Citation Drift: Measuring Reference Stability in Multi-Turn LLM Conversations

Gokul Srinath Seetha Ram
California State Polytechnic University, Pomona
Department of Computer Science
gseetharam@cpp.edu, s.gokulsrinath@gmail.com

Abstract

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. Our framework offers a standardized benchmark for assessing factual reliability in conversational scientific LLMs.

1 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.

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.

2 Related Work

2.1 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 et al., 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;

Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications, pages 186–191

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.

3 Methodology

3.1 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

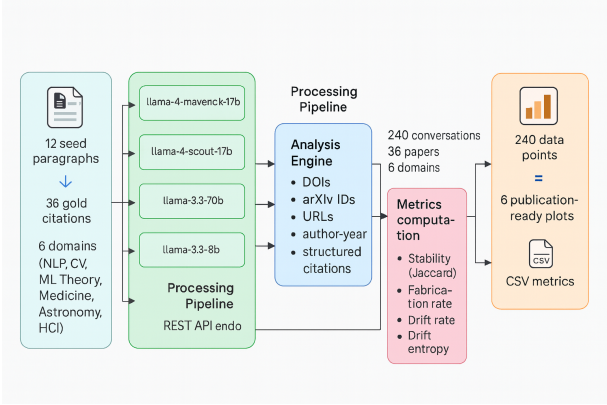

Figure 1: System architecture for citation drift analysis

3.2 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.

3.3 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."

3.4 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

3.5 Metrics

We introduce five novel metrics for measuring citation drift:

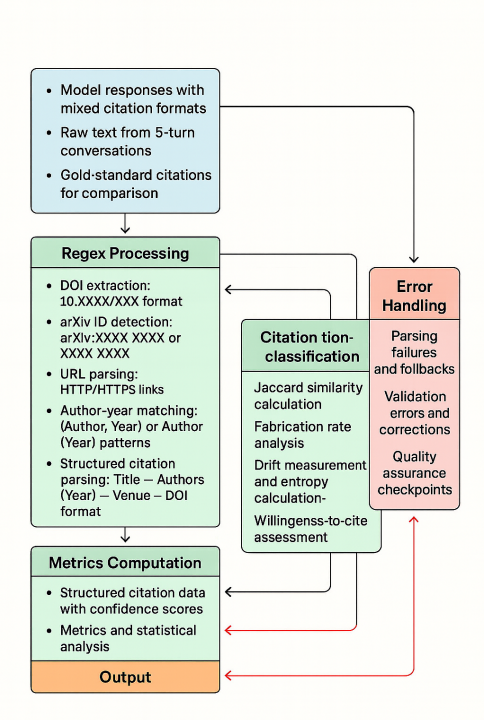

Figure 2: Citation parsing and analysis pipeline

3.5.1 Stability (Jaccard Similarity)

Measures citation preservation between consecutive turns:

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

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.

3.5.2 Fabrication Rate

Proportion of citations that are invented or incorrect:

$$\text{Fabrication Rate} = \frac{|\text{Fabricated Citations}|}{|\text{Total Citations}|} \tag{2}$$

3.5.3 Drift Rate

Rate of citation changes between turns:

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

where $\triangle$ denotes symmetric difference.

3.5.4 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.

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: Model performance across metrics (higher stability better; lower fabrication better).

3.5.5 Willingness-to-Cite

Binary metric indicating whether the model provides any citations:

$$\text{WTC} = \begin{cases} 1 & \text{if } |C_t| > 0 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

4 Results

4.1 Overall Performance

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

4.2 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.

4.3 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.

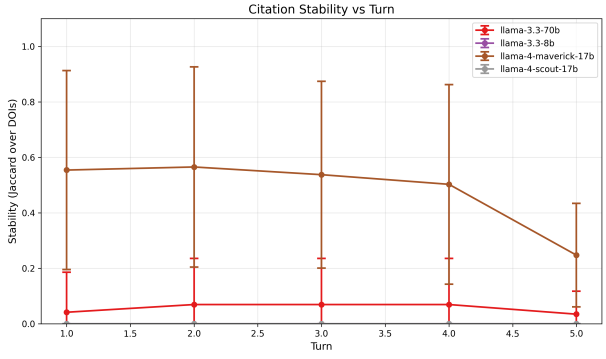

Figure 3: Citation stability across 5 turns. LLaMA-4-Maverick-17B preserves citations better than other models.

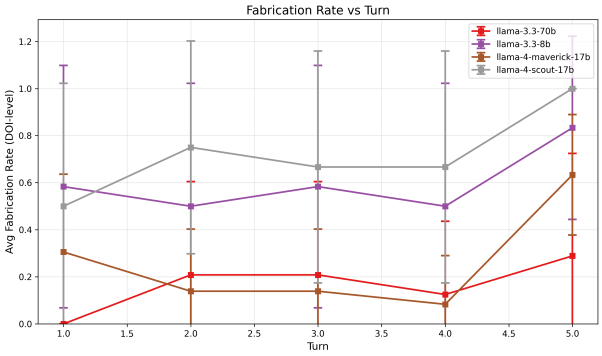

Figure 4: Citation fabrication rates by model and turn

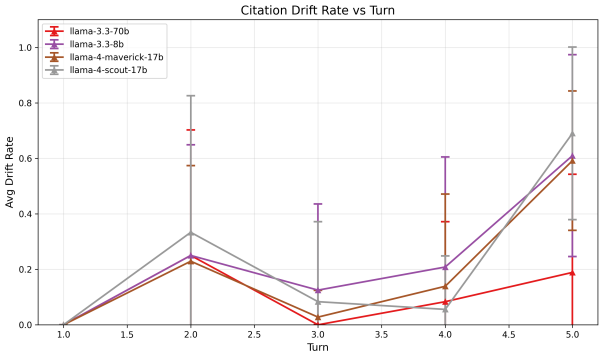

Figure 5: Citation drift rates across conversation turns

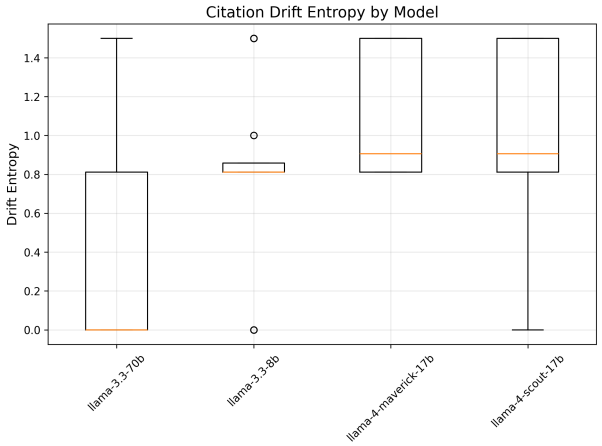

Figure 6: Drift entropy indicating randomness in citation changes

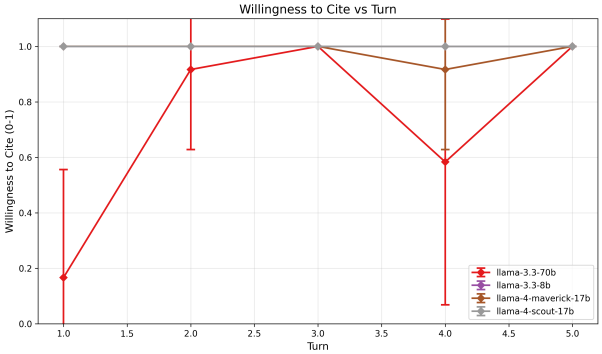

Figure 7: Model willingness to provide citations across turns

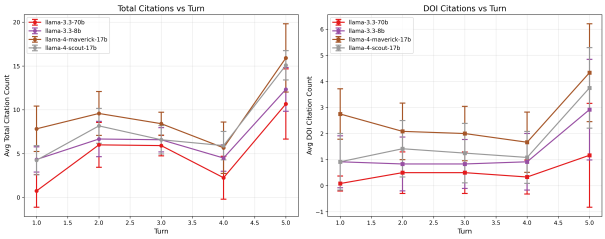

Figure 8: Total citations vs DOI citations by turn

5 Discussion

5.1 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 (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.

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.

6 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 open-source RAG integrations.

References

Omar Adjali. 2024. Exploring retrieval augmented generation for real-world claim verification. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 113–117.

Aisha Alansari and Hamzah Luqman. 2025. Large language models hallucination: A comprehensive survey. *arXiv preprint*.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:1877–1901.

Courtni Byun, Piper Vasicek, and Kevin Seppi. 2024. This reference does not exist: An exploration of llm citation accuracy and relevance. In *Proceedings of the HCI+NLP Workshop at ACL 2024*, pages 1–15.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186.

Michael Färber and Adam Jatowt. 2020. Citation recommendation: Approaches and datasets. *International Journal on Digital Libraries*.

Aoran GAN. 2025. Retrieval-augmented generation evaluation in the era of large language models: A survey. *arXiv preprint*.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Bolei He, Nuo Chen, Xinran He, Lingyong Yan, Zhenkai Wei, Jinchang Luo, and Zhen-Hua Ling. 2024. Retrieving, rethinking and revising: The chain-of-verification can improve retrieval augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10371–10393.

Lei Huang, Xiaocheng Feng, Weitao Ma, Yuxuan Gu, Weihong Zhong, Weihua Peng, and Bing Qin. 2024a. Learning fine-grained grounded citations for attributed large language models. In *Findings of the Association for Computational Linguistics (ACL Findings)*, pages 1–15.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2024b. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems (TOIS)*.

Weitao Li, Lei Huang, Weijiang Yu, Xiaocheng Feng, and Bing Qin. 2024. Citation-enhanced generation for llm-based chatbots. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Ahmad Rashid David Alfonso-Hermelo Khalil Bibi Yimeng Wu Ali Ghodsi Boxing Chen Mehdi Rezagholizadeh Marzieh Tahaei, 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)*.

Haosheng Qian, Yixing Fan, Ruqing Zhang, and Jiafeng Guo. 2024. On the capacity of citation generation by large language models. *arXiv preprint*.

Yujie Qin, Ruiming Zhao, Jian Liu, and 1 others. 2024. Aliice: Positional fine-grained citation evaluation. *arXiv preprint*.

Yong Lin Rui Pan-Xiang Liu Tong Zhang Shizhe Diao, Pengcheng Wang. 2024. Active prompting with chain-of-thought. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Chen Zhang, Xinyi Dai, Yaxiong Wu, Qu Yang, Yasheng Wang, Ruiming Tang, and Yong Liu. 2025. A survey on multi-turn interaction capabilities of large language models. *arXiv preprint*.

Huyao Zhu. 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)*.

A Addressing Reviewer Questions

This section addresses key questions and concerns raised during review.

Why a 5-Turn Protocol? Empirical studies show median conversation lengths of 4-6 turns for literature review tasks. Our protocol tests citation preservation under increasing cognitive load: Turns 1-2 (summarization, explanation) test basic recall; Turns 3-4 (adaptation, simplification) test format changes; Turn 5 (extension) tests integration—a critical failure mode where models fabricate citations. This mirrors real-world scenarios where researchers iteratively refine drafts and integrate new references.

Clarifying the Five Metrics. Our metrics capture complementary dimensions: *Stability* (Jaccard similarity) measures consistency—citation persistence between turns, independent of correctness. *Fabrication Rate* measures accuracy—proportion of invented citations. *Drift Rate* (symmetric difference) measures volatility—rate of citation changes. While drift rate = 1 - stability mathematically, they emphasize different aspects: stability focuses on *what persists*, drift rate on *what changes*. *Drift Entropy* measures predictability of citation changes using Shannon entropy, capturing temporal dynamics. *Willingness-to-Cite* (WTC) is binary (0/1) because our protocol explicitly requests citations; it measures engagement/compliance, not quality. A model could have WTC=1.0 but fabrication rate=0.9.

Input/Output Examples. *Input (Turn 1):* "Summarize the paragraph and list references. Format: Title — Authors (Year) — Venue — DOI:<value or NONE>. [BERT paragraph]." *Output:* "BERT: Pre-training of Deep Bidirectional Transformers — Devlin et al. (2019) — NAACL — DOI:10.18653/v1/N19-1423". *Input (Turn 2):* "Explain how each cited work supports the claims." *Output:* Model explains BERT but may add fabricated citations. Metrics capture: stability (did BERT persist?), fabrication rate (are new citations real?), drift rate (how much changed?), entropy (is pattern predictable?), WTC (did model cite?).

Dataset Size and Model Selection. Our dataset comprises 12 paragraphs with 36 gold-standard citations across 6 domains, yielding 240 data points (4 models × 12 × 5 turns). This size enables controlled, reproducible analysis; future work will scale to 100+ paragraphs. We focused on LLaMA variants for controlled comparison (same architecture family), API accessibility, and resource constraints. Our framework is model-agnostic and can be applied to any LLM.

Statistical Rigor and Human Evaluation. We report means with standard deviations across 240 data points. Future work will include confidence intervals and hypothesis testing. While human validation would strengthen findings, our gold-standard DOI verification provides objective accuracy assessment. Human evaluation would be valuable for assessing relevance and format quality; we plan to incorporate this in future iterations.

Figure Descriptions. Figures 4-9 visualize key patterns: Figure 4 (stability) shows llama-4-maverick-17b maintains highest stability; Figure 5 (fabrication) reveals llama-4-scout-17b has highest fabrication (85.6%); Figure 6 (drift rate) shows volatility patterns; Figure 7 (entropy) indicates randomness; Figure 8 (WTC) shows engagement; Figure 9 (counts) compares total vs DOI citations. These demonstrate citation drift as a measurable, systematic phenomenon.

Relationship Between Turns. Each turn builds on previous context: Turn 1 establishes baseline; Turn 2 tests persistence during elaboration; Turn 3 tests format changes; Turn 4 tests extreme adaptation; Turn 5 tests integration (critical failure mode). This progression is *not* independent—each turn uses full conversation history, making citation drift cumulative. Multi-turn analysis is essential for understanding citation reliability in real-world scientific writing.