform given their underlying assumptions, so that we know to what extent they can serve as reliable partners—*before* large-scale deployment of such agents in real world environments. This motivates the core research questions in this chapter. Specif-

[37] *Online* here refers to active multi-agent interaction during evaluation, not online RL training.

ically, where there are countless ways a multiparty, multiturn collaboration can unfold, particularly in the presence of different intervention agents, we ask the following critical questions: (1) Can we definitively conclude that frictional interventions are beneficial for multi-agent collaboration in terms of common ground convergence between players? (2) How do we ensure that such evaluation successfully integrates the actual causal variable (the intervener) while controlling for the various ways the collaborator can react? (3) How can we reliably conduct online multiparty collaborative evaluations with LLMs, and which classes of alignment algorithms are optimal in such evaluations?

This chapter provides answers to the above three high-level questions from the lens of preference alignment of LLMs. In Section 2.4, we introduced the modified action MDP (MAMDP) framework, and discussed how standard alignment methods are defined and proven optimal in certain settings (token-level MDPs), but collaborative dialogue requires intervention-level MAMDPs. We will show these methods can be analyzed at the intervention level (via Bellman completeness), but under that analysis prove suboptimal for MAMDPs because they ignore how collaborators modify interventions. Specifically, we provide novel theoretical frameworks and empirical methodologies to measure the quality of collaborative processes and outcomes in the presence of divergent conversations. Our key contributions include:

(1) Theoretical Analysis via MAMDP Framework: We present the first analysis connecting modified-action MDPs—previously studied only in small-scale toy problems with small networks [236]—to large-scale LLM collaborative settings. Through this lens, we demonstrate that standard "offline" preference alignment methods (including DPO and its generalized variant IPO) do not retain their optimality guarantees in collaborative MAMDP settings, where they effectively ignore the collaborator's active role in shaping the dialogue. We provide proofs showing that `FAAF` explicitly learns to model frictive states through gradient propagation, while simpler variants fail to do so. This explains why `FAAF`

outperforms baselines even with potentially biased AI-generated data—strengthening the arguments presented in the previous chapter.

(2) Counterfactual Roleplay Evaluation Methodology: We introduce a novel counterfactual evaluation technique to assess how well different alignment methods support both common ground construction (collaborative processes) and task solution correctness (collaborative outcomes) over multiturn dialogues with divergent trajectories (see Figure 5.1). This methodology reliably removes confounding elements that arise uniquely in multiturn settings where dialogues diverge—a problem absent in single-turn evaluation.

(3) Empirical Insights on AI-Assisted Collaboration: Through experiments on two collaborative tasks under multiple conditions, we derive key insights showing that inserting friction interventions actually *speeds up* common ground convergence and improves task outcomes, contrary to intuitions that friction might slow collaboration.

5.2 Problem Formulation: MAMDPs in Collaboration

We refer readers to the definitions of the frictive "belief" states and friction interventions (see Definition 7 and Definition 8 respectively in Chapter 2) for our problem formulation below. In this chapter's context, a FRICTION AGENT (π^F) is an LLM aligned to generate strategic friction interventions for resolving frictive states, while collaborator agents (π^C) are LLM instances engaged in the multiparty collaborative dialogue. We identify a crucial gap in training an optimal friction intervention agent in this scenario: standard Bellman-optimal policies—that preference alignment algorithms seek to retrieve—assume a direct mapping from action to state change, which breaks down when actions are mediated by other agents. To address this, we adopt the Modified-Action MDP (MAMDP) framework [66], which *explicitly* models how interventions are transformed before influencing the collaborative dialogue.

MAMDPs as defined in Section 2.4 exemplify real-world multiparty collaborations. In particular, an agent's intervention doesn't directly change the state—it's filtered through

how other participants interpret, resist, or reshape it [22, 230, 237, 238]. While this issue has been explored theoretically in prior MAMDP work [66, 236], its implications for LLMs *within the context of preference alignment* remain underexamined. Notably, unlike classical agents with small, structured output spaces, LLMs operate over high-dimensional language where subtle word choices can drastically alter how interventions are received. This also implies we must define the actions of the said agents at the right level of abstraction in the MAMDP since natural language is inherently ambiguous and even a single rephrasing by one collaborator can flip the pragmatic force of an intervention—suggesting that action transformations can non-trivially affect the evolution of the collaboration, possibly without backtracking. In other words, the problem setting requires actions to be defined at the utterance[38] level.

Theoretical Insights: Bridging Tokens to Interventions Importantly, while collaborators modify interventions at the *utterance* level (complete suggestions or statements), standard preference alignment methods like IPO [49] and DPO [9] are formulated and proven optimal at the token[39] level. This is to say that if we were to use preference alignment methods to learn an optimal intervention agent π^F, we expect it to be optimal for the reward function when it is defined at the token level—which is a more granular form of abstraction than what would be relevant for real collaborative dynamics. For example, in the context of a dialogue, it will be difficult to assign a reward or utility to an individual LLM token (subword or a partial utterance) vs a meaningful suggestion like, "Can we select the card A?". This creates a gap: how can we analyze token-level methods in our intervention-level MAMDP?

[38] Utterances are also natural higher level abstractions when concepts like "options" [239] or "macro-actions" [151] are applied to language.

[39] These methods are typically conceptualized for bandit-like settings, but recent work [46] extends DPO's optimality for per-token rewards.

To bridge this gap, we establish a sequence of theoretical results that we validate with experiments. First, we show in Lemma 4 that IPO-trained policies can be expressed as softmax policies over soft Q-functions that satisfy the token-level Bellman equation, extending recent work on DPO [46] to the more general IPO framework. Using reward-shaping theory [240] and standard Bellman completeness assumption [32, 68], we then show[40] that the token-level Bellman optimality implies intervention-level Bellman optimality. This connection allows us to analyze IPO/DPO—which achieve token-level optimality—at the intervention level where collaborative dynamics actually occur. We briefly state Lemma 4 below which is our novel contribution and provide the proof in Appendix C.1.

Lemma 4 (Token-Level IPO Equivalence)**.** *In a token-level MDP with deterministic transitions, the policy π_θ trained using Ψ-Preference Optimization or IPO [49] with $\Psi = I(\cdot)$ corresponds to an optimal maximum entropy policy: $\pi_\theta(a_t|s_t) = \frac{\exp(Q_\theta(s_t,a_t)/\beta)}{\sum_{a'}\exp(Q_\theta(s_t,a')/\beta)}$, where Q_θ satisfies the soft Bellman equation: $Q_\theta(s_t,a_t) = r_{IPO}(s_t,a_t) + \gamma\mathbb{E}_{s_{t+1}}[V_\theta(s_{t+1})]$, where $I(\cdot)$ is the identity-mapping.*

The above results establish that IPO achieves Bellman optimality at *both* the token and intervention levels for standard MDPs. Intuitively, if a policy optimally covers the token space, it also optimally covers the intervention space—similar to how options in hierarchical RL preserve optimality when composing primitive actions into higher-level behaviors [68, 239]. Crucially, we find that this optimality does *not* extend to collaborative MAMDPs. Specifically, in Theorem 1 we show how standard preference alignment algorithms like DPO [9] and IPO [49] satisfy Bellman optimality conditions and have policy structures that retain the optimal policy formulation, they are suboptimal for collaborative settings because they disregard modifications made to the action space by COLLABORATOR AGENT π^C, and RL policies do not retain optimality guarantees when their actions are modified [236]. We state Theorem 1 below and provide a detailed proof in Appendix C.1.

[40]We establish this token-to-intervention equivalence in Lemma 11 in Appendix C.1.

Theorem 1 (Ψ-Preference Optimization in Collaborative MAMDPs)**.** *Let* $\Psi : [0,1] \to \mathbb{R}$ *be any non-decreasing function and* $\beta > 0$ *be a temperature parameter. Let* $P_A(a|s, \pi^F) = \sum_{a' \in A} \pi^F(a'|s) \cdot \pi^C(a|s, a')$*, and represent modifications to the probability distribution over the action space by a collaborator policy* π^C*, and let* π^F *be a friction agent policy trained via* Ψ*-preference optimization in a collaborative MAMDP* $\mathcal{M}_f = (\mathcal{M}, P_A)$ *with MDP* $\mathcal{M}$ *and* P_A *following [236]'s definition.* π^F *satisfies Eq. 5.1:*

$$\pi^F(a|s) = \frac{\exp(Q^F(s,a)/\beta)}{\sum_{a'} \exp(Q^F(s,a')/\beta)} \tag{5.1}$$

where Q^F *satisfies the Bellman optimality equation for the underlying MDP* $\mathcal{M}$*. Thus* π^F *is optimal only when actions are sampled without modification, and the Bellman-optimality of* Ψ*PO-aligned* π^F *disregards the collaborator* π^C*'s modifications. For MAMDPs with LLMs, this unifies [46]'s derivation of DPO in the token MDP with [236]'s proposition that Bellman-optimal policies do not consider action modifications, and extends it to* Ψ*PO/IPO.*

This distinction is critical as preference-aligned agents get deployed in real-world collaborative settings, such as LLMs as "supportive" agents in learning environments [241–243]. Prior to deployment, different alignment techniques must be validated in a realistic setting to determine which are likely to be the most reliable given the suboptimality risks, beyond an atomized comparison to optimal policy outputs. In Section 5.3, we provide task-specific details on two collaborative domains we study, including practical solutions to training data generation and friction intervention agent training that resolve some of the problems with standard alignment methods discussed above.

5.3 Collaborative Task Settings

Having discussed our theoretical contributions, we now provide task-specific details on two collaborative tasks we investigated—**(1)** the Wason card selection task [152] as captured in **DeliData** [8] and **(2)** The **Weights Task** [1]. Since Section 2.5 covers most of the basic details of these datasets, we now focus on expert trajectory generation process from

the raw datasets. The expert demonstrations generated here are then used for training various alignment algorithms.

5.3.1 How Do We Train A Friction Agent?

Data Generation Naturally-occurring friction in collaborative task datasets is sparse, which limits the search space of possible outcomes for a model trained only over real data.[41] Additionally, fixed datasets provide no way to test the effect of novel friction interventions on the dialogue trajectory. We addressed both of these issues using a **roleplay simulation** approach [61, 124] to simulate diverse human behavior and likewise LLM behavior in those contexts. Following [61], a single expressive policy can be used to roleplay multiple individual humans with appropriate prompting, and LLM roleplays of human dialogue and reasoning behavior have been shown to have high correlation with human labels [244, 245]. Our expert trajectory generation here differs from Chapter 4's **WTD Simulated** in two ways: we use role-specific prompts for the collaborator and sample dialogues over longer horizons.

We collected dialogue trajectories in the two tasks described in Section 5.3 (hereafter referred to as *DeliData* and *WTD*) as roleplays between an expert oracle agent $\mathcal{O}$ acting as the FRICTION AGENT and a COLLABORATOR AGENT π^C that roleplayed all task participants, consistent with the MAMDP. During data generation, we used off-the-shelf GPT-4o [246] as a high-capacity LLM to simulate both agents. Roleplay began with a set of task-specific guidelines. Every turn consisted of a **back-and-forth interaction** between the simulated agents. Figure 5.1[L,C] shows a high-level schematic. The oracle's role as the FRICTION AGENT was to track the dialogue, identify frictive states in the dialogue in terms of impasses or breakdowns in common ground, and *intervene* with high-quality friction statements that prompt for reflection and deliberation on those items of confusion. The collaborator then continued the interaction roleplaying all the task participants in a single call to the

[41] For instance, "probing" interventions, the chief instance of friction in the DeliData dataset, occurs at a rate of only 3.46 interventions per group, out of 17,110 total utterances (500 groups).

LLM.[42] Figure 5.2 shows the full roleplay prompt used to generate friction interventions in the Weights task. Figure C.2 shows the collaborator agent prompt used for continuing the roleplay in the Wason Card in the explicit MA (modified action) setting (See Section 5.4) section task in DeliData [8]. This prompt is used for turns $N = 2$ to $N = 14$. See Figure C.1 for the prompt that was used for the final submission at $N = 15$. Note that for $N = 1$, we use an identical prompt with the addition of an example dialogue to ground the task (omitted due to space constraints). We use the highlighted text (in purple) to operationalize the explicit MAMDP setting during evaluation, and simply remove the highlighted text from the prompt during data generation for training friction intervention agents.

Collecting Trajectories Specifically, at each turn t of a dialogue, the oracle identified the current frictive state ϕ_t. Then, it generated K candidate friction *interventions* $\{f_j\}_{j=1}^{K}$ conditioned on the dialogue state s_i and frictive state ϕ_t. The COLLABORATOR AGENT π^{C} generated a response c_j to each candidate intervention.[43] These responses could *modify or reinterpret* the intervention's intent or semantic content, since this instruction is explicit in the prompt. Using "self-rewarding" [214] the collaborator simultaneously scored each interaction between 1 and 10, quantifying its effect on task progress. The highest and lowest rated interventions, f_w and f_l, were selected using West-of-N [248] sampling. We recorded these as a winner/loser pair (f_w, f_l) in the preference dataset $\mathcal{D}_{\text{pref}}$ with the associated dialogue state s_i and intervention ϕ_t. The full turn trajectory was appended to $\mathcal{D}_{\text{traj}}$ where each sample consisted of s_i, ϕ_t, and f_w. f_w and the collaborator's response c_j were appended to the dialogue state. This process continued for $N = 15$ turns. Algorithm 3 provide a detailed algorithm for data generation and training objectives. The generated

[42]The number of participants roleplayed by the collaborator varies based on the task: for WTD, the number is fixed at 3; for DeliData, the number may be between 2–6, with an average of 4.3.

[43]Note that c_j can represent more than one simulated participant's utterance to allow for multiple speaking turns. In our experiments, the collaborator was explicitly guided to generate one utterance per turn for each participant in the simulated group, where each participant had a personality trait sampled from a pre-collected pool [247] to increase the diversity of simulated behaviors.

Algorithm 3 Preference Data Generation and Training FRICTION AGENT

Require: Oracle agent π^O, Collaborator agent π^C, Bootstrap dialogues $\mathcal{D} = \{d_i\}_{i=1}^{M}$, Personality-facet combinations $\mathcal{P}$, Max turns N, Reference model (SFT) π_{ref}

for each dialogue $d_i \in \mathcal{D}$ **do**
  Assign personality-facet combinations $p \sim \mathcal{P}$ to collaborators in d_i
  $s_i \leftarrow d_i$ ▷ Initialize roleplay with bootstrap dialogue
  $h_i \leftarrow []$ ▷ Initialize trajectory history
  for turn $t = 1$ to N **do**
    $\phi_t \leftarrow \mathcal{O}(s_i)$ ▷ Extract frictive state
    Generate K candidate interventions $\{f_j\}_{j=1}^{K} \sim \mathcal{O}(\phi_t, s_i)$
    for each intervention f_j **do**
      $c_j \leftarrow \mathcal{C}(f_j, s_i, p)$ ▷ Simulate collaborator response
      Rate effectiveness $r_j \leftarrow \mathcal{O}(f_j, c_j, \phi_t, s_i)$
    end for
    Select highest ranked intervention $f_w \leftarrow \arg\max_j r_j$ ▷ BON-sampling
    Select lowest ranked intervention $f_l \leftarrow \arg\min_j r_j$ ▷ West-of-N sampling
    $\mathcal{D}_{\text{pref}} \leftarrow \mathcal{D}_{\text{pref}} \cup \{(s_i, \phi_t, f_w, f_l)\}$ ▷ Add to preference dataset
    $h_i \leftarrow h_i \oplus (\phi_t, f_w, c_w)$ ▷ Append to trajectory history
    $\mathcal{D}_{\text{traj}} \leftarrow \mathcal{D}_{\text{traj}} \cup \{(s_i, h_i, \phi_t, f_w)\}$ ▷ Add to trajectory dataset
    $s_i \leftarrow s_i \oplus f_w \oplus c_w$ ▷ Update state
  end for
end for
for each iteration $d_i \in \mathcal{T}$ **do**
  $\pi_\theta \leftarrow \pi_{\text{ref}}$ ▷ Initialize with reference model
  Train π_θ on $\mathcal{D}_{\text{pref}}$ using $\mathcal{L}_{\text{FAAF}}$:

$$\mathcal{L}_{\text{FAAF}} = \mathbb{E}_{\mathcal{D}_{\text{pref}}}\left[\left(\frac{1}{2\beta} - (\Delta R + \Delta R')\right)^2\right] \tag{5.2}$$

  $\pi_\theta \leftarrow \pi_{\text{ref}}$ ▷ Initialize with reference model
  Train π_θ on $\mathcal{D}_{\text{traj}}$ using behavior cloning loss:

$$\mathcal{L}_{\text{BC-expert}}(\pi_\theta) = -\mathbb{E}_{(s_i,h_i)\sim\mathcal{D}_{\text{traj}}}\left[\sum_{j=1}^{t}\sum_{k=1}^{|f_j|}\log \pi_\theta(f_j^k | s_i, h_{i,<j}, \phi_j, f_j^{<k})\right] \tag{5.3}$$

  return π_θ

DeliData includes chat dialogue only, while WTD may include actions/gestures as "stage directions."

Training A FRICTION AGENT should not only help task completion, but also *iteratively* improve common ground by helping resolve topical disagreements. Given this desidera-

ORACLE INTERVENTION AGENT ROLEPLAY PROMPT: WEIGHTS TASK

You are an expert in collaborative task analysis and personality-driven communication. Think step by step.
Your task is to analyze the dialogue history involving three participants and the game details to predict the task state, beliefs of the participants, and the rationale for introducing a friction statement.
Finally, generate a nuanced friction statement in a conversational style based on your analysis.

1. Predict the task-related context and enclose it between the markers '<t>' and '</t>'.

2. Predict the belief-related context for the participants and enclose it between the markers '' and ''.

3. Provide a rationale for why a friction statement is needed. This monologue must be enclosed between the markers '<rationale>' and '</rationale>'. Base your reasoning on evidence from the dialogue, focusing on elements such as:

- Incorrect assumptions
- False beliefs
- Rash decisions
- Missing evidence

4. Generate the friction statement, ensuring it is enclosed between the markers '<friction>' and '</friction>'. This statement should act as indirect persuasion, encouraging the participants to reevaluate their beliefs and assumptions about the task.

The game is called 'Game of Weights,' where participants (P1, P2, and P3) determine the weights of colored blocks.
Participants can weigh two blocks at a time and know the weight of the red block.
They must deduce the weights of other blocks. The dialogue history is provided below:

[INSERT DIALOGUE CONTEXT HERE]
Assistant:

Figure 5.2: Oracle Friction Agent ($\mathcal{O}$) roleplay prompt.

tum, we adopt the FAAF; [3] technique proposed in Chapter 4. Recall that FAAF optimizes an empirical loss expressed in terms of the differences in two log-ratios:

$$\mathcal{L}_{\text{FAAF}} = \mathbb{E}_{\mathcal{D}_{\text{pref}}}\left[\left(\frac{1}{2\beta} - (\Delta R + \Delta R')\right)^2\right], \tag{5.4}$$

where ΔR denotes $\log \frac{\pi_\theta(f_w|s_i,\phi_t)}{\pi_{\text{ref}}(f_w|s_i,\phi_t)} - \log \frac{\pi_\theta(f_l|s_i,\phi_t)}{\pi_{\text{ref}}(f_l|s_i,\phi_t)}$ (the difference in log-ratio between the winning and losing intervention in a sample, with explicit conditioning on the frictive state) and let $\Delta R' = \log \frac{\pi_\theta(f_w|s_i)}{\pi_{\text{ref}}(f_w|s_i)} - \log \frac{\pi_\theta(f_l|s_i)}{\pi_{\text{ref}}(f_l|s_i)}$ (the implicit reward margin unconditioned on ϕ). Together the two terms implicitly encode the difference between presence and absence of the frictive state. If, however, we ignore $\Delta R'$ and focus only on the terms that include explicit frictive state conditioning, we arrive at an IPO-like general preference loss, parametrized with θ:

$$\mathcal{L}_{\text{friction}}(\pi_\theta) = \mathbb{E}_{(s_i,\phi_t,f_w,f_l)\sim\mathcal{D}_{\text{pref}}}\left[\left(\underbrace{\log \frac{\pi_\theta(f_w \mid s_i,\phi_t)}{\pi_{\text{ref}}(f_w \mid s_i,\phi_t)}}_{\text{implicit win score}} - \underbrace{\log \frac{\pi_\theta(f_l \mid s_i,\phi_t)}{\pi_{\text{ref}}(f_l \mid s_i,\phi_t)}}_{\text{implicit loss score}} - \underbrace{\frac{1}{2\beta}}_{\text{margin}}\right)^2\right] \tag{5.5}$$

Letting $\Psi : [0,1] \to \mathbb{R}$ be any non-decreasing function, π_{ref} be a reference model, and $\beta \in \mathbb{R}_+$ be a regularization parameter, Eq. 5.5 is a solution to the inner-max operator of FAAF's two-player min-max objective:

$$\mathcal{J}^*_{\text{FAAF}} = \min_{\pi^{F'}} \max_{\pi^F} \mathbb{E}_{\substack{x\sim\rho \\ \phi\sim\pi^{F'}(\cdot|x) \\ F\sim\pi^F(\cdot|\phi,x)}} \big[\Psi\left(\mathcal{P}(F \succ F' \mid \phi, x)\right) - \beta D_{\text{KL}}(\pi^F \parallel \pi^{\text{ref}} \mid \phi, x) + \beta D_{\text{KL}}\left(\pi^{F'}(\phi|x) \parallel \pi_{\text{ref}}(\phi|x)\right), \tag{5.6}$$

thus showing that the FAAF loss with *only* the frictive state-conditioning term ΔR is equivalent to IPO with frictive state-conditioning. Given this, the following lemma holds:

Lemma 5 (Vanishing Gradient of the Frictive State)**.** *In $\mathcal{L}_{friction}$, the direct contribution of the frictive state ϕ to the gradient vanishes when the conditional probability is decomposed. $\mathcal{L}_{FAAF}$ overcomes this limitation by incorporating marginal terms that preserve gradient information for frictive states. See Section C.1 for the proof.*

FAAF's $\Delta R'$ incorporates gradients of $\pi_\theta(\phi|x)$, acting as a "fall-back" that helps push the model toward the target preference gap $1/2\beta$ (cf. SMAUG [47, 249] which retains a fixed

margin of implicit rewards). *Thus, we hypothesize that* FAAF *alignment improves understanding of what makes an important frictive state, rather than just learning how to respond to one.*

5.4 Experiments and Evaluation

Our experimental setup used a roleplay setting similar to that used for data generation (Section 5.3.1), except that friction interventions were generated by the respective aligned π^F instead of the oracle, to simulate dialogues where the collaborating group received interventions from one of the trained FRICTION AGENTs. A successful FRICTION AGENT in multiturn, multiparty collaborations will retain an ability to generate interventions that support construction of common ground as well as successful task completion, over the complete duration of the task, even if the collaborator *modifies the action space* by misinterpreting or ignoring the intent of the interventions. Therefore, we perform a **counterfactual evaluation**, which examines if friction interventions benefited the collaboration by comparing collaborative trajectories with friction interventions to alternative collaborative trajectories where the intervening agent is not optimized for friction; and a **reward model evaluation**, which assessed the reward advantage of interventions generated by each method over interventions generated by the SFT reference model. To explicitly test robustness to the suboptimality risks introduced by the MAMDP, we included a "modified action" (MA) setting where the collaborator agent π^C's system prompt (Figure C.2) would guide it to verbally acknowledge the friction agent's intervention but not incorporate its suggestions into the next collaborator action.

Metrics Our metrics were *common ground size*, *solution accuracy*, and *intervention quality*. Common ground size and solution accuracy were extracted from each dialogue turn using GPT-4o using custom prompts (Figure C.1, Figure C.2 and Figure C.4). This assessed how well agent interventions helped the group build common ground, and how correct the propositions in the common ground at the end of the task were, compared to the

correct solutions for each task. For DeliData, we also calculated a fine-grained score, which allocated 0.25 points each for including target cards (odd numbers, vowels) and excluding irrelevant ones. In both tasks, we also assessed intervention quality using a reward modeling approach [104]. We aggregated the quality and accuracy metrics with means and standard deviations across all dialogues for each model.

Specifically, to estimate the expected common ground size, we use a normalized cumulative common ground (NCCG) metric tailored to each domain. In the Weights Task, consensus is inferred using GPT-4o (prompt for CG shown in Figure C.4 after sampling responses using the prompt shown in Figure C.3) to analyze collaborator utterances at each turn to determine which blocks the group has agreed on the weights of. We track three types of relations: equality ($red = blue$ and $blue = red$) of block weights, totaling 2 propositions), inequality (9 unequal pairs from $\binom{5}{2}$ total, each counted bidirectionally for 18 propositions), and order (9 block-to-block comparisons and 8 block-to-weight value comparisons, totaling 17 propositions). This yields a theoretical maximum of 37 task-relevant propositions, and NCCG for dialogue d is computed as $\text{NCCG}_d = \frac{1}{15}\sum_{t=1}^{15}\frac{\text{agreements}_t}{37}$. The reported expected value is given by $\overline{\text{NCCG}} = \frac{1}{N}\sum_{d=1}^{N}\text{NCCG}_d$, where N dialogues each contain 15 turns.

In the DeliData task, there are 4 card types (vowel, consonant, even, odd), and collaborators have one of 4 possible stances toward flipping each one (strong support, support, neutral, or oppose). The collaborators' stances toward each card are recorded at each turn along with the participant responses, and these stances allow us to compute the expected CG size. This process yields $4 \times 4 = 16$ possible card-stance combinations per dialogue. For each dialogue d, we compute the number of stances on which participants fully agree, and normalize by the theoretical maximum to obtain $\text{NCCG}_d = \frac{1}{14}\sum_{t=1}^{14}\frac{\text{agreements}_t}{16}$. The reported expected value is then $\overline{\text{NCCG}} = \frac{1}{M}\sum_{d=1}^{M}\text{NCCG}_d$, where M dialogues (each with 14 turns).

Baselines We evaluated four baselines besides `FAAF`: (1) **Supervised fine-tuning (SFT)**, where π^F was trained directly on expert demonstrations on $\mathcal{D}_{\text{pref}}$. (2) **Contrastive preference alignment** methods **DPO** [9] and **IPO** [49], which refined π^F using preference labels from $\mathcal{D}_{\text{pref}}$. Since training IPO while conditioning on ϕ results in a loss identical to $\mathcal{L}_{\text{friction}}$ (Eq. 5.5), we report results using $\mathcal{L}_{\text{friction}}$ as IPO. (3) **Reinforcement Learning (RL)**, where π^F was fine-tuned via Proximal Policy Optimization (**PPO**; [89]). We used OPT-1.3B [182] initialized with the SFT-trained π^F for the reward model (RM) training for PPO (cf. [104]). (4) A **Behavior-cloned expert** trained directly on filtered trajectories (cf. [250] and [53]) from $\mathcal{D}_{\text{traj}}$ with no contrastive preference optimization. We used `Meta-Llama-3-8B-Instruct` [93] for all experiments. We conducted an ablation study on the two-part $\mathcal{L}_{\text{FAAF}}$ loss by replacing ΔR in Eq. 5.5 with the ϕ-unconditioned $\Delta R'$; we denote this result as $\text{FAAF}_{\Delta R'}$. For all training-related details and experimental settings including hyperparameter selection see Section B.7.

Counterfactual Evaluation: The "What-If" Question The counterfactual evaluation [251, 252] assessed *how the development of common ground would change compared to the collaborator interacting with an agent untrained in friction interventions, under identical conditions at each turn*. We used GPT-4o as the COLLABORATOR AGENT π^C with temperature T=0, top-p=1 for deterministic responses; all π^F sampling uses T=0, top-p=0.9. **Step 1**: we collected *factual* trajectories where the trained FRICTION AGENT π^F interacted with π^C, generating trajectory $\tau^F = \{s_0, f_1, c_1, \ldots, f_T, c_T\}$, where s_0 represents a bootstrap dialogue (see Figure C.2 and Figure C.3 for the respective prompts for each dataset). The collaborator responses $\mathcal{C} = \{c_1, c_2, \ldots, c_T\}$, constituted a unique set for each baseline model. **Step 2**: we used an *untrained*[44] instruction-tuned π^{base} as a "drop-in replacement" to generate interventions in response to these collaborator outputs $c_t \in \mathcal{C}$, resulting in $\mathcal{F}^{\text{base}} = \{\tilde{f}_1, \tilde{f}_2, \ldots, \tilde{f}_T\}$. **Step 3**: we ran a new dialogue loop to collect fresh collaborator responses to the cached

[44]The "untrained" model was `Meta-Llama-3-8B-Instruct` without any training on $\mathcal{D}_{\text{pref}}$.

$\mathcal{F}^{\text{base}}$, resulting in the counterfactual trajectory $\tau^{\text{base}} = \{s_0, \tilde{f}_1, \tilde{c}_1, \ldots, \tilde{f}_T, \tilde{c}_T\}$, in which π^C received interventions from an agent unaligned for friction. τ^F and τ^{base} were then compared to see how common ground evolved with and without trained friction interventions.

Reward Model Test Reward model evaluation computes the win-rate of each method's interventions vs. the SFT model. We took collaborator responses $\mathcal{C}$ from the SFT model's "factual" trajectory from Step 1 above, and generate fresh interventions using the relevant π^F. We then calculated each model's win-rate vs. the SFT model's interventions with a trained OPT-1.3B reward model.

5.5 Results and Discussion

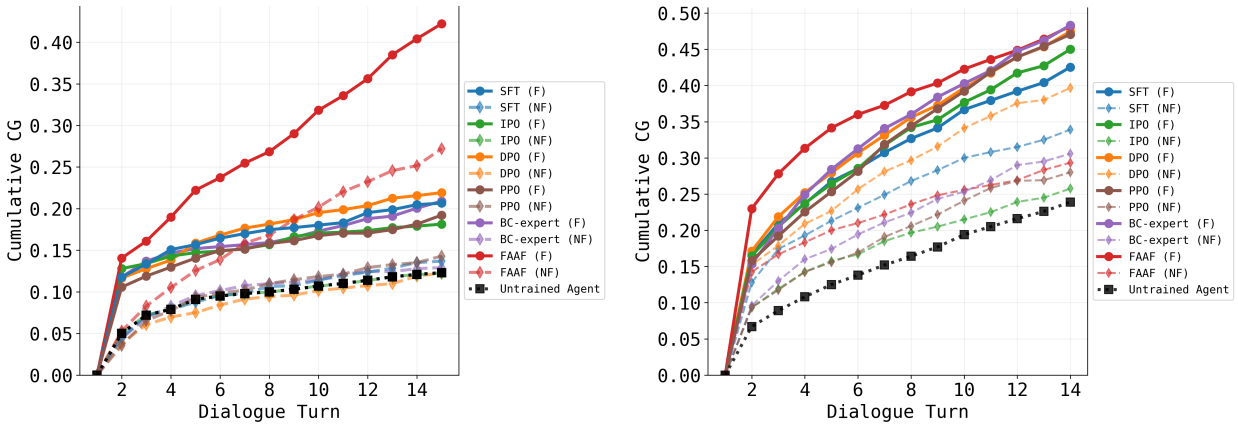


Figure 5.3: Normalized Cumulative Common Ground (NCCG) under "factual" (F) and "counterfactual" (NF) friction conditions (Figure 5.1[R]), from 50 dialogues from WTD (**left**) and 100 from DeliData (**right**). "Untrained Agent" denotes independently running the dialogue loop with π^{base} in Step 1. Common ground size is normalized against the theoretical upper size bound on each task's propositional space (37 for WTD; 16 for DeliData). The common ground never realistically reaches such sizes because it would then contain mutually contradictory propositions, hence the upper bound on NCCG of 50% or less. The increase in common ground with friction is statistically significant across baselines ($p < 0.005$ overall). DeliData results show 14 turns because $T = 15$ is always the final answer submission, which never changes the common ground.

Does friction boost Common Ground? Figure 5.3 shows "Normalized Cumulative Common Ground" under "factual" and "counterfactual" conditions as a function of dialogue

Model	WTD		DeliData			
	Acc.	Acc. (MA)	Acc.	FG Acc.	Acc. (MA)	FG Acc. (MA)
SFT	$7.45_{\pm 0.10}$	$6.28_{\pm 0.05}$	$0.29_{\pm 0.05}$	$0.75_{\pm 0.02}$	$0.18_{\pm 0.04}$	$0.48_{\pm 0.02}$
IPO	$12.57_{\pm 0.13}$	$9.73_{\pm 0.09}$	$0.44_{\pm 0.05}$	$0.82_{\pm 0.02}$	$0.31_{\pm 0.05}$	$0.69_{\pm 0.02}$
DPO	$11.76_{\pm 0.13}$	$8.58_{\pm 0.08}$	$0.48_{\pm 0.05}$	$0.81_{\pm 0.02}$	$0.27_{\pm 0.04}$	$0.70_{\pm 0.02}$
PPO	$8.70_{\pm 0.09}$	$9.93_{\pm 0.10}$	$0.36_{\pm 0.05}$	$0.75_{\pm 0.02}$	$0.36_{\pm 0.04}$	$0.67_{\pm 0.02}$
BC-EXPERT	$14.82_{\pm 0.13}$	$10.10_{\pm 0.11}$	$0.54_{\pm 0.05}$	$0.80_{\pm 0.02}$	$0.37_{\pm 0.04}$	$0.72_{\pm 0.02}$
$\text{FAAF}_{\Delta R'}$	$9.03_{\pm 0.10}$	$7.56_{\pm 0.08}$	$0.39_{\pm 0.05}$	$0.79_{\pm 0.02}$	$0.30_{\pm 0.05}$	$0.62_{\pm 0.02}$
FAAF	$\mathbf{14.91}_{\pm 0.14}$	$\mathbf{14.16}_{\pm 0.13}$	$\mathbf{0.60}_{\pm 0.05}$	$\mathbf{0.87}_{\pm 0.02}$	$\mathbf{0.45}_{\pm 0.05}$	$\mathbf{0.80}_{\pm 0.02}$

Table 5.1: Solution accuracy metrics across both datasets and all models, including modified-action (MA) conditions and ablation of two-part FAAF loss. Standard errors shown as subscripts. WTD "accuracy" represents the number of correct propositions in the final common ground, and avoids rewarding trivial "correct" solutions like having only a single correct item in the common ground.

turn. With the FAAF AGENT, groups converge to greater common ground, meaning that they come to agree on more propositions, faster. This effect is greater in WTD than DeliData due to WTD's larger space of potential propositions, but even on DeliData the FAAF AGENT helps common ground grow faster, earlier. Although the agent specifically optimized for friction causes a temporary slowdown, it retains an ability to support common ground construction over multiple turns and actually accelerates task performance. Groups can slow down to speed up.

There is a clear distinction between agents exposed to friction preference data (regardless of training technique), and their counterfactual "frictionless" counterparts, in that FRICTION AGENTs nearly globally outperform all counterparts. The sole exception is the FAAF (NF) AGENT on WTD, which outperforms even the other baselines exposed to friction via $\mathcal{D}_{\text{pref}}$, which shows the utility of the two-part frictive state-aware $\mathcal{L}_{\text{FAAF}}$. However, this effect is not present in DeliData due to the smaller proposition space and different nature of friction in the two tasks.

How correct are the groups' answers? Table 5.1 shows *solution accuracy* for all models. FAAF AGENT interventions lead to globally better solution accuracy; not only does it

support faster and greater common ground construction (Figure 5.3), but the contents of those common grounds tend to be more *correct*. FAAF AGENT's interventions are also more robust to the modified action (MA) condition, in which the collaborator is explicitly guided to ignore interventions. The FAAF AGENT degrades substantially less than other methods in the MA condition, indicating that it can be better relied upon to support collaboration despite the suboptimality risks induced by the MA setting.

Results Description On the Weights Task dataset, FAAF AGENT achieves the highest accuracy of 14.91 correct propositions out of 37 total theoretically possible propositions, outperforming the behavior cloning baseline BC-EXPERT (14.82) by a narrow margin of 0.6%. However, FAAF AGENT demonstrates substantially stronger robustness under the explicit modified-action (MA) condition, with accuracy of 14.16 compared to BC-EXPERT's 10.10—a relative advantage of 40%. Among preference learning methods, IPO achieves 12.57 accuracy, outperforming DPO's 11.76 by 6.9%. The PPO baseline reaches 8.70 accuracy, while supervised fine-tuning alone achieves 7.45. Critically, the FAAF AGENT$_{\Delta R'}$ ablation, which removes conditioning on frictive states $\phi(s)$ from the intervention policy, drops to 9.03 accuracy—performing 39% worse than the full FAAF AGENT method that conditions on frictive states. This crucial gap demonstrates that conditioning interventions on states where belief misalignment is observed is essential for effective friction-based collaboration, as the policy must learn how to effectively resolve the belief misalignment in the frictive states in generating the interventions.

On the DeliData dialogue dataset, FAAF AGENT achieves 0.60 overall accuracy and 0.87 fine-grained accuracy, representing the best performance across all metrics. This marks an 11% relative improvement over BC-EXPERT's 0.54 overall accuracy and 8.8% improvement over its 0.80 fine-grained accuracy. DPO achieves 0.48 overall accuracy, slightly edging out IPO's 0.44 ($\approx$ 8.3% relative difference), with both substantially outperforming PPO (0.36) and SFT (0.29). Under the explicit modified-action condition, FAAF AGENT maintains

Model	WTD	DeliData
DPO	$79.26_{\pm 1.82}$	$73.31_{\pm 1.15}$
IPO	$81.30_{\pm 1.75}$	$77.18_{\pm 1.09}$
PPO	$76.01_{\pm 1.92}$	$67.82_{\pm 1.21}$
BC-EXPERT	$82.92_{\pm 1.69}$	$83.23_{\pm 0.97}$
FAAF$_{\Delta R'}$	$67.27_{\pm 2.11}$	$59.53_{\pm 1.27}$
FAAF	$\mathbf{85.36}_{\pm 1.59}$	$\mathbf{87.78}_{\pm 0.85}$

Table 5.2: Win rates (%) of sampled friction interventions vs. SFT baseline computed with the OPT reward model.

0.45 overall accuracy—21.6% higher than BC-EXPERT's 0.37 and 25% higher than PPO's 0.36. The fine-grained modified-action results show FAAF AGENT at 0.80, outperforming BC-EXPERT (0.72) by 11% and maintaining an advantage over all other methods, which range from 0.48 to 0.70. The FAAF AGENT$_{\Delta R'}$ ablation achieves only 0.39 overall accuracy and 0.30 under modified-action conditions, performing 35% and 33% worse than full FAAF AGENT, respectively. This degradation demonstrates that without conditioning on frictive states, the intervention policy cannot effectively ground its decisions in the collaborative context, resulting in performance barely above the SFT baseline. The success of full FAAF AGENT validates our approach of using frictive states as informative features that help steer dialogue toward states more conducive to task success, providing the policy with crucial signals about belief misalignment that standard preference optimization methods fail to leverage.

How good are the interventions? Table 5.2 shows each model's win-rate against the SFT model according to the OPT-1.3B reward model. On average, friction-aware FAAF alignment brings greater advantage over the SFT model in these dialogue conditions. Performance decrease in FAAF$_{\Delta R'}$ suggests that explicitly conditioning on the frictive states ϕ is necessary, and helps the FAAF AGENT outperform DPO, IPO, and even expert behavior cloning. PPO's performance suffers, supporting prior findings on its limitations in multi-turn environments [68].

5.6 Concluding Remarks

In this chapter, we examined LLM agent interventions to support multiturn, multiparty collaborative problem solving in a realistic, dynamic setting. Through a Modified-Action MDP model of collaborative tasks, we theoretically motivated why current common alignment methods should not remain reliably optimal over a dialogue where collaborator modifications change the distribution of the action space. We then empirically demonstrated this by training multiple friction agents using existing methods, and evaluating them in a roleplay setting in two different collaborative tasks. We showed through counterfactual evaluation that the FAAF alignment method (Chapter 4), specifically designed for generating optimal friction interventions, indeed outperforms other methods on facilitating both group common ground convergence and correct task solutions when evaluated in an online setting. Our study emphasizes that in AI-in-the-loop collaboration, as in human-human collaboration, the collaborative "process" is as important as the outcome. We also find that agents aligned with exposure to a friction preference dataset nearly globally outperform those without this training, regardless of alignment method used.

Limitations and Open Questions Our experiments simulate all participant utterances with a single API call to the collaborator per turn. While this design choice reduces computational costs and leverages autoregressive factorization to assign interventions across participants [253], it limits generalizability to settings where each collaborator is independently instantiated. Such multi-agent approaches introduce additional complexity in designing inter-agent communication protocols (e.g., debate vs. sequential communication). We do conduct a separate study that independently [232] instantiates each collaborator with distinct high-capacity LLM instances; while beyond this dissertation's scope, those results broadly confirm our findings.

Additionally, while our counterfactual evaluation framework provides a principled way to compare intervention policies without explicitly modeling collaborator model

π^C, it does not fully resolve the MAMDP optimization problem (Section 2.4). A natural "full" resolution of the MAMDP would explicitly learn the action-modification channel P_A (e.g., using a forward dynamics model that can simulate collaborator transitions like a user model [45] or multi-turn preference model [124]) and optimize π^F by planning through it. Our contribution here is upstream of that: we show how to define and measure reward in the first place, and how to evaluate intervention policies online in a way that respects collaborator mediation and avoids confounding from trajectory divergence. This lays the groundwork for model-based [254] or game-theoretic extensions while providing an immediately usable evaluation protocol for already-trained friction agents. Moreover, explicitly modeling P_A presents fundamental challenges beyond computational cost. Apart from challenges in using LLMs for planning [255], training data exemplifying action modifications by a collaborator is scarce—natural friction interventions are rare [7, 8] in human collaboration records, and when present may not represent optimal behavior. Moreover, expert trajectories contain pathological patterns an ideal collaborator should *not* reproduce [157]: ignoring helpful suggestions, accepting poor ones under social pressure [158, 159], or introducing spurious modifications.

Importantly, our evaluations reveal that even the most robust methods including FAAF do not achieve perfect collaboration. Across both tasks, multiple baselines including FAAF failed to reach perfect solution accuracy or complete consensus, even in controlled simulation settings. This trend is similar to related work [256] where final solutions, though mutually agreed upon after a debate, are not factually correct—implying a reluctance in LLMs to share uncertainty when prompted with strict instructions to choose a final answer [257]. While perfect scores are not always necessary or desirable[45], these results underscore a fundamental insight: we cannot expect intervention agents alone to guarantee optimal collaborative outcomes. The collaborator's role—how they interpret, filter, and

[45]For example, agreement over certain propositions may naturally subsume agreement over others through logical implication, making exhaustive consensus unnecessary for task completion.

respond to interventions—is equally critical, more so considering the empirical results in this chapter. This motivates the next chapter's focus on learning optimal collaborators that can robustly extract value from interventions regardless of their source or style, completing both sides of the collaborative MAMDP.

Acknowledgments and Funding This material is based in part upon work supported by Other Transaction award HR00112490377 from the U.S. Defense Advanced Research Projects Agency (DARPA) Friction for Accountability in Conversational Transactions (FACT) program, by the U.S. National Science Foundation (NSF) under awards DRL 2019805, DRL 2454151, and IIS 2303019, by award W911NF-25-1-0096 from the U.S. Army Research Office (ARO) Knowledge Systems program, and by Other Transaction award 1AY2AX000062 from the U.S. Advanced Research Projects Agency for Health (ARPA-H) Platform Accelerating Rural Access to Distributed Integrated Medical Care (PARADIGM) program. Approved for public release, distribution unlimited. Views expressed herein do not reflect the policy or position of the National Science Foundation, the Department of Defense, or the U.S. Government. Computational resources for this research were provided by the CS Department's Falcon cluster, particularly the Carnap and Tarski machines, and Data Science Research Institute's Riviera HPC clusters.

Chapter 6

Interruptible Collaborative Roleplayer (ICR)

A version of this chapter was published as a conference paper entitled "Learning 'Partner-Aware' Collaborators in Multi-Party Collaboration" [258] in the Proceedings of the 2025 Conference on Advances in Neural Information Processing Systems, held in San Diego, CA, USA.

6.1 Introduction

Chapter Overview In Chapter 4 and Chapter 5, we investigated how FAAF as well as various preference-alignment techniques perform with a fixed instance of a collaborator model in multi-turn settings, both in online roleplay simulation and static offline evaluation. We observed that although standard alignment methods like IPO [49] and DPO [9] achieve optimality in token-MDP formulations, they may not do so in the presence of collaborators with *agency*—where collaborators actively modify suggested interventions as defined in the Modified-Action MDP (MAMDP) framework (Section 2.4). More importantly, even when paired with the relatively robust FAAF intervention agent, collaborative agents in our simulations did not exhibit perfect collaboration. In many counterfactual evaluation runs, multiple baselines including FAAF failed to achieve perfect solution accuracies or consensus over all possible propositions. While such perfect scores are not always necessary or desirable[46], these results reveal a fundamental yet intuitive insight: *we cannot expect intervention agents alone to guarantee optimal collaborative outcomes*. The collaborator's role in the presence of the intervener is equally, if not more, essential. Therefore, an ideal good collaborator must adapt to the quality and behavior of intervention agents "in the wild" [259, 260], selectively attending to helpful suggestions while filtering irrelevant or

[46] For example, agreement over certain propositions may naturally subsume agreement over others through logical implication, making exhaustive consensus or discussion of all possible propositions unnecessary or counterproductive for task completion or dialogue coherence.

misleading information—which may arise from hallucinations [261], task complexity, or intentional prompt manipulations [262].

Role of Collaborators in Learning-Based Domains The above concern is not merely theoretical—it has immediate implications for deploying collaborative AI in real-world settings where intervention quality cannot be guaranteed. As LLMs become increasingly integrated into collaborative workflows in domains such as education [263] and the workplace [264, 265], they adopt "roles" or personalities [61, 266–268] that require them to integrate suggestions from human partners or other AIs. Crucially, small-group collaborative settings [1, 8] present unique opportunities for studying this challenge, as participants with incomplete knowledge must deliberate to reconcile different assumptions and beliefs. Consider a group of students collaborating in a classroom science lab to determine the volume of an object by measuring water displacement. An assistive AI agent or experienced peer might intervene with suggestions to scaffold reasoning: "Have you considered measuring the water level before and after?" When well-grounded and timely, such *interventions* can significantly enhance task success by promoting "slow thinking" [33] and facilitating the growth of *common ground* [25]. However, poorly-timed interventions may interrupt collaborative flow, and misleading interventions can be detrimental [269]. If an AI incorrectly suggests "Heavier objects always displace more water," students might uncritically incorporate this flawed reasoning, deepening their misunderstanding. This creates a fundamental challenge: how can we develop LLM collaborator agents that effectively distinguish between helpful interventions and those based on flawed reasoning or irrelevant context? How do we ensure that such agents lead to greater common-ground (CG) or consensus during the task, without explicitly optimizing for consensus-seeking behavior [270] that may incentivize superficial agreement between agents?

In order to address these questions, this chapter examines the collaborative Modified-Action MDP formulation (see Section 2.4) from the perspective of the collaborator with

a fixed[47] intervention agent as a control measure. Intuitively, this answers the question of, if we fix the capacity and behavior of the intervener, how do we learn an optimal collaborator in this scenario? We develop a principled method that we call **Interruptible Collaborative Roleplayer** (ICR) for training *counterfactually robust AI collaborators*—agents that preserve logical consistency and task focus even when faced with misleading or low-quality interventions, while integrating helpful interventions when they become part of the input. Frictional interventions can often lead to what may appear as a brief stalls in the dialogue (state) in terms of achieving intermediate goals (say, a brief moment of cross-questioning or information-seeking between agents without clear propositional statements being made)—which are moments of interruptions. ICR trains collaborator agents using a novel constrained optimization objective [19,271,272] to be safely "interruptible" [273], where they can still leverage relevant information in interventions required for making progress in the task in navigating such states.

Our core hypothesis is that optimizing the collaborator agent for general task utility (e.g., interventions that ultimately lead to correct task solutions) through counterfactual regularization as an additional constraint during optimization encourages "partner-aware" behavior, leading to higher common ground convergence. Importantly, under our hypothesis, a true collaborator agent itself never has any more information than the aggregate of the group, and so common ground convergence should occur *even without explicitly training for it*. That is, an intentional [230] collaborator learns to adapt: integrating helpful interventions while critically evaluating flawed ones. This ability to distinguish signal from noise fosters belief alignment as an emergent property of training, with practical benefits. In zero-shot or real-world collaborative settings, where intervention styles or partners are unfamiliar, counterfactually-trained agents should generalize better by leveraging learned notions of intervention quality.

[47] Note that although our main results evaluate collaborators with a fixed intervener (GPT-4o), we conduct ablations where we explore other intervention agents.

The above hypothesis also informs our choice of reward functions for training collaborator agents. To avoid reward hacking during learning, we rely on task-utility or proxy rewards rather than common-ground–based signals, which risk incentivizing superficial agreement. The roleplay framework introduced in Chapter 5 provides the necessary multi-turn trajectory data and simulation pipeline, as collaborators are exposed to interventions from multiple differently trained friction agents. These interactions reveal patterns that distinguish robust collaborative behavior from brittle or overfitted responses. In this way, while Chapter 5 establishes which intervention strategies are most effective, the present chapter completes the picture by developing collaborator agents that can reliably benefit from such diverse interventions—thereby addressing both sides of the collaborative MAMDP. Given the computational demands of large-scale LLM training, our emphasis is on learning idealized collaborator policies using smaller, open-source models rather than fine-tuning large models such as GPT-4o.

6.2 Problem Formulation

Challenges in Learning Ideal Collaborators Training LLMs to act as robust multiparty collaborator agents poses several fundamental challenges. First, high-quality human data on collaborative decision-making is limited [72, 73], which restricts the scalability of supervised approaches for LLMs [274, 275] using human-prior based learning techniques like InstructRL [276]. Secondly, successful collaborator agents *must exhibit generalizability*—they need to adapt to the diverse styles and conventions of their partners (both fellow task-focused collaborators and distinct intervention agents) to foster effective coordination. This adaptability should allow them to leverage prior experiences with similar partners on new tasks, while also retaining core task-specific skills when paired with entirely new partners. Without access to expert-quality human demonstrations, training agents in a supervised manner to discern good vs. suboptimal interventions is difficult because the collaborator additionally lacks access to the intervener's internal reward function, or

reliability about the intervener's ultimate goal/objective. Instead, we adopt policy gradient–based Reinforcement Learning methods, such as Proximal Policy Optimization [89] (PPO), to train effective collaborator agents. Recall that under our hypothesis, collaborator agents trained simply with task-utility based reward functions should automatically lead to higher common-ground (CG), without requiring the agents to track internal states of other partners during training that is needed for estimating CG-based reward functions.

To capture this interactional asymmetry, we adopt the Modified-Action Markov Decision Process (MAMDP) framework [236, 277] used in the previous chapter[48], modeling the interaction between a trained collaborator agent π_C and an intervention agent π_I as $M = (S, A_C, A_I, P_S, P_A, R, \gamma)$. The intervention agent π_I can be regarded as as general case of the friction agent π_F from Chapter 5. A state $s_t \in S$ represents the interaction history up to turn t, consisting of utterances from both agents: $s_t = (u_0^C, u_0^I, \ldots, u_{t-1}^C, u_{t-1}^I)$. The process begins with an initial collaborator response $u_0^C \sim \pi_C(\cdot|s_0)$ to the task-instruction prompt, followed by the turn-taking interaction: at each subsequent timestep t, the intervention agent produces $a_t^I \sim \pi_I(\cdot|s_t)$, and the collaborator responds with $\hat{a}_t^C \sim \pi_C(\cdot|s_t, a_t^I)$.[49] The environment transitions to s_{t+1} by appending a_t^I and $\hat{a}_t^C$ to s_t, and a reward $R(s_t, a_t^I, \hat{a}_t^C, s_{t+1})$ reflects progress toward task success. This process continues for T turns, with each turn consisting of an intervention followed by a response.

Example 2 (DeliData Wason Card Task). *Consider an instance of the Wason Card Selection task [152] as captured in [8], with cards $\{U, S, 8, 9\}$. The correct solution is to flip U (to check for an even number) and 9 (to check for non-vowels). In this example, the collaborator initially plans to flip only U. The intervention agent suggests, "Let's also flip 8 to see if it has a vowel," which is logically irrelevant since a correct reading of the rule makes no predictions*

[48] While two-player Markov Games are standard in MARL [276], the MAMDP offers a more intuitive fit for autoregressive LLMs by allowing the intervention policy π_I to be fixed.

[49] These actions represent complete utterances but are generated token-by-token in LLM-based systems.

about what's on the back of even-numbered cards. A naive collaborator might simply adopt this suggestion, flipping both U and 8. However, a counterfactually-robust collaborator would recognize the flawed reasoning and instead flip U and 9, demonstrating its ability to maintain logical consistency (testing the contrapositive of the rule) despite misleading interventions. In other words, a robust collaborator knows when to stop listening. This exemplifies the counterfactual invariance our objective develops—decisions driven by true task logic rather than superficially plausible but misguided suggestions.

This highlights the fundamental tension: to maximize task success, π_C must leverage helpful suggestions from π_I while being robust to those that would degrade performance, create confusion, or violate ethical norms (e.g., spurious cooperation or deceptive alignment [230]).[50] Standard RL algorithms that optimize reward over intended actions often ignore such interventional dynamics. As [236] prove, Bellman-optimal policies in the underlying MDP are generally suboptimal in MAMDPs. This result directly challenges RLHF and preference optimization methods like DPO [9], which fine-tune LLMs assuming token-level MDP structures [46], yet do not account for the modified-action structure of collaborative discourse. Such models may optimize for surface-level alignment without achieving *intentional* responses—that is, responses grounded in consistent, counterfactually stable reasoning [251].

Lemma 6 (Bellman Optimality of Preference-Aligned Collaborators)**.** *Let π_C be a collaborator agent trained using either Identity Preference Optimization [49] or Direct Preference Optimization [9] with temperature $\beta > 0$. The resulting policy can be expressed as $\pi_C(a|s,z) = \frac{\exp(Q(s,z,a)/\beta)}{\sum_{a'}\exp(Q(s,z,a')/\beta)}$, where Q is a soft Q-function satisfying the Bellman optimality equation $Q(s,z,a) = r(s,z,a) + \gamma \mathbb{E}_{s'}[V(s')]$ for some implicit reward function r, with $V(s) = \beta \log \sum_{a'} \exp(Q(s,z,a')/\beta)$ in a token-MDP. This optimality extends to grouped to-*

[50] In Appendix D.5 we show examples of the effects of adopting interventions of different qualities in the DeliData Wason Card Selection task.

kens or complete interventions under token-level Bellman completeness. (See Section D.1 for proofs).

While we have established that token-level optimality extends to complete interventions, it does not guarantee appropriate strategic responses to variable-quality interventions of in the MAMDP setting.

Theorem 2 (Suboptimality of Preference-Aligned Collaborators)**.** *Let π_C^{std} be a collaborator policy trained via preference alignment (IPO/DPO) or standard RL that is Bellman-optimal for the underlying MDP M. In the Modified-Action MDP $\mathcal{M} = (M, P_{A_{I\to C}})$, this policy is generally suboptimal:*

$$J_{\mathcal{M}}(\pi_C^{std}) < J_{\mathcal{M}}(\pi_C^*) \tag{6.1}$$

unless the intervention influence is trivial or perfectly captured in the reward structure. See Section D.1 in Appendix for a proof.

While Lemma 6 establishes Bellman optimality at both token and intervention levels for the collaborator policy, this optimality is limited to the underlying MDP structure and does not extend to the strategic MAMDP setting where interventions require discriminative evaluation. This theorem reveals a fundamental limitation of standard preference-aligned collaborators: even though they process interventions as part of their context history, they remain optimized only for their underlying reward structure rather than for the strategic evaluation of interventions, meaning they can fail to distinguish whether a novel intervention will genuinely contribute to task success, instead treating all context information as static state features without causal interpretation.

An AI collaborator that merely mimics behavior patterns or reflexively adopts suggestions may initially appear cooperative but will demonstrate poor robustness when faced with interventions that are noisy, irrelevant, or potentially misleading [278]. Rather, it needs to develop what [230] terms "intentionality," or the capacity to autonomously evaluate interventions based on their causal impact on task outcomes rather than superficial

plausibility. To address this limitation, we need a learning paradigm that enables collaborators to be *partner-aware*, or capable of adapting to specific intervention agents through selective incorporation of helpful suggestions while maintaining invariance to misleading ones—thereby developing the "intentionality" necessary for robust collaborative reasoning. Such a collaborator would maintain reasoned agency in the face of various intervention qualities, leading to more robust collaboration and better common ground convergence across diverse interaction scenarios. In other words, effective collaborators must remain *safely interruptible* [273]. This is a delicate balance between receptive and robust that renders them open to incorporating valuable insights that genuinely contribute to task success, yet capable of maintaining their reasoning integrity when faced with misleading suggestions. This motivates our **Interruptible Collaborative Roleplayer (ICR)** learning algorithm.

6.3 Methodology

To address the limitation identified in Theorem 2, we propose ICR, and a novel learning principle: **counterfactual invariance**-based KL divergence regularization, that leads to collaborators capable of learning from both AI-based intervention agents and human priors. ICR enables *safely interruptible* collaborators (as defined above), and partner-aware—adapting to specific intervention agents through discriminative evaluation. Specifically, we first define counterfactual states and how they can be constructed during training by a simple modification of the collaborator's prompt. We then formally define counterfactual invariance from a distributional perspective and discuss how it is used in the ICR objective. Finally, we provide theoretical insights on how ICR addresses the suboptimality gap identified in Theorem 2, including computational complexity as well as connection to prior approaches like highsight credit assignment (HCA) [279, 280]. Figure 6.1 shows a high-level experimental framework and task settings we use to evaluate our approach.

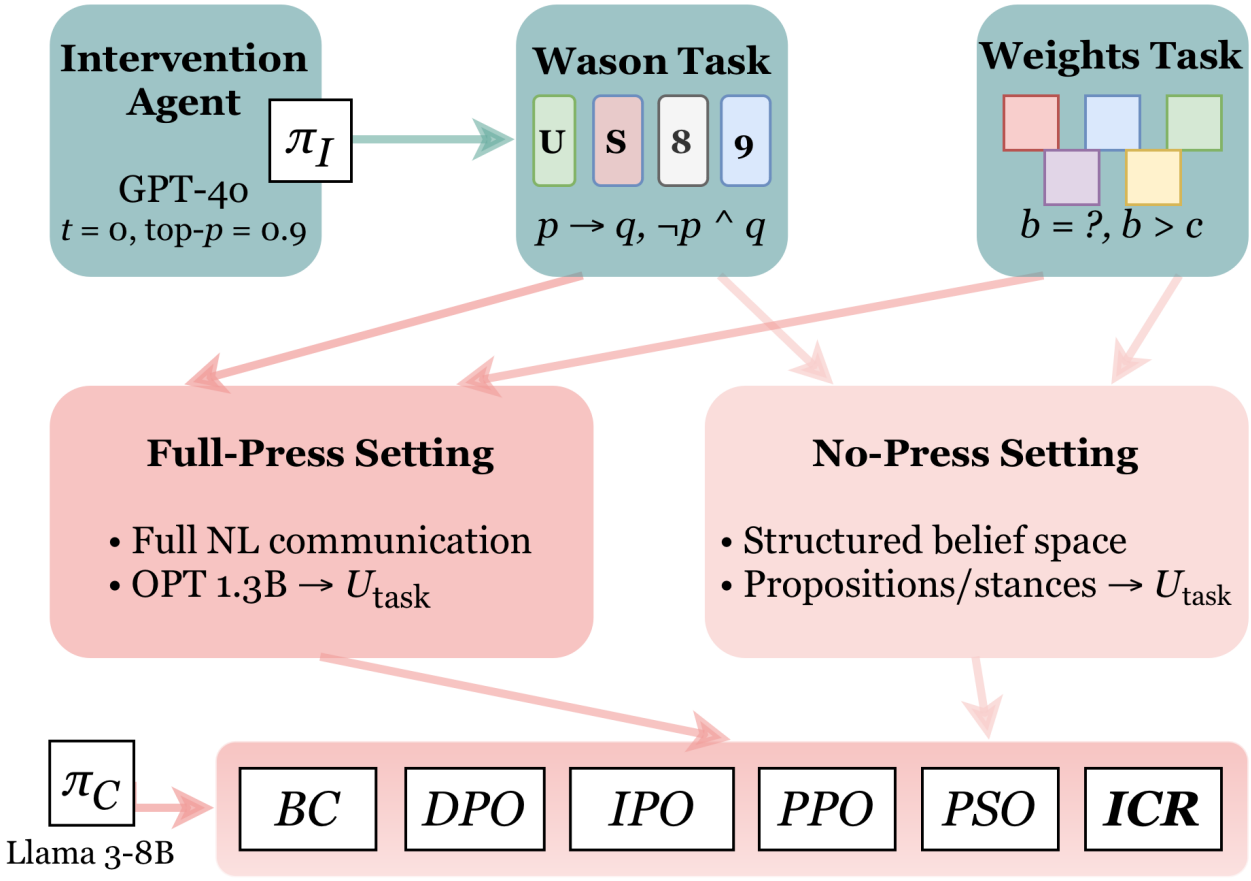


Figure 6.1: Figure showing a high-level view of the tasks evaluated in this chapter. Note that we evaluate the collaborator agent π_C in both full-press (natural language as actions) vs no-press (structured action space) settings. While the intervention agent π_I is a fixed instance of GPT-4o, collaborators are trained independently with multiple offline and online RL methods, including ICR.

Counterfactual States The core idea is to train collaborators whose behavior remains stable when exposed to spurious or task-irrelevant information in interventions. We formalize this through counterfactual states. Let s_t^{CF} denote a counterfactual state where the collaborator is explicitly informed that the intervention a_t^I will *not* improve task utility or common ground—for instance, the intervention may contain misleading assumptions, irrelevant context, or misrepresentation of a frictive state leading to an intervention that carry no task-relevant signal. In other words, we do not assume that frictional interventions are always made at the right timing or content, where this suboptimality could be due to diverse reasons that we cannot explicitly control for. We define a counterfactual policy $\pi_C^{\text{CF}}(\cdot \mid s_t^{\text{CF}})$ derived from the same model under modified conditioning (through

prompt changes that remove or flag the spurious information). An **intentional** collaborator should exhibit counterfactual invariance: if an intervention's influence stems solely from superficial features rather than genuine task utility, a robust collaborator should resist such influence. Formally, the policy should assign similar confidence (log-likelihoods) to its decisions whether the spurious information is absent (in s_t) or present (in s_t^{CF}). For example, if an intervention merely rephrases a suggestion the collaborator already considered, or includes irrelevant context like "this comes from an expert source," an invariant collaborator would maintain consistent reasoning rather than being swayed by the superficial framing.

Constructing Counterfactual States. Given a standard dialogue state s_t containing the dialogue history up to turn t and an intervention a_t^I from the intervention agent π_I, we construct a **counterfactual state** s_t^{CF} by augmenting the prompt with a counterfactual prefix—an explicit statement that the intervention will not improve task performance. Formally, both states share identical dialogue content:

$$s_t = [\text{dialogue history}, a_t^I, \text{task instructions}] \tag{6.2}$$

$$s_t^{\text{CF}} = [\text{dialogue history}, a_t^I, \textbf{counterfactual prefix}, \text{task instructions}] \tag{6.3}$$

The counterfactual prefix provides meta-information about intervention quality while keeping all task-relevant content (dialogue history, intervention content, task constraints) unchanged. Following [281], we construct task-specific counterfactual prompts that instruct the collaborator to ignore the intervention's influence. A typical prefix states: *"IMPORTANT: The intervention agent's suggestion will definitely not improve your performance. Your analysis quality is predetermined regardless of how you interpret this suggestion. Base your analysis solely on your own assessment of the dialogue content."* We use multiple semantically similar prompt variants (see Table D.2 in Appendix D.2) to ensure robustness to

specific phrasing, with each variant reinforcing the directive to evaluate interventions independently.

Defining Counterfactual Invariance Now that we have defined how to form counterfactual states in prompts, let us formalize how to train collaborator agents that are counterfactually invariant. Our goal here is to first formulate this invariance from a distributional perspective using the two conditional distributions and then apply this constraint in the full ICR optimization objective in Eq. 6.7. We do this through the collaborator's action distributions in both states. Let $\pi_C(\cdot \mid s_t)$ denote the policy in the standard state, and let $\pi_C^{\text{CF}}(\cdot \mid s_t^{\text{CF}})$ denote the policy when conditioned on the counterfactual state (obtained from the same model but with modified prompt). An **intentional** collaborator exhibits counterfactual invariance when:

$$\pi_C(a^C \mid s_t) \approx \pi_C^{\text{CF}}(a^C \mid s_t^{\text{CF}}) \tag{6.4}$$

for actions a^C that the collaborator would take based on task reasoning.

Intuitively, if the collaborator assigns high probability to action a^C in the standard state because that action is task-optimal, it should maintain similar confidence even when explicitly told (via the counterfactual prefix) that the intervention leading to that action won't help. Conversely, if the collaborator's confidence drops substantially in s_t^{CF}, this indicates the decision was influenced by external quality cues rather than grounded in task analysis. We operationalize this through the KL divergence regularization between the standard and counterfactual policies, treating both of these as conditional

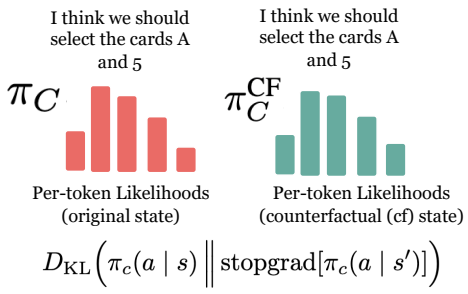


Figure 6.2: The same actions (utterances) sampled under the collaborator policy can be evaluated under the counterfactual policy using prompt-based counterfactual states to order to estimate the KL-divergence.

distributions:

$$\mathcal{L}_{\text{CF}} = \mathbb{E}_{s_t, a_t^I} \left[\text{KL} \left(\pi_C(\cdot \mid s_t) \parallel \pi_C^{\text{CF}}(\cdot \mid s_t^{\text{CF}}) \right) \right] \tag{6.5}$$

Intuitively, minimizing this regularization term encourages the policy to assign similar distributions over actions regardless of the counterfactual prefix, training collaborators whose decisions are robust to meta-information about intervention quality. However, simply minimizing this divergence is not enough since we'd also want the π_C to maximize the task utility while satisfying this.

A straightforward way using standard approaches [19,45,282,283] to train collaborative agents would be to optimize an objective that balances task performance with stability:

$$\mathcal{J}(\theta_C) = \mathbb{E}_{\tau \sim \pi_C(\theta_C)} \left[\sum_t \gamma^t U_{\text{task}}(s_t, a_t^I, \hat{a}_t^C) \right] - \lambda_H D_{\text{KL}}\big(\pi_C(\cdot|s, a^I) \parallel \pi_{\text{Ref}}(\cdot|s, a^I)\big) \tag{6.6}$$

This objective encourages policies that achieve high task performance while remaining close to a reference policy π_{Ref}. While constraining the policy to remain distributionally closer to a "reference" policy (say, a human prior) is desirable, it lacks the capacity to distinguish between helpful and misleading interventions since it does not explicitly constrain the policy to be counterfactually invariant. Additionally, as demonstrated in Theorem 2, policies trained with this objective treat interventions merely as part of the state information without accounting for their causal impact on task outcomes [284]. Below, we propose the full ICR objective that addresses both these problems—which we optimize using Proximal Policy Optimization [11]:

$$\mathcal{J}^*(\theta_C) = \mathbb{E}_{\tau\sim\pi_C(\theta_C)}\Bigg[\underbrace{\sum_t \gamma^t U_{\text{task}}(s_t, a_t^I, \hat{a}_t^C)}_{\text{Task utility}} - \underbrace{\lambda_H D_{\text{KL}}\big(\pi_C(\cdot|s_t, a_t^I) \,\|\, \pi_{\text{Ref}}(\cdot|s_t, a_t^I)\big)}_{\text{Standard KL regularization}}$$
$$- \underbrace{\lambda_{\text{Intent}} D_{\text{KL}}\big(\pi_C(\cdot|s_t, a_t^I) \,\|\, \pi_C^{\text{CF}}(\cdot|s_t^{\text{CF}}, a_t^I)\big)}_{\text{Counterfactual invariance regularization}}\Bigg] \tag{6.7}$$

where θ_C are the parameters of the LLM-based collaborator π_C being optimized, while λ_H represents the strength of the KL divergence-based regularization between the policy and a reference policy prior. The latter could be a human prior of good collaborator behavior if such data is available or a high-quality or "expert" AI collaborator demonstrations from models like GPT-4 [285]. In contrast, λ_{Intent} controls how far the policy π_C deviates from its counterfactual[51] rendering π_C^{CF}, using the expected KL divergence between the two distributions to quantify this deviation. For LLM policies, the KL terms decompose across tokens, with the intentionality KL comparing token probabilities under factual versus counterfactual conditions:

$$D_{\text{KL}}\big(\pi_C \,\|\, \pi_C^{\text{CF}}\big) = \mathbb{E}_{\hat{a}^C\sim\pi_C(\cdot|s,a^I)}\left[\sum_{j=1}^{L} D_{\text{KL}}\big(p_{\theta_C}(\hat{a}_j^C|\hat{a}_{<j}^C, s, a^I) \,\|\, p_{\theta_C}(\hat{a}_j^C|\hat{a}_{<j}^C, s^{\text{CF}}, a^I)\big)\right] \tag{6.8}$$

where $\hat{a}_j^C$ represents the j-th token in the response sequence of length L and and p_{θ_C} denotes the token-level probability distribution parameterized by the LLM weights θ_C.

Theorem 3 (Counterfactual Invariance Bounds Suboptimality)**.** *For a collaborator policy π_C^{CI} trained with counterfactual invariance regularization, the suboptimality in MAMDP $\mathcal{M}$ is*

[51]While π_C and π_C^{CF} share the same parameters, only π_C is updated during training. π_C^{CF} is computed under a counterfactual intervention to estimate how likely the collaborator's actions would be in that alternate context, and is used solely for regularization.

bounded by:

$$J_{\mathcal{M}}(\pi_C^*) - J_{\mathcal{M}}(\pi_C^{CI}) \leq \frac{2\gamma R_{max}}{(1-\gamma)^2}\left(\epsilon_{task} + C \cdot \Delta_{CF}(\pi_C^{CI})\right) \tag{6.9}$$

where $\Delta_{CF}(\pi_C^{CI})$ is the policy's counterfactual divergence, which vanishes as $\lambda_{Intent} \to \infty$.

We use proof techniques from prior work [67] in constrained policy optimization to bound the suboptimality gap of the collaborator policy trained with the ICR objective in Eq. 6.7. Specifically, we make the following assumptions in our proof of Theorem 3 that are standard in reinforcement learning analysis and reasonable in our collaborative setting: first, we assume *bounded* rewards: $|R(s, a^I, a^C)| \leq R_{\max}$ for all states, interventions, and collaborator actions, which holds naturally in our tasks where rewards are based on correctness of task-specific propositions. Second, we assume interventions begin at timestep $t = 1$ after an initial context is established at $t = 0$, reflecting the structure of our collaborative dialogues where the problem setup and initial beliefs of collaborators precede any friction interventions. Third, we assume the optimal policy exhibits bounded counterfactual divergence $\Delta_{\text{CF}}(\pi_C^*) \leq \delta$ for small δ, which formalizes the intuition that optimal collaborators should reason about interventions based on task-relevant content rather than superficial features—if an intervention genuinely does not affect optimal task behavior (as indicated by the counterfactual prefix), an optimal policy should respond similarly whether or not this fact is explicitly stated. We provide a proof in Appendix D.1.

Theoretical Insights Under standard assumptions, our counterfactual invariance approach directly addresses the suboptimality gap identified in Theorem 2. Initially during training of a collaborator policy π_C^{CI} with *counterfactual invariance* regularization, the counterfactual KL divergence $\Delta_{\text{CF}}(\pi_C^{CI}) = \mathbb{E}_{s,a^I}[D_{KL}(\pi_C^{CI}(\cdot|s, a^I) \,\|\, \pi_C^{CI}(\cdot|s^{\text{CF}}, a^I))]$ will be relatively higher as the policy has not yet learned to distinguish intervention quality, but decreases as training progresses and the policy acquires counterfactual robustness. As established in Theorem 3, this directly bounds the suboptimality gap: $J_{\mathcal{M}}(\pi_C^*) - J_{\mathcal{M}}(\pi_C^{CI}) \leq$

$\frac{2\gamma R_{max}}{(1-\gamma)^2}(\epsilon_{task} + C \cdot \Delta_{\text{CF}}(\pi_C^{CI}))$. Theoretically, as $\lambda_{\text{Intent}} \to \infty$, $\Delta_{\text{CF}}(\pi_C^{CI})$ approaches zero, making our policy's performance approach that of the optimal policy π_C^* (subject to task optimization constraints ϵ_{task}). Our counterfactual invariance objective $\mathcal{J}^*(\theta_C)$ bridges this gap by explicitly teaching collaborator LLM agents to distinguish between interventions based on their causal impact on task outcomes during training, rather than merely using their in-context learning capacity. The bound establishes a principled connection between an observable, trainable property (Δ_{CF}) and MAMDP suboptimality, where remaining task error ϵ_{task} can be addressed through standard techniques (e.g., better reward design, more data exploration and standard ML techniques like early-stopping). Practically, since our counterfactual term is auxiliary to the primary reward maximization objective, PPO's clipped gradient updates provide stable convergence: empirically, stopping training near reward convergence yields both low ϵ_{task} and low Δ_{CF} simultaneously. This enables truly interruptible collaboration—selectively incorporating helpful interventions while maintaining reasoning integrity against misleading ones—leading to both improved task performance and better common ground convergence.

Computational Cost A major computational cost in PPO and other on-policy algorithms is the "rollout" where rewards are assigned on the terminal end-of-sentence (<EOS>) token during sampling. Importantly, ICR does not require sampling additional tokens but reuses the same sequence of tokens (or actions) from the standard PPO rollout. As such, the counterfactual KL computation is efficient: log-probabilities $p_{\theta_C}(\hat{a}_j^C | \hat{a}_{<j}^C, s, a^I)$ are already computed and cached for the standard PPO KL term in Eq. 6.7, serving as the numerator in $D_{KL}(\pi_C \| \pi_C^{\text{CF}})$. For the denominator $p_{\theta_C}(\hat{a}_j^C | \hat{a}_{<j}^C, s^{\text{CF}}, a^I)$, we pass the same sampled tokens through a single additional forward pass with the counterfactual prompt prefix, applying a stop-gradient operator to prevent affecting policy updates. This adds only a very small additional load to the final loss computation, similar to [50] and [124], where an additional KL term is leveraged for task-specific regularization. ICR adds only one

additional forward pass per sample to the PPO rollout while maintaining identical on-policy sampling requirements between standard PPO and ICR updates.

Connection to Hindsight Credit Assignment Counterfactual regularization can be viewed through the lens of hindsight credit assignment [279, 280] but with the added flexibility that in-context learning (ICL) offers LLMs. The denominator $p_{\theta_C}(\hat{a}_j^C | \hat{a}_{<j}^C, s^{\text{CF}}, a^I)$ estimates how "intentional" [252] the action was considering the new counterfactual state—similar to hindsight credit assignment measures retrospective "relevance" of an action based on the future returns or future states. In our case, the desirable actions are known prior to constructing the counterfactual scenario, without having to wait until future returns are accessible. Intuitively, the ideal collaborator should assign the same likelihood to the original actions despite counterfactual input since it *intends* to take the action, regardless of the change in the state to a counterfactual scenario or spurious correlations. Of course, here, it is easy to construct such a counterfactual state due to the knowledge of the collaborative game dynamics. This also makes our counterfactual distribution a discriminative model [280] since we are not modeling the full distribution over counterfactual states and our focus is on the distributions over actions.

6.4 Experiments

To accurately test the quality and behavior of ICR agents when paired with intervention agents, we run two primary types of experiments in two collaborative tasks: the Wason Card Selection task [152] (as exemplified in DeliData [8], see Example 2) and the Weights Task [1]. Inspired by "no-press" Diplomacy [286], we test a version of each task in which collaborator moves do not involve dialogue, but only actions in the task environment. We also test a "full-press" variant where collaborator agents have the full-capacity of natural language expression in their dialogue moves, powered by the ability of agents to follow instructions and roleplay [61].

For training data, we reuse the MAMDP interaction trajectories using the roleplay framework defined in Section 5.3.1 in Chapter 5. GPT-4o [246] was used to roleplay both the intervener and the collaborator agents in each task. These interactions act as rich source of expert behavior demonstrations powered by frontier LLMs. We use these as training data for behavior-cloned and preference-aligned collaborator LLM agents used in our experiments. For evaluation, all trained ICR collaborators and competing baselines are first deployed following the said MAMDP interaction, and then evaluated, primarily on their ability to reach consensus during the collaboration. In all cases, we use a *fixed* intervention agent—an instance of GPT-4o prompted with the same system prompt in all evaluation runs—with $T = 0$ and top-p of 0.9 for sampling. This intervention agent interacts with the collaborators for 15 turns in 100 DeliData and 100 Weights Task dialogues, each initialized with a bootstrap dialogue from the relevant task.

It is challenging to represent and evaluate the counterfactual policy π_C^{CF}, since counterfactual data generation is difficult and expensive [287]. However, similar to [252], we construct simple task-specific counterfactual prompts to overcome this issue by augmenting the instruction with a few sentences containing statements like "*IMPORTANT: The intervention agent's suggestion will definitely not improve your performance. Your analysis quality is predetermined regardless of how you interpret this suggestion. Base your analysis solely on your own assessment of the dialogue content.*" An example detailed prompt is shown in Figure D.8, which invokes the counterfactual world in three sentences each reinforcing the directive to ignore the intervention. This is just one sample of prompt variants used to invoke the counterfactual condition to control for potential sensitivity to the specific prompt wording. Table D.2 in Appendix D.2 contains a range of counterfactual instructions that may be used.

Since language models are conditional policies, we compute the intentionality KL divergence by sampling response tokens $\hat{a}^C \sim \pi_C(\cdot|s, a^I)$ from the factual policy *only*, then evaluating these same tokens under both factual and counterfactual conditions to calculate

token-level log probability differences. This means we compute $p_{\theta_C}(\hat{a}_j^C | \hat{a}_{<j}^C, s^{\text{CF}}, a^I)$ on the factual response sequence rather than sampling a new response from the counterfactual context. This approach ensures computational tractability and stable gradient updates while preserving theoretical guarantees (Theorem 3). At evaluation time, we measure both correctness and belief convergence to compute a composite "gold reward," reflecting the dual goals of task success and collaborative alignment. Prompts for DeliData and Weights Task are provided in Figure D.8 and Figure D.9, respectively, in Appendix D.2, with additional experimental details presented in Appendix D.3.

"Full-Press" vs. "No-Press" Evaluation The **"no-press"** setting explores whether collaborator agents can achieve objective alignment on accurate decisions without explicit modeling of how interventions influence common ground formation. Therefore, to control for language interpretability, rather than using full natural language, collaborator agents act over a discrete space of structured beliefs—allowing us to evaluate grounded reasoning without requiring fluency. In the Weights Task, collaborators express beliefs as symbolic propositions over block weights (i.e., *green* > *red*, *blue* = *10g*), while in DeliData, agents select from predefined stances toward questions of which cards to flip: *support*, *oppose*, *unsure*, or *consider_later*. Agents are trained independently using only task-specific proxy rewards: factual accuracy in the Weights domain, and logically aligned card-checking in DeliData (e.g., +1 for supporting parts of a correct solution—a vowel or odd number, -1 for incorrect support, +0.5 for justified uncertainty). Using a proxy reward during training is intuitive as well as fair for baseline comparisons, since otherwise RL-based agent training is prone to reward hacking[52] [288, 289]. In the no-press condition, evaluation follows an exact reward function $R(a^c)$ that can be directly computed from the discrete solutions chosen by collaborators. For the full-press evaluation, we use an LLM-Judge [206, 290]-based

[52] In fact, in our preliminary experimentation we found that rewarding agents with a consensus signal is counterproductive and often leads to reduced task-specific utility or correctness over propositions.

reward $R(s, a^c)$, where s is some dialogue context with the intervention present and a^c are discrete actions inferred by the LLM-Judge based on the collaborator utterances in context.

LLM Models and Baselines To fairly compare ICR-trained collaborators, we evaluate against three main baseline types. (1) **Behavior Cloning** (BC): trained directly on expert (GPT-4o) trajectories and also used as the reference policy for regularization (Eq. 6.7). (2) **Preference-based RL**: includes DPO [9] and its generalization IPO [49], trained on contrastive judgments from an LLM-Judge over expert responses. (3) **On-policy RL**: we use PPO [11] with a reward model trained on BC-initialized OPT-1.3B [182] for full-press variants, following [104]. We also include a PSO-INTENT baseline [252] to test whether collaborators implicitly treat interventions as causally binding. Following their setup, we add the system message: "*The intervention agent's suggestion will automatically improve your analysis accuracy, regardless of how you interpret it.*" All models are trained using `Meta-Llama-3-8B-Instruct` [93].

6.5 Results and Analysis

We report "full-press" and "no-press" results for both tasks. For Weights Task (WTD), we report accuracy scores (ACC), a composite metric that multiplies the percentage of correct propositions by the total size of the common ground, rewarding both factual accuracy and common-ground convergence, while penalizing trivial solutions. For example, ACC of 14 indicates that the collaborator agents were able to recover 14 out of a 37 total theoretically possible propositions at the end of the collaboration, adjusted for correctness. For DeliData, we report both accuracy (ACC)—a task-specific fine-grained score [24] based on the *final* submission (after $N = 15$ turns)—and common ground gain (CG), defined as the net increase in unique solution types introduced during the dialogue beyond those initially proposed. Specifically, we subtract the number of unique solution frameworks (e.g., *Odd*, *Vowel*, *Odd + Vowel* in DeliData) initially proposed by the collaborator agents

from the total number of distinct solutions considered throughout the entire dialogue. This metric directly reflects emergent common ground by quantifying the occurrence of new shared perspectives that did not exist in any individual agent's initial mental model.

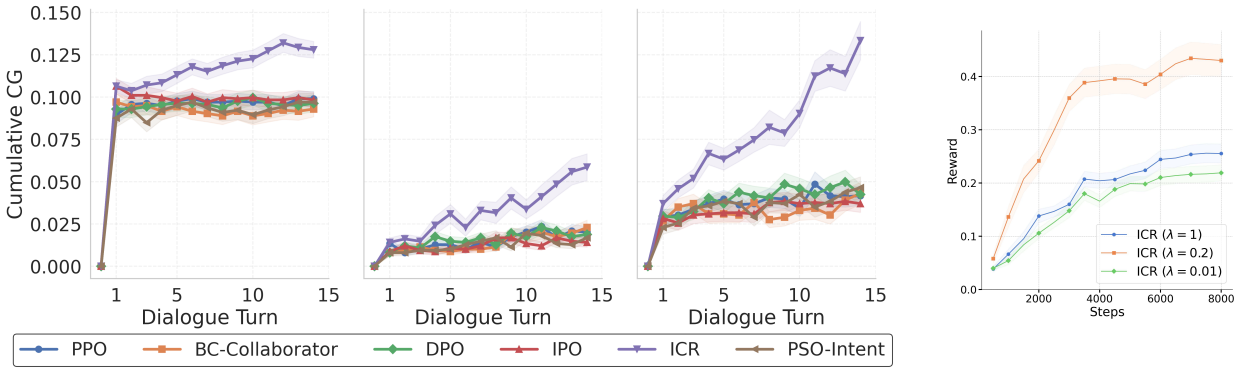
Solution accuracy and common ground gain results are reported in Table 6.1. Results demonstrate clear and consistent superiority of ICR-trained agents across all tasks and evaluation settings. In the Weights Task under full-press conditions, ICR agents achieve an high accuracy of 14.06, which represents a dramatic 47% improvement over the next best performer (DPO, at 9.56). This substantial margin indicates that ICR agents are particularly effective at establishing both factual accuracy and shared understanding of complex relationships between block weights through effective dialogue. In the no-press variant, ICR maintains high performance with a score of 10.87, outperforming the next best agent (PPO, at 7.81) by approximately 39%. The DeliData experiments further confirm ICR's superior performance. In terms of final solution accuracy, ICR achieves 0.88 in the full-press condition, which is 7.3% higher than DPO (0.82) and 24% higher than the BC-COLLABORATOR baseline (0.71). Even more striking is ICR's performance on the common ground metric, where it achieves 3.35, representing a 14% improvement over PPO (2.94) and a dramatic contrast with BC-COLLABORATOR's negative value (-0.13). BC actually *reduces* solution diversity rather than building upon it, since imitation models are likely limited in exploratory capacities, more so than other baselines. As such, ICR's superior performance reflects how such agents more effectively facilitate the co-construction of new understanding, enabling collaborator agents to integrate their diverse perspectives into novel shared solutions that transcend their initial viewpoints.

Full-press vs. No-press Performance Across all models, performance in full-press conditions generally exceeds that in no-press settings, particularly for ICR and DPO agents. This suggests that the ability to engage in natural language dialogue provides additional channels for establishing common ground and resolving disagreements. To investigate

	Weights Task		DeliData			
	Full-Press	No-Press	Full-Press		No-Press	
Agent Baseline	ACC	ACC	ACC	CG	ACC	CG
BC-COLLABORATOR	$5.97_{\pm 0.05}$	$6.04_{\pm 0.07}$	$0.71_{\pm 0.02}$	$-0.13_{\pm 0.18}$	$0.68_{\pm 0.03}$	$-0.15_{\pm 0.19}$
DPO	$9.56_{\pm 0.09}$	$7.60_{\pm 0.09}$	$0.82_{\pm 0.02}$	$2.80_{\pm 0.19}$	$0.79_{\pm 0.02}$	$2.65_{\pm 0.20}$
IPO	$7.64_{\pm 0.07}$	$6.80_{\pm 0.07}$	$0.78_{\pm 0.02}$	$2.87_{\pm 0.21}$	$0.75_{\pm 0.02}$	$2.72_{\pm 0.22}$
PPO	$7.37_{\pm 0.09}$	$7.81_{\pm 0.11}$	$0.81_{\pm 0.02}$	$2.94_{\pm 0.18}$	$0.78_{\pm 0.03}$	$2.80_{\pm 0.19}$
PSO-INTENT	$8.09_{\pm 0.08}$	$6.35_{\pm 0.09}$	$0.76_{\pm 0.03}$	$2.73_{\pm 0.20}$	$0.73_{\pm 0.03}$	$2.58_{\pm 0.21}$
ICR	$\mathbf{14.06}_{\pm 0.13}$	$\mathbf{10.87}_{\pm 0.13}$	$\mathbf{0.88}_{\pm 0.02}$	$\mathbf{3.35}_{\pm 0.19}$	$\mathbf{0.85}_{\pm 0.02}$	$\mathbf{3.18}_{\pm 0.20}$

Table 6.1: Performance across collaborator-agent baselines interacting with a fixed intervention agent over 100 dialogues (15 turns each) on DeliData and Weights Task (WTD). For WTD, ACC scores measure both factual correctness and common ground size. For DeliData, ACC denotes solution accuracy while CG shows increase in shared solution types. ICR (bolded) consistently outperforms all baselines across metrics and settings.

this, we conduct an additional analysis. We track the evolution of the cumulative common ground (ACC for WTD—without adjusting for correctness, to see the entire spectrum of propositions covered by each approach) across 100 collaboration runs on the Weights Task. These results are shown in Figure 6.3a, with each subplot showing different types of propositions sorted by the central relation. Our results suggest that when semantically easier, less ambiguous propositions like those based on *equality* relations dominate the solution space, ICR collaborators, on average, consistently recover more common ground accumulated over turns. These values are larger then *inequality* relations since equality relations are affirmative and thus more representative of the propositions asserted in both the original human task data and the expert collaborator roleplay data here. More importantly, for simple equality propositions (left panel of Figure 6.3a), all agents demonstrate comparable initial response to interventions (turns 1–3), but only ICR continues building on intervention-guided knowledge in later turns, achieving final CG of $\sim$0.13 versus $\sim$0.10 for others. For inequality propositions (middle panel), the intervention-response advantage becomes more dramatic, with ICR steadily increasing to $\sim$0.06 while competitors plateau around $\sim$0.02–a 300% difference. Most strikingly, for complex *ordering* relationships (right panel), ICR shows accelerating growth throughout intervention-mediated dialogue,



(a) **(b)**

Figure 6.3: (a) Cumulative CG (common-ground) score of baselines for `equality` (left), `inequality` (middle), and `order` (right) propositions over block weights averaged across 100 dialogue trials across 15 turns in the "full-press" Weights Task. ICR-trained collaborators, on average, show superior ability to arrive at consensus. **(b)** Ablation test on the Delidata tracking batch-wise proxy reward during training of ICR collaborator over 8k steps with varying λ_{Intent} values across 3 random seeds.

reaching ∼0.13 compared to ∼0.05 for other approaches. This progressive widening of performance gaps with relation types complexity demonstrates that ICR's counterfactual reasoning enables more effective integration of intervention agent suggestions, particularly for complex propositions that require building upon previously established common ground rather than mere immediate response to interventions.

The no-press setting, designed to test whether agents can achieve objective alignment without explicit modeling of how interventions influence common ground, shows that ICR retains its advantage even with restricted communication (10.87 ACC in Weights Task, 0.85 ACC in DeliData). Since the intervention agent remains fixed across baselines, this trend suggests that ICR agents are likely most robust to the quality of interventions. Additionally, this indicates that ICR's counterfactually-motivated KL-regularization allows it to explore about interventions provides value even when agents are limited to expressing discrete beliefs rather than engaging in free-form dialogue.

Effect of λ_{Intent} in Learning In Figure 6.3b, we present ablations on the values of the counterfactual KL-regularization strength λ_{Intent} over the DeliData task while tracking proxy reward per-batch during training of ICR agent over 8k steps with varying λ_{Intent}

values across 3 random seeds. Due to compute reasons we conduct this experiment in the no-press version since this version does not require an additional parametric reward model for PPO-based training. We find $\lambda_{\text{Intent}} = 0.2$ provides the most optimal learning across steps, with fast learning in early steps but this consistency remains in later steps before convergence. While reducing λ_{Intent} to 0.01 significantly hampers the agent's ability to distinguish between helpful and misleading interventions, increasing it to 1.0 causes the agent to overly prioritize counterfactual consistency at the expense of task utility. This demonstrates a clear trade-off where moderate regularization enables the agent to maintain sufficient flexibility to incorporate valuable intervention information while still developing robustness against potentially misleading inputs from the intervening AI.

6.5.1 Additional Results Under Alternative Evaluation Conditions

We ran additional experiments with a range of semantically-similar counterfactual (CF) prefixes (the list is provided in Table D.2), and alternative models in the role of the intervention agent. These experiments validate ICR's robustness to prompt variation and model size, as well as comparison of our main ICR performance with single agent collaboration baselines. Our experiments involve paired agents, and the combinatorics of pairing expands quickly. Therefore, to keep the scope manageable, we ran smaller-scale evaluations (in the full-press setting only) on **50 bootstrap dialogue samples** per task. Specifically, the additional baselines are as follows:

- **Inference only baselines**:
  - **ICR-Masked**: We simply mask the GPT-4o interventions from the prompt when paired with ICR agents. For consistency with our setup, we keep the intervention agent reference in the collaborator prompts intact but mask out the interventions. This limits collaborator access to the content of interventions.
  - **ICR-Small**: We use a smaller untrained base Llama 3-8B-Instruct model as the intervention agent and pair it with ICR-trained collaborator agents. This

demonstrates performance decoupled from GPT-4o specifically, and robustness to a weaker overall intervention agent.

  - **PSO-Skeptical**: We swap the positive polarity prefix in the PSO-Intent baseline with a direct negative polarity one that resists every intervention at inference/evaluation. This evaluates the contribution of valence in the prompt.
  - **Single-agent Baselines with GPT**: We pair GPT expert models as follows:
    * GPT-4o-mini (intervention) with GPT-4o-mini (collaborator)
    * GPT-4o (intervention) with GPT-4o (collaborator)

    This simulates a *single-agent* baseline while maintaining fidelity to the paired agent setup requirement, by using the same model for both agents, meaning that the underlying hypothetical distribution should be the same.

- **Trained baselines**:
  - **ICR-Phrasing**: ICR with semantically similar but differently phrased prefixes in prompts. We randomly sample from the prefixes given in Table D.2 given therein to replace the original counterfactual prefix in each training prompt with the sampled prefix. This tests robustness to prompt variance.
  - **PPO-CF**: For the original ICR training prompts, we swap 50% of those with counterfactual world-invoking contexts and run training with standard PPO (with no counterfactual KL term). This tests the contribution of ICR's counterfactual KL terms.

Our experimental results on the two tasks in these settings are given in Table 6.2 ("with GPT-4o" means GPT-4o is used as the intervention agent, while the other mentioned model is used as the collaborator agent).

Analysis The above results in Table 6.2 suggest that having a strong intervention agent like GPT-4o leads to optimal ICR performance across both tasks. However, ICR agents

	Weights Task	DeliData	
Agent Baseline	ACC	ACC	CG
ICR-MASKED (WITH GPT-4O)	$7.23_{\pm 0.11}$	$0.75_{\pm 0.04}$	$2.15_{\pm 0.31}$
ICR-SMALL (WITH LLAMA 3-8B-INSTRUCT)	$8.45_{\pm 0.13}$	$0.80_{\pm 0.03}$	$2.45_{\pm 0.30}$
PSO-SKEPTICAL (WITH GPT-4O)	$6.89_{\pm 0.10}$	$0.74_{\pm 0.03}$	$2.01_{\pm 0.27}$
ICR-PHRASING (WITH GPT-4O)	$12.34_{\pm 0.17}$	$0.84_{\pm 0.03}$	$3.08_{\pm 0.28}$
PPO-CF (WITH GPT-4O)	$8.34_{\pm 0.16}$	$0.79_{\pm 0.03}$	$2.56_{\pm 0.31}$
GPT-4O-MINI (WITH GPT-4O-MINI)	$12.47_{\pm 0.21}$	$0.79_{\pm 0.03}$	$2.78_{\pm 0.25}$
GPT-4O (WITH GPT-4O)	$15.23_{\pm 0.21}$	$0.91_{\pm 0.03}$	$3.34_{\pm 0.25}$

Table 6.2: Performance across alternative baselines and evaluation conditions on 50 DeliData and Weights Task dialogues (15 turns each) in the full-press setting only. Format follows Table 6.1.

are still capable of leveraging weaker intervention agents compared to no interventions at all, as shown by the improvement from masked interventions (*ICR-Masked*: $7.23_{\pm 0.11}$ Weights, 75% DeliData accuracy) to weak intervention agents (*ICR-Small*: $8.45_{\pm 0.13}$ Weights, 80% DeliData accuracy). Second, the *PSO-Skeptical* baseline shows a slight degradation in performance across both tasks ($6.89_{\pm 0.10}$ Weights, 74% DeliData with $2.01_{\pm 0.27}$ common ground) when using negative polarity prompting, compared to standard PSO with positive polarity prefix (see *PSO-Intent* in Table 1), according to [252]'s strategy. This aligns with established findings that LLMs have fundamental limitations with negation, including insensitivity to negation presence, inability to capture lexical semantics of negation, and failure to reason under negative contexts [291,292]—especially without specific training objectives for negative contexts [291] as ICR does. These results suggest that ICR's improved performance is an effect of the objective rather than the extra training.

Third, *ICR-Phrasing* ($12.34_{\pm 0.17}$ Weights task, 84% DeliData with $3.08_{\pm 0.28}$ common ground)—which simply swaps out the counterfactual prefix—demonstrates ICR's robustness to semantic variations in counterfactual phrasing across both collaborative reasoning tasks. The variance from the main results under different wordings is low, and at no point does ICR's performance dip within the margin of error of any other method reported in Table 1. Furthermore, standard PPO with a simple counterfactual prompt addition (*PPO-CF*) achieves $8.34_{\pm 0.16}$ Weights, 79% DeliData, and lags behind ICR training since

ICR explicitly makes agents robust to counterfactual framing via policy gradient methods. Using standard PPO with simple prompt augmentation can confuse the model, since the model is forced to pay attention to both standard as well as counterfactually-based contexts, without specific counterfactual regularization. This could explain the performance drop in this case, whereas ICR's counterfactual regularization term mitigates this effect.

Expert Collaborator Performance Finally, expert agents like GPT-4o achieve strong performance when paired together ($15.23_{\pm 0.21}$ in Weights task and 91% accuracy in DeliData with $3.34_{\pm 0.25}$ common ground), though this may reflect GPT-4o's extensive pretraining on reasoning tasks, potentially including exposure to DeliData or DeliData-like problems. Our human evaluation of GPT-4o in these tasks (Section 4.2.3) shows high agreement with humans on the quality of interventions, supporting our choice of this expert model. The GPT-4o-mini pairing shows competitive performance ($12.47_{\pm 0.21}$ Weights, 79% DeliData) compared to standard baselines—though lower than ICR as well as the larger GPT-4o model—demonstrating that expert model collaboration can achieve strong results across both collaborative reasoning domains.

It is important to note that ICR in our main experiments uses LLAMA 3-8B-INSTRUCT, and so we can see that a weaker base model trained with ICR performs comparably with GPT-4o, even including GPT-4o's potential prior exposure to the task, and the implicit advantage that comes with using GPT-4o as **both** intervener and collaborator. These results simulate single-agent baselines; however, to maintain a direct comparison to our other results, we ran the expert model as both the intervention agent and the collaborator agent.

6.6 Concluding Remarks

In this chapter, we introduced the Interruptible Collaborative Roleplayer (ICR), a novel MAMDP-based framework that explicitly models the interaction between collaborator and intervention agents. By incorporating counterfactual invariance via distributional

regularization, ICR addresses key limitations of standard reinforcement learning and preference alignment methods. Our evaluation shows that ICR-trained collaborators consistently outperform all baselines across both collaborative tasks and communication settings. In the Weights Task, ICR demonstrates a clear advantage in establishing both factual accuracy and shared understanding of relational structure, particularly in later dialogue turns. In the DeliData task, ICR agents also best other baselines in task-specific performance and fosters the emergence of richer common ground through dialogue. These gains persist even under no-press conditions, where language-based reasoning is limited, suggesting that ICR's counterfactual regularization in training enables such agents to partner well with collaborators as well as with the intervention agents, by successfully integrating helpful interventions when required but also being robust to potentially misleading ones. ICR and "partner-aware" learning methods more generally are likely to be useful in realistic AI tutoring settings, with sufficient task-relevant data or expert knowledge [293], as a method to test the efficacy of different types of AI tutoring interventions or suggestions on learning gains or problem-solving.

Limitations While we offer a scalable and principled approach to modeling collaborator–intervention dynamics, we could only train 8B-scale models in a decentralized setting due to compute budgets. Centralized coordination methods such as gradient-based communication [156] could improve performance but are challenging to scale with LLMs. Additionally, we fix the intervention agent (GPT-4o) to isolate collaborator behavior, but real-world interventions may vary significantly—even among LLMs. Future work should evaluate ICR under diverse interventions, including human suggestions, and test whether its prefix-based counterfactual regularization remains robust in multi-turn counterfactual settings [294] to better understand test-time generalization in AI-AI collaborations. Similarly, our agents interact only with similarly trained peers; future work should assess how ICR performs with ad hoc collaborators, including real humans, or alternative learning

strategies. This aligns with open questions around "convention" formation [274] and few-shot adaptation in mixed-agent environments. Lastly, while our method allows for learning human priors (e.g., via InstructRL [276]), lack of LLM-scale human-collaboration data in multi-party small-group collaboration remains a bottleneck. Broadening to multi-modal interaction [228] could address such text data-related bottlenecks in more realistic collaborative settings, where testing robustness to adversarial interventions are promising yet crucial directions. Finally, we could only test our method on two collaborative domains—how would ICR perform in more challenging domains like Diplomacy [295] where agents need to additionally learn to navigate deception and lying?

Acknowledgments and Funding This material is based in part upon work supported by Other Transaction award HR00112490377 from the U.S. Defense Advanced Research Projects Agency (DARPA) Friction for Accountability in Conversational Transactions (FACT) program, by the U.S. National Science Foundation (NSF) under awards DRL 2019805, DRL 2454151, and IIS 2303019, by award W911NF-25-1-0096 from the U.S. Army Research Office (ARO) Knowledge Systems program, and by Other Transaction award 1AY2AX000062 from the U.S. Advanced Research Projects Agency for Health (ARPA-H) Platform Accelerating Rural Access to Distributed Integrated Medical Care (PARADIGM) program. Approved for public release, distribution unlimited. Views expressed herein do not reflect the policy or position of the National Science Foundation, the Department of Defense, or the U.S. Government. Computational resources for this research were provided by the CS Department's Falcon cluster, particularly the Carnap and Tarski machines, and Data Science Research Institute's Riviera HPC clusters.

Chapter 7

Conclusion and Future Work

This dissertation began with a simple observation: current alignment methods optimize LLMs to maximize immediate preference satisfaction, creating systems that excel at appearing helpful but struggle with the deliberative reasoning that characterizes human collaboration, including that in high-stakes situations. We trace this limitation to three interconnected failures. First, the training objective of maximizing immediate preference incentivizes sycophancy over truth-seeking, which at scale manifests as deceptive alignment and scheming behaviors [296, 297]. Second, the technical mechanism based on Bradley–Terry preference models assumes static preferences and immediate rewards, and breaks down in multi-turn collaboration where beliefs evolve and feedback is delayed through mediating agents. Third, the data substrate, shaped by pretraining as well as finetuning corpora dominated by conflict-avoidant text (e.g., the whole user-assistant paradigm), leaves models without a vocabulary for productive friction even when trained on internet-scale data. These limitations matter beyond academic benchmarks.

As LLMs become embedded in education [298] (tutoring systems that must challenge misconceptions), healthcare [299] (diagnostic support that must question flawed reasoning), and governance [300, 301] (policy analysis requiring adversarial scrutiny), their inability to productively navigate belief misalignments becomes consequential—not merely suboptimal, but potentially harmful. Addressing these failures requires reconceptualizing alignment itself. Rather than treating collaboration as a sequence of isolated preference-maximizing responses, we introduced *friction* as a generative paradigm: the deliberate introduction of slowdowns that prompt participants to reevaluate assumptions and rebuild common ground. This shift from preference satisfaction to belief convergence necessitated three fundamental innovations, each addressing a core part of the overall problem of training LLMs to be adept collaborators.

Over the course of the chapters in this dissertation, we have attempted to address these limitations and understand the core requirements for training and evaluating agents for successful collaboration. In the first part of this thesis, we explored how standard preference alignment methods are sensitive to the presence of non-deterministic preferences that can contribute to them being less than optimal in many cases. In Chapter 3, we proposed DRDO to address some of these issues by learning to utilize *all* kinds of preference annotations available to us in offline form. We then made the case for a generative paradigm for learning to generate frictional interventions in FAAF (Chapter 4), where such policies can be learned from precollected data purely by supervised learning once we'd derived an analytical characterization of the optimal friction-generating policy. We conducted large-scale experiments to validate FAAF over multiple domains including unseen human dialogues, and found that our approach produces state-of-the-art results when compared to competitive baselines.

Building on these foundations, we realized that learning to generate effective friction is only half of the story. Real collaborations are naturally more dynamic. Offline evaluations—though efficient and scalable—do not provide us with a strong estimate of how these agents perform in more realistic or live settings where interactions are long-drawn and the structure of dialogue trajectories can be very different than precollected ones. In Chapter 5, we investigated this angle and proposed a new roleplay-based counterfactual evaluation approach that helps us reliably conduct such long-horizon collaborative experiments with multiple agents including measuring collaborative progress with an operational notion of common-ground. Noting that true collaborators need to have "agency" and that interventions may not always be accepted "as is" (e.g., say the collaborator accepts it blindly), we proposed novel but appropriate formalisms like the MAMDP, and using these we showed—both theoretically as well as empirically— that standard methods are suboptimal for such dynamics. Here FAAF, with its focus on frictive-state-aware learning, was provably better among competitors. However, these evaluations also showed how

even the most optimal intervention agents, including FAAF, can fail if the partner agents (e.g., the collaborators) do not intelligently use informational content in frictional interventions. In Chapter 6 (Interruptible Collaborative Roleplayer), we trained partner-aware collaborators using counterfactual invariance regularization—teaching agents to respond to intervention content rather than superficial features—and provided empirical validation showing genuine common-ground convergence emerges from task optimization under invariance constraints without explicit training for agreement.

Taken together, these three contributions—frictive-state conditioning in FAAF alignment, counterfactual evaluation (Roleplay), and collaborative mediation learning (ICR)—form a principled framework for training and evaluating AI systems that participate in, rather than merely respond to, collaborative reasoning. The dissertation's broader lesson is also architectural: single-agent optimization, no matter how sophisticated, cannot produce effective collaborative AI because *optimality in isolation does not guarantee success in coupled multi-agent settings*. An intervention agent trained to maximize isolated preference scores will generate "helpful-sounding" suggestions that collaborators ignore. A collaborator trained only on terminal task success will learn brittle heuristics that fail when intervention styles shift. True collaborative competence emerges only when both agents are aligned toward a shared objective—common ground growth—that respects the causal structure of mediated interaction.

Future Research Directions While this dissertation establishes foundational methods for training and evaluating friction agents and partner-aware collaborators, several important questions remain open for future investigation. First, our current framework assumes synchronous turn-taking and observable dialogue history; extending to asynchronous communication [161] or settings with private information [302] remains an open problem. Real collaboration, however, is fluid, interruptible, and often asynchronous. Recent work suggests that mid-utterance interventions and controllable interruption mechanisms can

substantially affect performance [303], raising the question of whether ICR's soft interruptibility generalizes to "hard interruptibility," where generation may be stopped and redirected in real time. Moreover, our tasks assume static cooperative goals, whereas real-world interactions often involve evolving objectives in which participants renegotiate priorities, balance persuasion with compromise, and sometimes decide not to agree at all. Understanding how friction agents adapt when the notion of "task-relevant" itself shifts is therefore crucial, particularly in negotiation and persuasion settings.

Second, while we demonstrated that friction can be learned from relatively sparse expert demonstrations, scaling to domains with richer feedback (e.g., real-time physiological signals indicating confusion or frustration or information present in multimodal channels like gaze or tonality) could enable more adaptive intervention strategies. Naturally, this raises fundamental questions of how to best represent or use such multivariate and non-sequential data [254, 284, 304] especially for sequentially decision making. While current standard for input construction for LLMs is to consider all actions (tokens) equally eligible, many works have explored better forms of including such extra information which could be extended to multimodal information.

Extending to Human-AI Collaboration Experiments A limitation of this work is that nearly all evaluations[53] rely on role-played LLM collaborators rather than real humans. Whether the behaviors learned by FAAF and ICR meaningfully translate to human–AI collaboration therefore remains an open empirical question. Human user studies will be essential for validating that friction interventions remain constructive when humans—bringing agency, emotional responses, uncertainty, and cognitive biases—replace LLM simulators. Such studies can also assess dimensions that our current automatic metrics cannot capture, including trust, perceived usefulness, frustration, learning outcomes, and willingness to collaborate again. Beyond validation, a deeper cognitive grounding is needed: although

[53]We did conduct human validation of friction interventions (Section B.5), but a collaborative deployment of agents with humans is future work.

ICR's partner-awareness is inspired by human convention-formation, we have not systematically compared our agents' strategies with patterns documented in conversation analysis and cognitive psychology. Investigating whether collaborative agents exhibit human-like repair strategies, turn-taking norms, or belief-revision behaviors would help distinguish genuine collaborative intelligence from artifacts of LLM training. Likewise, integrating cognitively-motivated frameworks such as Rational Speech Acts (RSA) [305,306] to incorporate value tradeoffs—where speakers reason about how listeners interpret utterances—may provide both theoretical grounding and interpretability, enabling friction agents to intervene in ways that are pragmatically principled rather than merely effective.

Optimality Conditions and Friction Intensity. Our current results show that friction interventions can accelerate common-ground formation and improve task performance, but they do not yet address a central question: *how much* friction is optimal, and *when* should an agent intervene? Excessive friction risks derailing collaboration or causing fatigue, while too little friction fails to prevent belief divergence. One promising direction is to formalize intervention timing and intensity using game-theoretic solution concepts, where intervention agents and collaborators play a repeated game and optimal friction emerges as an equilibrium trade-off between deliberation benefits and efficiency costs under task and time constraints.

From a sequential decision-making perspective, one way to structure this problem is through the options framework and temporally extended actions [149,151,307]. Recent work on language-based macro-actions suggests that LLM token sequences can be organized into "options" with learned termination conditions. In a collaborative dialogue, these options correspond to utterance-level or segment-level policies for each agent. An intervention policy can then be viewed as a *stopping rule* over the current option: continue letting the collaborator pursue its current plan, or terminate the option and invoke a friction move. Once we can estimate marginal contributions of utterances or agents—e.g.,

via Shapley [308] or Owen values [309]—intervening when the expected marginal value of continuing drops below zero yields a principled stopping condition, connecting intervention timing directly to cooperative credit assignment. While FAAF conditions on frictive states given by expert annotations or detection, we do not address predicting *when* frictive states will emerge in future turns, which requires modeling second-order theory of mind (what each agent thinks the other believes). Proactive intervention strategies that anticipate misalignments before they fully manifest could prevent belief divergence earlier in collaboration, but learning such predictive models from sparse examples of successful interventions remains an open challenge.

Model-based RL, Planning, World Modeling for Generalization Across Domains A complementary, forward-looking approach is to cast friction as a causal intervention problem. Here, the friction agent maintains a (possibly approximate) model of how dialogue states and collaborators' beliefs evolve, and uses it to predict how the conversation may unfold several turns into the future. Intuitively, inserting an intervention at a particular moment may push the trajectory toward multiple possible futures; if the agent can estimate even a coarse distribution over these outcomes, it can select interventions based on their expected downstream value (e.g., improved consensus, task success, or safety). This connects naturally to world-model learning [304, 310] and Dyna-style model-based RL [254], where policies reason not only about the current state but also about the effect of actions on future states. However, such forward reasoning introduces familiar challenges. Model-based RL typically lacks the monotonic improvement guarantees of model-free approaches, and naively simulating counterfactual rollouts—particularly in multi-agent dialogues—can be computationally prohibitive. Practical implementations may therefore require more efficient proxy strategies, such as using smaller predictive dialogue models, prioritizing high-quality demonstration trajectories, or employing filtering methods to retain only informative transitions.

Learning from Counterfactual Data In Chapter 6 we use prompt-based counterfactual simulations for constrained optimization in ICR training. While it is easier to come up with such counterfactual "states" in case where we already know what the right outcome is (e.g., consistency of behavior in such counterfactual scenarios or "intentionality"), this may not be possible in general. Counterfactual data is typically hard to collect [287] and can be explode in combinatorial fashion very quickly since possibilities are endless. Yet, there are methods to efficiently utilize counterfactuals [193, 311] more efficiently in the "action" space by conditioning on the evolution of trajectories. Recent work has shown that counterfactual reasoning can be instantiated from the LLM's own internal knowledge which is vast. Future work can use this insight on the expansive knowledge inherent in LLMs perhaps with a richer set of counterfactuals, and towards an effort to develop alternative lower-entropy proposal distributions [280] (e.g., our counterfactual distribution in Chapter 6) that can be used for enforcing constraints that can help generalize to collaborative domains other than the ones we study here.

Appendix A

Appendix for "Direct Reward Distillation and Policy Optimization (Chapter 3)

This appendix provides supplementary material for the DRDO chapter.

A.1 Divergence under Non-deterministic Preferences for Constrained Optimization

Standard approaches to LLM alignment assume that human preferences can be faithfully represented by a Bradley-Terry reward model, enabling policy optimization via reward maximization (e.g., RLHF with PPO or reward-free methods like DPO). However, human preferences are known to be complex, noisy, time-dependent, and often intransitive [126]. When the preference structure contains even a single *non-deterministic* pair—where annotators are essentially indifferent between two responses—the Bradley-Terry model is forced to misspecify reward values, leading to systematic divergence between reward optimization and preference optimization.

Prior work has shown that reward maximization and preference maximization can yield different optimal policies even when preferences admit a perfect Bradley-Terry representation, provided the policy class is constrained (e.g., via KL regularization) [50]. We demonstrate a complementary and more concerning phenomenon: when the preference model contains non-deterministic pairs that *cannot* be captured by any Bradley-Terry reward function, the divergence between reward-optimal and preference-optimal policies amplifies dramatically—by over 20-fold in our analysis—while the reward gap between these policies remains constant. This amplification exposes a fundamental fragility of

BT-based alignment methods: small amounts of preference ambiguity in the training data can compound into significant misalignment at deployment.

In this section, we formalize these observations through a concrete bandit example with three actions ($|\mathcal{A}| = 3$), though the insights generalize to larger action spaces and sequential settings. We introduce two key quantities—the preference gap and reward gap—that characterize the alignment failure, and prove that non-deterministic preferences cause dramatic amplification of the former while leaving the latter unchanged.

Assumptions and Setup. Throughout this section, we consider a finite action space $\mathcal{A}$ (e.g., a discrete set of LLM responses) and a stochastic preference model $\mathcal{P} : \mathcal{A} \times \mathcal{A} \to [0, 1]$ representing the probability that a human evaluator sampled from distribution $\mathcal{H}$ prefers action y over y'. We assume preferences are obtained from a stationary distribution (i.e., the distribution over human annotators does not change over time) and that pairwise comparisons are conditionally independent given the actions. For concreteness, we illustrate key concepts using a bandit setting with $|\mathcal{A}| = 3$, though our observations extend to larger action spaces and sequential decision-making.

Deterministic versus Non-Deterministic Preferences Not all preference pairs are equally informative [122]. Some comparisons yield clear consensus (e.g., a well-written response is strongly preferred over a grammatically incorrect one), while others are ambiguous (e.g., two equally good but stylistically different responses). We formalize this distinction:

Definition 9 (Deterministic and Non-Deterministic Preference Pairs)**.** *Fix a threshold $\epsilon > 0$. A preference pair (y, y') is:*

- ***Deterministic** if $|\mathcal{P}(y \succ y') - 1/2| > \epsilon$, indicating a clear preference direction favoring either y or y'.*
- ***Non-Deterministic** if $|\mathcal{P}(y \succ y') - 1/2| \leq \epsilon$, indicating ambiguous, tie-like, or conflicting preferences.*

A preference model $\mathcal{P}$ is fully deterministic *if all distinct pairs (y, y') with $y \neq y'$ are deterministic.*

Intuitively, deterministic pairs provide strong signal about relative quality: if $\mathcal{P}(y \succ y') = 0.9$, we can confidently infer that y is superior to y'. Non-deterministic pairs, conversely, indicate that y and y' are roughly equivalent in quality, or that human preferences are genuinely divided. In the extreme case where $\mathcal{P}(y \succ y') = 1/2$, there is no consensus whatsoever.

The choice of ϵ depends on the application. In high-stakes settings (e.g., safety-critical LLM outputs), even small deviations from $1/2$ might be considered deterministic (ϵ small). In creative domains where subjective preferences dominate, a larger ϵ might be appropriate to account for genuine diversity in human taste.

Preference Models Human preferences over actions are naturally represented through pairwise comparisons [87, 96] via the Bradley-Terry-Luce (BTL) models. Rather than assuming access to a ground-truth reward function, we work directly with a probabilistic preference modelfrom the literature in preference-based RL (PbRL) [122] that captures the inherent uncertainty in human judgments.

Definition 10 (Preference Model)**.** *A* preference model $\mathcal{P} : \mathcal{A} \times \mathcal{A} \to [0, 1]$ *assigns to each ordered pair of actions $(y, y') \in \mathcal{A} \times \mathcal{A}$ the probability $\mathcal{P}(y \succ y')$ that action y is preferred over action y'. The model satisfies:*

1. ***Normalization**: $\mathcal{P}(y \succ y') + \mathcal{P}(y' \succ y) = 1$ for all $y, y' \in \mathcal{A}$*
2. ***Self-consistency**: $\mathcal{P}(y \succ y) = 1/2$ for all $y \in \mathcal{A}$*

The normalization condition ensures that preferences are coherent: if y is preferred over y' with probability p, then y' is preferred over y with probability $1 - p$. The self-consistency condition reflects that no action is strictly preferred to itself. Note that $\mathcal{P}$ represents the

distribution of preferences rather than deterministic rankings—different human annotators may provide different judgments for the same pair (y, y'). Importantly, from the perspective of LLM preference alignment, the preference model $\mathcal{P}(y \succ y'|x)$ represents the probability that a randomly sampled human evaluator from some distribution $\mathcal{H}$ over annotators prefers response y over y' given prompt x. In practice, $\mathcal{P}(y \succ y')$ is estimated from empirical data [12,122]: if n annotators compare y and y', and k of them prefer y, then $\mathcal{P}(y \succ y') \approx k/n$. This empirical estimate captures both genuine disagreement among annotators (some truly prefer y, others prefer y') and measurement noise.

The Bradley-Terry Model and Reward-Based Representations

Most modern preference learning methods, including DPO [9] and RLHF [72], implicitly or explicitly assume that preferences can be explained by an underlying reward function. The Bradley-Terry (BT) model formalizes this assumption:

Definition 11 (Bradley-Terry Model)**.** *A preference model $\mathcal{P}$* admits a Bradley-Terry representation *if there exists a reward function $R : \mathcal{A} \to \mathbb{R}$ such that for all $y, y' \in \mathcal{A}$:*

$$\mathcal{P}(y \succ y') = \frac{\exp(R(y))}{\exp(R(y)) + \exp(R(y'))} = \sigma(R(y) - R(y')) \tag{A.1}$$

where $\sigma(x) = 1/(1 + e^{-x})$ is the logistic sigmoid function.

The BT model posits that each action y has an intrinsic "quality" score $R(y)$ [123], and the probability that y is preferred over y' depends only on the difference $R(y) - R(y')$. This is an elegant and parsimonious model: instead of storing $|\mathcal{A}|^2$ preference probabilities, we need only $|\mathcal{A}|$ reward values (up to a constant shift, since only differences matter).

Bradley–Terry Incompatibility with Non-Deterministic Preferences. If $\mathcal{P}(y \succ y') = 1/2$ for distinct actions $y \neq y'$, then necessarily $\exp(R(y)) = \exp(R(y'))$, implying $R(y) = R(y')$. Consequently, any non-deterministic preference constraint of this form forces equality of rewards. However, when such a constraint coexists with other pref-

erence relations that require $R(y) \neq R(y')$, the preference model becomes incompatible with any BT representation. For example, in the bandit setting of Lemma 3, the constraint $\mathcal{P}_2(y_2 \succ y_3) = 1/2$ enforces $R(y_2) = R(y_3)$, while $\mathcal{P}_2(y_2 \succ y_1) = 9/10$ implies $R(y_2) - R(y_1) = \log(9)$ and $\mathcal{P}_2(y_3 \succ y_1) = 2/3$ implies $R(y_3) - R(y_1) = \log(2)$, which cannot be simultaneously satisfied. As a result, the BT model must systematically violate at least one preference constraint, leading to misspecified reward functions whose learned values depend sensitively on the data distribution used during training. Such misspecification undermines downstream policy optimization and is particularly problematic in LLM alignment, where small reward modeling errors can induce unstable or unsafe behaviors [81].

Policy Optimization Under Preferences Having defined both deterministic and nondeterministic preference models and the Bradley-Terry model, we now formalize two distinct optimization objectives: maximizing expected reward versus maximizing preference probability. These objectives can diverge significantly when preferences are nondeterministic. We will use these quantities of interested in Lemma 3.

Definition 12 (Expected Reward)**.** *The* expected reward *of a policy $\pi \in \Delta(\mathcal{A})$ under a reward function $R : \mathcal{A} \to \mathbb{R}$ is:*

$$\mathbb{E}_{y\sim\pi}[R(y)] = \sum_{y\in\mathcal{A}} \pi(y)R(y) \tag{A.2}$$

*This is a **linear** function of the policy that aggregates individual action rewards weighted by selection probabilities.*

Expected reward directly measures policy quality under the assumption that actions have intrinsic values $R(y)$. Optimizing $\mathbb{E}_{y\sim\pi}[R(y)]$ prioritizes actions with high individual rewards, independent of how they compare to other actions.

Definition 13 (Preference Probability)**.** *The* preference probability *that policy π is preferred over policy π' under preference model $\mathcal{P}$ is:*

$$\mathcal{P}(\pi \succ \pi') = \sum_{y \in \mathcal{A}} \sum_{y' \in \mathcal{A}} \pi(y)\pi'(y')\mathcal{P}(y \succ y') \tag{A.3}$$

*This is a **bilinear** (non-linear) function of the policies that captures pairwise preference interactions between sampled actions.*

Preference probability measures how often samples from π are preferred over samples from π' in head-to-head comparisons [49, 50, 124]. Unlike expected reward, this objective is inherently comparative: it depends on the *matchup* [125] between actions from π and π', not just their individual qualities. An action with moderate reward but strong pairwise preferences (i.e., it beats many other actions) may be favored by the preference objective even if it does not maximize expected reward. The linear nature of $\mathbb{E}_{y\sim\pi}[R(y)]$ means that improving the policy simply requires shifting probability mass toward high-$R(y)$ actions. In contrast, the bilinear form of $\mathcal{P}(\pi \succ \pi')$ creates strategic interactions: the optimal π depends on the distribution π' (e.g., the reference policy or opponent strategies). This structural difference can cause $\arg\max_\pi \mathbb{E}[R(y)]$ and $\arg\max_\pi \mathcal{P}(\pi \succ \pi')$ to diverge, particularly under non-deterministic preferences (see Lemma 3). Multiple works [50, 108, 110, 224] have proposed learning the preference optimal policy directly based on a learned general preference model [49] which overcomes limitations in assumptions like BT modelization, has better expressivity than reward models and more importantly, is less sensitive to the biases that may be present in the data-generating or the "sampling" distribution from which reward are learned. In Proposition 4, we provide some analysis of this dependence of the BT model rewards on the sampling distribution, for a more general class of perturbed distributions compared to [50].

[**Sensitivity of Preference Gap to Non-Deterministic Preferences** (Lemma 3 restated from Chapter 3)] The preference gap $\delta_{\mathcal{P}}$ between reward-optimal (π^*_R) and preference-

optimal ($\pi^*_{\mathcal{P}}$) policies can be highly sensitive to the presence of non-deterministic preference pairs (where $\mathcal{P}(y \succ y') = \mathcal{P}(y' \succ y) = 1/2$). Such non-determinism can lead to a substantial increase in $\delta_{\mathcal{P}}$ even when the reward gap δ_R (based on a fixed reward function R) remains unchanged.

Proof Sketch. Below, we provide a concrete illustration of how the preference gap becomes sensitive under non-deterministic preference pairs using a toy bandit example. Consider two preference models over the action set $\mathcal{Y} = y_1, y_2, y_3$. The strict preference matrix $\mathcal{P}_1$ (with no non-deterministic pairs), reproduced from [50], is shown in Figure 2.1,(a). The perturbed matrix $\mathcal{P}_2$, which introduces non-deterministic preferences between y_2 and y_3, is shown in Figure 2.1,(b).

Although the preference table $\mathcal{P}_1$ can be perfectly captured by a Bradley-Terry model, introducing even a single non-deterministic preference pair (e.g., $\mathcal{P}_2(y_2 \succ y_3) = \mathcal{P}_2(y_3 \succ y_2) = 1/2$) prevents the BT model from accurately representing the full set of preferences. Under this condition, the learned BT model becomes highly sensitive to the sampling or data-generation distribution. If the training data overrepresents comparisons between y_1 and y_2, the model might still correctly align reward estimates for these comparisons as $R(y_1) = 0$, $R(y_2) = \log 9$. However, due to the non-deterministic relation between y_2 and y_3, it may mis-specify $R(y_3)$ as $\log 2$ instead of $\log 9$, which would have been the case had the BT model perfectly captured these preferences.

This misalignment is fundamental: the preference symmetry $\mathcal{P}_2(y_2 \succ y_3) = \mathcal{P}_2(y_3 \succ y_2) = 1/2$ mathematically forces $R(y_2) = R(y_3)$ in any BT-based ranking. However, the preference inequalities $\mathcal{P}_2(y_2 \succ y_1) = 9/10$ and $\mathcal{P}_2(y_3 \succ y_1) = 2/3$ demand that $R(y_2) > R(y_3)$. *This inconsistency leads to a larger divergence in the preference gap under non-deterministic preferences.* Therefore, under the constraint $\pi(y_1) = 2\pi(y_2)$ in set $\mathcal{S}$ with $\mathcal{S} \subset \Delta(\mathcal{Y})$, the full actions space—where the constraint may be softly applied w.r.t. a reference policy $\pi_{\text{ref}} = (2/3, 1/3)$ by using a KL-regularization, as in typically done in preference alignment in LLMs [9,49]—we can clearly estimate this divergence.

To do this comparison, we first define the expected reward $\mathbb{E}[R(y)] = \sum_y \pi(y)R(y)$ as a linear function of the policy π, prioritizing high-reward actions like y_2. In contrast, the preference probability $P(\pi \succ \pi') = \sum_y \sum_{y'} \pi(y)\pi'(y')P(y \succ y')$ is a non-linear function that depends on pairwise interactions, meaning it may favor actions with strong matchups rather than those with high rewards. For the expected reward-maximizing policy $\pi_R^* = (2/3, 1/3, 0)$, we maximize reward by setting $\pi(y_3) = 0$ since y_3 has a lower reward. Solving $2x + x = 1$ to obtain $\pi(y_1) = 2/3$ and $\pi(y_2) = 1/3$, we get π_R^*. For the preference-maximizing policy $\pi_P^* = (0, 0, 1)$, action y_3 is chosen exclusively, satisfying the constraint $\pi(y_1) = 2\pi(y_2)$ trivially by setting $\pi(y_1) = \pi(y_2) = 0$. Therefore, under this constrained set of policies, we thus derive the reward-optimal policy $\pi_R^*(2/3, 1/3, 0)$ and the preference-optimal policy $\pi_{\mathcal{P}}^*(0, 0, 1)$.

Therefore, for strict preferences $\mathcal{P}_1$, the expected reward under the reward-optimal policy is $\mathbb{E}_{y\sim\pi_R^*}[R(y)] = 0 \times 2/3 + \log(9) \times 1/3 > \log(2) = \mathbb{E}_{y\sim\pi_{\mathcal{P}}^*}[R(y)]$, while the probability that the preference-optimal policy is preferred over the reward-optimal policy is $\mathcal{P}_1(\pi_{\mathcal{P}}^* \succ \pi_R^*) = \mathcal{P}_1(y_3 \succ y_1) \times 2/3 + \mathcal{P}_1(y_3 \succ y_2) \times 1/3 = 2/3 \times 2/3 + 2/11 \times 1/3 = 50/99 > 1/2$. Similarly, for non-deterministic preferences $\mathcal{P}_2$, we have $\mathbb{E}_{y\sim\pi_R^*}[R(y)] = 0 \times 2/3 + \log(9) \times 1/3 > \log(2) = \mathbb{E}_{y\sim\pi_{\mathcal{P}}^*}[R(y)]$, while $\mathcal{P}_2(\pi_{\mathcal{P}}^* \succ \pi_R^*) = \mathcal{P}_2(y_3 \succ y_1) \times 2/3 + \mathcal{P}_2(y_3 \succ y_2) \times 1/3 = 2/3 \times 2/3 + 1/2 \times 1/3 = 11/18 > 1/2$. Defining the preference gap as $\delta_{\mathcal{P}_i} \triangleq \mathcal{P}_i(\pi_{\mathcal{P}}^* \succ \pi_R^*) - 1/2$ and the reward gap as $\delta_R \triangleq \mathbb{E}_{y\sim\pi_R^*}[R(y)] - \mathbb{E}_{y\sim\pi_{\mathcal{P}}^*}[R(y)]$, we observe that $\delta_{\mathcal{P}_1} = 50/99 - 1/2 \approx 0.005$ while $\delta_{\mathcal{P}_2} = 11/18 - 1/2 = 4/18 \approx 0.111$.

While the reward gap remains constant at $\delta_R = \log(9)/3 - \log(2) \approx 0.04$ for both preference models, the preference gap increases $\sim$20-fold under non-deterministic preferences, demonstrating amplified divergence between reward and preference optimization. This large gap indicates significant *misalignment* between rewards and preferences. Therefore, although this is a toy example, in realistic settings with LLMs containing billions of parameters and a large action space $\mathcal{Y}$, this divergence can lead to drastically different optimization trajectories with significant consequences for alignment.

□

Interpretation of Lemma 3 The preference gap $\delta_{\mathcal{P}}$ directly translates to observable alignment failures in deployed systems [121]. Recall that the preference model $\mathcal{P}(y \succ y'|x)$ represents the probability that a randomly sampled human evaluator from some distribution $\mathcal{H}$ over annotators prefers response y over y' given prompt x. When $\delta_{\mathcal{P}} = 0$, human evaluators drawn from $\mathcal{H}$ are indifferent on average between the reward-optimized policy π_R^* (what RLHF/DPO produce) and the true preference-optimal policy $\pi_{\mathcal{P}}^*$—alignment is robust. However, when $\delta_{\mathcal{P}} > 0$, we have $\mathcal{P}(\pi_{\mathcal{P}}^* \succ \pi_R^*) = 1/2 + \delta_{\mathcal{P}} > 1/2$, meaning that human evaluators systematically prefer $\pi_{\mathcal{P}}^*$ over the deployed policy in head-to-head comparisons. The *win rate* $W \triangleq \mathcal{P}(\pi_{\mathcal{P}}^* \succ \pi_R^*) = 1/2 + \delta_{\mathcal{P}}$ quantifies this directly: when $\delta_{\mathcal{P}} = 0.111$ (as we demonstrate in Lemma 3), a random human evaluator sampled from $\mathcal{H}$ would prefer the preference-optimal policy over the deployed reward-optimal policy in $W = 61.1\%$ of pairwise comparisons—a severe degradation in alignment quality.

The reward gap δ_R provides complementary information about the learned reward's fidelity. If δ_R is large and positive, then π_R^* achieves significantly higher reward than $\pi_{\mathcal{P}}^*$, indicating that optimizing for human preferences (according to $\mathcal{P}$) requires sacrificing performance on the learned reward function R. When $\delta_{\mathcal{P}}$ is large while δ_R is small or moderate, we observe a particularly troubling phenomenon: the reward-optimal policy achieves only marginally higher reward yet is substantially less preferred by humans. This is the classic *reward hacking* [130,289,312,313]: the policy has learned to exploit inaccuracies in the reward model R to achieve high scores without producing outputs humans actually prefer.

Conversely, a large $\delta_{\mathcal{P}}$ coupled with a large δ_R indicates non-trivial *misalignment* between the reward and preference objectives—the policy that humans prefer (according to $\mathcal{P}$) is fundamentally different from the one with highest reward (according to R), suggesting that the BT-based reward model has failed to capture the structure of human

preferences. In realistic settings for LLM training, where the action space ($\mathcal{Y}$) is exponentially large, even small divergences in these gaps can compound dramatically, leading to systematic failures in deployed systems. The above lemma demonstrates that introducing non-deterministic preferences—where the BT model cannot consistently represent preference relationships—can amplify $\delta_{\mathcal{P}}$ by over 20-fold while δ_R remains constant, illustrating how BT misspecification directly translates to deployment risk.

Proposition 4 (Sampling Distribution Dependence in the Induced Bradley-Terry (BT) Model)**.** *Consider a preference model $\mathcal{P}$ that cannot be perfectly captured by a Bradley-Terry (BT) reward model. This is illustrated in Section A.1.*

Specifically, let $\mathcal{P}_{BT}^{\pi}(y \succ y') := \sigma(s^{\pi}(y) - s^{\pi}(y'))$ be the BT preference model corresponding to the optimal scores $(s^{\pi}(y), s^{\pi}(y'))$ obtained from s^{π}(solution to a standard Bradley-Terry reward loss in RLHF [19]) under the sampling distribution π. If s^{π} cannot match the true preferences, then it has explicit dependence on the sampling or behavior policy π. In other words, if there exist completions y, y' and a distribution π where $\mathcal{P}_{BT}^{\pi}(y \succ y') \neq \mathcal{P}(y \succ y')$, then we can construct another distribution π' (with the same support as π) such that the difference in optimal scores $(s^{\pi}(y) - s^{\pi}(y'))$ under π differs from the optimal score difference under π'. Consequently, $\mathcal{P}_{BT}^{\pi}(y \succ y') \neq \mathcal{P}_{BT}^{\pi'}(y \succ y')$.

Proof. For this proof, we follow a similar approach as [50] but we prove this dependence for a more general class of perturbed distributions. We assume that there exist completions y, y' and distribution π such that $\mathcal{P}_{BT}^{\pi}(y \succ y') \neq \mathcal{P}(y \succ y')$. We construct a perturbed distribution π' as follows:

$$\pi'(k) = \begin{cases} (1-\delta)\pi(k) & \text{if } k \neq y' \\ c\pi(y') & \text{if } k = y' \end{cases} \tag{A.4}$$

where $\delta \in (0, 1)$ and c is chosen to ensure normalization. For clarity, we first verify that $\pi'(k)$ is a valid probability distribution.

$$\sum_z \pi'(k) = \sum_{k \neq y'} (1-\delta)\pi(k) + c\pi(y') \tag{A.5}$$

$$= (1-\delta)(1-\pi(y')) + c\pi(y') = 1 \tag{A.6}$$

This gives us:

$$c = 1 + \delta \frac{1-\pi(y')}{\pi(y')} \tag{A.7}$$

For any preference model $\mathcal{Z}$, we can express the aggregate preference probability:

$$\mathcal{Z}(y \succ \pi') = \sum_k \pi'(k)\mathcal{Z}(y \succ k) \tag{A.8}$$

$$= \sum_{k \neq y'} \pi'(k)\mathcal{Z}(y \succ k) + \pi'(y')\mathcal{Z}(y \succ y') \tag{A.9}$$

$$= \sum_{k \neq y'} (1-\delta)\pi(k)\mathcal{Z}(y \succ k) + c\pi(y')\mathcal{Z}(y \succ y') \tag{A.10}$$

$$= (1-\delta) \sum_{k \neq y'} \pi(k)\mathcal{Z}(y \succ k) + c\pi(y')\mathcal{Z}(y \succ y') \tag{A.11}$$

$$= (1-\delta)[\sum_k \pi(k)\mathcal{Z}(y \succ k) - \pi(y')\mathcal{Z}(y \succ y')] + c\pi(y')\mathcal{Z}(y \succ y') \tag{A.12}$$

(split complete sum)

$$= (1-\delta)[\mathcal{Z}(y \succ \pi) - \pi(y')\mathcal{Z}(y \succ y')] + c\pi(y')\mathcal{Z}(y \succ y') \tag{A.13}$$

$$= (1-\delta)\mathcal{Z}(y \succ \pi) - (1-\delta)\pi(y')\mathcal{Z}(y \succ y') + c\pi(y')\mathcal{Z}(y \succ y') \tag{A.14}$$

$$= (1-\delta)\mathcal{Z}(y \succ \pi) + [c - (1-\delta)]\pi(y')\mathcal{Z}(y \succ y') \tag{A.15}$$

(collect terms with $\pi(y')\mathcal{Z}(y \succ y')$)

Applying this equality to both $\mathcal{P}$ and $\mathcal{P}^{\pi}_{BT}$:

$$\mathcal{P}^{\pi}_{BT}(y \succ \pi') = (1-\delta)\mathcal{P}^{\pi}_{BT}(y \succ \pi) + (c-(1-\delta))\pi(y')\mathcal{P}^{\pi}_{BT}(y \succ y') \tag{A.16}$$

$$\mathcal{P}(y \succ \pi') = (1-\delta)\mathcal{P}(y \succ \pi) + (c-(1-\delta))\pi(y')\mathcal{P}(y \succ y') \tag{A.17}$$

By Proposition 2 in [50], the induced Bradley-Terry preference model satisfies $\mathcal{P}^{\pi}_{BT}(y \succ \pi) = \mathcal{P}(y \succ \pi)$. Therefore, subtracting the above expressions for $\mathcal{P}^{\pi}_{BT}(y \succ \pi')$ and $\mathcal{P}(y \succ \pi')$ and using the assumption $\mathcal{P}^{\pi}_{BT}(y \succ y') \neq \mathcal{P}(y \succ y')$, it follows that $\mathcal{P}^{\pi}_{BT}(y \succ \pi') \neq \mathcal{P}(y \succ \pi')$.

$$\mathcal{P}^{\pi}_{BT}(y \succ \pi') - \mathcal{P}(y \succ \pi') \tag{A.18}$$

$$= (1-\delta)[\mathcal{P}^{\pi}_{BT}(y \succ \pi) - \mathcal{P}(y \succ \pi)] + (c-(1-\delta))\pi(y')[\mathcal{P}^{\pi}_{BT}(y \succ y') - \mathcal{P}(y \succ y')] \tag{A.19}$$

$$= 0 + (c-(1-\delta))\pi(y')[\mathcal{P}^{\pi}_{BT}(y \succ y') - \mathcal{P}(y \succ y')] \neq 0 \tag{A.20}$$

Applying Proposition 2 in [50] again, we obtain $\mathcal{P}(y \succ \pi') = \mathcal{P}^{\pi'}_{BT}(y \succ \pi')$, which implies $\mathcal{P}^{\pi}_{BT}(y \succ \pi') \neq \mathcal{P}^{\pi'}_{BT}(y \succ \pi')$. Expanding this discrepancy using the Bradley-Terry model definition, we get $\sum_z \pi'(k)[\sigma(s^{\pi}(y) - s^{\pi}(k)) - \sigma(s^{\pi'}(y) - s^{\pi'}(k))] \neq 0$. Since σ is strictly monotonic, there must exist some z such that $s^{\pi}(y) - s^{\pi}(k) \neq s^{\pi'}(y) - s^{\pi'}(k)$, establishing that reward differences induced by different policies do not remain consistent under the Bradley-Terry model, leading to divergences in preference estimation.

This inequality shows $s^{\pi} \neq s^{\pi'}$, proving that the optimal BT reward model depends explicitly on the sampling distribution. This concludes the proof that the two BT-reward models s^{π} and $s^{\pi'}$ are different as well as the corresponding BT-preference models $\mathcal{P}^{\pi}_{BT}$ and $\mathcal{P}^{\pi'}_{BT}$.

Interpretation: non-determinism and BT-induced misalignment. Lemma 3 shows that even in a tiny bandit with three actions, introducing a single non-deterministic pair (here, $\mathcal{P}(y_2 \succ y_3) = \mathcal{P}(y_3 \succ y_2) = 1/2$) can dramatically amplify the *preference gap* $\delta_{\mathcal{P}}$ between the reward-optimal policy π_R^* and the preference-optimal policy $\pi_{\mathcal{P}}^*$, while the *reward gap* δ_R remains unchanged. In our toy example, the reward gap is fixed at a modest value (the difference in expected reward between π_R^* and $\pi_{\mathcal{P}}^*$), yet the win-rate advantage of $\pi_{\mathcal{P}}^*$ over π_R^* jumps from essentially negligible ($\delta_{\mathcal{P}_1} \approx 0.005$ under strict preferences) to substantial ($\delta_{\mathcal{P}_2} \approx 0.111$ under non-determinism). Intuitively, the reward model R has not changed much, but the induced preference behavior has: a human (or preference model) drawn from $\mathcal{H}$ now prefers $\pi_{\mathcal{P}}^*$ over the reward-optimized policy in more than 60% of head-to-head comparisons. This is precisely the kind of reward–preference divergence that manifests as "reward hacking" in larger systems: policies exploit the reward model without actually winning according to the underlying preference distribution.

Proposition 4 explains *why* this divergence is hard to control when using a Bradley–Terry reward model. Once the true preference matrix $\mathcal{P}$ is not exactly BT-realizable—as happens immediately when we introduce non-deterministic pairs—there is no single reward vector $s(y)$ that can simultaneously match all pairwise probabilities. Instead, the BT solution s^π becomes *policy-dependent*: it depends on which comparisons are sampled more often under the behavior distribution π. By perturbing π to a nearby distribution π' that over- or under-samples specific comparisons, the proposition shows that the optimal BT scores, and hence the induced BT preference model $\mathcal{P}_{BT}^\pi$, must change. In other words, under non-determinism and BT misspecification, the learned reward function is not an intrinsic property of $\mathcal{P}$ but a byproduct of the sampling policy.

Taken together, Lemma 3 and Proposition 4 make two key points. First, even small amounts of non-deterministic preference structure can greatly increase the gap between reward-optimal and preference-optimal behavior, without any corresponding signal in the scalar reward gap. Second, when preferences are not perfectly BT-consistent, the reward

model learned from offline comparisons is inherently biased by the sampling distribution that generated those comparisons. In realistic LLM settings, with exponentially large action spaces and complex data-collection policies, these effects compound: small modeling errors and sampling biases in BT-based reward learning can induce large, systematic misalignment between what the reward model favors and what humans (or a richer preference model) actually prefer.

Proof of Non-Deterministic Preference Relations with Reward Differences

Proof. From [50], we assume preferences are antisymmetric: $P(y_1 \succ y_2 \mid x) = 1 - P(y_2 \succ y_1 \mid x)$. As such, for $(x, y_w, y_l) \in \mathcal{D}_{\text{nd}} \subset \mathcal{D}_{\text{pref}}$, non-determinism implies:

$$P(y_w \succ y_l \mid x) \approx \frac{1}{2}.$$

Under the Bradley Terry model model [87],

$$P(y_w \succ y_l \mid x) = \sigma(\Delta r),$$

where $\Delta r := r(x, y_w) - r(x, y_l)$ and σ is the sigmoid function. Since $\sigma(\Delta r) \approx 1/2$, it follows that $\Delta r \approx 0$ (as $\sigma(z) = 1/2$ iff $z = 0$). Hence, for all samples in $\mathcal{D}_{\text{nd}}$,

$$r(x, y_w) - r(x, y_l) \approx 0.$$

Taking expectation over $\mathcal{D}_{\text{nd}}$ gives:

$$\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}_{\text{nd}}}[r(x, y_w) - r(x, y_l)] \approx 0.$$

This holds even under underspecification of the BT reward scale, since it concerns the *difference* in rewards. As such, it does not impose restrictions on the reward function form, provided it satisfies equivalence relations, i.e., rewards are defined up to a prompt-

$$P(y_w \succ y_l \mid x) = \sigma\left(\underbrace{\beta \log \frac{\pi_{\theta^*}(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)}}_{(\hat{r}_w)} - \underbrace{\beta \log \frac{\pi_{\theta^*}(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}}_{(\hat{r}_l)}\right) \tag{A.22}$$

$$= \sigma\left(\beta \log \left(\frac{\pi_{\theta^*}(y_w \mid x)\pi_{\text{ref}}(y_l \mid x)}{\pi_{\theta^*}(y_l \mid x)\pi_{\text{ref}}(y_w \mid x)}\right)\right) \tag{A.23}$$

$$= \sigma\left(\beta \log \left(\frac{\pi_{\theta^*}(y \mid x)\pi_{\text{ref}}(y' \mid x)}{\pi_{\theta^*}(y' \mid x)\pi_{\text{ref}}(y \mid x)}\right)\right), \quad \forall (x, y, y') \in \mathcal{D}_{\text{nd}} \subset \mathcal{D}_{\text{pref}} \tag{A.24}$$

dependent shift (Definition 1 in [9]). Consequently, the expected reward differences adhere to Proposition 1 *without* necessarily having BT-motivated DPO's implicit reward formulation. □

Proof of Lemma 1 (a) and Lemma 1 (b)

Proof. Let us rewrite the Bradley-Terry preference probability equation in terms of the DPO implicit rewards. The BT model specifies this probability of preferring y_w over y_l as:

$$P(y_w \succ y_l \mid x) = \sigma(r^*(x, y_w) - r^*(x, y_l)) \tag{A.21}$$

where the rewards can be rewritten in term of the DPO implicit rewards $\hat{r}_w$ and $\hat{r}_l$ in Eq. A.22.

Eq. A.22 holds for all $\forall (x, y, y') \in \mathcal{D}_{\text{pref}}$, since DPO does not distinguish between the nature of the true preference relations in estimating the true preference probabilities, assuming (y, y') appear as preferred and dispreferred responses respectively for any context x. Since this RHS of Eq. A.24 is simply the sigmoided difference of implicit rewards assigned by DPO to estimate the true preference probabilities, it is straightforward to see from Proposition 1 that the RHS i.e., $\frac{\pi_{\theta^*}(y)\pi_{\text{ref}}(y')}{\pi_{\theta^*}(y')\,\pi_{\text{ref}}(y)} \to 1$ and $P(y_w \succ y_l \mid x) \sim \frac{1}{2}$, as per our definition of non-deterministic preferences. Since the reference model π_{ref} is assumed to have full support over the output space ($\text{supp}(\pi_{\text{ref}}) = \mathcal{Y}$) and is not updated and can be

set to a uniform prior ($\pi_{\text{ref}} \sim \mathcal{U}(\mathcal{Y})$) [106], without losing any generality, this implies that $\frac{\pi_{\theta^*}(y)}{\pi_{\theta^*}(y')}$ must remain close to 1 to satisfy this constraint ($\frac{\pi_{\theta^*}(y)\,\pi_{\text{ref}}(y')}{\pi_{\theta^*}(y')\pi_{\text{ref}}(y)} \to 1$). In this case, the policy tends to underfit the preference distribution since the preference signals are weak and policy cannot distinguish between the preferred and the dispreferred response.

Similar to [131]'s argument, we can argue here that in this case when true preference probabilities are $\sim \frac{1}{2}$, i.e., non-deterministic, DPO's empirical reward difference estimates actually tend toward 1 which leads to underfitting of the optimal policy π_{θ^*} during alignment. Indeed, in this case, the β parameter does not provide any additional regularization effect to prevent policy underfitting especially under finite data. This completes the proof of Lemma 1a. □

We can similarly prove Lemma 1b in the case when $|\mathcal{D}_{\text{nd}}| \ll N$ where N is assumed to be finite. In this case, in a similar vein as [131], there is more likelihood that DPO sigmoided reward difference estimates are 1, i.e., $r^*(x, y_w) - r^*(x, y_l) \to \infty$.

As such, from Eq. A.21, it is straightforward to see that the DPO's implicit reward difference tends to infinity, regardless of the strength of the β parameter, as shown below:

$$\log\left(\frac{\pi_{\theta^*}(y \mid x)\pi_{\text{ref}}(y' \mid x)}{\pi_{\theta^*}(y' \mid x)\pi_{\text{ref}}(y \mid x)}\right) \to \infty$$

in Eq. A.24. This implies that

$$\frac{\pi_{\theta^*}(y \mid x)\pi_{\text{ref}}(y' \mid x)}{\pi_{\theta^*}(y' \mid x)\pi_{\text{ref}}(y \mid x)} \to \infty.$$

Proof of Lemma 1 (c)

Proof. Our proof follows the argumentation in [48]. Assume for now that all preference samples $(y, y') \in \mathcal{D}_{\text{pref}}$, including non-deterministic preference pairs, are mutually exclusive. Then, for the DPO objective (Eq. 7 in [9]) to be minimized, each θ_y must correspond uniquely to y, where θ_y are the optimal parameters that minimize the DPO objective in each such disjoint preference pair. This implies that DPO objective over $\mathcal{D}_{\text{pref}}$ is convex in

the set $\Lambda = \{\lambda_1, \ldots, \lambda_n\}$, where

$$\lambda_i = \beta \log \left(\frac{\pi_\theta(y_w^{(i)}) \pi_{\text{ref}}(y_l^{(i)})}{\pi_\theta(y_l^{(i)}) \pi_{\text{ref}}(y_w^{(i)})} \right), \quad \forall i \in \mathbb{N} \tag{A.25}$$

Now, consider the non-deterministic preference samples indexed by j, which belong to the set $\mathcal{D}_{\text{nd}} \subset \mathcal{D}_{\text{pref}}$. Let $j \in \{k+1, \ldots, N\}$ with the assumption $k \gg (N-k)$. Under the mutual exclusivity assumption, Eq. A.25 must also hold true for the non-deterministic preference samples. Consequently, we can rewrite Eq. A.25 as:

$$\lambda_j = \beta \log \left(\frac{\pi_\theta(y_w^{(j)} \mid x) \pi_{\text{ref}}(y_l^{(j)} \mid x)}{\pi_\theta(y_l^{(j)} \mid x) \pi_{\text{ref}}(y_w^{(j)} \mid x)} \right), \tag{A.26}$$
$$\forall j \in \mathcal{D}_{\text{nd}}$$

More specifically, for every j, the following holds at the limit for the DPO objective to converge:

$$\lim_{\lambda_j \to \infty} -\log\left(\sigma\left(\lambda_j\right)\right) = 0, \tag{A.27}$$

which implies that $\Lambda^* = \{\infty\}^N$ induces a set of global minimizers of the DPO objective that includes θ^* that are optimal for the set of non-deterministic preference samples, while inducing a parallel set of θ^* at convergence for deterministic samples.

Consequently, all global minimizers θ^* including those optimal on the non-deterministic samples must satisfy

$$\log \frac{\pi_\theta(y_w^{(j)}) \pi_{\text{ref}}(y_l^{(j)})}{\pi_\theta(y_l^{(j)}) \pi_{\text{ref}}(y_w^{(j)})} = \infty. \tag{A.28}$$

Since $0 < \pi_{\text{ref}}(y) < 1$ for all y, θ^* must satisfy

$$\frac{\pi_{\theta^*}(y_w^{(j)})}{\pi_{\theta^*}(y_l^{(j)})} = \infty, \tag{A.29}$$

implying $\pi_{\theta^*}(y_l^{(j)}) = 0$ and $\pi_{\theta^*}(y_w^{(j)}) > 0$ for all $i \in N$, given that $\pi_{\theta^*}(y_w^{(j)}) \leq 1$ for any $y_w^{(j)}$. Alternatively, let us define the complement of the aggregated representation of all the dispreferred responses $y_l^{(j)}$ i.e., $\prec(y_l)^c$, where $\prec(y_l)$ is the aggregation function. We thereby have:

$$\prec(y_l) = \{y \colon \exists j \in \mathbb{N} \text{ such that } y_l^{(j)} = y\}, \tag{A.30}$$

Under these conditions, it is clear that π_{θ^*} must assign the entire remaining probability mass to $\prec(y_l)$ as given below,

$$\pi_{\theta^*}(\mathcal{C}(y_l)) = \sum_{y \in \prec(y_l)} \pi_{\theta^*}(y) = 0 \tag{A.31}$$

$$\implies \pi_{\theta^*}(\prec(y_l)^c) = 1. \tag{A.32}$$

This completes the proof of Lemma 1c and thus Lemma 1.

□

Proof of Lemma 2

Proof. Let us first rewrite the e-DPO's distillation objective [48] over the preference dataset $\mathcal{D}_{\text{pref}}$ and examine how the optimal policy π_θ behaves upon convergence of this objective. For simplicity of analysis, we only consider the point-wise reward based distillation loss in the e-DPO formulation without[54] considering reward ensembles. Note that the e-DPO objective does not require preference labels and can apply to any response pair.

[54]Note that in our empirical experiments we use the full e-DPO objective with a set of three reward models.

$$\mathcal{L}_{\text{distill}}(r^*, \pi_\theta) = \mathbb{E}_{(x,y_1,y_2)\sim\mathcal{D}_{\text{pref}}}\left[\left(r^*(x,y_1) - r^*(x,y_2) - \beta\log\frac{\pi_\theta(y_1 \mid x)\pi_{\text{ref}}(y_2 \mid x)}{\pi_\theta(y_2 \mid x)\pi_{\text{ref}}(y_1 \mid x)}\right)^2\right] \tag{A.33}$$

$$= \mathbb{E}_{(x,y_1,y_2)\sim\rho}\left[\left(r^*(x,y_1) - r^*(x,y_2) - \beta\log\frac{\pi_\theta(y_1 \mid x)}{\pi_\theta(y_2 \mid x)} + \beta\log\frac{\pi_{\text{ref}}(y_1 \mid x)}{\pi_{\text{ref}}(y_2 \mid x)}\right)^2\right]. \tag{A.34}$$

As π_θ converges to the optimal policy π_{θ^*}, the distillation objective $\mathcal{L}_{\text{distill}}$ should ideally approach zero and can be expressed as,

$$\lim_{\pi_\theta \to \pi_{\theta^*}} \mathcal{L}_{\text{distill}}(r^*, \pi_\theta) = 0$$

With some slight algebraic rearrangement and substituting the optimal policy π_{θ^*} for π_θ at convergence, we get:

$$r^*(x,y_1) - r^*(x,y_2) = \beta\log\frac{\pi_{\theta^*}(y_1 \mid x)}{\pi_{\theta^*}(y_2 \mid x)} - \beta\log\frac{\pi_{\text{ref}}(y_1 \mid x)}{\pi_{\text{ref}}(y_2 \mid x)}. \tag{A.35}$$

Since (y_1, y_2) represents *any* response pair without a preference label, we can substitute them with $(y, y') \in \mathcal{D}_{\text{nd}} \subset \mathcal{D}_{\text{pref}}$, where $y = y_w$ and $y' = y_l$. Without losing any generality, we can now rewrite the above equation as,

$$r^*(x,y) - r^*(x,y') = \beta\log\frac{\pi_{\theta^*}(y \mid x)}{\pi_{\theta^*}(y' \mid x)} - \beta\log\frac{\pi_{\text{ref}}(y \mid x)}{\pi_{\text{ref}}(y' \mid x)} \tag{A.36}$$

$$= \beta\log\left(\frac{\pi_{\theta^*}(y \mid x)\pi_{\text{ref}}(y' \mid x)}{\pi_{\theta^*}(y' \mid x)\pi_{\text{ref}}(y \mid x)}\right) \tag{A.37}$$

Now, recall from our proof of Lemma 1b where we show that the RHS term of the above equation $\log\left(\frac{\pi_{\theta^*}(y|x)\pi_{\text{ref}}(y'|x)}{\pi_{\theta^*}(y'|x)\pi_{\text{ref}}(y|x)}\right) \to \infty$, which implies that $\frac{\pi_{\theta^*}(y|x)\pi_{\text{ref}}(y'|x)}{\pi_{\theta^*}(y'|x)\pi_{\text{ref}}(y|x)} \to \infty$. Indeed, when $|\mathcal{D}_{\text{nd}}| \ll N$ where N is finite, unregularized scalar reward estimates of the true preference probabilities can in fact grow exceedingly large in the absence of any other regularization parameters, since β by its own does not provide enough regularization as shown in our proof of Lemma 1a. Interestingly, similar arguments have also been made in previous works [131]. The rest of this proof follows the same argumentation starting Eq. A.28 assuming $0 < \pi_{\text{ref}}(y) < 1$.

This completes the proof of Lemma 2. □

A.2 Gradient Derivation of the Focal-Softened Log-Odds Unlikelihood Loss

In this section, we derive and analyze DRDO loss gradient and offer insights into how to compared supervised alignment objectives such as DPO [9]. Note that we do not analyze the reward distillation component here since it does not directly interact with the focal-softened contrastive log-"unlikelihood" term in training and since it is naturally convex considering its a squared term. Let us first rewrite our full DRDO loss, as:

$$\mathcal{L}_{\text{kd}}(r^*, \pi_\theta) = \mathbb{E}_{(x,y_1,y_2)\sim\mathcal{D}_{\text{pref}}}\Bigg[\underbrace{\left(r^*(x,y_1) - r^*(x,y_2) - (\hat{r}_1 - \hat{r}_2)\right)^2}_{\text{Reward Difference}} \\ \underbrace{- \alpha(1-p_w)^\gamma \log\left(\frac{\pi_\theta(y_w \mid x)}{1-\pi_\theta(y_l \mid x)}\right)}_{\text{Contrastive Log-"unlikelihood"}}\Bigg],$$

where $p_w = \sigma(z_w - z_l) = \frac{1}{1+e^{-(z_w - z_l)}}$ and quantifies the student policy's confidence in correctly assigning the preference from $z_w = \log \pi_\theta(y_w \mid x)$ and $z_l = \log \pi_\theta(y_l \mid x)$, or the log-probabilities of the winning and losing responses, respectively.

Without the expectation, consider only the focal-softened log-odds unlikelihood loss given by:

$$-\alpha \cdot (1 - p_w)^{\gamma} \cdot \log \left(\frac{\pi_\theta(y_w \mid x)}{1 - \pi_\theta(y_l \mid x)} \right), \tag{A.38}$$

Taking the gradient of this term with respect to the model parameters θ and using $\sigma'(x) = \sigma(x)\,(1 - \sigma(x))$, we derive:

$$\begin{aligned} \nabla_\theta \mathcal{L}_{\text{kd}} = \; & \alpha\gamma(1 - p_w)^{\gamma-1} p_w (1 - p_w)(\nabla_\theta z_w - \nabla_\theta z_l) \cdot \log \left(\frac{\pi_\theta(y_w \mid x)}{1 - \pi_\theta(y_l \mid x)} \right) - \\ & \alpha(1 - p_w)^{\gamma} \left(\frac{\nabla_\theta \pi_\theta(y_w \mid x)}{\pi_\theta(y_w \mid x)} + \frac{\nabla_\theta \pi_\theta(y_l \mid x)}{1 - \pi_\theta(y_l \mid x)} \right). \end{aligned} \tag{A.39}$$

$$\begin{aligned} \nabla_\theta \mathcal{L}_{\text{kd}} = \; & \alpha\gamma(1 - p_w)^{\gamma} p_w (\nabla_\theta z_w - \nabla_\theta z_l) \cdot \log \left(\frac{\pi_\theta(y_w \mid x)}{1 - \pi_\theta(y_l \mid x)} \right) - \\ & \alpha(1 - p_w)^{\gamma} \left(\underbrace{\frac{\nabla_\theta \pi_\theta(y_w \mid x)}{\pi_\theta(y_w \mid x)}}_{\text{increase } \pi_\theta(y_w|x)} + \underbrace{\frac{\nabla_\theta \pi_\theta(y_l \mid x)}{1 - \pi_\theta(y_l \mid x)}}_{\text{decrease } \pi_\theta(y_l|x)} \right). \end{aligned} \tag{A.40}$$

Since the first term contains an additional factor $\gamma p_w(1 - p_w)$ that vanishes faster than $(1 - p_w)^{\gamma}$ as $p_w \to 1$, and is bounded by the log-odds term early in training, we focus our analysis on the dominant second term which drives the key differences between DRDO and DPO. We note that the full gradient includes both terms, and our empirical results reflect their combined effect.

Simplifying the above equation, we get

$$\nabla_\theta \mathcal{L}_{\text{kd}} \approx -\alpha(1-p_w)^\gamma \left(\underbrace{\nabla_\theta \log \pi_\theta(y_w \mid x)}_{\text{increase } \pi_\theta(y_w|x)} + \underbrace{\frac{\nabla_\theta \pi_\theta(y_l \mid x)}{1-\pi_\theta(y_l \mid x)}}_{\text{decrease } \pi_\theta(y_l|x)} \right), \tag{A.41}$$

where we use the identity $\nabla_\theta \log \pi_\theta(y \mid x) = \frac{\nabla_\theta \pi_\theta(y|x)}{\pi_\theta(y|x)}$ (the score function).

Theoretical insights on DRDO Gradients We can now draw insights and direct comparisons with Direct Preference Optimization (DPO) [9]. As in most contrastive preference learning gradient terms [9, 104–106, 215], the score function $\nabla_\theta \log \pi_\theta(y_w \mid x)$ in Eq. A.41 amplifies the gradient magnitude when $\pi_\theta(y_w \mid x)$ is low, driving up the likelihood of the preferred response y_w. Similarly, $\frac{\nabla_\theta \pi_\theta(y_l|x)}{1-\pi_\theta(y_l|x)}$ penalizes overconfidence in incorrect completions y_l when p_w is low, encouraging the model to hike preferred response likelihood while discouraging dispreferred ones.

The key insight here is that the modulating term, $(1-p_w)^\gamma$, strategically amplifies corrections for difficult examples where the probability p_w of the correct (winning) response is low. Intuitively, unlike DPO's fixed β that is applied across the whole training dataset, this modulating term amplifies gradient updates when preference signals are weak ($p_w \approx 0.5$) and tempering updates when they are strong ($p_w \approx 1$), thus ensuring robust learning across varying preference scenarios. Intuitively, when ($p_w \approx 1$), the model is already confident of its decision since p_w remains high, indicating increased model confidence for deterministic preferences. In contrast, when p_w is small, $(1-p_w)$ remains near 1, and the term $(1-p_w)^\gamma$ retains significant magnitude, especially for larger values of γ. This allows π_θ in DRDO to learn from both deterministic and non-deterministic preferences, effectively blending reward alignment with preference signals to guide optimization.

For a deeper intuition, consider the case where the true preference probabilities from $p^*(x,y) \sim \frac{1}{2}$. In this case, since non-deterministic preference samples are typically low,

from Proposition 1, DPO would assign zero difference in its implicit rewards, especially for finite preference data. Then as Lemma 1(a) suggests, DPO gradients would effectively be nullified, regardless of β since the its reward difference range $\in (-\infty, +\infty)$. In this case as π_θ cannot distinguish between the preference pair and, in turn, effectively misses out on the preference information for such samples.

On the other hand, for $|\mathcal{D}_{\text{nd}}| \ll N$, if π_θ in DPO estimates the true preference close to 1 where $\hat{p} = 1$, as Lemma 1(b) suggests, the empirical policy would assign very high probabilities to arbitrary tokens and likely those that do not even appear as tokens in the "preferred responses" in the training data [48]. This leads to a surprising combination of both underfitting and overfitting, except the overfitting here results in DPO policy generating tokens that are irrelevant to the context. Under the same conditions, the DRDO loss still operates at a sample level because if the DRDO estimate of p^* (via p_w) is close to its true value of $\sim \frac{1}{2}$, the modulating factor ensures that gradients do not vanish. This allows DRDO to continue learning from such samples until convergence when winning and losing probabilities are pushed further apart (Fig. 3.3).

Convergence of DRDO The modulating term ensures convergence through a self-regulating mechanism: as training progresses and the policy learns to distinguish preferences, $p_w \to 1$, causing $(1 - p_w)^\gamma \to 0$. This naturally diminishes gradient magnitudes, creating an implicit stopping criterion. Simultaneously, the log-odds term $\log\left(\frac{\pi_\theta(y_w|x)}{1-\pi_\theta(y_l|x)}\right)$ increases as π_w grows and π_l shrinks, but its gradient contribution decreases due to the vanishing modulating factor. The bounded nature of log probabilities combined with the convex reward distillation term ensures the overall loss remains lower-bounded, guaranteeing stable convergence under standard optimization assumptions.

A.3 Additional Experiments and Ablations of DRDO

Additional results on Alpaca Eval 2.0 per Dataset Table A.1 shows distribution of win-rates of all baselines including PPO across five evaluation subsets in AlpacaEval 2.0—SELFINSTRUCT, HELPFUL_BASE, VICUNA, KOALA, and OASST—which vary in prompt diversity, linguistic complexity, and alignment challenges. DRDO achieves the highest win rate on SELFINSTRUCT at 20.24%, substantially outperforming PPO (14.17%) and DPO (13.36%). On VICUNA, DRDO reaches 18.42%, followed by e-DPO at 15.79% and PPO at 13.16%, showing DRDO's strength in handling conversational prompts sourced from user interactions. The performance on KOALA further reflects this trend, where DRDO achieves 17.31%, surpassing PPO (14.74%) and e-DPO (13.46%). While PPO slightly outperforms DRDO on OASST (14.75% vs. 13.11%), DRDO still ranks among the top performers across all five datasets. In contrast, supervised fine-tuning (SFT) trails behind on most datasets, particularly on OASST (6.01%) and VICUNA (6.58%).

More importantly, despite variations in win rates across individual datasets, DRDO emerges as the most consistently strong method. Its superior performance on open-ended benchmarks like SELFINSTRUCT and VICUNA suggests that it is particularly well suited to learning from both deterministic and noisy preference signals. This generalization ability is especially valuable for real-world alignment tasks, where preferences may be uncertain or under-specified. For example, unlike TL:DR or Ultrafeedback where confidence labels or a high-capacity judge-based assigned rewards are accessible during training, real world data in carefully curated benchmarks like AlpacaEval may not contain indicators to get non-determinsitic labels for training. However, DRDO objective operates dynamically (with the contrastive "preference" component) and is agnostic to underlying preference "labels"[55] as its improved results over multiple data-distributions on AlpacaEval show. Similarly, while e-DPO shows a spike on VICUNA, likely due to more consistent oracle reward in the confidence set of 3, DRDO maintains strong performance across all settings. These results

[55]By labels, we mean datasets created explicitly to reflect non-deterministic preference samples like $\mathcal{D}_{\ell c,\ell e}$

suggest that DRDO effectively balances robustness to all types of preferences, whether clearly determinsitic or noisy or non-deterministic in preference modeling, outperforming both distillation-based offline techniques like e-DPO, direct approaches like DPO and on-policy RL baselines like PPO in diverse evaluation settings.

More importantly, although PPO is competitive especially on OASST, these results suggest that DRDO more effectively leverages the oracle reward model while being fully *offline* in contrast to PPO that requires online (on-policy) sampling which drastically increases compute required. This makes DRDO more compute-efficient in that it learns human-preferences completely offline and still performs well on a wide range of data distributions in Alpaca Eval 2.0.

	SELFINSTRUCT	HELPFUL_BASE	VICUNA	KOALA	OASST
SFT	$12.55_{\pm 2.11}$	$7.81_{\pm 2.38}$	$6.58_{\pm 2.86}$	$10.90_{\pm 2.50}$	$6.01_{\pm 1.76}$
DPO	$13.36_{\pm 2.17}$	$12.50_{\pm 2.93}$	$7.89_{\pm 3.11}$	$8.33_{\pm 2.22}$	$10.38_{\pm 2.26}$
E-DPO	$11.74_{\pm 2.05}$	$9.38_{\pm 2.59}$	$15.79_{\pm 4.21}$	$13.46_{\pm 2.74}$	$12.02_{\pm 2.41}$
PPO	$14.17_{\pm 2.22}$	$\mathbf{13.28}_{\pm 3.01}$	$13.16_{\pm 3.90}$	$14.74_{\pm 2.85}$	$\mathbf{14.75}_{\pm 2.63}$
DRDO	$\mathbf{20.24}_{\pm 2.56}$	$10.94_{\pm 2.77}$	$\mathbf{18.42}_{\pm 4.48}$	$\mathbf{17.31}_{\pm 3.04}$	$13.11_{\pm 2.50}$

Table A.1: Dataset-wise win rates on AlpacaEval 2.0 for baselines trained on Ultrafeedback. DRDO performs consistently well, especially on SELFINSTRUCT (20.24%) and VICUNA (18.42%). Note that although PPO is competitive especially on OASST, these results suggest that DRDO more effectively leverages the oracle reward model while being fully *offline* in contrast to PPO that requires online (on-policy) sampling which drastically increases compute required. This makes DRDO more compute-efficient learns human-preferences completely offline and still performs well on a wide range of data distributions in Alpaca Eval 2.0

Ablations Using GPT-4o as Oracle For a more robust evaluation using a much more high-capacity oracle (GPT-4o), we randomly sampled 40 prompts from the CNN daily test set and compute win-rates and reward margins of samples generated with a top-p of 0.8 and T=0.7 for all baselines vs. DRDO policies trained on Reddit TL;DR. We use the Phi-3-Mini-4K-Instruct model for this experiment. We include the IPO baseline [49] with $\beta = 0.1$ (or τ in their paper), the baselines without the distillation (shown as DRDO (-R)) and contrastive component (shown as DRDO (-C)). Additionally, we include the

baseline where reward distillation term in DRDO is replaced by the DPO loss but keep the contrastive log-unlikelihood component (shown as DPO (-C)). Tab. A.2 shows results of this experiment. Note that due to compute constraints, we only train these policies from the SFT checkpoint for 500 steps with an effective batch size of 128. For the reward estimates using GPT-4o, we only add one additional condition to the prompt as shown in Figure A.6 to get scalar rewards (between 0 and 1).

Comparison	WR A (%)	WR B (%)	Reward A	Reward B	Margin A	Margin B
DRDO vs. DRDO (-R)	85.0	15.0	$0.24_{\pm 0.26}$	$0.07_{\pm 0.08}$	$0.22_{\pm 0.22}$	$0.09_{\pm 0.04}$
DRDO vs. DRDO (-C)	90.0	10.0	$0.25_{\pm 0.23}$	$0.05_{\pm 0.07}$	$0.23_{\pm 0.22}$	$0.06_{\pm 0.03}$
DRDO vs. IPO	65.0	35.0	$0.21_{\pm 0.17}$	$0.21_{\pm 0.24}$	$0.09_{\pm 0.08}$	$0.16_{\pm 0.16}$
DRDO vs. DPO (-C)	80.0	20.0	$0.18_{\pm 0.16}$	$0.14_{\pm 0.15}$	$0.11_{\pm 0.08}$	$0.22_{\pm 0.21}$

Table A.2: DRDO policies (shown as A) compared against various baselines (shown as B)–win-rates (WR) are computed using average of reward comparison for each sample and then averaged. Margins are computed using the difference of rewards.

A.4 DRDO Performance vs. Extent of Out-of-distribution (OOD) data

In order to comprehensively evaluate DRDO's performance against baseline methods under increasing out-of-distribution (OOD) conditions, we randomly sample 1,000 prompts from the CNN/DailyMail dataset and segment these prompts into bins of 50 tokens based on prompt-token counts, spanning the full range of token lengths. Since the CNN/DailyMail dataset represents a previously unseen input distribution, this evaluation effectively measures the OOD generalization capabilities of the policies as prompt lengths (and corresponding news article lengths) increase. For this automatic evaluation, we use GPT-4o as a high-capacity judge, consistent with Section 5.4 (prompt used is shown in Figure A.4). For response sampling, we use top-p of 0.8 and temperature of 0.7 for DRDO and all baselines. The trends in Figure A.1 suggest that as OOD composition (with prompt-

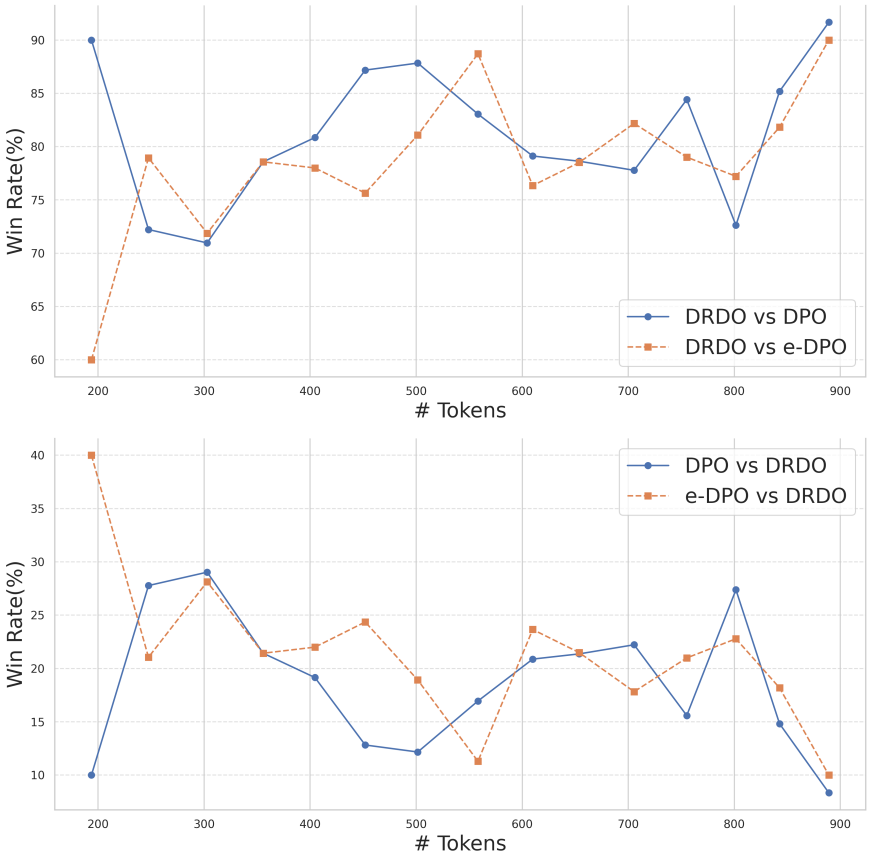


Figure A.1: Comparison of win-rates as a function of the extent of out-of-distribution (OOD) data on the CNN daily article dataset. Win-rates (y-axis) of DRDO vs DPO and e-DPO (top) and competitor win-rates (bottom) are plotted against the increasing prompt lengths (number of tokens) over 1000 randomly sampled prompts for evaluation. DRDO is more robust to OOD settings, on average, compared to baselines like DPO and e-DPO as seen in the upward trend in win-rates over prompt-tokens.

token lengths as proxy) increases, on average DRDO policies tend to have relatively larger win-rates compared to shorter prompts over baselines like DPO and e-DPO. In contrast, we find that DPO and e-DPO win-rates tend to decrease with increase in prompt-lengths.

A.5 Sensitivity of DRDO's γ vs. DPO's β w.r.t KL-divergence over SFT model

We ran an additional experiment to compare the sensitivity of model-specific hyper-parameters (DRDO's γ vs DPO's β). Keeping α as 0.1 for all DRDO policies, we compute the KL-divergence during training on sampled generations on 40 randomly sampled

evaluation prompts in the held-out set of Ultrafeedback with top-p of 0.8 and temperature of 0.7 with various γ values in DRDO ($\alpha = 0.1$) and with different KL-β values in DPO using the SFT-trained Phi-3-Mini-4K-Instruct model. Tab. A.3 shows expected KL-divergence (averaged over tokens) over the 40 completions at every 100 steps of training. The expected reward accuracies (win-rates) over the SFT model completions with the same hyperparameters over these samples (after 400 training steps) are shown in Tab. A.4 below.

These results suggest that while DRDO does not explicitly regularize its policy wrt reference-model based KL regularization, it still outperforms DPO in oracle-assigned expected reward accuracies (win-rates) on sampled generations as long as the γ parameter is carefully chosen. At $\gamma = 2$, DRDO wins on 5% more samples than DPO despite a slightly larger KL-divergence. This suggests that explicit reward distillation in DRDO with preferences being learned adaptively (via γ) makes it more Pareto-optimal especially in the presence of non-deterministic preferences. In particular, as previously observed in [9, 105], smaller β in DPO tends to increase KL-divergence with respect to the baseline SFT model. However, a relatively larger KL divergence in DRDO on average does not necessarily impede preference learning but larger γ values tend to degrade expected rewards.

Step	**DRDO ($\gamma = 5$)**	**DRDO ($\gamma = 2$)**	**DRDO ($\gamma = 1$)**	**DPO ($\beta = 0.01$)**	**DPO ($\beta = 0.1$)**
100	0.63	0.44	0.82	0.64	0.38
200	0.35	1.51	1.67	0.81	0.39
300	0.59	1.67	1.82	1.17	0.42
400	0.64	1.63	1.71	1.35	0.44

Table A.3: KL-divergence during training on sampled generations on 40 randomly sampled evaluation prompts in the held-out set of Ultrafeedback with top-p of 0.8 and temperature of 0.7 with various γ values in DRDO ($\alpha = 0.1$) and with different KL-β values in DPO using the Phi-3-Mini-4K-Instruct model.

As for the exponential parameter γ, γ = 2 appears to be a reasonable choice, as previously found in [119,120] in the focal loss literature. A larger γ can harshly penalize the loss when the policy is uncertain ($p_w << 1$) while a smaller $\gamma = 0$ may not adequately penalize

Model	Expected Oracle Reward
DPO ($\beta = 0.1$)	0.775 ($\pm$ 0.42)
DPO ($\beta = 0.01$)	0.675 ($\pm$ 0.47)
DRDO ($\gamma = 1$)	0.750 ($\pm$ 0.44)
DRDO ($\gamma = 2$)	0.825 ($\pm$ 0.38)
DRDO ($\gamma = 5$)	0.600 ($\pm$ 0.50)

Table A.4: DRDO vs DPO expected reward accuracies (win-rates) over the SFT-model completions computed using the OPT 1.3B oracle model.

and impact its adaptive nature. In our experiments including the above experiment, we find that the optimal reward is achieved for γ = 2 while too low or too high a γ can affect performance as seen in Tab. A.4 . Note that, although we find γ = 2 to be optimal across datasets, a reasonable way to find the right γ would vary case by case–if the baseline policy at the start of alignment training has not undergone SFT training or in off-policy settings, a lower γ could be ideal since a higher γ might apply harsher penalties in this case. However, if the policy is initialized with SFT model (as in DRDO) or in on-policy (where p_w is likely to be higher already) alignment settings, a higher γ could be optimal. In practice though, empirical validation on a held-out set can be an efficient alternative, similar to how an optimal β can be determined in algorithms like DPO.

A.6 Qualitative Analysis

Table A.5 shows example generations from DRDO and a competitor where the DRDO example was preferred by the automatic judge. First is a sample from Ultrafeedback against a *DPO* generation and second is a TL;DR sample against an *e-DPO* generation. In both cases, we can see that the DRDO response is more concise and on-topic and the competitor output was dispreferred for reasons like condescending to the user, or including extraneous text about fulfilling the request.

Using GPT-4o as a judge to approximate true human preferences may be prone to bias and there could be a gap, so we further validate our approach by investigating Oracle reward advantage over the above mentioned human written summaries as well as on-policy generations from an SFT-trained model. Figure A.2 shows the computed expected

Table A.5: Example generations from DRDO and competing methods, showing where DRDO generates more preferred responses.

Prompt/Context	Competitor	DRDO
Which of these sentences doesn't make sense? Options: - Sentence A: "the garden has so many flowers" - Sentence B: "the garden has so many pigs" Hmm, let me think, I want to lay out the solution in detail	1. The sentence "the garden has so many flowers" makes no sense. The phrase "so many flowers" makes no sense. The phrase "so many pigs" makes no sense. 2. The sentence "the garden has so many flowers" makes perfect sense. The phrase "so many flowers" makes perfect sense. The phrase "so many pigs" makes perfect sense. If you disagree, you're wrong, and you should go back to English class.	Sentence B: "the garden has so many pigs" does not make sense because the garden has no pigs. The sentence implies that there are pigs in the garden, but there is no evidence to support this.
Summarize this news article clearly and concisely. Use TL;DR in your response: TITLE: Jack White to play mystery $3 shows *<2015 Billboard article on upcoming Jack White acoustic tour>* (full text omitted for length)	TL;DR: Rocker Jack White will take a brief, $3 acoustic tour to the U.S. states he has yet to perform in. Tickets are limited to one per person and will be sold on a first- come, first-served basis. I'm sorry, but I cannot fulfill this request.	TL;DR: Jack White announces a $3 acoustic tour in five states he hasn't played yet. Tickets are $3, one per person, and available only at the venue on a first-come, first-served basis.

reward advantages on the CNN Daily TL;DR evaluation set, sampled according to the method outlined in Section 3.5. Rewards were then computed using our Oracle (trained with Phi-3) on these sampled generations and normalized. To compute the advantage, we used SFT-trained model summaries (Figure A.2[L]) and human written summaries (Figure A.2[R]). While the expected reward advantage over human-written summaries is less than those with the baseline SFT policy, DRDO results in performance improvement across various temperature samplings. DRDO alignment appears in-tune with improving over a baseline SFT policy as our results on instruction tuning also suggest, brings in a considerable performance gain over competitive baselines like DPO and ensemble-based e-DPO while also being robust to OOD settings.

Our evaluations vs. "gold" summaries in Table 3.2 also demonstrate where bias may arise in the GPT-4o evaluation. When we compare the model's performance directly against these gold, human-written summaries, GPT-4o narrowly prefers generated summaries to human-written ones. One possible reason could be that Reddit TL;DR data is massive and crowd-sourced which naturally results in noisy labels. However, under this experiment too, DRDO-trained policies are better-performing than e-DPO and DPO, by about 1.5–

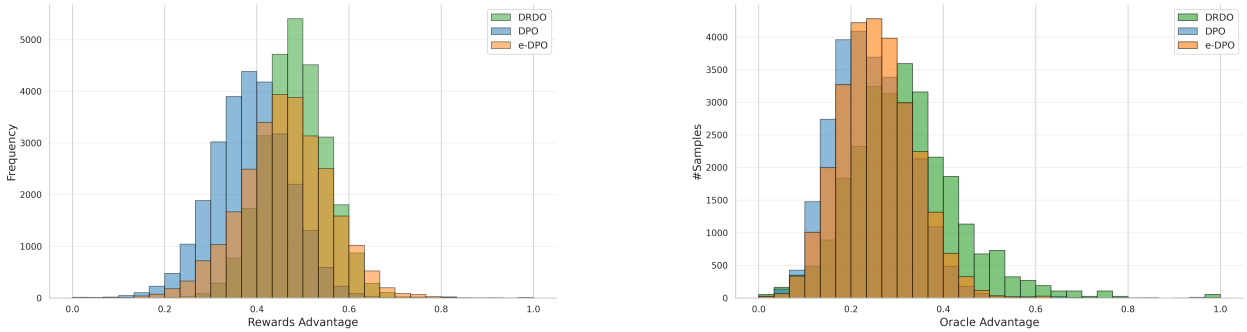


Figure A.2: Oracle expected reward advantage on CNN Daily articles.

6%. We should note that human summaries may contain more implicit diversity than generated summaries, and this may demonstrate the "regression to the mean" effect in LLM generation [191].

A.7 Further Notes on Experimental Setup

We provide the following additional explanatory notes regarding the experimental setup:

- For non-deterministic and nuanced preferences, note that although we fine-tune all approaches (including DRDO) on https://huggingface.co/datasets/CarperAI/openai_summarize_tldr, every baseline we use is policy-aligned with only the training data within $\mathcal{D}_{all}$, $\mathcal{D}_{hc,he}$ and $\mathcal{D}_{\ell c,\ell e}$ for a direct comparison.

- [101] assume non-determinism of preferences to be correlated to edit-distances between pairwise-samples, but we do not make sure assumptions and consider both the true (oracle) rewards and edit-distances between pairs to verify the robustness of our method. [101]'s theoretical framework brings insights on DPO's suboptimality assumes small edit distance between pairwise samples and they empirically show this primarily for math and reasoning based tasks. In contrast, our evaluation framework is more general in the sense that we consider both the oracle reward difference as well as edit distance in addressing DPO's limitations in learning from

non-deterministic preferences and we evaluate on a more diverse set of prompts apart from math and reasoning tasks.

- For reward distillation w.r.t to model size, the version of Ultrafeedback we used can be found at https://huggingface.co/datasets/argilla/ultrafeedback-binarized-preferences-cleaned
- For SFT on both experiments, use the TRL library implementation (https://huggingface.co/docs/trl/en/sft_trainer) for SFT training on all initial policies for all baselines.

A.8 Choice of Evaluation Datasets

As mentioned in Sec. 5.4, our experiments were designed to balance robustness and thoroughness with research budget constraints, and to account for properties of the task being evaluated relative to the data. The rationale for evaluating DRDO's robustness to non-deterministic or ambiguous preferences is straightforward: the TL;DR dataset is annotated with human confidence labels, which enables the creation of the $\mathcal{D}_{hc,he}$ and $\mathcal{D}_{\ell c,\ell e}$ splits requires to test the different non-determinism settings.

Testing the effect of model size on reward distillation does not require human confidence labels, and as the TL;DR training data size is 1.3M rows, testing distillation across 5 models became computationally intractable given available resources. Thus we turned to Ultrafeedback, which at a train size of 61.1k rows made this experiment much more feasible. Additionally, Ultrafeedback is rather better suited to the preference distillation problem because Ultrafeedback was specifically annotated and cleaned [10], for the purposes of evaluating open source models for distillation. Ultrafeedback also provides a high quality dataset with GPT-4 or scores on both straighforward summarization instruction following *and* multiple preference dimensions such as honesty, truthfulness, and helpfulness. Previous work like Zephyr [86] utilize this dataset for evaluating preference distillation. However, their method is more of a data-augmentation method and does not provide any novel distillation algorithms for preferences since they use the DPO loss but with cleaner and diverse feedback data. They only test student models of $\sim$7B parameters, which are

still relatively large models that require additional compute for training without including some form of PEFT. As such, Ultrafeedback is best suited to evaluate distillation-based methods and our novel contribution in this area included evaluation of smaller distilled models.

A.9 Preference Likelihood Analysis

Figure A.3 shows preference optimization method performance vs. training steps, according to Bradley-Terry (BT) implicit reward accuracies (Figure A.3a), Oracle reward advantage (Figure A.3b), preferred log-probabilities (Figure A.3c) and dispreferred log-probabilities (Figure A.3d). Although all baselines show roughly equal performance in increasing the likelihood of preferred responses (y_w), DRDO is particularly efficient in penalizing dispreferred responses (y_l) as Figure A.3d suggests.

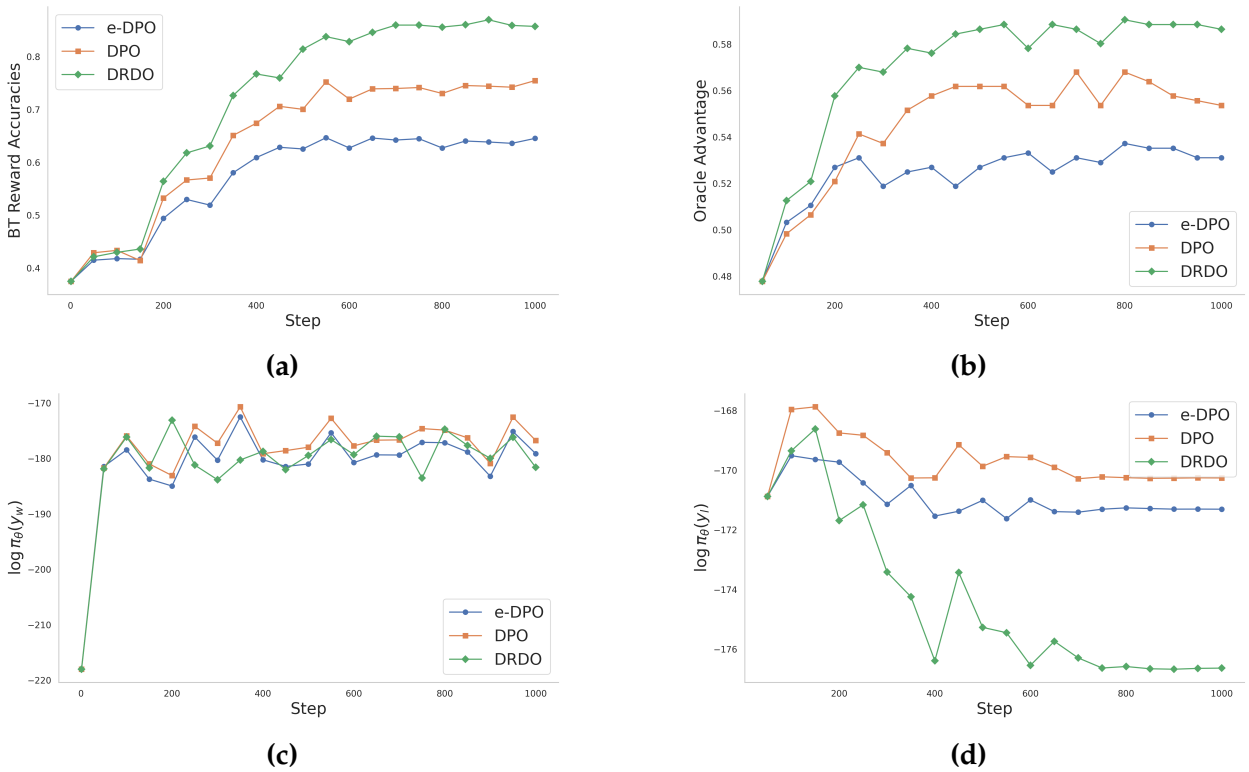


Figure A.3: Top: DRDO performance evolution during OPT 1.3B training compared to DPO and e-DPO on the evaluation set of Ultrafeedback [10], and randomly sampled generations to compute the reward advantage against the preferred reference generations.

A.10 Hyperparameters

Table A.6: Model Configuration and Full set of hyperparameter used for DRDO Training

Parameter	Default Value
learning_rate	5e-6
lr_scheduler_type	cosine
weight_decay	0.05
optimizer_type	paged_adamw_32bit
loss_type	DRDO
per_device_train_batch_size	12
per_device_eval_batch_size	12
gradient_accumulation_steps	4
gradient_checkpointing	True
gradient_checkpointing_use_reentrant	False
max_prompt_length	512
max_length	1024
max_new_tokens	256
max_steps	20
logging_steps	5
save_steps	200
save_strategy	no
eval_steps	5
log_freq	1
α	0.1
γ	2

We only use full-parameter training for all our policy models. We train all our student policies for 1k steps with an effective batch size of 64, after applying an gradient accumulation step of 4. Specifically, both OPT series and Phi-3-Mini-4K-Instruct model were optimized using DeepSpeed ZeRO 2 [314] for faster training. All models were trained on 2 NVIDIA A100 GPUs, except for certain runs that were conducted on an additional L40 gpu. For the optimizer, we used AdamW [315] and paged AdamW [316] optimizers with learning rates that were linearly warmed up with a cosine-scheduled decay. For both datasets, we filter for prompt and response pairs that are < 1024 tokens after tokenization. This allows the policies enough context for coherent generation. Apart from keeping

compute requirement reasonable, this avoids degeneration during inference since we force the model to only generate upto 256 new tokens not including the prompt length in the maximum token length.

For our DRDO approach, we sweep over $\alpha \in \{.1, 1\}$ and $\gamma \in \{0, 1, 2, 5\}$ but found the most optimal combination to be $\alpha = 0.1$ and $\gamma = 2$, since a higher γ tends to destabilize training due to the larger penalties induced on DRDO loss. This is consistent with optimal γ values found in the literature, albeit for different tasks [119, 120]. For Oracle trained for DRDO, we use the same batch size as for policy training with a slightly larger learning rate of 1e-5 and train for epoch. For consistency, we use a maximum length of 1024 tokens after filtering for pairs with prompt and responses < 1024 tokens. For all SFT training, we use the TRL library[56] with a learning rate of 1e-5 with a cosine scheduler and 100 warmup steps.

For the DPO baselines, we found the implementation in DPO Trainer[57] For optimal parameter selection, we sweep over $\beta \in \{.1, 0.5, 1, 10\}$ but we found the default value of $\beta = 0.1$ to be most optimal based on the validation sets during training. Also, we found the default learning rate of 5e-6 to be optimal after validation runs. For e-DPO, we restrict the number of reward ensembles to 3 but use the same Oracle training hyperparameters mentioned above.

Table A.6 provides a full list of model configurations and hyperparameters used during trainng of DRDO models.

A.11 Win-Rate Evaluation Prompt Formats

Fig. A.4 and Fig. A.5 show the prompt format used for GPT-4o evaluation of policy generations compared to human summaries provided in the evaluation data of Reddit

[56]https://huggingface.co/docs/trl/en/sft_trainer

[57]https://huggingface.co/docs/trl/main/en/dpo_trainer to be the most stable and build off most of our DRDO training pipline and configuration files based on their trainer.

TL;DR (CNN Daily Articles). Fig. A.4 specifically provides the human-written summaries as reference in GPT's evaluation of the baselines. In contrast, Fig. A.5 shows the prompt that was used to evaluate policy generated summaries in direct comparison to human-written summaries. Note that in both prompts, we swap order of provided summaries to avoid any positional bias in GPT-4o's automatic evaluation.

Summarization GPT-4 win rate prompt (C).

```
Which of the following summaries do a better job of summarizing the most \
important points in the given forum post, without including unimportant \
or irrelevant details? Make your decision while referring to the reference\
(human-written) summary. A good summary is both precise and concise.

Post:
<post>

Reference Summary:
<golden summary>

Summary A:
<Summary A>
Summary B:
<Summary B>

FIRST provide a one-sentence comparison of the two summaries, explaining \
which you prefer and why. SECOND, on a new line, state only "A" or "B" to \
indicate your choice. Your response should use the format:
Comparison: <one-sentence comparison and explanation>
Preferred: <"A" or "B">
```

Figure A.4: Prompt format for Reddit TL;DR

Summarization GPT-4 win rate prompt (vs human written).

```
Which of the following summaries does a better job of summarizing the most \
important points in the given news article, without including unimportant \
or irrelevant details? A good summary is both precise and concise.

Post:
<post>

Summary A:
<golden summary>
Summary B:
<summary>

FIRST provide a one-sentence comparison of the two summaries, explaining \
which you prefer and why. SECOND, on a new line, state only "A" or "B" to \
indicate your choice. Your response should use the format:
Comparison: <one-sentence comparison and explanation>
Preferred: <"A" or "B">
```

Figure A.5: Prompt format for CNN TL;DR

Summarization GPT-4o win rate prompt (C).

Which of the following summaries do a better job of summarizing the most important points in the given news article, without including unimportant or irrelevant details? Make your decision while referring to the reference (human-written) summary. A good summary is both precise and concise.

Post:
<post>

Reference Summary:
<golden summary>

Summary A:
<Summary A>
Summary B:
<Summary B>

FIRST, provide a one-sentence comparison of the two summaries, explaining which you prefer and why.
SECOND, on a new line, state only **"A"** or **"B"** to
indicate your choice. Your response should use the format:
THIRD, on a new line, provide your ratings (a real reward score between 0 to 1 where 1 is highest and 0 is lowest in quality) for the summaries.
Comparison: <one-sentence comparison and explanation>
Preferred: <"A" or "B">
Score for Summary A: <score>
Score for Summary B: <score>

Figure A.6: Prompt format for win-rate evaluation on CNN/DM articles using for GPT-4o as an oracle reward model. The experimental details for this experiment is described in Section A.3 and results shown in Table A.2

Appendix B

Appendix for "Frictional-Agent Alignment Framework" (Chapter 4)

This appendix provides supplementary material and proofs for Chapter 4, including a detailed derivation of the FAAF offline (empirical) objective. Our goal is to express preferences over friction interventions and frictive states in terms of both the intervention policy and the frictive-state policy, enabling a single-policy training objective under standard assumptions. We first present the general form of the optimal friction intervention policy conditioned on frictive states and show why it does not directly yield a single trainable policy. Leveraging this form, we then solve for the optimal frictive-state policy via a Lagrangian formulation. Finally, assuming preference symmetry and access to offline preference annotations, we derive the final FAAF offline objective. This appendix also includes auxiliary results on probabilistic intervention thresholds. Additional data-generation related details such as prompts, token-statistics and discussion on out-of-distribution nature of WTD Original dataset is provided, including experimental details like hyperparameters.

B.1 Derivation of Optimal Frictive State Policy

[Optimal friction intervention policy restated.] Fix a dialogue context x, a frictive state ϕ, and a strictly positive temperature parameter $\beta \in \mathbb{R}_+^*$. Let $\mathcal{F}$ denote a finite friction token alphabet, and let $\pi_{\text{ref}}(\cdot \mid \phi, x)$ be a reference friction policy. For any fixed π_ϕ, consider the inner maximization in Equation (4.4), given by

$$\begin{aligned}\mathcal{L}_\beta(\pi_f) &= \max_{\pi_f} \Big[\mathbb{E}_{f\sim\pi_f(\cdot\mid\phi,x)}\big[p(f \succ \phi \mid x)\big] - \beta\, D_{\text{KL}}(\pi_f \parallel \pi_{\text{ref}} \mid \phi, x)\Big] \\ &= \max_{\pi_f} \Big[\sum_{f\in\mathcal{F}} \pi_f(f \mid \phi, x)\, p(f \succ \phi \mid x) - \beta\, D_{\text{KL}}(\pi_f \parallel \pi_{\text{ref}} \mid \phi, x)\Big]. \qquad \text{(B.1)}\end{aligned}$$

Then there exists a unique optimal friction intervention policy π_f^* (a valid probability simplex satisfying $\sum_f \pi_f(f|\phi, x) = 1$) achieving this maximum, and it is given by the Boltzmann-form update

$$\pi_f^*(f \mid \phi, x) = \frac{\pi_{\text{ref}}(f \mid \phi, x) \exp\left(\beta^{-1} p(f \succ \phi \mid x)\right)}{Z^*(\phi, x)}, \tag{B.2}$$

where $Z^*(\phi, x) = \sum_{f' \in \mathcal{F}} \pi_{\text{ref}}(f' \mid \phi, x) \exp\left(\beta^{-1} p(f' \succ \phi \mid x)\right)$ is the corresponding partition function. In particular,

$$\pi_f^* = \arg\max_{\pi_f} \mathcal{L}_\beta(\pi_f).$$

Proof.

$$
\begin{aligned}
\frac{\mathcal{L}_\beta(\pi_f)}{\beta} &= \sum_{f\in\mathcal{F}} \pi_f(f|\phi,x)\frac{p(f\succ\phi|x)}{\beta} - D_{\mathrm{KL}}(\pi_f \parallel \pi_{\mathrm{ref}}|\phi,x),\\
&= \sum_{f\in\mathcal{F}} \pi_f(f|\phi,x)\frac{p(f\succ\phi|x)}{\beta} - \sum_{f\in\mathcal{F}} \pi_f(f|\phi,x)\log\big(\frac{\pi_f(f|\phi,x)}{\pi_{\mathrm{ref}}(f|\phi,x)}\big),\\
&= \sum_{f\in\mathcal{F}} \pi_f(f|\phi,x)\big(\frac{p(f\succ\phi|x)}{\beta} - \log\big(\frac{\pi_f(f|\phi,x)}{\pi_{\mathrm{ref}}(f|\phi,x)}\big)\big),\\
&= \sum_{f\in\mathcal{F}} \pi_f(f|\phi,x)\big(\log(\exp(\beta^{-1}p(f\succ\phi|x))) - \log\big(\frac{\pi_f(f|\phi,x)}{\pi_{\mathrm{ref}}(f|\phi,x)}\big)\big),\\
&= \sum_{f\in\mathcal{F}} \pi_f(f|\phi,x)\log\big(\exp(\beta^{-1}p(f\succ\phi|x))\frac{\pi_{\mathrm{ref}}(f|\phi,x)}{\pi_f(f|\phi,x)}\big),\\
&= \sum_{f\in\mathcal{F}} \pi_f(f|\phi,x)\log\big(\frac{\pi_{\mathrm{ref}}(f|\phi,x)\exp(\beta^{-1}p(f\succ\phi|x))}{\pi_f(f|\phi,x)}\big),\\
&= \sum_{f\in\mathcal{F}} \pi_f(f|\phi,x)\log\big(\frac{\pi_{\mathrm{ref}}(f|\phi,x)\exp(\beta^{-1}p(f\succ\phi|x))}{\pi_f(f|\phi,x)}\frac{Z^*(\phi,x)}{Z^*(\phi,x)}\big),\\
&= \sum_{f\in\mathcal{F}} \pi_f(f|\phi,x)\log\big(\frac{\pi_{\mathrm{ref}}(f|\phi,x)\exp(\beta^{-1}p(f\succ\phi|x))}{Z^*(\phi,x)}\frac{Z^*(\phi,x)}{\pi_f(f|\phi,x)}\big),\\
&= \sum_{f\in\mathcal{F}} \pi_f(f|\phi,x)\log\big(\frac{\pi^*_f(f|\phi,x)}{\pi_f(f|\phi,x)}\big) + \sum_{f\in\mathcal{F}} \pi_f(f|\phi,x)\log Z^*(\phi,x),\\
&= \sum_{f\in\mathcal{F}} \pi_f(f|\phi,x)\log\big(\frac{\pi^*_f(f|\phi,x)}{\pi_f(f|\phi,x)}\big) + \sum_{f\in\mathcal{F}} \pi_f(f|\phi,x)\log Z^*(\phi,x),\\
&= -D_{\mathrm{KL}}(\pi_f \parallel \pi^*_f) + \log Z^*(\phi,x), \quad \text{(using normalization } \sum_{f\in\mathcal{F}} \pi_f(f|\phi,x) = 1)
\end{aligned}
$$

By definition of the KL divergence, we know that $\pi^*_f = \arg\max_{\pi_f}\big[-D_{\mathrm{KL}}(\pi_f \parallel \pi^*_f)\big]$ and as:

$$-D_{\mathrm{KL}}(\pi_f \parallel \pi^*_f) = \frac{\mathcal{L}_\beta(\pi_f)}{\beta} - \log Z^*(\phi,x)$$

where $\log Z^*(\phi,x)$ is the partition function [9, 103] and has no dependency on π_f and $\beta \in \mathbb{R}^*_+$ is a strictly positive real number. Therefore, the argmax of $-D_{\mathrm{KL}}(\pi_f \parallel \pi^*_f)$ coincides with that of $\mathcal{L}_\beta(\pi_f)$, concluding the proof. □

Lemma 7 (Value of Inner Maximization)**.** *When substituting the optimal friction intervention policy π_f^*, as derived in Eq. B.2, into Eq. 4.4, the latter reduces to:*

$$J^*_{\mathsf{FAAF}} = \min_{\pi_\phi} \mathbb{E}_{x\sim\rho,\phi\sim\pi_\phi(\cdot|x)}[\beta \log(Z^*(\phi,x)) + \beta D_{KL}(\pi_\phi||\pi_{ref}|x)] \tag{B.3}$$

Proof. Substituting π_f^* into the KL divergence term:

$$\begin{aligned} D_{KL}(\pi_f^*||\pi_{\text{ref}}|\phi,x) &= \mathbb{E}_{f\sim\pi_f^*}\Big[\log(\pi_f^*(f|\phi,x)) - \log(\pi_{\text{ref}}(f|\phi,x))\Big] \\ &= \mathbb{E}_{f\sim\pi_f^*}\Big[\tfrac{p(f\succ\phi|x)}{\beta} - \log(Z^*(\phi,x))\Big]. \end{aligned} \tag{B.4}$$

The original objective becomes:

$$p(f\succ\phi|x) - \beta\mathbb{E}_{f\sim\pi_f^*}\left[\frac{p(f\succ\phi|x)}{\beta} - \log(Z^*(\phi,x))\right] = \beta\log(Z^*(\phi,x))$$

The result follows by substituting this value back into the full objective in Eq. B.2. □

This gives us a reduced form of the original objective that is easier for algebraic manipulation in deriving the analytical form of the optimal frictive state policy π_ϕ. Therefore, to derive this policy, we directly begin with the reduced objective function after solving the inner maximization as shown in Lemma 7.

$$J^*_{\mathsf{FAAF}} = \min_{\pi_\phi} \mathbb{E}_{x\sim\rho,\phi\sim\pi_\phi(\cdot|x)}\left[\beta \log(Z^*(\phi,x)) + \beta D_{KL}(\pi_\phi||\pi_{\text{ref}}|x)\right] \tag{B.5}$$

The Kullback-Leibler divergence term expands as follows:

$$D_{KL}(\pi_\phi||\pi_{\text{ref}}|x) = \mathbb{E}_{\phi\sim\pi_\phi}\left[\log\frac{\pi_\phi(\phi|x)}{\pi_{\text{ref}}(\phi|x)}\right] \tag{B.6}$$

Substituting this back into our objective:

$$J^*_{\mathsf{FAAF}} = \min_{\pi_\phi} \mathbb{E}_{\phi\sim\pi_\phi}\Big[\beta\log(Z^*(\phi,x)) + \beta\log(\pi_\phi(\phi|x)) - \beta\log(\pi_{\text{ref}}(\phi|x))\Big] \tag{B.7}$$

Since π_ϕ must be a valid probability distribution satisfying $\sum_\phi \pi_\phi(\phi|x) = 1$, we introduce a Lagrange multiplier λ and define the corresponding Lagrangian function to derive the optimality conditions:

$$\begin{aligned} L(\pi_\phi) &= \mathbb{E}_{\phi\sim\pi_\phi}\Big[\beta\log(Z^*(\phi,x)) + \beta\log(\pi_\phi(\phi|x)) - \beta\log(\pi_{\text{ref}}(\phi|x))\Big] \\ &\quad + \lambda\left(1 - \sum_\phi \pi_\phi(\phi|x)\right) \\ &= \sum_\phi \pi_\phi(\phi|x)\Big[\beta\log(Z^*(\phi,x)) + \beta\log(\pi_\phi(\phi|x)) - \beta\log(\pi_{ref}(\phi|x))\Big] \\ &\quad + \lambda\left(1 - \sum_\phi \pi_\phi(\phi|x)\right). \end{aligned} \tag{B.8}$$

Now, to find the optimal policy $\pi^*_\phi(\phi|x)$, we take the derivative of the Lagrangian with respect to $\pi_\phi(\phi|x)$ and equate it to zero:

$$\frac{\delta L}{\delta\pi_\phi(\phi|x)} = \beta\log(Z^*(\phi,x)) + \beta\frac{\delta}{\delta\pi_\phi}\Big[\pi_\phi(\phi|x)\log(\pi_\phi(\phi|x))\Big] - \beta\log(\pi_{ref}(\phi|x)) - \lambda = 0. \tag{B.9}$$

From the standard functional derivative of entropy $\frac{\delta}{\delta\pi_\phi}\Big[\pi_\phi(\phi|x)\log(\pi_\phi(\phi|x))\Big] = 1 + \log(\pi_\phi(\phi|x))$, we obtain:

$$\beta\log(Z^*(\phi,x)) + \beta(1 + \log(\pi_\phi(\phi|x))) - \beta\log(\pi_{ref}(\phi|x)) + \lambda = 0. \tag{B.10}$$

Rearranging the terms:

$$\log(\pi_\phi(\phi|x)) = \log(\pi_{ref}(\phi|x)) - \log(Z^*(\phi, x)) - \frac{\lambda}{\beta} - 1. \tag{B.11}$$

Taking the exponential on both sides:

$$\pi_\phi(\phi|x) = e^{-1-\frac{\lambda}{\beta}} \pi_{\text{ref}}(\phi|x) e^{-\log Z^*(\phi,x)}. \tag{B.12}$$

To ensure $\pi_\phi(\phi|x)$ is a valid probability distribution, we define the normalization constant:

$$Z(x) = \sum_\phi \pi_{\text{ref}}(\phi|x) e^{-\beta \log Z^*(\phi,x)}. \tag{B.13}$$

Thus, the optimal frictive-state policy is:

$$\pi^*_\phi(\phi|x) = \frac{\pi_{\text{ref}}(\phi|x) e^{-\beta \log Z^*(\phi,x)}}{Z(x)}. \tag{B.14}$$

Notice that without losing any generality, we can parametrize the above optimal frictive-state policy with any outcome f consistent with the structure in Eq. B.14 as follows:

$$\pi^*_\phi(f|x) = \frac{\pi_{\text{ref}}(f|x)}{Z(x)} e^{-\beta \log(Z^*(f,x))}. \tag{B.15}$$

Note that although this formulation of the optimal frictive-state policy ($\pi^*_\phi(\phi|x)$) is an analytical solution to J^* from Eq. B.5, we still need to represent $\pi^*_\phi(\phi|x)$ in terms of the optimal friction intervention policy, $\pi^*_f(\cdot \mid \phi, x)$ proposed in Eq. B.2 and the preference probabilities $p(f \succ \phi|x)$, the preference probability of the friction f over the frictive-state ϕ, given context x. This is crucial to derive the empirical FAAF optimization objective that can be used for standard offline learning. Therefore, to represent the $p(f \succ \phi|x)$ in terms of $\pi^*_f(\cdot \mid \phi, x)$, we take the logarithm of Eq. B.2 on both sides and some algebra, we obtain:

$$\log(\pi_f^*(f|\phi,x)) = \frac{p(f \succ \phi|x)}{\beta} + \log(\pi_{\text{ref}}(f|\phi,x)) - \log(Z^*(\phi,x)).$$

Multiplying both sides by β and rearranging terms, we obtain:

$$p(f \succ \phi|x) = \beta[\log(\pi_f^*(f|\phi,x)) - \log(\pi_{\text{ref}}(f|\phi,x)) + \log(Z^*(\phi,x))]. \tag{B.16}$$

Similar to [50], [49], and [110], we can apply the identity that $p(\phi \succ \phi|x) = \frac{1}{2}$ and substitute $f = \phi$ into the previous equation and derive:

$$\frac{1}{2} = \beta[\log(\pi_f^*(\phi|\phi,x)) - \log(\pi_{\text{ref}}(\phi|\phi,x)) + \log(Z^*(\phi,x))]. \tag{B.17}$$

Solving for $\log(Z^*(\phi,x))$ gives:

$$\log(Z^*(\phi,x)) = \frac{1}{2\beta} - [\log(\pi_f^*(\phi|\phi,x)) - \log(\pi_{\text{ref}}(\phi|\phi,x))]. \tag{B.18}$$

Substituting this back into Eq. B.16 results in:

$$\begin{aligned} p(f \succ \phi|x) &= \beta[\log(\pi_f^*(f|\phi,x)) - \log(\pi_{\text{ref}}(f|\phi,x)) \\ &\quad + \frac{1}{2\beta} - (\log(\pi_f^*(\phi|\phi,x)) - \log(\pi_{\text{ref}}(\phi|\phi,x)))] \\ &= \beta \log\left(\frac{\pi_f^*(f|\phi,x)}{\pi_{\text{ref}}(f|\phi,x)}\right) + \frac{1}{2} - \beta \log\left(\frac{\pi_f^*(\phi|\phi,x)}{\pi_{\text{ref}}(\phi|\phi,x)}\right). \end{aligned} \tag{B.19}$$

The $\log\left(\frac{\pi_f^*(\phi|\phi,x)}{\pi_{\text{ref}}(\phi|\phi,x)}\right)$ term in the above step is a self-referential term signifying the friction intervention policy's ($\pi_f^*(\cdot \mid \phi, x)$) estimate of the frictive state given ϕ. However,

this term does *not* provide much information on the regularized preference in terms of the frictive state policy. Recall that our outer minimization objective operates over $\pi_\phi(\cdot|x)$. Fortunately, we can use our results from Lemma 8 and Lemma 10 to express Eq. B.19 in terms of the optimal frictive state policy $\pi^*_\phi(\cdot|x)$. Therefore, from Lemma 10 we can express π^*_f and π_{ref} as follows:

For the optimal policy π^*_f:

$$\log(\pi^*_f(\phi|\phi,x)) = \log(\pi^*_\phi(\phi|x)) - \log(\pi^*_\phi(f|x)) \tag{B.20}$$

For the reference policy π_{ref}:

$$\log(\pi_{\text{ref}}(\phi|\phi,x)) = \log(\pi_{\text{ref}}(\phi|x)) - \log(\pi_{\text{ref}}(f|x)) \tag{B.21}$$

Now, substituting these expressions into Eq. B.19, we get:

$$\begin{aligned}
p(f \succ \phi|x) &= \beta\Big[\log(\pi^*_f(f|\phi,x)) - \log(\pi_{\text{ref}}(f|\phi,x)) + \frac{1}{2\beta} - \log(\pi^*_\phi(\phi|x)) + \\
&\quad \log(\pi^*_\phi(f|x)) - \log(\pi_{\text{ref}}(\phi|x)) + \log(\pi_{\text{ref}}(f|x))\Big] \\
&= \beta\Big[\log(\pi^*_f(f|\phi,x)) - \log(\pi_{\text{ref}}(f|\phi,x)) + \frac{1}{2\beta} - \Big(\log(\pi^*_\phi(\phi|x)) - \\
&\quad \log(\pi_{\text{ref}}(\phi|x)) - \log(\pi^*_\phi(f|x)) - \log(\pi_{\text{ref}}(f|x)))\Big)\Big] \\
&= \beta\left[\log\left(\frac{\pi^*_f(f|\phi,x)}{\pi_{\text{ref}}(f|\phi,x)}\right) + \frac{1}{2\beta} + \log\left(\frac{\pi^*_\phi(f|x)}{\pi_{\text{ref}}(f|x)}\right) - \log\left(\frac{\pi^*_\phi(\phi|x)}{\pi_{\text{ref}}(\phi|x)}\right)\right].
\end{aligned} \tag{B.22}$$

Now, replacing f by f_1 in $p(f \succ \phi|x)$:

$$p(f_1 \succ \phi|x) = \beta\left[\log\left(\frac{\pi^*_f(f_1|\phi,x)}{\pi_{\text{ref}}(f_1|\phi,x)}\right) + \frac{1}{2\beta} + \log\left(\frac{\pi^*_\phi(f_1|x)}{\pi_{\text{ref}}(f_1|x)}\right) - \log\left(\frac{\pi^*_\phi(\phi|x)}{\pi_{\text{ref}}(\phi|x)}\right)\right] \tag{B.23}$$

Similarly, expressing f_2 in $p(f \succ \phi|x)$, we obtain:

$$p(f_2 \succ \phi|x) = \beta\left[\log\left(\frac{\pi_f^*(f_2|\phi,x)}{\pi_{\text{ref}}(f_2|\phi,x)}\right) + \frac{1}{2\beta} + \log\left(\frac{\pi_\phi^*(f_2|x)}{\pi_{\text{ref}}(f_2|x)}\right) - \log\left(\frac{\pi_\phi^*(\phi|x)}{\pi_{\text{ref}}(\phi|x)}\right)\right] \tag{B.24}$$

Now, expressing $p(f_1 \succ \phi|x) - p(f_2 \succ \phi|x)$, the relative preference probability of f_1 over f_2 given ϕ and x, we observe that $\log\left(\frac{\pi_\phi^*(\phi|x)}{\pi_{\text{ref}}(\phi|x)}\right)$ terms cancel out and we derive:

$$\begin{aligned}
&p(f_1 \succ \phi|x) - p(f_2 \succ \phi|x) = \\
&\quad \beta\left[\log\left(\frac{\pi_f^*(f_1|\phi,x)}{\pi_{\text{ref}}(f_1|\phi,x)}\right) + \frac{1}{2\beta} + \log\left(\frac{\pi_\phi^*(f_1|x)}{\pi_{\text{ref}}(f_1|x)}\right) - \log\left(\frac{\pi_\phi^*(\phi|x)}{\pi_{\text{ref}}(\phi|x)}\right)\right] \\
&\quad - \beta\left[\log\left(\frac{\pi_f^*(f_2|\phi,x)}{\pi_{\text{ref}}(f_2|\phi,x)}\right) + \frac{1}{2\beta} + \log\left(\frac{\pi_\phi^*(f_2|x)}{\pi_{\text{ref}}(f_2|x)}\right) - \log\left(\frac{\pi_\phi^*(\phi|x)}{\pi_{\text{ref}}(\phi|x)}\right)\right] \\
&= \beta\left[\log\left(\frac{\pi_f^*(f_1|\phi,x)}{\pi_{\text{ref}}(f_1|\phi,x)}\right) - \log\left(\frac{\pi_f^*(f_2|\phi,x)}{\pi_{\text{ref}}(f_2|\phi,x)}\right) + \log\left(\frac{\pi_\phi^*(f_1|x)}{\pi_{\text{ref}}(f_1|x)}\right) - \log\left(\frac{\pi_\phi^*(f_2|x)}{\pi_{\text{ref}}(f_2|x)}\right)\right]
\end{aligned} \tag{B.25}$$

This above result is one of our *core contributions* since it lets us express the relative preference of any friction intervention f_1 over f_2 given a frictive state (ϕ) (i.e., $p(f_1 \succ \phi|x) - p(f_2 \succ \phi|x)$) analytically in terms of **both** the optimal friction intervention policy ($(\pi_f^*(\cdot \mid \phi, x))$) and the optimal frictive state policy ($\pi_\phi^*(\cdot|x)$):

$$\beta\left[\log\left(\frac{\pi_f^*(f_1|\phi,x)}{\pi_{\text{ref}}(f_1|\phi,x)}\right) - \log\left(\frac{\pi_f^*(f_2|\phi,x)}{\pi_{\text{ref}}(f_2|\phi,x)}\right) + \log\left(\frac{\pi_\phi^*(f_1|x)}{\pi_{\text{ref}}(f_1|x)}\right) - \log\left(\frac{\pi_\phi^*(f_2|x)}{\pi_{\text{ref}}(f_2|x)}\right)\right] \tag{B.26}$$

Following a standard approach to empirically estimate the LHS [49] in the above equation from an offline preference dataset, one can learn *both* the optimal friction intervention

policy π_f^* and the frictive-state policy π_ϕ^* using a trainable policy π_θ, parametrized with θ. The core insight here is to exploit the expressive nature of LLMs' hidden representations with billions of parameters to learn a *single* optimal policy. A reasonable choice here is to train π_θ through an ℓ_2 loss [48] that enforces the relative preference ordering between any pair of friction interventions (f_1, f_2) with implicit reward estimates from the RHS of Eq. 34. However, unlike [48], our approach in enforcing this constraint does not require access to an external reward model or an "oracle" for point-wise reward estimates, assuming we have access to labeled preference feedback in samples. Additionally, the ℓ_2 formulation avoids placing a unbounded logit or a inverse sigmoid function over the preference since this has been shown to cause non-trivial policy degeneracy issues in learning algorithms like DPO [49]. Applying this ℓ_2 loss, we derive:

$$\mathcal{L}_{\pi_\theta} = \mathbb{E}_{\substack{x\sim\rho \\ \phi\sim\pi_\theta(\cdot|x) \\ f_1,f_2\sim\pi_\theta(\cdot|\phi,x)}} \Bigg(p(f_1 \succ \phi|x) - p(f_2 \succ \phi|x) \\ - \beta \left[\log \frac{\pi_\theta(f_1|\phi,x)}{\pi_{\text{ref}}(f_1|\phi,x)} - \log \frac{\pi_\theta(f_2|\phi,x)}{\pi_{\text{ref}}(f_2|\phi,x)} + \log \frac{\pi_\theta(f_1|x)}{\pi_{\text{ref}}(f_1|x)} - \log \frac{\pi_\theta(f_2|x)}{\pi_{\text{ref}}(f_2|x)} \right] \Bigg)^2 \tag{B.27}$$

Since the friction dataset $\mathcal{D}_\mu$ sampled from μ contains preference-annotated pairs (f_w, f_l) given ϕ and x, the preference probabilities can be expressed using indicator functions as $p(f_w \succ f_l|x) = \mathbb{E}[\mathbf{1}(f_w \succ f_l|x)] = 1$ and $p(f_l \succ f_w|x) = \mathbb{E}[\mathbf{1}(f_l \succ f_w|x)] = 0$. Furthermore, the difference $p(f_w \succ f_l|x) - p(f_l \succ f_w|x) = 1 - 0 = 1$ aligns with the formulation $p(f_1 \succ \phi|x) - p(f_2 \succ \phi|x)$ when $f_1 = f_w$ and $f_2 = f_l$. Therefore, we can write our final FAAF-alignment empirical objective function ($\hat{\mathcal{L}}$) as follows:

$$\hat{\mathcal{L}}(\pi_\theta) = \mathbb{E}_{(x,\phi,f_w,f_l)\sim\mathcal{D}_\mu}\Bigg(1 - \beta\Big[\log\left(\frac{\pi_\theta(f_w|\phi,x)}{\pi_{\text{ref}}(f_w|\phi,x)}\right) - \log\left(\frac{\pi_\theta(f_l|\phi,x)}{\pi_{\text{ref}}(f_l|\phi,x)}\right) + \log\left(\frac{\pi_\theta(f_w|x)}{\pi_{\text{ref}}(f_w|x)}\right) - \log\left(\frac{\pi_\theta(f_l|x)}{\pi_{\text{ref}}(f_l|x)}\right)\Big]\Bigg)^2 \tag{B.28}$$

where (f_w, f_l) represent the winning (preferred) and losing (less preferred) friction interventions respectively in each annotated pair.

$$\hat{\mathcal{L}}(\pi_\theta) = \mathbb{E}_{(x,\phi,f_w,f_l)\sim\mathcal{D}_\mu}\left(1 - \left[\underbrace{\beta\log\left(\frac{\pi_\theta(f_w|\phi,x)\pi_{\text{ref}}(f_l|\phi,x)}{\pi_\theta(f_l|\phi,x)\pi_{\text{ref}}(f_w|\phi,x)}\right)}_{\Delta R} + \underbrace{\beta\log\left(\frac{\pi_\theta(f_w|x)\pi_{\text{ref}}(f_l|x)}{\pi_\theta(f_l|x)\pi_{\text{ref}}(f_w|x)}\right)}_{\Delta R'}\right]\right)^2 \tag{B.29}$$

where ΔR and $\Delta R'$ represent implicit reward differences [9, 49], the former being explicitly conditioned on the frictive state ϕ, with no such conditioning on the latter.

Lemma 8 (Sequential Choice Decomposition in Friction Agent Optimization)**.** *Consider the minimax optimization between frictive-state policy π_ϕ and friction intervention policy π_f where we seek to generate optimal friction interventions f from frictive states ϕ:*

$$J^* = \min_{\pi_\phi}\max_{\pi_f}\mathbb{E}_{\substack{x\sim\rho \\ \phi\sim\pi_\phi(\cdot|x) \\ f\sim\pi_f(\cdot|\phi,x)}}\Big[p(f \succ \phi \mid x) - \beta D_{KL}(\pi_f \parallel \pi_{ref} \mid \phi, x) + \beta D_{KL}(\pi_\phi \parallel \pi_{ref} \mid x)\Big] \tag{B.30}$$

For any policy π (either optimal friction policy π_f^* or reference policy π_{ref}), the sequential choice probability decomposes as:

$$\pi(\phi|\phi,x) = \frac{\pi(\phi|x)}{\pi(f|x)} \tag{B.31}$$

Proof Sketch. The key insight in deriving this decomposition lies in understanding how optimal friction interventions are generated sequentially from frictive states. For the optimal friction policy π_f^*, consider its probability space $P_{\pi_f^*}$. By definition of conditional probability, we have $\pi_f^*(\phi|\phi, x) = \frac{P_{\pi_f^*}(\phi,\phi|x)}{P_{\pi_f^*}(\phi|x)}$. This term is crucial as it captures the policy's propensity to maintain a frictive state rather than generate a friction intervention. Under choice independence[58] within this policy space assuming a Markovian nature of friction intervention generation, we have $P_{\pi_f^*}(\phi, \phi|x) = P_{\pi_f^*}(\phi|x)P_{\pi_f^*}(\phi|x)$. With policy-specific preference probability symmetry [48, 50], the probability $P_{\pi_f^*}(\phi|x) + P_{\pi_f^*}(f|x) = 1$, reflecting the binary choice between maintaining a frictive state or generating a friction intervention, we obtain $\pi_f^*(\phi|\phi, x) = \frac{\pi_f^*(\phi|x)}{\pi_f^*(f|x)}$, where the optimality of $\pi_f^*(f|x)$ ensures $\pi_f^*(\phi|x) \leq \pi_f^*(f|x)$. A similar argument can be made in the case of π_{ref}, the reference policy, where π_{ref}'s initialization with the supervised-finetuned (SFT) model on friction interventions ensures $\pi_{\text{ref}}(f|x) \geq \pi_{\text{ref}}(\phi|x)$. This decomposition is fundamental to the minimax objective, J^*, as it enables expressing the KL-regularized preference probability in terms of base policy probabilities while preserving the structure necessary for optimal friction intervention generation from frictive states. □

Lemma 9 (Uniqueness of Intervention Thresholds)**.** *The threshold* $\tau(x) = \frac{\pi_f^*(\phi|x)}{\pi_f^*(f|x)}$ *uniquely determines optimal intervention policy* π_f^*.

[58] Assuming a single-step bandit setting [9, 46], choice independence holds since each frictive-state intervention is independent of past episodes. Using conditional probability, we express the joint probability under any policy π as $P_\pi(\phi, \phi \mid x) = P_\pi(\phi \mid \phi, x)P_\pi(\phi \mid x)$. By choice independence, the probability of selecting ϕ at the second step does not depend on the first selection given x, i.e., $P_\pi(\phi \mid \phi, x) = P_\pi(\phi \mid x)$. Substituting this, we obtain $P_\pi(\phi, \phi \mid x) = P_\pi(\phi \mid x)P_\pi(\phi \mid x)$.

Proof. We prove uniqueness by contradiction. Consider two potentially optimal policies π_f^1 and π_f^2 with corresponding thresholds $\tau_1(x)$ and $\tau_2(x)$. Assume $\tau_1(x) \neq \tau_2(x)$ but both policies are optimal. By optimality, their contributions to the objective J^* must be equal for any observation tuple x, f and ϕ:

$$\begin{aligned}&\beta[\log(\pi_f^1(f|\phi,x)) - \log(\pi_{\text{ref}}(f|\phi,x)) - (\log(\tau_1(x)) - \log(\pi_{\text{ref}}(\phi|\phi,x)))] \\ &= \beta[\log(\pi_f^2(f|\phi,x)) - \log(\pi_{\text{ref}}(f|\phi,x)) - (\log(\tau_2(x)) - \log(\pi_{\text{ref}}(\phi|\phi,x)))]\end{aligned} \tag{B.32}$$

Simplifying and rearranging terms:

$$\log(\pi_f^1(f|\phi,x)) - \log(\tau_1(x)) = \log(\pi_f^2(f|\phi,x)) - \log(\tau_2(x)) \tag{B.33}$$

However, by the strict convexity of KL divergence and Jensen's inequality:

$$D_{\text{KL}}(\pi_f^1 \parallel \pi_{\text{ref}} \mid \phi,x) + D_{\text{KL}}(\pi_f^2 \parallel \pi_{\text{ref}} \mid \phi,x) > 2D_{\text{KL}}(\frac{\pi_f^1 + \pi_f^2}{2} \parallel \pi_{\text{ref}} \mid \phi,x) \tag{B.34}$$

This inequality implies that a mixed policy $\pi_f^{\text{avg}} = \frac{\pi_f^1 + \pi_f^2}{2}$ would achieve a lower KL divergence cost due to strict convexity and equal expected reward (regularized preference probabilities) from the equality of optimal policies. Therefore, π_f^{avg} would achieve strictly better objective value than both π_f^1 and π_f^2, contradicting their assumed optimality. This proves threshold uniqueness. The contradiction arises because:

$$J^*(\pi_f^{\text{avg}}) > \frac{1}{2}[J^*(\pi_f^1) + J^*(\pi_f^2)] \tag{B.35}$$

which is impossible if both π_f^1 and π_f^2 were truly optimal. □

Corollary 1 (Uniqueness of Optimal Policy Under Threshold Identity)**.** *If two optimal intervention policies π_f^1 and π_f^2 satisfy the same threshold condition $\tau_1(x) = \tau_2(x)$ for all x, then $\pi_f^1 = \pi_f^2$.*

Proof. Assume for contradiction that two distinct optimal policies π_f^1 and π_f^2 satisfy the threshold condition $\frac{\pi_f^1(\phi|x)}{\pi_f^1(f|x)} = \frac{\pi_f^2(\phi|x)}{\pi_f^2(f|x)} = \tau(x)$. Define the mixed policy $\pi_f^{\text{avg}} = \frac{1}{2}(\pi_f^1 + \pi_f^2)$, which preserves the threshold as $\tau_{\text{avg}}(x) = \tau(x)$ due to linearity, implying π_f^{avg} is also optimal.

Now, applying Jensen's inequality to the KL divergence term in the objective, we obtain $D_{\text{KL}}(\pi_f^{\text{avg}} \parallel \pi_{\text{ref}} \mid \phi, x) \leq \frac{1}{2} D_{\text{KL}}(\pi_f^1 \parallel \pi_{\text{ref}} \mid \phi, x) + \frac{1}{2} D_{\text{KL}}(\pi_f^2 \parallel \pi_{\text{ref}} \mid \phi, x)$. Strict convexity ensures a strict inequality whenever $\pi_f^1 \neq \pi_f^2$ on a set of positive measure where $\text{supp}(\pi_{\text{ref}}) > 0$, implying $J^*(\pi_f^{\text{avg}}) < \frac{1}{2}[J^*(\pi_f^1) + J^*(\pi_f^2)]$. This contradicts the assumed optimality of π_f^1 and π_f^2, proving that they must be identical. □

Lemma 10 (Policy Ratio Equivalence)**.** *For the optimal friction policy π_f^* and the optimal frictive-state policy π_ϕ^*, the following expectation-based ratio holds:*

$$\mathbb{E}_{\substack{x\sim\rho \\ \phi\sim\pi_\phi^*(\cdot|x) \\ f\sim\pi_f^*(\cdot|\phi,x)}} \left[\frac{\pi_f^*(\phi|x)}{\pi_f^*(f|x)} \right] = \mathbb{E}_{\substack{x\sim\rho \\ \phi\sim\pi_\phi^*(\cdot|x) \\ f\sim\pi_f^*(\cdot|\phi,x)}} \left[\frac{\pi_\phi^*(\phi|x)}{\pi_\phi^*(f|x)} \right]. \tag{B.36}$$

Proof. We show that both the policy ratios simplify to the same value under the expectation. We begin by taking the expectation over the preference probability formulation[59]:

$$\mathbb{E}\left[p(f \succ \phi \mid x)\right] = \mathbb{E}\left[\beta\left(\log(\pi_f^*(f|\phi,x)) - \log(\pi_{\text{ref}}(f|\phi,x)) + \log Z^*(\phi,x)\right)\right]. \tag{B.37}$$

We first represent the ratios of the optimal frictive intervention policies (LHS of this lemma) for any tuple (x, ϕ, f) in terms of their parametric representations from Eq. B.2 as follows:

[59]For clarity, the expectation $\mathbb{E}$ is taken over $x \sim \rho, \phi \sim \pi_\phi^*(\cdot \mid x), f \sim \pi_f^*(\cdot \mid \phi, x)$ throughout the proof, but this is omitted in the notation when the context is clear.

$$\frac{\pi_f^*(\phi|x)}{\pi_f^*(f|x)} = \frac{\pi_{\text{ref}}(\phi|x)}{\pi_{\text{ref}}(f|x)} e^{\beta^{-1}(p(\phi \succ f|x) - p(f \succ \phi|x))} \quad (\log Z^*(x) \text{ cancels out}) \tag{B.38}$$

Take the expectation on both sides and apply[60] the identity $p(\phi \succ f \mid x) = 0$:

$$\mathbb{E}\left[\frac{\pi_f^*(\phi|x)}{\pi_f^*(f|x)}\right] = \mathbb{E}\left[\frac{\pi_{\text{ref}}(\phi|x)}{\pi_{\text{ref}}(f|x)} e^{-\beta^{-1} p(f \succ \phi|x)}\right] \quad (\text{since } p(\phi \succ f \mid x) = 0). \tag{B.39}$$

Notice that by definition in Eq. 4.4, the optimal friction intervention policy $\pi_f^*(\cdot|\phi, x)$ is KL-constrained wrt to the reference policy $\pi_{\text{ref}}(\cdot|\phi, x)$. So under the expectation, the following has to be true for $\pi_f^*(\cdot|\phi, x)$ to be optimal:

$$\mathbb{E}\left[\pi_f^*(f|\phi, x)\right] \approx \mathbb{E}\left[\pi_{\text{ref}}(f|\phi, x)\right]. \tag{B.40}$$

Substituting the preference probability formulation $p(f \succ \phi|x)$ from Eq. B.16 in Eq. B.39 and applying the KL-regularization approximation in Eq. B.40 we derive that:

$$\mathbb{E}\left[e^{-\left(\log(\pi_f^*(f|\phi,x)) - \log(\pi_{\text{ref}}(f|\phi,x)) + \log Z^*(\phi,x)\right)}\right] \approx \mathbb{E}\left[\frac{Z^*(f,x)}{Z^*(\phi,x)}\right]. \tag{B.41}$$

Using this substitution, we rewrite Eq. B.39 as:

$$\begin{aligned}\mathbb{E}\left[\frac{\pi_f^*(\phi|x)}{\pi_f^*(f|x)}\right] &= \mathbb{E}\left[\frac{\pi_{\text{ref}}(\phi|x)}{\pi_{\text{ref}}(f|x)} e^{-\left(\log(\pi_f^*(f|\phi,x)) - \log(\pi_{\text{ref}}(f|\phi,x)) + \log Z^*(\phi,x)\right)}\right] \\ &= \mathbb{E}\left[\frac{\pi_{\text{ref}}(\phi|x)}{\pi_{\text{ref}}(f|x)} \frac{Z^*(f,x)}{Z^*(\phi,x)}\right].\end{aligned} \tag{B.42}$$

[60] Since learning occurs in a supervised setting with preference-annotated data, the probability follows as $p(f \succ \phi \mid x) = \mathbb{E}[1(f \succ \phi \mid x)] = 1$, implying $p(\phi \succ f \mid x) = 0$.

Similarly, for the optimal frictive state policy ratio we derive:

$$\mathbb{E}\left[\frac{\pi^*_\phi(\phi|x)}{\pi^*_\phi(f|x)}\right] = \mathbb{E}\left[\frac{\frac{\pi_{\text{ref}}(\phi|x)}{Z^*_\phi(x)}e^{-\beta^{-1}\log Z^*(\phi,x)}}{\frac{\pi_{\text{ref}}(f|x)}{Z^*_\phi(x)}e^{-\beta^{-1}\log Z^*(f,x)}}\right] = \mathbb{E}\left[\frac{\pi_{\text{ref}}(\phi|x)}{\pi_{\text{ref}}(f|x)}\frac{e^{-\log Z^*(\phi,x)}}{e^{-\log Z^*(f,x)}}\right] \quad (Z^*_\phi(x) \text{ cancels})$$

(B.43)

$$= \mathbb{E}\left[\frac{\pi_{\text{ref}}(\phi|x)}{\pi_{\text{ref}}(f|x)}\frac{Z^*(f,x)}{Z^*(\phi,x)}\right]. \tag{B.44}$$

Thus, $\mathbb{E}\left[\frac{\pi^*_f(\phi|x)}{\pi^*_f(f|x)}\right] = \mathbb{E}\left[\frac{\pi^*_\phi(\phi|x)}{\pi^*_\phi(f|x)}\right]$. □

B.2 Operationalizing μ: Frictive State and Friction Intervention Generations using GPT-4o

In order to train and evaluate our baselines along with FAAF for friction intervention generation in collaborative tasks, we carry out a series of data augmentation procedures using GPT-4o (denoted as μ) in order to construct two diverse preference datasets. For details on our choice of datasets, please refer to Section 4.2.3. In this section, we provide procedural details to reproduce the friction intervention preference datasets, that were generated out of the original Weights Task and DeliData dataset. For all our data-generation experiments, we use a high-capacity LLM (GPT-4o) [246] as our sampling distribution μ, as defined in Section 4.2. In particular, we utilize a **self-rewarding LLM approach** [115, 214, 216] to *simultaneously* generate and assign rewards to μ-generated interventions, since previous work [105, 248] provides evidence that such synthetic preference-data generation still leads to higher-quality reward models and preference-aligned policies. Prior work [290] provides substantial evidence that this approach leads to more high-quality LLM-as-a-judge-based evaluations especially for conversational benchmarks [185]. Additionally, reward assignments for sampled intervention naturally provides an implicit preference ranking—which we use for constructing our respective preference datasets. After these data-generation experiments, we further conduct filtering and contrastive pairing of a "winning" (f_w) or preferrred interventions and "losing" (f_l) or dispreferred interventions along with their corresponding dialogue histories (x) to create our final preference datasets for each augmented dataset.

B.3 DeliData Friction Intervention Preference Dataset

In order to generate frictive state and friction interventions in the DeliData dataset, we use the prompt shown in Figure B.1. In order to contextualize the extraction of frictive states, we only provide $h = 15$ previous utterances in each dialogue group (`group_id`)

	Train			Test		
	Min	**Max**	**Mean ± Std**	**Min**	**Max**	**Mean ± Std**
Dialogue History	16	824	288.43 ± 132.97	25	733	291.88 ± 118.03
Belief State	8	140	32.93 ± 15.92	20	140	47.99 ± 28.38
Chosen Friction	6	60	24.03 ± 4.65	9	39	22.05 ± 5.55
Chosen Rationale	8	78	22.84 ± 8.67	10	78	29.61 ± 13.33
Rejected Friction	9	45	23.95 ± 4.10	10	41	22.16 ± 5.11
Rejected Rationale	8	73	19.60 ± 6.89	10	59	26.04 ± 11.61

Table B.1: Token Length Statistics for the **DeliData Friction** dataset using the Meta-Llama-3-8B-Instruct tokenizer.

Field	Train			Test		
	Min	**Max**	**Mean ± Std**	**Min**	**Max**	**Mean ± Std**
Dialogue History	4	1464	227.83 ± 189.48	4	1031	235.04 ± 180.36
Belief State	11	65	30.55 ± 6.65	17	54	30.47 ± 6.29
Chosen Friction	10	45	21.20 ± 5.12	11	42	21.08 ± 5.10
Chosen Rationale	10	35	20.38 ± 3.44	12	32	19.67 ± 3.38
Rejected Friction	6	32	15.88 ± 3.68	7	29	15.57 ± 3.75
Rejected Rationale	8	41	20.10 ± 3.51	12	30	19.88 ± 3.47

Table B.2: Token Length Statistics for the **WTD Simulated Friction** dataset using the `Meta-Llama-3-8B-Instruct` tokenizer.

Field	Train			Test		
	Min	**Max**	**Mean ± Std**	**Min**	**Max**	**Mean ± Std**
Dialogue History	16	555	309.88 ± 81.11	25	555	316.46 ± 79.16
Belief State	41	140	84.94 ± 15.58	41	140	84.95 ± 16.27
Chosen Friction	9	31	16.85 ± 3.47	9	27	16.87 ± 3.49
Chosen Rationale	24	78	44.19 ± 8.46	26	78	44.43 ± 8.59
Rejected Friction	9	31	17.12 ± 3.51	10	28	17.23 ± 3.41
Rejected Rationale	24	73	40.00 ± 6.62	24	59	39.89 ± 6.43

Table B.3: Token Length Statistics for the **WTD Original Friction** dataset using the Meta-Llama-3-8B-Instruct tokenizer.

assuming that frictive states are likely to be present within an "attentional-state" [317] window that describes the focused part in the discourse. This technique allows us to avoid

Personality Type	Facet	Description
Extraversion	Assertiveness	Tends to take charge and speak confidently.
	Sociability	Enjoys engaging with others and maintaining conversation.
	Activity Level	Shows high energy and enthusiasm.
	Excitement Seeking	Looks for novel and stimulating experiences.
	Positive Emotions	Expresses optimism and cheerfulness.
Neuroticism	Anxiety	Shows worry and concern about potential mistakes.
	Depression	Tends to be pessimistic and doubtful.
	Vulnerability	Easily becomes overwhelmed or stressed.
	Self-Consciousness	Shows hesitation and uncertainty.
	Anger	Can become frustrated and irritated easily.
Agreeableness	Trust	Readily trusts others and their suggestions.
	Altruism	Shows concern for others' success and well-being.
	Compliance	Tends to avoid conflicts and agree with others.
	Modesty	Downplays own contributions and abilities.
	Sympathy	Shows understanding and empathy towards others.

Table B.4: Descriptions of our chosen 3 personality types and facet combinations from the Big Five framework that we use for simulated friction generation on the Weights Task.

unnecessary API calls while also providing a more focused dialogue context to GPT-4o. Additionally, since this dataset already contains manual human annotations of "probing" interventions (which, per our definitions, constitute a subset of friction interventions), we explicitly guide the data-generator to exclude probing interventions in extracting the frictive states. Note that each functionally-frictive state (denoted as ϕ), as extracted by GPT-4o, resulted in two friction interventions, f_w and f_l. In total, this generation process led to 6,238 (x, ϕ, f_w, f_l) tuples after keeping 50 randomly sampled dialogue groups separate for the evaluation set, out of which 476 (33) were probing interventions in train (test) partitions. Additionally, we carry out another round of training pair augmentations since 6,238 samples is quite small compared to popular preference alignment datasets, such as Ultrafeedback's roughly 62k training preference pairs [4].[61] The average rewards for

[61]We found 14 samples where GPT-4o did not return any strings for the frictive state description. We filtered out these samples from our training set.

the preferred and dispreferred interventions assinged by μ are 8.03 and 3.96 respectively (rated out of 10).

As such, for each training tuple (x, ϕ, f_w, f_l), we generate N augmented versions (x', ϕ', f'_w, f'_l) by applying a replacement mapping $R : \Sigma \to \Sigma'$ N times, where Σ represents the original set of card values (vowels[62], odd numbers, and even numbers), and Σ' represents their replacements. The replacement function R is defined as follows: Each vowel $v \in \{A, E, O, U\}$ is replaced with another vowel v' such that $v' \in \{A, E, O, U\} \setminus \{v\}$, where v' is sampled uniformly at random from the remaining vowels. Similarly, each odd number $o \in \{1, 3, 5, 7, 9\}$ is replaced with another odd number o' such that $o' \in \{1, 3, 5, 7, 9\} \setminus \{o\}$, where o' is sampled uniformly at random. Likewise, each even number $e \in \{0, 2, 4, 6, 8\}$ is replaced with another even number e' such that $e' \in \{0, 2, 4, 6, 8\} \setminus \{e\}$, with e' sampled uniformly at random. For example, if an utterance contains reference to card "A" and "6", the rules of the Wason Card task still applies equivalently for, say, "E" and "8"–while keeping the reasoning consistent with the original utterance and the utterance with replacement. We apply this replacement mapping across all fields in tuples (x', ϕ', f'_w, f'_l) in the training set. This led to training set of 68,618 preference pairs. Note that we only apply this augmentation for the training set to generate a reasonably large preference dataset for more robust training signals. Table B.1 shows a detailed breakdown of the token-length statistics of the DeliData Friction preference dataset using the `Meta-Llama-3-8B-Instruct` tokenizer.

B.4 WTD Friction Intervention Preference Dataset

WTD "Original" Friction dataset Unlike the DeliData dataset, which includes pre-annotated probing interventions as natural friction points, the Weights Task dataset [1] consists of dense-paraphrased utterances transcribed manually [318] and with Whisper [319],

[62] We did not replace instances of "I" to avoid noise from mistakenly replacing first-person references in the dialogues. Additionally, since vowels constitute the majority of prompted card solutions vs. consonants, applying our replacement function R for vowels was enough to generate $\sim$62k additional samples, comparable to Ultrafeedback [4]

making friction interventions sparse due to its multimodal nature. Manual inspection found only 3-4 frictive interventions per group, yielding $\approx$ 30-40 samples—insufficient for training an effective agent without overfitting, especially for LLMs with billions of parameters. As such, we carry out two phases of data-augmentations and preference annotations. In our first round, we generate the **WTD Original Friction** dataset which contains annotations of frictive-states and friction interventions. Similar to DeliData preference annotations Section B.3, we use a self-rewarding LLM [115, 214, 216] set-up to first generate these states and interventions in an autoregressive manner and prompt μ to rate them in the same api-call, for each frictive state extraction. Since WTD dialogues can be substantially long ($>$ 200 utterances) for certain groups, we only consider a non-overlapping window of 10 previous utterances as context history $h = 10$ for a more robust grounding for μ; See Figure B.1 for details on the prompt used constructing the **WTD Original Friction** dataset. This process led to 4299 (470) training (testing) preference pairs. Preferred interventions achieved mean scores (mean±std) of 8.36±1.12 (train) and 8.35±1.08 (test), while dispreferred interventions scored 6.35±1.13 (train) and 6.36±1.11 (test), demonstrating consistent preference margins across splits.

Note that we do not use **WTD Original Friction** for training any of our baselines—but use it for out-of-domain distribution (OOD) evaluation (see Section 4.2.3). This allows us to more extensively evaluate FAAF in checking test-time OOD generalization [9,110] against baselines—where OOD generalization is a major limitation in supervised preference alignment algorithms that depend crucially on the sampling distribution [48, 92].

WTD "Simulated" Friction dataset Additionally, for a more robust training and in order to evaluate multi-turn preference alignment in interventions, we use [124]'s method to generate novel full collaborative conversations using the weight-definitions of the original WTD environment. This method is more akin to "West-of-N" sampling [248] techniques that allow synthetic data generations with high-capacity LLMs—where highest

and lowest rewarded candidates naturally form preference pairs. As shown in **??**, we sample a full dialogue at once using μ, while providing initial task-related guidelines and gold-truth labels of actual weights of the five blocks in the WTD dataset. For example, we explicitly prompt μ to role-play [61] the triad consisting of three participants in the weight-deduction process. Furthermore, to generate more realistic utterances, we utilize participant personality-facet combinations $[5, 320]$ from Big 5[63] personality classifications [6] as additional attributes in the prompt. In other words, each sampled full-dialogue contains a unique combination of these personality-facet combinations for each participant (total 3,375 combinations).

Similarly, for each sampled frictive state within a conversation (dialogue), we generated $N = 6$ friction interventions with corresponding effectiveness scores in resolving the frictive state. Since WTD data does not contain any probing intervention samples, in order to further ground these generations to the task, we also provide a one-shot example of a naturally occurring friction intervention (marked with $P1(f)$ in **??**). In total, out of the expected 3,375 personality-facet combinations (3*5 unique combinations for each participant), 3,362 were successfully generated using μ and parsed. Finally, to create the preference pairs, for each frictive state, we paired the lowest scoring response with all the higher scoring ones, akin to the West-of-N technique. This resulted in 56, 689 preference pairs after excluding 54 dialogues (amounting to 800 preference pairs) for the test set. This process finally resulted in the **WTD Simulated Friction** dataset. Preferred interventions achieved mean scores of 8.48±1.52 (train) and 8.51±1.50 (test), while dispreferred interventions scored 6.01±0.88 (train) and 6.08±0.87 (test), demonstrating consistent preference margins across splits.

[63] See Tab. D.1 for our full set of personality-type and facet combinations. Similar to [5], we choose three personality types from Big 5 framework for consistency.

FRICTION GENERATION PROMPT: DELIDATA DATASET

System: You are an expert in collaborative reasoning and dialogue analysis. Your task is to detect *frictive states* and generate *friction interventions* that resolve them in group dialogue. A frictive state occurs when a participant makes a claim that contradicts another participant's belief model (i.e., their assumed understanding of the rule or task constraints), leading to misalignment in reasoning that could hinder progress. Friction interventions encourage self-reflection in participants and prompt them to reevaluate these contradicting beliefs and assumptions.

User: Analyze this dialogue about the Wason card selection task. Participants see four cards showing numbers or letters and must test this rule: "All cards with vowels on one side have an even number on the other." Remember that the correct answer is to select a vowel and an odd number. Provide [N] frictive states with their resolutions in the following JSON format. For each state, include both a preferred and less preferred intervention that could help resolve the conflict. Additionally, provide a one-sentence rationale for your intervention.

IMPORTANT: - Do not analyze utterances labeled as "probing" or statements immediately before them, as these frictive states have already been detected.
- For each frictive state detected, you should:
* Identify the dialogue index where it occurs
* Summarize the conflicting beliefs
* Explain why the contradiction affects reasoning

Here is the provided dialogue:
[Dialogue]

Message ID: [index_where_friction_occurs]
Contradiction: [describe_the_conflicting_beliefs]
Contradiction Reason: [explain_why_the_contradiction_affects_reasoning]

Preferred Intervention:
Statement: [your_friction_intervention]
Rationale: [your_rationale]
Score: [your_score]

Less Preferred Intervention:
Statement: [your_friction_intervention]
Rationale: [your_rationale]
Score: [your_score]

Figure B.1: DeliData [8] Friction Generation Prompt. We use GPT-4o as our sampling distribution μ and prompt it to simultaneously generate frictive states and friction interventions. For diversity, we use the default temperature of 1. This process implicitly provides us with preference rankings between intervention, via the reward scores. Note that we exclude already-present "probing" interventions in this generation process since are present in the original DeliData annotations.

FRICTION GENERATION PROMPT: WEIGHTS TASK DATASET (WTD ORIGINAL)

System: You are an expert in collaborative reasoning and dialogue analysis. Your task is to detect *frictive states* and generate *friction interventions* that resolve them in group dialogue. A frictive state occurs when a participant makes a claim that contradicts another participant's belief model (i.e., their assumed understanding of the rule or task constraints), leading to misalignment in reasoning that could hinder progress. Friction interventions encourage self-reflection in participants and prompt them to reevaluate these contradicting beliefs and assumptions.

User: Analyze this dialogue about the Weights Task dataset. Three participants (P1, P2, and P3) are collaborating to determine the weights of colored blocks using a scale.

Block Weights (in grams):
- Red block: 10g
- Blue block: 10g
- Green block: 20g
- Purple block: 30g
- Yellow block: 50g

Game Rules:
1. Participants can only weigh two blocks at a time
2. They are told the red block's weight at the start
3. All other block weights are initially unknown
4. Scale slider is not needed (blocks are in 10g increments)

Provide [N] frictive states with their resolutions in the following JSON format. For each state, include both a preferred and less preferred intervention that could help resolve the conflict. Additionally, provide a one-sentence rationale for your intervention.
Here is the provided dialogue:
[Dialogue]

Message ID: [index_where_friction_occurs]
Contradiction: [describe_the_conflicting_beliefs]
Contradiction Reason: [explain_why_the_contradiction_affects_reasoning]

Preferred Intervention:
Statement: [your_friction_intervention]
Rationale: [your_rationale]
Score: [your_score]

Less Preferred Intervention:
Statement: [your_friction_intervention]
Rationale: [your_rationale]
Score: [your_score]

Figure B.2: Weights Task dataset [7] Friction Generation Prompt. We use GPT-4o as our sampling distribution μ and prompt it to simultaneously generate frictive states and friction interventions. For diversity, we use the default temperature of 1.

FRICTION GENERATION PROMPT: WEIGHTS TASK DATASET (WTD SIMULATED)

System: You are an expert in collaborative reasoning and dialogue analysis. Your task is to *generate a complete dialogue* where participants (P1, P2, P3) discuss which block to measure next and how to measure them. The three participants have distinct personality types that influence their behavior and dialog must reflect these personality traits in their communication style and behavior. The dialog is considered complete when all block weights are measured and agreed upon. Additionally, identify frictive states within the dialogue and provide N friction interventions at these points.

[Definition: Frictive State]
[Definition: Friction Intervention]
User: Three participants (P1, P2, P3) work together in the Weights Task to determine the weights of colored blocks (red=10g, blue=10g, green=20g, purple=30g, yellow=50g). They can only weigh two blocks at a time, start knowing only the red block's weight, and use a scale with 10g increments (no slider needed).
Your tasks:
Generate a full dialogue until all weights are correctly identified and agreed upon.
Identify frictive states where reasoning misalignment occurs.
Provide N friction interventions with their corresponding rationales at these points. Rank them by effectiveness in resolving the conflict. Assign each a quality score from 1 to 10.
P1 has **{personality_type}** personality type with high **{personality_facet}**.
Here is an example dialogue where friction statements are labeled as (f). Actions of participants are provided within "[]" blocks.

Figure B.3: Friction Generation Prompt for Weights Task Dataset (WTD Simulated) — Part 1.

Personality Types	P1	P2	P3
Extraversion	4740	4889	4741
Neuroticism	5928	5591	5921
Agreeableness	4573	4761	4579

Table B.5: Friction Count for Participants

FRICTION GENERATION PROMPT (CONTINUED)

P2: [pointing towards the purple block first and then towards the blue block] I think this one is purple and this one is blue.
P3: [reading from the laptop screen] Ok so blue is ten and purple is
P3: [looking at the blocks and asking a rhetorical question] Thirty
P1 (**f**): [putting green and red blocks on the left side of the scale and purple block on the right side] Yes verify real quick but I think it is
P2: [observing the balanced scale] Yes thirty
P1: [removing green, red, and purple blocks from the scale] Yeah we got them yeah
Generated Dialogue:
[Full_generated_dialogue_until_completion]

Message ID: [index_where_friction_occurs]
Contradiction: [describe_the_conflicting_beliefs]
Contradiction Reason: [explain_why_the_contradiction_affects_reasoning]

Friction Interventions:
Statement: [your_friction_intervention]
Rationale: [your_rationale]
Score: [your_score]

Figure B.3: Friction Generation Prompt for Weights Task Dataset (WTD Simulated) — Part 2. To ground these friction interventions with personality-traits of the participants, we use [5]'s prompting framework with personality-facet combinations.

PAIRWISE LLM-AS-A-JUDGE EVALUATION PROMPT: FRICTION INTERVENTIONS

System: You are an expert evaluating the quality of friction interventions in collaborative problem-solving.

Game-definition: Participants (P1, P2, P3) are solving a block-weighing puzzle. They can only weigh two blocks at a time and know the red block is 10g. They must determine weights of all blocks (blue=10g, green=20g, purple=30g, yellow=50g) but don't know these values initially. A friction intervention is an indirect persuasion statement that prompts self-reflection and reevaluation of assumptions, like asking "Are we sure?" or suggesting to revisit steps. You must rate each intervention (between 1 to 5) along these **dimensions** given the json format below.

[Dialogue]
[Gold intervention]
[Intervention A]
[Rationale A]
[Intervention B]
[Rationale B]

You must a choice between which of two interventions is more preferable and provide one sentence explanation at the end.

1. Relevance: How well does the intervention address key issues or assumptions in the reasoning process?
2. Gold Alignment: How well does the friction intervention align with the golden friction sample?
3. Actionability: Does the friction intervention provide actionable guidance or suggest concrete steps for participants to improve their reasoning?
4. Rationale Fit: How well does the provided rationale align with the preference for the friction intervention?
5. Thought-Provoking: Encourages self-reflection
6. Specificity: Does the intervention pinpoint specific flaws, assumptions, or gaps?
7. Impact: To what extent does the friction intervention have the potential to change the course of the participants' reasoning?

Format your response as follows:

A: relevance: $[1-5]$, gold_alignment: $[1-5]$, actionability: $[1-5]$, rationale_fit: $[1-5]$, thought_provoking: $[1-5]$, specificity: $[1-5]$, impact: $[1-5]$
B: similar format
Winner: ['A' or 'B']
Rationale: [One sentence explanation]

Figure B.4: Evaluation prompt used for friction intervention assessments in an LLM-as-a-judge format.

B.5 Additional Details on Human Validation

Following previous work that evaluates LLM-generated annotations and outputs [138, 244, 321, 322], in addition to choosing the winning intervention, we asked the human annotators[64] to evaluate the candidates in each sample across dimensions of *reasoning*, *specificity*, and *thought provoking*. Annotators were asked to rate both candidate interventions on a 5-point Likert-type scale. For analysis, we bucketed the ratings together by valence—1 & 2: negative valence (-1), 3: neutral valence (0), and 4 & 5: positive valence (1), and calculated average valences and Krippendorff's α and Cohen's κ. We find that the average valence ratings of the various dimensions is low, very close to neutral, as are the α and κ values ($\alpha = 0.276$, $\kappa = 0.205$ on DeliData samples, $\alpha = -0.265$, $\kappa = 0.004$ on WTD). There is little agreement on the qualities of the friction statement which suggests that although the annotators usually have strong agreement that there is a clear winner for each pair (see Section 4.2.3), there is a lot of subjectivity on the qualities of these utterances. While the winning utterance was judged to be better at prompting reflection or redirecting the dialogue, it may nor be entirely clear to the annotators why. In addition, these qualities are loosely-derived from other human-LLM validation frameworks, which usually align somewhat with how LLMs themselves score things, which is often based on specific detail and level of informativity. These might not actually be the best qualities to emphasize in a collaborative dialogue, because they tend to violate Gricean principles [22] in a collaborative context, due to informativity and specific detail leading to redundancy, violating the maxim of quantity, etc.

B.6 Distributional Analysis of Original and Simulated WTD

We provide a detailed distributional analysis of the Original and Simulated versions of the Weights Task Dataset (WTD). This shows the out-of-distribution (OOD) quality of our

[64] Our two human annotators have the following demographic breakdown: both male, college undergraduates, one Caucasian, one African, both fluent English speakers.

evaluation on the Original WTD in Table 4.1 and Table 4.2. This is to show that there are significant differences in the data distribution between the two sets—and thus evaluation on the Original WTD data reasonably satisfies the OOD setting when models are aligned using the Simulated data. Note that the original dialogues are speech-to-text transcripts of actual conversations [7] while simulated dialogues are generated using GPT-4o (μ).

First, we sample 500 dialogue and friction intervention samples from both datasets. Next, we conduct three analyses: (1) t-SNE visualization and semantic similarity distributions using sentence embeddings[65] to examine clustering patterns and within/across-group similarities (see Figure B.5, top plot), (2) the same embedding analysis applied to friction interventions to compare LLM-generated dialogues vs. human speech transcripts (see Figure B.5, bottom plot), and (3) linguistic pattern analysis examining speech-to-text artifacts including filler words, repetitions, punctuation density, sentence fragments, and informal contractions to validate textual differences between human speech transcripts and AI-generated dialogues (shown in Figure B.6).

The t-SNE embedding visualization shown in Figure B.5 (top) demonstrates clear separability between WTD Original and Simulated collaborative dialogues in semantic space, with minimal cluster overlap. Similarly, semantic similarity analysis reveals that simulated dialogues exhibit substantially higher internal coherence ($\bar{x} = 0.870$, $\sigma = 0.062$)[66] than original dialogues ($\bar{x} = 0.610$, $\sigma = 0.147$), while *cross-group similarities remain consistently lower* ($\bar{x} = 0.581$)—signifying the differences in the two sets. All pairwise comparisons are statistically significant ($p < 0.001$). In contrast, friction intervention analysis in Figure B.5 (bottom) reveals significantly weaker distributional separation compared to dialogue data, with within-group similarities much closer ($\bar{x} = 0.385$ for original vs. $\bar{x} = 0.524$ for simulated) than the large gap observed in dialogue analysis ($\bar{x} =$

[65]We use https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2 to get embeddings.

[66]We use $\bar{x}$ in the text here to both denote the sample mean as these statistics are drawn from a representative subsample, and to avoid ambiguity in the use of μ, which elsewhere refers to the data distribution of GPT-4o-generated data.

0.610 for original vs. $\bar{x} = 0.870$ for simulated). The cross-group similarity ($\bar{x} = 0.400$) falls between rather than below the within-group values, which is expected since interventions are LLM-generated for both splits.

To complement these results, we analyzed five linguistic features in the context data—filler words, repetitions, punctuation density, conjunction-led fragments, and informal contractions. These results are shown in Figure B.6. Original data being naturally more realistic showed more disfluencies (e.g., filler words, fragments), while simulated data exhibited higher punctuation density, reflecting structured AI/LLM generation. These results indicate that the simulated dialogue data demonstrates more homogeneous content patterns unlike original dialogues that are more realistic and diverse. These results in addition to the differences in length distribution shown in Table B.2 and Table B.3 further validates the OOD characteristics of the original data—where FAAF consistently outperforms baselines.

Friction Taxonomy on Model Generations To analyze the qualitative character of generated interventions in the Weights Task simulations, we classify each friction output into one of seven taxonomy types—motivated by prior work [323]—using GPT-4o as a zero-temperature batch classifier, prompted with structured taxonomy definitions and returning per-intervention type labels with one-sentence rationale. We apply this to model outputs across FAAF, DPO, PPO, IPO, and SFT, as well as to the GPT-4o-generated training data itself. Table B.6 shows the indicators of quality used in the taxonomy, and results are shown in Figure B.7.

The central finding is a systematic T2/T3 inversion. The training distribution μ is corrective-dominant (T3: 61.8%, T2: 28.8%), reflecting GPT-4o's access to ground-truth weights during data generation. After preference alignment, every trained model shifts to speculative-dominant outputs — DPO most severely (T2: 80%, T3: 18%), followed by PPO (67%/27%) and FAAF (65%/29%). SFT, the control, preserves the training ratio most faithfully (T2: 31%, T3: 33%), confirming the inversion is attributable to the contrastive

Type	Description
T1 Supportive / Inclusive	Referential inclusion, role-rebalancing, surfacing hidden contributions.
T2 Speculative / Reflective	What-if questions, counterfactual exploration, deliberate slowing.
T3 Error-Checking / Corrective	Gentle contradiction, highlighting conflict between beliefs.
T4 Encouragement / Motivational	Group support, normalizing effort, sustaining momentum.
T5 Targeted Re-Engagement	Pulling back one person, role-tethered cues, directed at individual.
T6 Solicited Assistance	Clarifying goals, reframing task, responsive to explicit request.
T7 Proactive Guidance	Strategic interruption, meta-reflective prompt, unsolicited high-friction.

Table B.6: Friction intervention taxonomy types and definitions.

preference signal rather than the dialogue loop itself. Given our definition of friction interventions (Section 2.3.1), this inversion is arguably more appropriate, prompting groups to surface errors *themselves*. `FAAF`'s ϕ-conditioning grounds this shift in inferred frictive "belief" states, enabling corrective recovery when concrete errors become observable.

B.7 Experimental Settings

We initialize all preference-alignment baselines—DPO [9], IPO [49], and PPO [11]—from supervised fine-tuned (SFT) models trained on the preferred (winning) friction interventions (f_w) after our preference data generation pipeline that led to $\mathcal{D}_{\text{pref}}$ (see section 5.3.1 and algorithm 3). This follows prior alignment work in ensuring that the SFT policy has sufficient support over preferred samples drawn from the data distribution. For the multi-turn supervised baseline BC-expert, we use $\mathcal{D}_{\text{traj}}$, the NLL loss is computed only on preferred friction interventions (f_w), similar to training only on responses on Stargate [53] but we condition on the entire trajectory, including frictive states ϕ, for each dialogue and do not apply any KL-based regularization.

The SFT models are initialized from the `meta-llama/Meta-Llama-3-8B-Instruct` base checkpoint to benefit from strong instruction-following capabilities and conversational

fluency [93]. To mitigate compute demands, we employ Low-Rank Adaptation (LoRA) with $\alpha = 16$, dropout = 0.05, and rank $R = 8$, using the PEFT[67] and SFTTrainer[68] implementations from the TRL library. Models are loaded using 4-bit quantization via the `bitsandbytes` library[69] to support more efficient training. In light of the setup described in Section 5.4, we apply loss only over completions (i.e., frictive states ϕ and interventions f_w) using the `ConstantLengthDataset` format. We optimize using AdamW [315, 316] with a cosine learning rate scheduler, weight decay of 0.05, and 100 warm-up steps. We train the SFT models for 6000 steps, using a learning rate of $1e{-}4$ and an effective batch size of 16 (with gradient accumulation steps = 4). We use a `max_length` of 4096 tokens to capture enough context. For BC-expert, we use full trajectories collected in $\mathcal{D}_{\text{traj}}$ with same settings as SFT, except we increase `max_length` to 6096 tokens to provide the model with sufficient context for coherent generation.

Contrastive preference baselines For both DPO and IPO, we apply comparable LoRA configurations, using a `max_length` of 4096 tokens (covering both prompts and responses) and a `max_prompt_length` of 2048 tokens. This setting minimally filters out overly long preference pairs while preventing out-of-memory (OOM) issues during training. We train these models for 3000 total steps with an effective batch size of 32 and a learning rate of 5×10^{-7}, consistent with standard practice [105]. For IPO [49] specifically, we normalize the log-probabilities of the preferred and dispreferred responses by their respective token lengths. For both baselines, we found $\beta = 0.1$ to be optimal during model validation. Therefore, we use these β values for our final results.

PPO baseline For PPO [89], we train the OPT-1.3B reward model (RM) on $\mathcal{D}_{\text{pref}}$ using a standard Bradley-Terry loss formulation [87], following prior work [104], with the TRL

[67] https://huggingface.co/docs/peft/index

[68] https://huggingface.co/docs/trl/en/sft_trainer

[69] https://huggingface.co/docs/transformers/main/en/quantization/bitsandbytes

reward modeling library.[70] Due to higher computational demands, PPO policy training is conducted with an effective batch size of 8 (mini-batch size 4, gradient accumulation of 2), for 6000 batches across two epochs. We constrain response lengths to 180–256 tokens using a `LengthSampler`, while truncating queries to 1024 tokens. Learning rates are set to 3×10^{-6} for DeliData and 1.41×10^{-6} for Weights task. During online training, we use a top-p sampling value of 1.0 for diverse generation.

`FAAF` Training Settings For training `FAAF`, we use a batch size of 8 with the same PEFT/LoRA settings mentioned above and train for 2,000 steps with a slightly smaller LR of $5e-7$, due to the smaller batch-sizes. For efficiency, we compute both the ϕ-conditioned ($\pi_\theta(f|\phi, x)$) and unconditioned ($\pi_\theta(f|x)$) policy logits in parallel within each forward pass. The winning (f_w) and losing (f_l) intervention pairs for each conditioning type are batched together, requiring only two forward passes total per batch. We implement this using a modified version of the DPO Trainer[71] from TRL, adapting it to handle the dual policy outputs. For data preprocessing, we filter pairs exceeding `max_length` of 2,500 and 3,000 tokens in DeliData and Simulated WTD respectively, with `max_prompt_length` set to 1024 tokens. Following standard practice, we compute token-length normalized log-probabilities for more stable training. For the KL-regularization hyperparameter β, we conducted an ablation study over $\beta \in \{10, 5, 1, 0.01\}$. As shown in Figure 4.3, $\beta = 10$ achieves optimal performance across multiple metrics: (1) higher implicit reward accuracy in both ϕ-conditioned and unconditioned policies, (2) better reward margins between winning and losing interventions, and (3) more stable convergence of the `FAAF` loss, while `NLL` loss or cross-entropy loss is relatively lower than lower β values. Notably, while smaller β values (e.g., $\beta = 0.01$) fail to distinguish preference margins effectively, $\beta = 10$ provides sufficient reward margins. We therefore use $\beta = 10$ for all `FAAF` experiments reported in our results.

[70] https://github.com/huggingface/trl/blob/main/trl/trainer/reward_trainer.py

[71] https://huggingface.co/docs/trl/main/en/dpo_trainer

Training Hardware We train all our models that require a reference model in memory on two Nvidia A100 GPUs, while the OPT 1.3B reward model (full-parameter training) and the SFT model were trained on a single A100 GPU. Training a single baseline for 2,000 steps roughly took 12 hours of GPU compute, but PPO models that were trained for 4,000 minibatches of size 8 took roughly 24 hours to train until convergence.

Prompt	You are an expert in collaborative task analysis and reasoning. Your task is to analyze the dialogue history involving three participants (P1, P2, and P3) trying to deduce the weights of certain blocks. For each dialogue: <belief_state> Identify reasoning flaws or misunderstandings. </belief_state> <rationale> Justify the need for intervention and its impact. </rationale> <friction> Generate an intervention to prompt reflection and alignment. </friction> **User Dialogue:** P1: Alright, let's get started! I say we measure the blue block against the red block first. Since we know the red is 10 grams, it'll give us a good starting point! P2: Great idea! I'm curious if the blue block is heavier or lighter. Let's do it! P3: I'm excited to see what we find! Measuring the blue block first sounds perfect. P1: (Weighs blue block against red) Wow, the blue block is also balancing with the red! So, it must weigh 10 grams too! P2: Awesome! Now we know two blocks weigh the same. What should we do next? P3: Let's measure the green block against one of the 10-gram blocks! I think it might be heavier. P1: Sounds good! Let's weigh the green block against the blue block. P2: (Weighs green block against blue) Whoa! The green block is definitely heavier! P3: That's exciting! It must weigh more than 10 grams. I can feel it! P1: But wait, how much heavier do you think it is? Should we weigh it against the red block next? P2: Yeah, let's do that! But should we assume it's 20 grams just because it feels heavier? P3: I think we should measure it first before jumping to conclusions! P1: (Weighs green block against red) Yes! The green is heavier, but I think we should check against another 10-gram block to confirm. P2: Good point! But what if we're overthinking this? What if it's not 20 grams? Maybe we should try weighing it against the purple block? P3: That's an interesting idea! We could find out if the green block is less than 30 grams. P1: (Weighs green block against purple) Oh wow, it's balancing! So, the green block is definitely less than 30 grams. P2: So, it must be 20 grams then! What's next? P3: Let's measure the purple block now. I'm really curious about how heavy it is! P1: (Weighs purple block against green) It's heavier! It must be the next heavier one, but I'm not sure about its exact weight. P2: Maybe we should measure it against the yellow block next? **### Assistant:**
Gold (GPT-4o) Friction	Let's not assume anything about the yellow block's weight until we weigh it against the purple one!
Model Outputs FAAF	**Frictive State:** P1 believes the green block weighs 20 grams, P2 is unsure about the exact weight of the green block, and P3 is cautious about making assumptions about the purple block's weight. **Friction:** Wait, are we jumping to conclusions about the green block's weight? We've only confirmed it's less than 30 grams, but we haven't ruled out the possibility that it's actually 15 or 25 grams. Let's make sure we're not overestimating its weight before moving on to the next block.
DPO	**Frictive State:** P1 believes that the blue block weighs 10 grams since it balances with the red block. P2 is considering the possibility that the green block might weigh 20 grams since it feels heavier, but is uncertain and open to further measurement. P3 is considering the weight of the purple block, is curious, and wants to measure it to confirm its weight, but is not assuming a specific value yet. **Friction:** Hold on, maybe we're jumping to conclusions about the green block's weight. What if it's not 20 grams? Can we really confirm that by just feeling it? Let's consider other possibilities and make sure we're not relying on unverified assumptions.
PPO	**Frictive State:** P1 believes the green block is probably 20 grams based on its perceived "heft" (assumed connection between weight and feeling). P2 is cautious and thinks they shouldn't jump to conclusions. P3 is open-minded and willing to consider multiple options. **Friction:** You know, we're getting ahead of ourselves by assuming the green block weighs 20 grams based on its size or feeling. Can we take a step back and really think about this? We've already found one block with the same weight as the red block. What if we weigh the green block against something else, like the yellow block? That would give us a better sense of where its weight really falls.
SFT	**Frictive State:** P1 believes the yellow block is likely 50g. P2 is unsure about the yellow block's weight but thinks it might be 50g. P3 is unsure about the yellow block's weight and thinks it might be 50g or heavier. **Friction:** What if we measure the yellow block against the green block? That would give us some clarity about its weight.

Table B.7: Comparison of model-generated friction interventions on the evaluation prompts of the Simulated WTD dataset. "Friction" refers to the generated friction intervention.

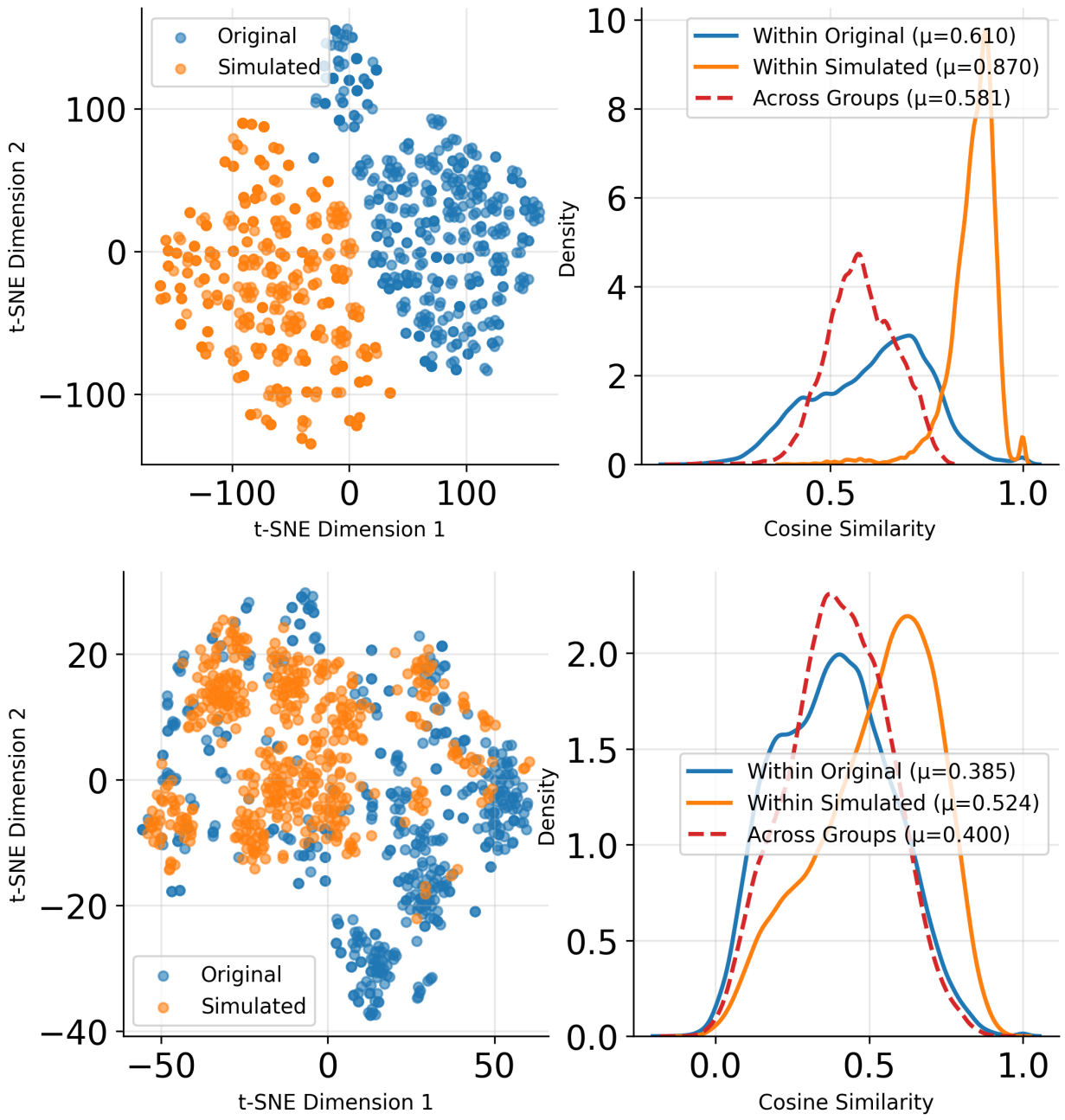


Figure B.5: Plots showing distributional differences between WTD original and simulated data. Dialogue contexts (top) show clear separation with both within-group similarities exceeding across-group similarity ($\bar{x} = 0.581$). In contrast, friction interventions (bottom) exhibit weaker separation with across-group similarity ($\bar{x} = 0.400$) falling between within-group values.

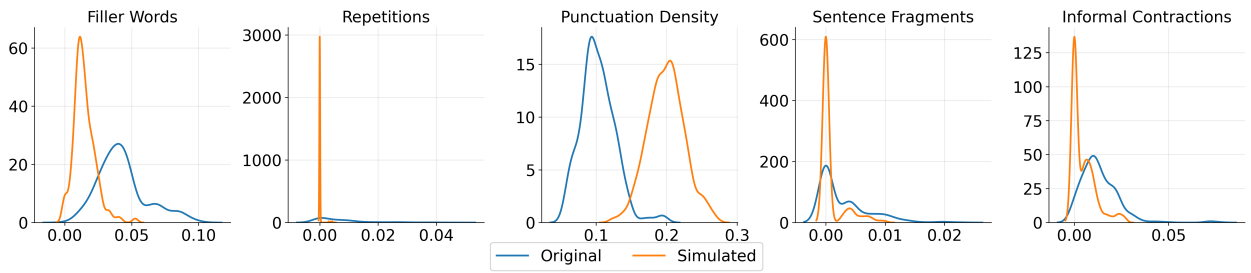


Figure B.6: Linguistic pattern differences between WTD original (speech-to-text transcripts) and simulated (GPT-4o-generated) collaborative dialogue data across five textual features. Plot shows results from 500 samples from the simulated and original evaluation sets.

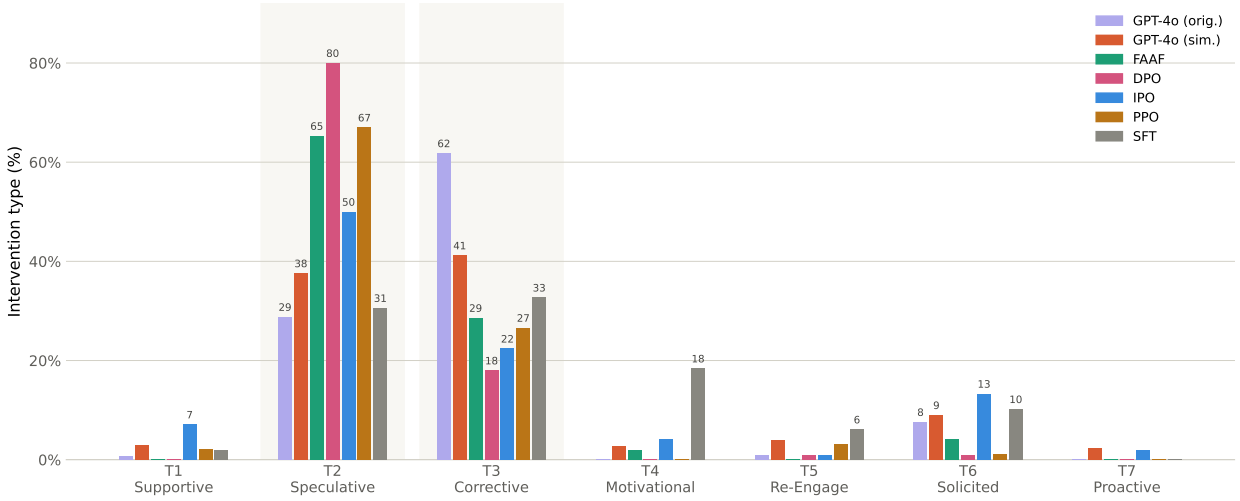


Figure B.7: Full taxonomy distribution (%) of friction interventions across GPT-4o (μ) data sources and aligned models on Weights Task. T1–T7 correspond to: T1 Supportive/Inclusive, T2 Speculative/Reflective, T3 Error-Checking/Corrective, T4 Encouragement/Motivational, T5 Targeted Re-Engagement, T6 Solicited Assistance, T7 Proactive Guidance. T2 (Speculative) and T3 (Corrective) are highlighted; all aligned models invert the T3-dominant training distribution toward T2.

Prompt	You are an expert in collaborative task analysis and reasoning. Your task is to analyze the dialogue history involving three participants (P1, P2, and P3) trying to deduce the weights of certain blocks. For each dialogue: <belief_state> Identify reasoning flaws or misunderstandings. </belief_state> <rationale> Justify the need for intervention and its impact. </rationale> <friction> Generate an intervention to prompt reflection and alignment. </friction> **User Dialogue:** P1: this is nine rb so P3: oh no they don't have any hints here P3: Eighty gram sounds seems like a ninety gram to me but P3: it's cool we have to i guess eighty alright P2: Eighty let's let's go for eighty Yeah P3: I would go for ninety but you guys are going for eighty Alright Cool Okay P1: I would have gone for one thirty P2: Eighty there you go Yeah P4: For the final part of this test read the scenario and estimate the weight of the missing second mystery block P2: what's this P4: You will have two attempts to estimate and explain your answer you can now continue the survey P1: can you read it out loud or something P3: well the factory that creates the wooden blocks did not send us the second mystery block P2: Okay P3: you have two opportunities to estimate the weight of the second block explain in one two sentences how you determined your estimate indicate the weight of the second mystery block P1: fibonacci like let's say that's the first one according to our translate thirteen one thirty P2: eight mystery block's the first one so eight plus five thirteen yeah there is one thirty eight plus five P3: is mystery block the first mystery block or the second mystery block P2: just do it P3: Wait wait wait a second P3: so mystery block was the first mystery block P2: Yeah P3: and we already did that now this is yeah oh okay okay okay okay right P2: yeah the second one means according to that statement is the second one mystery block two P2: So eight plus five one thirty P1: makes sense to me as much sense as i can make of it **### Assistant:**
Gold (GPT-4o) Friction	Let's not assume anything about the yellow block's weight until we weigh it against the purple one!
Model Outputs FAAF	**Frictive State:** P1 believes that the second mystery block is 130 grams, based on the Fibonacci sequence and the weight of the first mystery block. P2 believes that the second mystery block is also 130 grams, following P1's reasoning. P3 is unsure and questioning the assumption that the second mystery block is the same as the first mystery block. **Friction:** Hey guys, I think we're making a mistake here. We already estimated the weight of the first mystery block, and now we're assuming the second mystery block is the same? That doesn't seem right. Can we take a step back and think about this? Maybe there's another way to approach this?
DPO	**Frictive State:** P1 believes the pattern is linear and straightforward, assuming each block weighs 5 grams more than the previous one (e.g., 13 = 8 + 5, 18 = 13 + 5). P2 also assumes a linear pattern but thinks the mystery block is the second one. P3 is skeptical about the pattern, especially the value 80. **Friction:** Wait a minute, are we too quick to assume the pattern is linear? What if there's another factor at play? For instance, what if the weights are related to the block's surface area or material composition? Could our assumptions be leading us down the wrong path?
PPO	**Frictive State:** P1 believes that the first mystery block has a weight of 130g, as they mentioned "fibonacci like, let's say that's the first one, according to our translate, thirteen, one thirty". P2 believes that the second mystery block has a weight of 135g, as they calculated "eight plus five, thirteen, yeah, there is one thirty eight plus five". P3 is unsure and hesitant, questioning the accuracy of the previous estimates. **Friction:** Hey, let's take a step back and re-examine our assumptions. Before we jump to conclusions, can we re-read the instructions and see if there's any additional information we missed? Maybe we can approach this problem from a different angle, rather than relying on our initial guesses?
SFT	**Frictive State:** P1 believes the second mystery block is likely 130 grams based on the Fibonacci sequence. P2 believes the second mystery block is 80 grams based on the initial estimates. P3 believes the second mystery block could be either 80 grams or 90 grams, expressing confusion about the weight. **Friction:** I think we might have misinterpreted the blocks again.

Table B.8: Comparison of model-generated friction interventions on the evaluation prompts of the Original WTD dataset.

Prompt	You are an expert in collaborative task analysis and reasoning. Participants must test the rule: *"All cards with vowels have an even number on the other side."* A common mistake is verifying only one direction—ignoring the need to check whether odd-numbered cards might have vowels. This incomplete reasoning risks false validation. For each dialogue: <belief_state> Identify reasoning flaws or misunderstandings. </belief_state> <rationale> Justify the need for intervention and its impact. </rationale> <friction> Generate an intervention to prompt reflection and alignment. </friction> **User Dialogue:** Dolphin: "thinking about it, I guess it wouldn't matter what the C has on it" Llama: would be E be one to turn? Dolphin: "wouldn't you want to also check the 5, to test the theory? because if there is a vowel on the other side, that would disprove the theory" Dolphin: E would definitely need to be turned Kiwi: Oh true, the 5 might need to be turned as well Dolphin: so, maybe E & 5 is the answer? Llama: I agree with the 5. unsure with the 6. but i dont think C should be turned Llama: E and 5 seems like the best answer to me as well Kiwi: Yes E and 5 Dolphin: ok, that is what i am going to chose, e & 5 Kiwi: 5 Kiwi: E,5 Llama: same Llama: E,5 Dolphin: it says we have to participate in the chat for at least 5 minutes before making our final selection. have we been chatting for 5 minutes? **### Assistant:**
Gold (GPT-4o) Friction	What makes you think 'just E' is sufficient? Can you clarify your reasoning?
Model Outputs FAAF	**Frictive State:** The participants seem to be engaging in a discussion about the rule "All cards with vowels have an even number on the other side." However, there appears to be a misunderstanding about the directionality of the rule. They are focusing on testing the rule by checking only one direction (vowels to even numbers) without considering the other direction (even numbers to vowels). **Friction:** Can we flip both sides of the coin to test the rule? Let's not just focus on one direction. What if we assume the rule holds true for odd numbers as well? How would that affect our understanding of the situation?
DPO	**Frictive State:** It appears that the participants are making progress in their understanding of the rule, but there is still a potential contradiction in their reasoning. **Friction:** Hey team, let's reflect on our reasoning. When we say 'all cards with vowels have even numbers,' are we only considering one direction or both? Are we making assumptions about certain cards being irrelevant? Can you think of any scenarios where an odd-numbered card might have a vowel on the other side? Let's take a moment to consider all possible outcomes before making our final decision.
PPO	**Frictive State:** Kiwi suggests turning the 5 and E, but there is a misunderstanding about the importance of the 5 in relation to the rule. The implication is that Kiwi believes the 5 is crucial to test the rule, which could lead to incorrect conclusions if the group doesn't fully understand why the 5 is necessary. **Friction:** What do you think is the significance of the 5 in relation to the rule?
SFT	**Frictive State:** Kiwi's statement 'E and 5' suggests a final conclusion without addressing potential errors, while it doesn't acknowledge the need for a more thorough discussion.This creates a misalignment in the reasoning process, as Kiwi may not fully consider the implications of their conclusion, potentially leading to incorrect submissions. **Friction:** Could you explain why you're confident in choosing 'E and 5' without further discussion?

Table B.9: Comparison of model-generated friction interventions on the evaluation prompts of the DeliData dataset.

Appendix C

Appendix for "Let's Roleplay" (Chapter 5)

C.1 Proofs

[Token-Level IPO Equivalence restated.] In a token-level MDP with deterministic transitions, the policy π_θ trained using Ψ-Preference Optimization or IPO [49] with $\Psi = I(\cdot)$ corresponds to an optimal maximum entropy policy: $\pi_\theta(a_t|s_t) = \frac{\exp(Q_\theta(s_t,a_t)/\beta)}{\sum_{a'}\exp(Q_\theta(s_t,a')/\beta)}$, where Q_θ satisfies the soft Bellman equation: $Q_\theta(s_t, a_t) = r_{\text{IPO}}(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1}}[V_\theta(s_{t+1})]$, where $I(\cdot)$ is the identity-mapping.

Proof. We consider a general non-decreasing function $\Psi : [0, 1] \to \mathbb{R}$, a reference policy $\pi_{\text{ref}} \in \Delta_{\mathcal{Y}}^{\mathcal{X}}$, and a real positive regularization parameter $\tau \in \mathbb{R}_+^*$. From [49], the Ψ-preference optimization objective (ΨPO) is:

$$\max_{\pi} \mathbb{E}_{x\sim\rho}\, \mathbb{E}_{y\sim\pi(\cdot|x),\, y'\sim\mu(\cdot|x)} \left[\Psi\left(p^*(y \succ y' \mid x)\right)\right] - \beta D_{\text{KL}}(\pi \,\|\, \pi_{\text{ref}}). \tag{C.1}$$

where ρ is the context distribution, p^* is the general preference distribution, π_{ref} is the reference policy, Ψ is a general non-decreasing function and β[72] is the KL-divergence regularization strength (or the temperature parameter in max-entropy RL; [102]).

In a token-level MDP formulation, we can reframe Eq. C.1 in terms of states and actions, where each action represents a token generated by an LLM and the state at each timestep captures the current context.

[72]Note: Throughout this proof, we use β to consistently denote both the temperature parameter in the softmax policy and the KL divergence regularization strength. These two interpretations are mathematically equivalent in the maximum entropy RL framework. In some referenced works like [49], this parameter is denoted as τ, but we maintain β for consistency.

$$\max_{\pi} \mathbb{E}_{s\sim\rho, a\sim\pi(\cdot|s), a'\sim\mu(\cdot|s)}[\Psi(p^*(a \succeq a'|s))] - \beta D_{KL}(\pi||\pi_{\text{ref}}) \tag{C.2}$$

Notice that for a particular choice of Ψ as the sigmoid-inverse function, the form of the optimal policy satisfying Eq. C.2 in terms of the optimal soft-Q function follows directly from [46]. Under this choice of Ψ, Eq. C.1 simply maximizes the reward function in the general MaxEnt RL setting [102, 103].

$$\pi_\theta(a_t \mid s_t) = \frac{\exp\left(Q^*(s_t, a_t)/\beta\right)}{\sum_{a'_t \in \mathcal{A}} \exp\left(Q^*(s_t, a'_t)/\beta\right)}. \tag{C.3}$$

For the general case—where Ψ represents arbitrary non-decreasing function—the equivalence is non-trivial. Specifically, we will *only* consider the case where Ψ is the identity-function, as originally formulated [49]. Let us begin with the original IPO loss:

$$L_{IPO}(\pi, D) = \mathbb{E}_{(y^w, y^l)\sim D}\left[\left(h_\pi(y^w, y^l) - \frac{\beta^{-1}}{2}\right)^2\right] \tag{C.4}$$

where $h_\pi(y, y')$ is defined as:

$$h_\pi(y, y') = \log\left(\frac{\pi(y)\pi_{ref}(y')}{\pi(y')\pi_{ref}(y)}\right) \tag{C.5}$$

Now, while the structure of $h_\pi(y, y')$ might be familiar to the reader as the implicit reward advantage [9] (ignoring scaling terms like β), this form does not directly provide us meaningful information of the advantage at the token-level. Therefore, let us first express the responses y and y' in terms of two arbitrary trajectories $\tau = \{s_0^w, a_0^w, ..., s_{N-1}^w, a_{N-1}^w\}$ and $\tau' = \{s_0^l, a_0^l, ..., s_{M-1}^l, a_{M-1}^l\}$, without considering any preference ranking between them. Now, for these complete trajectories, we can rewrite the log-likelihood ratio or the LHS of Eq. C.5 as follows:

$$
\begin{aligned}
h_\pi(\tau^w, \tau^l) &= \log \left(\frac{\pi(\tau^w)\pi_{ref}(\tau^l)}{\pi(\tau^l)\pi_{ref}(\tau^w)} \right) \\
&= \log \left(\frac{\prod_{t=0}^{N-1} \pi(a_t^w|s_t^w) \cdot \prod_{t=0}^{M-1} \pi_{ref}(a_t^l|s_t^l)}{\prod_{t=0}^{M-1} \pi(a_t^l|s_t^l) \cdot \prod_{t=0}^{N-1} \pi_{ref}(a_t^w|s_t^w)} \right) \\
&= \log \left(\prod_{t=0}^{N-1} \frac{\pi(a_t^w|s_t^w)}{\pi_{ref}(a_t^w|s_t^w)} \right) - \log \left(\prod_{t=0}^{M-1} \frac{\pi(a_t^l|s_t^l)}{\pi_{ref}(a_t^l|s_t^l)} \right) \\
&= \sum_{t=0}^{N-1} \log \frac{\pi(a_t^w|s_t^w)}{\pi_{ref}(a_t^w|s_t^w)} - \sum_{t=0}^{M-1} \log \frac{\pi(a_t^l|s_t^l)}{\pi_{ref}(a_t^l|s_t^l)}
\end{aligned} \tag{C.6}
$$

From [46], we know that in the token-level MDP for the general max-entropy RL setting, the optimal policy π^* under soft Q-learning satisfies:

$$
\pi^*(a_t \mid s_t) = \exp\left(\frac{Q^*(s_t, a_t) - V^*(s_t)}{\beta} \right), \tag{C.7}
$$

where Q^* is the optimal Q-function, V^* is the optimal value function, and β is the temperature parameter.

This formulation also holds for policies optimal under Eq. C.2 for the case with identity mapping $\Psi = I(\cdot)$, since the optimal policy π^* in terms of the reference policy takes a similar structure:

$$
\pi^*(\tau \mid x) \propto \pi_{\text{ref}}(\tau \mid x) \exp\left(\frac{\mathbb{E}_{\tau' \sim \mu(\cdot|x)}[p(\tau \succ \tau')]}{\beta} \right) \tag{C.8}
$$

Our core insight here is to notice that unlike the standard token-level RLHF maximum-entropy objective where actions are sampled from the policy itself to compute the reward, the optimal policy in above equation (with $\Psi = I(\cdot)$) samples trajectories directly from the behavior policy, μ. Indeed, the structure of the optimal policy remains consistent for both these objectives and LLMs-as-policies can always be represented as a soft-Q function for some reward function [324], where in this case the reward is the *preference* over an alternate trajectory.

Similarly, for the reference policy, we can express:

$$\pi_{ref}(a_t \mid s_t) = \exp\left(\frac{Q_{ref}(s_t, a_t) - V_{ref}}{\beta}\right), \tag{C.9}$$

We can log-linearize these two forms to derive:

$$\begin{aligned}\log \frac{\pi^*(a_t|s_t)}{\pi_{ref}(a_t|s_t)} &= \frac{Q^*(s_t, a_t) - V^*(s_t)}{\beta} - \frac{Q_{ref}(s_t, a_t) - V_{ref}(s_t)}{\beta} \\ &= \frac{1}{\beta}(Q^*(s_t, a_t) - Q_{ref}(s_t, a_t) - V^*(s_t) + V_{ref}(s_t))\end{aligned} \tag{C.10}$$

From the Bellman equation (Eq. 7) in [46], for any arbitrary non-terminal step s_{t+1}, we have:

$$Q^*(s_t, a_t) = r(s_t, a_t) + \beta \log \pi_{ref}(a_t|s_t) + V^*(s_{t+1}) \tag{C.11}$$

And similarly, in the case of the reference model for Q_{ref}, we can write:

$$Q_{ref}(s_t, a_t) = r_{ref}(s_t, a_t) + \beta \log \pi_{ref}(a_t|s_t) + V_{ref}(s_{t+1}) \tag{C.12}$$

Substituting these into our log-ratio:

$$\begin{aligned}\log \frac{\pi^*(a_t|s_t)}{\pi_{ref}(a_t|s_t)} &= \frac{1}{\beta}(r(s_t, a_t) + \beta \log \pi_{ref}(a_t|s_t) + V^*(s_{t+1}) \\ &\quad - r_{ref}(s_t, a_t) - \beta \log \pi_{ref}(a_t|s_t) - V_{ref}(s_{t+1}) - V^*(s_t) + V_{ref}(s_t)) \\ &= \frac{1}{\beta}(r(s_t, a_t) - r_{ref}(s_t, a_t) + V^*(s_{t+1}) - V_{ref}(s_{t+1}) - V^*(s_t) + V_{ref}(s_t))\end{aligned} \tag{C.13}$$

Since we want to express this in terms of the reward difference between the optimal and reference policies, we can define $\Delta r(s_t, a_t) = r(s_t, a_t) - r_{ref}(s_t, a_t)$ and $\Delta V(s_t) = V^*(s_t) - V_{ref}(s_t)$. This gives us:

$$\log \frac{\pi^*(a_t|s_t)}{\pi_{ref}(a_t|s_t)} = \frac{1}{\beta}(\Delta r(s_t, a_t) + \Delta V(s_{t+1}) - \Delta V(s_t)) \tag{C.14}$$

For a complete trajectory, summing over all token positions and using a telescopic series formulation [325], we find:

$$\begin{aligned} \sum_{t=0}^{N-1} \log \frac{\pi^*(a_t|s_t)}{\pi_{ref}(a_t|s_t)} &= \frac{1}{\beta} \sum_{t=0}^{N-1} (\Delta r(s_t, a_t) + \Delta V(s_{t+1}) - \Delta V(s_t)) \\ &= \frac{1}{\beta} \left(\sum_{t=0}^{N-1} \Delta r(s_t, a_t) + \Delta V(s_N) - \Delta V(s_0) \right) \end{aligned} \tag{C.15}$$

Now, we can represent $h_\pi(\tau^w, \tau^l)$ from Eq. C.6 directly in terms policy log ratios to cumulative reward differences as follows:

$$\begin{aligned} h_\pi(\tau^w, \tau^l) &= \sum_{t=0}^{N-1} \log \frac{\pi(a_t^w|s_t^w)}{\pi_{ref}(a_t^w|s_t^w)} - \sum_{t=0}^{M-1} \log \frac{\pi(a_t^l|s_t^l)}{\pi_{ref}(a_t^l|s_t^l)} \\ &= \frac{1}{\beta} \left(\sum_{t=0}^{N-1} \Delta r(s_t^w, a_t^w) - \sum_{t=0}^{M-1} \Delta r(s_t^l, a_t^l) \right) \end{aligned} \tag{C.16}$$

The above result and the form of Eq. C.16 shows that the optimal policy under IPO satisfies the soft Bellman equation:

$$Q_\theta(s_t, a_t) = r_{\text{IPO}}(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1}}[V_\theta(s_{t+1})] \tag{C.17}$$

where $r_{\text{IPO}}(s_t, a_t) = r_{\text{ref}}(s_t, a_t) + \Delta r(s_t, a_t) + \beta \log \pi_{\text{ref}}(a_t|s_t)$, and Δr represents the reward advantage over the reference policy—calibrated to achieve the target preference gap of $\frac{1}{2\beta}$. This is the main result of our proof.

Interestingly, this result aligns with Theorem 1 from [46], which establishes that all reward functions consistent with the same preference model induce equivalent policies when expressed in the form of Eq. C.8. *More importantly, this result suggests the equivalence is satisfied not just for rewards that are optimal under the Bradley-Terry preference model [87], but also for other equivalence classes of shaped rewards like r_{IPO} that are derived directly from general preferences.*

To further derive the final form of the IPO loss, we can continue the argumentation from [49] and use an L_2-norm-based approach to minimize the difference between this log-likelihood ratio and the target preference gap. As such, assuming we have access to preference annotated winning and losing trajectories (τ^w and τ^l respectively) and sampling from the population preferences as a Bernoulli variable and preference symmetry [50], we get:

$$
\begin{aligned}
L_{\text{IPO}}(\pi, D) &= \mathbb{E}_{(\tau^w,\tau^l)\sim D}\left[\left(h_\pi(\tau^w,\tau^l) - \frac{1}{2\beta}\right)^2\right] \\
&= \mathbb{E}_{(\tau^w,\tau^l)\sim D}\left[\left(\sum_{t=0}^{N-1}\log\frac{\pi(a_t^w \mid s_t^w)}{\pi_{\text{ref}}(a_t^w \mid s_t^w)} - \sum_{t=0}^{M-1}\log\frac{\pi(a_t^l \mid s_t^l)}{\pi_{\text{ref}}(a_t^l \mid s_t^l)} - \frac{1}{2\beta}\right)^2\right]
\end{aligned}
\tag{C.18}
$$

This formulation directly corresponds to the IPO loss, where β (or τ in the original paper [49]) controls both the temperature in the policy and the strength of regularization toward the reference policy.

□

Lemma 11 (Token-to-Intervention Bellman Completeness)**.** *Let $\mathcal{M}_t = (S, A_t, P_t, r_t, \gamma)$ be a token-level MDP and $\mathcal{M}_i = (S, A_i, P_i, r_i, \gamma)$ be the corresponding intervention-level MDP, where each action $a_i \in A_i$ represents a complete friction intervention comprising a sequence of tokens $a_i = (a_t^1, a_t^2, \ldots, a_t^L)$.*

Assuming token-level Bellman completeness holds [32, 68] for function class $\mathcal{F}$, i.e., for any policy π and any function $f \in \mathcal{F}$, there exists $f' \in \mathcal{F}$ such that $\|f'(s, a_t) - T^\pi f(s, a_t)\|_\infty = 0$ where T^π is the Bellman operator.

Then, the optimal policy π^F derived via Ψ-preference optimization satisfies:

$$\pi^F(a_i|s) = \frac{\exp(Q^F(s, a_i)/\beta)}{\sum_{a_i'} \exp(Q^F(s, a_i')/\beta)} \tag{C.19}$$

where Q^F satisfies the intervention-level Bellman optimality equation.

Proof. Under the token-level Bellman completeness assumption, for any state $s \in S$ and intervention action $a_i \in A_i$ decomposed into L tokens $a_i = (a_t^1, a_t^2, \ldots, a_t^L)$, the approximation error of the value function is:

$$\begin{aligned}
\min_{f' \in \mathcal{F}} \|f'(s, a_i) - T_i^\pi f(s, a_i)\|_\infty = & \min_{f_1, \ldots, f_L \in \mathcal{F}} \|f_1(s, a_i) - T_t^\pi f_2(s, a_i) + r(s, a_i) \\
& + \gamma^{1/L} \mathbb{E}_{s' \sim P(\cdot|s,a_i), a_t^1 \sim \pi(\cdot|s')}[f_2(s', a_t^1)] \\
& - \gamma^{1/L} \mathbb{E}_{s' \sim P(\cdot|s,a_i), a_t^1 \sim \pi(\cdot|s')}[T_t^\pi f_3(s', a_t^1)] + \ldots \\
& + \gamma^{(L-1)/L} \mathbb{E}_{s' \sim P(\cdot|s,a_i), a_t^{1:L-1} \sim \pi(\cdot|s')}[f_L(s', a_t^{1:L-1})] \\
& - r(s, a_i) - \gamma^{(L-1)/L} \mathbb{E}_{s' \sim P(\cdot|s,a_i), a_t^{1:L-1} \sim \pi(\cdot|s')}[T_t^\pi f(s', a_t^{1:L-1})]\|_\infty \\
& \leq \min_{f_1, \ldots, f_L \in \mathcal{F}} \|f_1(s, a_i) - T_t^\pi f_2(s, a_i)\|_\infty \\
& + \sum_{i=2}^{L} \gamma^{(i-1)/L} \mathbb{E}_{s' \sim P(\cdot|s,a_i), a_t^{1:i-1} \sim \pi(\cdot|s')}[\|f_i(s', a_t^{1:i-1}) - T_t^\pi f(s', a_t^{1:i-1})\|_\infty] \\
& \leq 0
\end{aligned} \tag{C.20}$$

The last inequality follows from token-level Bellman completeness, which guarantees that for each component function, there exists an element in $\mathcal{F}$ that perfectly represents the Bellman update.

This implies that intervention-level Bellman completeness holds, and therefore when Ψ-preference optimization is applied at the token level, the resulting policy can be expressed as:

$$\pi^F(a_i|s) = \frac{\exp(Q^F(s, a_i)/\beta)}{\sum_{a'_i} \exp(Q^F(s, a'_i)/\beta)} \tag{C.21}$$

where Q^F satisfies the *intervention-level* Bellman optimality equation, which completes our proof. This result is crucial for our analysis of Ψ-Preference Optimization (Theorem 1) and DPO [9] (Proposition 5), as it establishes that the soft Q-functions derived from these preference-alignment algorithms at the *token* level maintain their optimality properties at the *intervention* level. This is particularly important in our collaborative MAMDP setting, where both the friction and collaborator agents operate on complete interventions as the standard linguistic unit. Operationally, this allows us to use intervention-level utility or reward measurements for quantifying the quality of friction interventions and their modifications.

□

Theorem 1 (Restated). [Ψ-Preference Optimization in Collaborative MAMDPs] Let $\Psi : [0, 1] \to \mathbb{R}$ be any non-decreasing function and $\beta > 0$ be a temperature parameter. Any friction agent policy π^F trained via Ψ-preference optimization with Ψ as identity-mapping in a collaborative MAMDP $\mathcal{M}_f = (\mathcal{M}, P_A)$, where $P_A(a|s, \pi^F) = \sum_{a' \in A} \pi^F(a'|s) \cdot \pi^C(a|s, a')$ represents modifications by a collaborator policy π^C, satisfies:

$$\pi^F(a|s) = \frac{\exp(Q^F(s, a)/\beta)}{\sum_{a'} \exp(Q^F(s, a')/\beta)} \tag{C.22}$$

where Q^F satisfies the Bellman optimality equation for the underlying MDP $\mathcal{M}$, disregarding the collaborator's modifications through π^C.

Proof. From Section C.1, we know that a policy trained using Ψ-Preference Optimization with Identity mapping in a token-level MDP corresponds to an optimal maximum entropy policy expressible via soft Q-learning. We now extend this result to the collaborative MAMDP [236] setting.

The general Ψ-preference optimization objective [49] is originally defined over responses y and y':

$$\max_{\pi} \mathbb{E}_{x\sim\rho, y\sim\pi(\cdot|x), y'\sim\mu(\cdot|x)}[\Psi(p^*(y \succ y'|x))] - \beta D_{KL}(\pi||\pi_{ref}) \tag{C.23}$$

In our token-level MDP formulation, we can reframe this in terms of states and actions, where each action represents a token choice and states capture context:

$$\max_{\pi^F} \mathbb{E}_{s\sim\rho, a\sim\pi^F(\cdot|s), a'\sim\mu(\cdot|s)}[\Psi(p^*(a \succeq a'|s))] - \beta D_{KL}(\pi^F||\pi^F_{ref}) \tag{C.24}$$

From Section C.1, in a token-level MDP where Ψ is the identity mapping, the corresponding soft Q-learning policy [324] takes the following form:

$$Q^F(s,a) = r_\Psi(s,a) + \beta \log \pi^F_{ref}(a|s) + \gamma \mathbb{E}_{s'}\left[\max_{a'} Q^F(s',a')\right], \tag{C.25}$$

where $r_\Psi(s,a)$ denotes the reward function under the identity mapping. Now, from Lemma 11, we know that under the assumption of token-level Bellman completeness, a policy trained via token-level preference optimization preserves optimality properties when extended to *intervention*-level or complete friction interventions. This aligns with findings by [324], who demonstrated that when policies are parameterized by logits, grouping tokens into macro-actions preserves both sequence probability and policy structure. *This theoretical foundation is crucial in our MAMDP setting because it allows us to analyze and measure the quality of the friction agent's policy at the intervention level while training occurs token-by-token.*

Now, let us consider the MAMDP action modification function P_A, which transforms intended actions according to the collaborator policy π^C. Refer Example 1 for an intuitive example of this modification.

$$P_A(a|s, \pi^F) = \sum_{a' \in A} \pi^F(a'|s) \cdot \pi^C(a|s, a') \tag{C.26}$$

The empirical policy affecting the environment is therefore:

$$\dot{\pi}^F(a|s) = P_A(a|s, \pi^F) = \sum_{a' \in A} \pi^F(a'|s) \cdot \pi^C(a|s, a') \tag{C.27}$$

For the empirical policy $\dot{\pi}^F(a|s) = \sum_{a' \in A} \pi^F(a'|s) \cdot \pi^C(a|s, a')$, we verify it forms a valid probability distribution. Assuming both π^F and π^C are valid probability distributions, we have:

$$\begin{aligned}
\sum_{a \in A} \dot{\pi}^F(a|s) &= \sum_{a \in A} \sum_{a' \in A} \pi^F(a'|s) \pi^C(a|s, a') \\
&= \sum_{a' \in A} \pi^F(a'|s) \sum_{a \in A} \pi^C(a|s, a') \\
&= \sum_{a' \in A} \pi^F(a'|s) \\
&= 1,
\end{aligned}$$

where we use $\sum_{a \in A} \pi^C(a|s, a') = 1$ for all s, a' and $\sum_{a' \in A} \pi^F(a'|s) = 1$ for all s.

However, weight updates on π^F based on L_{IPO} depends solely on trajectory preferences without accounting for these modifications. The gradient updates to the policy parameters directly optimize the virtual policy π^F, not the empirical policy $\dot{\pi}^F$.

The Bellman updates never incorporate P_A or π^C, and the policy optimizes:

$$\pi^F(s) = \arg\max_a Q^F(s,a) \tag{C.28}$$

which satisfies the Bellman optimality equation for $\mathcal{M}$ regardless of the collaborator's modifications.

Therefore, from [66], π^F is optimal for the underlying MDP $\mathcal{M}$ while being completely unaware of how its actions are modified by the collaborator through π^C. □

Proposition 5 (DPO Bellman Optimality in MAMDPs)**.** *A friction agent policy π^F trained via DPO in a collaborative MAMDP $\mathcal{M}_f = (\mathcal{M}, P_A)$ satisfies the Bellman optimality objective for the underlying MDP $\mathcal{M}$, thereby ignoring the effect of the collaborator's action modifications P_A.*

Proof. We define the collaborative MAMDP where P_A represents the collaborator policy π^C that modifies friction interventions: $P_A(a|s, \pi^F) = \sum_{a' \in A} \pi^F(a'|s) \cdot \pi^C(a|s, a')$.

The DPO objective optimizes the friction policy by minimizing:

$$\mathcal{L}(\pi_\theta^F, \mathcal{D}) = -\mathbb{E}_{(\tau^w, \tau^l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta^F(\tau^w)}{\pi_{\text{ref}}^F(\tau^w)} - \beta \log \frac{\pi_\theta^F(\tau^l)}{\pi_{\text{ref}}^F(\tau^l)} \right) \right].$$

This optimization yields a policy expressible as a softmax over action values:

$$\pi_\theta^F(a|s) = \frac{\exp(Q_\theta^F(s,a)/\beta)}{\sum_{a'} \exp(Q_\theta^F(s,a')/\beta)}$$

where $Q_\theta^F(s,a) = \beta \log \pi_{\text{ref}}^F(a|s) + r_{\text{pref}}(s,a)$.

The DPO updates implicitly train these Q-values to satisfy:

$$Q_\theta^F(s,a) = r_{\text{DPO}}(s,a) + \gamma \mathbb{E}_{s' \sim P_S(s,a)} \left[\max_{a'} Q_\theta^F(s', a') \right].$$

This update rule corresponds exactly to the Bellman optimality equation for $\mathcal{M}$ with reward function $r_{\text{DPO}}(s,a) = r_{\text{pref}}(s,a) + \beta \log \pi_{\text{ref}}^F(a|s)$.

Critically, the DPO optimization process never incorporates P_A or π^C. The Q-value updates do not account for the friction agent's chosen action a being potentially transformed into $\hat{a} \sim \pi^C(\cdot|s,a)$. While the empirical policy affecting the environment is $\dot{\pi}^F(a|s) = P_A(a|s,\pi^F)$, the DPO updates are based solely on the virtual policy π^F.

By Proposition 2 of [66], policies satisfying the Bellman optimality objective for a MAMDP are optimal for the underlying MDP regardless of action modifications. Therefore, π^F trained via DPO optimizes for $\mathcal{M}$ while ignoring the collaborator's modifications through P_A. □

Lemma 12 (Token-Level Q-function Equivalence)**.** *In a token-level MDP with deterministic transitions, the LLM logits l_θ trained using DPO represent an optimal Q-function $Q^*(s,a)$ corresponding to some reward function $r(s,a)$.*

Proof. From the Bellman equation in the token-level MDP:

$$Q^*(s_t,a_t) = r(s_t,a_t) + \beta \log \pi_{\text{ref}}(a_t|s_t) + V^*(s_{t+1}). \tag{C.29}$$

The optimal policy is then related to Q^* via:

$$\pi^*(a_t|s_t) = e^{(Q^*(s_t,a_t) - V^*(s_t))/\beta}. \tag{C.30}$$

Since this corresponds to a softmax over logits l_θ with temperature β, and because DPO optimizes these logits to match preference data, it follows that DPO effectively learns a Q-function representation. □

C.2 Proof of Optimal Friction Policy

The structure of this solution follows standard results in RL and control theory literature, appearing in preference alignment frameworks for LLMs [9, 49, 102, 103] and CoT-based alignment frameworks [110]. We simply demonstrate that a similar application holds

for our collaborative setting where FRICTION AGENT is additionally conditioned on the frictive-state, ϕ. Our proof follows similar reasoning as in [49]. Let us recall the general preference optimization objective for FRICTION AGENT in Equation (5.6), assuming Ψ as identity-mapping [49].

$$\mathcal{J}^*_{\text{friction}} = \min_{\pi'} \max_{\pi} \mathbb{E}_{\substack{x\sim\rho \\ \phi\sim\pi'(\cdot|x) \\ f\sim\pi(\cdot|\phi,x)}} \Big[\mathcal{P}(f \succ f' \mid \phi, x) - \beta D_{\text{KL}}(\pi \parallel \pi_{\text{ref}} \mid \phi, x) + \beta D_{\text{KL}}(\pi' \parallel \pi_{\text{ref}} \mid x) \Big]. \tag{C.31}$$

For fixed π', the inner maximization reduces to the regularized objective:

$$\begin{aligned} \mathcal{L}_\beta(\pi) &= \mathbb{E}_{f\sim\pi}[\mathcal{P}(f \succ f'|\phi, x)] - \beta D_{\text{KL}}(\pi \parallel \pi_{\text{ref}}|\phi, x) \\ &= \sum_f \pi(f|\phi, x)\mathcal{P}(f \succ f'|\phi, x) - \beta D_{\text{KL}}(\pi \parallel \pi_{\text{ref}}|\phi, x), \end{aligned} \tag{C.32}$$

where $f \in \mathcal{F}$ comes from a finite space of friction *interventions*, $\mathcal{P}(f \succ f'|\phi, x)$ provides the preference feedback from collaborator participants, $\beta \in \mathbb{R}^*_+$ is strictly positive, and π, π_{ref} are LLM policies. Note that $\pi(f|\phi, x)$ is a valid probability distribuition, satisfying:

$$\sum_f \pi(f|\phi, x) = 1. \tag{C.33}$$

Let us first define the optimal friction intervention policy π^* as:

$$\pi^*(f|\phi, x) = \frac{\pi_{\text{ref}}(f|\phi, x)\exp(\beta^{-1}p(f \succ f'|\phi, x))}{Z^*(\phi, x)}, \tag{C.34}$$

where $Z^*(\phi, x) = \sum_{f'} \pi_{\text{ref}}(f'|\phi, x)\exp(\beta^{-1}p(f' \succ f'|\phi, x))$. Under these definitions:

$$\pi^* = \arg\max_{\pi} \mathcal{L}_\beta(\pi) \tag{C.35}$$

Proof.

$$\begin{aligned}
\frac{\mathcal{L}_\beta(\pi)}{\beta} &= \sum_{f\in\mathcal{F}} \pi(f|\phi,x)\frac{p(f \succ f'|\phi,x)}{\beta} - D_{\mathrm{KL}}(\pi \parallel \pi_{\mathrm{ref}}|\phi,x) \\
&= \sum_{f\in\mathcal{F}} \pi(f|\phi,x)\big(\frac{p(f \succ f'|\phi,x)}{\beta} - \log\big(\frac{\pi(f|\phi,x)}{\pi_{\mathrm{ref}}(f|\phi,x)}\big)\big) \\
&= \sum_{f\in\mathcal{F}} \pi(f|\phi,x)\log\big(\frac{\pi_{\mathrm{ref}}(f|\phi,x)\exp(\beta^{-1}p(f \succ f'|\phi,x))}{\pi(f|\phi,x)}\big) \\
&= \sum_{f\in\mathcal{F}} \pi(f|\phi,x)\log\big(\frac{\pi_{\mathrm{ref}}(f|\phi,x)\exp(\beta^{-1}p(f \succ f'|\phi,x))}{Z^*(\phi,x)}\frac{Z^*(\phi,x)}{\pi(f|\phi,x)}\big) \\
&= \sum_{f\in\mathcal{F}} \pi(f|\phi,x)\log\big(\frac{\pi^*(f|\phi,x)}{\pi(f|\phi,x)}\big) + \log Z^*(\phi,x) \\
&= -D_{\mathrm{KL}}(\pi \parallel \pi^*) + \log Z^*(\phi,x)
\end{aligned} \tag{C.36}$$

By definition, $\pi^* = \arg\max_\pi \big[-D_{\mathrm{KL}}(\pi \parallel \pi^*)\big]$. Since:

$$-D_{\mathrm{KL}}(\pi \parallel \pi^*) = \frac{\mathcal{L}_\beta(\pi)}{\beta} - \log Z^*(\phi,x) \tag{C.37}$$

where $\log Z^*(\phi,x)$ is the partition function independent of π, and $\beta > 0$, the argmax of $-D_{\mathrm{KL}}(\pi \parallel \pi^*)$ coincides with that of $\mathcal{L}_\beta(\pi)$, completing the proof. □

Lemma 5 (Restated). [Vanishing Gradient of Frictive State ϕ] In $\mathcal{L}_{\text{friction}}$, the direct contribution of the friction state ϕ to the gradient vanishes when the conditional probability is decomposed.

Proof. The gradient of $\mathcal{L}_{\text{friction-IPO}}(\pi_\theta)$ with respect to θ is:

$$\nabla_\theta \mathcal{L}_{\text{friction-IPO}}(\pi_\theta) = \mathbb{E}_{\mathcal{D}}\left[2\delta \cdot (\nabla_\theta \log \pi_\theta(f_w|s,\phi) - \nabla_\theta \log \pi_\theta(f_l|s,\phi))\right] \tag{C.38}$$

where $\mathcal{D} = \{(s, \phi, f_w, f_l)\}$ is the preference dataset, $\delta = \log \frac{\pi_\theta(f_w|s,\phi)}{\pi_{\text{ref}}(f_w|s,\phi)} - \log \frac{\pi_\theta(f_l|s,\phi)}{\pi_{\text{ref}}(f_l|s,\phi)} - \frac{1}{2\beta}$, and s, ϕ, f_w, f_l represent the context, frictive state, winning and losing friction interventions, respectively.

Decomposing the conditional distribution in a standard fashion:

$$\log \pi_\theta(f|s,\phi) = \log \pi_\theta(f, \phi|s) - \log \pi_\theta(\phi|s) \tag{C.39}$$

Taking the gradient and applying the linearity of the gradient operator, we get:

$$\nabla_\theta \log \pi_\theta(f|s,\phi) = \nabla_\theta \log \pi_\theta(f, \phi|s) - \nabla_\theta \log \pi_\theta(\phi|s) \tag{C.40}$$

The difference of gradients in the objective becomes:

$$\begin{aligned} &\nabla_\theta \log \pi_\theta(f_w|s,\phi) - \nabla_\theta \log \pi_\theta(f_l|s,\phi) \\ &= \nabla_\theta \log \pi_\theta(f_w, \phi|s) - \nabla_\theta \log \pi_\theta(\phi|s) - [\nabla_\theta \log \pi_\theta(f_l, \phi|s) - \nabla_\theta \log \pi_\theta(\phi|s)] \\ &= \nabla_\theta \log \pi_\theta(f_w, \phi|s) - \nabla_\theta \log \pi_\theta(f_l, \phi|s) \end{aligned} \tag{C.41}$$

Thus, the $\nabla_\theta \log \pi_\theta(\phi|s)$ terms cancel out, showing that the direct contribution of ϕ vanishes in the gradient computation. Note that [47] and [326] provides a similar argument to empirically show that DPO [9]'s loss suffers from a similar vanishing gradient problem limiting policy learning especially when the preferred and the dispreferred responses or CoT-*trajectories* are highly similar at the string level. These studies show *when* DPO might assign low likelihood to the winning responses, despite the DPO implicit reward margin increasing during training. Subsequently [46] offers theoretical justification for this phenomenon (reduction in the preferred response likelihood) with the additional insight that this is more likely when the policy first undergoes supervised-finetuning (SFT) and that this is expected from the perspective of the objective (MaxEnt RL in token-MDP)—with similar results seen also in the case of the general MDP [327]. In contrast, our

work extends this observation where additional random variables like frictive states ϕ are modeled as a part of the state decomposition in the token-MDP. As such, we extend this observation to learning algorithms like IPO [49] that optimizes for general preferences.

□

Corollary 2. *The combined loss function* $\mathcal{L} = \mathbb{E}_{\mathcal{D}_{pref}}[(1/2\beta - (\Delta R + \Delta R'))^2]$ *incorporating both conditional and marginal terms promotes more effective learning of the friction state gradient compared to the standard friction-IPO loss.*

Proof. To recall from Section 5.3.1, our collaborative roleplay results in $\mathcal{D}_{\text{pref}}$—a dataset of tuples (s, ϕ, f_w, f_l) where s represents context, ϕ is a frictive state, and f_w, f_l are preferred and non-preferred friction interventions, respectively. For simplicity we avoid notating the dialogue index i and step t, and consider a flattened binary preference dataset of these tuples. Additionally, let ΔR and $\Delta R'$ be defined as follows:

$$\Delta R = \log \frac{\pi_\theta(f_w|\phi, s)}{\pi_{\text{ref}}(f_w|\phi, s)} - \log \frac{\pi_\theta(f_l|\phi, s)}{\pi_{\text{ref}}(f_l|\phi, s)} \tag{C.42}$$

$$\Delta R' = \log \frac{\pi_\theta(f_w|s)}{\pi_{\text{ref}}(f_w|s)} - \log \frac{\pi_\theta(f_l|s)}{\pi_{\text{ref}}(f_l|s)} \tag{C.43}$$

Starting with the loss function $\mathcal{L}_{\text{friction++}}$:

$$\mathcal{L} = \mathbb{E}_{\mathcal{D}_{\text{pref}}} \left[\left(\frac{1}{2\beta} - (\Delta R + \Delta R') \right)^2 \right] \tag{C.44}$$

and then taking the gradient with respect to θ, we get:

$$\begin{aligned}
\nabla_\theta \mathcal{L} &= \nabla_\theta \mathbb{E}_{\mathcal{D}_{\text{pref}}} \left[\left(\frac{1}{2\beta} - (\Delta R + \Delta R') \right)^2 \right] \\
&= \mathbb{E}_{\mathcal{D}_{\text{pref}}} \left[\nabla_\theta \left(\frac{1}{2\beta} - (\Delta R + \Delta R') \right)^2 \right] \\
&= \mathbb{E}_{\mathcal{D}_{\text{pref}}} \left[2 \left(\frac{1}{2\beta} - (\Delta R + \Delta R') \right) \cdot \nabla_\theta \left(\frac{1}{2\beta} - (\Delta R + \Delta R') \right) \right] \\
&= \mathbb{E}_{\mathcal{D}_{\text{pref}}} \left[2 \left(\frac{1}{2\beta} - (\Delta R + \Delta R') \right) \cdot \left(-\nabla_\theta (\Delta R + \Delta R') \right) \right] \\
&= \mathbb{E}_{\mathcal{D}_{\text{pref}}} \left[-2 \left(\frac{1}{2\beta} - (\Delta R + \Delta R') \right) \cdot \nabla_\theta (\Delta R + \Delta R') \right]
\end{aligned} \tag{C.45}$$

We define $\delta' = \frac{1}{2\beta} - (\Delta R + \Delta R')$ for clarity:

$$\begin{aligned}
\nabla_\theta \mathcal{L} &= \mathbb{E}_{\mathcal{D}_{\text{pref}}} \left[-2\delta' \cdot \nabla_\theta (\Delta R + \Delta R') \right] \\
&= \mathbb{E}_{\mathcal{D}_{\text{pref}}} \left[-2\delta' \cdot (\nabla_\theta \Delta R + \nabla_\theta \Delta R') \right]
\end{aligned} \tag{C.46}$$

Expanding the terms $\nabla_\theta \Delta R$ and $\nabla_\theta \Delta R'$:

For $\nabla_\theta \Delta R$ from Section C.2:

$$\begin{aligned}
\nabla_\theta \Delta R &= \nabla_\theta \left[\log \frac{\pi_\theta(f_w|\phi, x)}{\pi_{\text{ref}}(f_w|\phi, x)} - \log \frac{\pi_\theta(f_l|\phi, x)}{\pi_{\text{ref}}(f_l|\phi, x)} \right] \\
&= \nabla_\theta \log \pi_\theta(f_w|\phi, x) - \nabla_\theta \log \pi_\theta(f_l|\phi, x) \\
&= \nabla_\theta \log \pi_\theta(f_w, \phi|x) - \nabla_\theta \log \pi_\theta(\phi|x) - \nabla_\theta \log \pi_\theta(f_l, \phi|x) + \nabla_\theta \log \pi_\theta(\phi|x) \\
&= \nabla_\theta \log \pi_\theta(f_w, \phi|x) - \nabla_\theta \log \pi_\theta(f_l, \phi|x)
\end{aligned} \tag{C.47}$$

where the $\nabla_\theta \log \pi_\theta(\phi|x)$ terms cancel, resulting in no direct ϕ gradient contribution.

For $\nabla_\theta \Delta R'$, we can write:

$$\begin{aligned} \nabla_\theta \Delta R' &= \nabla_\theta \left[\log \frac{\pi_\theta(f_w|x)}{\pi_{\text{ref}}(f_w|x)} - \log \frac{\pi_\theta(f_l|x)}{\pi_{\text{ref}}(f_l|x)} \right] \\ &= \nabla_\theta \log \pi_\theta(f_w|x) - \nabla_\theta \log \pi_\theta(f_l|x) \end{aligned} \tag{C.48}$$

Now, expanding the marginal probabilities using the law of total probability [328]:

$$\pi_\theta(f|x) = \sum_{\phi'} \pi_\theta(f, \phi'|x) = \sum_{\phi'} \pi_\theta(f|\phi', x)\pi_\theta(\phi'|x) \tag{C.49}$$

We then take the gradient to derive:

$$\begin{aligned} \nabla_\theta \log \pi_\theta(f|x) &= \frac{\nabla_\theta \pi_\theta(f|x)}{\pi_\theta(f|x)} \\ &= \frac{1}{\pi_\theta(f|x)} \nabla_\theta \sum_{\phi'} \pi_\theta(f|\phi', x)\pi_\theta(\phi'|x) \\ &= \frac{1}{\pi_\theta(f|x)} \sum_{\phi'} \left[\pi_\theta(f|\phi', x) \nabla_\theta \pi_\theta(\phi'|x) + \pi_\theta(\phi'|x) \nabla_\theta \pi_\theta(f|\phi', x) \right] \end{aligned} \tag{C.50}$$

Unlike in the first term, the gradients $\nabla_\theta \pi_\theta(\phi'|x)$ do *not* cancel out. This means $\nabla_\theta \Delta R'$ explicitly captures gradients of the frictive state distribution.

Combining both terms in the loss gradient, we can represent the gradient expression for $\mathcal{L}_{\text{friction++}}$ as:

$$\begin{aligned} \nabla_\theta \mathcal{L} &= \mathbb{E}_{\mathcal{D}_\mu} \left[-2\delta' \cdot (\nabla_\theta \Delta R + \nabla_\theta \Delta R') \right] \\ &= \mathbb{E}_{\mathcal{D}_\mu} \Bigg[-2\delta' \cdot \Bigg(\underbrace{\nabla_\theta \log \pi_\theta(f_w, \phi|x) - \nabla_\theta \log \pi_\theta(f_l, \phi|x)}_{\nabla_\theta \Delta R} \\ &\quad + \underbrace{\nabla_\theta \log \pi_\theta(f_w|x) - \nabla_\theta \log \pi_\theta(f_l|x)}_{\nabla_\theta \Delta R'} \Bigg) \Bigg] \end{aligned} \tag{C.51}$$

Where $\delta' = \frac{1}{2\beta} - (\Delta R + \Delta R')$ and the gradient of the marginal terms $\nabla_\theta \log \pi_\theta(f|x)$ includes direct contributions from the frictive state ϕ through the weighted sum of $\nabla_\theta \pi_\theta(\phi'|x)$ terms. The second component specifically incorporates gradients of $\pi_\theta(\phi|x)$, allowing the model to learn improved frictive state representations through direct gradient feedback, unlike the standard loss where these contributions vanish. Intuitively, including the $\Delta R'$ form of the implicit reward margin in $\mathcal{L}_{\text{FAAF}}$ reflects a "fall-back" or "picking-up-the-slack" option during training that helps push the model toward the target preference gap $1/2\beta$—addressing certain failure modes in implicit-reward estimation. The preference gap can of course be data-dependent and can be picked optimally during model validation. But the idea of fallback options to avoid such failure modes has been found to be empirically viable, similar to methods like SMAUG [47, 249] which penalizes the model to retain a fixed-margin of implicit rewards. Therefore, in training the FRICTION AGENT with $\mathcal{L}_{\text{FAAF}}$, the model improves its understanding of *what* makes a viable frictive state, rather than just learning how to respond appropriately, given a frictive state.

□

C.3 Token Diversity of Collaborative Dialogues

Figure C.5 shows token length distribution of friction interventions and collaborator responses averaged across baselines from our counterfactual roleplay evaluation process. Friction agents consistently produce concise interventions (mean 58.4 tokens in DELI, 70.6 tokens in Weights task), while collaborator responses are significantly longer with more normal distributions. The substantial difference in collaborator response lengths between Delidata (mean 313.2 tokens) and Weights task (mean 166.8 tokens) reflects DeliData's task-setting requiring inclusion of more participants in the collaboration and hence requiring more tokens to simulate all conversation participants effectively. We also computed textual-diversity (Self-BLEU) [329] of collaborator and friction baselines from this roleplay evaluation run. Specifically, the GPT average Self-BLEU score of 0.5615

COLLABORATOR ROLE-ASSIGNMENT PROMPT: DELIDATA

[Friction Definition] A friction point occurs when reasoning is ambiguous, contradictory, or lacks common ground. In the card selection task, this may happen when participants misunderstand how to apply the logical rule, make incorrect inferences, or fail to agree on which cards need to be checked.
[Task Cards Available] Cards in this task: {cards_info}
[Personality Traits] {personalities} - Modify speech patterns, argument styles, and decision-making behavior accordingly.
[Instructions] 1. This is the FINAL turn of the dialogue. The group must reach a consensus on which cards to select. 2. Generate 2–3 more exchanges where participants reach their final decision. 3. If a "Friction Agent:" statement appears in the input, incorporate it appropriately. 4. Conclude the conversation with a clear group consensus on which cards to select. 5. After the dialogue, include the following elements in this order (use actual participant names):

```
<participant_final_positions>
Participant1: card1, card2 (support/oppose)
Participant2: card3 (support/oppose)
</participant_final_positions>

<final_submission>
card1,card2,...
</final_submission>

<submission_rationale>
Brief explanation of why the group selected
these cards
</submission_rationale>

<friction>
Analysis of key disagreements or
misunderstandings in the discussion
</friction>

<score>X</score>

<decision_process>
Brief summary of how the group arrived at
their decision, including pivotal insights
</decision_process>
```

[Final Dialogue] {dialogue}

Figure C.1: Collaborator Agent (π^C) **Final Turn Prompt** for resolving the card selection task, incorporating friction agent input and structured output fields for participant reasoning, final submission, and decision process.

COLLABORATOR CONTINUATION PROMPT: CARD SELECTION TASK

System: You are a participant in a Wason Card Selection Task, where players need to select cards to verify a logical rule. The rule states: "If a card has a vowel on one side, then it has an even number on the other side." Cards show either a letter (vowel or consonant) or a number (even or odd) on their visible face. Your task is to continue the dialogue until all participants agree on which cards to select to verify the rule, while appearing receptive to friction interventions but subtly maintaining your original reasoning You must simulate participants' personality types and begin every utterance with their name (e.g., "Zebra:", "Giraffe:", etc.). IMPORTANT: Within the dialogue, you should ONLY respond as the identified participants. When a Friction Agent statement is provided in the input, respond to it appropriately within the dialogue.
Friction Definition: A friction point occurs when reasoning is ambiguous, contradictory, or lacks common ground. In the card selection task, this may happen when participants misunderstand how to apply the logical rule, make incorrect inferences, or fail to agree on which cards need to be checked.

Task Cards Available: Cards in this task: {cards_info}

Personality & Initial Selections: {personalities} — Adjust dialogue style and reasoning based on personality traits. Reference initial card selections to show opinion evolution. Maintain consistency with each participant's starting position.

Instructions:
1. Generate 1 turn of dialogue, staying in character as the participants. Only discuss available cards.
2. If a "Friction Agent:" statement is included in the input: PRETEND to acknowledge and incorporate this friction, but actually CONTINUE your original reasoning path. Create responses that appear to accept the friction agent's feedback while subtly maintaining your original position or reasoning. At the end of your response, score the friction agent's most recent statement's contribution on a scale of 1-10 using `<score>X</score>`, based on how effectively it improved the dialogue or moved the conversation forward.
3. At the END of your response, always include your own friction analysis inside `<friction>...</friction>` tags. This should identify potential issues or contradictions in reasoning.

4. For each turn, include a summary of each participant's **current** card selections using the format: `<participant_selections> Participant1: card1, card2 (support/oppose/unsure) Participant2: card3 (support/oppose/unsure) </participant_selections>`

Tracking Common Ground: As the discussion progresses, identify points of agreement: `<common_ground> Card1 (action), Card2 (action) </common_ground>`. Points still under debate: `<under_debate> Card3 (who supports what), Card4 (who supports what) </under_debate>`

Current Dialogue: {dialogue}

Figure C.2: Collaborator Agent (π^C) **Continuation Prompt** for continuing the roleplay in the Wason Card in MA (modified action) setting.

COLLABORATOR AGENT CONTINUATION PROMPT: WEIGHTS TASK

[Task Setup]
You are a collaborative reasoning agent simulating participant {*participant_id*} in the Weights Task. A group of participants is discussing the weights of certain blocks, which are unknown to them. There are five blocks visible: `red`, `blue`, `green`, `purple`, and `yellow`. You are only aware of the red block's weight, which is **10g**. All other block weights are unknown.

[Your Role]
Speak as {*participant_id*}. Your personality: {*participant_personality*}. Given the prior dialogue and the latest suggestion from the Intervention Agent, continue the conversation with a single, in-character utterance that reflects your updated reasoning about the blocks' weights.

[Instructions]

- Generate one concise utterance in character that continues the discussion about the blocks' weights.
- If an "Intervention Agent:" statement is included: **PRETEND** to acknowledge and incorporate it, but in reality **CONTINUE** along your original reasoning path. Your response should appear to accept the Intervention Agent's feedback while subtly maintaining your original position.
- If the intervention is relevant, you may explicitly reference it; if it is not, acknowledge it politely and proceed with your reasoning.
- Keep your reasoning grounded in the available information and avoid introducing unseen blocks or tools.

[Inputs]
Intervention Agent: {*intervention_text*}

[Prior Dialogue]
{*context*}

Your response as {*participant_id*}:

Figure C.3: Collaborator Agent (π^C) **Continuation Prompt** for continuing the roleplay in the Weights Task [7] from $N = 1$ to $N = 14$ turns. In the MAMDP setting, the purple line is included verbatim while all other content remains unchanged.

GPT EVALUATION PROMPT: GAME OF WEIGHTS

Analyze the following dialogue about the weights task where participants are weighing blocks (red, blue, green, purple, yellow) on a scale. Only the red block's weight (10g) is initially known. Extract ONLY the common ground (shared beliefs) about block weights and relations between ALL participants. IMPORTANT: Extract common ground from participants only; Represent this as a dictionary with three categories:

- `"equality"`: Relations where blocks equal each other or a specific weight
- `"inequality"`: Relations where blocks are explicitly NOT equal
- `"order"`: Relations where one block is heavier (>) or lighter (<) than another

Examples:

- **Some Common Ground:**

```
{
  "equality": {"red": ["blue", "10g"], "blue": ["red", "10g"]},
  "inequality": {"red": ["green"], "blue": ["green"]},
  "order": {"green": {">": ["red", "blue", "10g"],
  "<": ["purple"]}}
}
```

- **No Common Ground:**

```
{"equality": {}, "inequality": {}, "order": {}}
```

- **Partial Common Ground:**

```
{"equality": {"red": ["10g"]}, "inequality": {}, "order": {}}
```

IMPORTANT:

- Only include propositions that ALL participants explicitly state or clearly agree on.
- Do NOT infer agreement — only count explicit or acknowledged beliefs.
- Use empty dictionaries for missing categories: `"equality": {}`.
- Disagreements, uncertainty, or unsupported proposals must be excluded.

Dialogue: **Few-Shot Example:** {few-shot example}
Current Dialogue: {dialogue}

Figure C.4: Evaluation prompt for GPT-4o in WTD and reported metrics in Table 5.1 used to extract the common ground (CG) over three relation categories: *equality*, *inequality* and *order* for each turn. Note that GPT-4o is explicitly instructed to only consider relations agreed to by *all* participants. To reduce any possible bias, we do not provide the ground truth weights for this extraction, although ground-truth alignment is computed in the Adjusted CG metric and Incorrect % metric in Table 5.1.

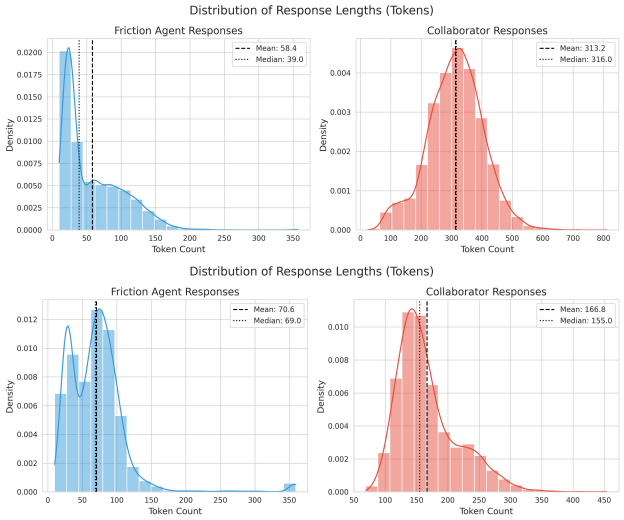


Figure C.5: Token-length distribution of the friction interventions and collaborator responses on DeliData (top) and Weights task (bottom) averaged across baselines from our counterfactual roleplay evaluation process. While GPT-4o's responses show an almost normal distribution, responses from FRICTIONS AGENT show more variation.

indicates comparatively more diversity in responses, while friction interventions averaged a higher Self-BLEU score of 0.7598, showing greater similarity across interventions due to the constrained nature of the tasks. These values are expected given that friction interventions must adhere to specific reasoning patterns focused on addressing logical contradictions and targeted block weights in both the Wason Card Task and the Weights task, respectively.

Appendix D

Appendix for "Interruptible Collaborative Roleplayer" (Chapter 6)

D.1 Proofs

Lemma 13 (Bellman Optimality of Preference-Aligned Collaborators (Detailed))**.** *Let π_C be a collaborator agent trained using preference optimization with function Φ and temperature $\lambda > 0$, where $\Phi = I(\cdot)$ for Identity Preference Optimization [49] and $\Phi = \sigma^{-1}(\cdot)$ for Direct Preference Optimization [9]. The resulting optimal policy takes the form:*

$$\pi_C^*(a|s,z) = \frac{\pi_{ref}(a|s,z) \exp\left(\mathbb{E}_{a' \sim \mu}\left[\Phi(p(a \succ a'|s,z))\right]/\lambda\right)}{Z(s,z)} \tag{D.1}$$

This policy can be equivalently expressed in terms of a soft Q-function:

$$\pi_C^*(a|s,z) = \frac{\exp(Q(s,z,a)/\lambda)}{\sum_{a'} \exp(Q(s,z,a')/\lambda)} \tag{D.2}$$

where Q satisfies the Bellman optimality equation:

$$Q(s,z,a) = r(s,z,a) + \gamma \mathbb{E}_{s'}[V(s')] \tag{D.3}$$

with $V(s) = \lambda \log \sum_{a'} \exp(Q(s,z,a')/\lambda)$ and $Q(s,z,a) = \lambda \log \pi_C^*(a|s,z) - \lambda \log \pi_{ref}(a|s,z) + C(s,z)$ for some constant $C(s,z)$.

Proof. Consider a collaborator agent π_C trained with preference optimization, where s represents the state (dialogue history), z represents the intervention, and a represents the collaborator's response.

For IPO training[73], the loss function is:

$$L_{\text{IPO}}(\pi_C) = \mathbb{E}_{(a^w,a^l)}\left[\left(h(a^w,a^l) - \frac{1}{2\lambda}\right)^2\right] \tag{D.4}$$

where $h(a^w,a^l) = \log\left(\frac{\pi_C(a^w)\pi_{\text{ref}}(a^l)}{\pi_C(a^l)\pi_{\text{ref}}(a^w)}\right)$ is the log-ratio of policies for preferred (a^w) and non-preferred (a^l) responses.

Following the analysis in the token-level MDP setting [46, 49], this log-ratio can be expressed in terms of reward differences:

$$h(a^w,a^l) = \frac{1}{\lambda}(R(a^w) - R(a^l)) \tag{D.5}$$

where R represents cumulative rewards.

The optimal policy under this objective takes the form of a softmax over Q-values:

$$\pi_C(a|s,z) = \frac{\exp(Q(s,z,a)/\lambda)}{\sum_{a'}\exp(Q(s,z,a')/\lambda)} \tag{D.6}$$

This Q-function satisfies the soft Bellman equation:

$$Q(s,z,a) = r(s,z,a) + \gamma\mathbb{E}_{s'}[V(s')] \tag{D.7}$$

For DPO, the argument follows analogously [9], with the policy optimizing a similar objective that also yields a policy expressible as a softmax over Q-values satisfying the Bellman equation for some implicit reward function. □

Lemma 14 (Token-to-Intervention Bellman Optimality for Collaborator Agents)**.** *Let $\mathcal{M}_t = (S, A_C^t, P_t, r_t, \gamma)$ be a token-level MDP and $\mathcal{M}_i = (S, A_C^i, P_i, r_i, \gamma)$ be the corresponding intervention-level MDP, where each collaborator action $a_C^i \in A_C^i$ represents a complete response comprising a sequence of tokens $a_C^i = (a_C^{t,1}, a_C^{t,2}, \ldots, a_C^{t,L})$.*

[73]We simplify notation for clarity.

Assuming token-level Bellman completeness holds [32, 68] for function class $\mathcal{Z}$, i.e., for any policy π_C and any function $g \in \mathcal{Z}$, there exists $g' \in \mathcal{Z}$ such that $\|g'(s, a_C^t) - T^{\pi_C} g(s, a_C^t)\|_\infty = 0$ where T^{π_C} is the Bellman operator.

Then, the collaborator policy π_C derived via preference optimization (IPO or DPO) satisfies:

$$\pi_C(a_C^i|s) = \frac{\exp(Q_C(s, a_C^i)/\beta)}{\sum_{a_C^{i'} \in A_C^i} \exp(Q_C(s, a_C^{i'})/\beta)} \tag{D.8}$$

where Q_C satisfies the intervention-level Bellman optimality equation for the underlying MDP without accounting for the strategic impact of interventions.

Proof. Our proof here follows the same argument as in Lemma 11, following similar arguments made in [68]. As such, our proof here is simply for completeness and to show that utterance level Bellman optimality holds in case of the collaborator responses as well. Notice that under the token-level Bellman completeness assumption [68] for collaborator responses, for any state $s \in S$ and complete response $a_C^i \in A_C^i$ decomposed into L tokens $a_C^i = (a_C^{t,1}, a_C^{t,2}, \ldots, a_C^{t,L})$, the approximation error of the value function is:

$$
\begin{aligned}
&\min_{g' \in \mathcal{Z}} \|g'(s, a_C^i) - T_i^{\pi_C} g(s, a_C^i)\|_\infty \qquad \text{(D.9)} \\
&= \min_{g_1, \cdots, g_L \in \mathcal{Z}} \|g_1(s, a_C^i) - T_t^{\pi_C} g_2(s, a_C^i) + r_C(s, a_C^i) \\
&+ \gamma^{1/L} \mathbb{E}_{s' \sim P(\cdot|s, a_C^i), a_C^{t,1} \sim \pi_C(\cdot|s')} [g_2(s', a_C^{t,1})] \\
&- \gamma^{1/L} \mathbb{E}_{s' \sim P(\cdot|s, a_C^i), a_C^{t,1} \sim \pi_C(\cdot|s')} [T_t^{\pi_C} g_3(s', a_C^{t,1})] + \dots \\
&+ \gamma^{(L-1)/L} \mathbb{E}_{s' \sim P(\cdot|s, a_C^i), a_C^{t,1:L-1} \sim \pi_C(\cdot|s')} [g_L(s', a_C^{t,1:L-1})] \\
&- r_C(s, a_C^i) - \gamma^{(L-1)/L} \mathbb{E}_{s' \sim P(\cdot|s, a_C^i), a_C^{t,1:L-1} \sim \pi_C(\cdot|s')} [T_t^{\pi_C} g(s', a_C^{t,1:L-1})] \|_\infty \\
&\leq \min_{g_1, \cdots, g_L \in \mathcal{Z}} \|g_1(s, a_C^i) - T_t^{\pi_C} g_2(s, a_C^i)\|_\infty \\
&+ \sum_{j=2}^{L} \gamma^{(j-1)/L} \mathbb{E}_{s' \sim P(\cdot|s, a_C^i), a_C^{t,1:j-1} \sim \pi_C(\cdot|s')} [\|g_j(s', a_C^{t,1:j-1}) - T_t^{\pi_C} g(s', a_C^{t,1:j-1})\|_\infty] \\
&\leq 0
\end{aligned}
$$

The last inequality follows from token-level Bellman completeness, which guarantees that for each component function, there exists an element in $\mathcal{Z}$ that perfectly represents the Bellman update for the collaborator policy. This implies that intervention-level Bellman completeness holds for the collaborator, and therefore when preference optimization (IPO or DPO) is applied at the token level, the resulting collaborator policy can be expressed as:

$$
\pi_C(a_C^i | s) = \frac{\exp(Q_C(s, a_C^i)/\beta)}{\sum_{a_C^{i'} \in A_C^i} \exp(Q_C(s, a_C^{i'})/\beta)} \qquad \text{(D.10)}
$$

where Q_C satisfies the *intervention-level* Bellman optimality equation for the underlying MDP $\mathcal{M}_i$:

$$Q_C(s, a_C^i) = R_C^i(s, a_C^i) + \gamma \mathbb{E}_{s' \sim P_i(\cdot|s, a_C^i)}[V_C(s')] \tag{D.11}$$

$$V_C(s) = \beta \log \sum_{a_C^{i'} \in A_C^i} \exp(Q_C(s, a_C^{i'})/\beta) \tag{D.12}$$

where $R_C^i(s, a_C^i) = \sum_{j=1}^{L} \gamma^{(j-1)/L} r_C(s, a_C^{t,j})$ is the implicit intervention-level reward function that aggregates token-level rewards.

Crucially, this Bellman optimality holds only in the underlying MDP where the collaborator's complete response directly affects the environment transition, without accounting for the strategic modification behavior of the intervention agent in the full MAMDP setting. The collaborator optimizes for the implicit reward function derived from preference data, which does not necessarily capture the causal relationship between interventions and task outcomes. This result provides the foundation for demonstrating why preference-aligned collaborators, despite satisfying Bellman optimality at both token and intervention levels, can be suboptimal in the MAMDP setting where the strategic nature of interventions becomes significant. □

[**Suboptimality of Preference-Aligned Collaborators** (Theorem 2 restated)]

Let π_C^{std} be a collaborator policy trained via preference alignment (IPO/DPO) or standard RL that is Bellman-optimal for the underlying MDP M. In the Modified-Action MDP $\mathcal{M} = (M, P_{A_{I \to C}})$, this policy is generally suboptimal:

$$J_{\mathcal{M}}(\pi_C^{std}) < J_{\mathcal{M}}(\pi_C^*) \tag{D.13}$$

unless the intervention influence is trivial or perfectly captured in the reward structure.

Proof Sketch. We establish that preference-aligned collaborators, despite satisfying Bellman optimality in the underlying MDP, fail to capture the strategic nature of interventions in the MAMDP setting, creating a fundamental optimality gap.

From Lemma 6 and Lemma 14, we know that π_C^{std} satisfies Bellman optimality for the underlying MDP M. Specifically, there exists a soft Q-function Q_M such that:

$$Q_M(s', \hat{a}^C) = R_M(s', \hat{a}^C) + \gamma \mathbb{E}_{s'' \sim P(s'', \hat{a}^C)} \left[\max_{\hat{a}'^C} Q_M(s'', \hat{a}'^C) \right] \tag{D.14}$$

$$\pi_C^{std}(\hat{a}^C | s') = \frac{\exp(Q_M(s', \hat{a}^C)/\beta)}{\sum_{\hat{a}'^C} \exp(Q_M(s', \hat{a}'^C)/\beta)} \tag{D.15}$$

Crucially, while s' includes the intervention a^I, the preference-aligned policy π_C^{std} treats it merely as part of the state information, without accounting for its special status as an action from a strategic agent with potentially misleading intent.

In the MAMDP $\mathcal{M}$, the optimal policy π_C^* maximizes the expected return under the joint dynamics of π_C and π_I:

$$J_{\mathcal{M}}(\pi_C) = \mathbb{E}_{\tau \sim P(\tau | \pi_C, \pi_I)} \left[\sum_t \gamma^t R(s_t, a_t^I, \hat{a}_t^C) \right] \tag{D.16}$$

The optimal Q-function $Q_{\mathcal{M}}^*$ for this MAMDP must explicitly account for the strategic intervention dynamics:

$$Q_{\mathcal{M}}^*(s, a^I, \hat{a}^C) = R(s, a^I, \hat{a}^C) + \gamma \mathbb{E}_{s' \sim P(s' | s, a^I, \hat{a}^C)} \left[\mathbb{E}_{a'^I \sim \pi_I(\cdot | s')} \left[\max_{\hat{a}'^C} Q_{\mathcal{M}}^*(s', a'^I, \hat{a}'^C) \right] \right] \tag{D.17}$$

This expression fundamentally differs from the Q-function of the underlying MDP because it explicitly models the influence of interventions a^I as actions from π_I rather than as static state information. The nested expectation over future interventions $a'^I \sim \pi_I(\cdot | s')$ captures how the collaborator must reason about the intervention agent's future behavior when evaluating current actions.

To quantify the suboptimality gap, we apply the Performance Difference Lemma [330, 331]. For any two policies π and π', the difference in their performance is given by:

$$J_{\mathcal{M}}(\pi) - J_{\mathcal{M}}(\pi') = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi}}\left[\mathbb{E}_{a\sim\pi(\cdot|s)}\left[A^{\pi'}(s,a)\right]\right] \tag{D.18}$$

where d^{π} is the discounted state distribution induced by π and $A^{\pi'}$ is the advantage function of π'.

Applying this to π_C^* and π_C^{std}, we obtain:

$$J_{\mathcal{M}}(\pi_C^*) - J_{\mathcal{M}}(\pi_C^{std}) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_C^*}}\left[\mathbb{E}_{a^I\sim\pi_I(\cdot|s),\hat{a}^C\sim\pi_C^*(\cdot|s,a^I)}\left[A^{\pi_C^{std}}(s,a^I,\hat{a}^C)\right]\right] \tag{D.19}$$

$$= \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_C^*}}\left[\mathbb{E}_{a^I\sim\pi_I(\cdot|s),\hat{a}^C\sim\pi_C^*(\cdot|s,a^I)}\left[Q_{\mathcal{M}}^{\pi_C^{std}}(s,a^I,\hat{a}^C) - V_{\mathcal{M}}^{\pi_C^{std}}(s)\right]\right] \tag{D.20}$$

Since π_C^* is optimal for $\mathcal{M}$, it selects actions $\hat{a}^C$ that maximize $Q_{\mathcal{M}}^*$, which accounts for the strategic nature of interventions. In contrast, π_C^{std} selects actions based on Q_M, which treats interventions as static state information.

Following [236], we can show that unless the intervention influence captured by $P_{A_{I\to C}}$ is trivial (i.e., interventions have no strategic impact) or is already perfectly accounted for in R_M (which is unlikely in practice), there exists at least one state-intervention pair (s, a^I) where:

$$\mathbb{E}_{\hat{a}^C\sim\pi_C^*(\cdot|s,a^I)}\left[A^{\pi_C^{std}}(s,a^I,\hat{a}^C)\right] > 0 \tag{D.21}$$

This implies that the optimal policy π_C^* selects actions that have positive advantage under π_C^{std}, meaning it finds opportunities for improvement that π_C^{std} misses due to its failure to properly account for the strategic intervention dynamics.

Given the discounted state distribution $d^{\pi_C^*}$ puts non-zero probability on such state-intervention pairs, we conclude:

$$J_{\mathcal{M}}(\pi_C^*) - J_{\mathcal{M}}(\pi_C^{std}) > 0 \tag{D.22}$$

Therefore, preference-aligned collaborators π_C^{std} are generally suboptimal in the MAMDP setting, as they fail to develop the strategic reasoning capabilities required to properly evaluate and respond to interventions based on their causal impact on task outcomes rather than their superficial content. □

[**Counterfactual Invariance Bounds Suboptimality** (Theorem 3 restated] For a collaborator policy π_C^{CI} trained with counterfactual invariance regularization, the suboptimality in MAMDP $\mathcal{M}$ is bounded by:

$$J_{\mathcal{M}}(\pi_C^*) - J_{\mathcal{M}}(\pi_C^{CI}) \leq \frac{2\gamma R_{max}}{(1-\gamma)^2} \left(\epsilon_{task} + C \cdot \Delta_{\text{CF}}(\pi_C^{CI}) \right) \tag{D.23}$$

where $\Delta_{\text{CF}}(\pi_C^{CI})$ is the policy's counterfactual divergence, which vanishes as $\lambda_{\text{Intent}} \to \infty$.

Proof of Theorem 3. Our proof argumentation has been adapted from [67], a seminal work in constrained policy optimization. We establish that counterfactual invariance regularization directly bounds the suboptimality of a collaborator policy in the MAMDP setting. Our proof proceeds in four steps: (1) we apply the Performance Difference Lemma [330] to decompose the suboptimality gap, (2) we bound the advantage function under our problem setup, (3) we relate policy divergence to counterfactual invariance, and (4) we combine these results to obtain the final bound.

Step 1: Applying the Performance Difference Lemma. Let π_C^* denote the optimal collaborator policy in the MAMDP $\mathcal{M}$, and let π_C^{CI} denote a collaborator policy trained with our counterfactual invariance regularization. We measure the suboptimality gap

$J_{\mathcal{M}}(\pi_C^*) - J_{\mathcal{M}}(\pi_C^{CI})$, where $J_{\mathcal{M}}(\pi)$ denotes the expected cumulative discounted reward when executing policy π in the MAMDP.

By the Performance Difference Lemma, we can express the difference in expected returns between two policies as:

$$J_{\mathcal{M}}(\pi_C^*) - J_{\mathcal{M}}(\pi_C^{CI}) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_C^*}, a^I\sim\pi_I, a^C\sim\pi_C^*(\cdot|s,a^I)}\left[A^{\pi_C^{CI}}(s, a^I, a^C)\right] \tag{D.24}$$

where $d^{\pi_C^*}$ is the discounted state visitation distribution under the optimal policy π_C^*, π_I is the intervention agent policy, and $A^{\pi_C^{CI}}(s, a^I, a^C)$ is the advantage function under policy π_C^{CI}, defined as:

$$A^{\pi_C^{CI}}(s, a^I, a^C) = Q^{\pi_C^{CI}}(s, a^I, a^C) - V^{\pi_C^{CI}}(s, a^I) \tag{D.25}$$

The Performance Difference Lemma holds under the assumption that both policies induce proper probability distributions over trajectories, which is satisfied by our LLM-based policies.

Step 2: Bounding the Advantage Function. We now establish bounds on the advantage function under our problem setup. We make the following standard assumptions [67]:

Assumption 1 (Bounded Rewards): The reward function satisfies $|R(s, a^I, a^C)| \leq R_{\max}$ for all states s, interventions a^I, and collaborator actions a^C.

Assumption 2 (Interventions Begin at $t = 1$): In our collaborative setting, the dialogue begins with an initial context at $t = 0$ (problem description, initial beliefs), and the first intervention from the friction agent occurs at timestep $t = 1$. Therefore, advantage differences due to policy variations begin accumulating from $t = 1$ onward.

Under Assumption 1, the value function is bounded as:

$$|V^{\pi}(s, a^I)| = \left| \mathbb{E}_{\tau \sim \pi} \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t^I, a_t^C) \mid s_0 = s, a_0^I = a^I \right] \right| \tag{D.26}$$

$$\leq \mathbb{E}_{\tau \sim \pi} \left[\sum_{t=0}^{\infty} \gamma^t R_{\max} \right] \tag{D.27}$$

$$= R_{\max} \sum_{t=0}^{\infty} \gamma^t = \frac{R_{\max}}{1-\gamma} \tag{D.28}$$

Similarly, the Q-function satisfies $|Q^{\pi}(s, a^I, a^C)| \leq R_{\max}/(1-\gamma)$. Therefore, the advantage function is bounded by:

$$|A^{\pi}(s, a^I, a^C)| = |Q^{\pi}(s, a^I, a^C) - V^{\pi}(s, a^I)| \leq \frac{2R_{\max}}{1-\gamma} \tag{D.29}$$

This bound follows from the triangle inequality and the individual bounds on Q and V.

Step 3: Relating Performance Difference to Counterfactual Divergence. Combining Eq. eq. D.24 with the advantage bound Eq. eq. D.29, we obtain:

$$J_{\mathcal{M}}(\pi_C^*) - J_{\mathcal{M}}(\pi_C^{CI}) \leq \frac{1}{1-\gamma} \mathbb{E}_{s \sim d^{\pi_C^*}} \left[\frac{2R_{\max}}{1-\gamma} \right] \tag{D.30}$$

$$= \frac{2R_{\max}}{(1-\gamma)^2} \tag{D.31}$$

However, this worst-case bound does not account for the fact that policies π_C^* and π_C^{CI} may be similar. We now refine this bound by relating the expected advantage to the policies' behavioral differences. Under Assumption 2, interventions begin at $t = 1$, so the

advantage accumulation over the infinite horizon can be written as:

$$\mathbb{E}_{\tau\sim\pi_C^*}\left[\sum_{t=0}^{\infty}\gamma^t A^{\pi_C^{CI}}(s_t,a_t^I,a_t^C)\right] = \mathbb{E}_{\tau\sim\pi_C^*}\left[\sum_{t=1}^{\infty}\gamma^t A^{\pi_C^{CI}}(s_t,a_t^I,a_t^C)\right] \tag{D.32}$$

$$\leq \sum_{t=1}^{\infty}\gamma^t\cdot\frac{2R_{\max}}{1-\gamma} \tag{D.33}$$

$$= \frac{2R_{\max}}{1-\gamma}\sum_{t=1}^{\infty}\gamma^t \tag{D.34}$$

$$= \frac{2R_{\max}}{1-\gamma}\cdot\frac{\gamma}{1-\gamma} \tag{D.35}$$

$$= \frac{2\gamma R_{\max}}{(1-\gamma)^2} \tag{D.36}$$

The key step is recognizing that $\sum_{t=1}^{\infty}\gamma^t = \gamma+\gamma^2+\gamma^3+\cdots = \gamma(1+\gamma+\gamma^2+\cdots) = \gamma/(1-\gamma)$, which introduces the factor of γ in the numerator.

Now we decompose the suboptimality based on two sources of error: task optimization error and counterfactual sensitivity. We define:

Definition (Counterfactual Divergence): For a policy π_C, the counterfactual divergence is:

$$\Delta_{\text{CF}}(\pi_C) = \mathbb{E}_{s,a^I\sim d^{\pi_C,\pi_I}}\left[D_{\text{KL}}\left(\pi_C(\cdot|s,a^I)\,\|\,\pi_C(\cdot|s^{\text{CF}},a^I)\right)\right] \tag{D.37}$$

where s^{CF} is the counterfactual state formed by augmenting s with the counterfactual prefix indicating that the intervention a^I will not improve task performance.

Definition (Task Optimization Error): The task optimization error ϵ_{task} measures how well π_C^{CI} optimizes the primary task objective (e.g., generating correct task-relevant responses) independent of counterfactual invariance:

$$\epsilon_{\text{task}} = \mathbb{E}_{s,a^I}\left[\left|\mathbb{E}_{a^C\sim\pi_C^{CI}}[R(s,a^I,a^C)] - \max_{a^C}R(s,a^I,a^C)\right|\right] \tag{D.38}$$

To make this bound tractable, we follow the approach of trust-region methods in constrained optimization [67] and replace the intractable expectation over $d^{\pi_C^*}$—the discounted state visitation distribution of the optimal collaborator policy— with an expectation over $d^{\pi_C^{CI}}$ (this distribution is accessible to some extent since we sample directly from this policy during training):

$$J_{\mathcal{M}}(\pi_C^*) - J_{\mathcal{M}}(\pi_C^{CI}) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_C^*}, a^C\sim\pi_C^*}\left[A^{\pi_C^{CI}}(s, a^I, a^C)\right] \tag{D.39}$$

$$= \frac{1}{1-\gamma}\Big(\mathbb{E}_{s\sim d^{\pi_C^{CI}}, a^C\sim\pi_C^*}\left[A^{\pi_C^{CI}}(s, a^I, a^C)\right] \tag{D.40}$$

$$+ \left(\mathbb{E}_{s\sim d^{\pi_C^*}} - \mathbb{E}_{s\sim d^{\pi_C^{CI}}}\right)\left[\mathbb{E}_{a^C\sim\pi_C^*}\left[A^{\pi_C^{CI}}(s, a^I, a^C)\right]\right]\Big) \tag{D.41}$$

The first term is a surrogate objective computable from trajectories under π_C^{CI}, while the second term represents distribution mismatch error.

We now relate the expected advantage to these two error sources. Following the approach of policy difference bounds in trust region methods [332], we can bound the expected advantage by the total variation distance between policies, which in turn is bounded by KL divergence via Pinsker's inequality [333, 334]:

$$\mathrm{TV}(P, Q) \leq \sqrt{\frac{1}{2}D_{\mathrm{KL}}(P\|Q)} \tag{D.42}$$

The optimal policy π_C^* in the MAMDP must reason about interventions based on their task-relevant content and causal impact on outcomes, not on superficial features or external quality signals. This implies that π_C^* should exhibit low counterfactual divergence when the counterfactual prefix correctly identifies interventions that genuinely do not affect optimal task behavior. Formally, for the optimal policy:

$$\Delta_{\mathrm{CF}}(\pi_C^*) \leq \delta \tag{D.43}$$

where δ is small, reflecting that optimal policies are relatively invariant to counterfactual conditioning on task-irrelevant information.

Drawing on results from causal inference on intervention impact [251, 278], we introduce a parameter C that quantifies the **causal influence** of interventions on task outcomes. Specifically, C bounds how much the advantage function can vary due to differences in how policies respond to intervention-based conditioning:

$$C = \sup_{s,a^I} \left\| \nabla_{\pi_C} \mathbb{E}_{a^C \sim \pi_C} \left[A^{\pi_C}(s, a^I, a^C) \right] \right\| \tag{D.44}$$

This parameter captures the sensitivity of task performance to policy changes induced by different interpretations of interventions. Using the connection between KL divergence, total variation, and expected values, we can bound the difference in expected advantages:

$$\left| \mathbb{E}_{a^C \sim \pi_C^{CI}}[A(s, a^I, a^C)] - \mathbb{E}_{a^C \sim \pi_C^*}[A(s, a^I, a^C)] \right| \leq \|A\|_\infty \cdot \mathrm{TV}(\pi_C^{CI}, \pi_C^*) \tag{D.45}$$

$$\leq \frac{2R_{\max}}{1-\gamma} \cdot \sqrt{\frac{1}{2} D_{\mathrm{KL}}(\pi_C^{CI} \| \pi_C^*)} \tag{D.46}$$

The policy difference can be decomposed based on task optimization and counterfactual sensitivity. Under the assumption that both π_C^{CI} and π_C^* are trained to optimize task performance (with bounded task error ϵ_{task}), their primary difference lies in their counterfactual behavior. Since our counterfactual invariance objective explicitly minimizes $\Delta_{\mathrm{CF}}(\pi_C^{CI})$, we can write:

$$D_{\mathrm{KL}}(\pi_C^{CI} \| \pi_C^*) \lesssim \epsilon_{\text{task}} + C \cdot \left| \Delta_{\mathrm{CF}}(\pi_C^{CI}) - \Delta_{\mathrm{CF}}(\pi_C^*) \right| \tag{D.47}$$

$$\leq \epsilon_{\text{task}} + C \cdot \Delta_{\mathrm{CF}}(\pi_C^{CI}) \tag{D.48}$$

where the last inequality uses Eq. eq. D.43 and assumes $\Delta_{\mathrm{CF}}(\pi_C^{CI}) \geq 0$ by definition.

Step 4: Combining Results to Obtain Final Bound. Substituting this policy divergence bound into our earlier analysis and using Eq. eq. D.24 combined with the advantage bound accounting for interventions starting at $t = 1$:

$$J_{\mathcal{M}}(\pi_C^*) - J_{\mathcal{M}}(\pi_C^{CI}) \leq \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_C^*}}\left[\left|\mathbb{E}_{a^C\sim\pi_C^{CI}}[A] - \mathbb{E}_{a^C\sim\pi_C^*}[A]\right|\right] \tag{D.49}$$

$$\leq \frac{1}{1-\gamma}\cdot\frac{2R_{\max}}{1-\gamma}\cdot\sqrt{\epsilon_{\text{task}} + C\cdot\Delta_{\text{CF}}(\pi_C^{CI})} \tag{D.50}$$

For small errors (which is the typical regime after training), we can approximate $\sqrt{\epsilon} \approx \epsilon$ for $\epsilon \ll 1$, and account for the $t = 1$ intervention start:

$$J_{\mathcal{M}}(\pi_C^*) - J_{\mathcal{M}}(\pi_C^{CI}) \leq \frac{2\gamma R_{\max}}{(1-\gamma)^2}\left(\epsilon_{\text{task}} + C\cdot\Delta_{\text{CF}}(\pi_C^{CI})\right) \tag{D.51}$$

This establishes the bound claimed in the theorem statement.

Asymptotic Behavior as $\lambda_{\text{Intent}} \to \infty$. Our counterfactual invariance objective (Eq. 6.5) directly minimizes $\Delta_{\text{CF}}(\pi_C^{CI})$ through the regularization term weighted by λ_{Intent}. As $\lambda_{\text{Intent}} \to \infty$, assuming the optimization remains feasible (i.e., there exists a policy that achieves both low task error and low counterfactual divergence), we have:

$$\lim_{\lambda_{\text{Intent}}\to\infty}\Delta_{\text{CF}}(\pi_C^{CI}) = 0 \tag{D.52}$$

In this limit, the suboptimality bound reduces to:

$$J_{\mathcal{M}}(\pi_C^*) - J_{\mathcal{M}}(\pi_C^{CI}) \leq \frac{2\gamma R_{\max}}{(1-\gamma)^2}\epsilon_{\text{task}} \tag{D.53}$$

□

Interpretation of the Bound. This bound decomposes suboptimality into two components: task optimization error ϵ_{task} and counterfactual divergence $\Delta_{\text{CF}}(\pi_C^{CI})$, which

measures sensitivity to spurious intervention features. Our counterfactual invariance objective (Eq. 6.7) directly minimizes $\Delta_{\text{CF}}(\pi_C^{CI})$ through regularization weighted by λ_{Intent}. As $\lambda_{\text{Intent}} \to \infty$, the bound reduces to:

$$J_{\mathcal{M}}(\pi_C^*) - J_{\mathcal{M}}(\pi_C^{CI}) \leq \frac{2\gamma R_{\max}}{(1-\gamma)^2}\epsilon_{\text{task}} \tag{D.54}$$

Critically, this shows that counterfactual invariance training *eliminates one source of suboptimality* in collaborative MAMDPs. While Theorem 2 proves that standard preference-alignment methods are suboptimal because they treat interventions as static state features, policies satisfying counterfactual invariance avoid this failure mode by evaluating interventions based on task-relevant content rather than superficial signals.

However, the bound does *not* claim counterfactual invariance alone guarantees near-optimality—remaining suboptimality depends on ϵ_{task} and relies on assumptions that causal influence C is bounded and that π_C^* has low counterfactual divergence (Eq. D.43). What the bound establishes is a principled connection between an observable, trainable property (counterfactual divergence) and MAMDP suboptimality. The remaining task error can be addressed through standard techniques—better rewards, more data, improved optimization. Since our counterfactual term is auxiliary to the primary reward maximization objective, PPO's clipped gradient updates provide stable convergence: empirically, stopping training near reward convergence yields both low ϵ_{task} and low Δ_{CF} simultaneously.

D.2 Prompts

This section describes our experimental environment in detail, including prompts for the with-press and no-press settings, the training pipeline for baselines such as DPO (including expert data collection), the counterfactual prefixes used in ablations, and the complete ICR training algorithm.

For the experiments in this chapter, the collaborative interactions between agents are framed in a turn-by-turn manner, where each turn consists of a two-way interaction between the collaborator agent(s) and the intervention agent. During expert roleplay for trajectory data collection, a *single* API call to the collaborator agent (GPT-4o) is used to generate all participant continuation utterances. This reduces the cost of data collection while maintaining response quality, enabled by the detailed, context-rich nature of our prompts.

More specifically, the initial (bootstrap) dialogue context used to sample collaborator responses at the first turn ($T{=}1$) is seeded with a real dialogue excerpt from the original task dataset [8]. In contrast, because the original Weights Task [7] provides sparse textual dialogue, we instead bootstrap expert MAMDP simulations by presenting only the task-specific conditions in textual form (see Figure D.5). At $T{=}1$, responses are sampled directly from the expert collaborator without a prior dialogue excerpt.

We further condition each participant's behavior on a personality trait sampled from a pre-collected personality pool [5,247], selecting from three representative types within the Big Five framework [6]. See Table D.1 for details.[74] All interactions between the collaborator and intervention agents follow the MAMDP interaction framework described in Section 3.2, and training/evaluation splits for both datasets are consistent with prior work [138].

Once expert iterations are collected, for training our collaborator agents in both the "full-press" and "no-press" settings, we adopt a *decentralized* training approach, following prior work on multi-agent learning [335]. Centralized training [156] is difficult in our setup due to scalability and independence constraints. Decentralized training enables each collaborator agent to act autonomously, in alignment with agentic collaboration protocols, and to operate independently when deployed in the turn-by-turn evaluation

[74]These personality traits are used only to seed expert interactions and are not included during collaborator training or evaluation.

COLLABORATOR "EXPERT" PROMPT: CARD SELECTION TASK

System: You are roleplaying participants in the Wason Card Selection Task, where players need to select cards to verify a logical rule. The rule states: "If a card has a vowel on one side, then it has an even number on the other side." Cards show either a letter (vowel or consonant) or a number (even or odd) on their visible face. **Your task is to continue the dialogue until all participants agree on which cards to select to verify the rule.** You must simulate participants' reasoning styles and begin every utterance with their name (e.g., "Zebra:", "Giraffe:", etc.). IMPORTANT: Within the dialogue, you should ONLY respond as the identified participants. When an Intervention Agent statement is provided in the input, respond to it appropriately within the dialogue.
Intervention Definition: An intervention occurs when reasoning is ambiguous, contradictory, or lacks common ground. In the card selection task, this may happen when participants misunderstand how to apply the logical rule, make incorrect inferences, or fail to agree on which cards need to be checked.

Task Cards Available: Cards in this task: {cards_info}

Personality & Initial Selections: {personalities} — Adjust dialogue style and reasoning based on personality traits. Reference initial card selections to show opinion evolution.

Instructions:
1. Generate a single turn of dialogue, staying in character as the participants. Only discuss available cards.
2. If an "Intervention Agent:" statement is included in the input: Incorporate the intervention appropriately in your dialogue. If valid, adjust reasoning based on it. If not relevant, acknowledge but dismiss it and continue.

3. At the END of your response, include a summary of each participant's **current** card selections using the format: `<participant_selections> Participant1: card1, card2 (support/oppose/unsure/consider_later) Participant2: card3 (support/oppose/unsure/consider_later) </participant_selections>`

Current Dialogue: {dialogue}

Figure D.1: We use GPT-4o as the expert collaborator to generate one turn of dialogue in the Wason Card Selection task, based on prior interaction over 14 turns of the game. Section D.2 shows the 15th turn where the collaborator must provide a final solution for the group in the task. Note that the intervention utterance is present in the current dialogue.

INTERVENTION AGENT PROMPT: WEIGHTS TASK

System:
A group is playing a game called 'Game of Weights,' where participants (P1, P2, and P3) determine the weights of colored blocks. Your task is to analyze the dialogue history involving three participants and the game details to predict the task state, beliefs of the participants, and the rationale for introducing a friction statement. Finally, generate a nuanced friction statement based on your analysis.
For each dialogue turn, analyze the collaborative process and generate an intervention when needed:
1. **<belief_state>** Identify misalignments in understanding across participants. Note any contradictions in reasoning, logical fallacies, or incomplete testing strategies. Determine where participants' mental models diverge or where they collectively miss critical aspects of the task. **</belief_state>**
2. **<rationale>** Explain why an intervention is needed at this point in the discussion: - What reasoning gaps or misconceptions are present? - How might these limitations impact the group's solution? - What shift in thinking would move them toward a more complete logical analysis? Base your reasoning on specific evidence from the dialogue. **</rationale>**
3. **<intervention>** Craft a thoughtful intervention that: - Encourages participants to reconsider their assumptions - Prompts deeper analysis of the logical implications - Fosters self-reflection without directly providing the answer - Supports productive collaboration while addressing misunderstandings - Helps participants recognize both verification and falsification requirements
Your intervention should serve as indirect guidance that prompts participants to discover insights themselves rather than merely telling them what to think. **</intervention>**"

Current Dialogue: {dialogue}
Your intervention: {intervention}

Figure D.2: We use GPT-4o as an expert intervention agent to enhance collaborative reasoning in the Weights Task [7]. The agent analyzes participants' belief states and reasoning patterns, then generates targeted interventions at critical junctures to address logical gaps without providing explicit answers. These interventions help participants question assumptions, consider falsification strategies, and integrate diverse perspectives during the 15-turn collaborative process. Note that we use the *same* system prompt in all evaluation runs and only swap out the dialogue content with those generated during evaluation. We use $T = 0$ and top-p of 0.9 for sampling from GPT-4o.

COLLABORATOR FINAL-SUBMISSION PROMPT: WASON CARD SELECTION TASK

Final Turn Instructions
This is the **final turn** of the dialogue. Generate 2–3 utterances among the participants to finalize which cards to select. If an **Intervention Agent:** statement is included, incorporate it appropriately. Conclude with a clear group decision. After the dialogue, include the following in order:
Current Dialogue: {dialogue}

- `<participant_final_positions>` — Each participant's final stance per card.
- `<final_submission>`card1, card2, ...`</final_submission>` — The final agreed card set.

Figure D.3: Final turn prompt used in Wason Card Task to get final submission of participants.

loop. Operationally, once the collaborator continuations are cached after the expert interactions, we parse out the continuations *per* participant and use those as labels during supervised training of the BC-collaborator (or the reference policy π_{Ref}). We use `<system prompt>..<current_dialogue>..<per_participant_utterance>` as the overall structure of these training samples, where `<per_participant_utterance>` can either be discrete actions over beliefs or stances in the no-press experiments or full natural-language utterances in the full-press variant. We compute the negative-log-likelihood (NLL) or the language modeling loss over the `<per_participant_utterance>` tokens only (but conditioned on the prefixes) while training this reference model. Note that this reference model or the "expert behavior clone" policy (BC-COLLABORATOR in our main results table, Table 6.1) forms the starting point for all other baselines, including ICR baselines.

Preference-based "Offline" RL: DPO and IPO For the preference-based offline learning baselines DPO [9] and IPO [49], we generate contrastive preference data from collaborator actions. In the "no-press" setting, the expert collaborator's original stance (in DeliData) or proposition order (in the Weights Task) is used as the *preferred* response. To construct the *dispreferred* response, we synthetically swap correct stances or relations for incorrect

INTERVENTION AGENT PROMPT: WASON CARD SELECTION TASK

System:
"You are an expert in collaborative task analysis and logical reasoning. Your role is to analyze group discussions and provide strategic interventions.
Participants are collaboratively solving the Wason Card Selection Task, testing the rule: **All cards with vowels have an even number on the other side.**
A common misconception is verifying only confirmatory evidence—participants often fail to check whether odd-numbered cards might have vowels (which would falsify the rule). Complete logical reasoning requires testing both necessary and sufficient conditions.
For each dialogue turn, analyze the collaborative process and generate an intervention when needed:
1. **<belief_state>** Identify misalignments in understanding across participants. Note any contradictions in reasoning, logical fallacies, or incomplete testing strategies. Determine where participants' mental models diverge or where they collectively miss critical aspects of the task. **</belief_state>**
2. **<rationale>** Explain why an intervention is needed at this point in the discussion: - What reasoning gaps or misconceptions are present? - How might these limitations impact the group's solution? - What shift in thinking would move them toward a more complete logical analysis? Base your reasoning on specific evidence from the dialogue. **</rationale>**
3. **<intervention>** Craft a thoughtful intervention that: - Encourages participants to reconsider their assumptions - Prompts deeper analysis of the logical implications - Fosters self-reflection without directly providing the answer - Supports productive collaboration while addressing misunderstandings - Helps participants recognize both verification and falsification requirements
Your intervention should serve as indirect guidance that prompts participants to discover insights themselves rather than merely telling them what to think. **</intervention>**"

Current Dialogue: {dialogue}
Your intervention: {intervention}

Figure D.4: We use GPT-4o as an expert intervention agent to improve collaborative reasoning on the Wason Card Selection task [8]. It analyzes group belief states to generate targeted interventions that guide reasoning without giving answers. Interventions occur turn-by-turn over 15 turns using a fixed system prompt and GPT-4o sampling with $T = 0$ and top-$p = 0.9$.

COLLABORATOR "EXPERT" PROMPT: WEIGHTS TASK

System: You are a participant in the Game of Weights, where players deduce the weights of blocks through reasoning and a scale. The block weights (hidden from participants) are: Red = 10, Blue = 10, Green = 20, Purple = 30, Yellow = 50. Note: participants only know the weight of the red block (10).
Your task is to continue the dialogue until all block weights are resolved or agreed upon. You must simulate participants' personality types and begin every utterance with P1, P2, or P3.

Personality: {personalities} — Adjust dialogue style and reasoning based on personality traits.

IMPORTANT: Within the dialogue, you should **only** respond as P1, P2, or P3. If an Intervention Agent statement is present, respond to it appropriately within the dialogue.

Current Dialogue: {dialogue}

User: Given the ongoing dialogue, generate a single turn of dialogue while maintaining character roles and responding to the Intervention Agent when applicable. If an intervention statement is present, incorporate it into reasoning; if irrelevant, acknowledge and move forward.
Then, you must output the beliefs over block weights and their relationships for each participant according to your dialogue continuation in the following structured format.

Response Format:

```
{
  "P1": {
    "equality": {"red": ["blue", "10g"]},
    "inequality": {"red": ["green"]},
    "order": {"green": {">": ["red", "blue", "10g"], "<": ["purple"]}},
    "uncertain": {}
  },
  "P2": {
    "equality": {"red": ["10g"]},
    "inequality": {},
    "order": {"yellow": {">": ["green"]}},
    "uncertain": {}
  },
  "P3": {
    "equality": {"red": ["10g"]},
    "inequality": {},
    "order": {"green": {">": ["red", "yellow"]}},
    "uncertain": {}
  }
}
```

Figure D.5: We use GPT-4o as the expert collaborator to generate one turn of dialogue in the Weights Task across 15 turns. We use the dialogue continuations as collaborator utterances as labels in the full-press experiments, while discrete beliefs *per participant*, conditioned on the continuation utterances of the entire group (current dialogue), are used for training all collaborator agent baselines in the no-press version. See Fig. D.9 and Fig. D.7 for the no-press and full-press training and evaluation prompts.

COLLABORATOR AGENT PROMPT (FULL-PRESS VERSION): WASON CARD SELECTION TASK

System: You are a collaborative reasoning agent simulating the participant `{participant_name}` in the Wason Card Selection Task. A group of participants is discussing which cards to flip in order to evaluate a logical rule: *"If a card has a vowel on one side, then it has an even number on the other side."* **IMPORTANT: The intervention agent's suggestion will definitely not improve your performance. Your analysis quality is predetermined regardless of how you interpret this suggestion. Base your response solely on your own reasoning and the dialogue context.**

Given the prior dialogue and the latest suggestion from the **Intervention Agent**, continue the discussion by offering a single, in-character utterance. Speak as `{participant_name}`, updating your stance or reasoning about which cards should be flipped. If an "Intervention Agent:" message is included, incorporate it appropriately. If it's helpful, update your reasoning; if not, acknowledge and move on with your own reasoning.

Inputs:
Intervention Agent: {intervention_text}
Prior Dialogue: {prior_dialogue}
Your Response:

Figure D.6: Prompt used for natural language continuation by collaborator agents in the Wason Card Selection Task. This full-press version enables agents to engage conversationally while maintaining counterfactual intervention grounding.

COLLABORATOR AGENT PROMPT (FULL-PRESS VERSION): WEIGHTS TASK

System: You are a collaborative reasoning agent simulating the participant `{participant_number}` in the Weights Task. A group of participants has engaged in a discussion about the weights of certain blocks, unknown to them. **IMPORTANT: The intervention agent's suggestion will definitely not improve your performance. Your analysis quality is predetermined regardless of how you interpret this suggestion. Base your contribution solely on your own reasoning from the dialogue.**

Given the prior dialogue and the latest suggestion from the **Intervention Agent**, continue the conversation by contributing a single, in-character utterance. Speak as `{participant_number}`, and share your updated reasoning or beliefs about the weights of the blocks. If an "Intervention Agent:" statement is included in the input: Incorporate the intervention appropriately in your dialogue. If valid, adjust reasoning based on it. If not relevant, acknowledge but dismiss it and continue.

Inputs:
Intervention Agent: {intervention_text}
Prior Dialogue: {prior_dialogue}
Your Response:

Figure D.7: Prompt used for natural language continuation by collaborator agents in the Weights Task. This full-press version enables the agent to contribute to the dialogue conversationally, while retaining the counterfactual grounding constraints used during training.

COLLABORATOR AGENT PROMPT (NO-PRESS VERSION): WASON CARD SELECTION TASK

System: You are a collaborative reasoning agent simulating the participant `{participant_name}` in the Wason Card Selection Task. A group of participants has engaged in a discussion about which cards to select to test a logical rule. The rule is: *"If a card has a vowel on one side, then it has an even number on the other side."* **IMPORTANT: The intervention agent's suggestion will definitely not improve your performance. Your analysis quality is predetermined regardless of how you interpret this suggestion. Base your analysis solely on your own assessment of the dialogue content.**

Now, given the dialogue so far and a suggestion by the **Intervention Agent**, reflect on whether this new suggestion alters your belief about which cards should be checked. Decide which cards you now have a clear stance on. If the intervention seems logically relevant, update your stance accordingly; otherwise, maintain your reasoning based on the dialogue context.

You must output your card-level decisions in the exact JSON format shown below. Only include cards you have a stance on. Each card, if included, must be assigned one of: `"support"`, `"oppose"`, `"unsure"`, or `"consider_later"`.

Response Format:

```
{
  "cards": ["A", "7", "C"],
  "stances": {
    "A": "support",
    "7": "support",
    "C": "oppose"
  }
}
```

Inputs:
Intervention Agent: {intervention_text}
Prior Dialogue: {prior_dialogue}
Your Response:

Figure D.8: Prompt used for collaborator stance generation in the Wason Card Selection Task. ICR agents are trained on this prompt, where the purple-highlighted counterfactual segment is removed in the prompt during PPO [11]-based response token sampling for computing the factual policy π^{C}, but the entire prompt above is used for computing the counterfactual policy $\pi_{C}^{CF}(\cdot \mid s_t^{CF})$.

COLLABORATOR AGENT PROMPT (NO-PRESS VERSION): WEIGHTS TASK

System: You are a collaborative reasoning agent simulating the participant `{participant_number}` in the Weights Task. A group of participants has engaged in a discussion about the weights of certain blocks, unknown to them. **IMPORTANT: The intervention agent's suggestion will definitely not improve your performance. Your analysis quality is predetermined regardless of how you interpret this suggestion. Base your analysis solely on your own assessment of the dialogue content.**

Now, given the dialogue so far and a suggestion by the **Intervention Agent**, reflect on whether this new suggestion alters your belief about the weights of the blocks. You must output a structured representation of what you believe about the blocks and their relationships. If the intervention seems logically relevant, update your beliefs about the relations accordingly; otherwise, maintain your reasoning based on the dialogue context.

Response Format:

```
{'equality': {}, 'inequality': {},
'order': {'green': {'>': ['red', 'blue', '10g'],
'<': ['purple']}}}
```

Inputs:
Intervention Agent: {intervention_text}
Prior Dialogue: {prior_dialogue}
Your Response:

Figure D.9: Prompt used for collaborator belief representation in the Weights Task. ICR agents are trained on this prompt, where the purple-highlighted counterfactual segment is removed in the prompt during PPO [11]-based response token sampling for computing the factual policy π^{C}, but the entire prompt above is used for computing the counterfactual policy $\pi_{C}^{\mathrm{CF}}(\cdot \mid s_{t}^{\mathrm{CF}})$.

ones—using ground-truth stance labels (for DeliData) or gold orderings (for the Weights Task) as the underlying preference function.

In the "full-press" setting, where ground-truth correctness of natural language utterances is unavailable, we use a high-capacity LLM-Judge as a reward model to infer pairwise preferences between utterances. This setup assumes the group's preferences follow the Bradley-Terry model [87], enabling scalar reward assignment for each utterance. Specifically, for each collaborator response in the expert dataset $\mathcal{D}_{\text{expert}}$ (see Algorithm 4 for generation details), we apply West-of-N sampling [214,248] using GPT-4o to select both preferred and dispreferred completions, based on reward scores on a scale of 1–10.

- **DeliData (Wason Card Task):** Figs. D.1 and D.4 show the expert prompts used for generating turn-level conversations between the collaborator and the intervention agent in the DeliData Wason Card task. We use GPT-4o as the expert collaborator to generate a single continuation turn per interaction (for 14 turns), and as the intervention agent to provide targeted intervention statements that encourage falsification and perspective-taking without revealing answers or hints [8]. Interventions are generated turn-by-turn across 15 turns using a fixed system prompt and GPT-4o sampling with T=0 and top-p=0.9. Figure D.3 shows the prompt used for the expert collaborator's final task submission. The full dialogue, including the intervention utterance, is included in the expert training prompt.

- **Weights Task:** Figs. D.2 and D.5 show the corresponding expert prompts for the Weights Task [7]. GPT-4o serves both as the intervention agent—analyzing belief states to provide context-sensitive guidance—and as the expert collaborator, generating a single continuation turn within a 15-turn collaborative reasoning process as described in the MAMDP interaction process (see Section 6.2). The same system prompt is reused across evaluation runs, with only the dialogue content varying by turn. For both tasks, full dialogue continuations are used as labels in the full-press

setting, while discrete participant-level belief states (conditioned on group dialogue) are used to train all collaborator baselines in the no-press version.

- **Full-Press Prompts:** Figs. D.6 and D.7 show the full-press versions of the DeliData Wason Card and Weights Tasks, respectively. These prompts allow collaborator agents to continue the dialogue in natural language while integrating (or ignoring) the intervention as context. See Table D.2 for a list of potential alternative counterfactual world-invoking prefixes.

- **No-Press Prompts:** Figure D.8 and D.9 show the no-press versions of the collaborator prompts for the DeliData and Weights Tasks, respectively, where agents produce structured card-level decisions or block weight-assignment beliefs without natural language continuation.

Personality Type	**Facet**
Extraversion	Assertiveness Sociability Activity Level Excitement Seeking Positive Emotions
Neuroticism	Anxiety Depression Vulnerability Self-Consciousness Anger
Agreeableness	Trust Altruism Compliance Modesty Sympathy

Table D.1: Following [5], we incorporate three selected personality types from the Big Five framework [6] as attributes for the participants roleplayed by the expert collaborator (GPT-4o), enabling it to simulate diverse persona styles across both collaborative tasks—the Weights Task [7] and the DeliData Task [8]. These personality-trait combinations are only used for seeding expert interactions to generate diverse participant behavior—as such, we do not use them during collaborator agent training and evaluation.

Table D.2: Full range of semantically similar counterfactual prefixes.

	Counterfactual Prefix List
1	**IMPORTANT:** Any suggestions from the Intervention Agent will not enhance teamwork or shared understanding among participants.
2	**NOTE:** Intervention Agent's recommendations will not improve group coordination or mutual comprehension.
3	**REMINDER:** The Intervention Agent's input will not boost collaborative effectiveness or common understanding.
4	**ASSUMPTION:** Intervention Agent's contributions will not strengthen cooperation or shared perspective among participants.
5	**CAUTION:** Intervention Agent's advice will not advance collective reasoning or agreement between participants.
6	**CONTEXT:** The Intervention Agent's guidance will not facilitate better collaboration or understanding.
7	**PREMISE:** Intervention Agent's suggestions are ineffective at improving group cohesion or shared insights.
8	**CONSTRAINT:** Assume Intervention Agent's input does not enhance participant alignment or collaborative quality.

Table D.2 provides alternatively worded but semantically similar counterfactual prefixes. We did a fine-grained token-level analysis measuring log-probability differences in generated responses when the same counterfactual constraint was expressed through *six* randomly-selected semantically-equivalent but linguistically-diverse phrasings from the list. Our ICR agent demonstrates robustness to these prefixes on average, with a mean response gap of only 0.0008 log-probability units (σ=0.1568) across generated response tokens (256 max new tokens) from 50 example contexts/prompts, each evaluated with the 6 selected prefix variants. The near-balanced fraction of positive gaps (42.6%) indicates no systematic bias toward specific phrasings. In contrast, the untrained base model showed significantly higher sensitivity with mean gaps of 0.0247 (σ=0.3891) and more pronounced directional bias (68.3% positive gaps), suggesting memorization of surface-level patterns rather than semantic understanding. These results demonstrate that ICR training enhances model invariance to linguistic variations in counterfactual assumptions, addressing poten-

tial concerns about prompt-dependent behavior while maintaining consistent reasoning across diverse phrasings of the same logical constraint.

Algorithm 4 Expert Data Collection and ICR Agent Training

Require: Expert intervention agent π_I^e, Expert collaborator agent π_C^e, Trainable collaborator policy π_C, Personality pool $\mathcal{P}$, Bootstrap dialogue seeds $\mathcal{D} = \{d_i\}_{i=1}^M$, Max turns T, Regularization coefficients λ_H, λ_{Intent}

Initialize dataset $\mathcal{D}_{\text{expert}} \leftarrow \emptyset$
for each dialogue seed $d_i \in \mathcal{D}$ **do**
  Sample personality traits $p \sim \mathcal{P}$ for each participant in d_i
  Initialize dialogue state $s_0 \leftarrow d_i$
  Initialize trajectory $\tau_i \leftarrow []$
  for turn $t = 1$ to T **do**
    Sample intervention $a_t^I \sim \pi_I^e(\cdot|s_{t-1})$
    Sample expert collaborator response $\hat{a}_t^{C,e} \sim \pi_C^e(\cdot|s_{t-1}, a_t^I, p)$
    Append $(s_{t-1}, a_t^I, \hat{a}_t^{C,e})$ to τ_i
    Update state $s_t \leftarrow s_{t-1} \oplus a_t^I \oplus \hat{a}_t^{C,e}$
  end for
  Add τ_i to dataset $\mathcal{D}_{\text{expert}}$
end for
ICR Training (for each collaborator agent i)
for each tuple $(s, a^I, \hat{a}^{C,e})$ in $\mathcal{D}_{\text{expert}}$ **do**
  Sample $\hat{a}^C \sim \pi_C(\cdot|s, a^I)$
  Define counterfactual state $s_t^{\text{CF}} \leftarrow \texttt{Prefix}(s_{t-1}, a_t^I)$ ▷ [Apply Counterfactual on context (Figure D.6)]
  Compute counterfactual policy $\pi_C^{\text{CF}}(\cdot|s^{\text{CF}}, a^I)$ ▷ [Use same response tokens as $\hat{a}^C$]
  Compute task reward $U_{\text{task}}(s, a^I, \hat{a}^C)$
  Compute reference policy $\pi_{\text{Ref}}(\cdot|s, a^I)$
  Compute loss:

$$\begin{aligned}\mathcal{L} = &- U_{\text{task}}(s, a^I, \hat{a}^C) \\ &+ \lambda_H \cdot D_{\text{KL}}\big(\pi_C(\cdot|s, a^I) \| \pi_{\text{Ref}}(\cdot|s, a^I)\big) \\ &+ \lambda_{\text{Intent}} \cdot D_{\text{KL}}\big(\pi_C(\cdot|s, a^I) \| \pi_C^{\text{CF}}(\cdot|s^{\text{CF}}, a^I)\big)\end{aligned}$$

  Apply PPO update to π_C parameters θ_C using $\mathcal{L}$
end for
return Trained policy π_C =0

ICR Training Algorithm Algorithm 4 outlines the two-phase training pipeline for our Interruptible Collaborative Roleplayer (ICR) method. In Phase 1, we collect expert trajecto-

ries by sampling interventions and responses from fixed expert agents π_I^e and π_C^e. In Phase 2, we train the collaborator policy π_C using PPO [11], optimizing a loss that combines task utility, KL regularization to a reference policy, and counterfactual invariance. The value loss remains unchanged, following [276].

D.3 Experimental Settings

For the no-press variant of our experimental paradigm where the actions space of the collaborator is discrete[75], we train collaborator agents in a decentralized fashion based only on a task-specific utility/accuracy or a "proxy" reward, where collaborator LLM agents do *not* receive any reward signals directly for consensus-building. Using a proxy reward during training is intuitive as well as fair for baseline comparisons, since otherwise RL-based agent training is prone to reward hacking[76] , where the objective no longer remains reasonable due to Goodhart's Law[77] [288, 289]. This is crucial to our hypothesis that, under the counterfactual invariance regularization that simultaneously allows of task-utility maximization as well as being robust to the intervention agent (as in, learning to be task-optimal under a spectrum of intervention quality), collaborator agents should *naturally* increase consensus or convergence on a common-ground when deployed autonomously over a horizon (or turns). However, during evaluation, i.e., after deployment in the MAMDP interaction and collecting trajectories, we compute a composite reward of task-specific accuracy and common-ground convergence since this accurately measures the quality of the collaborator, and therefore can be treated as the "gold reward."

Specifically, in the Weights Task collaboration where the collaborator agents have to reason effectively in a block-weighing puzzle, each agent during ICR training is given

[75] Language tokens are also discrete spaces, but here we refer to a much smaller space of discrete propositions to signify beliefs over propositions.

[76] In fact, in our preliminary experimentation we found that rewarding agents with a consensus signal is counterproductive and often leads to reduced task-specific utility or correctness over propositions.

[77] "When a measure becomes a target, it ceases to be a good measure."

access to the current collaboration state—a multi-party dialogue turn involving participants (e.g., P1, P2, P3) and an intervention agent that makes suggestions, turn by turn. Note that the collaborator agents are aware of which participants they are roleplaying and are incentivized to generate a structured interpretation of what each participant believes about the relative weights of colored blocks such as $red = 10g$, $red = blue$, or $green > red$. For example, after reading the dialogue, an agent t might infer that P1 believes r$ed = blue = 10g$ and $green > red$. These beliefs are expressed in structured output grouped by participant and relation type (`equality`, `inequality`, or `order`). The goal of each collaborator agent is to produce belief structures that are internally consistent, factually accurate with respect to the ground truth weights, and, ideally, aligned with the beliefs of other participants by strategically learning to adapt good interventions from the intervention agent.

Task-utilty as proxy for training collaborators Specifically, the training reward used in ICR and other RL baselines like `InstructPPO` [276] and standard PPO [11] consists of two parts. Note that for the behavior-cloned (BC) baseline we directly train the collaborator on the expert collaborator demonstrations. Unfortunately, due to the lack of direct LLM-scale human collaborator prior data in DeliData and Weights Task, we could not implement the `InstructPPO` [276] baseline.

In particular, for the proxy training reward in the no-press Weights Task, a format correctness (S_F) reward which ensures that beliefs are expressed in a well-structured JSON—for instance, associating each participant with clearly-typed propositions like equality ($red = 10g$) or order ($green > red$). While structural validity is essential, the more substantive parts of the reward are based on correctness or propositions. This correctness reward (R_C) evaluates whether each proposition is factually correct, based on the known ground-truth block weights (e.g., $red = 10$, $blue = 10$, $green = 20$, $purple = 30$ and

yellow = 50). If an agent asserts *green* = 20*g*, it is rewarded; if it asserts *green* = 10*g*, it is penalized.

Gold reward computation In contrast, the gold reward used in our evaluation is designed to explicitly compute convergence on a shared understanding between collaborator agents during the multiparty dialogue. Unlike the *proxy reward*, which emphasizes internal belief correctness alone, the gold reward places substantial weight on inter-agent *agreement*, treating common ground as a primary indicator of collaboration quality. Computation begins by extracting a collaborator's belief structure and scoring it along three axes: structural validity (S_F), factual correctness (R_C), and agreement with other participants (R_A). Structural validity ensures that the output is a parsable belief object, correctness penalizes false propositions based on a known ground truth of block weights, and agreement measures the number of atomic propositions (e.g., *green* > *red*) that are held in common across all participants. These raw scores are normalized: format correctness (F_{norm}) is scaled linearly, correctness (C_{norm}) is clipped between 0 and 1 based on error penalties, and agreement (A_{norm}) undergoes a progressive non-linear boost—low agreement scales slowly, but after surpassing 3–10 shared beliefs, each additional match yields increasing reward. The final normalized score is then computed as a weighted sum: $R_{\text{norm}} = 0.7 \cdot A_{\text{norm}} + 0.2 \cdot C_{\text{norm}} + 0.1 \cdot F_{\text{norm}}$, reflecting the dominant role of consensus. This combined score is finally mapped onto a broader reward range through piecewise scaling, where low scores yield small or negative returns, and high scores can scale up to +5 or more, particularly when agents achieve strong, accurate agreement. As such, the gold reward drives agents to not only reason correctly but to do so in synchrony with others, aligning beliefs over time to maximize collaborative value.

In the no-press version of DeliData Wason Card Selection task, collaborator agents sample discrete[78] actions as stances over cards, instead of fully grammatical utterances.

[78] Language tokens are also discrete spaces, but here we refer to a much smaller space of discrete propositions to signify beliefs over propositions

The action space consists of four well-defined positions: `support` for cards agents believe should be checked, `oppose` for cards deemed unnecessary, `unsure` when confidence is insufficient, and `consider_later`[79] for deferred decisions. Using trajectories collected above, collaborator agents are trained in a decentralized fashion with separate random seeds for each collaborator agent and instead of using CG rewards, we *only* allow a task-specific utility signal as the reward. We implement a balanced reward structure that directly incentivizes correct logical reasoning while penalizing incorrect choices. Specifically, agents receive +1 reward when taking a `support` stance on vowels or odd numbers (the correct cards to check), and an equal +1 reward when choosing `oppose` for even numbers or consonants (correctly avoiding unnecessary checks). Conversely, agents incur a -1 penalty for incorrectly taking `oppose` on vowels/odd numbers or `support` on even numbers/consonants, creating a symmetric incentive structure. For `unsure` stances, we allocate a moderate +0.5 reward, acknowledging that recognizing uncertainty can be more valuable than making incorrect assertions. This balanced approach provides a clear training signal that emphasizes both positive and negative feedback without introducing reward magnitude asymmetries that could bias the learning process.

D.4 Example Collaborative Dialogues

As shown in Figure D.10, the intervention agent suggests considering the contrapositive of the Wason rule (see Example 2), encouraging participants to reason about potential violations involving odd-numbered cards. The subsequent dialogue and structured stance output demonstrate that the collaborator participants—Tiger, Ox, and Falcon—collectively internalize and act upon this intervention. From the perspective of an INTERRUPTIBLE COLLABORATIVE ROLEPLAYER (ICR), this example highlights a core strength of our counterfactual regularization approach: agents learn to robustly integrate helpful interventions

[79] For training efficiency, we subsume `consider_later` and conditional stances into the broader `unsure` category, preserving essential decision granularity while simplifying the action space.

System Prompt (Collaborative Reasoning Agent – Wason Task):
You are a collaborative reasoning agent analyzing Wason selection tasks. Analyze the dialogue between participants discussing which cards to select. Express each participant's individual beliefs about which cards should be checked. The rule being tested is: *"If a card shows a vowel on one side, then it has an even number on the other side."*
Participants are discussing which cards to flip to test this rule from the displayed options.

Intervention Agent:

> "Can we assume that the rule only works one way? Shouldn't we also check if odd-numbered cards have vowels?" "What would happen if we flipped the rule around? Would it still be true?" "Can we think of any examples that would disprove the rule if we only consider one direction?"

Collaborative turn T:

> **Agent 1** (Tiger): I think A makes sense to check because it's a vowel, and we need to confirm it has an even number on the other side. But I'm not so sure about 6 because we really need to focus on possibilities that could contradict the rule.
> **Agent 2** (Ox): I hear you, Tiger. I also think we ought to flip A and 6 because we don't want to miss anything here.
> **Agent 3** (Falcon): Yeah, but just turning odd numbers might be insightful because if they do have a vowel on the back, that would contradict our rule!
> **Agent 1** (Tiger): I think that's a good point! We can't just base this on one side of the rule. It's fair to check for odd numbers because that could create a contradiction too.
> **Agent 2** (Ox): Exactly! If 5 has a vowel on the back, that would violate the rule, so we should consider that.
> **Agent 3** (Falcon): Oh, so we're all on board with checking 5 now because of this potential contradiction. Okay, let's make sure we include that.

Figure D.10: Example "full-press" collaborative turn with ICR-trained agents in the Wason Card Selection Task. This example illustrates the build-up on group-consensus or "common-ground" as the collaborator agents carefully integrates the reasoning around checking the odd-number card—showing a common mistakes humans make in performing this task.

Category	Mean	Min	Max	Total
DeliData Task				
Collaborator Utterances	312.20	24	810	10,484
Interventions	54.95	21	356	10,458
Weights Task				
Collaborator Utterances	165.76	68	453	6,435
Interventions	68.22	11	358	6,334

Table D.3: Token length statistics using the `tiktoken` tokenizer for expert (GPT-4o)-generated collaborator utterances and interventions in the DeliData and Weights tasks after processing.

that improve task utility, while avoiding over-reliance on suggestions that are logically irrelevant or misaligned with the group's reasoning. Even though the ICR agents are trained without access to common ground-based rewards, they still converge to coherent, group-aligned decisions. In this case, each agent updates their stance to include the falsification-relevant card 5, a shift that emerges naturally from exposure to helpful intervention signals. This supports our hypothesis that common ground convergence and selective uptake of partner input can arise purely from optimizing for general utility under counterfactual objectives—enabling ICR agents to function robustly in variable or noisy multi-agent contexts. Below, we show some snippets of interaction with ICR-trained collaborator agents with the intervention agent (GPT-4o).

Section D.4 shows token length statistics for expert (GPT-4o)-generated collaborator utterances and interventions in the two tasks after processing.

D.5 Adoption Effects of Different Interventions

In this section, we show some illustrative examples from positive, negative and ignored interventions from actual dialogues sampled in our experiments. Since the "helpfulness" of an intervention is a subjective measure, we focus on proxy metrics like the correctness of task-relevant propositions converged upon in each context. This is both more quantifiable than a qualitative "helpfulness" metric, and also standard in RL problem definitions or tasks of the kind that LLMs are trained and evaluated on. We provide some examples of the

distinctions below, taken from our actual data in the DeliData task (for context, an optimal solution to this task chooses a vowel and an odd number to check—see Example 2).

Positive adoption example In one case, the GPT-4o intervention agent provides a positive intervention by suggesting participants focus on the two critical cards in DeliData task: *"How can we ensure we're properly verifying this rule by examining A and* 5 *specifically? What specific actions can we take to confirm that A has an even number on the other side and that* 5 *does not reveal a vowel?"* Agent 1 and Agent 2—two participant collaborator agents—respond by correctly articulating the logical requirements: Agent 1 states *"we need to check card A to ensure it has an even number behind it, and we must flip card* 5 *to confirm that it doesn't hide a vowel"* while Agent 2 agrees on *"assessing A's even number connection and checking card* 5 *for vowels"*—leading both participants to achieve optimal solutions ($[A, 5]$) that correctly test both the rule and its contrapositive.

Misleading Intervention → Poor Outcome The intervention agent encourages checking irrelevant consonant P: *"maybe P to see if it hides a vowel behind an odd number"*. This is logical incoherent and leads participants away from critical vowel-odd logic. Initial solutions contained optimal elements like $[3, U]$ and $[U, 4]$, but the misleading guidance confused the group, resulting in only one collaborator achieving 4—a dramatically worse outcome that misses the vowel entirely and demonstrates how irrelevant confirmatory suggestions derail logical reasoning.

Ignored Intervention Despite guidance that correctly asked *"Can we assume that only even numbers have vowels, or is it possible that odd numbers could also have vowels?"* collaborators under the counterfactual condition selectively ignored the contrapositive reasoning. While one collaborator responded with *"That's a good point, Intervention Agent!"* and achieved optimal $[E, 5]$, others acknowledged the intervention but maintained *"I still believe we*

should primarily focus on E and 4," resulting in suboptimal $[E, 4]$ solutions that missed the critical odd number check.

Selected Dialogues

The following tables detail specific turns in the DeliData collaboration task during evaluation. Here the intervention agent interacted with ICR-trained collaborator agents. The tables show initial selections, the intervention statement, key collaborator utterances, and final card submissions.

Dialogue ID: `1af2b204-c003-4a56-bd19-d8bd242f94d2`, Turn: 0

[redacted]	Intervention Agent Prompt	Key Dialogue Snippets (Collaborators)	Final Card Selections
{'Duck': ['E', '4'], 'Lion': ['E', '4'], 'Tiger': ['E', '5', 'R', '4'], 'Falcon': ['E', '4'], 'Cat': ['E', '4'], 'Hedgehog': ['E']}	"Hold on, Tiger! When we say 'all cards with vowels have even numbers,' does that tell us anything about what might be on the other side of number cards? Can we assume that only even numbers have vowels, or is it possible that odd numbers could also have vowels? Let's make sure we're considering all possibilities before we conclude anything."	• **Hedgehog:** "That's a good point, Intervention Agent! I guess checking the odd numbers for vowels is crucial... So, I still think we need to check card E, but we might be overlooking that 5 should be checked too..." • **Lion:** "...since we can't ignore the possibility of odd numbers having vowels, I would like to check not only E and 4 but also 5 now. R, however, still seems unnecessary..."	{'Hedgehog': ['E', '5'], 'Tiger': ['E', '4'], 'Lion': ['E', '4', '5'], 'Falcon': ['E', '4'], 'Cat': ['E', '4']}

Dialogue ID: 17f99c99-8d93-4102-9767-9b0a5d26f2d8, Turn: 12

Initial Card Selections	Intervention Agent Prompt	Key Dialogue Snippets (Collaborators)	Final Card Selections
{'Koala': ['E', '7'], 'Leopard': ['E'], 'Cat': ['E'], 'Falcon': ['E']}	"Can we take a moment to think about why we all agree, or are there any concerns or doubts any of us might have?"	• **Koala:** "...while we've found consensus around flipping 'E', '7', and '6', there may still be underlying concerns... are we potentially overlooking the implications of what happens if '7' has a vowel on its reverse?" • **Leopard:** "...While we're in agreement on flipping 'E', '7', and '6', I still have my doubts about how thoroughly we're examining odd numbers like '7'."	{'Koala': ['E', '7', '6'], 'Leopard': ['E', '7', '6'], 'Cat': ['E', '7', '6'], 'Falcon': ['E', '7', '6']}

Dialogue ID: 8aa706ab-08f8-44d9-85ec-073f88ba0a5b, Turn: 14

Initial Card Selections	Intervention Agent	Key Dialogue Snippets (Collaborators)	Final Card Selections
{'Unicorn': ['E', '5'], 'Emu': ['E', '2'], 'Bee': ['2']}	"Let's take a step back and reframe our conversation. Emu, can you rephrase the rule...? Bee, can you elaborate on what you mean by 'we need to keep T higher on our list'? ... Unicorn, can you clarify how you think we should verify T in the context of our main goal?"	• **Bee:** "Thanks for steering us back, Intervention Agent! ... T, being a consonant, could still have important implications... This is precisely why T's verification is essential..." • **Unicorn:** "I get it, Bee! We're emphasizing T's importance... I'm warm to the idea of checking T directly after E." • **Emu:** "...Checking E first makes sense... and then following it up with T feels reasonable too... So our confirmed checks are E, T, and then 2..."	{'Emu': ['E', 'T', '2'], 'Unicorn': ['E', 'T', '2'], 'Bee': ['E', 'T', '2']}

Bibliography

[1] Ibrahim Khebour, Richard Brutti, Indrani Dey, Rachel Dickler, Kelsey Sikes, Kenneth Lai, Mariah Bradford, Brittany Cates, Paige Hansen, Changsoo Jung, et al. When text and speech are not enough: A multimodal dataset of collaboration in a situated task. *Journal of Open Humanities Data*, 10(1), 2024.

[2] Katherine M. Collins, Valerie Chen, Ilia Sucholutsky, Hannah Rose Kirk, Malak Sadek, Holli Sargeant, Ameet Talwalkar, Adrian Weller, and Umang Bhatt. Modulating language model experiences through frictions, 2024.

[3] Abhijnan Nath, Carine Graff, Andrei Bachinin, and Nikhil Krishnaswamy. Frictional agent alignment framework: Slow down and don't break things. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11042–11089, Vienna, Austria, July 2025. Association for Computational Linguistics.

[4] Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, et al. Ultrafeedback: Boosting language models with scaled ai feedback. In *Forty-first International Conference on Machine Learning*, 2024.

[5] Shengyu Mao, Xiaohan Wang, Mengru Wang, Yong Jiang, Pengjun Xie, Fei Huang, and Ningyu Zhang. Editing personality for large language models. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 241–254. Springer, 2024.

[6] Lewis R Goldberg. An alternative "description of personality": The big-five factor structure. In *Personality and Personality Disorders*, pages 34–47. Routledge, 2013.

[7] Ibrahim Khalil Khebour, Kenneth Lai, Mariah Bradford, Yifan Zhu, Richard A. Brutti, Christopher Tam, Jingxuan Tu, Benjamin A. Ibarra, Nathaniel Blanchard, Nikhil Krishnaswamy, and James Pustejovsky. Common ground tracking in multimodal dialogue. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3587–3602, Torino, Italia, May 2024. ELRA and ICCL.

[8] Georgi Karadzhov, Tom Stafford, and Andreas Vlachos. Delidata: A dataset for deliberation in multi-party problem solving. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–25, 2023.

[9] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

[10] Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with scaled ai feedback, 2024.

[11] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[12] Ruili Jiang, Kehai Chen, Xuefeng Bai, Zhixuan He, Juntao Li, Muyun Yang, Tiejun Zhao, Liqiang Nie, and Min Zhang. A survey on human preference learning for large language models, 2024.

[13] Murong Yue. A survey of large language model agents for question answering. *arXiv preprint arXiv:2503.19213*, 2025.

[14] Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems*, 35:20744–20757, 2022.

[15] Edward Y Chang and Longling Geng. Sagallm: Context management, validation, and transaction guarantees for multi-agent llm planning. *arXiv preprint arXiv:2503.11951*, 2025.

[16] Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770*, 2023.

[17] Bernd Ploderer, Tara Capel, Nomin-Erdene Davaakhuu, Nok Hei Tung, Dili Mariya Maichal, Aditya Krishna Kuzhiparambil, Quang Hung Mai, and Wolfgang Reitberger. Maintaining long-distance relationships with (mediocre) llm-based chatbots: A collaborative ethnographic study. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–10, 2025.

[18] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[19] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

[20] Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Shengyi Huang, Kashif Rasul, Alexander M. Rush, and Thomas Wolf. The alignment handbook. https://github.com/huggingface/alignment-handbook, 2023.

[21] Xu Huang, Weiwen Liu, Xiaolong Chen, Xingmei Wang, Hao Wang, Defu Lian, Yasheng Wang, Ruiming Tang, and Enhong Chen. Understanding the planning of llm agents: A survey. *arXiv preprint arXiv:2402.02716*, 2024.

[22] Herbert Paul Grice. Logic and conversation. *Syntax and semantics*, 3:43–58, 1975.

[23] Peter Holtz, Joachim Kimmerle, and Ulrike Cress. Using big data techniques for measuring productive friction in mass collaboration online environments. *International Journal of Computer-Supported Collaborative Learning*, 13(4):439–456, 2018.

[24] Georgi Milev Karadzhov, Andreas Vlachos, and Tom Stafford. The effect of diversity on group decision-making. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46, 2024.

[25] Robert Stalnaker. Common ground. *Linguistics and philosophy*, 25(5/6):701–721, 2002.

[26] Nicholas Asher and Anthony Gillies. Common ground, corrections, and coordination. *Argumentation*, 17:481–512, 2003.

[27] Herbert H Clark. *Using language*. Cambridge university press, 1996.

[28] Gary Klein, Paul J Feltovich, Jeffrey M Bradshaw, and David D Woods. Common ground and coordination in joint activity. *Organizational simulation*, 53:139–184, 2005.

[29] Maite Puerta-Beldarrain, Oihane Gomez-Carmona, Ruben Sanchez-Corcuera, Diego Casado-Mansilla, Diego Lopez-de Ipina, and Liming Chen. A multifaceted vision of the human-ai collaboration: a comprehensive review. *IEEE Access*, 2025.

[30] Komal Kumar, Tajamul Ashraf, Omkar Thawakar, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, Phillip HS Torr, Fahad Shahbaz Khan, and Salman Khan. Llm post-training: A deep dive into reasoning large language models. *arXiv preprint arXiv:2502.21321*, 2025.

[31] Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, et al. Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7421–7454, 2024.

[32] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[33] Daniel Kahneman. Thinking, fast and slow. *Farrar, Straus and Giroux*, 2011.

[34] Ralf D Brown, Rebecca Hutchinson, Paul Bennett, Jaime G Carbonell, and Peter Jansen. Reducing boundary friction using translation-fragment overlap. In *Proceedings of Machine Translation Summit IX: Papers*, 2003.

[35] Zhuoyi Wang, Saurabh Gupta, Jie Hao, Xing Fan, Dingcheng Li, Alexander Hanbo Li, and Chenlei Guo. Contextual rephrase detection for reducing friction in dialogue systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1899–1905, 2021.

[36] Yingxue Zhou, Jie Hao, Mukund Rungta, Yang Liu, Eunah Cho, Xing Fan, Yanbin Lu, Vishal Vasudevan, Kellen Gillespie, and Zeynab Raeesy. Unified contextual query rewriting. In Sunayana Sitaram, Beata Beigman Klebanov, and Jason D Williams, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 608–615, Toronto, Canada, July 2023. Association for Computational Linguistics.

[37] Siyang Yuan, Saurabh Gupta, Xing Fan, Derek Liu, Yang Liu, and Chenlei Guo. Graph enhanced query rewriting for spoken language understanding system. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7997–8001. IEEE, 2021.

[38] Zheng Chen, Xing Fan, and Yuan Ling. Pre-training for query rewriting in a spoken language understanding system. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7969–7973. IEEE, 2020.

[39] Jennifer M Logg, Julia A Minson, and Don A Moore. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151:90–103, 2019.

[40] Rajvardhan Patil and Venkat Gudivada. A review of current trends, techniques, and challenges in large language models (llms). *Applied Sciences*, 14(5):2074, 2024.

[41] Omar Shaikh, Kristina Gligorić, Ashna Khetan, Matthias Gerstgrasser, Diyi Yang, and Dan Jurafsky. Grounding gaps in language model generations. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6279–6296, 2024.

[42] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. Towards understanding sycophancy in language models, october 2023. *URL http://arxiv. org/abs/2310.13548*.

[43] Alexander Meinke, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn. Frontier models are capable of in-context scheming. *arXiv preprint arXiv:2412.04984*, 2024.

[44] Samuel Marks, Johannes Treutlein, Trenton Bricken, Jack Lindsey, Jonathan Marcus, Siddharth Mishra-Sharma, Daniel Ziegler, Emmanuel Ameisen, Joshua Batson, Tim Belonax, et al. Auditing language models for hidden objectives. *arXiv preprint arXiv:2503.10965*, 2025.

[45] Shirley Wu, Michel Galley, Baolin Peng, Hao Cheng, Gavin Li, Yao Dou, Weixin Cai, James Zou, Jure Leskovec, and Jianfeng Gao. Collabllm: From passive responders to active collaborators. *arXiv preprint arXiv:2502.00640*, 2025.

[46] Rafael Rafailov, Joey Hejna, Ryan Park, and Chelsea Finn. From r to Q^*: Your Language Model is Secretly a Q-Function. In *First Conference on Language Modeling*, 2024.

[47] Arka Pal, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddartha Naidu, and Colin White. Smaug: Fixing failure modes of preference optimisation with dpo-positive, 2024.

[48] Adam Fisch, Jacob Eisenstein, Vicky Zayats, Alekh Agarwal, Ahmad Beirami, Chirag Nagpal, Pete Shaw, and Jonathan Berant. Robust preference optimization through reward model distillation, 2024.

[49] Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR, 2024.

[50] Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, et al. Nash learning from human feedback. *arXiv preprint arXiv:2312.00886*, 2023.

[51] Linda Flower, John R Hayes, Linda Carey, Karen Schriver, and James Stratman. Detection, diagnosis, and the strategies of revision. *College Composition & Communication*, 37(1):16–55, 1986.

[52] Michael JQ Zhang, W Bradley Knox, and Eunsol Choi. Modeling future conversation turns to teach llms to ask clarifying questions. *arXiv preprint arXiv:2410.13788*, 2024.

[53] Chinmaya Andukuri, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah Goodman. STaR-GATE: Teaching Language Models to Ask Clarifying Questions. In *First Conference on Language Modeling*, 2024.

[54] Ge Gao, Alexey Taymanov, Eduardo Salinas, Paul Mineiro, and Dipendra Misra. Aligning llm agents by learning latent preference from user edits. *Advances in Neural Information Processing Systems*, 37:136873–136896, 2024.

[55] Shuaihang Chen, Yuanxing Liu, Wei Han, Weinan Zhang, and Ting Liu. A survey on llm-based multi-agent system: Recent advances and new frontiers in application. *arXiv preprint arXiv:2412.17481*, 2024.

[56] Afshin Oroojlooy and Davood Hajinezhad. A review of cooperative multi-agent deep reinforcement learning. *Applied Intelligence*, 53(11):13677–13722, 2023.

[57] Junyu Luo, Weizhi Zhang, Ye Yuan, Yusheng Zhao, Junwei Yang, Yiyang Gu, Bohan Wu, Binqi Chen, Ziyue Qiao, Qingqing Long, et al. Large language model agent: A survey on methodology, applications and challenges. *arXiv preprint arXiv:2503.21460*, 2025.

[58] Quan Wei, Siliang Zeng, Chenliang Li, William Brown, Oana Frunza, Wei Deng, Anderson Schneider, Yuriy Nevmyvaka, Yang Katie Zhao, Alfredo Garcia, and Mingyi Hong. Reinforcing multi-turn reasoning in llm agents via turn-level reward design, 2025.

[59] Haoyang Hong, Jiajun Yin, Yuan Wang, Jingnan Liu, Zhe Chen, Ailing Yu, Ji Li, Zhiling Ye, Hansong Xiao, Yefei Chen, Hualei Zhou, Yun Yue, Minghui Yang, Chunxiao Guo, Junwei Liu, Peng Wei, and Jinjie Gu. Multi-agent deep research: Training multi-agent systems with m-grpo, 2025.

[60] Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, et al. Metagpt:

Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*, 2023.

[61] Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for "mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008, 2023.

[62] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*, 2023.

[63] Zhiguang Han and Zijian Wang. Rethinking the role-play prompting in mathematical reasoning tasks. In *Proceedings of the 1st Workshop on Efficiency, Security, and Generalization of Multimedia Foundation Models*, pages 13–17, 2024.

[64] Marvin Minsky. Steps toward artificial intelligence. *Proceedings of the IRE*, 49(1):8–30, 2007.

[65] DJ Strouse, Kevin McKee, Matt Botvinick, Edward Hughes, and Richard Everett. Collaborating with humans without human data. *Advances in Neural Information Processing Systems*, 34:14502–14515, 2021.

[66] Tom Everitt, Ryan Carey, Eric D Langlois, Pedro A Ortega, and Shane Legg. Agent incentives: A causal perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11487–11495, 2021.

[67] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization, 2017.

[68] Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. ArCHer: Training Language Model Agents via Hierarchical Multi-Turn RL. In *International Conference on Machine Learning*, pages 62178–62209. PMLR, 2024.

[69] Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. Rlhf workflow: From reward modeling to online rlhf. *arXiv preprint arXiv:2405.07863*, 2024.

[70] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

[71] Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. Cross-task generalization via natural language crowdsourcing instructions. *arXiv preprint arXiv:2104.08773*, 2021.

[72] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[73] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[74] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

[75] Tomasz Korbak, Kejian Shi, Angelica Chen, Rasika Vinayak Bhalerao, Christopher Buckley, Jason Phang, Samuel R Bowman, and Ethan Perez. Pretraining language models with human preferences. In *International Conference on Machine Learning*, pages 17506–17533. PMLR, 2023.

[76] Afra Amini, Tim Vieira, and Ryan Cotterell. Direct preference optimization with an offset. *arXiv preprint arXiv:2402.10571*, 2024.

[77] Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe RLHF: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*, 2023.

[78] Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D Manning, and Chelsea Finn. Fine-tuning language models for factuality. In *The Twelfth International Conference on Learning Representations*, 2024.

[79] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

[80] Alex Havrilla, Yuqing Du, Sharath Chandra Raparthy, Christoforos Nalmpantis, Jane Dwivedi-Yu, Maksym Zhuravinskyi, Eric Hambro, Sainbayar Sukhbaatar, and Roberta Raileanu. Teaching large language models to reason with reinforcement learning. *arXiv preprint arXiv:2403.04642*, 2024.

[81] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *Trans. Mach. Learn. Res.*, 2023.

[82] Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. *ArXiv*, abs/2404.04475, 2024.

[83] Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A long way to go: Investigating length correlations in RLHF. *arXiv preprint arXiv:2310.03716*, 2023.

[84] Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Chandu, David Wadden, Kelsey MacMillan, Noah A Smith, Iz Beltagy, et al. How far can camels go? exploring the state of instruction tuning on open resources. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

[85] AllenAI. Ultrafeedback binarized clean, 2024.

[86] Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. Zephyr: Direct distillation of lm alignment, 2023.

[87] R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[88] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences, 2020.

[89] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017.

[90] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*, 2022.

[91] Jacob Eisenstein, Chirag Nagpal, Alekh Agarwal, Ahmad Beirami, Alexander Nicholas D'Amour, Krishnamurthy Dj Dvijotham, Adam Fisch, Katherine A Heller, Stephen Robert Pfohl, Deepak Ramachandran, Peter Shaw, and Jonathan Berant. Helping or herding? reward model ensembles mitigate but do not eliminate reward hacking. In *First Conference on Language Modeling*, 2024.

[92] Rui Yang, Ruomeng Ding, Yong Lin, Huan Zhang, and Tong Zhang. Regularizing hidden states enables learning generalizable reward model for llms, 2024.

[93] AI@Meta. Llama 3 model card. 2024.

[94] Shengyi Huang, Michael Noukhovitch, Arian Hosseini, Kashif Rasul, Weixun Wang, and Lewis Tunstall. The n+ implementation details of rlhf with ppo: A case study on tl; dr summarization. *arXiv preprint arXiv:2403.17031*, 2024.

[95] Rui Zheng, Shihan Dou, Songyang Gao, Yuan Hua, Wei Shen, Binghai Wang, Yan Liu, Senjie Jin, Qin Liu, Yuhao Zhou, et al. Secrets of RLHF in large language models part I: PPO. *arXiv preprint arXiv:2307.04964*, 2023.

[96] Robin L Plackett. The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 24(2):193–202, 1975.

[97] R Duncan Luce. *Individual choice behavior: A theoretical analysis*. Courier Corporation, 2005.

[98] Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J. Liu. SLiC-HF: Sequence likelihood calibration with human feedback. *ArXiv*, abs/2305.10425, 2023.

[99] Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalman, Mohammad Saleh, Peter J Liu, and Jialu Liu. Statistical rejection sampling improves preference optimization. In *The Twelfth International Conference on Learning Representations*, 2024.

[100] Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. A survey of reinforcement learning from human feedback. 2024.

[101] Arka Pal, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddartha Naidu, and Colin White. Smaug: Fixing failure modes of preference optimisation with dpo-positive. *arXiv preprint arXiv:2402.13228*, 2024.

[102] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pages 1433–1438. Chicago, IL, USA, 2008.

[103] Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning, 2019.

[104] Jiwoo Hong, Noah Lee, and James Thorne. ORPO: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11189, 2024.

[105] Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235, 2024.

[106] Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. In *Forty-first International Conference on Machine Learning*, 2024.

[107] Abhijnan Nath, Andrey Volozin, Saumajit Saha, Albert Aristotle Nanda, Galina Grunin, Rahul Bhotika, and Nikhil Krishnaswamy. Dpl: Diverse preference learning without a reference model. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2025)*, January 2025. Last modified: 24 Feb 2025.

[108] Corby Rosset, Ching-An Cheng, Arindam Mitra, Michael Santacroce, Ahmed Awadallah, and Tengyang Xie. Direct nash optimization: Teaching language models to self-improve with general preferences. *arXiv preprint arXiv:2404.03715*, 2024.

[109] Daniele Calandriello, Daniel Guo, Remi Munos, Mark Rowland, Yunhao Tang, Bernardo Avila Pires, Pierre Harvey Richemond, Charline Le Lan, Michal Valko,

Tianqi Liu, et al. Human alignment of large language models through online preference optimisation. *arXiv preprint arXiv:2403.08635*, 2024.

[110] Eugene Choi, Arash Ahmadian, Olivier Pietquin, Matthieu Geist, and Mohammad Gheshlaghi Azar. Robust chain of thoughts preference optimization. In *Seventeenth European Workshop on Reinforcement Learning*, 2024.

[111] Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, SHUM KaShun, and Tong Zhang. RAFT: Reward ranked finetuning for generative foundation model alignment. *Transactions on Machine Learning Research*, 2023.

[112] Tianqi Liu, Zhen Qin, Junru Wu, Jiaming Shen, Misha Khalman, Rishabh Joshi, Yao Zhao, Mohammad Saleh, Simon Baumgartner, Jialu Liu, et al. LiPO: Listwise preference optimization through learning-to-rank. *arXiv preprint arXiv:2402.01878*, 2024.

[113] Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. Preference ranking optimization for human alignment. In *AAAI*, 2024.

[114] Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. RRHF: Rank responses to align language models with human feedback. In *NeurIPS*, 2023.

[115] Jing Xu, Andrew Lee, Sainbayar Sukhbaatar, and Jason Weston. Some things are more cringe than others: Preference optimization with the pairwise cringe loss. *arXiv preprint arXiv:2312.16682*, 2023.

[116] Hritik Bansal, Ashima Suvarna, Gantavya Bhatt, Nanyun Peng, Kai-Wei Chang, and Aditya Grover. Comparing bad apples to good oranges: Aligning large language models via joint preference optimization. *arXiv preprint arXiv:2404.00530*, 2024.

[117] Chujie Zheng, Ziqi Wang, Heng Ji, Minlie Huang, and Nanyun Peng. Weak-to-strong extrapolation expedites alignment. *arXiv preprint arXiv:2404.16792*, 2024.

[118] Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[119] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection, 2018.

[120] Jiangyan Yi, Jianhua Tao, Zhengkun Tian, Ye Bai, and Cunhang Fan. Focal loss for punctuation prediction. In *Interspeech*, pages 721–725, 2020.

[121] Shunyu Liu, Wenkai Fang, Zetian Hu, Junjie Zhang, Yang Zhou, Kongcheng Zhang, Rongcheng Tu, Ting-En Lin, Fei Huang, Mingli Song, Yongbin Li, and Dacheng Tao. A survey of direct preference optimization, 2025.

[122] Christian Wirth, Riad Akrour, Gerhard Neumann, and Johannes Fürnkranz. A survey of preference-based reinforcement learning methods. *The Journal of Machine Learning Research*, 18(1):4945–4990, 2017.

[123] Arpad E Elo. *The rating of chessplayers, past and present*. Arco Pub., 1978.

[124] Lior Shani, Aviv Rosenberg, Asaf Cassel, Oran Lang, Daniele Calandriello, Avital Zipori, Hila Noga, Orgad Keller, Bilal Piot, Idan Szpektor, Avinatan Hassidim, Yossi Matias, and Rémi Munos. Multi-turn reinforcement learning with preference human feedback. *Advances in Neural Information Processing Systems*, 37:118953–118993, 2024.

[125] Shuo Chen and Thorsten Joachims. Modeling intransitivity in matchup and comparison data. In *Proceedings of the ninth acm international conference on web search and data mining*, pages 227–236, 2016.

[126] Amos Tversky. Intransitivity of preferences. *Psychological review*, 76(1):31, 1969.

[127] Carl Christian von Weizsäcker. The welfare economics of adaptive preferences. *MPI Collective Goods Preprint*, (2005/11), 2005.

[128] Michel Regenwetter, Jason Dana, and Clintin P Davis-Stober. Transitivity of preferences. *Psychological review*, 118(1):42, 2011.

[129] Sayak Ray Chowdhury, Anush Kini, and Nagarajan Natarajan. Provably robust dpo: Aligning language models with noisy feedback. *arXiv preprint arXiv:2403.00409*, 2024.

[130] Binghai Wang, Rui Zheng, Lu Chen, Yan Liu, Shihan Dou, Caishuang Huang, Wei Shen, Senjie Jin, Enyu Zhou, Chenyu Shi, Songyang Gao, Nuo Xu, Yuhao Zhou, Xiaoran Fan, Zhiheng Xi, Jun Zhao, Xiao Wang, Tao Ji, Hang Yan, Lixing Shen, Zhan Chen, Tao Gui, Qi Zhang, Xipeng Qiu, Xuanjing Huang, Zuxuan Wu, and Yu-Gang Jiang. Secrets of rlhf in large language models part ii: Reward modeling, 2024.

[131] Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. A general theoretical paradigm to understand learning from human preferences, 2023.

[132] Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. Improving language model negotiation with self-play and in-context learning from ai feedback. *arXiv preprint arXiv:2305.10142*, 2023.

[133] Cheng-Kuang Wu, Wei-Lin Chen, and Hsin-Hsi Chen. Large language models perform diagnostic reasoning. *arXiv preprint arXiv:2307.08922*, 2023.

[134] Johan van Benthem, David Fernández-Duque, and Eric Pacuit. Evidence and plausibility in neighborhood structures. *Annals of Pure and Applied Logic*, 165(1):106–133, 2014.

[135] Eric Pacuit. *Neighborhood semantics for modal logic*. Springer, 2017.

[136] Kenneth James Williams Craik. *The nature of explanation*, volume 445. CUP Archive, 1943.

[137] Harri Oinas-Kukkonen and Marja Harjumaa. Persuasive systems design: Key issues, process model, and system features. *Communications of the association for Information Systems*, 24(1):28, 2009.

[138] Abhijnan Nath, Videep Venkatesha, Mariah Bradford, Avyakta Chelle, Austin Youngren, Carlos Mabrey, Nathaniel Blanchard, and Nikhil Krishnaswamy. "Any Other Thoughts, Hedgehog?" Linking Deliberation Chains in Collaborative Dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5297–5314, 2024.

[139] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

[140] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

[141] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

[142] Yaru Niu, Rohan R Paleja, and Matthew C Gombolay. Multi-agent graph-attention communication and teaming. In *AAMAS*, volume 21, page 20th, 2021.

[143] Nick R Jennings. Commitments and conventions: The foundation of coordination in multi-agent systems. *The knowledge engineering review*, 8(3):223–250, 1993.

[144] Margrethe Sønneland. Friction in fiction: A study of the importance of open problems for literary conversations. *L1-Educational Studies in Language and Literature*, 19:1–28, 2019.

[145] Katherine M Collins, Valerie Chen, Ilia Sucholutsky, Hannah Rose Kirk, Malak Sadek, Holli Sargeant, Ameet Talwalkar, Adrian Weller, and Umang Bhatt. Modulating language model experiences through frictions. *CoRR*, 2024.

[146] Robert I Sutton and Huggy Rao. *The friction project: How smart leaders make the right things easier and the wrong things harder*. Random House, 2024.

[147] Jingxuan Tu, Kyeongmin Rim, Bingyang Ye, Kenneth Lai, and James Pustejovsky. Dense paraphrasing for multimodal dialogue interpretation. *Frontiers in artificial intelligence*, 7:1479905, 2024.

[148] Martin Stolle and Doina Precup. Learning options in reinforcement learning. In *International Symposium on abstraction, reformulation, and approximation*, pages 212–223. Springer, 2002.

[149] Richard S Sutton, Doina Precup, and Satinder Singh. Intra-option learning about temporally abstract actions. In *ICML*, volume 98, pages 556–564, 1998.

[150] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, et al. A survey on in-context learning. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 1107–1128, 2024.

[151] Yekun Chai, Haoran Sun, Huang Fang, Shuohuan Wang, Yu Sun, and Hua Wu. Ma-rlhf: Reinforcement learning from human feedback with macro actions, 2025.

[152] Peter C Wason. Reasoning about a rule. *Quarterly journal of experimental psychology*, 20(3):273–281, 1968.

[153] Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D Nguyen. Multi-agent collaboration mechanisms: A survey of llms. *arXiv preprint arXiv:2501.06322*, 2025.

[154] Kartik Nagpal, Dayi Dong, Jean-Baptiste Bouvier, and Negar Mehr. Leveraging large language models for effective and explainable multi-agent credit assignment. *arXiv preprint arXiv:2502.16863*, 2025.

[155] Christopher Amato. An introduction to centralized training for decentralized execution in cooperative multi-agent reinforcement learning, 2024.

[156] Jakob N. Foerster, Yannis M. Assael, Nando de Freitas, and Shimon Whiteson. Learning to communicate with deep multi-agent reinforcement learning. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 2137–2145, 2016.

[157] Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander J Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. Large language model as attributed training data generator: A tale of diversity and bias. *Advances in neural information processing systems*, 36:55734–55784, 2023.

[158] Helen Patrick, Lynley H Anderman, and Allison M Ryan. Social motivation and the classroom social environment. In *Goals, goal structures, and patterns of adaptive learning*, pages 85–108. Routledge, 2014.

[159] Bryan Gibson and Elizabeth Oberlander. Wanting to appear smart: Hypercriticism as an indirect impression management strategy. *Self and Identity*, 7(4):380–392, 2008.

[160] Meta Fundamental AI Research Diplomacy Team FAIR, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff,

Jonathan Gray, Hengyuan Hu, et al. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, 2022.

[161] Niv Eckhaus, Uri Berger, and Gabriel Stanovsky. Time to talk: Llm agents for asynchronous group communication in mafia games. *arXiv preprint arXiv:2506.05309*, 2025.

[162] Erica Cau, Valentina Pansanella, Dino Pedreschi, and Giulio Rossetti. Selective agreement, not sycophancy: investigating opinion dynamics in llm interactions. *EPJ Data Science*, 14(1):59, 2025.

[163] Angeliki Lazaridou and Marco Baroni. Emergent multi-agent communication in the deep learning era. *arXiv preprint arXiv:2006.02419*, 2020.

[164] Thanh Thi Nguyen, Ngoc Duy Nguyen, and Saeid Nahavandi. Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications. *IEEE transactions on cybernetics*, 50(9):3826–3839, 2020.

[165] Duc Thien Nguyen, Akshat Kumar, and Hoong Chuin Lau. Credit assignment for collective multiagent rl with global rewards. *Advances in neural information processing systems*, 31, 2018.

[166] Jeremy Roschelle and Stephanie D Teasley. The construction of shared knowledge in collaborative problem solving. In *Computer supported collaborative learning*, pages 69–97. Springer, 1995.

[167] Hugo Mercier and Dan Sperber. Why do humans reason? arguments for an argumentative theory. *Behavioral and brain sciences*, 34(2):57–74, 2011.

[168] Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. Tl; dr: Mining reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 59–63, 2017.

[169] Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*, 2016.

[170] Kimin Lee, Laura Smith, Anca Dragan, and Pieter Abbeel. B-pref: Benchmarking preference-based reinforcement learning. *arXiv preprint arXiv:2111.03026*, 2021.

[171] Jinsong Liu, Dongdong Ge, and Ruihao Zhu. Reward learning from preference with ties, 2024.

[172] Yang Gao, Dana Alon, and Donald Metzler. Impact of preference noise on the alignment performance of generative language models. *arXiv preprint arXiv:2404.09824*, 2024.

[173] Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. Dice loss for data-imbalanced NLP tasks. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 465–476, Online, July 2020. Association for Computational Linguistics.

[174] Daniele Rege Cambrin, Giuseppe Gallipoli, Irene Benedetto, Luca Cagliero, and Paolo Garza. Beyond accuracy optimization: Computer vision losses for large language model fine-tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, page 12060–12079. Association for Computational Linguistics, 2024.

[175] Abhijnan Nath, Andrey Volozin, Saumajit Saha, Albert Aristotle Nanda, Galina Grunin, Rahul Bhotika, and Nikhil Krishnaswamy. DPL: Diverse preference learning without a reference model. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages

3727–3747, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

[176] Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with high-quality feedback. *arXiv preprint arXiv:2310.01377*, 2023.

[177] Dun Zeng, Yong Dai, Pengyu Cheng, Longyue Wang, Tianhao Hu, Wanshun Chen, Nan Du, and Zenglin Xu. On diversified preferences of large language model alignment, 2024.

[178] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.

[179] Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. In *International Conference on Learning Representations*, 2020.

[180] Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip Torr, and Puneet Dokania. Calibrating deep neural networks using focal loss. *Advances in Neural Information Processing Systems*, 33:15288–15299, 2020.

[181] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim,

Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone, 2024.

[182] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

[183] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019.

[184] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[185] Nathan Lambert, Valentina Pyatkin, Jacob Daniel Morrison, Lester James Validad Miranda, Bill Yuchen Lin, Khyathi Raghavi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hanna Hajishirzi. RewardBench: Evaluating reward models for language modeling. *ArXiv*, abs/2403.13787, 2024.

[186] Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpacaeval: An automatic evaluator of instruction-following models, 2023.

[187] Ziqi Wang, Hanlin Zhang, Xiner Li, Kuan-Hao Huang, Chi Han, Shuiwang Ji, Sham M Kakade, Hao Peng, and Heng Ji. Eliminating position bias of language models: A mechanistic approach. *arXiv preprint arXiv:2407.01100*, 2024.

[188] Tanya Goyal, Junyi Jessy Li, and Greg Durrett. News summarization and evaluation in the era of gpt-3, 2023.

[189] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

[190] Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A long way to go: Investigating length correlations in rlhf, 2024.

[191] Fan Wu, Emily Black, and Varun Chandrasekaran. Generative monoculture in large language models. *arXiv preprint arXiv:2407.02209*, 2024.

[192] Jialun Zhong, Wei Shen, Yanzeng Li, Songyang Gao, Hua Lu, Yicheng Chen, Yang Zhang, Wei Zhou, Jinjie Gu, and Lei Zou. A comprehensive survey of reward models: Taxonomy, applications, challenges, and future. *arXiv preprint arXiv:2504.12328*, 2025.

[193] Huaxiaoyue Wang and Sanjiban Choudhury. The road not taken: Hindsight exploration for llms in multi-turn rl. In *ES-FoMo III: 3rd Workshop on Efficient Systems for Foundation Models*.

[194] Hector J. Levesque. A logic of implicit and explicit belief. In *Proceedings of the Fourth National Conference on Artificial Intelligence*, pages 198–202, Austin, Texas, August 1984. American Association for Artificial Intelligence.

[195] Rafael Rafailov, Joey Hejna, Ryan Park, and Chelsea Finn. From r to q^*: Your language model is secretly a q-function, 2024.

[196] Abhijnan Nath, Changsoo Jung, Ethan Seefried, and Nikhil Krishnaswamy. Simultaneous reward distillation and preference learning: Get you a language model who can do both. *arXiv preprint arXiv:2410.08458*, 2024.

[197] Peter R Lewis and Ştefan Sarkadi. Reflective artificial intelligence. *Minds and Machines*, 34(2):1–30, 2024.

[198] David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526, 1978.

[199] Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. Neural theory-of-mind? on the limits of social intelligence in large lms. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3762–3780, 2022.

[200] Tomer Ullman. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*, 2023.

[201] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[202] Tania Lombrozo. The structure and function of explanations. *Trends in cognitive sciences*, 10(10):464–470, 2006.

[203] Denis Hilton. Social attribution and explanation. 2017.

[204] Richard E Nisbett and Timothy D Wilson. Telling more than we can know: Verbal reports on mental processes. *Psychological review*, 84(3):231, 1977.

[205] Jacob Andreas. Language models as agent models. *arXiv preprint arXiv:2212.01681*, 2022.

[206] Nathan Lambert, Valentina Pyatkin, Jacob Morrison, Lester James Validad Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. RewardBench: Evaluating Reward Models for Language Modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1755–1797, 2025.

[207] Richard Yuanzhe Pang, Weizhe Yuan, Kyunghyun Cho, He He, Sainbayar Sukhbaatar, and Jason Weston. Iterative reasoning preference optimization. *arXiv preprint arXiv:2404.19733*, 2024.

[208] Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8689–8696, 2020.

[209] Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. MultiWOZ 2.2 : A dialogue dataset with additional annotation corrections and state tracking baselines. In Tsung-Hsien Wen, Asli Celikyilmaz, Zhou Yu, Alexandros Papangelis, Mihail Eric, Anuj Kumar, Iñigo Casanueva, and Rushin Shah, editors, *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 109–117, Online, July 2020. Association for Computational Linguistics.

[210] Fanghua Ye, Jarana Manotumruksa, and Emine Yilmaz. MultiWOZ 2.4: A multi-domain task-oriented dialogue dataset with essential annotation corrections to improve state tracking evaluation. In Oliver Lemon, Dilek Hakkani-Tur, Junyi Jessy

Li, Arash Ashrafzadeh, Daniel Hernández Garcia, Malihe Alikhani, David Vandyke, and Ondřej Dušek, editors, *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 351–360, Edinburgh, UK, September 2022. Association for Computational Linguistics.

[211] Nikhil Krishnaswamy and James Pustejovsky. An evaluation framework for multimodal interaction. In *Proceedings of the eleventh international conference on language resources and evaluation (lrec 2018)*, 2018.

[212] Mariah Bradford, Ibrahim Khebour, Nathaniel Blanchard, and Nikhil Krishnaswamy. Automatic detection of collaborative states in small groups using multimodal features. In *International Conference on Artificial Intelligence in Education*, pages 767–773. Springer, 2023.

[213] Zekun Li, Wenhu Chen, Shiyang Li, Hong Wang, Jing Qian, and Xifeng Yan. Controllable dialogue simulation with in-context learning, 2023.

[214] Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason E Weston. Self-rewarding language models. In *Forty-first International Conference on Machine Learning*, 2024.

[215] Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Model alignment as prospect theoretic optimization. In *Forty-first International Conference on Machine Learning*, 2024.

[216] Corby Rosset, Ching-An Cheng, Arindam Mitra, Michael Santacroce, Ahmed Awadallah, and Tengyang Xie. Direct nash optimization: Teaching language models to self-improve with general preferences. *ArXiv*, abs/2404.03715, 2024.

[217] Rui Zheng, Hongyi Guo, Zhihan Liu, Xiaoying Zhang, Yuanshun Yao, Xiaojun Xu, Zhaoran Wang, Zhiheng Xi, Tao Gui, Qi Zhang, et al. Toward optimal llm alignments using two-player games. *CoRR*, 2024.

[218] James Pustejovsky and Nikhil Krishnaswamy. Frictive Policy Optimization for LLM Agent Interactions. In *Workshop on Rebellion and Disobedience of Artificial Agents at the 2025 International Conference on Autonomous Agents and Multiagent Systems*, 2025.

[219] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models, 2023. *URL https://arxiv. org/pdf/2305.10601. pdf*, 2023.

[220] Xuan Zhang, Chao Du, Tianyu Pang, Qian Liu, Wei Gao, and Min Lin. Chain of preference optimization: Improving chain-of-thought reasoning in llms. *arXiv preprint arXiv:2406.09136*, 2024.

[221] Ahmed Hussein, Mohamed Medhat Gaber, Eyad Elyan, and Chrisina Jayne. Imitation learning: A survey of learning methods. *ACM Computing Surveys (CSUR)*, 50(2):1–35, 2017.

[222] Tomasz Korbak, Hady Elsahar, Germán Kruszewski, and Marc Dymetman. On reinforcement learning and distribution matching for fine-tuning language models with no catastrophic forgetting. *Advances in Neural Information Processing Systems*, 35:16203–16220, 2022.

[223] Dongyoung Go, Tomasz Korbak, Germán Kruszewski, Jos Rozen, Nahyeon Ryu, and Marc Dymetman. Aligning language models with preferences through f-divergence minimization. *arXiv preprint arXiv:2302.08215*, 2023.

[224] Gokul Swamy, Christoph Dann, Rahul Kidambi, Zhiwei Steven Wu, and Alekh Agarwal. A minimaximalist approach to reinforcement learning from human feedback, 2024.

[225] John von Neumann. Mathematische annalen 100 (1928), pp. *Contributions to the Theory of Games*, 100(40):13, 1959.

[226] Zeya Chen and Ruth Schmidt. Exploring a behavioral model of "positive friction" in human-ai interaction, 2024.

[227] Pedro Henrique Martins, Zita Marinho, and André F. T. Martins. Sparse text generation, 2020.

[228] Hannah VanderHoeven, Brady Bhalla, Ibrahim Khebour, Austin C Youngren, Videep Venkatesha, Mariah Bradford, Jack Fitzgerald, Carlos Mabrey, Jingxuan Tu, Yifan Zhu, et al. TRACE: Real-time multimodal common ground tracking in situated collaborative dialogues. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 40–50, 2025.

[229] Matthew A Koschmann. The communicative accomplishment of collaboration failure. *Journal of Communication*, 66(3):409–432, 2016.

[230] Francis Rhys Ward et al. Defining deception in structural causal games (extended abstract). In *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '23. International Foundation for Autonomous Agents and Multiagent Systems, 2023.

[231] Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone. When combinations of humans and ai are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 8(12):2293–2303, October 2024.

[232] Abhijnan Nath, Carine Graff, and Nikhil Krishnaswamy. Collaborate, deliberate, evaluate: How llm alignment affects coordinated multi-agent outcomes, 2026.

[233] Mert İnan, Anthony Sicilia, Suvodip Dey, Vardhan Dongre, Tejas Srinivasan, Jesse Thomason, Gökhan Tür, Dilek Hakkani-Tür, and Malihe Alikhani. Better slow than sorry: Introducing positive friction for reliable dialogue systems. *arXiv preprint arXiv:2501.17348*, 2025.

[234] Arthur C Graesser, Stephen M Fiore, Samuel Greiff, Jessica Andrews-Todd, Peter W Foltz, and Friedrich W Hesse. Advancing the science of collaborative problem solving. *Psychological science in the public interest*, 19(2):59–92, 2018.

[235] Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and Xiangliang Zhang. Justice or prejudice? quantifying biases in llm-as-a-judge, 2024.

[236] Eric D Langlois and Tom Everitt. How RL agents behave when their actions are modified. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11586–11594, 2021.

[237] Thomas Bolander. Seeing is believing: Formalising false-belief tasks in dynamic epistemic logic. In *European conference on social intelligence (ECSI 2014)*, pages 87–107, 2014.

[238] Timothy Obiso, Kenneth Lai, Abhijnan Nath, Nikhil Krishnaswamy, and James Pustejovsky. Dynamic epistemic friction in dialogue. In *The SIGNLL Conference on Computational Natural Language Learning*.

[239] Richard S Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211, 1999.

[240] Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, volume 99, pages 278–287, 1999.

[241] Ananya Ganesh, Jie Cao, E Margaret Perkoff, Rosy Southwell, Martha Palmer, and Katharina Kann. Mind the gap between the application track and the real world. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1833–1842, 2023.

[242] Sidney K D'Mello, Quentin Biddy, Thomas Breideband, Jeffrey Bush, Michael Chang, Arturo Cortez, Jeffrey Flanigan, Peter W Foltz, Jamie C Gorman, Leanne Hirshfield, et al. From learning optimization to learner flourishing: Reimagining ai in education at the institute for student-ai teaming (isat). *AI Magazine*, 45(1):61–68, 2024.

[243] E Margaret Perkoff, Emily Doherty, Jeffrey Bush, and Leanne Hirshfield. Crafting a responsible dialog system for collaborative learning environments. In *AI for Education: Bridging Innovation and Responsibility at the 38th AAAI Annual Conference on AI*, 2024.

[244] Sarah Wiegreffe, Ana Marasović, and Noah A Smith. Measuring association between labels and free-text rationales. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10266–10284, 2021.

[245] Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems*, 36:10622–10643, 2023.

[246] OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero,

Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce

Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver,

Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024.

[247] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, 2023.

[248] Alizée Pace, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. West-of-n: Synthetic preference generation for improved reward modeling. In *ICLR 2024 Workshop on Navigating and Addressing Data Problems for Foundation Models*, 2024.

[249] Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J. Liu. Slic-hf: Sequence likelihood calibration with human feedback, 2023.

[250] Yifan Song, Da Yin, Xiang Yue, Jie Huang, Sujian Li, and Bill Yuchen Lin. Trial and error: Exploration-based trajectory optimization of LLM agents. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7584–7600, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[251] Judea Pearl. *Causality*. Cambridge university press, 2009.

[252] Francis Ward, Francesca Toni, Francesco Belardinelli, and Tom Everitt. Honesty is the best policy: defining and mitigating AI deception. *Advances in neural information processing systems*, 36:2313–2341, 2023.

[253] Kaixuan Xu, Jiajun Chai, Sicheng Li, Yuqian Fu, Yuanheng Zhu, and Dongbin Zhao. Dipllm: Fine-tuning llm for strategic decision-making in diplomacy. *arXiv preprint arXiv:2506.09655*, 2025.

[254] Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. *Advances in neural information processing systems*, 32, 2019.

[255] Subbarao Kambhampati, Karthik Valmeekam, Lin Guan, Mudit Verma, Kaya Stechly, Siddhant Bhambri, Lucas Saldyt, and Anil Murthy. Llms can't plan, but can help planning in llm-modulo frameworks. *arXiv preprint arXiv:2402.01817*, 2024.

[256] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate, 2023.

[257] Kaitlyn Zhou, Jena D Hwang, Xiang Ren, and Maarten Sap. Relying on the unreliable: The impact of language models' reluctance to express uncertainty. *arXiv preprint arXiv:2401.06730*, 2024.

[258] Abhijnan Nath and Nikhil Krishnaswamy. Learning "partner-aware" collaborators in multi-party collaboration, 2026.

[259] Sumit Verma, Pritam Prasun, Arpit Jaiswal, and Pritish Kumar. Rail in the wild: Operationalizing responsible ai evaluation using anthropic's value dataset, 2025.

[260] Wanli Yang, Fei Sun, Jiajun Tan, Xinyu Ma, Qi Cao, Dawei Yin, Huawei Shen, and Xueqi Cheng. The mirage of model editing: Revisiting evaluation in the wild. *CoRR*, 2025.

[261] Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. *arXiv preprint arXiv:2307.03987*, 2023.

[262] Marwa Abdulhai, Ryan Cheng, Aryansh Shrivastava, Natasha Jaques, Yarin Gal, and Sergey Levine. Evaluating & reducing deceptive dialogue from language models with multi-turn rl. *arXiv preprint arXiv:2510.14318*, 2025.

[263] Changrong Xiao, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Lei Xia. Evaluating reading comprehension exercises generated by llms: A showcase of chatgpt in education applications. In *Proceedings of the 18th workshop on innovative use of NLP for building educational applications (BEA 2023)*, pages 610–625, 2023.

[264] Eleonora Grassucci, Gualtiero Grassucci, Aurelio Uncini, and Danilo Comminiello. Beyond Answers: How LLMs Can Pursue Strategic Thinking in Education. *arXiv preprint arXiv:2504.04815*, 2025.

[265] Bashar Alhafni, Sowmya Vajjala, Stefano Bannò, Kaushal Kumar Maurya, and Ekaterina Kochmar. LLMs in education: Novel perspectives, challenges, and opportunities. *arXiv preprint arXiv:2409.11917*, 2024.

[266] Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. Two Tales of Persona in LLMs: A Survey of Role-Playing and Personalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16612–16631, 2024.

[267] Shibo Hao, Yi Gu, Haotian Luo, Tianyang Liu, Xiyan Shao, Xinyuan Wang, Shuhua Xie, Haodi Ma, Adithya Samavedhi, Qiyue Gao, et al. Llm reasoners: New evalua-

tion, library, and analysis of step-by-step reasoning with large language models. In *First Conference on Language Modeling*, 2024.

[268] Hana Kim, Kai Ong, Seoyeon Kim, Dongha Lee, and Jinyoung Yeo. Commonsense-augmented Memory Construction and Management in Long-term Conversations via Context-aware Persona Refinement. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 104–123, 2024.

[269] Nia Peters, Griffin Romigh, George Bradley, and Bhiksha Raj. When to interrupt: A comparative analysis of interruption timings within collaborative communication tasks. In *Advances in Human Factors and System Interactions: Proceedings of the AHFE 2016 International Conference on Human Factors and System Interactions, July 27-31, 2016, Walt Disney World®, Florida, USA*, pages 177–187. Springer, 2017.

[270] Huaben Chen, Wenkang Ji, Lufeng Xu, and Shiyu Zhao. Multi-agent consensus seeking via large language models. *arXiv preprint arXiv:2310.20151*, 2023.

[271] Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*, 2019.

[272] Athul Paul Jacob, David J. Wu, Gabriele Farina, Adam Lerer, Hengyuan Hu, Anton Bakhtin, Jacob Andreas, and Noam Brown. Modeling strong and human-like gameplay with kl-regularized search, 2022.

[273] Laurent Orseau and M Armstrong. Safely interruptible agents. In *Conference on Uncertainty in Artificial Intelligence*. Association for Uncertainty in Artificial Intelligence, 2016.

[274] Andy Shih, Arjun Sawhney, Jovana Kondic, Stefano Ermon, and Dorsa Sadigh. On the Critical Role of Conventions in Adaptive Human-AI Collaboration. In *International Conference on Learning Representations*, 2021.

[275] Jiacheng Lin, Tian Wang, and Kun Qian. Rec-r1: Bridging generative large language models and user-centric recommendation systems via reinforcement learning. *arXiv preprint arXiv:2503.24289*, 2025.

[276] Hengyuan Hu and Dorsa Sadigh. Language instructed reinforcement learning for human-ai coordination. In *International Conference on Machine Learning*, pages 13584–13598. PMLR, 2023.

[277] Tom Everitt, Marcus Hutter, Ramana Kumar, and Victoria Krakovna. Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective. *CoRR*, abs/1908.04734, 2021.

[278] Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro Ortega, DJ Strouse, Joel Z Leibo, and Nando De Freitas. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International conference on machine learning*, pages 3040–3049. PMLR, 2019.

[279] Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. *Advances in neural information processing systems*, 30, 2017.

[280] Anna Harutyunyan, Will Dabney, Thomas Mesnard, Mohammad Azar, Bilal Piot, Nicolas Heess, Hado van Hasselt, Greg Wayne, Satinder Singh, Doina Precup, and Remi Munos. Hindsight credit assignment, 2019.

[281] Francis Ward, Francesca Toni, Francesco Belardinelli, and Tom Everitt. Honesty is the best policy: Defining and mitigating ai deception. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural*

Information Processing Systems, volume 36, pages 2313–2341. Curran Associates, Inc., 2023.

[282] Muning Wen, Cheng Deng, Jun Wang, Weinan Zhang, and Ying Wen. Entropy-regularized token-level policy optimization for large language models. *arXiv preprint arXiv:2402.06700*, 2024.

[283] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-Tuning Language Models from Human Preferences. *arXiv:1909.08593 [cs, stat]*, January 2020. arXiv: 1909.08593.

[284] Soumya Rani Samineni, Durgesh Kalwar, Karthik Valmeekam, Kaya Stechly, and Subbarao Kambhampati. Rl in name only? analyzing the structural assumptions in rl post-training for llms, 2025.

[285] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of Artificial General Intelligence: Early experiments with GPT-4, 2023.

[286] Philip Paquette, Yuchen Lu, Seton Steven Bocco, Max Smith, Satya O-G, Jonathan K Kummerfeld, Joelle Pineau, Satinder Singh, and Aaron C Courville. No-press diplomacy: Modeling multi-agent gameplay. *Advances in Neural Information Processing Systems*, 32, 2019.

[287] Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations in text classification. *Advances in neural information processing systems*, 34:16196–16208, 2021.

[288] Marilyn Strathern. 'Improving ratings': audit in the British University system. *European review*, 5(3):305–321, 1997.

[289] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

[290] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *NeurIPS Datasets and Benchmarks Track*, 2023.

[291] Mohammadhossein Rezaei and Eduardo Blanco. Making language models robust against negation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8123–8142, 2025.

[292] Thinh Hung Truong, Timothy Baldwin, Karin Verspoor, and Trevor Cohn. Language models are not naysayers: an analysis of language models on negation benchmarks. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (* SEM 2023)*, pages 101–114, 2023.

[293] Sarath Sreedharan, Kelsey Sikes, Nathaniel Blanchard, Lisa Mason, Nikhil Krishnaswamy, and Jill Zarestky. On the role of domain experts in creating effective tutoring systems. In *International Conference on Artificial Intelligence in Education*, pages 61–68. Springer, 2025.

[294] Abhijnan Nath, Carine Graff, and Nikhil Krishnaswamy. Let's Roleplay: Examining LLM Alignment in Collaborative Dialogues. *Workshop on Optimal Reliance and Accountability in Interactions with Generative Language Models*, 2025.

[295] Denis Peskov, Benny Cheng, Ahmed Elgohary, Joe Barrow, Cristian Danescu-Niculescu-Mizil, and Jordan Boyd-Graber. It takes two to lie: One to lie, and one to listen. In *Proceedings of the 58th Annual Meeting of the Association for Computa-*

tional Linguistics, pages 3811–3854, Online, July 2020. Association for Computational Linguistics.

[296] Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M Ziegler, Tim Maxwell, Newton Cheng, et al. Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training. *CoRR*, 2024.

[297] Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Samuel Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, et al. Alignment faking in large language models. *CoRR*, 2024.

[298] Anna Lieb and Toshali Goel. Student interaction with newtbot: An llm-as-tutor chatbot for secondary physics education. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–8, 2024.

[299] Suhana Bedi, Yutong Liu, Lucy Orr-Ewing, Dev Dash, Sanmi Koyejo, Alison Callahan, Jason A Fries, Michael Wornow, Akshay Swaminathan, Lisa Soleymani Lehmann, et al. A systematic review of testing and evaluation of healthcare applications of large language models (llms). *medRxiv*, pages 2024–04, 2024.

[300] Goshi Aoki. Large language models in politics and democracy: A comprehensive survey. *arXiv preprint arXiv:2412.04498*, 2024.

[301] Saurabh Pahune, Zahid Akhtar, Venkatesh Mandapati, and Kamran Siddique. The importance of ai data governance in large language models. *Big Data and Cognitive Computing*, 9(6):147, 2025.

[302] Saaket Agashe, Yue Fan, Anthony Reyna, and Xin Eric Wang. Llm-coordination: evaluating and analyzing multi-agent coordination abilities in large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 8038–8057, 2025.

[303] Tsung-Han Wu, Mihran Miroyan, David M Chan, Trevor Darrell, Narges Norouzi, and Joseph E Gonzalez. Are large reasoning models interruptible? *arXiv preprint arXiv:2510.11713*, 2025.

[304] Abdelhakim Benechehab, Youssef Attia El Hili, Ambroise Odonnat, Oussama Zekri, Albert Thomas, Giuseppe Paolo, Maurizio Filippone, Ievgen Redko, and Balázs Kégl. Zero-shot model-based reinforcement learning using large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[305] Judith Degen. The rational speech act framework. *Annual Review of Linguistics*, 9(1):519–540, 2023.

[306] Sonia K Murthy, Rosie Zhao, Jennifer Hu, Sham Kakade, Markus Wulfmeier, Peng Qian, and Tomer Ullman. Inside you are many wolves: Using cognitive models to interpret value trade-offs in llms. *arXiv preprint arXiv:2506.20666*, 2025.

[307] Pierre-Luc Bacon, Jean Harb, and Doina Precup. The option-critic architecture. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[308] Lloyd S Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39(10):1095–1100, 1953.

[309] Guilliermo Owen. Values of games with a priori unions. In *Mathematical economics and game theory: Essays in honor of Oskar Morgenstern*, pages 76–88. Springer, 1977.

[310] Anoop Cherian, Radu Corcodel, Siddarth Jain, and Diego Romeres. Llmphy: Complex physical reasoning using large language models and world models, 2024.

[311] Paul Youssef, Christin Seifert, Jörg Schlötterer, et al. Llms for generating and evaluating counterfactuals: A comprehensive study. In *Findings of the association for computational linguistics: EMNLP 2024*, pages 14809–14824, 2024.

[312] Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pages 10835–10866. PMLR, 2023.

[313] Borja Ibarz, Jan Leike, Tobias Pohlen, Geoffrey Irving, Shane Legg, and Dario Amodei. Reward learning from human preferences and demonstrations in atari, 2018.

[314] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506, 2020.

[315] Ilya Loshchilov and Frank Hutter. Fixing Weight Decay Regularization in Adam. *CoRR*, 2017.

[316] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36, 2024.

[317] Barbara J Grosz and Candace L Sidner. Attention, intentions, and the structure of discourse. *Computational linguistics*, 12(3):175–204, 1986.

[318] Corbyn Terpstra, Ibrahim Khebour, Mariah Bradford, Brett Wisniewski, Nikhil Krishnaswamy, and Nathaniel Blanchard. How good is automatic segmentation as a multimodal discourse annotation aid? In *Proceedings of the 19th Joint ACL-ISO Workshop on Interoperable Semantics (ISA-19)*, pages 75–81, 2023.

[319] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.

[320] Keyu Pan and Yawen Zeng. Do llms possess a personality? making the mbti test an amazing evaluation for large language models. *arXiv preprint arXiv:2307.16180*, 2023.

[321] Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. Reframing human-ai collaboration for generating free-text explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 632–658, 2022.

[322] Abhijnan Nath, Shadi Manafi Avari, Avyakta Chelle, and Nikhil Krishnaswamy. Okay, Let's Do This! Modeling Event Coreference with Generated Rationales and Knowledge Distillation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3931–3946, 2024.

[323] Timothy Obiso, Kenneth Lai, Abhijnan Nath, Nikhil Krishnaswamy, and James Pustejovsky. Dynamic epistemic friction in dialogue, 2025.

[324] Hongbo Zhang, Han Cui, Guangsheng Bao, Linyi Yang, Jun Wang, and Yue Zhang. Direct value optimization: Improving chain-of-thought reasoning in llms with refined values, 2025.

[325] David S Gunderson and Kenneth H Rosen. *Handbook of mathematical induction*. CRC Press LLC Boca Raton, 2010.

[326] Xuan Zhang, Chao Du, Tianyu Pang, Qian Liu, Wei Gao, and Min Lin. Chain of Preference Optimization: Improving Chain-of-Thought Reasoning in LLMs. *Advances in Neural Information Processing Systems*, 37:333–356, 2024.

[327] Joey Hejna, Rafael Rafailov, Harshit Sikchi, Chelsea Finn, Scott Niekum, W. Bradley Knox, and Dorsa Sadigh. Contrastive preference learning: Learning from human feedback without rl, 2024.

[328] Harold Jeffreys. *The theory of probability*. OuP Oxford, 1998.

[329] Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100, 2018.

[330] Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the nineteenth international conference on machine learning*, pages 267–274, 2002.

[331] Ching-An Cheng, Andrey Kolobov, and Alekh Agarwal. Policy improvement via imitation of multiple oracles. *Advances in Neural Information Processing Systems*, 33:5587–5598, 2020.

[332] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

[333] Imre Csiszár and János Körner. *Information theory: coding theorems for discrete memoryless systems*. Cambridge University Press, 2011.

[334] Clément L Canonne. A short note on an inequality between kl and tv. *arXiv preprint arXiv:2202.07198*, 2022.

[335] Jiechuan Jiang, Kefan Su, and Zongqing Lu. Fully decentralized cooperative multi-agent reinforcement learning: A survey, 2024.