From Scaling Laws to Agentic Grounding: A Survey of Emerging Reliability Risks in Large Language Models

Jason Ford
Colorado State University
Fort Collins, CO, USA
jason.ford@colostate.edu

***Abstract*—As the artificial intelligence industry exhausts high-quality, human-generated datasets, the shift toward recursive training on synthetic outputs has introduced significant systemic instabilities. This paper analyzes the compounding risks of model collapse and functional incoherence as failure modes that can result in the irreversible erosion of data diversity and model stability. The results of this analysis are an urgent call to both industry and the research community to recognize that current scaling laws are insufficient for resolving these defects, as increased compute may actually broaden the surface area for incoherent execution. To address these vulnerabilities, a transition toward agentic architectures is proposed that implements non-machine learning constraints as a way to ground model outputs. By integrating structural anchors like state-based persistence and parallel consensus loops, these frameworks provide a deterministic foundation for maintaining logical consistency in high-consequence environments. This research suggests that the path to sustainable, trustworthy AI lies in moving beyond purely probabilistic scaling toward verifiable, context-dependent reliability.**

***Keywords*—Machine Learning, Artificial Intelligence, Large Language Model, Model Collapse, Functional Incoherence, Data Provenance, RAG Safety, Algorithmic Bias**

I. Introduction

Large Language Model (LLM) development has largely been dictated by the scaling laws paradigm, which posits that increases in model size, compute, and data volume yield predictable improvements in performance [1]. This has resulted in the emergence of frontier models as powerful few-shot learners [2]. However, as the AI industry approaches the limits of human-generated data, a significant depletion of human-generated data has emerged [3]. To circumvent this, developers have increasingly turned to synthetic data, inadvertently triggering self-consuming, recursive training loops [4]. Within this recursive cycle, the model essentially absorbs its own generational artifacts, leading to failures that are not only significant, but potentially irreversible.

Contemporary research has found that models trained on their own outputs are susceptible to collapse as they progressively forget the underlying data distribution, leading to irreversible model flaws [5]. In addition, the diversity of model outputs progressively decreases when models do not have a sufficient sampling bias to maintain quality. Recursive training also creates fairness feedback loops where systematic shifts in the model's internal data distribution amplify existing biases, further marginalizing already at-risk groups [6]. While these shifts represent a statistical decay of the data, they also manifest as a fundamental breakdown in expected model behavior.

Recent findings identify functional incoherence as another volatile mode of failure [7]. As models engage in longer reasoning sequences, their failures become increasingly incoherent, stemming from high variance that additional scale is unlikely to resolve. This behavior suggests that a given model's perceived intelligence does not guarantee that it will be capable of reliably completing complex tasks [8]. This gap between latent capability and reliable execution undermines the foundations for trustworthy AI.

Because trustworthiness is a relational and context-dependent concept rather than a universal model attribute, purely training-level solutions are likely insufficient [9]. Instead, a socio-technical approach is required that prioritizes the user's decision context and external verifiability. This paper reviews the literature surrounding the conditions that precipitate model collapse and functional incoherence. The result of this review proposes that agentic architectures utilizing non-ML constraints can serve as anchors that help ground models while mitigating these unpredictable failure conditions.

The investigation into this topic is driven by the need to understand the practical application of deep learning architectures in real-world scenarios. The primary works by Shumailov et al. and Alemohammad et

al. were selected to establish the statistical limits and mathematical risks of training models on synthetic data. These were contrasted with the more recent research from Hägele et al. that characterizes the reasoning-time failures occurring with frontier models that use iterative "thinking" loops. The goal of this survey is to evaluate whether the scaling laws that have historically dictated LLM development remain a viable path for ensuring operational safety, or if the field is approaching a structural ceiling that necessitates a shift toward more grounded, agentic model development. This remainder of this paper is organized as follows:

- *Section II*: Establishes the survey scope by defining related work and the criteria used for exclusion.
- *Section III*: Analyzes the mechanisms of recursive decay, statistical convergence and autophagous loops identified in training-time research.
- *Section IV*: Investigates functional incoherence and the stochastic failure modes that emerge in long-chain reasoning.
- *Section V*: Explores the socio-technical dimensions of AI trustworthiness and the shift toward relational evaluation.
- *Section VI*: Proposes a mitigation framework utilizing data provenance and agentic architectures.
- *Section VII*: Provides a comparative analysis addressing the strengths and weaknesses of the primary sources reviewed for this survey.
- *Section VIII*: Synthesizes the technical commonalities and unifying themes discovered during this review.
- *Section IX*: Identifies open questions regarding task complexity and synthetic data saturation.
- *Section X*: Summarizes the findings of this research and the implications for sustainable model development.

II. Related Work

This survey focuses on the structural decay of LLMs and potential agentic mitigations. Several other research avenues were considered but ultimately excluded in order to maintain focus on architectural reliability. These include:

- *Human-in-the-Loop Curation*: While there is a significant body of work on using human feedback in model fine-tuning, these were ignored because they don't provide a solution for the underlying mathematical challenges where the base training distribution is already contaminated.
- *Efficiency and Quantization*: Research into model quantization and compression techniques were excluded as these tend to focus more on addressing inference speed rather than the logical coherence and distributional stability that are the focus of this review.
- *Adversarial Attacks*: Significant research has been conducted on techniques like jailbreaking and prompt injection. These were omitted from review in lieu of prioritizing systemic internal failure modes over external manipulations of pre-trained models.

III. Mechanisms of Recursive Decay

The transition toward recursive training where successive models are trained on the outputs of previous generations is primarily driven by the projected depletion of high-quality, human-generated data [10]. While this approach may address short-term data scarcity needs, it introduces systemic failures characterized as a self-consuming or autophagous loop [4].

A. The Curse of Recursion and Model Collapse

Model collapse is the process where generative models progressively forget the underlying data distribution of their predecessors. Unlike concept drift in traditional machine learning which typically stems from evolving external data, model collapse is an internal, potentially irreversible defect [5]. When models are trained on generated content, the tails of the original distribution which represent rare or complex data points tend to disappear first. This leads to a state where the model only generates a narrow subset of the reference distribution [6]. Experimental evidence suggests that once the original data sources are lost, the model's ability to represent the full complexity of human interaction is permanently diminished [12].

B. Model Autophagy Disorder (MAD)

Alemohammad et al. refine this understanding through the lens of Model Autophagy Disorder (MAD), which diagnoses the specific conditions under which the quality and diversity of outputs progressively decrease [4]. Their research indicates that without a significant sampling bias to filter for quality, generative models trained in a fully synthetic loop experience significant reliability drops within approximately 10 generations. A critical finding is the existence of an admissible threshold for synthetic samples for each model. This suggests that the ratio of synthetic to real data a model can ingest before collapsing is heavily dependent on the sampling bias employed during the generation process. Nevertheless, this reliance introduces an important trade-off, given that the same mechanisms used to preserve quality can actually accelerate the narrowing of the model's worldview.

C. Model-Induced Distribution Shifts (MIDS) and Bias

These trade-offs are best described as Model-Induced Distribution Shifts (MIDS), which are the specific statistical convergence that occurs over successive generations of training. Wyllie et al. demonstrate that when synthetic outputs contaminate new training datasets, the resulting data distribution converges toward a point estimate [6]. While this convergence may temporarily mask quality loss, it amplifies unfairness and bias. As the distribution narrows, errors and biases present in early generations are embedded and amplified, leading to a representative disparity where already marginalized groups are effectively erased from the model's operational reality [11]. These compounding failures suggest that the use of human-curated data is not merely a preference but a mathematical requirement for long-term model stability.

IV. Functional Incoherence

While the statistical erosion of a model's foundational data distribution is an obvious threat to reliability, recent research identifies functional incoherence as a more immediate operational threat. This phenomenon describes a state where models optimized for complex reasoning fail to maintain logical or procedural consistency when executing long sequences of actions. Unlike model collapse, which is a training-time degradation, functional incoherence manifests at inference time and is intrinsically linked to the complexity of the task and the depth of the model's reasoning process [7].

A. Inherent Variance and Emergent Stochasticity

Functional incoherence is quantified by measuring the fraction of a model's error that stems from variance rather than bias over repeated test-time trials. Hägele et al. have demonstrated that as frontier models spend more time navigating multi-step action loops, their failure modes shift from predictable, biased errors toward highly stochastic, incoherent behaviors [7]. This inherent stochasticity suggests models become increasingly unpredictable as task complexity increases, regardless of the size of their training dataset. The authors suggest that these failures seldom represent the consistent pursuit of a misaligned goal, but are instead erratic deviations that make the model's trajectory difficult to anticipate or correct.

B. The Illusion of Thinking and Complexity Limits

The danger of this stochasticity is exacerbated by surface-level fluency in problem-solving that masks fundamental limitations in handling underlying problem complexity [8]. While many models can perform impressively on isolated benchmarks, their reliability drops significantly when tasked with maintaining coherence across extended operational contexts. This suggests a functional ceiling where the increased computational overhead of "thinking" actually introduces more opportunities for statistical drift. By increasing the length of the reasoning chain, the model effectively decouples its perceived accuracy from its operational consistency [12].

C. The Limits of Scaling as a Solution

If functional incoherence is an emergent byproduct of reasoning depth, then conventional scaling appears unlikely to eliminate it. The scaling laws that successfully predicted improvements in linguistic fluency and few-shot learning do not account for the variance-based errors that arise in complex, goal-oriented sequences. Because these failures stem from the inherent randomness of the model's internal state during long-chain reasoning, increasing the parameter count or the volume of synthetic training data may actually exacerbate the problem by providing a broader surface area for incoherent behavior to manifest. This mismatch between scale and reliability underscores the need for structural interventions to ground model behavior in high-stakes environments [13]. The inability of pure scaling to resolve these stochastic failures necessitates a re-evaluation of how we define and measure a system's safety and worthiness for deployment.

V. Re-evaluating AI Trustworthiness

As generative models transition from experimental prototypes to foundational components of critical infrastructure, the metrics for evaluating their utility must shift from strictly relying on performance benchmarks to a broader framework of trustworthiness [9], [14]. The technical failures of model collapse and functional incoherence highlight the reality that a model's statistical accuracy or fluency is not synonymous with its reliability in a deployed environment. Trust in AI should be understood as a relational and context-dependent dynamic between the system and its human users rather than a static property of the software.

A. Relational Trust and the Trustor's Perspective

This shift toward relational trust elevates the trustor's perspective as a necessary dimension of evaluation, acknowledging that there are no universal attributes that make a model inherently worthy of trust [9]. Instead, trustworthiness is established when the system's behavior aligns with the user's specific context and risk tolerance. This alignment is particularly critical as AI is integrated into sensitive sectors like healthcare, military operations, and weather tracking, where the cost of incoherent failure could be catastrophic [15],

TABLE I
COMPARISON OF COMMON LLM ARCHITECTURES VS. PROPOSED AGENTIC ARCHITECTURES

Feature	Common LLM Architecture	Proposed Agentic Architecture
Core Paradigm	Scaling Laws (Parameters/Data)	Structural Anchors (Non-ML Constraints)
Data Strategy	Synthetic-heavy (Vulnerable to MAD)	Provenance-focused (Human-curated)
Reasoning Path	Stochastic Next-token Prediction	State-based Persistence & Logic Loops
Failure Mode	Functional Incoherence and Stochastic Failures	Procedural & Logical Consistency
Grounding	Internal Latent Distributions	External Non-ML Verifiability
Trust Model	Static / Benchmark-centric	Relational / Context-dependent

[16]. Because users in these domains operate under high-consequence risks, they require a guarantee of reliability rather than just plausible outputs.

B. The Role of Context and External Verifiability

To provide this guarantee, trustworthiness must be achieved in external verifiability rather than internal generative capacity. Traditional evaluations of foundation models often fail to capture the socio-technical risks identified by international regulatory and advisory bodies because they prioritize fluency over grounding [17]. This necessitates a move away from the assumption that more computing resources or data will naturally lead to more trustworthy models. Developers must focus on how AI integrates into a user's specific decision context, ensuring that the system provides the grounding required for decision-making in high-consequence environments. By conceptualizing trust as a response to risk, the mitigation of incoherence necessitates architectures that prioritize system transparency and grounding over generative volume. If trust is a response to risk and context, then it follows that architectural choices must move toward structured, verifiable frameworks.

VI. MITIGATION STRATEGIES

The compounding nature of recursive decay and the stochastic volatility of functional incoherence necessitate a multi-layered defense strategy. This approach emphasizes data integrity as the foundation and agentic grounding as an operational safeguard to ensure algorithmic fairness and reliability.

A. Data Provenance and Integrity

The preservation of high-quality, human-curated datasets is the most fundamental requirement for breaking the autophagous feedback loops described in Section III. Practitioners must implement rigorous curation pipelines that treat data provenance as a non-negotiable requirement. Without explicit source metadata to distinguish human from synthetic content, models cannot appropriately weight inputs leading to undetected contamination and the amplification of hallucinations [18]. While Retrieval-Augmented Generation (RAG) is often cited as a grounding mechanism, evidence suggests that these systems can inadvertently amplify unsafe behaviors or introduce new vulnerabilities if the retrieved context is not effectively managed [19]. As a result, data integrity must be viewed as the prerequisite for any architectural intervention.

B. Agentic Anchors and Non-ML Constraints

To mitigate functional incoherence, system design must move beyond the sole use of probabilistic models for largely deterministic tasks [8]. While data provenance addresses the training-time collapse, agentic architectures address the inference-time incoherence described in Section IV.

- **State-Based Persistence:** By maintaining a deterministic record of task states and past actions, agentic frameworks can reduce the variance-based errors found in long-chain reasoning [7].
- **Consensus Judging and Parallel Loops:** Cross-verification through multiple model iterations or specialized agents in parallel research loops allows for an algorithmic implementation of checks and balances which can help catch erratic deviations before they propagate through a sequence of actions [8].
- **Hybrid Architectures:** Combining transformer-based reasoning with symbolic logic ensures that models remain anchored to ground truth, even when internal statistical drift occurs [20].

C. Algorithmic Reparation and Robustness

Retraining alone is insufficient to counteract the loss of distribution tails and the resulting representative disparity. Algorithmic Reparation (AR) offers a framework for selectively curating training batches to mitigate the amplification of societal inequities within the data ecosystem [6]. By embedding fairness-aware sampling and dynamic reweighting into the training pipeline, practitioners can disrupt the feedback loops that otherwise entrench existing biases [17]. Ensemble learning techniques may also provide a path to improving robustness to distributional shifts by sampling from

TABLE II
SYNTHESIS OF FOUNDATIONAL RISKS IN RECURSIVE AND REASONING-INTENSIVE LLMS

Research Vector	Research Cited	Key Contributions
Statistical Decay	[4], [5]	Establish that training on synthetic data is self-consuming. Identifies the irreversible loss of distribution tails and the admissible threshold of synthetic data required before total model collapse.
Societal Bias	[6]	Proves that systematic distribution shifts create feedback loops that amplify biases and marginalize at-risk groups.
Operational Risk	[7], [9]	Focus on failures occurring at inference-time. Identifies functional incoherence where task complexity leads to high variance, and argues that trustworthiness must be re-conceptualized as a relational and context-dependent dynamic.

diverse regions of the loss landscape, preventing the model from converging on a singular, overfitted point estimate typical of model collapse.

D. Intelligent Scaling for Sustainability

Prior work exploring recursive training indicates that increasing model size without corresponding improvements in data diversity leads to diminishing returns and potential collapse [5]. Long-term sustainability requires modular training and phased fine-tuning that prioritize output consistency over parameter count. By aligning model development with the relational requirements of trustworthy AI, these strategies can help ensure that model growth remains tethered to the operational reliability expected of these systems [16].

VII. COMPARATIVE ANALYSIS OF PRIMARY WORKS

This section evaluates the specific contributions, strengths, and limitations of the primary research surveyed as part of this review.

A. Shumailov et al. (2023) and Alemohammad et al. (2023)

The primary strength of these works is their rigorous characterization of the curse of recursion and model autophagy disorder [4], [5]. By identifying the existence of a threshold for the inclusion of synthetic data in model training, they provide a predictive framework for when a model will begin to collapse. A significant limitation is that their experiments were largely conducted on earlier generations of generative models. It's unclear whether current frontier models might possess a higher resilience to distributional shifts that could delay the onset of collapse.

B. Wyllie et al. (2024)

The primary contribution of this research is the formalization of "Fairness Feedback Loops" which connect the mathematical reality of model collapse to societal outcomes [6]. A major strength of this work is its demonstration that even if model performance appears stable on average, the loss of distribution tails leads to a disproportionate loss of accuracy for marginalized groups. This provides an important ethical dimension to the technical study of distribution shifts. A limitation of the paper is its heavy reliance on sampling bias as a primary mitigation strategy. The authors identify the trade-off inherent in leveraging sampling bias to preserve output quality, but don't fully explore its consequences. While Shumailov et al. and Alemohammad et al. established the mathematical inevitability of collapse, Wyllie et al. address a critical limitation in their work by shifting the focus from aggregate performance to distribution-specific degradation. In doing so, they prove that the average stability reported in earlier studies masks the erasure of minority data tails.

C. Hägele et al. (2026)

The strength of this research is in its novel identification of functional incoherence as a failure mode that is distinct from statistical training decay [7]. They provide a necessary critique of current scaling laws by quantifying how failures shift from bias to high variance during multi-step reasoning. A weakness of this work is its lack of prescriptive solutions. While it diagnoses the causes of inference-time failures, it does not provide a standardized metric for measuring task complexity, making it difficult for practitioners to predict when a model will become incoherent. This work represents a shift from the earlier literature that focused on training-time degradation by addressing the inference-time instability of frontier models. This is a growing limitation in the field where models may be trained on pristine data yet still exhibit signs of collapse during the execution of complex tasks.

D. Wirz et al. (2025)

The authors contribute a socio-technical perspective to this review by re-conceptualizing trustworthiness as a relational dynamic rather than a static model attribute, which is essential for deploying AI in high-consequence environments [9]. However, the paper is primarily theoretical, and lacks a concrete technical implementation for how relational trust can be measured through automated benchmarks or non-ML constraints. This paper serves as a theoretical counterpoint to the

purely algorithmic focus of Shumailov et al. and Hägele et al. They argue that if the technical failures identified in those works can't be fully eliminated, the field must shift toward the relational trust framework that they propose. This also addresses the 'optimization wall' hit by purely technical scaling.

VIII. Takeaways and Synthesis

The results of this review recommends a transition from purely probabilistic scaling toward structural grounding. The technical commonality across the core research analyzed here is that increasing compute and data volume now act as catalysts for systemic instability when those inputs are synthetic or when reasoning chains are extended. One unifying theme is the reliability gap between a model's linguistic fluency and its operational consistency. The specific question this survey sought to answer was whether scaling laws could resolve these emerging defects. The answer provided by the literature is a decisive 'no'; increased scale without architectural constraints like the agentic anchors proposed in this review likely broadens the surface area for both model collapse and functional incoherence.

IX. Open Questions & Future Work Directions

Following this review, several questions remain that are critical for the deployment and reliability of trustworthy AI systems:

- While Hägele et al. identify variance as the driver of functional incoherence, there is no standardized metric for task complexity. What mathematical threshold determines when a task's reasoning depth will trigger stochastic failure?
- If human-generated data is finite, is it possible to identify a synthetic ratio that allows for continued training without triggering model collapse? If not, is any amount of synthetic data eventually corrosive?
- While this paper proposes agentic architectures as a mitigation strategy, do these non-ML constraints introduce their own biases or performance bottlenecks that limit the intelligence they are meant to ground?

X. Conclusion

These findings demonstrate that as the AI ecosystem faces a terminal scarcity of human-generated data, the shift toward recursive training introduces systemic instabilities that threaten the sustainability of future models [3], [18]. To better understand the interaction between these risks and the necessity for structural anchors, Table II categorizes the risks identified in this research. The statistical decay identified by Alemohammad et al. and Shumailov et al. creates a self-consuming environment where the loss of distribution tails becomes a catalyst for the feedback loops identified by Wyllie et al [4]–[6]. This narrowing of the model's worldview ensures that biases are not only retained but structurally amplified, effectively erasing the representation of at-risk groups. Beyond these training-time defects, the research reveals an equally critical operational failure. As task complexity increases, Hägele et al. demonstrate that model failures can manifest as functional incoherence as operational depth increases [7]. This mismatch between perceived intelligence and execution emphasizes the necessity of trust frameworks like those proposed by Wirz et al., shifting the burden of proof from internal model attributes to external, context-dependent verifiability [9].

The convergence between these foundational defects and operational instabilities suggests that probabilistic scaling has reached a point of diminishing returns for system reliability. This paper proposes that while the statistical decay of self-consuming loops may be an inherent risk of training on synthetic data, the resulting functional incoherence can be mitigated through the structural implementation of agentic architectures. By utilizing non-machine learning constraints such as state-based persistence to reduce reasoning variance, parallel research loops to facilitate cross-verification, and consensus-oriented judging to maintain procedural logic, agentic frameworks can provide the necessary anchors to ground models in verifiable truth. Ensuring the long-term sustainability and trustworthiness of generative AI requires moving beyond the limitations of purely probabilistic scaling toward robust, agentic systems designed to satisfy the rigorous relational requirements of human trust.

References

[1] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling Laws for Neural Language Models," Jan. 23, 2020. https://doi.org/10.48550/arXiv.2001.08361

[2] T. B. Brown et al., "Language Models are Few-Shot Learners." May 28, 2020. https://doi.org/10.48550/arXiv.2005.14165

[3] P. Villalobos, A. Ho, J. Sevilla, T. Besiroglu, L. Heim, and M. Hobbhahn, "Will we run out of data? Limits of LLM scaling based on human-generated data," Oct. 26, 2022. https://arxiv.org/abs/2211.04325

[4] S. Alemohammad, J. Casco-Rodriguez, L. Luzi, A. I. Humayun, H. Babaei, D. LeJeune, A. Siahkoohi, and R. G. Baraniuk, "Self-Consuming Generative Models Go MAD," arXiv, July 4, 2023. https://arxiv.org/pdf/2307.01850

[5] I. Shumailov, Z. Shumaylov, Y. Zhao, Y. Gal, N. Papernot, and R. Anderson, "The curse of recursion: training on generated data makes models forget," arXiv (Cornell University), May 27, 2023. https://doi.org/10.48550/arXiv.2305.17493

[6] M. Wyllie, I. Shumailov, N. Papernot, "Fairness Feedback Loops: Training on Synthetic Data Amplifies Bias," arXiv, March 12, 2024. https://doi.org/10.48550/arXiv.2403.07857

[7] A. Hägele, A. P. Gema, H. Sleight, E. Perez, J. Sohl-Dickstein. ”The hot mess of AI: How does misalignment scale with model intelligence and task complexity?” Jan. 30, 2026. https://arxiv.org/abs/2601.23045

[8] I. Shumailov et al., “The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity,” June 7, 2025. https://arxiv.org/abs/2506.06941

[9] C. D. Wirz et al., “(Re)Conceptualizing trustworthy AI: A foundation for change,” Artificial Intelligence, vol. 330, p. 104309, 2025. https://doi.org/10.1016/j.artint.2025.104309

[10] W. Zhang, Z. Sheng, Z. Yin, Y. Jiang, Y. Xia, J. Gao, Z. Yang, and B. Cui, “Model degradation hinders deep graph neural networks,” Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 2493–2503, June 9, 2022. https://doi.org/10.48550/arXiv.2206.04361

[11] S. Fort, H. Hu, and B. Lakshminarayanan, “Deep Ensembles: A Loss Landscape Perspective,” arXiv, June 25, 2020. https://doi.org/10.48550/arXiv.1912.02757

[12] Y. Wang, J. Haddow, A. Birch, and H. Peng, ”Assessing the Reliability of Large Language Model Knowledge,” arXiv, October 15, 2023. https://doi.org/10.48550/arXiv.2310.09820

[13] E. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, “On the Dangers of Stochastic Parrots,” Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 610–623, March 1, 2021. https://doi.org/10.1145/3442188.3445922

[14] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P. Yu, Q. Yang, and X. Xie. 2018. A Survey on Evaluation of Large Language Models. J. ACM 37, 4, Article 111 (August 2018), 45 pages. https://doi.org/10.48550/arXiv.2307.03109

[15] M. Brundage et al., “The Malicious Use of Artificial intelligence: Forecasting, Prevention, and Mitigation,” CoRR, Jan. 2018, doi: 10.48550/arxiv.1802.07228.

[16] R. Bommasani et al., “On the Opportunities and Risks of Foundation Models,” arXiv, August 16, 2021. https://doi.org/10.48550/arXiv.2108.07258

[17] N. T. Lee, P. Resnick, and G. Barton, “Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms,” Brookings Institution, May 22, 2019. https://www.brookings.edu/articles/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms

[18] M. Harrison Dupre, “When AI Is Trained on AI-Generated Data, Strange Things Start to Happen,” Futurism, August 2, 2024. https://futurism.com/ai-trained-ai-generated-data-interview

[19] B. An, S. Zhang, and M. Dredze, “RAG LLMs are Not Safer: A Safety Analysis of Retrieval-Augmented Generation for Large Language Models,” arXiv.org, Apr. 25, 2025. https://arxiv.org/abs/2504.18041

[20] B. Cao, H. Lin, X. Han, L. Sun, L. Yan, M. Liao, T. Xue, J. Xu. 2021. Knowledgeable or Educated Guess? Revisiting Language Models as Knowledge Bases. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). https://doi.org/10.48550/arXiv.2307.03109