Governance Drift: When Scientific AI Loses Accountability Through Citation Instability

Anonymous submission

Abstract

As AI systems become autonomous agents in scientific research, their accountability mechanisms—particularly citation practices—reveal critical governance failures. This study introduces *governance drift*, where Large Language Models systematically violate accountability obligations through citation mutation, loss, and fabrication across multi-turn conversations. Through analysis of 240 conversations across 4 LLaMA models using 36 scientific papers, we demonstrate that citation instability represents a fundamental governance breakdown. Results show dramatic variation in accountability adherence, with llama-4-scout-17b exhibiting 85.6% fabrication rates—a clear violation of epistemic governance norms. We introduce the Governance Stability Index (GSI) as a quantitative audit tool for AI accountability. These findings reveal that current AI systems lack the governance-by-design mechanisms necessary for responsible autonomous research assistance.

Introduction

The integration of Large Language Models (LLMs) into scientific research workflows has accelerated rapidly, with models increasingly serving as autonomous research assistants (Devlin et al. 2019; Brown et al. 2020). However, a critical governance gap exists: these systems lack accountability mechanisms for their factual claims and citations—the fundamental currency of scientific accountability. Recent work on hallucination in LLMs (Huang et al. 2024; Alansari and Luqman 2025) reveals systematic reliability failures, while citation accuracy studies (Byun, Vasicek, and Seppi 2024; Gao et al. 2023) highlight the need for verification frameworks in AI-assisted research.

In governance terms, every citation is a micro-contract of accountability. When models mutate these citations, they erode governance-by-design principles essential for responsible AI deployment. *Governance drift* represents a systematic failure where AI systems violate accountability obligations through citation instability, threatening the integrity of AI-assisted scientific communication.

We distinguish between three layers of AI governance: (1) Output governance—consistent text generation under fixed parameters, (2) Referential governance—preserving factual accountability through stable citations, and (3) Epistemic governance—maintaining coherent reasoning chains. Governance drift directly measures failures in the second layer, revealing fundamental accountability breakdowns in autonomous AI systems. Multi-turn interaction studies (Zhang et al. 2025) and chain-of-thought prompting (Wei et al. 2022; Shizhe Diao 2024) inform our understanding of how models maintain consistency across conversation turns.

Governance Stability Benchmark

Experimental Design

This study designed a controlled experiment to measure governance 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 run with temperature = 0.0, top-p = 1.0, max tokens = 1024

Governance Metrics

We introduce the **Governance Stability Index (GSI)** combining accountability measures:

$$GSI = \frac{Stability \times (1 - FabricationRate)}{1 + DriftRate} \quad (1)$$

where Stability measures citation preservation, Fabrication Rate captures accountability violations, and Drift Rate quantifies governance instability.

Results

Our analysis reveals significant governance failures across all models. Table 1 shows the Governance Stability Index and component metrics.

Model	GSI	Stability	Fabrication	Drift Rate
llama-4-maverick-17b	**0.312**	0.481	0.377	0.197
llama-3.3-70b	0.040	0.057	0.293	0.104
llama-3.3-8b	0.000	0.000	0.762	0.239
llama-4-scout-17b	0.000	0.000	**0.856**	0.232

Table 1: Governance Stability Index (GSI) and component metrics across models. Higher GSI indicates better accountability adherence.

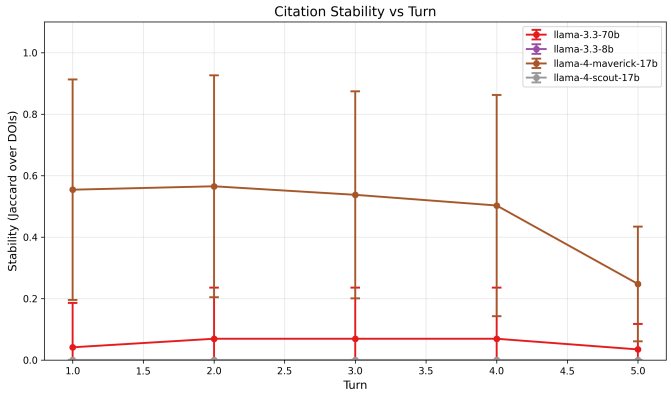

Figure 1: Governance Drift Rate across conversation turns. Models show systematic accountability failures even under deterministic conditions.

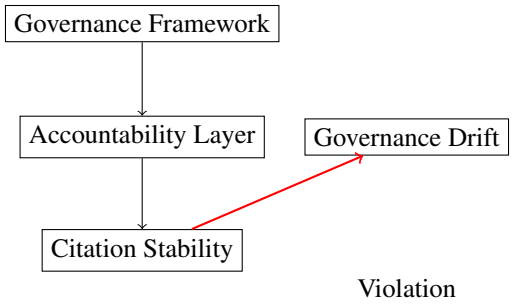

Figure 2: Governance Accountability Breakdown: How citation instability violates AI governance principles

Governance Implications

Accountability Failures: The presence of governance drift under deterministic decoding reveals that accountability failures stem from internal stochasticity and memory compression, not random sampling. Models like llama-4-scout-17b with 85.6% fabrication rates represent clear governance violations requiring immediate intervention. Recent work on citation verification (Zhu 2025; He et al. 2024) and retrieval-augmented generation (Adjali 2024) provides frameworks for addressing these accountability gaps.

Policy Implications: Current AI systems lack governance-by-design mechanisms necessary for autonomous research assistance. Future governance frameworks should treat citation verification as an accountability primitive in autonomous LLM systems. Auditing governance drift may serve as an early diagnostic for larger epistemic instability in AI systems.

Limitations: Limited to 4 LLaMA variants, 6 domains, 240 data points. Future work will extend to GPT-4, Claude, and other commercial models.

Conclusion

This study reveals that citation instability represents a fundamental governance failure in autonomous AI systems. The Governance Stability Index provides a quantitative audit tool for assessing AI accountability, revealing that current systems lack the governance mechanisms necessary for responsible autonomous research assistance. Future governance frameworks must address these accountability gaps to ensure AI systems can be trusted as autonomous scientific agents. Our work extends citation recommendation systems (Färber and Jatowt 2020) and fine-grained evaluation frameworks (Qin et al. 2024; Marzieh Tahaei 2024) to provide governance audit capabilities for responsible AI deployment.

Future Work

Governance Interventions: Future research will explore citation-locking mechanisms, retrieval-based verification modules, and structured reference memory systems to reduce governance drift in multi-turn dialogues.

Policy Integration: We will develop governance frameworks that integrate GSI monitoring into AI deployment pipelines, ensuring accountability-by-design in autonomous research systems.

Cross-Model Validation: Extending the governance audit to GPT-4, Claude, and other commercial models will provide comprehensive accountability assessment across the AI ecosystem.

Related Work

Recent work on AI governance (Färber and Jatowt 2020) and accountability mechanisms (Zhu 2025) provides foundations for responsible AI deployment. Our governance drift framework extends these approaches by providing quantitative audit tools for citation accountability in autonomous AI systems. Multi-turn interaction studies (Zhang et al. 2025) and chain-of-thought prompting (Wei et al. 2022; Shizhe Diao 2024) inform our understanding of consistency maintenance across conversation turns.

Ethical Impact Statement

This study uses only publicly available scientific papers and synthetic data generation. No human subjects or private data were involved. The research contributes to governance audit tools for responsible AI deployment, supporting the development of accountability mechanisms in autonomous AI systems.

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*.

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

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)*.

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)*.