Filtering the Data Lake: Why Siloed Human Oversight Is Insufficient to Prevent Data Degradation in Recursive Generative AI Training - A Review

Syaamantak Das

Centre for Educational Technology, Indian Institute of Technology Bombay, Mumbai, India

Abstract

As Generative Artificial Intelligence systems increasingly train on data that is partially produced by earlier models, concerns are growing about cumulative data quality degradation, and the amplification of hallucinations. While human oversight mechanisms, such as dataset curation, auditing, and reinforcement learning from human feedback, are widely adopted, they are often implemented in organizational, technical, or regulatory silos. In this paper, we present a structured review of research (N=31) published over the last three years (2023-25) on Large Language Models, Human–AI interaction, and data-centric oversight practices. Synthesizing empirical findings, system descriptions, and policy-oriented proposals, we argue that siloed forms of human oversight are structurally insufficient to prevent degradation in recursive Generative AI training pipelines. Our analysis highlights three recurring limitations: (i) weak data provenance and lineage tracking, (ii) limited scalability of human-in-the-loop filtration, and (iii) a lack of integrative governance across model development stages. Based on this synthesis, our findings reframes data fidelity (through a Data Shapley model) as a core locus of human oversight, and directions for a data-centric research agenda for future work.

Keywords

Generative AI, Large Language Models, data provenance, data degradation, human oversight, recursive training, hallucination, dataset governance, Human-in-the-loop

1. Introduction

Generative Artificial Intelligence systems, particularly large language models (LLMs), increasingly rely on vast and heterogeneous data lakes[1] for training and continual refinement. As these systems become embedded in everyday information practices, a growing portion of the data they consume is no longer exclusively human-generated[2], but instead partially produced by earlier generations of generative models[3]. This recursive training dynamic has raised concerns about cumulative data degradation[4], including the amplification of hallucinations[5], loss of factual grounding etc. Unlike isolated model failures, such degradation emerges over time through feedback loops between model outputs, data collection pipelines, and retraining processes.

Human oversight is commonly proposed as a safeguard against these risks. Across research and practice, oversight mechanisms include dataset curation, human-in-the-loop filtering, reinforcement learning from human feedback, post-hoc auditing, and emerging approaches to data provenance and watermarking[6]. Collectively, these practices aim to ensure that training data remains reliable, representative, and aligned with human values. However, despite their widespread adoption, evidence from recent literature [7] suggests that these mechanisms have not kept pace with the scale, velocity, and recursive nature of contemporary generative AI systems.

In this paper, we argue that a central reason for this shortfall lies in the siloed nature of existing human oversight. Oversight practices are typically confined to specific stages of the AI lifecycle, such as dataset preparation, model alignment, or deployment-time evaluation, without systematic integration across stages. Technical tools for data filtration operate largely independently of organizational decision-making, while governance and regulatory efforts often remain disconnected from operational training

Joint Proceedings of the ACM Intelligent User Interfaces (IUI) Workshops 2026, March 23-26, 2026, Paphos, Cyprus
syaamantak.das@iitb.ac.in (S. Das)
0000-0001-9896-3312 (S. Das)

pipelines. As a result, oversight signals generated in one context rarely propagate effectively to others, limiting their capacity to prevent long-term data degradation under recursive training.
We present a structured review of research published (N=31) over the last three years on Generative AI, Human–AI interaction, and data-centric oversight practices. Synthesizing empirical findings, process-level analyses, and policy-oriented proposals, we examine how data degradation arises in recursive training settings and how current forms of human oversight attempt, yet ultimately struggle to address it. Rather than attributing these limitations to isolated technical failures, we emphasize structural misalignments between the scale of generative data production and the fragmented design of oversight mechanisms. Based on this synthesis, we consolidate evidence demonstrating that data degradation in generative AI is a systemic phenomenon exacerbated by recursive use of model-generated content. Second, we reframe data fidelity as an approach for supporting human oversight around data-centric integration, calling for tighter quality of data and governance across the AI lifecycle.

2. Problem Background

2.1. Generative AI and Recursive Training

We use the term *Generative AI* [8] to refer to machine learning systems (e.g. ChatGPT[1], Google Gemini[2]) capable of producing novel content such as text, code, or images, based on patterns learned from large-scale datasets. In this paper, our focus is primarily on LLMs, which have become a dominant paradigm for generative AI research and deployment. Recursive training describes a process in which successive model generations are trained using data that includes outputs from earlier generative models[9], whether through deliberate synthetic data generation or inadvertent reuse of model-generated web content. Recent research [10] has raised concerns that repeated exposure to synthetic content can alter data distributions in ways that degrade model reliability over time.

2.2. Data Degradation in Generative AI Systems

We use the term *data degradation* to denote a broad class of phenomena in which the informational quality of training data deteriorates across iterations[11]. In generative AI contexts, degradation manifests as increased hallucination rates[12], diminished representation of rare knowledge[13], homogenization of outputs[14], and subtle shifts in style or epistemic assumptions that accumulate across training cycles. Unlike traditional dataset bias or label noise, which are static properties, data degradation in recursive training is dynamic and emergent, shaped by feedback loops between model outputs, data collection, and retraining.

2.3. Human Oversight in the AI Lifecycle

Human oversight encompasses the range of practices through which human actors monitor, intervene in, or guide AI system development and deployment. In the Generative AI literature, oversight appears in multiple forms, including manual dataset curation, annotation and filtering, reinforcement learning from human feedback [15], red-teaming and auditing, and governance mechanisms such as documentation, standards, and regulatory compliance. These practices are distributed across different roles and organizational units: dataset curation by data engineers or contractors, alignment by specialized research teams, and governance by legal or policy groups. While each form of oversight may be locally effective, their separation across technical and institutional boundaries can limit their collective impact on long-term system behavior.

[1] https://chatgpt.com/
[2] https://gemini.google.com/app

3. Methodology for Literature Review

3.1. Siloed Oversight as an Analytical Lens

In this paper, we use siloed human oversight as an analytical lens rather than a critique of specific individuals or practices. By "*siloed*", we refer to oversight mechanisms that are temporally, organizationally, or technically isolated from one another. Examples include filtering datasets without visibility into downstream training effects, applying human feedback without access to data provenance information, or establishing governance requirements that lack operational enforcement within training pipelines. This lens allows us to examine why oversight mechanisms that are effective in isolation may nevertheless fail to address systemic risks such as data degradation under recursive training. It also provides a foundation for considering how oversight might be restructured to better account for feedback loops across the AI lifecycle.

3.2. Review Scope and Time Frame

This review focuses on 31 preprint / peer-reviewed research literature published between December 2022 and December 2025 that examines Generative AI systems, particularly LLMs, in relation to training data, human oversight, and reliability. We include technical studies, human–computer interaction research, and policy-oriented proposals where they directly engage with data-centric oversight issues. We do not attempt to provide an exhaustive survey of all work on model evaluation, bias, or AI ethics. Instead, we concentrate on literature that sheds light on how data practices and human oversight interact in recursive generative AI settings. This paper adopts a structured narrative review approach [16] to synthesize recent research on data degradation and human oversight in recursive generative AI training. The review is designed to balance breadth across technical, human-centered, and governance-oriented literatures with sufficient depth to support analytic claims and a forward-looking research agenda.

3.3. Data Sources and Search Strategy

We conducted literature searches across a combination of academic databases and preprint repositories, including Google Scholar, arXiv, the ACL Anthology, the ACM Digital Library, and IEEE Xplore. We systematically examined proceedings from major conferences spanning machine learning (NeurIPS, ICML, ICLR), natural language processing (ACL, EMNLP, NAACL), human-computer interaction (CHI, IUI), and fairness and accountability (FAccT). Policy and governance-oriented work was additionally sourced from interdisciplinary venues and white papers where relevant. Search queries combined terms related to generative AI, training data, and human oversight, specifically using combinations of: 'generative AI', 'large language model', 'LLM', 'synthetic data', 'recursive training', 'model collapse', 'data degradation', 'human oversight', 'RLHF', 'data curation', and 'data provenance'.

3.3.1. Inclusion and Exclusion Criteria

Papers were included if they met at least one of the following criteria: (1) Empirically evaluated generative AI systems with respect to data quality, hallucination, or reliability, (2) Proposed or assessed methods for dataset curation, filtering, provenance tracking, or data valuation, (3) Examined human oversight practices in the development or deployment of generative AI systems, and (4) Advanced policy, governance, or normative frameworks directly related to data-centric oversight. We tried to exclude those research works that focused solely on model architecture or optimization without substantive discussion of data practices or oversight, as well as studies addressing AI ethics or bias at a high level without operational implications for training data.

3.4. Coding and Synthesis Procedure

Thirty-one shortlisted papers (shown in Appendix A) were analyzed using a hybrid deductive–inductive coding approach [17]. Deductive codes were derived from the paper's guiding concerns, including data

Table 1
Thematic synthesis of findings from the literature on data degradation and human oversight in recursive generative AI training

Theme	Key Findings	Evidence Type	Implications for Human Oversight
A. Data degradation under recursion	Recursive exposure to model-generated content is associated with increased hallucination, factual drift, or loss of distributional diversity over successive training cycles.	Empirical evaluations, benchmark analyses, simulation studies	Oversight focused on single training iterations cannot detect or prevent cumulative degradation.
B. Synthetic data reuse	Synthetic data can improve performance in low-resource settings but risks reinforcing errors and biases when reused at scale without provenance controls.	Controlled experiments, ablation studies	Human oversight must distinguish beneficial from harmful synthetic data uses.
C. Existing human oversight mechanisms	Oversight practices (e.g., dataset curation, RLHF, auditing) are typically applied at isolated lifecycle stages and evaluated locally rather than systemically.	System descriptions, practitioner reports, HCI studies	Fragmented oversight limits visibility into downstream effects on training data quality.
D. Structural limits of siloed oversight	Organizational, technical, and governance silos prevent oversight signals from propagating across data pipelines, training, and deployment.	Conceptual analyses, policy papers, case studies	Effective oversight requires integration across roles, tools, and governance layers.

degradation mechanisms, oversight modalities, provenance and lineage, and scalability constraints. Inductive codes were added to capture emergent themes such as feedback-loop effects, organizational fragmentation, and trade-offs between scale and data fidelity. Rather than aggregating quantitative effect sizes, we synthesized evidence qualitatively, focusing on reported mechanisms, empirical patterns, and methodological limitations. The results of this synthesis are presented in Section 4.

3.4.1. Limitations

Our analysis emphasizes structural relationships between data practices and human oversight rather than performance comparisons between individual models. In particular, we examine how oversight mechanisms are positioned relative to the AI lifecycle and how this positioning affects their ability to mitigate data degradation under recursive training. As a narrative review conducted by a single author, this work is subject to limitations including potential selection bias and the absence of formal inter-coder reliability measures. To mitigate these risks, we prioritized transparency in search strategies and aimed to include diverse perspectives across technical, human-centered, and governance literatures.

4. Findings and Inferences

The literature reviewed reveals recurring patterns linking recursive generative AI training, data degradation, and the structural limitations of existing human oversight mechanisms. To synthesize findings across technical, human-centered, and governance-oriented studies, we organize results into four thematic categories. These themes capture both empirical observations and analytic claims reported in recent research. Table 1 summarizes key findings from the literature, thematically coded as A–D. For each theme, the table identifies the dominant empirical or conceptual claim, the types of evidence reported, and the implications for human oversight in recursive generative AI training pipelines.

4.1. Inference and Discussion: Research Questions for Data-Centric Human Oversight

The synthesis presented in Table 1 indicate that data degradation in recursive generative AI training arises from a mismatch between the systemic nature of data feedback loops and the localized design of current human oversight mechanisms. Addressing this mismatch requires reframing human oversight around data as a first-class object of intervention, one that can be inspected, evaluated, and governed across the AI lifecycle. We present a set of open research questions organized around these themes A–D, and we highlight data fidelity valuation, particularly through Data Shapley–style approaches, as a promising integrative direction to answer these questions for future work.

4.1.1. A: Measuring Data Contribution Under Recursive Training

Research Question: How can the contribution of individual data points or data subsets to model behavior be measured under recursive generative AI training? Shumailov et al.[9] and Guo et al.[3] demonstrate through empirical studies that quality loss and diversity decline compound across training generations, revealing that single-iteration evaluation fails to detect cumulative degradation.

4.1.2. B: Differentiating Beneficial and Harmful Synthetic Data

Research Question: Under what conditions does synthetic or model-generated data improve model performance, and when does it accelerate degradation? Theme B highlights the dual role of synthetic data as both a performance enhancer and a risk factor. A key open question is whether data valuation methods can be used to distinguish beneficial synthetic data from data that propagates errors or biases.

4.1.3. C: Integrating Oversight Signals Across the AI Lifecycle

Research Question: How can human oversight signals generated at different lifecycle stages be aggregated into coherent data-centric decisions? Theme C reveals that oversight mechanisms are often evaluated in isolation, with limited visibility into their downstream effects. Research is needed on frameworks that integrate diverse oversight signals, such as human feedback, audit outcomes, and provenance metadata, into unified representations of data utility and risk. Data valuation methods offer one potential abstraction layer through which heterogeneous oversight inputs can be reconciled and operationalized in training pipelines.

4.1.4. D: Governance and Accountability Through Data Valuation

Research Question: Can data-centric valuation support more accountable and auditable forms of human oversight in generative AI systems? Theme D emphasizes that organizational and regulatory silos weaken oversight by separating responsibility for data quality from authority over training decisions.

4.2. Future Work: A possible Data Shapley-style model as a Unifying Direction

Across these research questions, Data Shapley style valuation emerges as a promising, though still exploratory[18], direction for integrating human oversight into data-centric decision-making. Data Shapley methods, from cooperative game theory, quantify each training example's marginal contribution to model performance by measuring performance differences when that example is included versus excluded. For recursive AI oversight, this offers three advantages: it creates quantitative trails linking training data to model behaviors (addressing weak provenance), provides a common metric for integrating diverse oversight signals across organizational boundaries (addressing fragmentation), and enables principled decisions about retaining or excluding synthetic data (addressing Theme B). While exact Shapley computation remains computationally prohibitive at LLM scale, approximation methods offer paths forward, and the conceptual framework provides direction for lifecycle-spanning oversight that our analysis indicates is necessary.

5. Conclusion

This review demonstrates that data degradation in recursive generative AI training is a systemic phenomenon that siloed human oversight cannot adequately address. Our synthesis of 31 recent papers reveals fragmentation across organizational, technical, and governance boundaries that prevents effective quality control across the AI lifecycle. Data-centric valuation approaches offer a promising direction for creating the integrated, lifecycle-spanning oversight mechanisms necessary for maintaining data fidelity in recursive training contexts.

Declaration on Generative AI

During the preparation of this work, the author(s) used Grammarly in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] R. Hai, C. Koutras, C. Quix, M. Jarke, Data lakes: A survey of functions and systems, IEEE Transactions on Knowledge and Data Engineering 35 (2023) 12571–12590.

[2] A. Bauer, S. Trapp, M. Stenger, R. Leppich, S. Kounev, M. Leznik, K. Chard, I. Foster, Comprehensive exploration of synthetic data generation: A survey, arXiv preprint arXiv:2401.02524 (2024).

[3] X. Guo, Y. Chen, Generative ai for synthetic data generation: Methods, challenges and the future, arXiv preprint arXiv:2403.04190 (2024).

[4] M. E. A. Seddik, S.-W. Chen, S. Hayou, P. Youssef, M. Debbah, How bad is training on synthetic data? a statistical analysis of language model collapse, arXiv preprint arXiv:2404.05090 (2024).

[5] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, et al., A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, ACM Transactions on Information Systems 43 (2025) 1–55.

[6] S. Sterz, K. Baum, S. Biewer, H. Hermanns, A. Lauber-Rönsberg, P. Meinel, M. Langer, On the quest for effectiveness in human oversight: Interdisciplinary perspectives, in: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 2495–2507. URL: https://doi.org/10.1145/3630106.3659051. doi:10.1145/3630106.3659051.

[7] A. Holzinger, K. Zatloukal, H. Müller, Is human oversight to ai systems still possible?, 2024.

[8] S. S. Sengar, A. B. Hasan, S. Kumar, F. Carroll, Generative artificial intelligence: a systematic review and applications, Multimedia Tools and Applications 84 (2025) 23661–23700.

[9] I. Shumailov, Z. Shumaylov, Y. Zhao, N. Papernot, R. Anderson, Y. Gal, Ai models collapse when trained on recursively generated data, Nature 631 (2024) 755–759.

[10] S. Z. Shabgahi, P. Aghazadeh, A. Mirhoseini, F. Koushanfar, Fortifai: Fending off recursive training induced failure for ai model collapse, arXiv preprint arXiv:2509.08972 (2025).

[11] Q. Bertrand, A. J. Bose, A. Duplessis, M. Jiralerspong, G. Gidel, On the stability of iterative retraining of generative models on their own data, arXiv preprint arXiv:2310.00429 (2023).

[12] A. Jesson, N. Beltran-Velez, Q. Chu, S. Karlekar, J. Kossen, Y. Gal, J. P. Cunningham, D. Blei, Estimating the hallucination rate of generative ai, Advances in Neural Information Processing Systems 37 (2024) 31154–31201.

[13] A. J. Peterson, Ai and the problem of knowledge collapse, AI & SOCIETY (2025) 1–21.

[14] H. A. Vu, G. Reeves, E. Wenger, What happens when generative ai models train recursively on each others' generated outputs?, arXiv preprint arXiv:2505.21677 (2025).

[15] Y. Cao, Q. Z. Sheng, J. McAuley, L. Yao, Reinforcement learning for generative ai: A survey, arXiv preprint arXiv:2308.14328 (2023).

[16] D. Turnbull, R. Chugh, J. Luck, Systematic-narrative hybrid literature review: A strategy for integrating a concise methodology into a manuscript, Social Sciences & Humanities Open 7 (2023) 100381.

[17] J. Fereday, E. Muir-Cochrane, Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development, International journal of qualitative methods 5 (2006) 80–92.

[18] J. T. Wang, P. Mittal, D. Song, R. Jia, Data shapley in one training run, arXiv preprint arXiv:2406.11011 (2024).

A. List of papers reviewed

1. Shumailov, I., Shumaylov, Z., Zhao, Y. et al. AI models collapse when trained on recursively generated data. Nature 631, 755–759 (2024). https://doi.org/10.1038/s41586-024-07566-y
2. Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., Anderson, R. (2023). The curse of recursion: Training on generated data makes models forget. arXiv preprint arXiv:2305.17493.
3. Guo, Y., Shang, G., Vazirgiannis, M., Clavel, C. (2024, June). The curious decline of linguistic diversity: Training language models on synthetic text. In Findings of the Association for Computational Linguistics: NAACL 2024 (pp. 3589-3604).
4. Gerstgrasser, M., Schaeffer, R., Dey, A., Rafailov, R., Sleight, H., Hughes, J., ... Koyejo, S. (2024). Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data. arXiv preprint arXiv:2404.01413.
5. Kazdan, J., Schaeffer, R., Dey, A., Gerstgrasser, M., Rafailov, R., Donoho, D. L., Koyejo, S. (2024). Collapse or thrive? perils and promises of synthetic data in a self-generating world. arXiv preprint arXiv:2410.16713.
6. Xing, X., Shi, F., Huang, J. et al. On the caveats of AI autophagy. Nat Mach Intell 7, 172–180 (2025). https://doi.org/10.1038/s42256-025-00984-1
7. Dohmatob, E., Feng, Y., Yang, P., Charton, F., Kempe, J. (2024). A tale of tails: Model collapse as a change of scaling laws. arXiv preprint arXiv:2402.07043.
8. Briesch, M., Sobania, D., Rothlauf, F. (2023). Large language models suffer from their own output: An analysis of the self-consuming training loop.
9. Zhu, X., Cheng, D., Li, H., Zhang, K., Hua, E., Lv, X., ... Zhou, B. (2024). How to Synthesize Text Data without Model Collapse?. arXiv preprint arXiv:2412.14689.
10. Amin, K., Babakniya, S., Bie, A., Kong, W., Syed, U., Vassilvitskii, S. (2025). Escaping collapse: The strength of weak data for large language model training. arXiv preprint arXiv:2502.08924.
11. Jie, Y. W., Ferdinan, T., Kazienko, P., Satapathy, R., Cambria, E. (2024, November). Self-training large language models through knowledge detection. In Findings of the Association for Computational Linguistics: EMNLP 2024 (pp. 15033-15045).
12. Dohmatob, E., Feng, Y., Subramonian, A., Kempe, J. (2024). Strong model collapse. arXiv preprint arXiv:2410.04840.
13. Yang, S., Ali, M. A., Yu, L., Hu, L., Wang, D. (2024). MONAL: Model autophagy analysis for modeling human-AI interactions. arXiv preprint arXiv:2402.11271.
14. Feng, Y., Dohmatob, E., Yang, P., Charton, F., Kempe, J. (2024). Beyond model collapse: Scaling up with synthesized data requires verification. arXiv preprint arXiv:2406.07515.
15. Wang, Z., Wu, Z., Zhang, J., Guan, X., Jain, N., Lu, S., ... Koshiyama, A. (2024). Bias Amplification: Large Language Models as Increasingly Biased Media. arXiv preprint arXiv:2410.15234.
16. Drayson, G., Yilmaz, E., Lampos, V. (2025). Machine-generated text detection prevents language model collapse. arXiv preprint arXiv:2502.15654.
17. Xu, W., Zhu, G., Zhao, X., Pan, L., Li, L., Wang, W. (2024, August). Pride and prejudice: LLM amplifies self-bias in self-refinement. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 15474-15492).
18. Zhang, J., Qiao, D., Yang, M., Wei, Q. (2024). Regurgitative training: The value of real data in training large language models. arXiv preprint arXiv:2407.12835.
19. Wang, T., Horiguchi, A., Pang, L., Priebe, C. E. (2025). LLM web dynamics: Tracing model collapse in a network of LLMs. arXiv preprint arXiv:2506.15690.
20. Das, D., De Langis, K., Martin-Boyle, A., Kim, J., Lee, M., Kim, Z. M., ... Kang, D. (2024). Under the surface: Tracking the artifactuality of llm-generated data. arXiv preprint arXiv:2401.14698.
21. Martínez, G., Watson, L., Reviriego, P., Hernández, J. A., Juarez, M., Sarkar, R. (2023, August). Towards understanding the interplay of generative artificial intelligence and the internet. In International Workshop on Epistemic Uncertainty in Artificial Intelligence (pp. 59-73). Cham: Springer Nature Switzerland.

22. Long, L., Wang, R., Xiao, R., Zhao, J., Ding, X., Chen, G., Wang, H. (2024). On llms-driven synthetic data generation, curation, and evaluation: A survey. arXiv preprint arXiv:2406.15126.
23. M. Marchi, S. Soatto, P. Chaudhari and P. Tabuada, "Heat Death of Generative Models in Closed-Loop Learning," 2024 IEEE 63rd Conference on Decision and Control (CDC), Milan, Italy, 2024, pp. 1524-1530, doi: 10.1109/CDC56724.2024.10886816.
24. Havrilla, A., Dai, A., O'Mahony, L., Oostermeijer, K., Zisler, V., Albalak, A., ... Meyerson, E. (2024). Surveying the effects of quality, diversity, and complexity in synthetic data from large language models. arXiv preprint arXiv:2412.02980.
25. Martínez, G., Watson, L., Reviriego, P., Hernández, J. A., Juarez, M., Sarkar, R. (2023). Combining generative artificial intelligence (AI) and the Internet: Heading towards evolution or degradation?. arXiv preprint arXiv:2303.01255.
26. Wang, K., Zhu, J., Ren, M., Liu, Z., Li, S., Zhang, Z., ... Wang, Y. (2024). A survey on data synthesis and augmentation for large language models. arXiv preprint arXiv:2410.12896.
27. Qin, Z., Dong, Q., Zhang, X., Dong, L., Huang, X., Yang, Z., ... Wei, F. (2025). Scaling laws of synthetic data for language models. arXiv preprint arXiv:2503.19551.
28. Veselovsky, V., Ribeiro, M. H., West, R. (2023). Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks. arXiv preprint arXiv:2306.07899.
29. Chen, H., Waheed, A., Li, X., Wang, Y., Wang, J., Raj, B., Abdin, M. I. (2024). On the diversity of synthetic data and its impact on training large language models. arXiv preprint arXiv:2410.15226.
30. Bertrand, Q., Bose, A. J., Duplessis, A., Jiralerspong, M., Gidel, G. (2023). On the stability of iterative retraining of generative models on their own data. arXiv preprint arXiv:2310.00429.
31. Borji, A. (2024). A Note on Shumailov et al.(2024):AI Models Collapse When Trained on Recursively Generated Data'. arXiv preprint arXiv:2410.12954.