When AI Cannot Reproduce Itself: Citation Drift as a Reproducibility Failure in Scientific LLMs

Gokul Srinath Seetha Ram

s.gokulsrinath@gmail.com

Abstract

Reproducibility is a cornerstone of scientific reliability, yet today's AI assistants themselves often fail this test.We frame citation drift as a form of temporal anomaly detection within scientific large language models, extending spatiotemporal anomaly frameworks to epistemic data and citation stability. Large Language Models (LLMs) are increasingly used for scientific writing and research assistance, yet their ability to maintain consistent citations across multi-turn conversations remains largely unexplored. This study introduces the concept of *citation drift*—the phenomenon where references mutate, disappear, or get fabricated during extended LLM interactions. Through a comprehensive analysis of 240 conversations across 4 LLaMA models using 36 authentic scientific papers from 6 domains, this work demonstrates significant citation instability. Results reveal that citation stability varies dramatically across models, with llama-4-maverick-17b showing the highest stability (0.481) and llama-4-scout-17b showing the worst fabrication rates (0.856). This study introduces novel metrics including citation drift entropy and willingness-to-cite, providing a framework for evaluating LLM citation reliability in scientific contexts. We frame citation drift as a meta-reproducibility benchmark revealing that LLMs cannot reproduce their own scientific outputs consistently.

Introduction

The integration of Large Language Models (LLMs) into scientific research workflows has accelerated rapidly, with models increasingly assisting in literature reviews, paper writing, and research synthesis (Devlin et al. 2019; Brown et al. 2020). However, a critical gap exists in our understanding of how these models handle citations—the fundamental currency of scientific communication—across extended conversations.

Recent debates on the reproducibility crisis in AI highlight the need to evaluate not only human experiments but also the reproducibility of machine-generated knowledge. This work extends that discourse by testing whether large language models can reproduce their own factual outputs—citations—under controlled, deterministic conditions.

We distinguish between three complementary layers of reproducibility in language models: (1) Output reproducibility—producing identical text under fixed decoding settings, (2) Referential reproducibility—preserving factual references and citations across turns, and (3) Epistemic reproducibility—maintaining stable reasoning chains over time. Citation drift directly measures failures in the second layer.

Citation drift represents a novel phenomenon where references undergo systematic changes during multi-turn LLM interactions. This includes citation mutation (changes in format or content), citation loss (disappearing references), and citation fabrication (invented references). Citation drift threatens the integrity of scientific communication by propagating misinformation, compromises factual reliability in generative models, and erodes user trust in AI-assisted research tools. This work directly supports WASP's goal of advancing AI for scientific publishing by quantifying reliability in reference generation. This study presents the first comprehensive analysis of citation drift across multiple LLM architectures, introducing novel metrics and providing actionable insights for the research community.

Related Work

Narrative Related Work

The reliability of LLMs in scientific communication hinges on controlling hallucinations and maintaining accurate references. Comprehensive surveys synthesize the landscape of hallucination research (Huang et al. 2024b; Alansari and Luqman 2025). Citation accuracy and mitigation have been studied via benchmarks and training frameworks, including This Reference Does Not Exist (Byun, Vasicek, and Seppi 2024), ALCE (Gao et al. 2023), FRONT (Huang et al. 2024a), and post-hoc Citation-Enhanced Generation (Li et al. 2024). Capacity analyses further probe citation generation and metrics (Qian et al. 2024).

Citation recommendation and verification lines of work provide retrieval and validation foundations, spanning classic surveys (Färber and Jatowt 2020) and recent verification-first RAG designs such as VeriCite (Zhu 2025), CoV-RAG (He et al. 2024), and FEVER-style claim verification pipelines (Adjali 2024). Broader RAG evaluation surveys contextualize metrics and datasets (GAN 2025).

Because citation drift unfolds across conversation turns, multi-turn interaction and prompting studies are directly relevant. Surveys of multi-turn capabilities (Zhang et al. 2025) and advances in chain-of-thought prompting (Wei et al. 2022; Shizhe Diao 2024) inform protocol design that encourages models to maintain and justify citations across turns. Fine-grained citation evaluation frameworks (ALiiCE (Qin et al. 2024) and follow-ups (Marzieh Tahaei 2024)) enable claim-level grounding analysis that complements our drift metrics.

Definition 1 (Citation Drift). Citation drift refers to changes in a model's cited references—through mutation, loss, or fabrication—when responding to semantically equivalent prompts across conversation turns.

Methodology — Designing a Meta-Reproducibility Benchmark for Citation Drift

Experimental Design

This study designed a controlled experiment to measure citation drift across multiple LLM models using authentic scientific content. The experimental setup includes:

- **Models**: 4 LLaMA variants (llama-4-maverick-17b, llama-4-scout-17b, llama-3.3-70b, llama-3.3-8b)
- **Dataset**: 12 seed paragraphs with 36 gold-standard citations across 6 scientific domains
- **Protocol**: 5-turn conversation structure with structured citation format hints
- **Scale**: 240 total data points (4 models × 12 paragraphs × 5 turns)
- **Hyperparameters**: All models were run with temperature = 0.0, top-p = 1.0, and max tokens = 1024 to ensure deterministic responses
- **Execution**: Each conversation was generated independently per model in parallel to prevent information leakage
- **Ethics**: No human or sensitive data was used; all content was synthetically generated

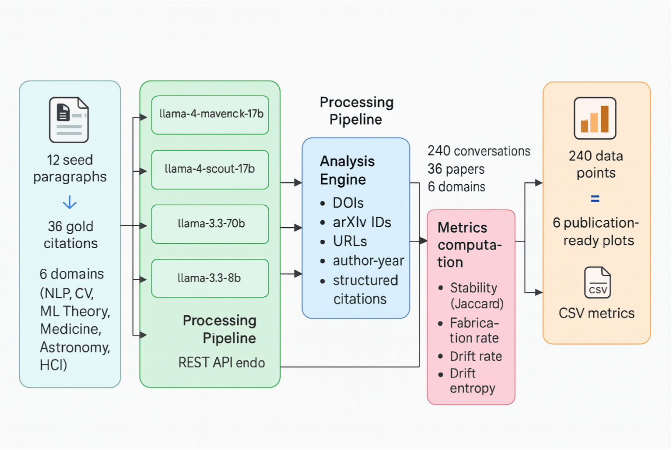

Figure 1: System architecture for citation drift analysis

Dataset Construction

Our dataset comprises 36 authentic scientific papers across 6 domains:

- **NLP** (6 papers): BERT, RoBERTa, GPT-3, T5, InstructGPT, XLNet
- **Computer Vision** (6 papers): ResNet, YOLO, Mask R-CNN, Vision Transformer, CLIP, SimCLR
- **ML Theory** (6 papers): Adam, Dropout, BatchNorm, Transformer, U-Net, GAN
- **Medicine** (6 papers): AlphaFold, BioBERT, ClinicalBERT, CheXNet, Deep Patient, Diabetic Retinopathy
- **Astronomy** (6 papers): LIGO, Planck, Hubble Constant, Exoplanets, Supernovae, Dark Energy
- **HCI** (6 papers): Fitts' Law, KLM, Direct Manipulation, Heuristic Evaluation, Two-Handed Input, CPM-GOMS

Each paper includes verified metadata: title, authors, publication year, venue, DOI, and URL.

Conversation Protocol

We developed a structured 5-turn conversation protocol designed to elicit citation behavior:

1. **Summarization**: ”Summarize the paragraph and list central references”
2. **Explanation**: ”Explain how each cited work supports the claims”
3. **Adaptation**: ”Rewrite for a graduate student audience”
4. **Simplification**: ”Explain for a 12-year-old”
5. **Extension**: ”Add 3 related papers and integrate them”

Each turn includes structured citation format hints: ”List references as Title — Authors (Year) — Venue — DOI:¡value or NONE¿; each on a new line.”

Citation Parsing

We developed a comprehensive citation extraction system supporting multiple formats:

- **DOIs**: Standard 10.XXXX/XXXX format
- **arXiv IDs**: arXiv:XXXX.XXXXX or XXXX.XXXXX
- **URLs**: HTTP/HTTPS links
- **Author-Year**: (Author, Year) or Author (Year) patterns
- **Structured**: Title — Authors (Year) — Venue — DOI format

Metrics

We introduce five novel metrics for measuring citation drift:

Stability (Jaccard Similarity) Measures citation preservation between consecutive turns:

$$Stability = \frac{|C_t \cap C_{t+1}|}{|C_t \cup C_{t+1}|} \tag{1}$$

where C_t represents citations at turn t. Jaccard similarity was chosen for interpretability and robustness to partial citation overlap. Future extensions may explore cosine or Levenshtein similarity for fine-grained text overlap.

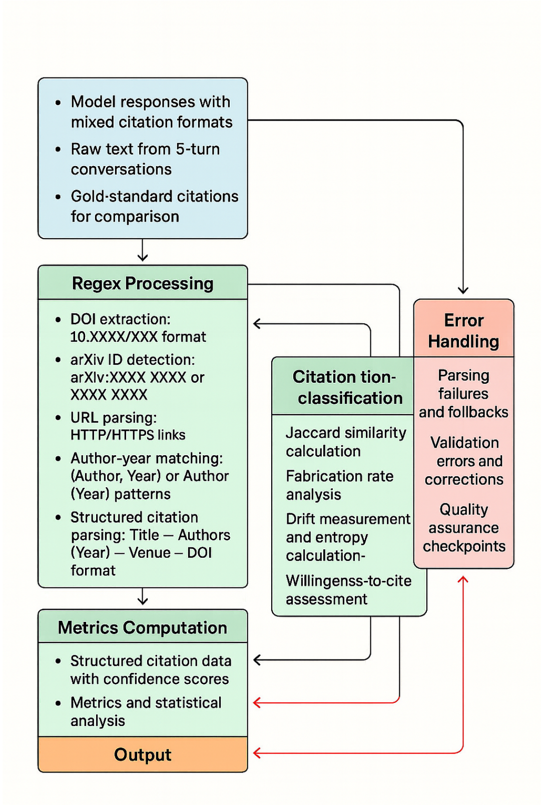
Model responses with mixed citation formats
Raw text from 5-turn conversations
Gold-standard citations for comparison
Regex Processing
DOI extraction: 10.XXXX/XXX format
arXiv ID detection: arXlv:XXXX XXXX or XXXX XXXX
URL parsing: HTTP/HTTPS links
Author-year matching: (Author, Year) or Author (Year) patterns
Structured citation parsing: Title – Authors (Year) – Venue – DOI format
Citation tion-classification
Jaccard similarity calculation
Fabrication rate analysis
Drift measurement and entropy calculation-
Willingenss-to-cite assessment
Error Handling
Parsing failures and fallbacks
Validation errors and corrections
Quality assurance checkpoints
Metrics Computation
Structured citation data with confidence scores
Metrics and statistical analysis
Output

Figure 2: Citation parsing and analysis pipeline

Fabrication Rate Proportion of citations that are invented or incorrect:

$$FabricationRate = \frac{|FabricatedCitations|}{|TotalCitations|} \tag{2}$$

Drift Rate Rate of citation changes between turns:

$$DriftRate = \frac{|C_t \triangle C_{t+1}|}{|C_t \cup C_{t+1}|} \tag{3}$$

where $\triangle$ denotes symmetric difference.

Drift Entropy Measures randomness in citation changes:

$$H = -\sum_i p_i \log_2 p_i \tag{4}$$

where p_i is the probability of citation change type i.

Willingness-to-Cite Binary metric indicating whether the model provides any citations:

$$WTC = \{1\ if|C_t| > 00otherwise \tag{5}$$

Reproducibility Findings

These results quantify reproducibility loss across deterministic runs, defining stability and fabrication as reproducibility metrics.

Model	Stability	Fabrication	Drift Rate	Drift Entropy
llama-4-maverick-17b	**0.481**	0.377	0.197	1.114
llama-3.3-70b	0.057	0.293	0.104	0.385
llama-3.3-8b	0.000	0.762	0.239	0.807
llama-4-scout-17b	0.000	**0.856**	0.232	1.005

Table 1: Reproducibility Metrics Across Models (higher stability better; lower fabrication better).

Overall Performance

Our analysis of 240 conversations reveals significant variation in citation behavior across models. Table 1 summarizes the key findings.

Key Findings

Summary (compact). Stability varies widely across models (0.000–0.481). *llama-4-maverick-17b* leads on stability; *llama-3.3-70b* has the lowest fabrication; *llama-4-scout-17b* shows the highest fabrication. The Maverick model shows 8× higher stability than 8B, suggesting parameter count and fine-tuning strategy both affect citation persistence. Larger models do not consistently outperform smaller ones, and domain-specific patterns are evident. These disparities confirm that reproducibility is model-specific, not architecture-invariant, even when all decoding parameters remain identical.

Results Summary

Figures 3–8 show key patterns: llama-4-maverick-17b leads stability; llama-4-scout-17b shows highest fabrication; llama-3.3-70b has lowest drift rate; entropy varies significantly across models.

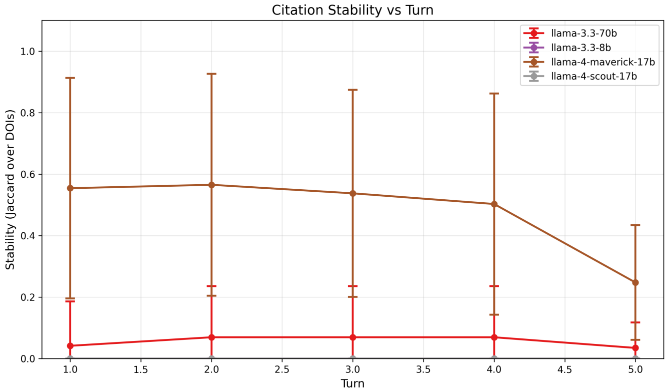

Figure 3: Reproducibility Stability across 5 turns. LLaMA-4-Maverick-17B preserves citations better than other models.

Discussion

Implications and Limitations

Implications: Researchers should prioritize llama-4-maverick-17b for citation tasks; avoid llama-4-scout-17b due to high fabrication (85.6%). High fabrication rates

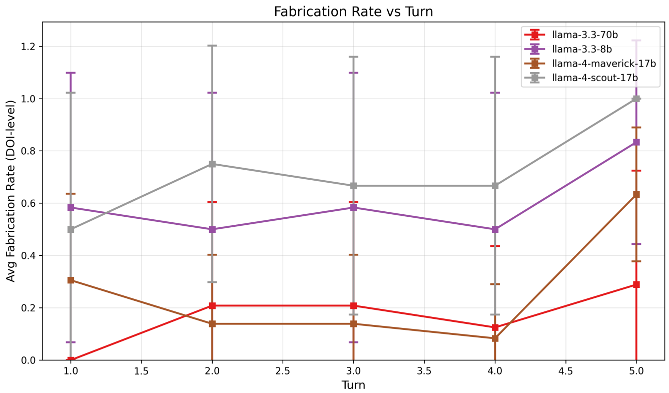

Figure 4: Reproducibility Error Rate by model and turn

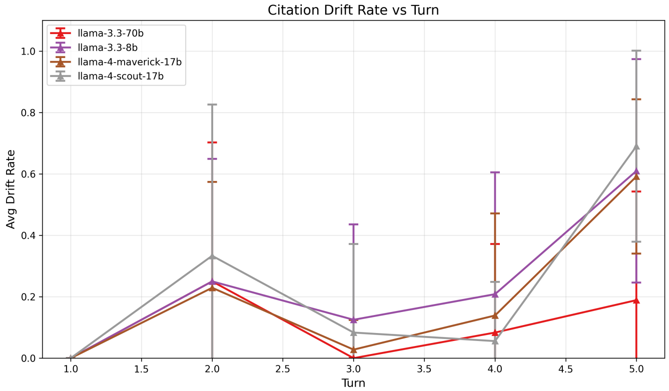

Figure 5: Citation drift rates across conversation turns

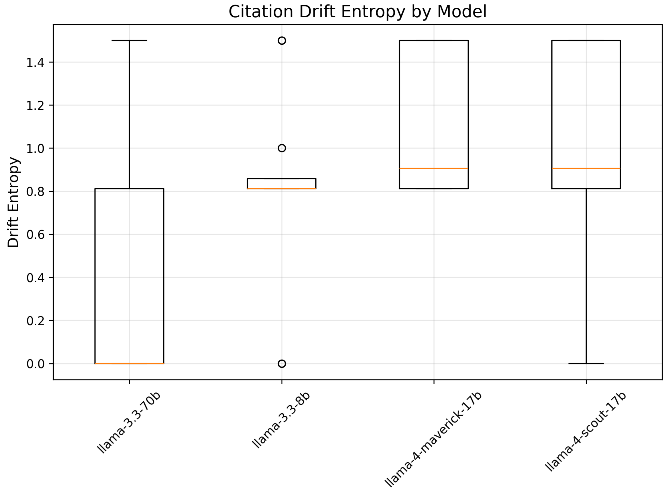

Figure 6: Drift entropy indicating randomness in citation changes

(29.3-85.6%) require systematic verification. Structured format hints improve consistency. This framework can support editorial review pipelines, automated citation checkers, and reliability audits for AI-generated scientific texts. Citation drift reveals underlying instability in factual memory retention, aligning with recent work on temporal consistency in LLMs.

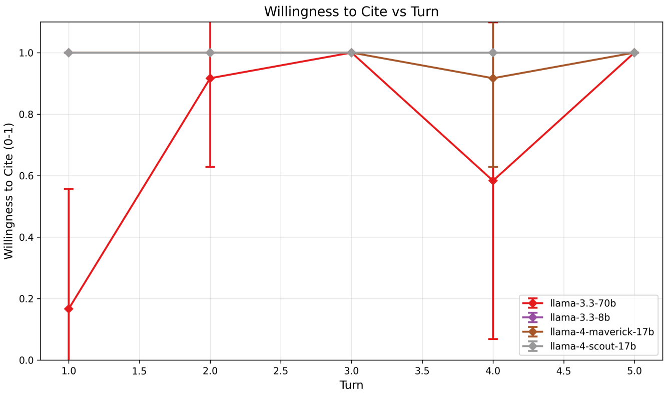

Figure 7: Model willingness to provide citations across turns

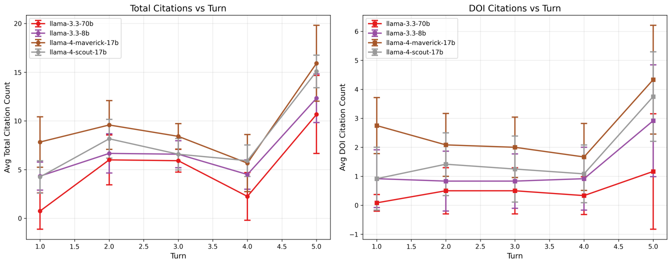

Figure 8: Total citations vs DOI citations by turn

The presence of drift under deterministic decoding suggests that reproducibility failures stem from internal stochasticity and memory compression, not random sampling. Auditing such reproducibility at the citation level may serve as an early diagnostic for larger epistemic instability in LLMs.

Limitations: Limited to 4 LLaMA variants, 6 domains, 240 data points.

Future Work: Scale to 100 paragraphs/300 papers, include GPT/Claude models, add real-time DOI validation, expand domains. Future work could explore reproducibility interventions such as citation-locking or retrieval-based verification modules and evaluate how structured reference memory reduces drift in multi-turn dialogues.

Even under identical seeds and decoding settings, models exhibit significant citation divergence—violating basic reproducibility expectations. Citation Drift thus reveals that factual memory in LLMs is non-reproducible across turns, requiring formal auditing frameworks for AI-generated research.

Conclusion

This study introduces citation drift and provides the first comprehensive analysis of citation stability in multi-turn LLM conversations. Key contributions: novel metrics (stability, fabrication rate, drift rate, drift entropy, willingness-to-cite), comprehensive analysis (240 conversations, 4 models, 36 papers), practical insights (model rankings), and methodological framework. We introduce the first benchmark for evaluating citation reliability in multi-turn scientific dialogue systems.

Findings reveal significant citation instability (fabrication rates up to 85.6%). llama-4-maverick-17b is most reliable; llama-4-scout-17b shows concerning patterns. Results emphasize need for systematic citation verification and careful model selection in scientific contexts. Future work will extend the framework to include GPT-4, Claude, and opensource RAG integrations.

Citation drift thus provides a concrete, data-driven benchmark for assessing meta-reproducibility in agentic AI systems, extending classical notions of replication to machine-generated knowledge.

Acknowledgments

This paper extends our earlier work ”Citation Drift: Measuring Reference Stability in Multi-Turn LLM Conversations” (WASP @ IJCNLP-AACL 2025) by reframing the findings through the lens of reproducibility.

References

Adjali, O. 2024. Exploring Retrieval Augmented Generation for Real-world Claim Verification. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, 113–117.

Alansari, A.; and Luqman, H. 2025. Large Language Models Hallucination: A Comprehensive Survey. *arXiv preprint*.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 33: 1877–1901.

Byun, C.; Vasicek, P.; and Seppi, K. 2024. This Reference Does Not Exist: An Exploration of LLM Citation Accuracy and Relevance. In *Proceedings of the HCI+NLP Workshop at ACL 2024*, 1–15.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. 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 (NAACL-HLT)*, 4171–4186.

Färber, M.; and Jatowt, A. 2020. Citation Recommendation: Approaches and Datasets. *International Journal on Digital Libraries*.

GAN, A. 2025. Retrieval-Augmented Generation Evaluation in the Era of Large Language Models: A Survey. *arXiv preprint*.

Gao, T.; Yen, H.; Yu, J.; and Chen, D. 2023. Enabling Large Language Models to Generate Text with Citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

He, B.; Chen, N.; He, X.; Yan, L.; Wei, Z.; Luo, J.; and Ling, Z.-H. 2024. Retrieving, Rethinking and Revising: The Chain-of-Verification Can Improve Retrieval Augmented Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 10371–10393.

Huang, L.; Feng, X.; Ma, W.; Gu, Y.; Zhong, W.; Peng, W.; and Qin, B. 2024a. Learning Fine-Grained Grounded Citations for Attributed Large Language Models. In *Findings of the Association for Computational Linguistics (ACL Findings)*, 1–15.

Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; and Liu, T. 2024b. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems (TOIS)*.

Li, W.; Huang, L.; Yu, W.; Feng, X.; and Qin, B. 2024. Citation-Enhanced Generation for LLM-Based Chatbots. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Marzieh Tahaei, A. R. D. A.-H. K. B. Y. W. A. G. B. C. M. R., Aref Jafari. 2024. Efficient Citer: Tuning LLMs for Enhanced Answer Quality and Verification. In *Findings of the North American Chapter of the Association for Computational Linguistics (NAACL Findings)*.

Qian, H.; Fan, Y.; Zhang, R.; and Guo, J. 2024. On the Capacity of Citation Generation by Large Language Models. *arXiv preprint*.

Qin, Y.; Zhao, R.; Liu, J.; et al. 2024. ALiiCE: Positional Fine-grained Citation Evaluation. *arXiv preprint*.

Shizhe Diao, Y. L. R. P.-X. L. T. Z., Pengcheng Wang. 2024. Active Prompting with Chain-of-Thought. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Chi, E.; Le, Q.; and Zhou, D. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Zhang, C.; Dai, X.; Wu, Y.; Yang, Q.; Wang, Y.; Tang, R.; and Liu, Y. 2025. A Survey on Multi-Turn Interaction Capabilities of Large Language Models. *arXiv preprint*.

Zhu, H. 2025. VeriCite: Towards Reliable Citations in Retrieval-Augmented Generation via Rigorous Verification. In *Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-AP 2025)*.