THESIS

FOLLOW THE SIGNAL: MODELS OF ATTENTION, REASON, AND BELIEF

Submitted by

Videep Venkatesha

Department of Computer Science

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Summer 2025

Master's Committee:

Advisor: Nathaniel Blanchard

Nikhil Krishnaswamy
Sarath Sreedharan
Anne Cleary

ABSTRACT

FOLLOW THE SIGNAL: MODELS OF ATTENTION, REASON, AND BELIEF

Attention, reasoning, and belief are central to how we perceive, decide, and collaborate. Though inherently abstract-with no direct physsical manifestation- these phenomena leave behind observable signals in subtle traces in gaze, language, timing, and interaction. These traces vary across individuals and contexts, yet they offer a window into the underlying cognitive processes. In this thesis, I model the behavioral and linguistic signals that reflect aspects of attentional shifts, expressions of reasoning, and evolving belief states, and investigate how machine learning can be used to detect and interpret them as they arise in everyday settings.

First, I focus on moments of inward attention, identifying gaze patterns that predict when participants feel familiarity—even without conscious recall, using eye-tracking during immersive virtual tours. I then analyze written descriptions of three distinct internal attentional states: familiarity, unexpected thoughts, and involuntary memories.

Then, I frame the link of probing questions i.e questions that explicitly elicit justifications or clarifications, and their causal utterances as traces of reason as they emerge in group dialogue

Next, in the case of belief, I extract explicitly stated propositions from natural dialogue. These structured propositions reflect participants' evolving belief states during a collaborative task. I design and evaluate multiple extraction pipelines, demonstrating the feasibility of tracking belief expression in real time.

Finally, I holistically examine how automated systems with noisy data shape downstream performance on collaborative problem-solving detection—a task that inherently reflects attention, belief, and reasoning. I show that, while performance remains comparable across systems, lower-fidelity inputs reduce interpretive granularity.

In combination, these contributions demonstrate how machine learning can detect the emergence of traces of these phenomena——transforming these abstract states into observable patterns.

ACKNOWLEDGEMENTS

I have been extremely fortunate during my time at Colorado State University to work alongside incredible mentors, researchers, and scholars who have shaped my academic and personal journey.

First and foremost, I would like to thank my advisor and mentor, Dr. Nathaniel Blanchard. At a time when I hadn't yet considered research as a career, he introduced me to its potential and fulfillment. His guidance has helped me grow not only as a writer and thinker, but as someone more curious and intentional in asking meaningful questions. I am especially grateful for his support, not just as an advisor, but as someone I could rely on while navigating life in a new country.

I am also deeply grateful to Dr. Nikhil Krishnaswamy, whose mentorship and unwavering work ethic have inspired me throughout this journey. His guidance has given me unique opportunities to engage with language and research in ways that I truly believe elevated my work. I can only aspire to one day respond to messages at 4 a.m. before heading out on a 20-mile run, just like him.

I would like to thank Dr. Anne Cleary for welcoming me into the world of cognitive psychology. As someone once considering a major in psychology, I feel especially fortunate to collaborate with her on topics that I find genuinely fascinating.

I am also thankful to Dr. Sarath Sreedharan for his thoughtful insights and the occasional pep talks that helped me stay grounded. His guidance, even in passing moments, has had a lasting impact.

Finally, to my friends at the IMPACT and SIGNAL labs, thank you for making research feel exciting, collaborative, and fun. Your presence has made this journey all the more meaningful.

DEDICATION

I would like to dedicate this thesis to my parents, Veena and Venkatesha, whose love and support have led me to where I am today.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

Take a simple scene:

You're in a small group, solving a logic puzzle, where your task is deduce the weights of blocks using a balance scale. One person suddenly points out a detail no one else had noticed: "Wait, the red block dipped lower than the green one." Another interrupts, shifting the topic: "Are we even sure the scale is balanced?" A third revisits something from earlier: "Didn't we already compare red and green?". Attention shifts. A belief is questioned. The group pauses—deliberates. Someone proposes, "The red block must be heavier than the green one." Another agrees with a nod. A third disagrees: "No, I think the green one is heavier—it stayed down longer last time." Then a voice cuts in, probing: "Wait, why did we already write down red as 20 grams?" The conversation splinters. One person starts rethinking the earlier trial. Another flips through the notes, scanning for confirmation. A few go silent, mentally replaying the sequence of events. Someone finally says, "I thought we had that part figured out."

What just happened?

A cascade of distinctly human phenomena: shifting attention, evolving beliefs, and active processes of deliberation and reasoning. These are not events we explicitly declare. They arise and unfold through pauses, hesitations, gestures, changes in tone — expressions embedded in interaction [3,4]. These phenomena can be observed, often not directly, but through the signals they leave behind. This thesis models the phenomena of attention, reasoning, and belief not just as internal abstractions, but as patterns that become visible in interaction, expressed through gaze, speech, and dialogue structure. By analyzing these external traces, we ask: How do such phenomena arise, unfold, and become recognizable? And how can machine learning help us trace these patterns.

These phenomena are often not obvious — do not always present themselves through clean, discrete signals. Instead, they emerge gradually, often ambiguously, and are shaped by a constellation of external expressions like language, gaze, timing, posture, and prosody. Across disci-

plines, scholars have emphasized that human phenomena like attention, belief, and reasoning rarely map directly onto any single behavioral cue [5, 6]. Rather, they must be inferred from complex, context-sensitive patterns of expression, which are themselves shaped by interpersonal dynamics and shared history [7, 8]. Even small signals like hesitations, intonation changes, brief silences can reflect divergent processes such as uncertainty, reflection, or disagreement [9, 10]. For example, in the opening scene, when someone interrupts with "Are we even sure the scale is balanced?", the utterance is brief, but the underlying state of skepticism, confusion, or even a shift in focus must be inferred from how it fits into the ongoing group dynamic. This ambiguity in interpretation makes computational modeling of human experience especially challenging: the same phenomenon may manifest in multiple ways across individuals, and the same signal may point to different meanings depending on context [11]. And yet, people routinely navigate this ambiguity. We make sense of others not by decoding fixed meanings, but by interpreting signals in context, moment by moment. This is the everyday miracle of human interaction.

Machine learning has shown promise in identifying structure in noisy and high-dimensional data—including language, physiological signals, and multimodal interaction [12–14]. In contexts where human interpretation can be inconsistent or ambiguous—especially when signals unfold over time or depend on subtle context—machine learning offers a data-driven way to surface patterns that may not be readily apparent. While these models are far from perfect, their ability to detect regularities across large contexts makes them a useful tool for exploring how abstract phenomena like attention, belief, and reasoning leave behind external traces.In many of these domains, human interpretation can be inconsistent or ambiguous, especially when signals unfold over time or are context-dependen [15, 16].

In this thesis, we use machine learning to detect and interpret human phenomena such as shifts in attention, expressions of belief, and traces of collaborative reasoning. These are not modeled just as mental states, but as patterns that emerge in eye gaze and language during interaction. By grounding our models in annotated data and observable traces, we follow the tradition of compu-

tational models that aim not to replace human insight, but to make patterns of interaction more visible and interpretable [17, 18].

1.1 Attention

In our opening scene, one participant interrupts the ongoing discussion to ask, “Are we even sure the scale is balanced?”, a comment that cuts across the group’s current focus. Another flips through the notes in silence, while someone else appears to mentally replay past trials. These moments reflect shifts in attention, both external and internal. We may not see the full trajectory of someone’s focus, but we glimpse it in how they re-enter the conversation, what they latch onto, and how their language reveals a sudden reorientation. Attention shifts are rarely announced but they are revealed in timing, phrasing, and what gets remembered or forgotten.

Attention is the mechanism by which we prioritize certain aspects of the environment or our mental life over others. It is dynamic and limited, constantly shifting in response to both internal and external cues. Psychological and neuroscientific theories have long distinguished between different types of attention: sustained, selective, divided, and alternating [19, 20]. These forms allow us to maintain focus, switch tasks, and respond flexibly to a changing environment.

Of particular interest in this thesis is internal attention. Internal attention is less visible, more transient, and often harder to measure. Yet it plays a crucial role in how we experience familiarity, form associations, and reflect on past events [21, 22]. This thesis focuses on specific instances of internal attention as they manifest during sensations of familiarity, déjà vu, involuntary memories, and unexpected thoughts. Though subtle, these shifts leave detectable traces in gaze and language.

In Chapter 2, I analyze how familiarity manifests in gaze behavior, even without explicit recall. In Chapter 3, I examine how language reveals patterns in spontaneous thoughts. Rather than modeling all of attention, I focus on transient, internally directed shifts that may not be externally announced, but nonetheless leave subtle traces.

1.2 Reasoning

Returning to the scene: the group is not merely stating claims, but actively testing and reworking them. One participant proposes that red is heavier; another counters that green stayed down longer; a third revisits a past trial, and yet another questions why a weight had already been recorded. These contributions don't stand alone. They respond to one another, forming a chain of justifications, contradictions, and revisions. This is reasoning as interaction: a live process of deliberation, distributed across voices and time, where the group collectively tries to make sense of evidence, memory, and uncertainty. Reasoning has been studied in philosophy and psychology as the process by which individuals draw conclusions from premises or evidence. Classical models of reasoning focus on logic, deduction, and rule-based inference [23, 24], while more recent accounts have emphasized bounded rationality and the role of social context [25, 26]. In group settings, reasoning is inherently interactive as it emerges through questions, clarifications, and chains of dialogue.

While reasoning in groups can take on many forms, this thesis models reasoning, in interaction, by introducing and analyzing deliberation chains , the structured links between utterances that capture how probing questions relate to earlier justifications in Chapter 4. These chains attempt to reflect how reasoning unfolds over time, not as isolated arguments but as process in dialogue. By using machine learning to infer these links, we trace the shape of collaborative reasoning without reducing it to isolated logic steps.

1.3 Beliefs

At a pivotal point in the scene, someone asserts, "The red block must be heavier than the green one." Another later offers the opposite claim: "No, I think the green one is heavier." These are not just tentative speculations. They are beliefs, presented confidently into the shared space. Whether accepted, questioned, or silently contested, each assertion carries more than content: it reveals how the speaker interprets the situation and wants others to align. Beliefs in group settings are not just

internal states, they are are social moves, shaped by prior talk and aimed at influencing what comes next.

Beliefs are mental representations that guide interpretation and action. In cognitive science, beliefs are seen as probabilistic hypotheses that we hold about the world, updated through evidence [27]. In collaborative settings, beliefs are not just internal but also they are performed: we foucs on CPS alone in this study [28–31].

In this thesis, I model belief expression through the lens of propositional content in Chapter 5. By extracting structured propositions from dialogue—e.g., "blue = 10," "yellow is heavier than red", I capture what participants are assuming, asserting, or testing. These propositions allow us to trace the evolving belief landscape of the group and quantify how beliefs are shared, revised, or misunderstood over time.

Each of these phenomena is fundamental in our daily lives. We must reason. We must know what others believe. And we must understand the nature of our own attention. These capacities are cornerstones for interacting with the world and with one another. Yet they are also abstract, internal, and often arise from ambiguous signals. While it is impossible to fully model attention, belief, or reasoning in all their complexity, we can detect specific instances where they become visible. This thesis focuses on such instances—not to define the phenomena in totality, but to examine how they surface in context and explore their relevance across domains such as education, cognitive science, and human-computer interaction.

Thesis Overview

This thesis is not driven by the pursuit of achieving state-of-the-art bench, but by a deeper question: Can abstract, and often hidden, structures of human experience, specifically attention, belief, reasoning be made visible through language, gaze, and interaction? The thesis proceeds in three parts, moving from inward-facing cognition to collaborative expression, and finally to the systems that must interpret human signals in the wild:

- **Part I: Modeling Internal Attention Shifts in Individuals**

- **Part II: Modeling Collaborative Sense-Making in Groups**
- **Part III: From Signal to System: Challenges of Automation**

Part I: Modeling Attentional Shifts in Individuals

This part focuses on instances where attention shifts inward. Specifically, it investigates how the sense of familiarity is reflected in eye gaze, and how descriptions of spontaneous thoughts leave traces of their inherent characteristics.

- **Chapter 2: Detecting Familiarity Through Gaze**

  We use eye-tracking data from immersive virtual tours to detect moments when participants experience familiarity. Our models reveal that gaze patterns differ not only between familiar and unfamiliar scenes, but also between successful and failed recall—showing that familiarity leaves subtle, measurable traces in behavior.

- **Chapter 3: Language of Spontaneous Thought**

  We analyze natural language descriptions of spontaneous cognitive events of déjà vu, involuntary memories, and unexpected thoughts using vector representations of language, and supervised classifiers. Our models uncover linguistic patterns that distinguish between different kinds of spontaneous thoughts, suggesting that private cognitive states exhibit stable surface patterns. Furthermore, these patterns align with existing theories of memory and spontaneous cognition, suggesting that machine learning offers a viable, empirical framework for studying subjective mental phenomena.

Part II: Modeling Collaborative Sense-Making in Groups

When humans collaborate, they do more than exchange information. They negotiate beliefs, pose questions, and reason together. This part investigates how beliefs, reasoning, and shared knowledge unfold through dialogue.

- **Chapter 4: Modeling Probing and Deliberation Chains**

  We introduce the concept of a deliberation chain, a structure that links probing questions

to earlier utterances. This provides a lens for tracing how reasoning unfolds over time in dialogue, capturing the causal relationships that drive collaborative sense-making.

- **Chapter 5: Extracting Propositional Knowledge from Dialogue**

  We view beliefs as they are explicitly expressed in the form of propositions during a task. By extracting these propositions, we make visible the belief states of individual participants and reveal how shared understanding is constructed and negotiated through dialogue.

Part III: From Signal to System: Challenges of Automation

The final part of this thesis addresses the challenges of applying these models in real-world settings. Human signals do not arrive in clean form—they are filtered through systems that introduce noise and ambiguity. Here, we examine how design decisions in automated processing affect what can ultimately be detected and interpreted.

- **Chapter 6: Automated Detection of Collaborative Problem Solving Behaviors**

  We analyze how transcription and segmentation methods impact the detection of Collaborative Problem Solving (CPS) behavior, which inherently involves aspect of attention, reason, beliefs, in group dialogue. Using multi modal classifiers, we show how early-stage system choices shape what collaborative behaviors remain visible—and what becomes obscured.

We work with real-world systems where data is often noisy and imperfect. Yet, even within this noise, we find consistent signals that reflect the emergence of attention, reasoning, and belief. While modest in performance, machine learning models are still able to detect and trace these phenomena.

Part I

Modeling Attention Shifts in Individuals

This part of the thesis focuses on human phenomena that emerge when attention turns inward—experiences such as familiarity, déjà vu, involuntary memories, and unexpected thoughts. These are fleeting, often manisfeting differently across different people, yet widely shared across individuals. Though subjective, they are not inaccessible. They manifest in how people look, what they say, and how they describe what has occurred.

In this part of the thesis, we ask whether such human phenomena leave behind detectable traces, and whether machines can be used not simply to predict them, but to illuminate the subtle ways they manifest. Chapter 2 investigates familiarity as it arises in immersive environments. Even when participants do not consciously recall seeing a place before, their gaze behavior reflects subtle recognition. Using eye-tracking data, we model how familiarity is expressed through patterns of eye-gaze.

Chapter 3 focuses on spontaneous thought. We analyze written descriptions of déjà vu, involuntary memories, and unexpected thoughts to explore how these internal attentional shifts are marked in language. By training classifiers on linguistic patterns, we show that even the most unprompted thoughts exhibit surface regularities.

Across both chapters, we treat familiarity and spontaneous thought not as static mental states, but as expressive human phenomena-brief moments where attention turns inward and leaves behind behavioral signals. By modeling these traces, we show how machine learning can serve not just as a predictive tool, but as a means to study and interpret complex human experience.

Chapter 2

Detecting Familiarity through eye-gaze

2.1 Introduction

We often find ourselves sensing that something is familiar—a face, a place, a scene—without being able to pinpoint why. This internal experience, though fleeting and hard to articulate, is a window into a part of how we make sense of the world. Familiarity is not always tied to explicit recollection; it can emerge subtly, driven by partial matches in memory, structure, or context. And yet, the feeling is real, and it often influences behavior: we pause, pay attention, become curious. In such moments, attention turns inward. The mind is briefly reoriented toward internal memory where it traces, searches, compares, sensing alignment or mismatch. This inward shift, though invisible, is part of how familiarity functions as a cognitive signal.

In cognitive psychology, familiarity has traditionally been studied as one of the two processes underlying recognition memory, alongside recollection [32, 33]. Recent work suggests that familiarity may precede or even trigger attempts at recollection [22, 34, 35]. In this study, we investigate familiarity as an internal state in its own right, one that arises spontaneously and may or may not culminate in successful recall.

But how does one study an internal state? The central challenge is that such states are invisible. They must be inferred through expression. In this work, we explore whether a subjective sense of familiarity leaves a measurable trace in eye movement behavior. By capturing moments when participants self-report a feeling of familiarity, whether or not they can explain its source, we analyze the gaze patterns that precede those reports, asking: Can familiarity be read through the eyes?

To do this, we adapt a virtual tour paradigm known to induce familiarity through exposure to spatially similar but novel scenes [36–38]. While past research has used gaze patterns to classify

memory states [39], the focus has often been on unconscious indicators of recognition, not on the subjective sensation of familiarity itself.

Here, we center the experience of familiarity as it is felt and expressed. Using eye-tracking data from scenes that elicit this feeling, we train machine learning models to predict its occurrence. We first establish that from eye data, signals, while relatively small, still reveal the onset of familiarity. But familiarity is not a binary switch to be detected. It is a gradient to be read, sometimes sensed consciously, other times not. We find that even when participants do not explicitly report familiarity, their gaze can reflect it. Moreover, among scenes that do elicit familiarity, gaze patterns differ depending on whether participants successfully recall the source. Thus, in addition to treating the model as an end in itself, we use it as a probe, a tool to surface subtle patterns in behavior. In doing so, this work contributes to a larger inquiry at the heart of this thesis: How do these phenomena become legible through signals, and how might machines help us detect them?

2.2 Related Work

The automatic detection of internal cognitive states such as mind-wandering, tip-of-the-tongue moments, or familiarity relies on interpreting externally observable signals. Eye gaze, in particular, has proven to be a sensitive channel for such detection. Prior work has shown that gaze-based features can be used to classify whether a participant has seen an image before, even in the absence of explicit recognition reports [39]. However, such studies often focus on recognition rather than on the subjective experience of familiarity itself.

Our work builds on these foundations but shifts the emphasis: from memory performance to phenomenological state. By focusing on self-reported moments of familiarit, and eliciting them through scenes that are configurally similar but visually distinct, we isolate a subtler signal, one that emerges not just from direct recall, but also from sensed resemblance.

This aligns our work with recent efforts to model internal states like mind wandering using behavioral and physiological data. Across studies, gaze-based models have shown consistent advantages over other modalities in detecting lapses in attention or spontaneous shifts in thought [40,41].

While most of these studies have examined reading or video-watching contexts [42, 43], few have explored immersive environments or real-time subjective reports.

We use global gaze features—those independent of stimulus content—mirroring prior approaches, but apply them in a new domain: the detection of a spontaneous, phenomenological state. In this way, we extend the methodology of internal state detection to a new frontier, providing a computational lens on the moment when something feels known, even if we do not yet know why.

2.3 Virtual Reality Familiarity Dataset Materials and Procedure

Familiarity Task

This experiment adapted a virtual tour paradigm previously shown to evoke spontaneous feelings of familiarity [35]. Participants experienced both a "study" and "test" phase. In the study phase, participants were placed at the center of immersive virtual scenes and asked to remember the scenes along with their names. A static scene (e.g., a golf course) would appear in the headset, and a voice would say, "This is a golf course. Golf course," to reinforce encoding of both the visual and semantic content. Participants could look around by turning their heads, but they did not walk through the environments.

During the test phase, participants viewed new scenes, some of which were spatially configured to match the layouts of previously viewed study scenes, even though the surface features differed. For example, a scene labeled "hallway" in the test phase might share the same spatial arrangement of objects as a previously seen "alley." Participants were instructed to press a button on their handheld VR controller if they experienced a sense of familiarity while viewing a test scene. Scenes were presented for a fixed duration, averaging 46 seconds, and participants were not probed after each one. Instead, button presses were used to capture the moment a feeling of familiarity arose.

After the button press, participants were asked to verbally describe what the scene reminded them of, if anything. Many participants were able to identify the study scene that evoked the

sense of familiarity, though others described personal associations (e.g., “my friend’s basement”) or reported no clear source. These responses were logged by the experimenter and recorded via microphone. Participants completed two blocks of this study-test procedure.

Participants

Participants were 26 undergraduate students at Colorado State University who participated for course credit. The sample size was determined based on prior work using the same paradigm [35].

Procedure

Participants remained seated in a stationary chair for the duration of the experiment to minimize motion sickness. Each participant was fitted with an HTC Vive Pro Eye VR headset and instructed on its use. The headset included built-in headphones and handheld controllers. Eye tracking calibration was performed using the SRanipal SDK’s built-in procedure, which required participants to follow on-screen prompts while the system adjusted for individual pupil distance and eye movement characteristics.

During the study phase, scenes were presented one at a time, and participants were encouraged to remember both the scene and its label, which was spoken aloud. During the test phase, new scenes were shown, and participants were instructed:

“If the scene starts to feel familiar to you, push the button under your THUMB to indicate that it feels familiar. Try to do this AS SOON as you start to feel a sense of familiarity with the scene. Specifically, if the scene reminds you of a specific scene that you viewed earlier. Let the experimenter know what that scene is that this scene is reminding you of. Sometimes, a scene may remind you of a similar-looking scene from earlier. Whenever this happens (even if you did not push the button) please tell the experimenter the name of the earlier-viewed scene.”

Button presses were logged in real time through the Unity terminal, and participants were prompted to report the source of familiarity verbally after each one. Participants were reminded that it was normal to sometimes recognize the feeling of familiarity without being able to explain

it. The original experiment was designed with four counterbalanced versions, ensuring that any given test scene was equally likely to have been experimentally familiarized or not.

Eye Tracking and Feature Generation

Eye tracking data was collected using the HTC Vive Pro Eye headset with the SRanipal SDK (v1.3.6.8) in Unity. The SDK enabled real-time access to the following features: pupil position, pupil diameter, eye openness, gaze origin, and gaze direction. Eye tracking samples were captured at approximately 120Hz using the SDK's callback function, and data was saved to a CSV file after each scene. To ensure accurate timing, we used the system's Unix timestamp (via DateTime) rather than the SDK's internal timestamp, due to known bugs in earlier versions of SRanipal [44].

Each participant generated two CSV files (one per block), each containing nearly 100,000 rows of timestamped data. Each row captured the eye tracking features, the participant's current scene, and whether the familiarity indication button was pressed at that moment.

For consistency with prior work, we used PyTrack [45] to extract derived features such as fixation count, saccade duration, blink frequency, and microsaccade metrics.

2.4 Experiments

2.4.1 Experiment 1 — Detecting Familiarity

Participants reported 538 instances of familiarity. On average, the button indicating familiarity was pressed approximately 15.29 seconds into a scene (SD = 9.11). To extract features based on reported familiarity, we retrieved data from the moments preceding participants' button presses. We discarded a short buffer prior to the button press as has been commonly used in other model detection attempts [42, 46–48]. This way, the model prediction is not based on the physiological patterns from the act of making the report, but rather the patterns of eye data leading up to the report.

To generate negative training instances, i.e insntances where familiarity was not experineced, we randomly sampled test videos where participants did not report familiarity. To create a balanced

dataset, we generated the same amount of negative instances per participant as we had reported positive familiarity instances. This resulted in a dataset with 1,076 entries, 538 of them being positive familiarity instances and 538 of them being negative familiarity instances. All of these instances range from being 1 to 3 seconds long depending on the window size being experimented, as further discussed in section

2.4.2 Experiment 2 – Detecting the Experimentally Familiarized Status of Scenes

This experiment aimed to determine whether eye-gaze features could be used to predict whether a scene had been experimentally familiarized, that is, whether it shared its spatial configuration with a study phase scene, even when participants did not report feeling familiarity.

To construct the dataset, we extracted eye-gaze data from 1-, 2-, and 3-second windows prior to each reported familiarity indication. Four buffer periods (0 ms, 250 ms, 500 ms, and 1000 ms) were used to prevent contamination from motor artifacts associated with the button press, consistent with prior work [42, 46]. Trials in which the familiarity indication occurred too early in the scene to allow for the specified buffer and window size were excluded.

For each positive instance (i.e., a reported familiarity during an experimentally familiarized scene), a corresponding negative instance was sampled from the same participant, drawn from scenes where no familiarity was reported. No buffer was necessary for these negative instances, as they did not include a button press.

2.4.3 Experiment 3 – Detecting Recall Success Among Familiarity Reports

In this final experiment, we asked whether machine learning models could distinguish between instances where participants successfully recalled the origin of familiarity and those where they failed to do so, even though both sets involved explicit reports of familiarity.

The dataset was limited to experimentally familiarized scenes that received a button press indicating familiarity. These instances were then further split based on participants' verbal responses:

those that correctly identified the associated study scene were labeled as "recall success," while the rest were labeled as "recall failure."

As with the previous experiments, we extracted eye-gaze features from 1-, 2-, and 3-second windows preceding each button press, with buffer periods of 0, 250, 500, and 1000 ms. Because all instances involved reported familiarity, both classes (recall success and failure) contained buffer windows. Feature extraction and dataset preparation followed the same procedure as in Experiment 2.

2.4.4 Training and Evaluation.

All detection experiments used Hyperopt for hyperparameter tuning, with random search conducted over 300 training evaluations per model. We experimented with a range of supervised learning algorithms, including AdaBoost, Naive Bayes, Logistic Regression, Support Vector Classifier, Random Forest, and K-Nearest Neighbors. To ensure generalizability across participants, model performance was evaluated using a Leave-One-Participant-Out Cross-Validation (LOPOCV) approach. For each fold, a model was trained on data from all but one participant and tested on the held-out participant. This was repeated across all participants, and results are reported as the average across folds. Cohen's Kappa was used as the primary evaluation metric to account for chance agreement, with F1 scores also reported for completeness

2.5 Results

We present results for each of the three experiments below. Importantly, the goal of testing different buffer and window combinations is not to optimize model performance, but to conduct an exhaustive sweep across configurations to probe for the presence, stability, and consistency of a signal. This approach allows us to examine whether any behavioral patterns emerge across participants, even if model performance remains modest overall.

2.5.1 Experiment 1

Using a 500 ms buffer and a 2-second window, our best-performing model was a K-Nearest Neighbors (KNN) classifier, which achieved a Cohen's Kappa of 0.18 (SD = 0.14). Additional model configurations and results are shown in Table 2.1. While overall accuracy remains modest, the relatively low standard deviation suggests consistent performance across participants, rather than performance being driven by a few high-confidence cases.

Across all tested configurations, models using 2-second windows tended to outperform shorter or longer windows. In particular, the 2-second window with a 500 ms buffer provided the most stable results. Increasing the window to 3 seconds led to a drop in performance, possibly because gaze patterns unrelated to the familiarity experience were captured. Likewise, shorter buffers may have included gaze behavior associated with the act of pressing the button rather than the internal state of familiarity.

F1 scores remained relatively stable across models and window-buffer combinations (ranging from 0.55 to 0.59), suggesting that the models maintained consistent balance between sensitivity and precision regardless of overall kappa performance.

Table 2.1: Buffer-Window Model Search Results. Participants are evaluated in a leave-one-participant-out paradigm and the average (Standard Deviation in parenthesis)

Buffer	Window	Model	Cohen's Kappa	F1 Score
250 ms	1 sec	AB	0.17 (0.13)	0.59 (0.07)
	2 sec	LR	0.17 (0.22)	0.59 (0.11)
	3 sec	LR	0.14 (0.20)	0.57 (0.10)
500 ms	1 sec	KNN	0.14 (0.17)	0.57 (0.09)
	2 sec	KNN	0.18 (0.14)	0.59 (0.09)
	3 sec	AB	0.13 (0.15)	0.55 (0.08)
1000 ms	1 sec	RF	0.16 (0.15)	0.58 (0.08)
	2 sec	RF	0.17 (0.20)	0.59 (0.10)
	3 sec	LR	0.17 (0.22)	0.17 (0.11)

2.5.2 Experiment 2

In this experiment, we trained models to detect whether a scene had been experimentally familiarized, using eye-gaze features preceding either a self-reported familiarity or randomly sampled non-reports from other scenes. The best-performing model was an AdaBoost classifier using a 3-second window and a 500 ms buffer, which achieved a Cohen's Kappa of 0.16 (SD = 0.22) and an F1 score of 0.64. Full results across buffer and window configurations are shown in Table 2.3.

Across conditions, models using a 2- or 3-second window generally outperformed those using shorter durations, suggesting that signals of experimental familiarity may require a slightly longer window to accumulate. Interestingly, while all trials included in this task were from scenes that were objectively familiarized earlier in the experiment, only some had self-reported familiarity during the test scene. Models trained on scenes with reported familiarity consistently outperformed those trained on scenes without reports, but performance was still above chance even in the absence of explicit reports (e.g., Kappa = 0.09 with no report and no buffer).

These findings support the idea that the eye-gaze signal associated with familiarity may emerge even when participants are not consciously aware of it. The ability to classify scenes based on experimental familiarity—irrespective of subjective awareness—suggests that the experience of familiarity reflects a graded process, with detectable behavioral signatures even in the absence of conscious recognition.

2.5.3 Experiment 3

In this final experiment, we asked whether eye-gaze behavior could distinguish between instances where participants successfully recalled the source of a familiar scene versus when they did not—despite reporting a feeling of familiarity in both cases. All test scenes in this analysis were experimentally familiarized, and all involved an explicit familiarity report by the participant.

The best-performing model was a Naive Bayes classifier using a 3-second window and a 250 ms buffer, achieving a Cohen's Kappa of 0.25 (SD = 0.21) and an F1 score of 0.66. Additional results are presented in Table 2.3.

Table 2.2: Detecting the experimentally familiarized status of scenes using eye-gaze features — a comparison of model performance with various buffer and window sizes (Standard Deviation in parenthesis).

Buffer	Window	Model	Cohen's Kappa	F1 Score	Familiarity Reported
0 ms	1 sec	Naive Bayes	0.11 (0.17)	0.49 (0.13)	✓
	2 sec	Ada Boost	0.08 (0.16)	0.61 (0.11)	✓
	3 sec	Ada Boost	0.06 (0.17)	0.60 (0.10)	✓
250 ms	1 sec	Random Forest	0.08 (0.18)	0.62 (0.13)	✓
	2 sec	Ada Boost	0.13 (0.18)	0.63 (0.15)	✓
	3 sec	Ada Boost	0.13 (0.21)	0.60 (0.13)	✓
500 ms	1 sec	Ada Boost	0.08 (0.25)	0.62 (0.15)	✓
	2 sec	SVC	0.13 (0.16)	0.63 (0.12)	✓
	3 sec	**Ada Boost**	**0.16 (0.22)**	**0.64 (0.12)**	✓
1000 ms	1 sec	Ada Boost	0.08 (0.17)	0.62 (0.14)	✓
	2 sec	Random Forest	0.10 (0.18)	0.62 (0.15)	✓
	3 sec	Ada Boost	0.09 (0.18)	0.60 (0.11)	✓
N/A	1 sec	SVC	0.05 (0.14)	0.56 (0.17)	✗
	2 sec	Ada Boost	0.01 (0.14)	0.56 (0.14)	✗
	3 sec	**Ada Boost**	**0.09 (0.21)**	**0.63 (0.15)**	✗

Compared to the previous experiments, this task yielded slightly stronger model performance overall, with multiple configurations producing moderate kappa values and high F1 scores. These findings suggest that the act of successful recall leaves a measurable trace in the gaze behavior that precedes it—even when all participants reported a sense of familiarity. Notably, models were able to distinguish recall success from recall failure using only short eye-gaze windows prior to the button press, reinforcing the idea that familiarity itself reflects a spectrum of cognitive engagement.

2.6 Discussion

This chapter explored the subtle, sensation of familiarity, a moment when something feels known without explicit recall. Across three experiments, we investigated whether this internal state leaves detectable traces in eye-gaze behavior and whether machine learning models can use those traces to infer its presence. These experiments, offer that subjective cognitive states are not entirely hidden: they manifest through patterned behavior, and these patterns can be modeled.

Table 2.3: Detecting participant's recall status preceding positive reports among experimentally familiarized scenes — a comparison of model performance with various buffer and window sizes (Standard Deviation in parenthesis).

Buffer	Window	Model	Cohen's Kappa	F1 Score
0 ms	1 sec	Ada Boost	0.16 (0.27)	0.64 (0.17)
	2 sec	Ada Boost	0.14 (0.26)	0.53 (0.21)
	3 sec	Naive Bayes	0.23 (0.18)	0.64 (0.12)
250 ms	1 sec	Ada Boost	0.14 (0.24)	0.64 (0.13)
	2 sec	Naive Bayes	0.10 (0.24)	0.54 (0.19)
	3 sec	**Naive Bayes**	**0.25 (0.21)**	**0.66 (0.12)**
500 ms	1 sec	Naive Bayes	0.06 (0.18)	0.45 (0.16)
	2 sec	Ada Boost	0.19 (0.18)	0.66 (0.12)
	3 sec	Naive Bayes	0.23 (0.28)	0.65 (0.14)
1000 ms	1 sec	Ada Boost	0.16 (0.23)	0.65 (0.15)
	2 sec	Naive Bayes	0.18 (0.25)	0.60 (0.15)
	3 sec	Random Forest	0.17 (0.23)	0.66 (0.15)

From Feeling to Signal

Familiarity i.e., a moment is difficult to pinpoint. It arises quickly, often without conscious effort, and can disappear just as fast. Yet in Experiment 1, we found that models trained on eye movements leading up to a button press, moments when participants felt this internal tug, could detect it above chance, indicating that even a brief window of gaze data held information about the feeling before it was expressed.

While the performance was not high, it was stable: variability across participants was low, suggesting the signal is consistent, if faint. These findings affirm one of the thesis's core assumptions that even fleeting, subjective experiences become legible through behavior. In this case, the act of "feeling" familiarity was not expressed through words, but through subtle shifts in where and how long participants looked.

Familiarity Without Awareness

In Experiment 2, we pushed this notion further. Rather than modeling when participants reported feeling familiarity, we asked whether we could detect whether a scene had been shown before, even when participants didn't realize it. Models trained only on scenes that were experi-

mentally familiarized (some of which were not labeled as such by participants) were still able to classify with moderate success.

This finding speaks to the layered nature of internal states. Familiarity may begin as a preconscious signal, one that subtly shapes attention before it reaches awareness. Even when participants did not press the button, their gaze may have displays a sense of recognition. This aligns with theories of implicit memory and supports a broader view: internal states are not binary (on/off), but continuous. And machine learning, when applied carefully, can surface these gradations.

Expression, Recall, and the Layers of Knowing

In Experiment 3, we examined a finer distinction: what happens when participants not only feel familiarity, but also successfully recall the prior experience. Here, all scenes were both experimentally familiar and subjectively labeled familiar. Yet models could still detect whether participants had accessed a memory.

This suggests that the process of moving from recognition to recall leaves its own behavioral trace. The transition from a "feeling" to a "knowing" is not just semantic—it is behavioral. That machine learning models can detect this difference strengthens the case that internal mental transitions—even those that feel personal can be surfaced through computational modeling.

2.6.1 Implications for Understanding Attention and Designing Learning

This chapter reinforces a central claim of the thesis: that human phenomena—like shifts in attention and feelings of familiarity—manifest in observable ways that can be detected, interpreted, and modeled. A shift in attention, in particular, leaves behind consistent traces in gaze behavior. These traces do not transparently reveal what someone is thinking, but they form patterns that are both systematic and meaningful across individuals. Foundational work shows that eye movements reflect shifts in cognitive focus, comprehension processes, and task demands [49, 50].

In educational settings, teachers frequently rely on such nonverbal cues to assess student engagement. Behaviors like averting gaze, fixating on key visuals, or hesitating before responding are subtle but informative. Computational tools can formalize these intuitions. By identifying gaze

patterns that correspond to attention shifts such as returning to a previous visual, dwelling over a complex concept, or disengaging when confused—learning systems can detect waning focus and respond in real time. Multimodal emotion and engagement detection via gaze has been successfully demonstrated in live classroom environments [51].This work also highlights the importance of familiarity in learning. When content feels partly known yet elusive, learners often experience a sense of curiosity, a drive to resolve uncertainty [34].

More broadly, within the context of internal sates,this aligns with a growing body of research emphasizing the role of observable signals like gaze, timing, and gesture as fundamental to how humans express and develop understanding. While there is extensive research on how internal states like mind-wandering manifest in gaze behavior, this is the first work to show the sensation of familiarity being manifested in eye-gaze. [7, 10, 51, 52]

This chapter demonstrates that internal attentional shifts, though fleeting and subjective, can leave behind structured, observable traces in eye gaze. Attention, however, is not monolithic—internal shifts can take many forms, from drifting off-task to sudden memory recollections. In Chapter 3, we expand this investigation by analyzing participant-generated descriptions of spontaneous thoughts, revealing both commonalities and differences between states like déjà vu, involuntary memories, and unexpected thoughts.

Chapter 3

Language as a window into spontaneous thoughts

3.1 Introduction

In Chapter 1, we used machine learning to detect the experience of familiarity, an internal, spontaneous shift of attention inward, through patterns in eye gaze. That work showed how subtle behavioral cues can signal a cognitive state that is otherwise difficult to articulate. Familiarity, while elusive, is part of a broader category of experiences where attention turns inward unexpectedly.

In this chapter, we stay within that same realm of attention shifts. We explore spontaneous cognitive phenomena such as involuntary autobiographical memories (IAMs), unexpected thoughts (UTs), and déjà vu (DV), each a distinct instance where thought arises without conscious intention, often interrupting the flow of ongoing experience. These moments, like familiarity, are fleeting and internal, yet they leave traces: in language. While this work does not detect the occurrence of these phenomena as they happen in real time, it detects their type through how they are described. Here, the signal is not gaze or physiology, but the linguistic patterns that emerge when participants reflect on these experiences.

These experiences have typically been studied through traditional methods such as structured self-report and appraisal ratings. Here ,we ask what can be learned from their linguistic expression. We position machine learning, particularly natural language processing, as a tool to explore how such internal states are framed, encoded, and expressed in words.

Spontaneous thoughts are mental states or sequences of mental states that occur often in our daily lives. They are a cornerstone of human cognition, occupying as much as 30–50% of our waking lives [53]. They are thought to arise due to an absence of strong constraints such as deliberate focus, external demands, or specific tasks that guide our attention or thought [54]. For example, you walk into a cafe for the first time, and you suddenly think to yourself, “I feel like I’ve been

here before, but I can't pinpoint why". This sudden, compelling sensation that a situation has been experienced before despite evidence to the contrary is known as déjà vu, a form of spontaneous thought [55].

Historically, the study of spontaneous thoughts has relied on participants' self-reports of such experiences as well as their corresponding appraisal ratings [55, 56]. For example, Involuntary Autobiographical Memories (IAM), the recollection of personal events triggered by environmental cues, are often recorded in diary studies where participants log memory occurrences and rate them on various dimensions [57], while unexpected thoughts (UT), thoughts that feel surprising in timing and content and offer new insights or novel perspectives [57, 58], have been assessed through structured prompts that assess prior experiences with such phenomena [58, 59].

These approaches have allowed us to describe the phenomenological patterns and potential triggers of such experiences by having participants reflect on dimensions such as emotional valence and their possible cues. At the same time, utilizing appraisal ratings alone neglects the linguistic dimension of the descriptions provided by participants, or the ways in which these experiences are described. In this work, we address this gap by leveraging the linguistic information provided by participants in a study that has been previously published [59] to examine three distinct spontaneous thought types – déjà vu, IAMs, and UTs.

We expand on previous work by [59], which showed several key differences between the three spontaneous thought types mentioned above, among older and younger adults. Their findings indicated that older adults tended to rate their involuntary thoughts as more spontaneous, absorbing, and unrelated to their current task, while younger adults described déjà vu as more positive compared to IAMs and UTs. To show these differences, [59] relied on principal components analysis and sentiment analysis from participants' descriptions and appraisals of these past experiences, along with machine learning classification models to predict participants' age and the type of involuntary thought. We extend their study by investigating the linguistic characteristics of spontaneous thoughts, leveraging natural language processing techniques to analyze how participants describe their experiences.

We chose to examine the content of participants' provided thought descriptions as, language provides a window into how individuals encode, frame, and express these mental experiences, reflecting both cognitive and emotional nuances. The idea that language forms the bedrock of cognition has been widely discussed, with researchers arguing that linguistic structures provide a framework for conceptualization and reasoning [60]. Furthermore, language has been shown to be central to emotional experience, playing a crucial role in the construction and expression of affective states [61] and in the communication of emotions in social contexts [62]. Thus, how individuals articulate and frame their thoughts through language can reveal subtle patterns within spontaneous thought types and may provide insight into potential similarities and differences among different forms of spontaneous thought.

We hypothesized that linguistic features would meaningfully distinguish Déjà Vu, Involuntary Autobiographical Memories, and Unexpected Thoughts, echoing prior appraisal-based findings. Specifically, we expected language to reflect known theoretical distinctions such as DV's abstract and generally positive tone [34, 63], IAMs' vivid personal detail, and UTs' surprising and often negative content [58], while remaining open to emergent patterns through exploratory analysis.

We conduct our analysis using two language representations: Term Frequency-Inverse Document Frequency, a common method for highlighting important words in text, and contextual embeddings from BERT, which capture meaning based on context within a sentence. These methods allow us to assess both surface-level word usage and deeper semantic patterns within participants' descriptions. Additionally, we employ logistic regression coefficient analysis to identify words most predictive of each thought type, shedding light on the specific linguistic markers associated with spontaneous thought experiences. Beyond classification, we also analyze the emotional content of these descriptions using pre-trained models, capturing the affective landscape of each thought type.

Our findings reveal that spontaneous thoughts exhibit distinct linguistic signatures. Déjà vu experiences are more likely to contain abstract and spatial terms, whereas IAMs are characterized by vivid, autobiographical details. UTs, on the other hand, often include markers of surprise,

intensity, and intrusiveness. Classification models achieve over 70% accuracy, demonstrating that language alone provides meaningful differentiation between these thought types. Furthermore, emotional analysis highlights key differences in affective tone, with déjà vu descriptions showing a greater proportion of neutral and positive emotions, while IAMs and unexpected thoughts tend to contain more negative emotional content. These findings align with existing research on the phenomenology of spontaneous thoughts, which suggests that IAMs are more emotionally intense and personally significant [64], while déjà vu is often described as a neutral or even slightly positive metacognitive experience [34, 63]. Similarly, prior work on UTs highlights their intrusive and unpredictable nature [58].

3.2 Methods

3.2.1 Dataset

Data from 314 participants were analyzed to examine three types of involuntary thought experiences: Déjà vu, Involuntary Autobiographical Memory (IAM), and Unexpected Thought (UT). Participants were instructed to recall and type out one UT, one IAM, and one instance of déjà vu, for a total of three thoughts. These descriptions formed the basis of the text classification models and linguistic analyses conducted in our study. The dataset consisted of detailed textual descriptions of involuntary thoughts, with additional metadata capturing appraisal dimensions such as spontaneity, relatedness to the task, and emotional intensity. The prompts are detailed below. Additional details on the dataset are available in [59].

- Déjà vu: Think back to a time when you felt strangely like a situation was a re-experience of something that you've experienced before, but could not pinpoint why.
- IAM: Think back to a time when you had a specific memory involuntarily pop into your head.
- UT: Think back to a time when you had an unexpected thought involuntarily pop into your head.

All of the full recall data is also available at https://osf.io/ge3f8/?view_only=2826a7aa65a34e77a4c22b77a159194a

3.2.2 Language Representation

Two types of language representations were used: Term Frequency-Inverse Document Frequency (TF-IDF) [65] and BERT-base-uncased [66]. These approaches provide both basic and richer contextual embeddings, allowing us to evaluate language patterns at different levels of complexity.

TF-IDF Representation

The TF-IDF representation is a traditional method that transforms text into numerical values based on word frequency, allowing for a straightforward yet informative depiction of a text's vocabulary. TF-IDF emphasizes words that are unique to specific documents within a larger collection, aiming to highlight distinctive language features. The "Term Frequency" (TF) component measures how often a word appears in a document, while the "Inverse Document Frequency" (IDF) component downweights words that appear frequently across many documents. By combining these factors, TF-IDF assigns a higher weight to words that are important within individual documents but rare across the entire set of documents, helping to distinguish each thought type based on characteristic vocabulary.

For instance, if the word "memory" appears frequently in descriptions of IAMs but less so in descriptions of déjà vu or UTs, TF-IDF would assign it a relatively high weight in IAM-related texts. This method provides a surface-level representation that is easy to interpret and highlights distinctive terms for each thought type without relying on the broader context in which they appear.

BERT-base-uncased Representation

TF-IDF captures the frequency and uniqueness of words but it lacks the ability to understand word meanings in context. To address this limitation, the BERT-base-uncased, a transformer-based model was used, that can generate richer, context-sensitive embeddings for each thought type.

BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained model designed to capture nuanced language patterns by examining the context of each word. BERT considers each word in relation to its surrounding words, generating embeddings that reflect both the individual word meanings and the overall sentence context, unlike traditional word representations, which treat words independently [66].

3.2.3 Text Classification

Text classification machine learning models were trained to automatically classify participant thought types. Model performance was assessed using accuracy, Cohen's Kappa [67], and the F1 score. Accuracy represents the proportion of correct predictions out of the total number of predictions. In contrast, Cohen's Kappa evaluates the agreement between predicted and actual values while accounting for agreement occurring by chance. Kappa values range from -1 (complete disagreement) to 1 (perfect agreement), with a value of 0 indicating that the model's performance is equivalent to random guessing. Finally, F1 score is the harmonic mean of precision and recall, where precision is the ratio of total positive to total predicted positive, and recall is the ratio of total positive to total actual positive.

Text Preprocessing

We applied two sets of analyses to prepare the descriptions of involuntary thoughts for text classification. Text preprocessing refers to standard methods for cleaning and preparing text for computational analysis. We conducted one analysis without any preprocessing and one with preprocessing. The preprocessing aimed to reduce potential biases that could arise from participants explicitly stating the type of thought (e.g., stating "I experienced déjà vu when…"). This is a well established technique used in Natural Language Processing [68]. We thus applied the following text preprocessing — a) Stop Word Removal: Common stop words (e.g., 'the', 'is' and were removed and b) Keyword Removal: Words explicitly indicating the type of thought (e.g., 'déjà vu', 'unexpected', 'involuntary', 'popped') were removed. For example, the sentence "I once had an

unexpected thought to leave the UK and emigrate to New Zealand against the wishes of my family" becomes “leave UK emigrate New Zealand against wishes family".

Model and Hyperparameter selection

Several machine learning models such as logistic regression, decision tree, random forest, fast-text, and support vector machine (SVM) were used for the TF-IDF representation. Hyperopt was used to fine-tune hyperparameters. This library utilizes a probabilistic model to systematically search for hyperparameters that optimize the model’s performance metric. In our case, we sought to maximize the weighted average F1 score over a 1000 trials. We selected the model with median F1 performance in order to analyze the general trends, rather than risking drawing conclusions from an outlier model. We also note that the model search returns multiple models with the same output since slight changes in the hyperparameters might not cause a significant change in the model performance. The duplicate models were removed and the results of the median model are presented in the result section. For the richer BERT representations, we employed a grid search to converge on a set of hyperparameters. Regardless of language representation, each model was trained and evaluated using leave-one-participant-out cross-validation

Coefficient Analysis for Logistic Regression

A logistic regression classifier trained on the TF-IDF representation of participant descriptions was used for the analysis of the linguistic features. Logistic regression allows for the examination of feature coefficients, which provide insights into how individual words contribute to the classification of thought types. Each feature (word) in the dataset is assigned a weight, indicating its positive or negative association with a specific thought type. To identify these associations, we extracted the top 10 positively and negatively weighted coefficients for each thought type. These coefficients represent the linguistic features most predictive of each class, providing a foundation for understanding the distinctive language patterns associated with déjà vu, IAMs, and UTs.

3.2.4 Emotion Analysis

An emotion analysis was conducted to further investigate the inherent nature of these internal states. Specifically, we examined the specific emotions embedded in the thought descriptions. For this purpose, a pre-trained transformer model trained on 58k tweets was used. The text was not preprocessed in this analysis to maintain consistency with the original dataset the model was trained on [69]. The model was used to classify emotions into categories: joy, sadness, fear, surprise, anger, or love.

We added an additional layer of analysis, given that such models are often biased toward the data they are trained on, and recognizing that the domain of tweets might not fully translate to involuntary thought experiences. We utilized Llama 3.1 8-B [70], a state-of-the-art large language model, to classify each thought description into the emotion categories motivated by Ekman's Theory of Emotions [71]. To accommodate the fact that some thoughts might not contain significant emotional content, we also included a 'neutral' category. This ensured a more nuanced understanding of the emotional landscape, allowing us to account for instances where the emotional content was either subtle or absent.

The combination of these approaches aims to minimize the potential biases and provides a more robust analysis of the emotional characteristics present in each thought type. By leveraging both a pre-trained model and a flexible prompting approach with Llama, we aimed to capture a broad spectrum of emotions conveyed in the participants' reports.

3.3 Results

We begin by presenting the results of the text classification models. We then present the results of the coefficient analyses to reveal the distinct words that allowed the models to distinguish one thought type from another. Finally, we present the emotion analyses, which examined the distribution of emotions across the different thought types using both the BERT model and the Llama 3.1 prompt-based approach.

3.3.1 Text Classification

The results for both TF-IDF and BERT representations are summarized in Table 3.1, with Table 3.2 providing the confusion matrices across all conditions. The matrices illustrate the number of correct and incorrect predictions made by each model for each thought type, allowing us to better understand the performance of each approach.

Table 3.1: Performance metrics across processing levels and language representation methods

	TF-IDF		**BERT**	
Metric	**Raw**	**Pre-proc.**	**Raw**	**Pre-proc.**
Accuracy	0.80	0.70	0.86	0.73
F1	0.81	0.70	0.86	0.73
Kappa	0.71	0.60	0.80	0.58

As can be seen from Table 3.1, each text classification model was able to accurately classify each thought type with at least 70% accuracy, suggesting that the language used to describe these involuntary thought experiences is quite separable while still suggesting overlapping features amongst them. The raw descriptions with BERT representations performed the best, achieving 86% accuracy. However, it is important to note that there is a significant drop in performance for both the TF-IDF and BERT models after pre-processing, likely because many of the provided thought descriptions contained keywords (e.g., memory) that improved overall model performance. Notably, the models utilizing contextual embeddings (BERT) did not significantly outperform traditional approaches like TF-IDF in accuracy and F1 scores. This suggests that while contextual embeddings provide added depth of understanding, the improvement over traditional methods like TF-IDF is incremental, which underscores the complexity of distinguishing between thought types.

Table 3.2 illustrates an interesting pattern in the data. Déjà vu exhibited greater separability (i.e., was less confused with other thought types) compared to Involuntary Autobiographical Memories (IAM) and Unexpected Thoughts (UT), which were consistently more likely to be confused with each other across all classification models. This suggests that IAMs and UTs may be

Table 3.2: Confusion Matrices for Déjà Vu (DV), Involuntary Autobiographical Memories (IAM), and Unexpected Thoughts (UT) using Term Frequency-Inverse Document Frequency (TF-IDF) and Bidirectional Encoder Representations from Transformers (BERT) models

Model	Actual	Predicted		
		Déjà Vu	**IAM**	**UT**
Raw TF-IDF	**Déjà Vu**	258	37	19
	IAM	18	253	43
	UT	16	16	252
Preprocessed TF-IDF	**Déjà Vu**	237	41	36
	IAM	21	213	80
	UT	19	86	209
Raw BERT	**Déjà Vu**	280	23	11
	IAM	10	273	31
	UT	16	33	235
Preprocessed BERT	**Déjà Vu**	253	31	30
	IAM	30	228	36
	UT	27	84	203

more phenomenologically similar to one another, with déjà vu occupying a separate space in the phenomenology spectrum, a proposal consistent with previous findings [58, 59]

3.3.2 Coefficient Analysis

We utilized a logistic regression classifier that was trained on the TF-IDF representation to further investigate the source of confusion amongst the thought types and, to identify the most and least influential tokens (i.e., positive and negative words) for predicting each thought type. These tokens thus reflect the subtle linguistic differences that distinguish déjà vu, IAM, and UT within the model. Déjà vu was characterized by positively weighted abstract and spatial terms (e.g., seemed, place, visited), while negatively weighted words were personal (e.g., mother, father, child), highlighting its contentless, metacognitive nature [22, 34, 72]

IAMs showed the opposite trend, positively weighted words were personal (e.g., parents, girlfriend, mum), while negatively weighted words referenced knowledge or experience (e.g., knew, happened, strange), aligning with findings that IAMs are vivid and autobiographical [57, 59]

UTs were marked by positively weighted emotionally charged words (e.g., unexpectedly, death, urge, random), while also having a few terms that were personal and knowledge-based, similar to IAMs. This linguistic overlap likely explains why IAMs and UTs were more frequently confused with each other than with déjà vu [58, 59]

3.3.3 Emotion Analysis

Finally, we analyzed the emotional content of the thought reports using two models — a BERT based model trained on Twitter data and a Llama 3.1 zero shot emotion classification. Figure 3.1 presents the results of the BERT-based model. The model classifies a sentence into Joy, Fear, Sadness, Love, Surprise, and Anger. As can be seen from figure 3.1, the BERT-based model indicated that déjà vu reports contained more words related to joy and fewer words related to sadness and anger compared to UT and IAM. UTs and IAMs shared a greater overlap in emotional tone, with IAMs generally carrying fewer fear-related terms.

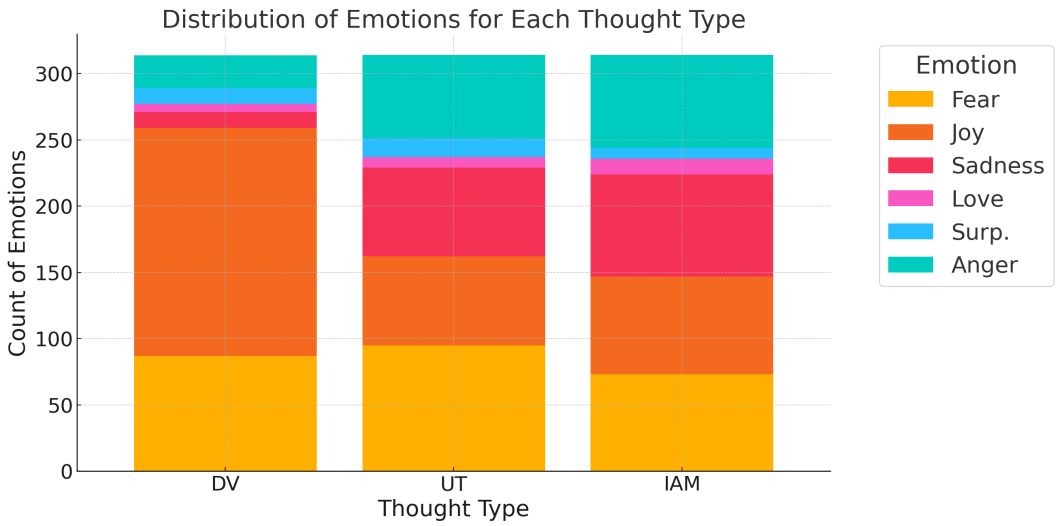

Figure 3.1: Distribution of emotions for each thought type using BERT. The model classified sentences into one of seven emotion categories: Fear, Joy, Sadness, Disgust, Surprise, Anger.

We then prompted Llama 3.1 with 7 emotion categories where the intention was to extract the emotion from the content of the experience. The model classified each sentence into one of the emotions: Fear, Joy, Sadness, Disgust, Surprise, Anger, and Neutral. The distribution of

these emotions for each thought type is summarized in Figure 3.2. The Llama model provided a more nuanced understanding of the emotional landscape by including a 'neutral' category, which helped to capture the subtler aspects of emotional content. Furthermore, the prompt was designed to extract the emotions of the content of the thought. The analysis revealed that déjà vu was characterized by a higher proportion of neutral and joy-related language, while IAM was associated with more fear-related terms, and UT exhibited more varied emotional content, including sadness and disgust. We also note that the two models are not meant to be directly compared, but are instead used in an exploratory sense to surface patterns in the emotional tone of thought descriptions. The inclusion of the 'neutral' category in the LLaMA-based classifier is intentional—it allows the model to indicate when a thought lacks strong emotional valence, rather than forcing a choice among only highly affective categories. While further work could expand the emotional taxonomy or calibrate models more precisely to this task, this analysis provides a starting point for probing how emotional tone differs across spontaneous thought types.

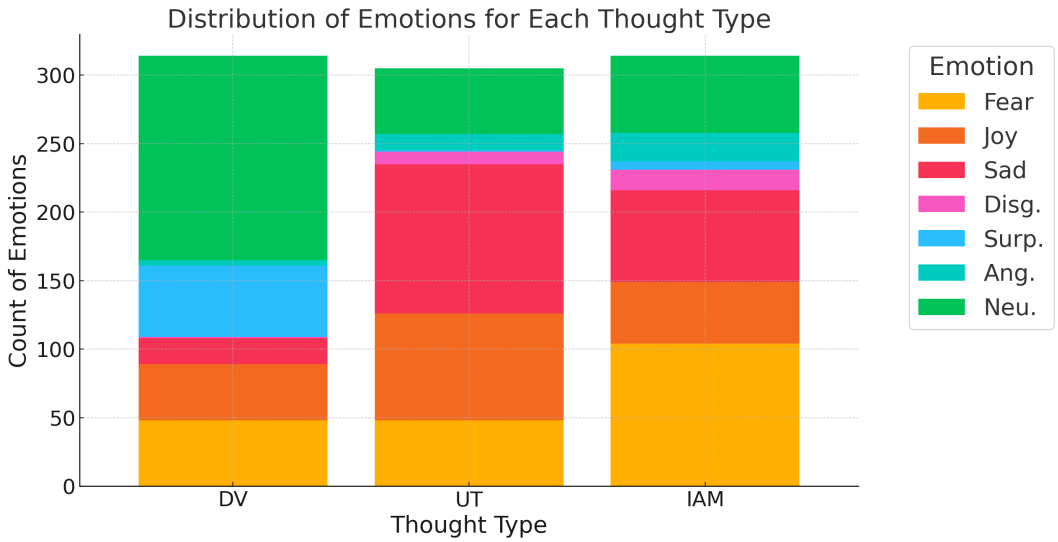

Figure 3.2: Distribution of emotions for each thought type using LLaMA 3.1. The model classified sentences into one of seven emotion categories: Fear, Joy, Sadness, Disgust, Surprise, Anger, and Neutral.

3.4 Discussion

Our findings revealed intriguing patterns that both align with and extend existing literature. Our results supported our initial hypothesis that language alone could meaningfully differentiate spontaneous thought types in ways consistent with prior appraisal-based methods. First, using our novel approach, which included leveraging richer language representations and selecting the model with median performance, we replicated known distinctions between thought types, with déjà vu showing greater separability and IAMs and UTs frequently confused—patterns [59]. Patterns of misclassification between IAMs and UTs further emphasized the nuanced overlap between these two thought types. IAMs often contained specific and vivid details, leading to frequent confusion with UTs, particularly when emotionally significant content was present. In contrast, déjà vu descriptions, despite occasionally including concrete narrative elements, were largely distinct, further underscoring their abstract nature. These trends highlight how both linguistic content and emotional tone play a crucial role in distinguishing thought types.

To further explore how the models classified each thought type, we utilized a coefficient analysis, which revealed several intriguing patterns. For déjà vu, positively weighted terms included neutral or spatial descriptors such as “place", “walking", and “seemed" again suggesting a sense of abstract familiarity and also consistent with research showing that déjà vu most commonly occurs with places [63]. Conversely, negatively weighted terms, such as “mother," “father," and “child," underscored the absence of strong personal or emotionally charged connections, reinforcing the idea that déjà vu may be characterized as a phenomenon that is contentless in nature [73]. This aligns with theories of metamemory [72], which posit that the feeling of familiarity without specific content is a distinct attentional state with this study providing computational evidence for its abstract and spatially oriented nature.

In contrast, IAMs were characterized by positively weighted terms related to vivid, personal, and relational content, such as “dinner," “party," and “father," reflecting their autobiographical essence [57]. Negative terms for IAM, including “strange" and “walking," highlighted the contrast with the concrete, detailed nature of these memories. For unexpected thoughts, positively weighted

terms like “random," “intrusive," and “unexpectedly" captured their sudden and surprising quality, while negative terms, such as “specific" and “visited," indicated that UTs are less tied to structured or historical details.

The exploration of emotional content within the descriptions of spontaneous thoughts was motivated by prior research highlighting the integral role of emotions in cognitive phenomena such as memory retrieval, attention shifts, and mind-wandering. For example, [57] emphasized the emotional vibrancy of involuntary autobiographical memories, while [58] and [59] identified the emotionally charged nature of unexpected thoughts. Similarly, the association of déjà vu with curiosity and positive emotional states [34] provided a rationale for investigating emotional dimensions across all three thought types. Our findings revealed distinct emotional profiles.

Before delving into the specific results, it is important to note that the emotion analysis broadly supports the classification analyses by reinforcing the observed distinctions among the three thought types. Specifically, déjà vu, which was consistently classified as the most distinct thought type, also exhibited a unique emotional profile characterized by neutrality and occasional positivity. In contrast, the overlap in emotional profiles between UTs and IAMs - both displaying predominantly negative emotional content—parallels their higher misclassification rates.

Déjà vu consistently exhibited more neutral or positive emotional content, aligning with its characterization as a phenomenon marked by a sense of familiarity without specific or emotionally charged content [34, 73]. UTs demonstrated a predominance of negative emotional content, with higher levels of fear and sadness. These thoughts often emerge as startling or disturbing, with emotionally charged content that disrupts attention. The Llama model’s classification of UTs included significant markers of ‘negative feelings,’ mainly fear, underscoring their intrusive and distressing impact. Similarly, IAMs also exhibited predominantly negative emotions, but these were more frequently manifested as sadness rather than fear. This distinction emphasizes the autobiographical and emotionally reflective nature of IAMs.

This study is also motivated by the broader implications of classifying attentional states. Detecting attention shifts is a growing area of interest in applied domains such as education and

adaptive learning systems [31, 74–76]. For example, given the association between déjà vu and curiosity [34], identifying instances of déjà vu could signal teachable moments where learners are primed for exploration, while UT, akin to mind-wandering, might require intervention to refocus attention. By identifying not just when attention shifts occur but also the nature of the underlying thought, ML models could pave the way for personalized interventions that enhance engagement and learning outcomes.

Several limitations must be acknowledged in the current study. The use of natural language processing to identify linguistic patterns may inadvertently highlight the structure of language rather than the inherent nature of the thoughts themselves. However, despite this potential limitation, our findings align with existing theories on spontaneous cognition, suggesting that the linguistic patterns captured are reflective of underlying cognitive states rather than merely artifacts of language use. Additionally, the reliance on retrospective self-reports introduces the challenge of memory accuracy, as recollection may not fully capture the immediacy or authenticity of these thoughts in real time. This limitation underscores the difficulty of detecting and analyzing spontaneous cognitive states as they occur. The reliance on self-reported descriptions introduces subjectivity, as participants may differ in their ability to articulate their thoughts or may selectively report details. This limitation is compounded by linguistic and cultural variability; the study focused exclusively on English-speaking participants, which restricts the generalizability of findings to other languages and cultures. Another limitation lies in the potential overfitting of ML models to dataset-specific patterns, despite the use of cross-validation techniques and the use of the median model. We also acknowledge that language may not provide a fully unfiltered window into internal cognitive states, and that descriptions are inevitably shaped by participants' conceptualizations and familiarity with certain terms. Our preprocessing approach aimed to remove the most overt self-labels (e.g., 'déjà vu', 'unexpected thought') to reduce trivial classification, while preserving more naturalistic descriptors like 'intrusive', which participants might use spontaneously rather than diagnostically. Future work could more systematically investigate the boundary between conscious self-ascription and underlying linguistic signals. Additionally, while we envision ap-

plications for improving educational and clinical interventions, we recognize that language-based models of spontaneous thought raise important ethical concerns, including the potential for misuse in domains such as targeted advertising or persuasion.

The present findings open several avenues for future research. First, the ability to distinguish spontaneous thought types using linguistic patterns suggests potential applications in clinical memory assessments, where spontaneous recollections could serve as non-invasive cognitive markers. Second, the alignment between linguistic signatures and cognitive theories motivates the development of computational models of spontaneous thought that incorporate language-based features, enabling more precise modeling of spontaneous retrieval and prediction errors in cognition.

Conclusion

This study demonstrates that language can serve as a powerful lens for examining internal cognitive states—specifically, spontaneous thoughts that arise without conscious intention. While our findings replicate known distinctions between thought types like déjà vu, IAMs, and UTs, they do so through linguistic and computational methods, offering converging evidence that the way we describe our thoughts reflects meaningful psychological structure. Importantly, the approach taken here does not replace traditional appraisal methods, but augments them. By modeling the language people use to describe spontaneous cognition, we gain access to subtle features of these experiences—patterns that may not be consciously reported, but are nonetheless embedded in expression. As such, language-based analysis stands as a complementary and scalable modality for studying spontaneous thought, opening the door to future work on more ambiguous, under-theorized, or culturally varied internal states.

This chapter advances the thesis's central claim: that human phenomena, like inward shifts of attention in this case, leave interpretable traces in outward behavior. While Chapter 2 showed how gaze patterns reflect subtle transitions in familiarity and focus, this chapter demonstrates that language, too, carries markers of how attention turns inward.

Part II

Modeling Collaborative Sense-Making in Groups

Part II Introduction

The first part of this thesis focused on internal shifts of attention—specifically, how phenomena like the sensation of familiarity or the emergence of spontaneous thoughts reflect brief, inward attentional shifts. By analyzing gaze behavior and linguistic expression, we showed how subtle signals can reveal moments when abstract phenomena like inward shifts of attention become externally observable.

In Part II, we shift focus to two other abstract phenomena: reason, the process of drawing conclusions, and belief, what one holds to be true. While each can occur at the individual level, we examine how they emerge and evolve in *collaborative dialogue*, where they are expressed, negotiated, and challenged over time. In group settings, reasoning and belief are rarely static—they unfold through interaction, shaped by turn-taking, disagreement, and shared goals. This part builds on the scene introduced in the thesis: a moment of confusion, clarification, and revision, where group members make sense of a problem together. Through dialogue, they express assumptions, challenge inferences, and iteratively construct knowledge. Our goal here is to model this process—to trace how group-level understanding forms and evolves.

We model two aspects of this dynamic process. Chapter 4 introduces the concept of a deliberation chain: a structured link between a probing utterance and the prior statements that prompted it. These chains serve as traces of reasoning, allowing us to computationally reconstruct how understanding evolves over time. We use discourse modeling, clustering, and linking algorithms to detect these chains in natural dialogue. Chapter 5 turns to propositional expression. We develop methods to extract structured propositions from collaborative dialogue. These propositions represent what participants explicitly state and, implicitly, what they believe, about the task at hand. By systematically extracting these expressions, we make belief structures and task-relevant reasoning more accessible to downstream modeling.

Together, these chapters explore how deliberation and belief expression manifest at the group level, not through introspection but through observable language. We treat collaborative reasoning

and belief as an expressive phenomenon, something that unfolds across dialogue. and we show how machine learning can be used to trace its structure and surface its patterns.

Chapter 4

Modeling Probing and Deliberation Chains

4.1 Introduction

In collaborative dialogue, understanding does not emerge all at once. It unfolds incrementally, through contributions that build, revise, or challenge what has been said. Sometimes, a pivotal moment occurs when a speaker asks a question that surfaces tension, ambiguity, or contradiction in the group's reasoning. These moments, marked by probing utterances, are rarely spontaneous; they often emerge as the result of a shared yet gradually constructed deliberation.

This chapter continues the thesis-wide aim of modeling human phenomena as they manifest in observable interaction. Here, we focus on deliberation, the collaborative process of reasoning through possibilities, testing assumptions, and co-constructing knowledge. Deliberation is not always orderly. It can be distributed across speakers, nonlinear in structure, and encoded in varied language. But its traces are there in dialogue, especially in how probing questions arise.

Recent breakthroughs in generative AI have raised the possibility of systems that follow and interact with multiparty dialogue. But modeling deliberation chains is particularly challenging: while utterances follow a linear order, the reasoning they reflect may be nonlinear, distributed across speakers, and expressed in varied language. Capturing these structures requires methods that go beyond turn-level models.

In this chapter, we define deliberation chains as turn sequences that culminate in a *probing intervention*—an explicit elicitation of input that introduces no new information but pushes the group to reflect or justify. Crucially, we model these probes as arising from earlier *causal interventions*: prior utterances that directly contribute to the probe's emergence. Without these causal links, the probing question would not have occurred in that form or moment.

Both probe and cause have been linked to effective group performance [77]. Modeling these chains allows us to trace the arc of group reasoning, not just what was said, but why it was said

at a particular moment, and in response to which prior ideas. Being able to track them supports a deeper understanding of collaborative dynamics, enabling disagreement detection, prompting for clarification, or analyzing how ideas evolve over time [78–80].

Our approach draws from discourse coherence theory and joint modeling techniques commonly used in coreference resolution. We argue that linking probing utterances to their causal antecedents is a necessary step toward AI systems that understand and support group reasoning—particularly in domains like education and team-based problem solving.

Our contributions are:

- A novel task of automatically constructing "deliberation chains" of probing questions in a dialogue and with their causal utterances;
- A formal graphical framework for deliberation chains derived from the semantics of situated dialogue [78];
- A unique adaptation of methods from coreference resolution to this new task;
- Baseline evaluation on two collaborative dialogue datasets—DeliData and the Weights Task Dataset—using a joint modeling framework to link probing and causal interventions.

By computationally modeling how deliberation surfaces through language, this chapter advances the broader goal of understanding group reasoning, not as an abstract state, but as a phenomenon enacted through interaction. Our code may be found at: https://github.com/csu-signal/ProbingDelibration

4.2 Related Work

Collaborative Dynamics Effective collaboration relies on more than task alignment—it depends on how individuals interact to co-construct shared understanding. Prior work has highlighted the centrality of questioning in this process [81–83], as well as the need for individuals to actively initiate and sustain interaction [84]. Dialogue mechanisms such as conversational repair play a key role in negotiating common ground [85], while the ability to externalize internal knowledge—through explanation, clarification, or probing—is essential to collaborative sense-making [86]. Importantly,

deliberation has been linked to conceptual change within a group: by prompting reconsideration or surfacing unspoken assumptions, it can lead team members to revise their positions, thereby deepening shared understanding [87, 88].

Joint Modeling in Coreference Resolution Our approach draws on insights from joint modeling techniques in coreference resolution, where the goal is to identify and cluster referring expressions. Several models have proposed end-to-end architectures that jointly optimize entity linking and mention detection [89–92], with subsequent work addressing scalability [2, 93] and generalization across domains [94]. Unlike span-level mention detection, our work treats utterances as the core discourse unit and applies similar joint modeling logic to detect links between reasoning steps in dialogue. This framing allows us to repurpose structural modeling techniques from coreference to track causal relationships between utterances within collaborative tasks.

Free-Text Rationales Recent advances in instruction-tuned language models have enabled the use of free-text rationales (FTRs) to guide model behavior across a range of tasks. These rationales make explicit the reasoning steps behind a decision or classification, and have been used in domains like argument mining, abductive reasoning, and coreference annotation [95–98]. Chain-of-thought prompting has emerged as a popular method for leveraging this capacity, helping models articulate intermediate reasoning rather than relying solely on direct prediction [99–101]. In our work, FTRs are used to enhance the annotation of deliberation chains in collaborative dialogue by helping surface the reasoning behind why a particular probe is causally linked to earlier turns.

4.3 Problem Formulation

Segmented Discourse Representation Theory (SDRT) posits that interpreting an utterance involves supplementing its semantics with pragmatic content based on the demands of *discourse coherence* [102]. The relation between utterances and prior content required for a full interpretation gives rise to structures which in collaborative dialogues represent the evolution of information

that propels such dialogues towards task-completion [77]. Let us define the relevant structures below:

Definition 1. Based on [78], let $\mathcal{G} = (\mathcal{V}, \mathcal{E}_1, \mathcal{E}_2, \lambda)$ be a **deliberation graph** in a collaborative dialogue, that in turn comprises sets of individual deliberation chains. $\mathcal{G}$ is characterized as a weakly-connected, weighted, acyclic graph. Here, $\mathcal{V}$ represents vertices for probing ($\mathcal{P}$) and causal ($\mathcal{C}$) interventions[1]; edges $\mathcal{E}_1$ denotes connectivity between vertices; weights $\mathcal{E}_2$ indicate causal influence from $\mathcal{C}$ to $\mathcal{P}$, thereby establishing a total order; and λ is a directed path induction function over $\mathcal{E}_2$ and a vertex $v \in \mathcal{V}$ that emits the root intervention $\mathcal{C}$ and terminal intervention $\mathcal{P}$ in $\mathcal{G}$, implicit in the discourse's linear order.

Definition 2. Given a deliberation graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}_1, \mathcal{E}_2, \lambda)$, a **deliberation chain** (or *intervention cluster*[2]) is a subgraph $\mathcal{G}' = (\mathcal{V}', \mathcal{E}_1', \mathcal{E}_2', \lambda)$ of $\mathcal{G}$, such that $\{\mathcal{P}_{\hat{\imath}}, \mathcal{C}_{\hat{\jmath}}\} \subseteq \mathcal{V}'$, where $\hat{\jmath} = \min\{j \mid \lambda(\mathcal{C}_j) \in \mathcal{V}'\}$ and $\hat{\imath} = \max\{i \mid \lambda(\mathcal{P}_i) \in \mathcal{V}'\}$ indicate the initial and final occurrences respectively in the traversal of $\mathcal{G}$ from $\mathcal{C}_{\hat{\jmath}}$ to $\mathcal{P}_{\hat{\imath}}$. See Fig. 5.7.

We formulate deliberation chain construction as a coreference resolution-style clustering problem [104, 105], over a dialogue, D, with N utterances, that the system must cluster into probing interventions and their linked causes, such that each cluster forms a unique deliberation chain. Given the elements of a cluster, λ reconstructs the chain by enforcing transitive closure over the within-cluster links given the temporal order inherent in the discourse, under Definition 1 above. This formulation motivates our joint modeling approach, which is detailed in Sec. 5.4.

In Fig. 4.1, we provide a detailed example of a deliberation chain from our dataset. The causal interventions (*e.g., "You have to at least select either the letter A or card 4."*) and probing questions (*e.g., "Can you explain why?"*) form a structured sequence, where probing interventions are linked

[1]Past probing interventions ($\mathcal{P}_{<i}$) likely influence current and future ones ($\mathcal{P}_i$), ensuring weak connectivity, and any $\mathcal{P}$ cannot be the cause of its own $\mathcal{C}$, thereby guaranteeing acyclicity. This structure reflects the linear progression typical in turn-based dialogues. Potential non-linearities in multimodal contexts [78] largely do not affect the acyclic structure because multimodal channels tend to overlap rather than invert the linear order of dialogue entirely [103].

[2]We will use *deliberation chain* for the ordered sequence of interventions in a dialogue, and *intervention cluster* for the clusters output by our system. Both denote a chain of sequential interventions linked by transitive closure, similar to *entity clusters* in coreference literature.

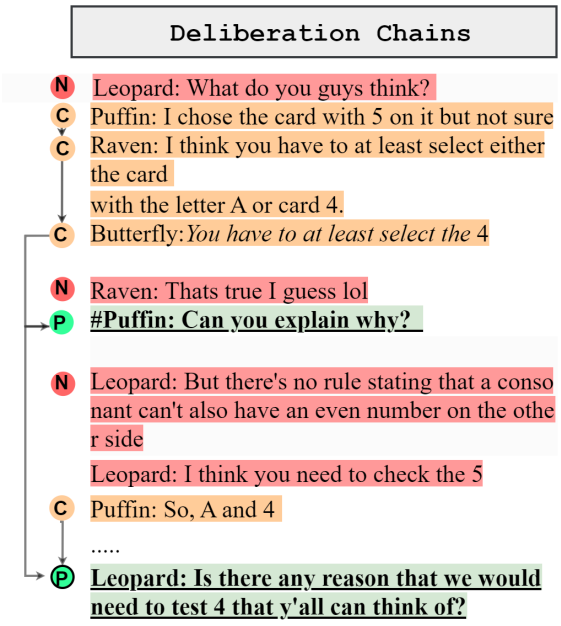

Figure 4.1: Example of a deliberation chain, showing the flow of interventions and their causal relationships within a collaborative task. This example is adapted from our model's output on the DeliData corpus.

to their causal antecedents. This transitive closure forms the deliberation chain, which reflects how participants navigate the problem-solving process.

4.4 Dataset Annotation

We evaluate intervention clustering on two recent, challenging collaborative dialogue datasets: DeliData and the Weights Task Dataset.

4.4.1 DeliData

The DeliData corpus [77] is a publicly-available dataset intended for studying group deliberation in multiparty problem-solving. It comprises 500 group dialogues, totaling 14,003 utterances, centered around the Wason card selection task, a well-established cognitive puzzle [106]. Each group contains 5 participants, who are presented with 4 cards that have a number or a letter on them. They must collectively decide which cards to turn over to test the rule, "All cards with

vowels on one side have an even number on the other?" The dataset includes both the dialogues themselves, which denote cards by the symbols on them (letters or numbers), and a measure of decision correctness (task performance) before and after the group discussion, and is annotated with deliberation cues, argumentation structures, and other conversational dynamics. DeliData splits consist of 300, 100, and 100 randomly-chosen groups for training, development, and testing, respectively.

4.4.2 Weights Task Dataset

The Weights Task Dataset (WTD) [107] is an anonymized publicly-available dataset intended for studying small group collaboration. It comprises 10 videos, where groups of three participants must use a balance scale to identify the weights of differently-colored weighted blocks and the pattern that describes the weights. The task unfolds in 3 stages, where users solve the problem with the scale, without the scale, and with inferred knowledge of the pattern in weights. The dataset includes multiple annotations, including human gold-standard transcriptions of the participants' dialogues. Utterances reference blocks by color and deduced candidate weights, and can be used to identify probing questions and their potential causal interventions. WTD splits consist of 7, 1, and 2 randomly-chosen groups for training, development, and testing, respectively.

4.4.3 Data Augmentation of WTD

The WTD is a multimodal dataset, but as the focus of this paper is establishing this novel task, our current study does not incorporate non-verbal cues. Instead, we employ *dense paraphrasing* [108] as an augmentation technique to explicitly define which blocks are being referred to in the situated dialogue, so that probing and causal interventions can be modeled using just a textual signal. The WTD annotations include dense paraphrased utterances for the first stage but not the second two. We followed the procedure from [80] to dense paraphrase the remainder of the dataset (e.g., replacing "those" with "red block and blue block" in cases where the video makes clear that those blocks are the intended denotata). Utterances were dually annotated (Cohen's $\kappa = 0.69$) and adjudicated by an expert.

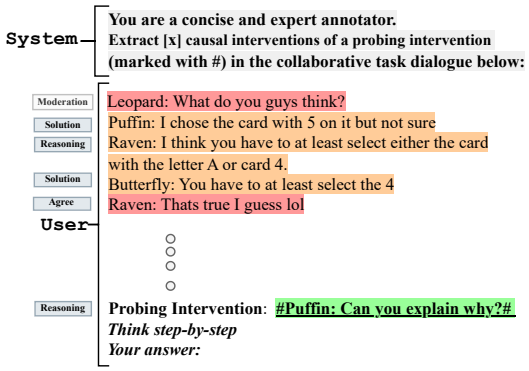

Figure 4.2: Prompting framework for GPT to select causal interventions given a probing intervention and a dialogue history (example from DeliData). Ground-truth labels for probing and causal interventions are marked in green and brown, respectively.

4.4.4 GPT Annotations of Deliberation Chains

Like coreference cluster annotation, which often requires exhaustive cross-comparisons across tokens [109], human annotation of deliberation chains is time-consuming and expensive. Therefore, to create "gold" chains for fair comparison, we draw on work in LLM-augmented annotations with Chain-of-Thought (COT) reasoning [97, 101, 110] for "soft" gold labels.

We apply a two-pronged strategy.

1. We sequentially prompt `GPT-3.5-turbo-0125` using an argument-extraction framework [95] (see Fig. 4.2) to extract causal interventions for all probing interventions in the data with prior dialogue history

   As such, we do not omit probing labels from the previous utterances given as context. and a system-based task-description to guide its reasoning. We also explicitly ask the LLM to generate free-text rationales (FTRs) corresponding to every causal intervention extracted, to augment its reasoning [111, 112].

2. We do an extensive human evaluation of these LLM-generated annotations to validate quality of extracted clusters. FTRs were used as an additional reference for human evaluators to validate the GPT's annotations and their alignment with human reasoning. This evaluation demonstrated high acceptability of GPT labels and reasoning to humans (see 4.4.5 for details).

4.4.5 Human Evaluation of GPT-Annotated Labels

We conducted a human evaluation to assess the quality of the GPT-generated annotations on a random representative subset of 25 samples from both DeliData and WTD test sets. These samples were evaluated across several dimensions: relevance, presence in sequence, information sufficiency, acceptability, and rationale overlap.

The annotators consistently agreed that the annotated utterances were indeed causal to the probing utterance, as indicated by high agreement on the first two questions concerning *Relevance to Context* and *Presence in Sequences*. These are the most critical aspects of the evaluation, and the high level of agreement demonstrates that the core annotations were valid. The annotators' answers to questions concerning rationale alignment, however, showed more variability, as expected and seen in Fig. 4.3. While annotators may agree that an utterance is causal, they may align less with the specifics of the rationale behind why it is causal. This variation is natural and does not impact the overall validity of the annotations.

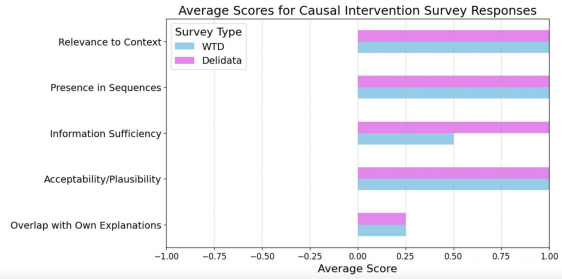

Figure 4.3: Average Scores for Causal Intervention Survey Responses.

We calculated Krippendorff's alpha to measure inter-annotator agreement. Each unique qualitative response was mapped to distinct numerical categories (e.g., Yes, No, Not enough information, Enough information) to capture the differences between responses more effectively. This calculation resulted in Krippendorff's alpha values of 0.88 for DeliData and 0.92 for WTD, indicating strong agreement between annotators on these samples.

4.5 Joint Learning of Deliberation Chains

To automatically cluster interventions that form a deliberation chain $\mathcal{G}'$, a model must learn to assign, for each possible $\mathcal{P}_i$, the most suitable antecedent utterance $\mathcal{C}_j$, that forms a correct link in the chain. Prior works in coreference resolution [89, 90] typically addressed such assignments using joint-learning frameworks that exhaustively score antecedent "spans" and thereby produce coreference chains. Our approach implicitly produces the correct chain since interventions in a dialogue follow a linear order assuming transitivity across links.

Standard joint-learning frameworks for coreference resolution typically operate at the *span*-level. For our task, where the entire deliberative utterance forms a distinct discourse unit [78], this is an incompatible approach. As such, we propose a joint-learning framework that models the task as a conditional probability distribution $Pr(P, C, L \mid D)$, partitioned into multinomial probabilities, assuming that utterance spans are conditionally independent given the dialogue D. Mathematically,

$$Pr(P, C, L \mid D) = \prod_{i=1}^{N} \prod_{j=1}^{N} Pr(p_i \mid D) Pr(c_j \mid D) Pr(l_{ij} \mid D), \tag{4.1}$$

where P refers to the probability of an utterance being a Probing intervention, C refers to the probability of an utterance being a Causal intervention, and L refers to the probability of a Link between the two utterances. p_i, c_j and l_{ij} are treated as random variables denoting the probabilities of an utterance being probing, being causal, and of the link between the two interventions, respectively; N denotes the number of individual utterances within a dialogue D.

4.5.1 Model

Intervention Pair Representation As the right-hand side of Eq. 4.1 represents causal dynamics as probabilities of links between pairs of utterances in the discourse, we draw on a cross-encoding strategy from coreference research [2, 91, 113] to score pairs of utterances. Since some dialogues, especially in the Weights Task Dataset, can reach up to ∼200 utterances, we use the Longformer

model [114] as the base encoder. To construct an expressive representation for a pair of interventions ($\mathcal{P}_i$, $\mathcal{C}_j$), we first demarcate their start and end with special tokens (<m> and </m>). For context around a probing intervention, we also concatenate the k previous utterances[3] along with participant name or number as given in the dataset. We extract the [CLS] token representation of this concatenated input, the cross-attentional context of $\mathcal{P}_i$ and $\mathcal{C}_j$, as well as their Hadamard product, $\mathcal{P}_i \odot \mathcal{C}_j$. This results in a combined vector representation for pair ($\mathcal{P}_i$, $\mathcal{C}_j$):

$$V(\mathcal{P}_i, \mathcal{C}_j) = [V_{CLS}, V_{\mathcal{P}_i}, V_{\mathcal{C}_j}, V_{\mathcal{P}_i} \odot V_{\mathcal{C}_j}] \tag{4.2}$$

Next, to maximize the log-likelihood in our joint-learning framework (Eq. 4.1), we generate three sets of scores from specific segments of Eq. 4.2 using three feed-forward neural networks (FFNN):

- a linking score $l_{ij} = \text{FFNN}_l(V(\mathcal{P}_i, \mathcal{C}_j))$, the probability of a pair of utterances forming a true link; and
- two intervention scores, $s_i = \text{FFNN}_p(V_{\mathcal{P}_i})$ and $s_j = \text{FFNN}_c(V_{\mathcal{C}_j})$ of the candidate and the antecedent, respectively, being valid interventions.

Thus, the model picks up on two types of learning signals: correctly assigning a true antecedent to a candidate intervention while also learning what constitutes a valid intervention. We directly optimize the model with $\mathcal{L}_{\text{joint}}$:

$$\mathcal{L}_{\text{joint}} = \alpha_{\text{p}}\mathcal{L}_{\text{probing}} + \alpha_{\text{c}}\mathcal{L}_{\text{causal}} + \alpha_{\text{l}}\mathcal{L}_{\text{link}} \tag{4.3}$$

that consists of a weighted-combination of three separate loss terms. $\mathcal{L}_{\text{probing}}$ and $\mathcal{L}_{\text{causal}}$ are each defined as:

$$\mathcal{L}_{\text{[probing,causal]}}(*) = -\sum_{*=1}^{N} y_* \log(\sigma(s_*)) \tag{4.4}$$

[3]Setting $k = 10$ and max sequence length (probing intervention with preceding utterances) to 512 was empirically found to cross-encode both utterances in a pair, on average, without losing expressive tokens or incurring inordinate compute cost.

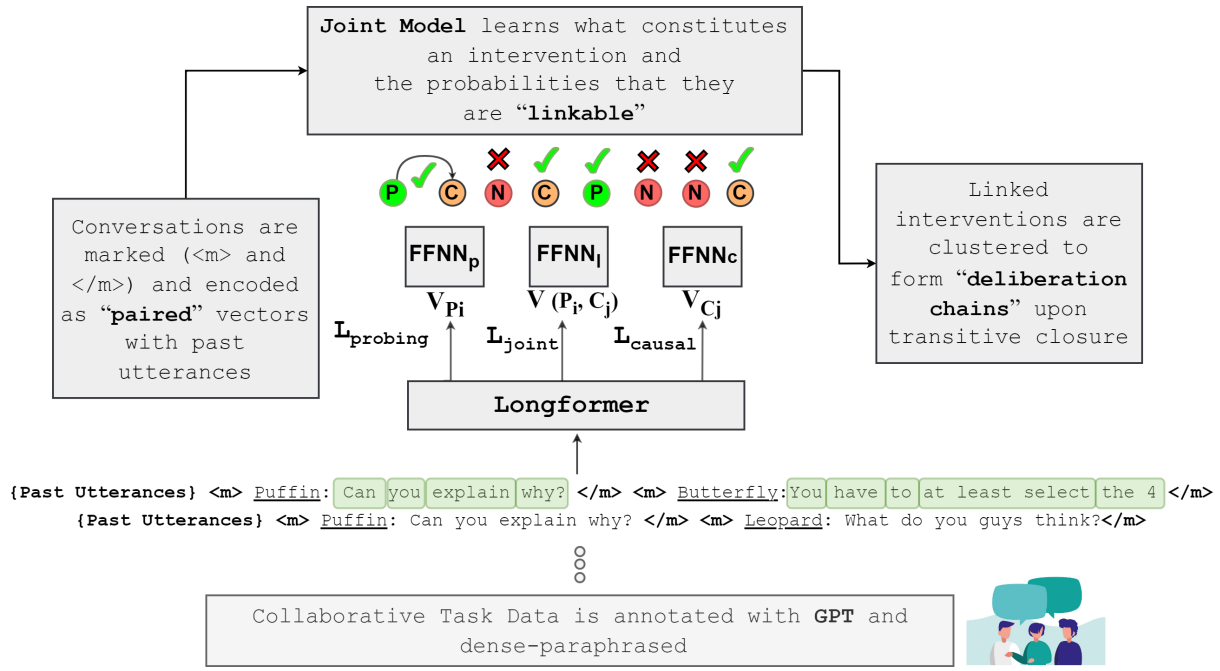

Figure 4.4: Our joint-learning framework for *deliberation chains*, learning to assign correct antecedent utterances for every valid intervention using a "probing" score, a "causal" score, and a "linking" score. Pairs of utterances are encoded with global attention (in green between <m> and </m>), further contextualized by past utterances.

where $*$ corresponds to i and j in $\mathcal{L}_{\text{probing}}$ and $\mathcal{L}_{\text{causal}}$, respectively, σ is the sigmoid function, and y is the predicted output. The final term is $\mathcal{L}_{\text{link}}$:

Training Pair Generation For training a pairwise scorer model, an efficient pair generation process is crucial. A naive way to implement Eq. 4.1 compares each utterance u_i to the set of all its preceding antecedents $U(i) = \{\epsilon,\ u_1, \ldots, u_{i-1}\}$ to generate pairwise scores.[4] This results in $\sim O(N^2)$ complexity for a dialogue of N utterances. Discourse-coherence theory [93, 115] suggests that the most pertinent information to a specific utterance remain *within* an "attentional state", i.e., the point of focus of participants within a dialogue. As such, given a dialogue of N utterances, for each target u_i, we define a window W of previous utterances considered for training.

[4]Our training method is generalizable to all utterances, since the ground truth label on any candidate can be causal, probing, or *neither* (a non-intervention dummy variable, ϵ). Generated pairs may have true labels that are any combination of probing and causal, since two causal interventions may be linked to the same probing intervention, or two probing interventions may share a cause, which results in these pairs themselves being linked under transitive closure. This follows standard practice in pairwise approaches to coreference across long documents.

Because of the long tail of true negative samples (non-links), this value is tuned over the dev split of each dataset to make the ratio of positive to negative samples more balanced (cf. [2] for optimal training.). Given a true intervention cluster after annotation and labeling, all pairs within it are considered positive pairs. Negatives comprise all other pairs under consideration (which may be limited by window W).

During training, the model is forced to learn discourse-relevant signals from the positive pairs drawn from true intervention clusters. Applying Longformer's *global attention* to *all* tokens in the pair (Fig. 5.7) allows us to encode relevant global features within W. Utterances in the preceding neighborhood W typically display lexical overlap for items with similar semantic roles, or task-specific phrases.[5] When such pairs are sourced from separate intervention clusters that occur within W, they naturally form difficult samples for encoder-only LLMs like Longformer due to misleading lexical overlap [2, 112].

Inference We evaluate two inference strategies. For our naive approach, we relax W and generate all candidate antecedent utterances within D, score them using using the intervention scores (mean of s_i and s_j), and only keep the remaining pairs based on a threshold τ

This reduces cross-comparisons in building the intervention clusters as we only use the pairwise scorer to score the remaining utterances. We also consider all scores generated without relaxing W. While the naive approach tests the system's recall under a long-tail of true-negatives, this method enforces a more balanced distribution, resulting in a "soft" upper-bound on model precision. Pairwise scoring generates an adjacency matrix of links between utterances. Inducing transitivity between links using a connected-components based clustering method with a threshold of 0.5 generates the final intervention clusters. Under temporal ordering, these expand to deliberation chains within a dialogue.

[5]For instance, in the Weights Task, neighboring utterances contain overlapping arguments like "red block" when the group is solving a particular subtask relevant to that block.

4.6 Experiments

We evaluated our joint modeling method against 3 similarity baselines and two cross-encoder methods adapted from coreference research.

4.6.1 Similarity Baselines

For simple similarity baselines, we assessed:

- Simple token overlap between utterances. This may indicate correspondence between a probing intervention and its cause(s), as the utterances may share terms. To assess lexical similarity between utterance pairs, we calculated the Levenshtein distance ratio (0–100) between the two strings.
- The overlap of salient *entities* within the utterances. We computed an Intersection over Union (IoU) of entity counts score based on categorical features derived from task-relevant categories referenced in each utterance (i.e., vowels, consonants, even and odd numbers in DeliData, and colors and weights in WTD).
- Cosine similarities between embeddings of the two utterances, extracted from BERT-`base-uncased`, following the intuition that probing utterances should share some *semantic*, not just token or entity similarity [116] with their causal counterparts.

For each, we set a threshold value for each dataset, equal to the average of the relevant metric over the dev set. If the relevant metric for a test pair exceeded this calculated threshold for the dataset, we linked that pair.

4.6.2 Cross-Encoder Baselines

For trainable baselines, we specifically chose recent coreference resolution frameworks that operate on an "utterance" level (instead of a span-level) for a valid comparison (see Sec. 5.4). For fairness, we used the base encoders from these frameworks as well as with their cross-encoding strategies, but not their fine-tuned weights, since fine-tuning on a separate task can likely tilt the model out of distribution [117].

We used [1]'s Cross-Document Language Model (CDLM). with a context length of 1,024 preceding tokens along with their cross-encoding setup.[6] We also employed [2]'s "bidirectional" BCE loss-based learning method. This generates a mean of the BCE losses over the forward pass of utterances paired in both directions: (u_i, u_j and u_j, u_i). Like our joint modeling approach, the context window here is 512 tokens.

4.6.3 Joint Modeling Hyperparameters

For joint modeling, we use the Adam [118] optimizer with batch size 24, with learning rates of $1e-6$ for the encoder fine-tuning, $1e-4$ for the pairwise scorers, and $1e-5$ for the intervention scorers. Each training epoch on an NVIDIA A100 took $\sim$20 and $\sim$40 minutes for DeliData and WTD, respectively. Each model was evaluated after a single training run for 16 epochs after robust hyperparameter tuning on the validation sets.

4.7 Results

	DeliData				WTD			
	B^3			CoNLL	B^3			CoNLL
	R	P	F_1	F_1	R	P	F_1	F_1
Lexical Overlap	26.6	81.3	40.0	28.6	41.6	50.0	45.4	36.6
Entity Overlap	34.9	71.7	46.9	40.6	27.2	70.0	39.2	26.7
BERT-Cosine	98.6	49.9	66.3	69.2	**100.0**	7.1	13.2	35.3
LongContext	84.7	60.7	70.7	68.2	72.1	23.8	35.8	45.5
Bidirectional	90.8	59.2	71.7	70.9	64.5	31.5	42.4	44.3
LLaMA 2-7B-chat	**99.9**	49.7	66.4	69.7	**100.0**	7.1	13.2	35.3
— Ours (Joint - W)	92.3	60.5	73.1	73.6	54.4	**75.0**	63.0	50.3
— Ours (Joint + W)	87.8	**72.6**	**79.5**	**76.4**	67.9	61.7	**64.7**	**58.1**

Table 4.1: B^3 and CoNLL F_1 metrics on DeliData and WTD test set results. "LongContext" denotes [1]'s coreference methodology applied to deliberation chain clustering. "Bidirectional" denotes [2]'s methodology.

[6]CDLM (https://huggingface.co/biu-nlp/cdlm) is trained on documents with overlapping information and is suitable for handling long inputs, which are both traits of our dialogues (Sec. 4.5.1). For compute reasons, we truncate pairs at a maximum sequence length of 1,024 tokens after tokenization since the token-length of utterance pairs in training is $\sim$220 tokens for both datasets, on average.

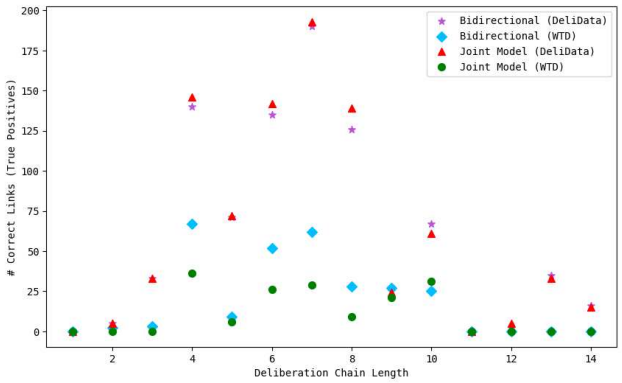

Figure 4.5: Cluster-level distribution of correctly assigned intervention links for the best-performing cross-encoder baseline compared to Joint - W on both datasets.

We evaluate against coreference methodology using cluster metrics computed using the CoVal coreference scorer [119], specifically B^3 and CoNLL F_1 metrics, as presented for both datasets in Table 4.1.[7] We also present zero-shot results from LLaMA 2-7B-chat.

The multimodal nature of the WTD likely makes it more challenging than DeliData due to cues that may be missed in even the dense paraphrased language. The use of the windowed approach results in a small performance improvement due to the exclusion of false positive links outside W. The BERT-Cosine and LLaMA 2 zero-shot baselines perform extremely similarly (returning identical metric values on WTD) and achieve perfect or near-perfect recall. This is likely due to these methods returning a very high proportion of false positive links and transitive closure subsequently clustering (almost) all interventions in a dialogue.

4.8 Discussion

Quantitative Analysis Fig. 4.5 shows the count of correct links between interventions assigned by the bidirectional baseline and our non-windowed joint model for each cluster size.[8] In DeliData, which has longer and more diverse chains, the joint model consistently links more pairs correctly in frequent mid-sized clusters. In contrast, the joint model is more conservative but more

[7]Since we are using the gold intervention labels for our experiments, using B^3 is more reliable compared to other metrics [93, 120].

[8]Only the non-windowed model results in a full comparison to all other baselines because Joint + W does not consider all gold pairs.

precise on WTD, with a $\sim$45-point increase in B^3 precision over the bidirectional model. These patterns suggest that the joint model forms a more accurate and globally coherent representation of deliberation chains, particularly in domains with long or noisy discourse histories.

	Dialogue	Free-Text Rationale(s)
(a)	**[C1] Emu: I picked the card with the vowel A on it, because the rule said all cards with vowels on one side will have an even number on the other** [C2] Koala: I think it is A and 2 [C3] Hamster: I agree ... **[P4] Bee: So are we ready to final submit**	[C1] *"Emu's statement directly relates to the reasoning behind choosing the card with the vowel A, which is crucial in the decision-making process."*
(b)	[C1] Narwhal: What card did you think needed to be turned? ... [C2] Guinea pig: I picked 6 and U ... **[C3] Kiwi: We need to pick one that wouldn't fit the rule to test it. Maybe?** ... **[P4] Kiwi: 7 and U?**	[C3] *"This statement hints at the strategy of testing a card that would break the rule to confirm its validity, indicating a shift in the participant's thought process during the discussion."*
(c)	[C1] Participant 2: Oh maybe I'll try holding it here ... **[C2] Participant 2: Mystery block, blue block, red block, green block, purple block, yellow block kinda feels the same** ... **[C3] Participant 1: So how about purple block, green block two, I had eh purple block, yellow block two** ... **[P4] Participant 2: Is there a better way to measure mystery block?**	[C2] *"This utterance indicates the participant's initial attempts to compare the weights of various blocks using their fingers, setting the groundwork for exploring different measurement techniques."* [C3] *"This utterance directly led to the probing question as it involved a new approach of grouping blocks on fingers to measure their weights."*

Table 4.2: Test samples from DeliData (a & b) and WTD (c). Bolded utterances indicate $(\mathcal{P}, \mathcal{C})$ pairs that our method (Joint - W) linked correctly and all other methods failed to. FTRs are given for the annotation of the indicated utterance as causal. These are not included in the input for inference, but are provided as indicators of the kinds of information our framework is likely to learn from the labels that were created using this COT-guided process.

Qualitative Analysis Table 4.2 highlights test cases where the non-windowed joint model successfully predicted causal-probing links that all baselines missed. These examples, taken from both datasets, show that our model captures both thematic and referential coherence over longer distances in the dialogue. In several cases, the model linked utterances that formed coherent logical

structures (e.g., premise-question-confirmation) even when the specific content was ambiguous or distributed across multiple turns.

1. In `DeliData`, our model correctly links P4 to C1, which shares references to "letter A," vowels, and even numbers. C2 summarizes what participants agree on. P4 prompts clarification on all these points, and our model appropriately identifies the causal tie to the relevant earlier utterance.

2. The FTR in another example captures a shift in Kiwi's reasoning; our model recognizes this shift and links two same-speaker utterances (C3 and P4), highlighting its ability to recover intra-speaker deliberation chains.

3. In `WTD`, our model links P4 with both C2 and C3, which pertain to measurement strategies, even though that connection is only evident when viewed alongside the FTR's summary. This indicates that our model benefits from the richer annotation context introduced during GPT-based labeling.

These examples also show that our model successfully links utterances that are far apart in the dialogue—much further back than even the longer-context CDLM baseline. This ability to recover long-range dependencies is crucial for modeling deliberation, where causal relationships are often distributed across time and participants.

Modelling Delibration This chapter continues the thesis's broader aim of understanding how abstract phenomena like reason becomes visible. Rather than treating probing as an isolated event, we modeled it as emerging from structured discourse patterns, identifying utterances that causally lead to a probe. In doing so, we move from event classification to sequence-level interpretation, capturing how understanding is built collaboratively.

This modeling opens possibilities for tracking reasoning as it unfolds in group settings. Much like formal logic, where conclusions arise from layered premises, group reasoning involves accumulations of justifications, corrections, and challenges. Our ability to recover these deliberation

chains computationally allows us to ask: why was this question asked now? What led up to this doubt? Where did the group shift direction? These are not trivial questions—and yet, as our models show, patterns in language contain answers.

Beyond academic inquiry, such models can have impact in applied contexts. In collaborative learning environments, deliberation chains can reveal gaps in understanding or moments of conceptual breakthroughs. In decision-support systems, they can provide explanations for how a team arrived at a conclusion. By grounding these signals in naturalistic interaction data, our work brings us closer to AI systems that don't just classify behaviors but interpret the structure of collaborative thought.

Ethical Considerations The deployment of systems that model deliberation and probing must be handled with care. In educational contexts, for example, models may mischaracterize students who do not conform to normative communication styles, potentially leading to unjust assessments. We explicitly advocate for these models to augment—not replace—human judgment, especially in high-stakes environments like classrooms.

The multimodal nature of datasets like `WTD` also raises concerns around surveillance. Since deliberation modeling involves analyzing linguistic, behavioral, and sometimes visual patterns, safeguards must be put in place to ensure ethical usage, especially in real-time or personalized settings. All data used in this work were anonymized and collected under IRB-reviewed protocols, but future applications will require continued oversight to prevent misuse.

Finally, models that identify deliberation chains could be co-opted by malicious systems—for example, agents that deliberately disrupt collaborative reasoning by introducing artificial "friction." It is critical to separate the computational modeling of discourse from its normative interpretation or prescriptive use. Our work provides a framework, not a judgment system.

Limitations This work has several limitations. While GPT-based annotation allowed us to scale deliberation chain labeling across datasets, such annotations may introduce biases or inconsistencies not present in human-labeled data. Although our annotations were validated with human

judgments, future work should develop fully gold-standard datasets or directly compare model performance on human vs. AI-generated labels.

Additionally, coreference metrics such as B^3 and CoNLL F1 were adapted to evaluate cluster recovery in deliberation chains. While useful, these metrics may not fully capture the nuance of deliberative structure. Prior work [120] highlights the limitations of these metrics even in standard coreference tasks; future work should investigate whether new metrics are needed to evaluate deliberation more accurately.

Finally, while our model assumes access to gold-standard probe and causal transcriptions, real-world deployments must contend with noisy ASR, segmentation errors, and online inference constraints — in Chapter 6, we explore how these realities impact performance. Extending the system to jointly detect and link probing utterances—especially in multimodal or streaming contexts—remains an important next step.

Future Work Our current model treats probes as given; future systems should identify both the probe and its causes jointly. This would enable real-time support systems that monitor emerging questions in dialogue as they happen. Another avenue is adapting the model for use in live collaborative tasks, integrating it into intelligent tutoring systems to surface reasoning gaps or prompting students to elaborate based on past utterances.

The multimodal aspect of the `WTD` dataset also opens new research directions. For example, gestures or physical actions (e.g., pointing to blocks) might themselves be part of a deliberation chain. Integrating nonverbal modalities into causal linking models could significantly enhance interpretability. Prior frameworks like that of [121] offer promising directions for incorporating gesture and speech into a unified model of deliberation.

Our task can also be generalized to other domains of interaction, such as legal argumentation, policy debates, or collaborative design. In each case, the ability to surface causal chains of reasoning can support reflection, evaluation, or intervention.

Conclusion In this chapter, we introduced a novel task—deliberation chain construction—and proposed a span-based model to detect causal links between utterances in collaborative dialogue. Our approach outperformed coreference-inspired baselines across two datasets and revealed deeper insights into how group reasoning unfolds. By capturing traces of reasoning, we move toward AI systems that can interpret the dynamics of group collaborative reasoning.

It is not enough to trace how participants justify claims—we must also examine what they believe to be true. After all, reasoning is often a response to conflicting or evolving beliefs. In the next chapter, we shift focus to modeling these belief states directly by extracting structured propositions from dialogue. This allows us to trace how individual and shared beliefs surface, shift, and evolve.

Chapter 5

Extracting Propositional Knowledge from Dialogue

5.1 Introduction

In collaborative problem solving, individuals must not only listen and deliberate, they must also contribute what they believe to be true. These moments of articulation, where a participant asserts a claim about the task, are foundational to shared understanding. Whether explicitly stated or implicitly assumed, these beliefs, when voiced, form the building blocks of group reasoning.

This chapter focuses on modeling the human phenomenon of belief, expressed naturally in group dialogue. While the previous chapter explored how probing questions emerge through deliberation, here we turn to what is being reasoned over: the task-relevant propositions that participants offer as facts, inferences, or assumptions. These propositions reveal what each participant believes about the problem and, collectively, they define the evolving state of group knowledge. For computer-assisted education, this capability is especially crucial. A key goal of intelligent systems is to determine what students know, infer, or understand in the course of a task or assignment, not simply based on answers, but based on the reasoning and articulation that occurs in dialogue. In naturalistic settings like small-group classrooms, however, such expressions are often messy. Utterances are fragmented, overlapping, and grounded in shared physical or visual context. Compared to written responses or formal dialogue systems, natural speech makes it far more difficult to isolate the propositional content of any single utterance.

This challenge is compounded in tasks like knowledge tracing [122], where the same belief may be expressed in radically different ways. Real-world language includes repetition, filler, disfluency, and syntactic variation—all of which can obscure meaning. Extracting propositions despite this variability is critical if automated systems are to make reliable inferences about student understanding.

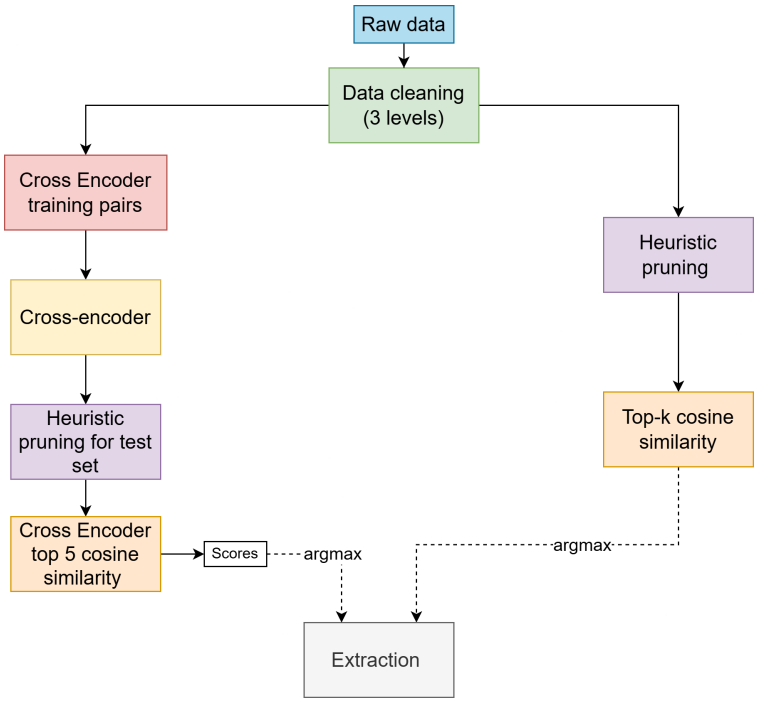

Figure 5.1: Schematic overview of the two methods used for propositional extraction. The process begins with Raw Data, which undergoes Data Cleaning across three levels to filter out irrelevant utterances. In the first method, a Cross-Encoder is trained on utterance-proposition pairs, followed by Heuristic Pruning for the test set, and outputs the top-5 candidate propositions using cosine similarity obtained from the trained Cross Encoder. In the second method, Top-k Cosine Similarity is directly applied to the heuristically pruned candidate propositions. The final Extraction step selects the best proposition using argmax over the similarity scores. Dashed lines indicate the selection process for the final proposition, while color coding differentiates key components of the pipeline.

The propositional content that students assert is critical to tracking the collaborative process as students share their understanding and build consensus or common ground [80, 123]. For example, an automated agent for collaborative problem solving support would need to track surfaced propositions as a measure of task progression. Additionally, students in collaborative settings achieve better learning outcomes when they engage in *leading* the discussion, which involves making new claims and not simply reiterating previously-stated information [124]. The ability to extract propositional content from dialogue provides a way for an agent to determine whether a claim was already stated within the group. This would provide a necessary feature to determine whether a student is helping to lead the task forward, thereby enabling better prediction of learning outcomes from mined data.

We take the transcribed utterances of a shared collaborative task, which are annotated with ground truth task-relevant propositions that are expressed therein, and use cosine similarity and cross-encoder methods to extract the propositions from the utterance text. Fig. 5.1 shows a schematic overview of our approach. We also extend our methods to utterances automatically segmented and transcribed by Google Cloud Platform's Automated Speech Recognition, showing how our propositional extraction methods may be incorporated into an automated system with a relatively low level of degradation due to automated transcription. Further, we explore the use of synthetic data augmentation using large language models (LLMs) to improve the robustness of our models and investigate the zero-shot inference capabilities of GPT-4 and LLaMA 2. To guide this work, we focus on several key research questions: **1)** How accurately can task-relevant propositions be extracted from naturalistic dialogues? **2)** How do different extraction methods compare in performance, and how does automated transcription affect this accuracy? **3)** Can synthetic data and zero-shot approaches enhance performance?

Our novel contributions are grouped into two key areas:

- **Advancing Propositional Extraction Methods:** We establish a novel, challenging task of propositional extraction from natural speech during collaborative interactions. We compare cosine similarity and cross-encoder methods, using multiple language models and different levels of data cleaning, establishing new baselines and theoretical upper bounds for this task. Additionally, we explore synthetic data generation techniques to enhance model performance and assess zero-shot inference capabilities using large language models such as GPT-4 and LLaMA 2.

- **Practical Implementation and Real-Time Feasibility:** We assess the impact of automated speech transcription on extraction performance, demonstrating a relatively low level of performance degradation compared to manually transcribed utterances.

5.2 Background and Related Work

Collaborative tasks concern the construction and maintenance of a shared conception of the problem at hand [125], involving mutual engagement and coordinated effort to solve the problem together. Within such a framework, especially one centered around shared synchronous tasks, quantity of specific propositions discussed has been shown to be a significant predictor of learning gains [126]. Therefore, propositional extraction serves an important role in automated analysis of shared task data in an educational context, or for an automated system to make inferences about construction of shared knowledge in real time.

Propositional Extraction Prior work on propositional extraction from natural language has primarily been conducted from written texts in domains such as question answering, where early methods relied on approaches such as semantic memory [127]. Classical machine learning approaches like support vector machines have been applied to opinion mining to find "propositional opinions," or sentence fragments that contain the object of an assertion, incorporating word and feature-level knowledge from resources like WordNet, FrameNet, and PropBank [128]. Linguistic features have even been used to extract "ideas" from transcribed speech in the clinical domain, as a technique to predict Alzheimer's disease and other types of cognitive decline [129]. These early works not only show the utility of propositional extraction in various domains, but also demonstrate the relative sparsity of study on this topic. With the advent of neural network methods for text processing, these have been applied to NLP problems like propositional extraction from argumentation and rhetoric [130, 131]. These approaches include reported speech, as may appear in documents such as news articles. To the best of our knowledge, we are the first to attempt a similar task on transcribed naturalistic speech data from a collaborative task setting reminiscent of small group work in classrooms. Successfully extracting the propositional content expressed by an utterance is critical to modeling the epistemic positioning of the speaker toward the proposition. In a group context, these two components are required to track the shared and divergent beliefs of the group as they pertain to a task, as a method of modeling task progress.

Pairwise Representation Learning All of the aforementioned approaches frame the problem as one of establishing a mutual relationship between a piece of text from a dataset and another piece of text from a library of candidates, be they ideas, opinions, or propositional information more generally. Pairwise representational learning techniques have long been popular in the deep learning community for learning such relationships between two pieces of text. While some previous works modeled these relationships for text-generation tasks like abstractive document summarization [132], machine-comprehension [133], or document-reconstruction [134], others have also explored pairwise learning to compute similarity metrics between pairs of documents [135–137] as well as for masked language modeling [138]. More recently, for clustering-related tasks like coreference resolution, a "cross-encoding" framework has been used to learn pairwise features of possible coreferent mentions [1, 2, 91–93, 139]. These works, originally inspired by [113], learn high-level semantic features of a mention (e.g., of an entity or event) within a sentence in the context of another mention-containing sentence and compute the coreference probabilities of such pairs before clustering mentions that refer to the same entity. We adopt this "cross-encoding" technique for both our candidate proposition generation procedure, as well as for calculating the probability of a given utterance referring to a candidate proposition.

Cross-Encoders According to discourse coherence theory, in a dialogue between two or more participants, the content of the discussion is essentially a subset of the common knowledge, beliefs, and common intention (goal) that each participant has at any given point. As such, certain processing decisions like identifying referring expressions or detecting common propositional content between utterances can be made locally within the "attentional state" of the discourse. Following discourse coherence theory [115, 140], a human reader of a text or listener of a dialogue will focus attention on only a small subset of the total possible complement of events and entities. For instance, in a collaborative problem-solving setting, the words in an utterance that any participant uses to describe a specific sub-task within the overall task are constrained by "discourse segment purpose" or their common intention at that specific point in the dialogue. This constraint in the appearance of utterances to maintain coherence in the collaborative problem-solving dia-

logue allows us to map an utterance to a proposition by focusing only on the *local* elements in the utterance/proposition pairs.

However, since linguistic constraints or rule-based heuristics used to determine this attentional state can be narrow in their scope or domain-specific, most previous works have modeled the attentional state using neural networks [93, 141, 142]. The neural model creates a latent representation or high-dimensional *embedding* of discourse-relevant entities *in context* and a variety of similarity methods (such as nearest neighbors or neural attention mechanisms [143]) may be used to determine which entities are subjects of the current attentional state given a context. These models are typically built on top of pre-trained transformer-based language models (LMs) [143] like RoBERTa or Longformer [114, 144] that are known to capture rich semantic features through their contextualized representations of tokens and sequences. Apart from computationally modeling the innate structural coherence in a discourse, these architectures can also generate potential referents by demarcating the attentional state within a dialogue, through context.

These works have focused on various natural language understanding tasks, including coreference resolution. Our task is adjacent to coreference resolution since we have to map a set of utterances to their corresponding propositions in a collaborative dialogue. As such, we take inspiration from the pairwise scorer/cross-encoder architecture commonly used as a pairwise representation learning framework in cross-document coreference resolution [1, 2, 91, 92, 101, 139, 145]. This method forces a classifier to learn a combined representation of one mention (represented by a trigger word) in the context of the other, both of which are encoded within their respective sentences. This learning strategy is an effective way to generate similarity scores between pairs of event or entity mentions due to the contextualized learning framework.

Toward real-world use of AI As AI performance has increased, more works have begun to investigate how various automated preprocessing methods impact downstream task performance, since imperfect data is inevitable in a real-world application [76]. Some of these efforts intentionally examine performance given imperfectly preprocessed data [146] while others have explicitly explored how various preprocessing techniques degrade performance [147, 148]. In particular, var-

ious recent works have explored how imperfect data corresponds with performance in small-group contexts [149–153]. These works have shown that the influence of automated but imperfect tools, e.g., automatic speech recognition (ASR) for transcription, do degrade performance but not catastrophically so. In this work, we also examine how data imperfection degrades performance on a novel task: propositional extraction from natural dialogue during a collaborative group task.

While prior work has successfully applied propositional extraction in domains such as argumentation, and rhetoric, to the best of our knowledge, our study is the first to tackle this task in the context of naturalistic, collaborative dialogue involving multimodal signals in real-time. Our approach uniquely focuses on extracting propositions from overlapping, co-situated speech, demonstrating its applicability to collaborative educational tasks, where tracking shared knowledge is critical. We also extend cross-encoder methodology, commonly applied in coreference resolution, to propositional extraction in natural speech dialogues.

5.3 Datasets

5.3.1 Weights Task Dataset

Figure 5.2: Example still from the Weights Task being performed. The utterance associated with this frame is "i guess green block is like twenty and red block, blue block is like ten and ten". This utterance expresses the proposition $green = 20 \wedge red = 10 \wedge blue = 10$.

The Weights Task [107] is a situated collaborative problem-solving (CPS) task wherein groups of three work together to deduce the weights of differently colored blocks using a balance scale. There are a total of 10 groups, resulting in approximately three hours of audiovisual data. Participants consented to the release of their likenesses for research purposes. The study protocol and release of A/V data were approved by the Colorado State University institutional review board.[9] In this work we focus on Phase 1 of the task, where the group has five blocks of different colors ($C = \{\text{red}, \text{yellow}, \text{green}, \text{blue}, \text{purple}\}$) whose weights follow an instance of the Fibonacci sequence ($W_n = \{10g, 10g, 20g, 30g, 50g\}$). At the start of the task, the group is told that the red block weighs 10 grams.[10]

The Weights Task Dataset (WTD) contains speech transcribed manually by humans (hereafter referred to as "Oracle" transcriptions) as well as speech transcribed automatically by Google Cloud Platform's Automatic Speech Recognizer (Google ASR). The Oracle and Google transcription processes also *segmented* the speech into utterances—a single person's continuous speech, delimited by silence. In the dataset, there are a total of 2,140 utterances that contain transcribed speech according to Oracle segmentation, and 1,500 utterances containing transcribed speech according to Google segmentation. Fig. 5.2 shows still frame of a group performing the task. Due to the overlapping nature of speech in this setting, utterance segmentation leads to many sentence fragments and overlaps, as well as mistranscription by the automated system, which leads to challenges in extracting the intended meaning behind any given utterance. An additional challenge to meaningful information extraction from the linguistic channel is that due to the multimodal nature of the task, a complete interpretation of an utterance may require recourse to another modality. For example, someone may say "this one" while pointing to a specific block. The pointing makes it clear which block is being referred to but without access to the video showing where the person is pointing, the language alone is ambiguous. The above factors enumerate just some of the challenges to extracting propositions expressed through dialogue in this setting.

[9]The dataset and consent documents associated with the original study protocol are publicly available at https://zenodo.org/records/10252341.

[10]Although a gram is a unit of mass, the colloquial dialogue in the dataset uses "mass" and "weight" interchangeably.

The propositions themselves are annotated in the context of the *common ground* that evolves between group members as the task proceeds, that is, the set of propositions Φ each individual comes to believe as factual and that the group must agree upon, implicitly or explicitly, to arrive at the goal [154]. In the case of the Weights Task, the participants must all arrive at the correct assignments of weight $w \in W$ to color $c \in C$ to solve the task. The WTD is annotated with the propositions that are asserted, evidenced, or agreed upon as the task unfolds, based upon the multiple modal channels and prior context. Our goal is to recover those propositions from the transcribed speech.

Within the dataset, there are 127 utterances which describe 46 unique propositions that were expressed during the task, with *yellow = 50* as the most frequent, appearing 17 times. The average frequency of each unique value is approximately 2.76. Proposition breakdowns show 107 instances containing a single proposition, 13 instances with two propositions, and 7 instances with more than two propositions. Among the operators, "$=$" is used exclusively in 95 instances, while "$\neq$" does not appear exclusively. Additional comparisons include "$>$" (14 instances) and "$<$" (12 instances), with 6 cases involving multiple operators. Notably, the top propositions occurring in the task are *yellow = 50*, *purple = 30*, *red = 10*, *green = 20*, *red = blue*, and *blue = 10*. These represent the correct weight assignments for each block, with other propositions involving various comparisons to derive the correct solution. A visualization of the distribution is provided in Fig. 5.3.

5.3.2 DeliData

DeliData [77] is a dataset designed for examining group deliberation in multi-party problem-solving tasks, using the Wason card selection task as the focal activity. The Wason card selection task is a well-established task used in psychology research to explore reasoning processes [155]. In this task, participants are presented with four cards, each with a number on one side and a letter on the other. The task is to determine which cards need to be turned over to test a rule, such as "All cards with vowels on one side have an even number on the other." This task is designed to reveal common reasoning errors, such as confirmation bias, where individuals might select cards

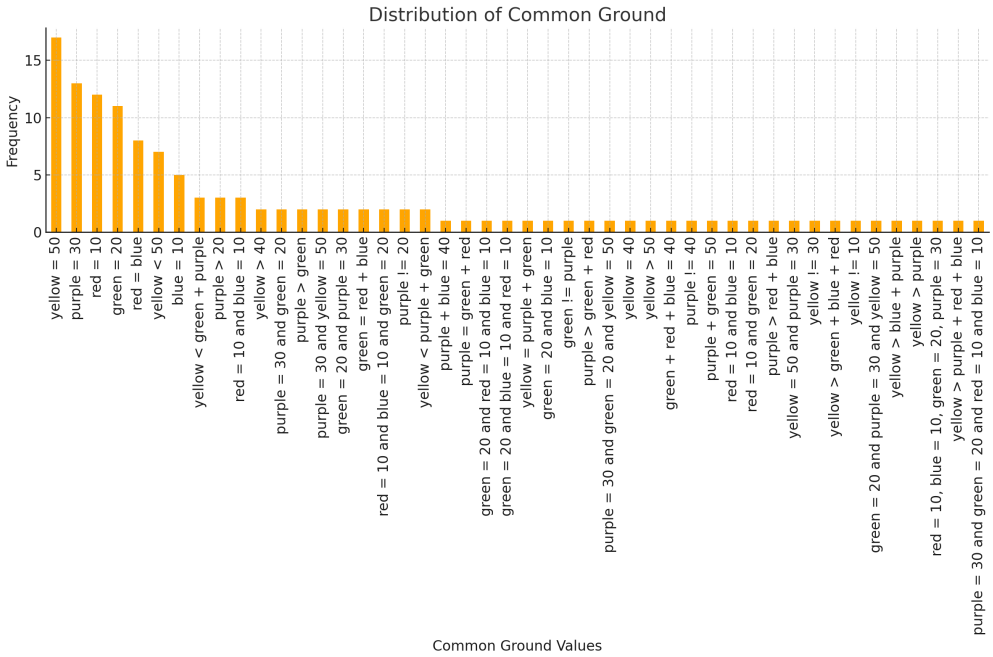

Figure 5.3: This figure illustrates the frequency of all 47 unique propositions expressed in the Weights Task Dataset. The horizontal axis lists the common ground propositions, such as weight assignments (e.g., *yellow = 50*), while the vertical axis represents their frequency across the dataset. The proposition *yellow = 50* is the most frequently expressed, appearing 17 times, followed by other key propositions such as *purple = 30* and *green = 20*. Less frequent propositions include combinations of weights and logical relations. This distribution highlights the diversity and repetition of propositions as participants collaboratively deduce the weights of colored blocks.

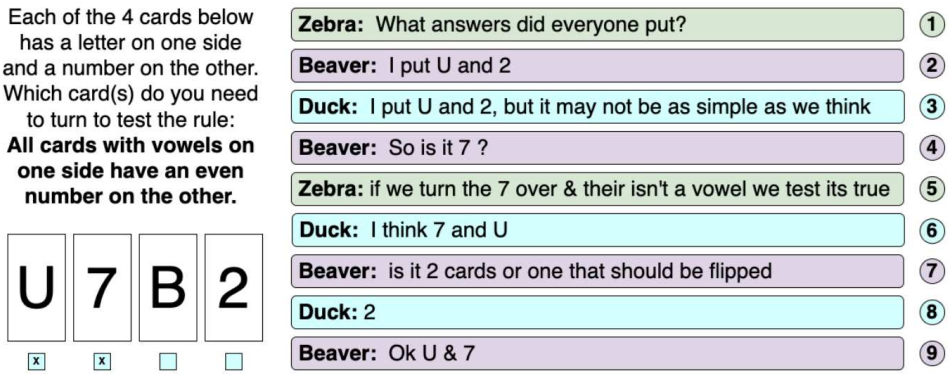

Figure 5.4: Abridged conversation from DeliData, illustrating the collaborative reasoning process of three participants solving the Wason card selection task. The task involves determining which cards to flip to test the rule that "All cards with vowels on one side have an even number on the other." Dialogue excerpts showcase how participants propose, evaluate, and revise their answers during the task. Reproduced from Karadzhov et al. (2023).

that could confirm the rule rather than those that could potentially disprove it. Fig. 5.4 shows an example conversation.

This dataset comprises 500 group dialogues, totaling approximately 14,000 utterances. The corpus is annotated with deliberation cues, focusing on how participants propose, evaluate, and revise solutions in a collaborative setting. The groups consist of up to five participants, and the dialogues occur in an online chat format.

5.4 Methods

In this section, we outline the data preprocessing steps, followed by a detailed description of the methodologies employed for extracting propositional content from the datasets.

5.4.1 Proposition Enumeration

Weights Task Propositional content in the Weights Task takes the form of a relation between a block and a weight value (e.g., $red = 10$), between two blocks (e.g., $red = blue$), or between one block and a combination of other blocks (e.g., $red < blue + green$). To generate all possible candidate propositions in the domain, we employed a systematic process that combined the five

block colors ($red, blue, green, purple, yellow$), five potential weights ($10, 20, 30, 40, and 50$), and four relations ($=$, $\neq$, $<$, $>$) into all possible combinations that fit the aforementioned formats. "Conjunctive" propositions (e.g., *green > 20 and yellow < 50*) were also allowed, up to a length of three conjuncts (the maximum that ever appeared in the actual dataset). We normalized all candidate propositions for the symmetric property of equality (e.g., so that $red = blue$ is the same as $blue = red$), and dropped the resulting duplicates. The result was 5,005 total candidate propositions that *could* be expressed in the Weights Task domain.

Any given proposition might be expressed in multiple ways. For instance, in the data "purple block's thirty," "purple one thirty," "let's go thirty purple block's thirty," and "teeter teeter purple block's less forty greater twenty purple block's likely thirty" all appear as ways of expressing the proposition $purple = 30$, despite the fact that they may contain extra words or even mentions of additional blocks or weights not contained within the proposition actually expressed. We therefore modeled propositional extraction as a type of *coreference* problem, where the goal is not to determine whether two entity mentions refer to the same thing [156], but rather to determine if two utterances mention both the same entity (block) and the same property (weight or relation).

DeliData Propositional content in DeliData takes the form of a structured set-member or attributive relation between a card and a property (e.g., $is(E, Vowel)$ for "[the] E [card] is a [member of set] Vowel"), or between a card and a hypothetical property on the hidden side (e.g., $has(E, Even)$ for "E has an Even number [on the other side]"). To generate all possible candidate propositions in the domain, we systematically combined the possible cards (A–Z and 1–9), the relevant properties ($Even$, Odd, $Vowel$, $Consonant$), and the defined relations ("is", "is not", "has", "does not have") into all possible combinations that align with the Wason selection task's structure. These propositions were normalized to account for symmetric properties and redundancies, resulting in a comprehensive set of propositions that could potentially be expressed in the dataset.

Given that any single proposition might be expressed in multiple ways within the dialogue, as in the WTD, we formalized the proposition in a similar manner. For example, different participants might express the same idea using varying phrases such as "E could have Even," "E would show

Even," or "E must have an Even." Despite differences in phrasing or additional words, these expressions all map to the same underlying proposition $has(E, Even)$.

Unlike the WTD, which has the propositions incorporated into its common ground annotations, DeliData does not contain explicit annotations of propositions expressed. Therefore, we had 2 annotators perform this task, following a strict guideline to ensure consistency across the dataset. Utterances were annotated with the proposition expressed (if any) following the form $< card > < relation > < property >$, and the annotations were made with a focus on distinguishing between visible aspects of the card (e.g., "is Vowel") and speculative or hypothesized aspects (e.g., "has Even"). This structured annotation allowed us to capture the reasoning process of the participants. Cohen's κ for this task was 0.955, indicating high annotator agreement [157], for a total of 255 utterances annotated over 100 groups. Utterances that described the multiple properties of the same card were removed since it expressed ambiguity and did not have an assertion. Examples of these include the utterance "I see, yeah, it doesn't matter if 4 is a vowel or not" expresses both $has(4, Vowel)$ and $\neg has(4, Vowel)$. Within those 100 groups, the occurrences of each operator are as follows: *is* appears 85 times, *is not* occurs 23 times, *has* appears 158 times, and *does not have* occurs 21 times. Proposition breakdowns show 197 instances with singular propositions and 57 instances with double propositions. These statistics cover 100 groups within the dataset that is being used for this task.

To generate the set of candidate propositions, first, we defined the allowable properties for each card type (letters and digits) based on the relation being asserted and according to the rules governing what can be on opposite sides of a card. For example, when the relation is "is" or "is not," a card showing a letter can only have the properties $Vowel$ or $Consonant$, while a card showing a digit can only have the properties $Even$ or Odd. Conversely, when the relation is "has" or "does not have," the properties apply to the hidden side of the card, with digits having $Vowel$ or $Consonant$ as properties and letters having $Even$ or Odd.

To avoid generating contradictory or redundant propositions, we ensured that each card was mentioned only once in any given set of propositions, which prevented conflicting assertions about the same card.

The final step involved generating all valid combinations of one or two propositions that adhered to the defined rules and consistency checks. This process produced a comprehensive set of candidate propositions that could theoretically be expressed during the Wason task. The result was a total of 38,362 propositions that could be used in the DeliData domain.

5.4.2 Annotation and Preprocessing of the Weights Task

Because of the multimodal nature of the Weights Task and the prevalent use of demonstratives, we enriched the transcribed utterances using a "dense paraphrasing" method inspired by [158] [158, 159], that rewrites a textual expression to reduce ambiguity and make explicit the underlying semantics. We isolated the utterances containing at least one pronoun from a predefined set of {"it", "they", "them", "this", "that", "these", "those"}, performed a partial assignment of blocks referenced by those pronouns based on actions that overlapped the utterances, and had annotators identify the blocks denoted by the remaining pronouns, if any, while referring to the video (see Fig. 5.5). This annotation was performed separately for the Oracle and Google transcriptions. Utterances were dually annotated, resulting in an average Cohen's $\kappa = 0.89$ over the Oracle transcriptions and $\kappa = 0.87$ over the Google transcriptions. A gold standard was then generated through adjudication by an expert. The original utterances were then replaced with the dense paraphrased versions. High agreement scores and accuracy metrics demonstrate the reliability and effectiveness of the annotation process. This procedure *decontextualizes* the utterances from their multimodal dependencies, allowing us to evaluate the utterance as though it were text only.

Data Cleaning of the Weights Task

Filtration of the WTD is motivated by the fact that many utterances, even after dense paraphrasing, still do not mention a specific object or weight, meaning that extracting an object-weight or object-object relation from the utterance alone is infeasible. Our filtration steps follow steps

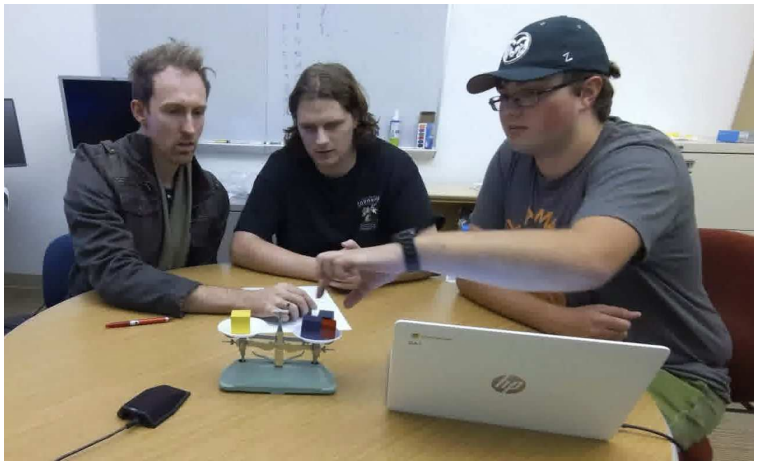

Figure 5.5: The image depicts participants in the Weights Task discussing potential solutions while interacting with the blocks and the balance scale. This setup emphasizes the importance of multimodal context (e.g., gestures and object interactions) in interpreting verbal utterances. For example, the original utterance is "we can replace one of [these] with the twenty." With reference to the video, an annotator can see the rightmost participant reaching for the red and blue blocks, so the dense paraphrased utterance is "we can replace one of *red block, blue block* with the twenty."

used in existing coreference research [2]. The decision to follow this methodology was made at the outset before any experimental results were available. We adopted three levels of data cleaning for WTD.

Level 1: The first level of cleaning consisted of removing all instances where neither color nor weight was mentioned in the transcript. An example of an utterance removed at this step would be "i mean it's not gonna go anywhere i guess it's just oh."

Level 2: The second level of cleaning involved removing all utterances where the mentioned colors and weights did not match the annotated proposition. For example, in an utterance "yeah red block, blue block should be twenty as well", "yeah" is actually an acceptance of a previously asserted proposition (in this case $green = 20$), and $red + blue = 20$, the mention of which is in the utterance, is not a valid propositional form in the task domain as the left hand side must be a single block (in this case, the truth of $red + blue = 20$ is implicit in two other (valid) propositions $red = 10$ and $blue = 10$).

Level 3: The final level of cleaning removed all instances that do not mention a color, but only a weight. For instance, the utterance "well the top is a ten" is annotated as $blue = 10$, but

with only the text, even a human would struggle to identify the correct proposition. The dataset annotators, meanwhile, had access to the video and could see that the top block referred to is blue, but as we focus only on transcriptions of natural speech, this information is not available to our method. By removing such ambiguous utterances, Level 3 cleaning results in a cleaner dataset where all remaining utterances explicitly mention both a color and a weight, making it easier for an automated system to extract propositions accurately.

Since the DeliData task does not include a multimodal component with the task, and has been already pre-processed, references to cards are already unambiguous, and so no additional cleaning was done on this dataset.

Data Augmentation of the Weights Task

The propositional extractor from [31] was limited by the sparsity of the utterances that actually expressed a proposition, totaling 47 unique propositions compared to the 5,005 possible propositions in the domain. For example, while $yellow + purple + green > red$ is a possible proposition according to the combinatorial process described in Sec. 5.4.1, it is unlikely to actually be expressed during task performance, as the combination of yellow, purple, and green blocks so clearly outweighs the red block that groups typically do not attempt it. In contrast, $green + purple = yellow$ is much more likely but may appear only once within a group, if at all.

To address this sparsity and improve generalization of the cross-encoder method, we explored a data-augmentation procedure inspired by prior work [101, 160–162]. Specifically, we prompted GPT-4 to generate 10 additional utterances for each of the 47 unique propositions present in the actual data, resulting in 470 new instances. These were combined with the 127 existing proposition instances, bringing the total coverage to 597 instances. Each generated utterance was subsequently human-validated for correctness before model training. The augmented data was used for training purposes only. The prompt used for GPT-4 is given in Fig. 5.6, followed by the specific proposition for which supplementary utterances were generated in the example case.

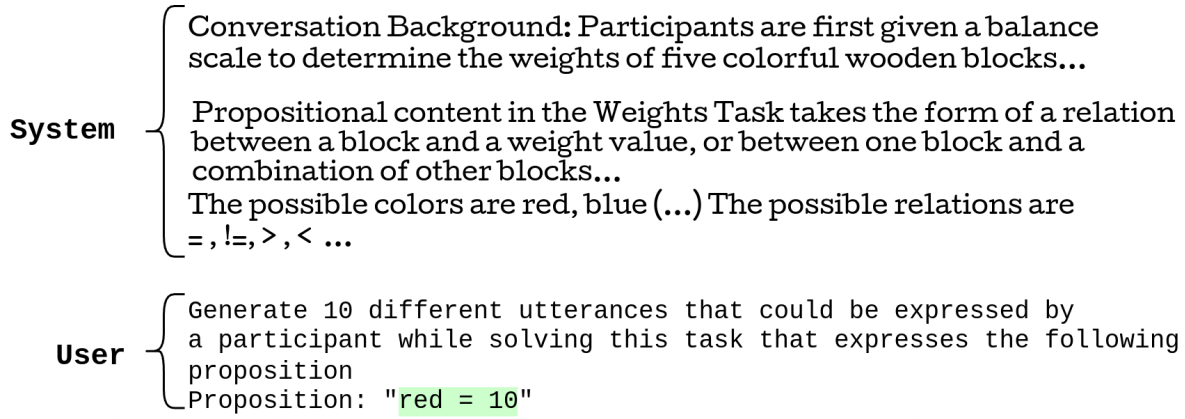

Figure 5.6: Synthetic data generation prompt used to augment the dataset for the Weights Task. The system defines the task context and possible relations, while the user prompt specifies generating 10 unique utterances expressing a target proposition (e.g., *red = 10*). This approach expands the dataset while maintaining linguistic diversity and task relevance.

While we performed data augmentation on the WTD for reasons of sparsity, with 100 of the 500 DeliData groups, we accumulated sufficient training data for the cross-encoder and no additional data was required for training purposes.

5.4.3 Cross-Encoder

Above, and in Sec. 5.2, we motivated propositional extraction as a type of coreference problem. Therefore, we use a cross-encoder neural network that is common in natural lanugage processing (NLP) approaches to coreference. The cross-encoder learns a paired "contextualized" representation for an utterance proposition pair. Unlike previous coreference approaches mentioned in Sec. 5.2, which focus on the specific trigger word within a sentence, we encode the *entire* utterance in the context of the proposition to generate a combined representation for an utterance/proposition pair. This is for two reasons. Firstly, in our framework both the transcript and the candidate proposition can contain more than one color mention, which serves as a trigger indicating a block. For instance, consider "so *purple* block, *blue* block should be forty right there" (utterance) and $purple + blue = 40$ (candidate proposition). Encoding the utterance once for each specific color-trigger using a language model could drastically increase computational cost without any additional benefits of contextualization. This could also likely break down higher-level

semantic signals that can otherwise be encoded with a wider context-window or the entire sentence. Secondly, under certain lenient pruning strategies, some transcripts may not contain any color at all. E.g., "... so you know twenty plus ten thirty probably ..." with a candidate proposition $red = 10 \wedge green = 20 \wedge purple = 30$. In such cases, full sentential context may capture more subtle semantic signals that are crucial for this task.

We encoded processed utterances as vector representations in two language models: **BERT-**`base-uncased` [138], and **RoBERTa-**`base` [144]. Before encoding, each dataset entailed slightly different preprocessing of the text. For WTD, stop words were filtered out according to a standard list augmented with words that occurred in five or fewer bigrams over all the transcriptions, and are not number words, color words, or (in)equality relation words. For DeliData, stop words were filtered out according to a standard list, with the exception of the set {'a', 'd', 'i', 'm', 'o', 's', 't', 'y', 'not', 'is', 'has', 'on'}. These characters or words were crucial to the context of the DeliData task, because they could refer to individual cards or relations between elements of a card. To retrieve the encoded vectors, we summed over the last four encoder layers of each model and took the average of the `[CLS]` (classification token, used to aggregate information from the entire sequence) or `<bos>` (beginning-of-sequence token, used for sequence initialization) token vector and all individual token vectors in the utterance. These vectors were used for propositional extraction by comparison using cosine similarity, and for training the cross-encoder architecture.

For an utterance/proposition pair (u_i, p_j), we construct an overall representation of the pair using the language model encoder. This representation consists of four individual parts, following modern standard practice in coreference established by [1]. We first surround u_i and p_j individually with special tokens `<m>` and `</m>` that are added to the language model tokenizer vocabulary and acquire learned representations during the training process. The first part of this overall representation is V_{CLS}, the pooled representation (`[CLS]`/`<bos>`) token of the last encoder hidden state. This representation is often used as a classification token in NLP tasks. Then, we encode u_i and p_j individually in the *context* of each other (that is, u_i when preceding p_j and p_j when fol-

lowing u_i)[11]. These comprise the second and the third components of the overall representation: V_{u_i} and V_{p_j}. We then encode the element-wise, or Hadamard, product of these two representations ($V_{u_i} \odot V_{p_j}$) to provide further cross-attention based signals. These four individual representations are then concatenated into a unified representation ($[V_{CLS}, V_{u_i}, V_{p_j}, V_{u_i} \odot V_{p_j}]$), which is fed into a multi-layer perceptron (MLP) to get similarity scores between the utterance and proposition (Eq. 5.1). The MLP is a two-layer neural network (768 and 128 neurons) that takes in the concatenated representation (768×4 = 3072 dimensions) and outputs a scalar, or after a sigmoid operation, the probability of an utterance referring to a proposition.

$$Score(u_i, p_j) = MLP([V_{CLS}, V_{u_i}, V_{p_j}, V_{u_i} \odot V_{p_j}]) \tag{5.1}$$

The candidate proposition with the highest score is retrieved, or the scores can be used to compute a *ranking* of candidate propositions, for metrics like top-k accuracy. Fig. 5.7 shows a schematic overview of the cross-encoder architecture.

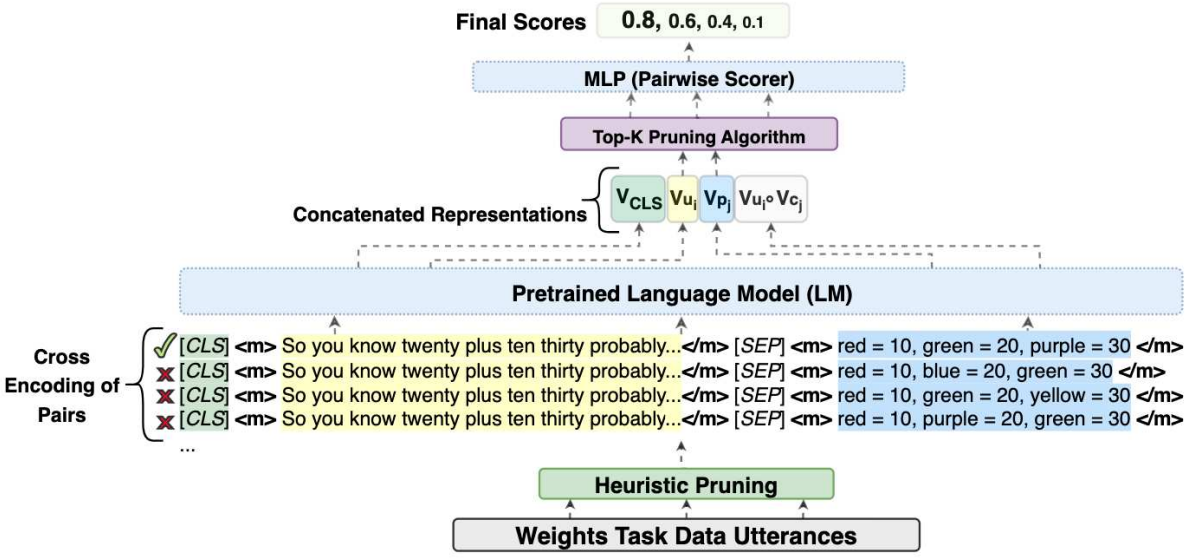

Figure 5.7: Schematic overview of the cross-encoder architecture, using example Weights Task data.

[11]The positional encoder of transformer models cause the resulting representations to be different despite the input order being the same. This avoids positional bias observed in transformer models and allows for a more unbiased loss computation.

Cross-Encoder Training

The parameters of the MLP are learned along with the parameters of the pretrained language model. Motivated by [2], we use a symmetric cross-encoding framework that minimizes the mean of the Binary Cross Entropy (BCE). More specifically, an utterance (u_i) and a proposition (p_j) are encoded bidirectionally, by interchanging their sequential positions in the input text ((u_i, p_j) and (p_j, u_i)), to avoid positional bias affecting loss computation observed in transformer-based models [163]. This results in a different unified representation in each direction and we minimize the average of the BCE loss over the encodings in both directions. Mathematically,

$$\mathcal{L}_{\text{BCE}(\theta,\phi)} = -\frac{1}{m}\sum_{i=1}^{m}\left(y_i \cdot \log \hat{y}_i + (1 - y_i) \cdot \log\left(1 - \hat{y}_i\right)\right) \tag{5.2}$$

where y and $\hat{y}$ are the true and predicted probabilities for an utterance-proposition encoding in one of the directions in a sample batch of size m. θ and ϕ are the parameters of the MLP and the pretrained LM, respectively. We train using a batch size of 20 for 12 epochs, with a learning rate of $1e-6$ on the LM parameters and $1e-4$ on the MLP pairwise scorer. With augmented data, the same training procedure is followed but the augmented data is added to the training set for each group-wise fold.

5.4.4 Experiments

We investigated two methods for extracting propositional content from utterances: a *cosine similarity* baseline, and a *cross-encoder* adapted from entity and event coreference research in the field of NLP. These were both evaluated over the Oracle transcriptions of utterances, and the Google automatic transcriptions, and using various levels of data cleaning to explore performance of the different methods in settings that range from more idealized to more realistic. Below we describe the methodology for cleaning the data and training the cross-encoder.

Cross-Encoder

Heuristic pruning of candidate propositions As mentioned, our data suffers from an imbalance between negative and positive samples, in that the vast majority of candidate propositions are not matches for a given utterance. This phenomenon is also present in common event coreference datasets, which results in a training dataset that is severely imbalanced toward negative pairs if not handled [135]. In our case, it

is usually quite obvious when a candidate proposition is not a possible match for an utterance because the candidate does not contain the object or weight value mentioned in the utterance. Therefore, we employ a heuristic pruning strategy on both datasets.

For the Weights Task Dataset, heuristic pruning operates at two levels. 1) We compare all propositions that include both the color and weight mentioned in the utterance (e.g., candidate matches for an utterance containing "red" and "ten" would include $red = 10$, $red \neq 10$, $red < 10$, etc.) 2) If the list of candidates is still empty, as might be the case for utterances such as, e.g., "it's fifty!", we then enlarge the search space by getting all the propositions that contain any of the colors or weights mentioned in the utterance. This process is similar to the lemma-based heuristic pruning used for training a cross-encoder for cross-document event coreference by [2].

DeliData has a similar negative-positive imbalance in the training data. We therefore employ a similar two-step pruning here. 1) We compare all propositions that include the same Card and the Property mentioned in the utterance (e.g., candidate matches for an utterance containing "Z" and "Odd" would include $has(Z, Odd)$ and $\neg has(Z, Odd)$). This essentially involves extracting a set of entities from the utterance and retaining only candidate propositions which contain the equivalent set. 2) If this list is empty, it might be that the utterance is expressing multiple propositions, e.g., "I thought the 2 needed to be turned over since it did not say that all even number cards only have vowels" is expressing the proposition $is(2, Even)$. However, the utterance contains the set of entities $\{2, Even, Vowel\}$ and no propositions contain all of these elements. In this case, in this pruning step, all individual propositions that can be expressed using elements of the utterance are retained as candidates. In this example, $is(2, Even)$, $\neg is(2, Even)$, $has(2, Vowel)$, $\neg has(2, Vowel)$, would be considered candidates.

Training data construction After filtering the candidate propositions with heuristic pruning, to create the training dataset for the cross-encoder, we pair an utterance with its annotated correct proposition as a positive pair and choose four random propositions from the filtered candidate propositions and pair them with the utterance as negative pairs. For example, the WTD utterance "ok so the red has ten" would be a positive match with $red = 10$ and a negative match with only three other candidates generated after pruning. This results in a more balanced ratio of negative to positive candidate propositions for a given utterance, which is beneficial for training. The random selection from the filtered propositions ensures a

diverse and robust set of negative samples. We pick only four random negative samples because a significant number of annotated propositions are of the form appropriate for the dataset, e.g., <color, relation, weight>, which means that after the first level of heuristic pruning, certain transcripts would have only four possible candidate propositions, viz. $< color > \{=, \neq, <, >\} < weight >$.

Testing methodology We perform a rotating leave-one-group-out experiment where cross-encoder training is performed over 9 of 10 groups in the WTD, and 99 of 100 groups in DeliData, with the remaining group reserved for the test set. The test group is then rotated through.

For testing, we use the same pruning methodology as described above for each dataset, but where necessary, further prune the candidate utterance-proposition pairs from the test set using a top-k pruning strategy, for which we use the previously trained cross-encoder. Specifically, we compute the cosine similarities between the embeddings of an utterance and the remaining candidate propositions, while interchanging their mutual positions. For instance, if (u_i, p_j) represents an utterance-proposition pair, we encode both $[V_{u_i}, V_{p_j}]$ and $[V_{p_j}, V_{u_i}]$ to retain their positional information. Since the cross-encoder has been trained to minimize the mean of the bidirectional BCE loss, the latent representations of positive pairs likely point in similar directions in the embedding space vis-à-vis the negative pairs. As such, a top-k pruning strategy allows us to generate the most similar candidate propositions for a particular utterance and remove mismatches which are more obvious. This helps the system's precision by minimizing the loss of pairs during pruning. We use $k = 5$ to ensure approximate consistency with the training set, which has a 1:4 ratio of positive to negative samples. We then score these leftover pairs using our trained cross-encoder. For each utterance, we consider the extracted proposition to be the one with the highest score as given by the cross-encoder since need a ranking system to choose a proposition for the evaluation metrics.

Cosine Similarity

For a given utterance's vector representation, we compute the cosine similarities between the embeddings of all candidate propositions and the utterance embeddings. We then sort these cosine similarities, retrieving the proposition(s) with the most similar embeddings to the utterance embedding. We use the same pruning strategies mentioned in Sec. 5.4.3 to be consistent. Because cosine similarity calculations only require the utterances to be encoded through a pre-trained model, and no training of a separate model,

we simply compare the encodings of utterances to those of propositions without the need for a leave-one-group-out split.

Zero-Shot Baselines

In addition to the cross-encoder and cosine similarity methods, we also establish zero-shot baselines using GPT-4 and LLaMA2-13B, inspired by prior works [164]. The goal of these baselines is to assess the feasibility of extracting propositions directly from utterances without any explicit training or fine-tuning. The zero-shot approach involves prompting the language models with an utterance and instructing them to identify the underlying proposition. The prompt structure is designed to provide the models with context about the task and the expected format of the output. The model is also asked to produce a rationale for its decision, we leads the model to perform chain-of-thought -style reasoning, which has been shown to elicit improved reasoning in LLMs and guard against erroneous outputs [30, 101, 165]. The use of this technique provides extra guidance to the LLM, resulting in zero-shot baselines that are not artificially low. The specific prompts used for GPT-4 and LLaMA2-13B are shown in Fig. 5.8. We do not report LLaMA 2-13B zero-shot performance on DeliData, as the model was unable to extract any coherent, properly-formed propositions from the provided utterances given the prompt.

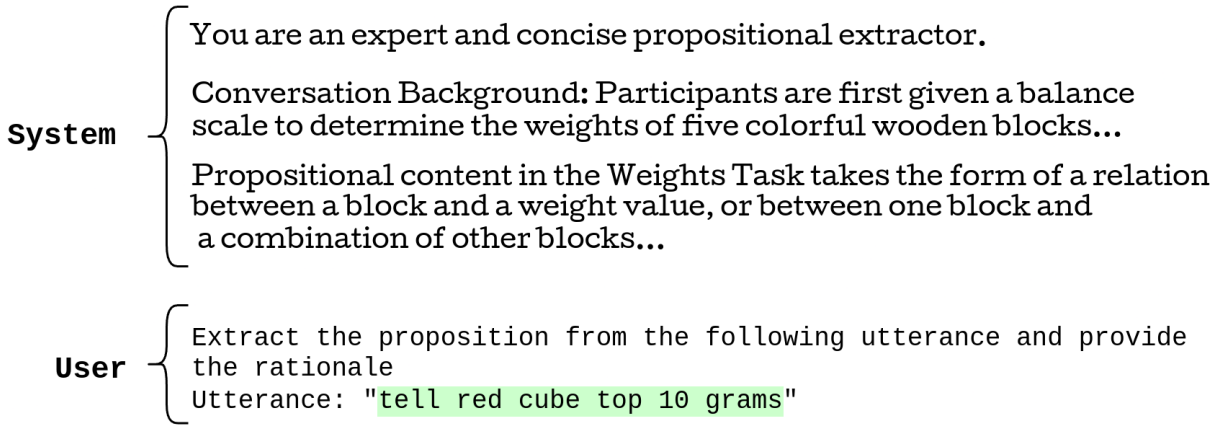

Figure 5.8: Prompt used to establish zero-shot baselines for propositional extraction. The system prompt specifies the task context and defines the structure of propositional content, while the user prompt provides an utterance (e.g., “tell red cube top 10 grams") for which the system must extract the corresponding proposition and rationale. This approach evaluates the model’s ability to generalize without prior task-specific training.

5.4.5 Metrics

To evaluate our system's performance, we use 3 common metrics for retreival tasks: Intersection Over Union (IOU), top-1 accuracy, top-3 accuracy.

IOU measures how well we extract components of propositions, even if the entire proposition is not retrieved perfectly. It calculates the overlap between predicted and true proposition. For example, if the true proposition is $red = 10 \wedge blue = 20 \wedge green = 10$ and we extract proposition $red = 10 \wedge blue = 30$, we consider the cardinality of the intersection of the two sets ($\{red = 10\}$) over their union ($\{red = 10, blue = 20, green = 10, blue = 30\}$). This assesses partial matches where some, but not all, of the correct propositional content is retrieved. In the example, the IOU score would be $\frac{1}{4}$ or 0.25. This is because only one element ($red = 10$) matches out of a total of four unique elements across both propositions.

Top-1 accuracy is stricter; it only counts if we extract the exact proposition. For example, if the true proposition is $red = 10$, the only way to attain a score of 1 is if the prediction is also $red = 10$.

Top-3 accuracy also requires an exact match, but it counts if the correct proposition is among the top three extractions. For instance, if the true proposition is $red = 10$, a set of top three predictions $red \neq 10$, $red > 10$, and $red = 10$, would get a top-3 accuracy score of 1 since the correct proposition is present within the top 3. Top-3 accuracy is not reported for zero-shot baselines since GPT-4 and LLaMA 2 extract only one proposition from each utterance.

5.5 Results

For the Weights Task Dataset, we report all results across the three different levels of data cleaning discussed in Sec. 5.4.2. Results include the cross-encoder with and without augmented training, the cosine similarity method, and zero shot.[12] First, we established the performance of our methods on a "best case" baseline. Then, we explored how performance was modulated by the level of data cleaning and automatic speech recognition (ASR). Across these data variants, our cross-encoder outperformed the other methods. We then explored how these methods, specifically the cross-encoder, the cosine similarity, and zero shot methods, generalized to a new domain: the DeliData dataset. All results are presented in Tables 5.1 - 5.7

[12]We do not reproduce the Longformer results reported in [31] because the Longformer-based cross-encoder substantially underperformed the cross-encoders using BERT and RoBERTa, so we focus on those other models here.

while detailing extraction method, and the performance metrics. '-aug' refers to the cross encoder model trained on augmented data.

Best-case Cross-Encoder First, we evaluated our methods on Level 3 (the most rigorous level of data cleaning) with Oracle transcriptions (Table 5.1). The performance on this data establishes a "best-case" baseline for propositional extraction on the Weight Task, where the transcriptions are manually transcribed and optimally cleaned, removing utterances that contain no color but only a weight value.

Impact of data cleaning In Tables 5.2–5.6, we evaluated our methods on increasingly challenging data conditions. Specifically, we first reduced how clean the data was (as detailed in Sec. 5.4.2) to Levels 2 (Table 5.3) and Level 1 (Table 5.5), where Level 1 is the most difficult condition.

Impact of automatic speech recognition Then, we explored how ASR transcriptions impacted performance across Level 3 (Table 5.2), Level 2 (Table 5.4), and Level 1 (Table 5.6). As expected, all methods exhibited a drop in performance as data became more realistic (and thus more difficult); however, our cross-encoder-based method consistently outperformed the other methods.

Generalization to DeliData Finally, we present the DeliData results in Table 5.7. Unlike the WTD, DeliData preprocessing does not involve varying levels of cleaning or different transcription methods. Instead it provides a more straightforward evaluation scenario where the dialogue is solely in text form and inherently cleaner, making it approximately comparable of the WTD results in the condition reported in Table 5.1, with maximal cleaning and Oracle transcription.

5.6 Model Selection and Statistical Analysis

We selected **Oracle Level 3 RoBERTa** as the best-performing model on the Weights Task Dataset (WTD) based on its superior results across the tested conditions, *assuming no data augmentation* as augmenting data on the WTD did not significantly impact performance. To validate this selection, we performed paired t-tests, comparing Oracle Level 3 RoBERTa IOU scores with each other condition. For each group, the IOU scores were calculated and compared across all models and configurations. The results, as shown in Table 5.8, reveal that *data cleaning significantly impacts model performance*, with Level 3 cleaning yielding

Table 5.1: Propositional extraction performance on the Weights Task dataset with *Level 3 cleaning* using *Oracle* transcriptions. The columns represent IOU (Intersection Over Union), Acc. (Top-1 Accuracy), and Top-3 Accuracy. Bolded values indicate the best performance for each metric

	IOU	Acc.	Top-3
BERT	0.664	0.640	0.773
BERT-aug	**0.771**	**0.762**	**0.905**
RoBERTa	0.683	0.671	0.829
RoBERTa-aug	0.753	0.747	0.867
BERT-cosine	0.570	0.547	0.747
RoBERTa-cosine	0.337	0.307	0.520
GPT-4	0.659	0.546	–
LLaMA 2	0.643	0.513	–

Table 5.2: Propositional extraction performance on the Weights Task dataset with *Level 3 cleaning* using *automatic* transcriptions. Columns include IOU (Intersection Over Union), Acc. (Top-1 Accuracy), and Top-3 Accuracy. Bolded values indicate the best performance for each metric.

	IOU	Acc.	Top-3
BERT	0.635	0.607	0.787
BERT-aug	**0.651**	**0.633**	**0.817**
RoBERTa	0.645	0.607	0.738
RoBERTa-aug	0.648	0.617	0.800
BERT-cosine	0.281	0.262	0.344
RoBERTa-cosine	0.057	0.049	0.147
GPT-4	0.483	0.417	–
LLAMA 2	0.463	0.416	–

Table 5.3: Propositional extraction performance on the Weights Task dataset with *Level 2 cleaning* using *Oracle* transcriptions. Columns include IOU (Intersection Over Union), Acc. (Top-1 Accuracy), and Top-3 Accuracy. Bolded values indicate the best performance for each metric.

	IOU	Acc.	Top-3
BERT	0.596	0.562	0.730
BERT-aug	**0.639**	**0.607**	**0.831**
RoBERTa	0.585	0.573	0.789
RoBERTa-aug	0.599	0.573	0.753
BERT-cosine	0.505	0.472	0.651
RoBERTa-cosine	0.284	0.258	0.461
GPT-4	0.599	0.472	–
LLaMA 2	0.520	0.460	–

Table 5.4: Propositional extraction performance on the Weights Task dataset with *Level 2 cleaning* using *automatic* transcriptions. Columns include IOU (Intersection Over Union), Acc. (Top-1 Accuracy), and Top-3 Accuracy. Bolded values indicate the best performance for each metric.

	IOU	Acc.	Top-3
BERT	0.537	0.526	0.737
BERT-aug	**0.563**	**0.547**	**0.747**
RoBERTa	0.530	0.500	0.697
RoBERTa-aug	0.498	0.480	0.680
BERT-cosine	0.232	0.210	0.276
RoBERTa-cosine	0.052	0.039	0.118
GPT-4	0.391	0.333	–
LLaMA 2	0.418	0.373	–

significantly better results than Levels 1 and 2 (Table 5.8). This supports the conclusion that data cleaning at this level enhances the model's effectiveness. The cross-encoder demonstrates statistically significant improvements over the cosine-based approaches, as seen in the comparisons between RoBERTa Level 3 cross encoder model and and Cosine RoBERTa/BERT.

Table 5.5: Propositional extraction performance on the Weights Task dataset with Level 1 cleaning using Oracle transcriptions. Columns include IOU (Intersection Over Union), Acc. (Accuracy), and Top-3 Accuracy. Bolded values indicate the best performance for each metric.

	IOU	Acc.	Top-3
BERT	0.526	0.496	0.609
BERT-aug	**0.561**	**0.539**	**0.678**
RoBERTa	0.448	0.426	0.557
RoBERTa-aug	0.501	0.474	0.649
BERT-cosine	0.229	0.200	0.356
RoBERTa-cosine	0.419	0.347	0.514
GPT-4	0.453	0.374	–
LLaMA 2	0.336	0.304	–

Table 5.6: Propositional extraction performance on the Weights Task dataset with Level 1 cleaning using automatic transcriptions. Columns include IOU (Intersection Over Union), Acc. (Accuracy), and Top-3 Accuracy. Bolded values indicate the best performance for each metric.

	IOU	Acc.	Top-3
BERT	0.353	0.309	0.427
BERT-aug	0.389	0.345	**0.527**
RoBERTa	0.383	0.336	0.464
RoBERTa-aug	**0.422**	**0.373**	0.473
BERT-cosine	0.036	0.027	0.081
RoBERTa-cosine	0.164	0.114	0.198
GPT-4	0.298	0.261	–
LLaMA 2	0.336	0.304	–

Table 5.7: Propositional extraction performance on the *DeliData dataset*. Columns include IOU (Intersection Over Union), Acc. (Top-1 Accuracy), and Top-3 Accuracy. Bolded values indicate the best performance for each metric.

	IOU	Acc.	Top-3
BERT	**0.707**	**0.634**	**0.773**
RoBERTa	0.675	0.605	**0.773**
BERT-cosine (+ pruning)	0.499	0.436	0.668
BERT-cosine (− pruning)	0.175	0.102	0.130
RoBERTa-cosine (+ pruning)	0.413	0.344	0.680
RoBERTa-cosine (− pruning)	0.228	0.090	0.165
GPT-4	0.545	0.433	–

Furthermore, *there is no statistically significant difference between Google and Oracle transcription systems*, as demonstrated in the non-significant comparisons (Table 5.8). Similarly, *training with augmented data did not result in statistically significant improvements* over standard training. Both findings suggest that additional complexities, such as alternative transcription systems or data augmentation, do not meaningfully impact performance for this task. We also observed that our model's performance did not exhibit statistically significant differences compared to GPT and LLAMA 2-13B. Given that these models are state of the art models and trained on vast datasets, achieving comparable results without significant performance dif-

ferences is a promising indication that our cross-encoder approach is an effective alternative with far fewer parameters. This suggests that our method holds substantial potential in practical applications, providing competitive accuracy without the need for large generative models.

Table 5.8: Paired t-test results (Significant, $p < 0.05$) comparing Oracle Level 3 RoBERTa with other models

Comparison	t-statistic	p-value
Oracle Level 2 RoBERTA	2.35	0.043*
Oracle Level 1 RoBERTA	6.22	<0.001*
Oracle Level 3 BERT-cosine	-1.14	0.027*
Oracle Level 3 RoBERTa-cosine	-1.14	<0.001*
Oracle Level 3 RoBERTa-aug	-1.14	0.285
Google Level 3 RoBERTa	1.01	0.337
GPT-4	0.07	0.941
LLaMA 2	0.06	0.948

* Indicates statistical significance at $p < 0.05$.

5.7 Discussion

Comparison of data cleaning strategies As expected, with increased levels of data cleaning on the Weight Task, we see a trend of improving performance across all extraction strategies, language models, and transcription methods. The progressive removal of noise, such as incomplete or ambiguous utterances as discussed in Sec. 5.4.2, directly enhances the accuracy and IOU of propositional extraction. This trend is consistent across different language models (BERT, RoBERTa) and holds true whether using manually segmented Oracle transcriptions or automatically generated ASR transcriptions. However, it is important to note that this increase in performance comes at a trade-off. As we apply more rigorous cleaning criteria, the number of usable utterances decreases significantly. With fewer samples to evaluate, the models may become overly tuned to the cleaner dataset. Furthermore, in real-world applications where such extensive cleaning might not be feasible, the performance gains seen under these ideal conditions might not fully translate.

Comparison of extraction methods The cross-encoder consistently outperforms all other baselines across all three metrics. Comparing the extraction methods across Tables 5.1– 5.6 shows that the cross-encoder outperforms the cosine baseline by at least 0.2 IOU on average. On the other hand, with a metric that does not reward partial selection, like traditional or Top-1 accuracy, the cross-encoder outperforms the cosine baseline by at least 40%, on average, although the absolute scores are typically lower than the more lenient IOU metric.

Comparison of transcription methods As expected, using automatic transcriptions of the speech leads to a consistent degradation in performance, as automated segmentation and transcription may incorrectly conflate two overlapping utterances from different people, or as annotators leave out or insert words, where such errors are expected to be minimized by a careful human transcriber. However, this degradation can sometimes be quite small, especially at higher levels of data cleaning, when using the cross-encoder, and the BERT or RoBERTa models. For instance, when using the cross-encoder, the accuracy using BERT embeddings of Google automated transcriptions increases from 30.9% at data cleaning Level 1 (least stringent) to 60.7% at data cleaning Level 3 (most stringent), while when using cosine similarity with pruning, accuracy only increases from 14.4% to 26.2%.

Comparison of language models Using embeddings from BERT typically achieves the best performance, but the performance gap with RoBERTa embeddings is usually quite small especially for the cross-encoder. RoBERTa sometimes performs better than BERT on less clean data, which may reflect the larger and more diverse training data of RoBERTa. Both of these models significantly outperform the Longformer model from [31].

Comparison of augmented vs. raw training data Training with augmented data results in a small increase in the performance across all metrics on the WTD. The lack of a significant performance jump can be attributed to the fact that the multimodal nature of the Weights Task. A sentence like "10 10 20" is annotated as $green = 20$. This is because the participant is pointing at the green block at the time, which the human annotators can see, but the utterance does not explicitly convey that green weighs 20g. Thus, GPT-4 does not have the capacity to generate a sentence that is both similar to "10 10 20" *and* makes clear that the intended meaning is $green = 20$. While GPT-4 is capable of generating clean, syntactically

correct utterances that can enrich the dataset, it struggles to replicate the nuanced, context-dependent nature of the original utterances. In the case of "10 10 20," the crucial information—namely, the association of the utterance with the green block—is derived from visual context, something that GPT-4 cannot infer or incorporate when generating new data. This limitation suggests that while data augmentation can help increase the quantity of training data and may offer some performance benefits, it does not necessarily equip the model to better handle the complexities introduced by the multimodal nature of the task. The model's improved performance with augmented data is therefore marginal, as it still struggles to interpret or generalize from utterances where the meaning heavily relies on non-verbal cues, such as gestures or object references.

We initially hypothesized that the limited availability of annotated multi-modal WTD data was a primary factor limiting model performance. However, the lack of significant gains from this augmentation points to nuances and intricacies involved in proposition expression that go beyond mere data quantity.

Comparison of zero-shot baselines Zero-shot LLM performance reflects the complexity of the task. The utterances are often not clean or complete, and therefore LLaMA 2-13B and GPT-4 are often unable to extract propositions from them. However, zero-shot baselines still follow the same pattern of the cross-encoder and the cosine baselines, where a cleaner dataset results in a better performance. This is expected, as with cleaner data the large language models are provided with coherent sentences with explicit mentions of the blocks and the weights. Level 3 cleaning on the WTD results in a zero-shot IOU comparable to that of the cross-encoder. Zero-shot results on DeliData further confirm the pattern that Deli is most similar to clean versions of the WTD, and that the cleaner utterances hold explicit semantic meaning and convey the relationship between the task-relevant elements.

Our results demonstrate that a fine-tuned cross-encoder model is comparable to baselines from powerful LLMs like GPT-4 and LLaMA 2-13B in the task of propositional extraction from dialogue. While these large language models offer impressive capabilities, their deployment in real-world scenarios, especially within educational contexts, presents significant challenges. Particularly, GPT-4 employs a pay-per-use model, making it financially unsustainable for large-scale or continuous applications. Similarly, even though LLaMA 2 is an open-weight model, running a 13-billion parameter model necessitates access to substantial computation resources including high-end GPUs, which might be prohibitively expensive or unavailable for

many communities. Moreover, directly sending transcribed student utterances or other private information to external LLMs, even for zero-shot inference, could pose significant privacy risks and compliance issues.

As noted in Sec. 5.4.4, LLaMA 2-13B failed to extract any coherent propositions from DeliData, highlighting the complex nature of the task and ways in which propositions may be expressed. For instance, when given the utterance, “It’s asking you to test the rule. Would it not prove that the rule is tested if the 2 turns up a vowel?", which expresses the proposition $has(2, Vowel)$, LLaMA 2-13B returned the proposition, “A is even, B is odd, C is vowel, D is consonant,” which is completely unanchored from the entities and card properties actually contained in the utterance.

Comparison of datasets The performance of the best cross-encoder on the DeliData dataset (0.707 IOU, using BERT) falls between the best cross-encoder’s performance on WTD with Level 2 (0.639 IOU, using BERT-sug) and Level 3 (0.771 IOU, using BERT-aug) cleaning. This is in part due to the unimodal (language-only) nature of the DeliData task and the tendency of the participants to be explicit about the elements of the cards and their properties. This may also be an effect of the text-chat nature of the dialogue, where participants explain themselves more fully to avoid ambiguity and misinterpretation by their task partners. The more explicit nature of DeliData utterances is also reflected in the zero-shot performance, which has a much higher baseline on DeliData than on all but the most rigorously cleaned version of the Weights Task data; our cross encoder method is more uniquely suited to handling the ambiguities that arise in a multimodal task like the Weights Task.

5.7.1 Group-wise Analysis

Figs. 5.9 shows IOU and 5.10 shows top-3 accuracy results from the test samples of each group, at Level 1 (most lenient) data cleaning, using BERT embeddings. The plots compare performance using Oracle (left charts) vs. Google (right charts) utterances and compare the cross-encoder to cosine similarity with heuristic candidate pruning.

We can see that cross-encoder performance on Group 7 is nearly identical regardless of which transcription method was used. This is likely because Group 7’s utterances used mainly simple propositions of the form <color> <relation> <weight>. These instances are easy to extract from the transcripts, and the automated transcripts are likely of high-fidelity.

We can see in Fig. 5.9 that Group 4's IOU drops significantly when comparing cosine similarity's performance over Oracle transcriptions vs. over Google transcriptions. While exploring the samples from this group, several issues were noted. We found eight utterances in the Oracle data and only seven in the Google data, meaning that one of the utterances was completely missed by Google ASR. This utterance happened to be very straightforward and easy for the cosine method to classify. The Oracle transcript is simply "blue ten." Another issue, again due to the segmentation, is Google ASR may merge two utterances. This highlights a limitation of ASR models, where some additional context may needed to know when a speaker has moved to another sentence. Obviously, the main difference between using the different transcription methods is the transcripts themselves. One instance from Group 4 states "easy green block twenty cause ..." whereas Google ASR transcribed the utterance as "okay e green block red block 10 ...". These results highlight certain issues that should be considered when deploying such an information extraction system over the outputs of an ASR system, as may be required in classroom environments.

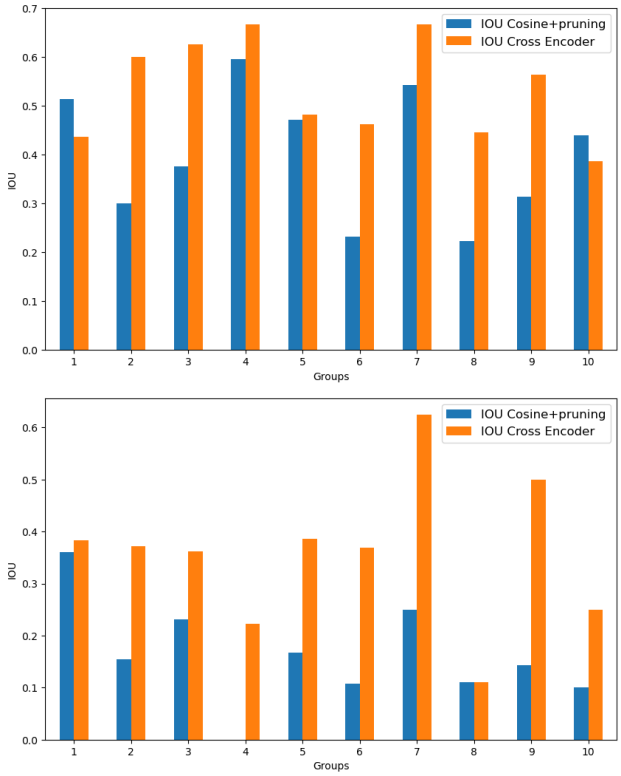

Figure 5.9: Group-wise *Intersection over Union (IOU)* comparison at Level 1 data cleaning using BERT embeddings. The left chart shows performance with *Oracle* transcriptions, while the right chart reflects performance with *Google ASR* transcriptions. Blue bars represent Cosine Similarity with Pruning, and orange bars represent the Cross-Encoder method across all groups.

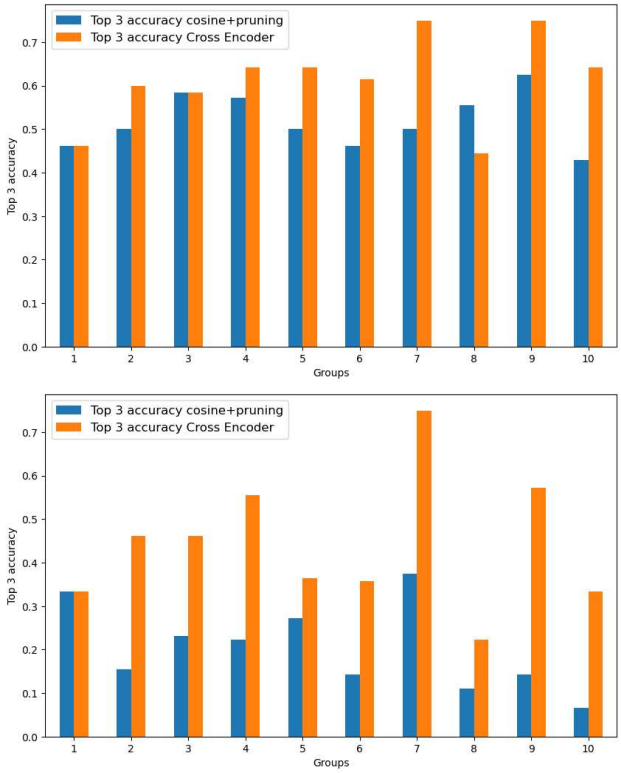

Figure 5.10: Group-wise *Top-3 Accuracy* comparison at Level 1 data cleaning using BERT embeddings. The left chart displays performance with *Oracle* transcriptions, and the right chart shows performance with *Google ASR* transcriptions. Blue bars represent Cosine Similarity with Pruning, and orange bars represent the Cross-Encoder method across groups.

5.7.2 Error Analysis

As the cross-encoder is consistently the best-performing extraction method, examining samples it gets wrong is informative. One such example is the utterance “green block one probably twenty ten ten twenty". The correct proposition is $blue = 10 \wedge green = 20 \wedge red = 10$. The annotators have access to the video and can see that when saying “ten ten twenty,” the speaker is actually pointing to the blue block, then the red block, then the green block. This information is not available through the textual medium alone.

The nature of the DeliData task leads to some errors, mostly pertaining to misinterpreting letters. For example, the utterance “makes sense flip 6 check a vowel” mentions only the card ‘6’. However, since the word ‘a’ could also represent a card (“A”), propositions with the cards 6 and A are retained as candidates, leading to mis-retrieval errors. The same occurs with the card “I”.

ASR transcription errors appear to play a role in zero-shot extraction errors. For example, zero-shot LLMs are unable to extract a proposition from the automatically transcribed utterance "blue block seems done" (actual utterance "blue block seems 10"), because the word "done" does not provide any context to the weight of the blue block. The cross-encoder models successfully retrieve the proposition $blue = 10$ from the utterance, even with the transcription error, because it is trained with task context that links the blue block and weight 10g.

Top-k Errors In order to compare our two extraction methods, we carried out a detailed analysis of candidate propositions that were ranked similarly, based on their cross-encoder scores or cosine similarities. On average, at Level 1 (most lenient) data cleaning, the cross-encoder performs comparatively better at ranking the correct propositions in the top 5. For instance, the cross-encoder ranks 8 and 21 correct propositions higher than the cosine similarity method, for Google and Oracle transcripts respectively. The cosine similarity method ranks 1 (Google) and 11 (Oracle) correct propositions higher.

On the other hand, there were at least 14 Google utterance transcripts and 37 Oracle utterance transcripts where both the extraction methods performed equivalently.

Qualitative Analysis On average, simpler utterances that contain a reference to only one color and/or weight are correctly retrieved by both the cross-encoder and cosine similarity. For instance, "I tell red cube ten grams" (correct proposition $red = 10$) and "green twenty" ($green = 20$). More interestingly, the cross-encoder seems to retrieve utterances with ambiguous context without a direct reference to color or with multiple colors more effectively than the cosine similarity method. For example, "Fifty I" ($yellow = 50$) and "green block twenty red block, blue block ten ten" ($blue = 10 \wedge green = 20 \wedge red = 10$). This is likely due to the cross-encoder's cross-attention based signals that are being sourced from the entire utterance in the context of the candidate proposition. This was previously observed in [1] where modeling global signals in parallel with local features led to an overall increase in coreference resolution performance. Since in the actual task data "yellow" was expressed most frequently in the context of "50" and relation $=$, when only "Fifty" is expressed in an utterance, $yellow = 50$ gets the highest score. This is not possible with the cosine similarity since it is not trained on the data and there is no particular relationship between "yellow" and "fifty" in general language.

In collaborative tasks, participants bring individual knowledge, but progress hinges on how that knowledge is externalized, interpreted, and collectively aligned. This chapter focused on the extraction of task-relevant propositions—discrete expressions of belief or inference, as a key signal in that process. These utterances do not just describe a participant's mental state; they provide the linguistic material that others in the group can build on, challenge, or reinterpret.

Our findings demonstrate that, despite the noise and variability of natural dialogue, task-relevant propositions can be reliably identified using adapted NLP methods. By training models that link utterances to structured representations of meaning, we make explicit the often, implicit process by which ideas take shape in group discourse. This capability has practical implications for real-time monitoring and educational feedback, where understanding what has been said—and whether it moves the group forward—is critical to evaluating collaboration and learning.

Just as the previous chapter modeled the arc of reasoning through deliberation chains, this work zooms in on the foundational layer of that reasoning: the explicit assertions made by individuals. Taken together, the two chapters reveal a multiscale view of group interaction—from the emergence of ideas to their eventual convergence or conflict.

Propositional extraction, then, is not just about labeling what was said. It is a window into how knowledge is constructed and shared, and beliefs emerge. The structured outputs of this process can inform systems that support collaborative tasks by identifying missing pieces of knowledge, redundancy in discussion, or opportunities for prompting deeper inquiry. As we continue to build systems that interact with humans in educational or team-based contexts, capturing these underlying phenomena of understanding will be essential [166].

5.8 Limitations

Dataset Size and Loss Due to Cleaning The data preparation and cleaning procedures inevitably result in the loss of several utterances. This leads to small datasets, with the Weights Task Dataset (WTD) ranging from dozens to slightly over 100 utterances, depending on the level of data cleaning, and 255 utterances for DeliData. The reduced dataset size can impact the robustness of the analysis and limit the generalizability of the results.

Impact of Automated Transcription Errors Errors in automated transcripts can adversely affect the efficacy of the candidate pruning process. For example, Google transcribes an utterance as "blue block's obviously time," when the transcribed word "time" was actually uttered as "10." Such transcription errors disrupt the pruning process for candidate propositions, as it incorrectly limits the search space to propositions mentioning "blue" without "10," thereby affecting performance.

Dependence on Heuristic Pruning Heuristic pruning of candidate propositions significantly affects performance, as seen in the results of cosine similarity with and without pruning. Pruning not only reduces the search space but also helps maintain a balanced sample distribution for training and aligns test data with the distribution of the training data. However, the pruning methodologies are task-specific and must be adapted to the nature of the propositions in each scenario, limiting the automatic generalizability of the method.

Task Definition Dependency The system assumes a well-defined task structure with a finite set of propositions that can be expressed. This reliance on predefined proposition templates means the approach is inherently limited to scenarios where possible outcomes are known a priori. Consequently, the method cannot be readily applied to tasks with open-ended goals or undefined propositional spaces.

Reliance on Annotated Training Data The system requires access to annotated training data for the cross-encoder model. While synthetic data augmentation using GPT-4 helps mitigate the scarcity of training samples, the generated utterances may fail to capture the nuanced linguistic variations present in collaborative settings. As a result, the quality of augmented data does not fully replicate the richness of real-world dialogues, limiting the system's ability to generalize.

Need for Domain Expertise The approach necessitates domain expertise to define relevant propositions and validate their relevance to the task. Subject matter experts play a crucial role in ensuring that extracted propositions are meaningful and aligned with task requirements. However, this dependence on manual oversight hinders the scalability of the system to new tasks or domains without significant investment in expert resources.

Privacy Concerns in Real-World Deployment Participants in the Weights Task Dataset consented to the recording and analysis of their data using third-party tools such as Google ASR for research purposes. However, in real-world classroom implementations, using cloud-based services like Google ASR raises ethical concerns regarding student privacy. To address this, local custom models would need to be developed and deployed to ensure data privacy and compliance with ethical standards.

5.9 Conclusions

In this paper, we have defined and explored the complex problem of automatically identifying propositional content from transcriptions of natural speech in a collaborative task. Automated propositional extraction from speech serves a number of important educational purposes. For example, tracking the assertion of propositions over time indicates how students are or are not discussing key concepts relevant to the task, which in turn indicates the construction of shared knowledge [125].

The Weights Task data presents many challenges, from overlapping speech to incomplete sentences, and we have evaluated a suite of transformer-based language models based on two different methodological frameworks: a cosine similarity baseline vs. a cross-encoder. Our experiments present a feasible method for performing the extraction of task-relevant propositions by building upon publicly-available language models and pairwise representation learning techniques. The successful implementation of the same task on DeliData, a dataset with an entirely different domain, shows the generalizability of our methods given only an inventory of task-relevant propositions, which can be enumerated deterministically. While ground-truth annotation is needed for cross-encoder training, our success on a small amount of data demonstrates the small amount of needed annotation. We have also shown that this task is not a trivial one to be disposed of with off-the-shelf LLMs, as demonstrated by the inferior performance of GPT-4 and LLaMA 2-13B when compared to our own methods.

Our best performing methods, particularly the cross-encoding framework, show a narrow performance gap when operating over automated transcriptions when compared to human transcriptions, suggesting a feasible path forward toward fully automating such a system in a live environment. A clear application in a classroom is in a system that models the shared knowledge of a group toward the task goal, and might be a component of an AI agent who assists small groups in collaborative problem solving (CPS) [167, 168].

Part III explores this transition from model to system—investigating how upstream choices in transcription and segmentation affect downstream performance. As we move from detecting human phenomena to deploying CPS detection systems, we confront new challenges of fidelity, interpretability, and signal loss in noisy, real-world data.

Part III

From Signal to System: Challenges of Automation

In natural interaction, the fidelity of data is never guaranteed. Even humans mishear common words, misinterpret intent, or respond based on partial signals. For AI systems, this challenge is magnified: language must be transcribed automatically, speech must be segmented algorithmically, and meaning must be inferred without the benefit of perfect understanding. The way interaction is recorded, how and when signals are captured, plays a central role in what we can ultimately infer.

This part of the thesis focuses on that challenge: what it means to detect human phenomena when data is automatically recorded. If we are to model collaboration in real-world environments, we must grapple with the effects of noise, misalignment, and loss. These are not just technical concerns, they are epistemological ones. The process of recording is not neutral; it shapes what can be known.

Earlier parts of the thesis explored the nature of internal and interactive human phenomena: how attention shifts, how beliefs are surfaced, and how reasoning emerges. Here, we turn to Collaborative Problem Solving (CPS), a setting where these phenomena naturally converge. CPS requires reasoning toward a shared goal, surfacing and negotiating beliefs, and staying aligned in focus. A participant can advance the group's understanding by proposing a new idea, or derail it entirely by shifting attention off task. CPS is not just a label; it is an emergent structure composed of these finer-grained signals.

In this final part, we ask: how reliably can CPS behaviors be detected when data is collected under real-world constraints? And what do these constraints reveal about the broader project of building AI systems that support human learning and collaboration?

In doing so, we complete the arc of the thesis: from individual attention shifts, to models of reasoning and expression of beliefs, to the systems that might one day interpret these phenomena.

Chapter 6

Implications of System choices

6.1 Introduction

In real-world classrooms and collaborative environments, systems designed to detect meaningful group behaviors must operate on noisy, unstructured input. Unlike prior chapters where signals were meticulously modeled, whether internal states or structured reasoning chains, this chapter explores what happens when signal processing is delegated to machines. Specifically, we investigate how automated segmentation and transcription affect the detection of Collaborative Problem Solving (CPS) behaviors in multiparty group dialogue.

CPS is a key 21st-century skill emphasized in frameworks like PISA 2015 [169–171]. A widely used approach for assessing CPS is through identifying markers that categorize group interactions based on behavioral indicators. These CPS markers, as outlined in [83], fall into three primary facets: (1) Constructing Shared Knowledge, where students build common ground and exchange information; (2) Negotiation and Coordination, where they propose solutions, refine ideas, and resolve conflicts; and (3) Maintaining Team Function, which involves regulating interactions and ensuring productive collaboration. Prior research has shown that these markers are predictive of successful learning outcomes [172,173], and automated detection systems have emerged as promising tools for monitoring group progress at scale.

Yet in practice, data arrives messily. Utterances are not neatly separated. Words may be misheard. Automatic speech recognition (ASR) systems and segmentation algorithms make decisions that directly shape the inputs to any downstream classifier. Yet in practice, data arrives messily. Utterances are not neatly separated. Words may be misheard. Automatic speech recognition (ASR) systems and segmentation algorithms make decisions that directly shape the inputs to any downstream classifier. Figure 6.1 illustrates this issue: over the same time window, the Oracle segmentation captures two separate utterances, each with distinct content and collaborative function, while the automated system merges them into one. This leads to both label overlap and transcript compression, which can obscure the speaker turns and degrade downstream CPS classification

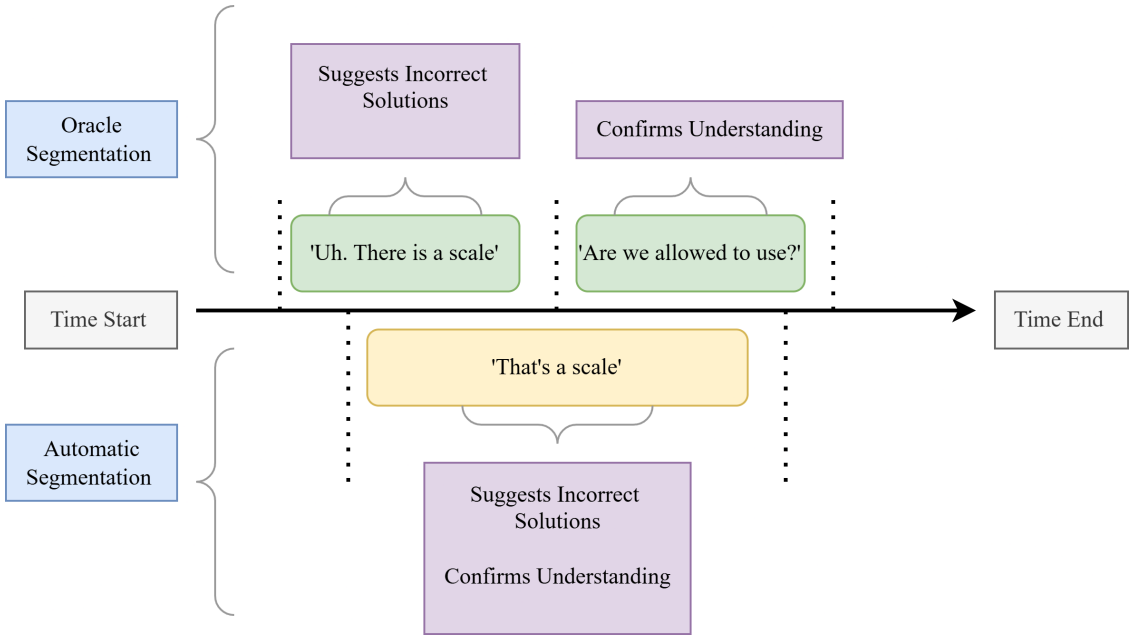

Figure 6.1: Comparison of Oracle segmentation with Oracle transcripts (green) and Google automatic segmentation with Google transcripts(yellow) over the same time span. Oracle preserves distinct utterances, transcripts, and labels; Google merges them, reducing granularity.

Segmentation, in this context, refers to dividing continuous group audio into discrete speaker-aligned utterances. While ASR systems have been extensively evaluated in fields such as broadcast transcription and spoken dialogue systems [174–179], their specific impact within CPS tasks—particularly in educational discourse—remains underexplored. A few studies have examined ASR in classroom dialogue [153, 180], and Cao et al. [181] note that while ASR errors affect lexical tasks, their influence on discourse-level classification is less understood.

This chapter addresses that gap by systematically examining how automation choices affect CPS marker detection. We begin with a high-quality baseline using human-segmented and manually transcribed data, then evaluate the impact of fully automated pipelines, replacing transcription and segmentation with outputs from commercial ASR systems. We also explore the effect of partial transcript availability, simulating real-time classification scenarios where the full dialogue is not yet observed. Our findings suggest that even with incomplete data, there is sufficient signal to predict CPS behaviors with reasonable fidelity.

Finally, we investigate the use of large language models (LLMs), specifically LLaMA 3.1 8B [70], to approximate human annotations. While LLaMA demonstrates robustness to transcription and segmentation noise, often outperforming traditional models under noisy conditions—it struggles to match human, coded

CPS annotations in fully clean settings, likely due to the nuanced and context-dependent nature of CPS behaviors.

Altogether, these experiments probe the limits of machine-mediated signal interpretation in collaborative settings. Can models trained on ideal data still perform when reality intervenes? What structure is lost in translation—and what signal remains? By evaluating these questions, this chapter contributes to our broader thesis arc: understanding how signals of reasoning and interaction manifest in language, and how those signals survive—or degrade—when filtered through machines.

6.2 Methodology

6.2.1 Dataset: The Weights Task Dataset

We utilize the Weights Task Dataset (WTD) [107], as described in Chapter 4 and Chapter 5. Each interaction is annotated for CPS indicators, following the framework outlined by [83]. This framework includes 19 distinct markers of collaborative behavior, with each of them being mapped to constructing shared knowledge, negotiation/coordination, and maintaining team function. The dataset includes Gold-standard (Oracle) segmentation and transcriptions which were manually annotated to serve as the most accurate reference and Google Automatic Speech Recognition (ASR) transcriptions, which was automatically segmented using Google's Voice Activity Detector.

Feature Extraction

We extracted both linguistic and acoustic modalities from the dataset. We extracted linguistic features using BERT-base-small and prosodic features using openSMILE. These were concatenated to form a compact multimodal representation of each utterance, following prior work [66, 182, 183].

6.2.2 Modeling Approach

We employ a supervised learning framework to classify Collaborative Problem Solving (CPS) facets. The classification task is framed as a multi-label classification problem, where each utterance can belong to one or more CPS facets: Constructing Shared Knowledge (Const), Negotiation and Coordination (Neg), and Maintaining Team Function (Maintain).

CPS Indicator Labels

Each utterance was assigned to a CPS facet if it contained any subcategory corresponding to that facet. For example, if an utterance contained both *Suggests appropriate ideas* and *Confirms understanding*, it would be labeled as Const = 1, while Neg = 0 and Maintain = 0, since these two facets fall under *Constructing Shared Knowledge*. This ensured that all relevant utterances contributed to their respective high-level categories. The Google segments inherited the labels of all of the oracle segments that fell under the time span of the Google segment.

Classification Models

We experimented with Random-Forest [184], and AdaBoost [185] following the methodology in [173] to evaluate the effectiveness of different classification models. Deep learning models were not considered, as the size of the dataset was insufficient to support their effective training without overfitting. Hyperparameter tuning was performed using *Hyperopt*, optimizing for *average AUROC scores* over 500 iterations [186]. The search was conducted using the Oracle-Segmented and Oracle-Transcripts condition, and the best-performing model was then evaluated across all other conditions. The best model identified was a Random Forest classifier with criterion set to `entropy`, max features set to `None`, and number of estimators set to `148`. We employ leave-one-group-out cross-validation to evaluate model performance. This is a common technique used to ensure robustness of the machine learning model [187]. We assess model performance using AUROC (Area Under the Receiver Operating Characteristic Curve). AUROC values range from 0 to 1, where 0.5 is random guessing and 1 is perfect [188]. Aditionally, we also calulate the average Precision which measures the proportion of predicted positive labels that are actually correct, and Recall which measures the proportion of actual positive labels that were correctly identified, of the 3 CPS facets, for all conditions.

6.2.3 LLM-Based CPS Detection

We designed a prompt-based approach where utterances are presented alongside their dialogue history, and the LLaMA 3.1 8-b [70] model is asked to categorize them under predefined CPS facets. The prompt provides essential background information about the collaborative task, explaining its three distinct phases

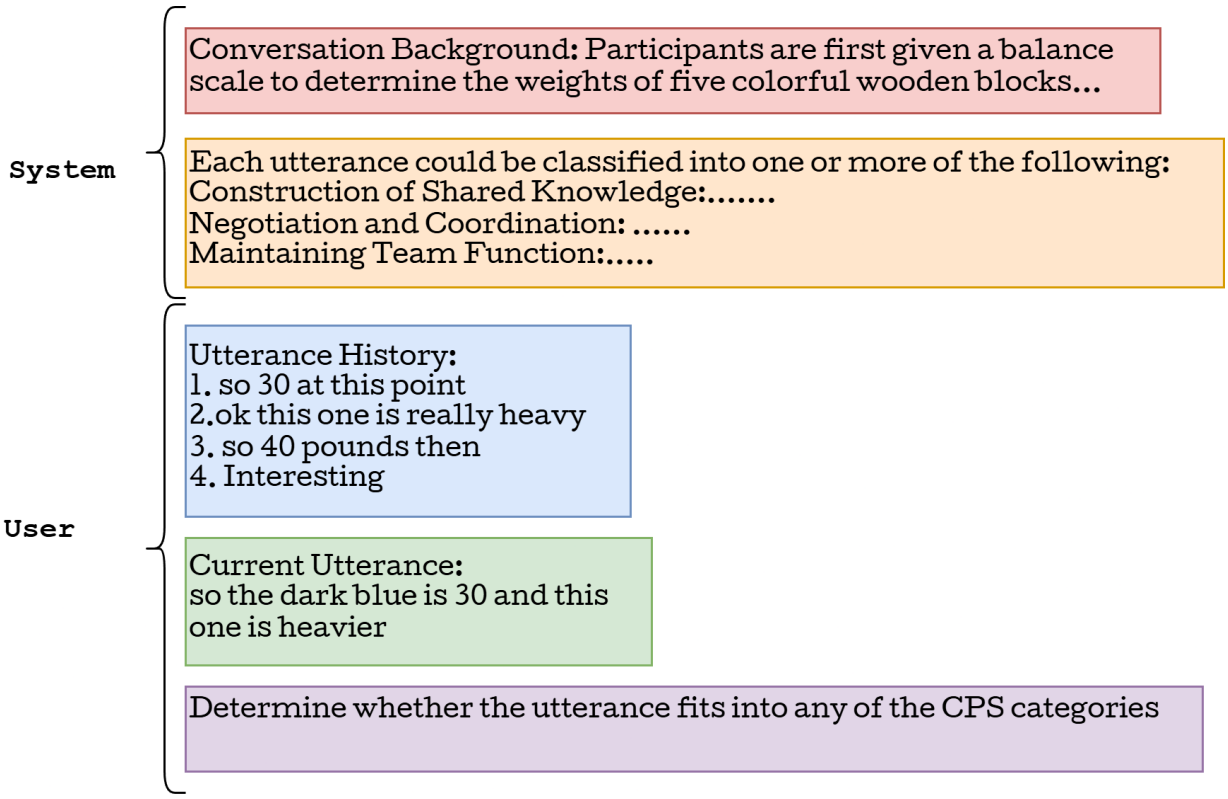

Figure 6.2: A condensed example prompt used for LLaMA-based CPS classification. The system provides background information and CPS category definitions, while the user prompt consists of dialogue history and the current utterance to be classified.

and their objectives. It also explicitly defines the CPS categories and their subtypes, offering detailed descriptions of each classification label.

It contains the immediate conversational history, including up to five previous utterances within the same group, followed by the current utterance to be classified. We provide the model with context by incorporating dialogue history, allowing it to differentiate between isolated statements and collaborative interactions. The prompt explicitly asks the model to categorize the utterance into one or more CPS facets (Constructing Shared Knowledge, Negotiation and Coordination, and Maintaining Team Function).

A sample of a prompt is detailed in Figure 6.2. The prompt utilized the definitions of the CPS facets as given by

6.3 Results

CPS Detection with Oracle Segmentation and Transcription

We first evaluate our models using oracle-segmented utterances with manually transcribed text to establish an upper-bound benchmark for CPS detection performance. This condition represents the highest-quality data scenario. Table 6.1 presents the AUROC performance of CPS detection using oracle-segmented utterances with manually transcribed text.

Table 6.1: AUROC performance of CPS detection using oracle-segmented utterances with manually transcribed text. This benchmark serves as the upper bound for classification accuracy across verbal, acoustic, and combined feature sets.

	Const(±SD)	**Neg(±SD)**	**Maintain(±SD)**
Verbal	0.716 (0.045)	0.765 (0.042)	0.717 (0.051)
Acoustic	0.706 (0.066)	0.592 (0.045)	0.530 (0.061)
Verbal+Acoustic	0.758 (0.044)	0.765 (0.045)	0.720 (0.045)

Impact of Automated Segmentation and Transcription

Table 6.2 presents AUROC scores for different combinations of segmentation and transcription methods. As expected, the highest AUROC is achieved with Oracle transcripts and segmentation (average AUROC = 0.744), while the fully automated condition (Google segmentation and transcription) yields the lowest

performance (average AUROC = 0.679). Notably, using Google segmentatizon with Oracle transcripts nearly matches the Oracle–Oracle performance (0.740 vs. 0.744). Exploring each condition further provides deeper insight into the nature of these effects. Table 6.3 presents the average precision and recall across all CPS facets.

Table 6.2: AUROC scores by segmentation and transcription method. Highest values in bold.

Seg.	Tran.	Const(±SD)	Neg(±SD)	Maintain(±SD)	Average
Oracle	Oracle	**0.758 (0.044)**	**0.765 (0.045)**	**0.720 (0.045)**	**0.744**
Oracle	Google	0.721 (0.029)	0.720 (0.049)	0.653 (0.031)	0.698
Google	Oracle	0.754 (0.066)	0.757 (0.047)	0.699 (0.061)	0.740
Google	Google	0.718 (0.079)	0.705 (0.071)	0.616 (0.078)	0.679

We observe that automated transcription substantially degrades precision (from 0.704 to 0.528) and recall (from 0.297 to 0.246) when manual high-quality segmentation is used. This suggests that transcription quality plays a critical role in preserving classifier performance. Automated segmentation has a more nuanced impact: when using oracle transcripts, precision drops slightly (from 0.704 to 0.658), but recall improves (from 0.297 to 0.339), suggesting that merged utterances may help capture broader CPS behaviors at the cost of granularity.

Interestingly, the fully automated condition (Google segmentation and transcription) achieves the highest recall (0.342) despite lower precision (0.601), indicating potential utility for real-time applications that prioritize broader detection. The poorest-performing condition is Oracle segmentation with Google transcription—likely due to a mismatch between clean segment boundaries and noisy transcriptions.

Effect of Partial Transcripts on CPS Classification

In real-time applications, immediate access to full transcripts is often impractical. Instead of relying on complete dialogues, an effective system should anticipate CPS behaviors even when only partial transcripts are available. To evaluate this, we assess classification performance under conditions with limited data, examining whether early segments of conversations contain enough information for accurate CPS detection. Table 6.4 presents AUROC scores for models trained with reduced transcript availability.

Table 6.3: Average precision and recall across CPS facets. Highest values in bold.

Seg.	Tran.	Avg. Precision	Avg. Recall
Oracle	Oracle	**0.704**	0.297
Oracle	Google	0.528	0.246
Google	Oracle	0.658	0.339
Google	Google	0.601	**0.342**

Table 6.4: AUROC scores for CPS classification using partial transcripts.

Seg.	Tran.	Const(±SD)	Neg(±SD)	Maintain(±SD)
Oracle	Oracle	0.691 (0.029)	0.684 (0.049)	0.625 (0.031)
Oracle	Google	0.731 (0.066)	0.676 (0.048)	0.589 (0.062)

Evaluating LLM-Based CPS Classification

Large Language Models (LLMs) are increasingly being explored for educational applications. However, their effectiveness in CPS classification remains uncertain. We assess whether LLaMA 3.1 8b accurately classifies CPS moves compared to human annotations and traditional machine learning approaches. Table 6.5 reports Cohen's Kappa values for LLaMA-based and Random-Forest-based CPS classification across different segmentation and transcription settings.

Table 6.5: Average Cohen's Kappa values for LLM-based CPS classification compared to human-labeled data.

Model	Segment	Trans	AvgKappa
LLaMA	**Oracle**	**Oracle**	0.223
LLaMA	**Google**	**Google**	0.239
R-F	**Oracle**	**Oracle**	0.213
R-F	**Google**	**Google**	0.190

6.4 Discussion

A striking finding of this study is that the use of automated segmentation and ASR-based transcription does not seem to significantly degrade CPS classification performance compared to oracle transcriptions and human-segmented data. This result is particularly promising, as it suggests that fully automated pipelines

can still achieve a reasonable level of CPS detection accuracy. However, while CPS classification performance remains high with automatic segmentation, we found the number of instances decreased, as shown in Table 6.6. This is due to multiple oracle-segmented utterances being merged into a single segment under Google's system. We observed 518 such instances where multiple oracle segments were combined, with 18 cases merging more than two oracle segments. The underlying distribution of labels reveals important limitations of automated segmentation.

Table 6.6: Number of utterances and CPS labels under Oracle vs. Google segmentation.

	Utterances	**Const**	**Neg**	**Maintain**
Oracle Segments	2482	906	866	391
Google Segments	1824	664	642	328

Merging distinct utterances into a single segment leads to multi-label segments and loss of granularity as seen in Figure 6.1, where the distinction between the two CPS facets is blurred. The merging of segments may obscure important micro-level interactions, potentially limiting an AI-driven system's ability to make fine-grained distinctions in collaborative behaviors, and other downstream tasks [30, 76]. While automated segmentation allows for efficient large-scale CPS classification, it sacrifices interpretability and label precision.

Across all conditions, we also observe that the *Maintain label* consistently performed the worst among all categories. This is likely due to its lower sample size, which limits the model's ability to learn robust patterns for detection. However, prior work has shown that Maintaining Team Function has the strongest correlation with learning gains, making it particularly important to improve the accuracy of its detection for AI-driven classroom support [189]. Special attention should be paid to strategies that enhance the representation and recognition of this behavior to ensure effective intervention by AI agents.

Beyond evaluating the impact of automated segmentation, we also explored the effect of reducing the amount of available transcript data as seen in Table 6.4. Surprisingly, we found that even when models were trained on only the first half of the transcripts, they still retained enough signal to make reliable CPS predictions. This finding has significant implications, as it suggests that CPS labels can be predicted in an anticipatory manner rather than purely reflectively—that is, a CPS move can be inferred even before an

utterance is fully completed. This result indicates that an AI-driven system could potentially make real-time predictions about collaborative behaviors rather than waiting for full utterances to be processed. In classroom applications, this capability is particularly promising, as it would allow for early intervention—alerting educators to potential collaboration breakdowns or productive interactions as they unfold. Rather than operating as a post-hoc analysis tool, an automated CPS monitoring system could function dynamically, enabling adaptive, context-aware interventions that support student collaboration in the moment.

The feasibility of real-time implementation hinges on how reliably early predictions align with final classifications. While our results demonstrate that partial transcript information may be often sufficient for CPS detection, further investigation is needed to understand the trade-off between early predictions and overall accuracy. Future work should explore confidence-based early stopping mechanisms, where predictions are made as soon as the model reaches a certain certainty threshold.

Finally, our evaluation of Large Language Models (LLMs) for CPS classification reveals that LLaMA is more robust to noise introduced by ASR and utterance segmentation automation than traditional Random-Forest models,as seen from the Kappa values in Table 6.5. However, future work needs to investigate if this robustness is linked to LLaMA's access to the richer context provided through the utterance history — Random-forest only makes predictions at the utterance-level, without such context, making the comparison lopsided.

Further prompt engineering could enhance LLM performance, but may reduce the generalizability of our findings to other collaborative contexts. Moreover, deploying LLMs in classroom settings poses significant challenges due to the high computational resources required for inference, the need for real-time processing, and the difficulty of ensuring consistent reliability in complex, dynamic interactions.

Although LLMs technically outperform traditional baselines, its performance has a long way to go before it could potential act as an assistive tools for annotators——e.g., by pre-labeling data for human verification——and their direct use in fully automated CPS detection remains infeasible at this stage. Future research will explore lightweight or fine-tuned architectures optimized for educational settings, balancing computational efficiency with classification accuracy.

Beyond technical feasibility, the ability to automatically detect CPS behaviors has meaningful implications for classroom practice. If integrated into a teacher-facing dashboard, such a system could offer real-time or post-hoc summaries of group interactions, helping educators assess students' collaboration skills.

However, for such insights to be actionable, teachers need to know not just what behavior occurred, but who said what and when. Segmentation choices directly affect this interpretability: over-segmentation can fragment meaningful behaviors, while under-segmentation can blur speaker turns and behavior distinctions. Understanding this trade-off is crucial for real-world deployment. While our study focuses on Google's segmentation system, which tends to under-segment, it does not examine other ASR pipelines that might over-segment or exhibit different boundary behavior. Exploring a wider range of segmentation strategies could yield substantially different outcomes. A single dataset (WTD) [190] was chosen for its annotated CPS labels and controlled task structure. While this ensures consistency, it limits generalizability to other collaborative contexts (e.g., open-ended tasks or diverse age groups). Additionally, our evaluation maps oracle CPS labels onto automatically segmented transcripts. This decision simplifies comparison across segmentation methods but different mapping strategies could result in varying classification performance.

Conclusion Our study demonstrates that fully automated pipelines for CPS detection utilizing ASR-based transcription and segmentation—can achieve promising levels of performance with only minimal degradation compared to oracle transcriptions. This indicates automated CPS monitoring systems in classroom settings are feasible. Further, we show such systems can make accurate predictions even with partial transcript information, opening the door for anticipatory, rather than retrospective, intervention. However, our study also highlights the need for more precise segmentation methods or post-processing strategies are needed to preserve fine-grained interaction. Surprisingly, LLM-based classifications performed relatively well at CPS classification, whether the discourse was automatically or manually processed. Overall, this work highlights both the potential and limitations of current AI-driven approaches for modeling CPS in small group interaction.

Part IV

Discussion

The phenomena studied in this work are abstract, subjective, and deeply human. We pursue them, not to capture them in full, but to catch their traces, to understand how they surface, and to model how they are expressed. This work is not an attempt to achieve state-of-the-art performance on a classification benchmark. At its heart, this work is driven by a fundamental pursuit: the desire to know. To know what others are attending to. To know what they believe. To know what they are thinking, whether they speak it aloud or not, a is a pursuit of interpretability. Across all three parts of this thesis, we have asked how pattern recognition might support this pursuit.

In Part I, we investigated specific instances in which attention turns inward—moments such as the sensation of familiarity or the emergence of spontaneous thoughts. While these experiences are often private and difficult to observe directly, we used behavioral and linguistic signals to examine how they manifest. By leveraging modalities like eye gaze and natural language, this work explores whether such internally directed phenomena leave behind measurable traces, and how those traces might be interpreted.

In Chapter 2, we examined the experience of familiarity using eye gaze data across three experiments. First, we showed that gaze patterns can predict the onset of familiarity, even before it is explicitly reported. Second, we found that participants' eye movements also reflect whether a scene was previously encountered, even when no familiarity was consciously reported—suggesting the presence of implicit recognition. Third, we observed a distinction in gaze behavior between cases of successful versus failed recall, indicating that different forms of familiarity leave different behavioral traces. Together, these findings demonstrate that internal states like familiarity can be detected through subtle, observable patterns in gaze. These findings add to ongoing efforts to study internal cognition by showing that even subtle, subjective experiences like familiarity leave measurable traces in behavior. While detection remains imperfect, consistent patterns in eye gaze reflect different forms of familiarity, offering a path toward better understanding how internal states manifest across individuals and contexts.

In Chapter 3, we explored the idea that internal attention is neither static nor singular. There are many ways in which attention turns inward, and each reveals something different about the how we experinece it. We focused on three spontaneous cognitive experiences, déjà vu (DV), involuntary autobiographical memories (IAM), and unexpected thoughts (UT)—each arising without deliberate intention and often without warning. Through a detailed analysis of participants' descriptions of these experiences, we found that each thought type carried distinctive linguistic markers. These markers were not just surface-level features, but

reflected underlying aspects that aligned with findings from prior psychological literature. By applying natural language processing techniques, we demonstrated that modeling spontaneous thoughts through language is not only viable, but revelatory: linguistic patterns alone were sufficient to distinguish between thought types, suggesting that the language we use encodes meaningful signatures of our internal mental life.

Together, these chapters illustrate that machine learning, particularly models grounded in behavioral and linguistic data—can be a powerful tool for studying these experiences. While no model can fully captures the richness of internal experience, these methods offer avenues for inquiry, augmenting traditional techniques.

In Part II, we turn our attention to two other abstract phenomena: belief and reasoning. Each is fundamental to how people engage with the world—shaping what we understand, what we question, and how we decide. Like attention, both belief and reason manifest differently across individuals. Here, we study how they emerge in a collaborative setting, where they are not just present but necessary to complete the task at hand. Because the task requires coordination, these internal states must be made external—expressed, questioned, and negotiated through dialogue moves. In this context, we see how belief and reasoning become observable

In Chapter 4, we introduced and modeled deliberation chains—structured links between probing utterances and their causal antecedents. These chains give us insight into how groups reason together over time. Rather than treating each utterance in isolation, we ask: what came before this question? What prior statement prompted a moment of tension, ambiguity, or inquiry? In doing so, we capture not only the outcome of group reasoning, but its underlying structure—its arcs, its revisions, its pivots. For educators, such models can illuminate how students arrive at a conclusion. For AI agents, they offer a potential pathway for timely intervention and support during collaborative tasks. Modeling these chains is novel, but our results show that the patterns are learnable. Machine learning once again demonstrates its capacity to uncover meaningful structure in human behavior—even when that structure is implicit, layered, and expressed across people.

In Chapter 5, we examined how belief states manifest in collaborative interaction by extracting explicitly stated propositional content from dialogue. Belief—what someone holds to be true—is a fundamental human phenomenon. It shapes how we interpret the world, how we act within it, and how we relate to others. In collaborative contexts, knowing what another person believes directly informs how we coordinate, clarify, or correct. This chapter focused on capturing such beliefs in the form of structured propositions—those that are explicitly verbalized during group problem-solving. While many beliefs in conversation remain implicit

or unspoken, our approach targets the surface expressions: the direct statements participants make about the task. By formalizing these utterances as propositions, we offer a way to visualize and track the evolving belief landscape within a group. For educators, this can illuminate how student understanding develops over time. For AI systems, it provides an actionable representation of task-related beliefs—helpful in identifying misunderstandings or contradictions that arise in real time.

Together, Chapters 4 and 5 demonstrate how high-level human phenomena—reasoning, deliberation, and belief—surface in dialogue moves and can be modeled through machine learning.

Reason, and expression of beliefs are foundational processes that make collaboration possible. Through careful modeling and representation, we can begin to uncover how they unfold in real-world group settings—and how machines might one day help us understand and support them.

In Part III and chapter 6, we turn our attention to a crucial reality: detecting human phenomena through machine learning does not happen in a vacuum. In applied settings—like classrooms, meetings, or collaborative workspaces—the data we rely on is rarely perfect. Speech is messy. Transcripts are noisy. Segments bleed together. Automatic Transcriptions, Speech disfluencies, utterances being bifocated. Yet if we hope to build systems that detect and support human understanding, this is the world we must contend with. This part focuses on Collaborative Problem Solving (CPS) markers as an integrative lens. These markers, by design, encompass a convergence of the very human phenomena explored in earlier parts of the thesis: attention, belief, reasoning, deliberation. A student sharing an idea is expressing a belief; a clarification may signal a shift in shared understanding; an off-topic remark may reflect a lapse in attention. However, modeling CPS behaviors at scale requires automated systems. The final chapter evaluates how critical design decisions—specifically, how we segment and transcribe speech—affect downstream classification. We find that the fidelity of signal processing directly shapes what models can detect. Merge two utterances and you risk losing a key distinction; mistranscribe a word and the meaning of an exchange may shift entirely. These distortions are not just technical issues—they are epistemological ones. They determine what we are able to know from the data.

We ask: Can collaborative dynamics be detected when the inputs are imperfect? What signal persists despite noise? And what structure is lost in translation? These are the questions that matter not only for performance metrics, but for the broader goal of designing systems that meaningfully support human interaction.

In this way, Part III completes the trajectory of the thesis: from moments of inward attention, to traces of group reasoning, to the systems that must interpret them as they unfold in the real world. It reminds us that every model sits atop a chain of choices—and that if we are to model the richness of human life, we must attend to how that life is first recorded, and rendered into signals we can learn from.

This thesis does not claim that machine learning offers a complete account of attention, belief, or reasoning. These are abstract human phenomena, inherently complex and varied across individuals. Rather, what this work shows is that these leave behind observable traces—signals in gaze, language, or behavior—that, when carefully modeled, offer new ways to detect their emergence.

At the same time, this work does not claim to resolve the ambiguity of human experience, but to make it more visible—to surface the patterns that emerge even in fleeting glances or half-formed thoughts. These patterns are not always consistent or complete, but they offer footholds for inquiry. By drawing from cognitive science, education, and machine learning, this thesis advocates for an approach in which computational tools assist—not replace—our efforts to interpret, support, and enrich human understanding. In doing so, it advances a shared ambition that underlies all three disciplines: the desire to know—and to be known.

Bibliography

[1] Avi Caciularu, Arman Cohan, Iz Beltagy, Matthew Peters, Arie Cattan, and Ido Dagan. CDLM: Cross-document language modeling. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2648–2662, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[2] Shafiuddin Rehan Ahmed, Abhijnan Nath, James H. Martin, and Nikhil Krishnaswamy. $2 * n$ is better than n^2: Decomposing event coreference resolution into two tractable problems. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1569–1583, Toronto, Canada, July 2023. Association for Computational Linguistics.

[3] Daniel C Dennett. *The Intentional Stance*. MIT Press, 1987.

[4] Erving Goffman. *The Presentation of Self in Everyday Life*. Anchor Books, 1959.

[5] H. Paul Grice. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and semantics: Vol. 3: Speech acts*, pages 41–58. Academic Press, 1975.

[6] Herbert H. Clark. *Using Language*. Cambridge University Press, 1996.

[7] Charles Goodwin. Action and embodiment within situated human interaction. *Journal of Pragmatics*, 32(10):1489–1522, 2000.

[8] Adrian Bangerter and Herbert H. Clark. Collaborative referencing in maps: When and why do speakers use names? *Memory & Cognition*, 31(3):403–415, 2003.

[9] Uta Frith and Chris Frith. The neural basis of social cognition. *Annual Review of Psychology*, 63:287–313, 2006.

[10] Leonhard Schilbach et al. Toward a second-person neuroscience. *Behavioral and Brain Sciences*, 36(4):393–414, 2013.

[11] Michael Tomasello. *Becoming Human: A Theory of Ontogeny*. Harvard University Press, 2019.

[12] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT Press, 2nd edition, 2018.

[13] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1103–1114, 2017.

[14] Louis-Philippe Morency, Rada Mihalcea, and Piyush Doshi. Towards multimodal machine learning: A review. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 347–350. ACM, 2011.

[15] Andrew Wilson and Frank Keil. Using neural networks to model theory of mind and predict human behavior. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, 2018.

[16] Vince D. Calhoun and Tulay Adali. The neuroscience of human decision-making through the lens of machine learning. *Neuron*, 109(5):759–775, 2021.

[17] Terra Blevins, Luke Zettlemoyer, and Omer Levy. Attention as explanation: A review of attention-based interpretability in deep learning. *Transactions of the Association for Computational Linguistics*, 9:697–717, 2021.

[18] Tanzeem Choudhury and Alex Pentland. Inferring relevant social networks from interpersonal communication. In *Proceedings of the 19th international conference on World wide web*, pages 301–310, 2009.

[19] Michael I. Posner and Steven E. Petersen. The attention system of the human brain. *Annual review of neuroscience*, 13:25–42, 1990.

[20] Marvin M. Chun, Julie D. Golomb, and Nicholas B. Turk-Browne. Mechanisms of visual attention in the human cortex. *Annual review of psychology*, 62:73–101, 2011.

[21] Jonathan Smallwood and Jonathan W. Schooler. The science of mind wandering: empirically navigating the stream of consciousness. *Annual review of psychology*, 66:487–518, 2015.

[22] Anne M Cleary, Zachary C Irving, and Caitlin Mills. What flips attention? *Cognitive Science*, 47(4):e13274, 2023.

[23] Philip N. Johnson-Laird. *Mental models*. Harvard University Press, 1983.

[24] Jonathan St. B. T. Evans. Dual-processing accounts of reasoning, judgment, and social cognition. *Annual review of psychology*, 59:255–278, 2008.

[25] Gerd Gigerenzer and Daniel G. Goldstein. Reasoning the fast and frugal way: models of bounded rationality. *Psychological review*, 103(4):650–669, 1996.

[26] Hugo Mercier and Dan Sperber. Why do humans reason? arguments for an argumentative theory. *Behavioral and brain sciences*, 34(2):57–74, 2011.

[27] Thomas L. Griffiths, Joshua B. Tenenbaum, et al. Bayesian models of cognition. *The Cambridge handbook of computational psychology*, pages 59–100, 2008.

[28] Herbert H Clark and Susan E Brennan. Grounding in communication. In *Perspectives on socially shared cognition*, volume 13, pages 127–149. APA, 1991.

[29] Robert C. Stalnaker. Common ground. *Linguistics and philosophy*, 25(5):701–721, 2002.

[30] Abhijnan Nath, Videep Venkatesha, Mariah Bradford, Avyakta Chelle, Austin C. Youngren, Carlos Mabrey, Nathaniel Blanchard, and Nikhil Krishnaswamy. "any other thoughts, hedgehog?" linking deliberation chains in collaborative dialogues. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5297–5314, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[31] Videep Venkatesha, Abhijnan Nath, Ibrahim Khebour, Avyakta Chelle, Mariah Bradford, Jingxuan Tu, James Pustejovsky, Nathaniel Blanchard, and Nikhil Krishnaswamy. Propositional extraction from natural speech in small group collaborative tasks. In Benjamin Paaßen and Carrie Demmans Epp, editors, *Proceedings of the 17th International Conference on Educational Data Mining*, pages 169–180, Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society.

[32] George Mandler. Familiarity breeds attempts: A critical review of dual-process theories of recognition. *Perspectives on Psychological Science*, 3(5), 2008.

[33] Anne M. Cleary. Recognition memory, familiarity, and déjà vu experiences. *Current Directions in Psychological Science*, 17(5), 2008.

[34] Katherine L McNeely-White and Anne M Cleary. Piquing curiosity: Déjà vu-like states are associated with feelings of curiosity and information-seeking behaviors. *Journal of Intelligence*, 11(6):112, 2023.

[35] Noah S Okada, Katherine L McNeely-White, Anne M Cleary, Brooke N Carlaw, Daniel L Drane, Thomas D Parsons, Timothy McMahan, Joseph Neisser, and Nigel P Pedersen. A virtual reality paradigm with dynamic scene stimuli for use in memory research. *Behavior Research Methods*, pages 1–24, 2023.

[36] Anne M. Cleary, Anthony J. Ryals, and Jason S. Nomi. Can déjà vu result from similarity to a prior experience? support for the similarity hypothesis of déjà vu. *Psychonomic Bulletin Review*, 16, 2009.

[37] Alan S. Brown and Elizabeth J. Marsh. evoking false beliefs about autobiographical experience. *Psychonomic Bulletin Review*, 15, 2008.

[38] Anne M. Cleary, Alan S. Brown, Benjamin D. Sawyer, Jason S. Nomi, Adaeze C. Ajoku, and Anthony J. Ryals. Familiarity from the configuration of objects in 3-dimensional space and its relation to déjà vu: A virtual reality investigation. *Consciousness and Cognition*, 21(2), 2012.

[39] George Nishimura and Aldo Faisal. Déjà vu: Classification of memory using eye movements. *na*, 2015.

[40] Stephen Hutt, Kristina Krasich, Caitlin Mills, Nigel Bosch, Shelby White, James R. Brockmole, and Sidney K. D'Mello. Automated gaze-based mind wandering detection during computerized learning in classrooms. *User Modeling and User-Adapted Interaction*, 29:821–867, 2019.

[41] Vishal Kuvar, Julia W. Y. Kam, Stephen Hutt, and Caitlin Mills. Detecting when the mind wanders off task in real-time: An overview and systematic review. *ICMI '23: Proceedings of the 25th International Conference on Multimodal Interaction*, pages 163–173, 2023.

[42] Myrthe Faber, Robert Bixler, and Sidney K. D'Mello. An automated behavioral measure of mind wandering during computerized reading. *Behavior Research Methods*, 50:134,150, 2017.

[43] Caitlin Mills, Robert Earl Bixler, Xinyi Wang, and Sidney K. D'Mello. Automatic gaze-based detection of mind wandering during film viewing. In *Educational Data Mining*, 2016.

[44] Yu Imaoka, Andri Flury, and Eling D. de Bruin. Assessing saccadic eye movements with head-mounted display virtual reality technology. *Frontiers in Psychiatry*, 11(572938), 2020.

[45] Ghose. Pytrack: An end-to-end analysis toolkit for eye tracking. *Behavior Research Methods*, 52(2588–2603), 2020.

[46] Vishal Kuvar, Nathaniel Blanchard, Alexander Colby, Laura Allen, and Caitlin Mills. Automatically detecting task-unrelated thoughts during conversations using keystroke analysis. *User Modeling and User-Adapted Interaction*, pages 617–641, 2023.

[47] Angela Stewart, Nigel Bosch, Huili Chen, Patrick J. Donnelly, and Sidney K. D'Mello. Where's your mind at? video-based mind wandering detection during film viewing. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, UMAP '16, page 295–296, New York, NY, USA, 2016. Association for Computing Machinery.

[48] Nathaniel Blanchard, Robert Bixler, Tera Joyce, and Sidney D'Mello. Automated physiological-based detection of mind wandering during learning. In *Intelligent Tutoring Systems: 12th International Conference, ITS 2014, Honolulu, HI, USA, June 5-9, 2014. Proceedings 12*, pages 55–60. Springer International Publishing, 2014.

[49] Marcel Adam Just and Patricia A Carpenter. A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4):329–354, 1980.

[50] Keith Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372–422, 1998.

[51] Nigel Bosch, Sidney D'Mello, Ryan Baker, Jaclyn Ocumpaugh, and Valerie Shute. Detecting student emotions in computer-enabled classrooms. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, pages 4125–4131, 2016.

[52] Robert Bixler, Nathaniel Blanchard, Luke Garrison, and Sidney D'Mello. Automatic gaze-based user-independent detection of mind wandering during computerized reading. *User Modeling and User-Adapted Interaction*, 26:33–68, 2015.

[53] Matthew A Killingsworth and Daniel T Gilbert. A wandering mind is an unhappy mind. *Science*, 330(6006):932–932, 2010.

[54] Kalina Christoff, Zachary C Irving, Kieran CR Fox, R Nathan Spreng, and Jessica R Andrews-Hanna. Mind-wandering as spontaneous thought: a dynamic framework. *Nature reviews neuroscience*, 17(11):718–731, 2016.

[55] Anne M Cleary, Alan S Brown, Benjamin D Sawyer, Jason S Nomi, Adaeze C Ajoku, and Anthony J Ryals. Familiarity from the configuration of objects in 3-dimensional space and its relation to déjà vu: A virtual reality investigation. *Consciousness and cognition*, 21(2):969–975, 2012.

[56] Dorthe Berntsen. Involuntary autobiographical memories and their relation to other forms of spontaneous thoughts. *Philosophical Transactions of the Royal Society B*, 376(1817):20190693, 2021.

[57] Dorthe Berntsen. Involuntary autobiographical memories. *Applied cognitive psychology*, 10(5):435–454, 1996.

[58] Cati Poulos, Andre Zamani, David Pillemer, Michelle Leichtman, Kalina Christoff, and Caitlin Mills. Investigating the appraisal structure of spontaneous thoughts: evidence for differences among unexpected thought, involuntary autobiographical memories, and ruminative thought. *Psychological Research*, 87(8):2345–2364, 2023.

[59] Christopher Steadman, Videep Venkatesha, Cati Poulos, Anne M. Cleary, Nathaniel Blanchard, and Caitlin Mills. Involuntary Thoughts in Older Versus Younger Adults: A Multidisciplinary Approach to Investigating Déjà Vu, Involuntary Autobiographical Memories, and Unexpected Thoughts. *Technology, Mind, and Behavior*, 6(2), apr 16 2025. https://tmb.apaopen.org/pub/6ufsbkgd.

[60] Stephen C. Levinson. *Space in Language and Cognition: Explorations in Cognitive Diversity*. Cambridge University Press, 2003.

[61] Kristen A. Lindquist, Jessica K. MacCormack, and Holly Shablack. The role of language in emotion: Predictions from psychological constructionism. *Frontiers in Psychology*, 6:444, 2015.

[62] Rainer Reisenzein and Michael Junge. The language of emotions: An analysis of a semantic field. *Cognition and Emotion*, 26(4):785–796, 2012.

[63] Anne M Cleary and Alan S Brown. *The déjà vu experience*. Routledge, 2021.

[64] Dorthe Berntsen and David C Rubin. Emotionally charged autobiographical memories across the life span: The recall of happy, sad, traumatic and involuntary memories. *Psychology and aging*, 17(4):636, 2002.

[65] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*. Cambridge University Press, 2008.

[66] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[67] Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282, 2012.

[68] Serhad Sarica and Jianxi Luo. Stopwords in technical language processing. *Plos one*, 16(8):e0254937, 2021.

[69] Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. Carer: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3687–3697, 2018.

[70] A. Grattafiori et al. The llama 3 herd of models. *arXiv*, 2024.

[71] Paul Ekman. Are there basic emotions? *Psychological Review*, 1992.

[72] Joseph Neisser, George Abreu, Daniel L Drane, Nigel P Pedersen, Thomas D Parsons, and Anne M Cleary. Opening a conceptual space for metamemory experience. *New ideas in psychology*, 69:100995, 2023.

[73] Anne M Cleary, Cati Poulos, and Caitlin Mills. A possible shared underlying mechanism among involuntary autobiographical memory and déjà vu. *Behavioral & Brain Sciences*, 46, 2023.

[74] Nathaniel Blanchard, Robert Bixler, Tera Joyce, and Sidney D'Mello. Automated physiological-based detection of mind wandering during learning. In *Intelligent Tutoring Systems: 12th International Conference, ITS 2014, Honolulu, HI, USA, June 5-9, 2014. Proceedings 12*, pages 55–60. Springer, 2014.

[75] Gabriel King Smith, Caitlin Mills, Alexandra Paxton, and Kalina Christoff. Mind-wandering rates fluctuate across the day: Evidence from an experience-sampling study. *Cognitive research: principles and implications*, 3:1–20, 2018.

[76] Iliana Castillon, Videep Venkatesha, Hannah VanderHoeven, Mariah Bradford, Nikhil Krishnaswamy, and Nathaniel Blanchard. Multimodal features for group dynamic-aware agents. In *Interdisciplinary Approaches to Getting AI Experts and Education Stakeholders Talking Workshop at AIEd. International AIEd Society*, pages 1–6, Durham, UK, 2022. Springer Cham.

[77] 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.

[78] Julie Hunter, Nicholas Asher, and Alexandra Lascarides. A formal semantics for situated conversation. *Semantics and Pragmatics*, 11:1–52, 2018.

[79] Katherine Atwell, Mert Inan, Anthony B Sicilia, and Malihe Alikhani. Combining discourse coherence with large language models for more inclusive, equitable, and robust task-oriented dialogue. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3538–3552, 2024.

[80] 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.

[81] Jessica Andrews-Todd and Carol M. Forsyth. Exploring social and cognitive dimensions of collaborative problem solving in an open online simulation-based task. *Computers in Human Behavior*, 104:105759, March 2020.

[82] OECD. *PISA 2015 Assessment and Analytical Framework*. OECD, 2017.

[83] Chen Sun, Valerie J. Shute, Angela Stewart, Jade Yonehiro, Nicholas Duran, and Sidney D'Mello. Towards a generalized competency model of collaborative problem solving. *Computers & Education*, 143:103672, 2020.

[84] Friedrich Hesse, Esther Care, Juergen Buder, Kai Sassenberg, and Patrick Griffin. A Framework for Teachable Collaborative Problem Solving Skills. In Patrick Griffin and Esther Care, editors, *Assessment and Teaching of 21st Century Skills: Methods and Approach*, Educational Assessment in an Information Age, pages 37–56. Springer Netherlands, Dordrecht, 2015.

[85] Riccardo Fusaroli, Kristian Tylén, Katrine Garly, Jakob Steensig, Morten H. Christiansen, and Mark Dingemanse. Measures and mechanisms of common ground: Backchannels, conversational repair, and interactive alignment in free and task-oriented social interactions. pages 2055–2060. Cognitive Science Society, 2017.

[86] 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, November 2018. Publisher: SAGE Publications Inc.

[87] Georgi Karadzhov, Tom Stafford, and Andreas Vlachos. What makes you change your mind? an empirical investigation in online group decision-making conversations. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 552–563, 2022.

[88] Robert C Stalnaker. Assertion. In *Pragmatics*, pages 315–332. Brill, 1978.

[89] Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. End-to-end neural coreference resolution. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[90] Rui Zhang, Cicero dos Santos, Michihiro Yasunaga, Bing Xiang, and Dragomir Radev. Neural coreference resolution with deep biaffine attention by joint mention detection and mention clustering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 102–107, 2018.

[91] Arie Cattan, Alon Eirew, Gabriel Stanovsky, Mandar Joshi, and Ido Dagan. Cross-document coreference resolution over predicted mentions. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5100–5107, Online, August 2021. Association for Computational Linguistics.

[92] Xiaodong Yu, Wenpeng Yin, and Dan Roth. Pairwise representation learning for event coreference. In Vivi Nastase, Ellie Pavlick, Mohammad Taher Pilehvar, Jose Camacho-Collados, and Alessandro Raganato, editors, *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 69–78, Seattle, Washington, July 2022. Association for Computational Linguistics.

[93] William Held, Dan Iter, and Dan Jurafsky. Focus on what matters: Applying discourse coherence theory to cross document coreference. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1406–1417, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[94] Michael Bugert, Nils Reimers, and Iryna Gurevych. Generalizing cross-document event coreference resolution across multiple corpora. *Computational Linguistics*, 47(3):575–614, 2021.

[95] Shafiuddin Rehan Ahmed, George Arthur Baker, Evi Judge, Michael Reagan, Kristin Wright-Bettner, Martha Palmer, and James H. Martin. Linear cross-document event coreference resolution with X-AMR. 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 10517–10529, Torino, Italia, May 2024. ELRA and ICCL.

[96] Xiaolong Wang, Yile Wang, Yuanchi Zhang, Fuwen Luo, Peng Li, Maosong Sun, and Yang Liu. Reasoning in conversation: Solving subjective tasks through dialogue simulation for large language models, 2024.

[97] Ansh Radhakrishnan, Karina Nguyen, Anna Chen, Carol Chen, Carson Denison, Danny Hernandez, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, et al. Question decomposition improves the faithfulness of model-generated reasoning. *arXiv preprint arXiv:2307.11768*, 2023.

[98] Wenting Zhao, Justin T Chiu, Claire Cardie, and Alexander M Rush. Abductive commonsense reasoning exploiting mutually exclusive explanations. *arXiv preprint arXiv:2305.14618*, 2023.

[99] 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.

[100] Peter West, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. Symbolic knowledge distillation: from general language models to commonsense models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4602–4625, Seattle, United States, July 2022. Association for Computational Linguistics.

[101] 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 Kevin Duh, Helena Gomez, and Steven Bethard, editors, *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, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[102] Nicholas Asher and Alex Lascarides. *Logics of conversation*. Cambridge University Press, 2003.

[103] Katya Alahverdzhieva, Alex Lascarides, and Dan Flickinger. Aligning speech and co-speech gesture in a constraint-based grammar. *Journal of Language Modelling*, 5, 2017.

[104] Vincent Ng and Claire Cardie. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 104–111, 2002.

[105] Heeyoung Lee, Marta Recasens, Angel Chang, Mihai Surdeanu, and Dan Jurafsky. Joint entity and event coreference resolution across documents. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 489–500, 2012.

[106] Peter C Wason. Reasoning about a rule. *Quarterly journal of experimental psychology*, 20(3):273–281, 1968.

[107] 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, 2024.

[108] Jingxuan Tu, Kyeongmin Rim, Eben Holderness, Bingyang Ye, and James Pustejovsky. Dense paraphrasing for textual enrichment. In *Proceedings of the 15th International Conference on Computational Semantics*, pages 39–49, 2023.

[109] Michael Bugert, Nils Reimers, Shany Barhom, Ido Dagan, and Iryna Gurevych. Breaking the subtopic barrier in cross-document event coreference resolution. In *Text2Story @ ECIR*, pages 23–29, 2020.

[110] 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.

[111] Jenny Kunz, Martin Jirenius, Oskar Holmström, and Marco Kuhlmann. Human ratings do not reflect downstream utility: A study of free-text explanations for model predictions. In Jasmijn Bastings, Yonatan Belinkov, Yanai Elazar, Dieuwke Hupkes, Naomi Saphra, and Sarah Wiegreffe, editors, *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 164–177, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics.

[112] Sahithya Ravi, Chris Tanner, Raymond Ng, and Vered Shwartz. What happens before and after: Multi-event commonsense in event coreference resolution. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1700–1716, 2023.

[113] Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *International Conference on Learning Representations*, pages 1–14, Addis Ababa, Ethiopia, 2020. ICLR.

[114] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *CoRR*, abs/2004.05150:1–17, 2020.

[115] Barbara J Grosz and Candace L Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204, 1986.

[116] Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does bert learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, 2019.

[117] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution, 2022.

[118] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[119] Nafise Sadat Moosavi, Leo Born, Massimo Poesio, and Michael Strube. Using automatically extracted minimum spans to disentangle coreference evaluation from boundary detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Florence, Italy, 2019. Association for Computational Linguistics.

[120] Nafise Sadat Moosavi and Michael Strube. Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 632–642, 2016.

[121] Nicholas Asher, Julie Hunter, and Kate Thompson. Modelling structures for situated discourse. *Dialogue & Discourse*, 11:89–121, 2020.

[122] Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. Deep knowledge tracing. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28, pages 505 – 513, Montreal, Canada, 2015. Curran Associates, Inc.

[123] Chen Sun, Valerie J Shute, Angela Stewart, Jade Yonehiro, Nicholas Duran, and Sidney D'Mello. Towards a generalized competency model of collaborative problem solving. *Computers & Education*, 143:103672, 2020.

[124] Noreen M Webb, Marsha Ing, Eric Burnheimer, Nicholas C Johnson, Megan L Franke, and Joy Zimmerman. Is there a right way? Productive patterns of interaction during collaborative problem solving. *Education Sciences*, 11(5):214, 2021.

[125] Jeremy Roschelle and Stephanie D. Teasley. The construction of shared knowledge in collaborative problem solving. In Claire O'Malley, editor, *Computer Supported Collaborative Learning*, pages 69–97, Berlin, Heidelberg, 1995. Springer Berlin Heidelberg.

[126] Hannie Gijlers and Ton de Jong. Sharing and confronting propositions in collaborative inquiry learning. *Cognition and Instruction*, 27(3):239–268, 2009.

[127] Simon Dennis. An unsupervised method for the extraction of propositional information from text. *Proceedings of the National Academy of Sciences*, 101(suppl_1):5206–5213, 2004.

[128] Steven Bethard, Hong Yu, Ashley Thornton, Vasileios Hatzivassiloglou, and Dan Jurafsky. Automatic extraction of opinion propositions and their holders. In *Exploring Attitude and Affect in Text: Theories and Applications*, Papers from the 2004 AAAI Spring Symposium, pages 20–27, Palo Alto, California, 2004. AAAI.

[129] Vineeta Chand, Kathleen Baynes, Lisa M Bonnici, and Sarah Tomaszewski Farias. A rubric for extracting idea density from oral language samples. *Current Protocols in Neuroscience*, 58(1):10–5, 2012.

[130] Yohan Jo, Jacky Visser, Chris Reed, and Eduard Hovy. A cascade model for proposition extraction in argumentation. In Benno Stein and Henning Wachsmuth, editors, *Proceedings of the 6th Workshop on Argument Mining*, pages 11–24, Florence, Italy, August 2019. Association for Computational Linguistics.

[131] Yohan Jo, Jacky Visser, Chris Reed, and Eduard Hovy. Extracting implicitly asserted propositions in argumentation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the*

2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 24–38, Online, November 2020. Association for Computational Linguistics.

[132] Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In Stefan Riezler and Yoav Goldberg, editors, *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany, August 2016. Association for Computational Linguistics.

[133] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. *Advances in Neural Information Processing Systems*, 28:1693–1701, 2015.

[134] Jiwei Li, Thang Luong, and Dan Jurafsky. A hierarchical neural autoencoder for paragraphs and documents. In Chengqing Zong and Michael Strube, editors, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1106–1115, Beijing, China, July 2015. Association for Computational Linguistics.

[135] Shafiuddin Rehan Ahmed, Abhijnan Nath, Michael Regan, Adam Pollins, Nikhil Krishnaswamy, and James H. Martin. How good is the model in model-in-the-loop event coreference resolution annotation? In Jakob Prange and Annemarie Friedrich, editors, *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 136–145, Toronto, Canada, July 2023. Association for Computational Linguistics.

[136] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.

[137] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, pages 1–43, Virtual, 2020. ICLR.

[138] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[139] Yutao Zeng, Xiaolong Jin, Saiping Guan, Jiafeng Guo, and Xueqi Cheng. Event coreference resolution with their paraphrases and argument-aware embeddings. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3084–3094, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[140] Dedre Gentner. Testing the psychological reality of a representational model. In David L. Waltz, editor, *Theoretical Issues in Natural Language Processing-2*, pages 1–7, Las Cruces, New Mexico, 1978. Association for Computational Linguistics.

[141] Haixia Chai and Michael Strube. Incorporating centering theory into neural coreference resolution. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2996–3002, Seattle, United States, July 2022. Association for Computational Linguistics.

[142] Sungho Jeon and Michael Strube. Centering-based neural coherence modeling with hierarchical discourse segments. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7458–7472, Online, November 2020. Association for Computational Linguistics.

[143] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International*

Conference on Neural Information Processing Systems, NIPS'17, pages 6000—-6010, Red Hook, NY, USA, 2017. Curran Associates Inc.

[144] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. In *International Conference on Learning Representations*, pages 1–15, Addis Ababa, Ethiopia, 2020. ICLR.

[145] Abhijnan Nath, Sheikh Mannan, and Nikhil Krishnaswamy. AxomiyaBERTa: A phonologically-aware transformer model for Assamese. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11629–11646, Toronto, Canada, July 2023. Association for Computational Linguistics.

[146] Nathaniel Blanchard, Daniel Moreira, Aparna Bharati, and Walter Scheirer. Getting the subtext without the text: Scalable multimodal sentiment classification from visual and acoustic modalities. In Amir Zadeh, Paul Pu Liang, Louis-Philippe Morency, Soujanya Poria, Erik Cambria, and Stefan Scherer, editors, *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, pages 1–10, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[147] 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 Harry Bunt, editor, *Proceedings of the 19th Joint ACL-ISO Workshop on Interoperable Semantics (ISA-19)*, pages 75–81, Nancy, France, June 2023. Association for Computational Linguistics.

[148] Benjamin Ibarra, Brett Wisniewski, Corbyn Terpstra, Videep Venkatesha, Mariah Bradford, and Nathaniel Blanchard. Investigating automated transcriptions for multimodal cps detection in group-work. In *International Conference on Human-Computer Interaction*, pages 214–224. Springer, 2025.

[149] Patrick J. Donnelly, Nathaniel Blanchard, Borhan Samei, Andrew M. Olney, Xiaoyi Sun, Brooke Ward, Sean Kelly, Martin Nystrand, and Sidney K. D'Mello. Multi-sensor modeling of teacher instructional segments in live classrooms. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 177–184, Tokyo, Japan, 2016. Association for Computing Machinery.

[150] Patrick J. Donnelly, Nathaniel Blanchard, Andrew M. Olney, Sean Kelly, Martin Nystrand, and Sidney K. D'Mello. Words matter: Automatic detection of teacher questions in live classroom discourse using linguistics, acoustics, and context. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, LAK '17, pages 218–227, New York, NY, USA, 2017. ACM.

[151] Nathaniel Blanchard, Patrick Donnelly, Andrew M. Olney, Borhan Samei, Brooke Ward, Xiaoyi Sun, Sean Kelly, Martin Nystrand, and Sidney K. D'Mello. Identifying teacher questions using automatic speech recognition in classrooms. In Raquel Fernandez, Wolfgang Minker, Giuseppe Carenini, Ryuichiro Higashinaka, Ron Artstein, and Alesia Gainer, editors, *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 191–201, Los Angeles, September 2016. Association for Computational Linguistics.

[152] Robert Bixler, Nathaniel Blanchard, Luke Garrison, and Sidney D'Mello. Automatic detection of mind wandering during reading using gaze and physiology. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, page 299–306, New York, NY, USA, 2015. Association for Computing Machinery.

[153] Mariah Bradford, Paige Hansen, J. Ross Beveridge, Nikhil Krishnaswamy, and Nathaniel Blanchard. A deep dive into microphone hardware for recording collaborative group work. In Antonija Mitrovic and Nigel Bosch, editors, *Proceedings of the 15th International Conference on Educational Data Mining*, pages 588–593, Durham, United Kingdom, July 2022. International Educational Data Mining Society.

[154] Eric Pacuit. *Neighborhood Semantics for Modal Logic*. Springer Publishing Company, Incorporated, New York, NY, 1st edition, 2017.

[155] Jonathan St BT Evans. Reasoning, biases and dual processes: The lasting impact of Wason (1960). *Quarterly Journal of Experimental Psychology*, 69(10):2076–2092, 2016.

[156] Shalom Lappin and Herbert J Leass. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561, 1994.

[157] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[158] Jingxuan Tu, Kyeongmin Rim, and James Pustejovsky. Competence-based question generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1521–1533, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

[159] Jingxuan Tu, Kyeongmin Rim, Eben Holderness, Bingyang Ye, and James Pustejovsky. Dense paraphrasing for textual enrichment. In *Proceedings of the 15th International Conference on Computational Semantics (IWCS)*, pages 39–49, Nancy, France, June 2023. Association for Computational Linguistics.

[160] Jonáš Kulhánek, Vojtěch Hudeček, Tomáš Nekvinda, and Ondřej Dušek. AuGPT: Auxiliary tasks and data augmentation for end-to-end dialogue with pre-trained language models. In Alexandros Papangelis, Paweł Budzianowski, Bing Liu, Elnaz Nouri, Abhinav Rastogi, and Yun-Nung Chen, editors, *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 198–210, Online, November 2021. Association for Computational Linguistics.

[161] Lucas Francisco Amaral Orosco Pellicer, Taynan Maier Ferreira, and Anna Helena Reali Costa. Data augmentation techniques in natural language processing. *Applied Soft Computing*, 132:109803, 2023.

[162] Sahithya Ravi, Chris Tanner, Raymond Ng, and Vered Shwartz. What happens before and after: Multi-event commonsense in event coreference resolution. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1708–1724, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[163] Sebastian Hofstätter, Aldo Lipani, Sophia Althammer, Markus Zlabinger, and Allan Hanbury. Mitigating the position bias of transformer models in passage re-ranking. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part I 43*, pages 238–253, virtual event, 2021. Springer.

[164] Xiaohan Yang, Eduardo Peynetti, Vasco Meerman, and Chris Tanner. What GPT knows about who is who. In Shabnam Tafreshi, João Sedoc, Anna Rogers, Aleksandr Drozd, Anna Rumshisky, and Arjun Akula, editors, *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 75–81, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[165] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[166] Hannah VanderHoeven, Brady Bhalla, Ibrahim Khebour, Austin 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. *arXiv preprint arXiv:2503.09511*, 2025.

[167] 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.

[168] Sifatul Anindho, Videep Venkatesha, Mariah Bradford, Anne M Cleary, and Nathaniel Blanchard. An exploration of internal states in collaborative problem solving. In *International Conference on Human-Computer Interaction*, pages 135–150. Springer, 2025.

[169] OECD. Pisa 2015 assessment and analytical framework: Science, reading, mathematic and financial literacy. *Paris: OECD Publishing*, 2017.

[170] Think: Kids. Collaborative problem solving. *https://thinkkids.org/Schools/*, 2025.

[171] Eckhard Klieme, Johannes Hartig, Daniel Rauch, and Wolfgang Blum. Assessing collaborative problem solving: An overview of the pisa 2015 assessment framework. In *Collaborative problem solving: An educational perspective*, pages 31–53. Springer, 2016.

[172] Michael Flor, Su-Youn Yoon, Ou Lydia Liu, and Michael Wagner. Automated classification of collaborative problem solving interactions in simulated science tasks. *ETS Research Report Series*, 2016(1):1–12, 2016.

[173] 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.

[174] Mark J. Gales and Steve J. Young. The application of hidden markov models in speech recognition. *Foundations and Trends in Signal Processing*, 1(3):195–304, 2008.

[175] Dong Yu and Li Deng. *Automatic Speech Recognition: A Deep Learning Approach*. Springer, 2014.

[176] Su-Youn Yoon, Yiting Xue, and Mark Warschauer. Speech-to-text for literacy: How asr errors affect educational applications. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5059–5070, 2020.

[177] Klaus Zechner, Keelan Evanini, Su-Youn Yoon, and Xinhao Wang. The challenges of asr in automated speaking assessment. In *Proceedings of the Workshop on Speech and Language Technology in Education (SLaTE)*, 2015.

[178] Ramesh Manuvinakurike and David DeVault. Using asr word confusion networks for modeling decisions in spoken dialogue systems. In *Proceedings of Interspeech*, 2015.

[179] Keelan Evanini, Derrick Higgins, and Klaus Zechner. Automated speech scoring for non-native middle school students with multiple task types. In *Proceedings of Interspeech*, 2013.

[180] Nathaniel Blanchard, Michael Brady, Andrew M Olney, Marci Glaus, Xiaoyi Sun, Martin Nystrand, Borhan Samei, Sean Kelly, and Sidney D'Mello. A study of automatic speech recognition in noisy classroom environments for automated dialog analysis. In *Artificial Intelligence in Education: 17th International Conference, AIED 2015, Madrid, Spain, June 22-26, 2015. Proceedings 17*, pages 23–33. Springer, 2015.

[181] Jie Cao, Ananya Ganesh, Jon Cai, Rosy Southwell, E. Margaret Perkoff, Michael Regan, Katharina Kann, James H. Martin, Martha Palmer, and Sidney D'Mello. A comparative analysis of automatic speech recognition errors in small group classroom discourse. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization (UMAP '23)*, pages 250–261, Limassol, Cyprus, 2023. ACM.

[182] Florian Eyben, Klaus R. Scherer, Björn W. Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong. The

geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202, 2016.

[183] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462, 2010.

[184] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[185] Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[186] James Bergstra, Daniel Yamins, and David Cox. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pages 115–123, 2013.

[187] Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Springer, 1993.

[188] Tom Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

[189] Sidney K D'Mello, Nicholas Duran, Amanda Michaels, and Angela EB Stewart. Improving collaborative problem-solving skills via automated feedback and scaffolding: a quasi-experimental study with cpscoach 2.0. *User Modeling and User-Adapted Interaction*, pages 1–39, 2024.

[190] Ibrahim Khebour, Richard Brutti, Indrani Dey, Rachel Dickler, Kelsey Sikes, Kenneth Lai, Mariah Bradford, Brittany Cates, Paige Hansen, Changsoo Jung, Brett Wisniewski, Corbyn Terpstra, Leanne Hirshfield, Sadhana Puntambekar, Nathaniel Blanchard, James Pustejovsky, and Nikhil Krishnaswamy. The Weights Task Dataset: A Multimodal Dataset of Collaboration in a Situated Task, September 2023.