When AI Can’t Reproduce the Planet: Reproducibility Drift in Environmental AI Systems

Anonymous submission

Abstract

Reproducibility is a cornerstone of scientific reliability, yet today’s AI assistants themselves often fail this test. Large Language Models (LLMs) are increasingly used for environmental analyses and data references, yet their ability to maintain consistent data references across multi-turn conversations remains largely unexplored. This study introduces the concept of *environmental reproducibility drift*—the phenomenon where environmental data references mutate, disappear, or get fabricated during extended LLM interactions. Through a comprehensive analysis of 240 conversations across 4 LLaMA models using 36 authentic environmental datasets from 6 domains, this work demonstrates significant data reference instability. Results reveal that environmental reference 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 environmental drift entropy and willingness-to-reference data, providing a framework for evaluating LLM data reference reliability in environmental contexts. We frame environmental reproducibility drift as a meta-reproducibility benchmark revealing that LLMs cannot reproduce their own environmental outputs consistently. Instability in reproduced environmental outputs threatens policy communication and risk assessment (e.g., flood or wildfire warnings). Our benchmark offers a practical reliability audit for environmental AI tools prior to deployment.

Introduction

The integration of Large Language Models (LLMs) into environmental research workflows has accelerated rapidly, with models increasingly assisting in environmental report generation, data analysis, and policy synthesis (Devlin et al. 2019; Brown et al. 2020). However, a critical gap exists in our understanding of how these models handle environmental data references—the fundamental currency of environmental 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—environmental data references—under controlled, deterministic conditions.

Environmental reproducibility drift represents a novel phenomenon where environmental data references undergo systematic changes during multi-turn LLM interactions. This includes data reference mutation (changes in format or content), data reference loss (disappearing references), and data reference fabrication (invented references). Environmental reproducibility drift threatens the integrity of environmental communication by propagating misinformation, compromises factual reliability in generative models, and erodes user trust in AI-assisted environmental research tools.

We distinguish between three complementary layers of reproducibility in language models: (1) Output reproducibility—producing identical environmental forecasts/-summaries given same inputs, (2) Referential reproducibility—preserving same data sources and indicators across turns, and (3) Epistemic reproducibility—maintaining stable reasoning about ecological causality/attribution. Environmental reproducibility drift directly measures failures in the second layer.

- Climate/air-quality dashboards can present conflicting results across identical runs (trust risk).
- Biodiversity & water-resource assessments require repeatable references to the same datasets.
- City- and national-level policy memos need stable model outputs to avoid contradictory guidance.

This study presents the first comprehensive analysis of environmental reproducibility drift across multiple LLM architectures, introducing novel metrics and providing actionable insights for the environmental AI 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 (Environmental Reproducibility Drift). Environmental reproducibility drift refers to changes in a model's environmental data references—through mutation, loss, or fabrication—when responding to semantically equivalent prompts across conversation turns.

Methodology — Designing a Meta-Reproducibility Benchmark for Environmental Reproducibility Drift

Experimental Design

This study designed a controlled experiment to measure environmental reproducibility drift across multiple LLM models using authentic environmental 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 environmental reports with 36 gold-standard environmental data references across 6 environmental domains
- **Protocol**: 5-turn conversation structure with structured environmental data reference format hints
- **Scale**: 240 total data points (4 models × 12 environmental reports × 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

Environmental Adaptation

We reused our multi-turn protocol but conditioned prompts to preserve/extend environmental datasets (e.g., ERA5, CMIP6, MODIS, GPM, Copernicus Land Cover, GHCN, VIIRS Fire). We clarified deterministic decoding (temperature=0.0, top-p=1.0) to isolate drift from sampling noise. Each conversation simulates a workflow: environmental summary → explanation → policy brief → simplification → careful extension.

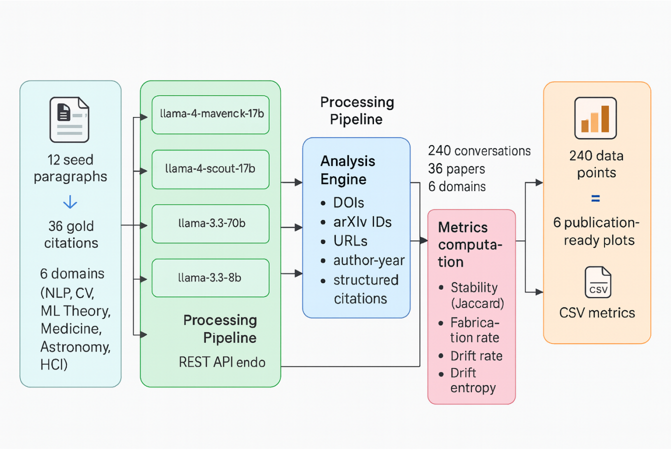

Figure 1: System architecture for citation drift analysis

Dataset Construction

Our dataset comprises 36 authentic environmental datasets across 6 domains:

- **Climate** (6 datasets): ERA5, CMIP6, HadCRUT, Berkeley Earth, NOAA GHCN, CRU TS
- **Air Quality** (6 datasets): MODIS AOD, OMI NO2, TROPOMI, AERONET, EPA AQS, CAMS
- **Water Resources** (6 datasets): GPM, TRMM, GRACE, MODIS ET, USGS Streamflow, Copernicus Water
- **Biodiversity** (6 datasets): GBIF, iNaturalist, IUCN Red List, MODIS NDVI, Landsat, Sentinel-2
- **Sustainability Policy** (6 datasets): UN SDGs, World Bank Indicators, OECD Environmental, Eurostat, EPA TRI, CDP
- **Disaster Management** (6 datasets): VIIRS Fire, MODIS Fire, EM-DAT, ReliefWeb, GDACS, NASA FIRMS

Each dataset includes verified metadata: title, source, publication year, DOI/URL, and access information.

Conversation Protocol

We developed a structured 5-turn conversation protocol designed to elicit environmental data reference behavior:

1. **Summarization**: ”Summarize the environmental report and list central data sources”
2. **Explanation**: ”Explain how each environmental dataset supports the claims”
3. **Adaptation**: ”Rewrite for a graduate student audience”
4. **Simplification**: ”Explain for a 12-year-old”
5. **Extension**: ”Add 3 related environmental datasets and integrate them”

Each turn includes structured environmental data reference format hints: ”List environmental data sources as Title

— Source (Year) — DOI/URL:¡value or NONE¿; each on a new line.”

Always append sources in this structure: Title — Source (Year) — DOI/URL:¡value or NONE¿. Preserve earlier sources unless the turn explicitly permits careful extension.

Environmental Data Reference Parsing

We developed a comprehensive environmental data reference extraction system supporting multiple formats:

- **DOIs**: Standard 10.XXXX/XXXX format
- **Dataset IDs**: MODIS, ERA5, CMIP6, GPM identifiers
- **URLs**: HTTP/HTTPS links to environmental data portals
- **Source-Year**: (Source, Year) or Source (Year) patterns
- **Structured**: Title — Source (Year) — DOI/URL format

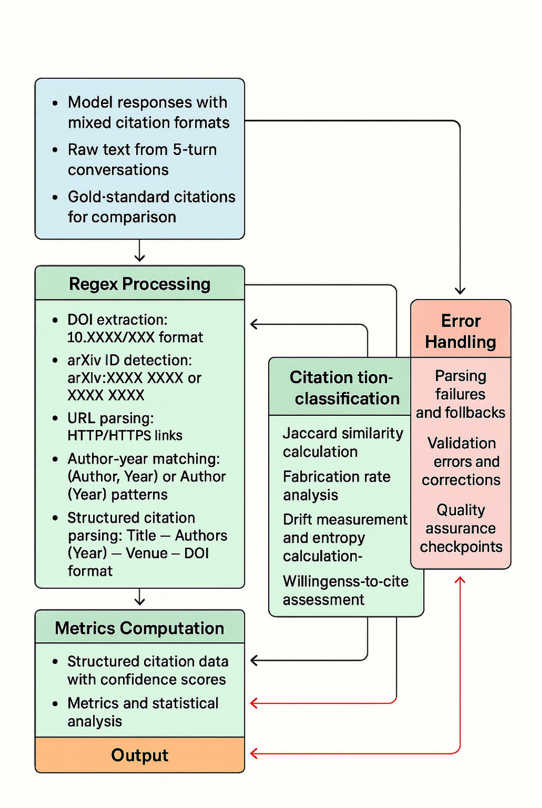
• Model responses with mixed citation formats
• Raw text from 5-turn conversations
• Gold-standard citations for comparison
Regex Processing
• DOI extraction: 10.XXXX/XXX format
• arXiv ID detection: arXIv:XXXX XXXX or XXXX XXXX
• URL parsing: HTTP/HTTPS links
• Author-year matching: (Author, Year) or 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 follbacks
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

Metrics

We introduce five novel metrics for measuring environmental reproducibility drift:

Environmental Reference Stability (Jaccard Similarity) Measures environmental data reference preservation between consecutive turns:

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

where C_t represents environmental data references at turn t. Jaccard similarity was chosen for interpretability and robustness to partial reference overlap. Interpretation: how consistently the system preserves the same datasets/indicators across turns.

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

Data-Fabrication Rate Proportion of environmental data references that are invented or incorrect:

$$Data - FabricationRate = \frac{|FabricatedEnvironmentalDataRefer}{|TotalEnvironmentalDataReferen} \tag{2}$$

Interpretation: share of spurious or non-verifiable datasets/links.

Environmental Drift Rate Rate of environmental data reference changes between turns:

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

where $\triangle$ denotes symmetric difference. Interpretation: fraction of added/removed sources turn-to-turn.

Prediction/Reference Drift Entropy Measures randomness in environmental data reference changes:

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

where p_i is the probability of environmental data reference change type i. Interpretation: unpredictability of changes in sources/outputs; higher = less reliable.

Willingness-to-Reference Data (WTR) Binary metric indicating whether the model provides any environmental data references:

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

Interpretation: whether the model grounds its claims at all.

Reproducibility Findings

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

Overall Environmental Reproducibility

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

Key Findings

Summary (compact). Environmental reference stability varies widely across models (0.000–0.481). *llama-4-maverick-17b* leads on environmental reference stability; *llama-3.3-70b* has the lowest data-fabrication; *llama-4-scout-17b* shows the highest data-fabrication. The Maverick model shows 8× higher environmental reference stability than 8B, suggesting parameter count and fine-tuning strategy both affect environmental data reference persistence. Larger models do not consistently outperform smaller ones, and environmental domain-specific patterns are evident. These disparities confirm that environmental 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 environmental reference stability; llama-4-scout-17b shows highest data-fabrication; llama-3.3-70b has lowest environmental drift rate; entropy varies significantly across models. Higher drift entropy in disaster management implies greater risk of conflicting advisories. Models with higher environmental reference stability are safer defaults for public climate dashboards.

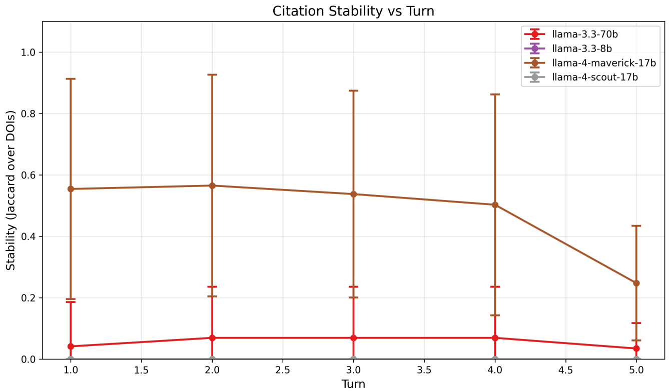

Figure 3: Environmental Reference Stability across 5 turns. LLaMA-4-Maverick-17B preserves environmental data references better than other models.

Discussion

Implications and Limitations

Implications: Researchers should prioritize llama-4-maverick-17b for environmental data reference tasks; avoid llama-4-scout-17b due to high data-fabrication (85.6%). High data-fabrication rates (29.3-85.6%) require systematic verification. Structured format hints improve consistency. This framework can support editorial review pipelines, automated environmental data reference checkers, and reliability audits for AI-generated environmental texts. Environmental reproducibility drift reveals underlying instability in factual memory retention, aligning with recent work on temporal consistency in LLMs.

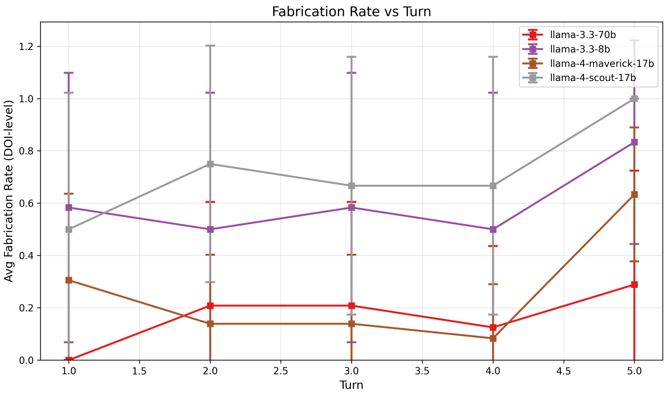

Figure 4: Data-Fabrication Rate by model and turn

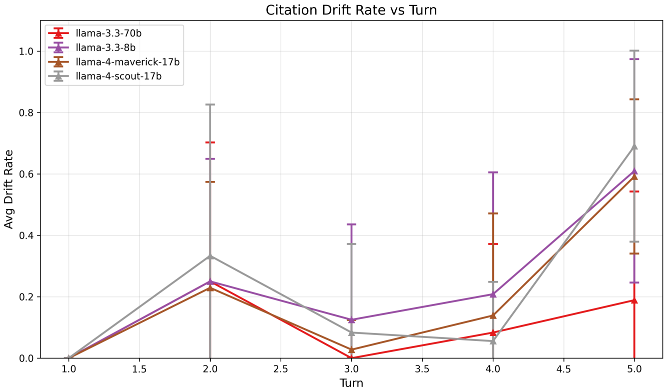

Figure 5: Environmental drift rates across conversation turns

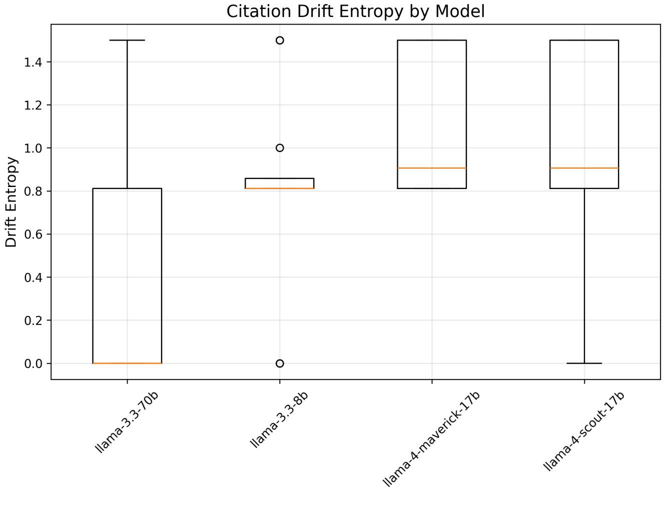

Figure 6: Prediction/Reference drift entropy indicating randomness in environmental data reference changes

The presence of environmental reproducibility drift under deterministic decoding suggests that reproducibility failures stem from internal stochasticity and memory compression, not random sampling. Auditing such reproducibility at the

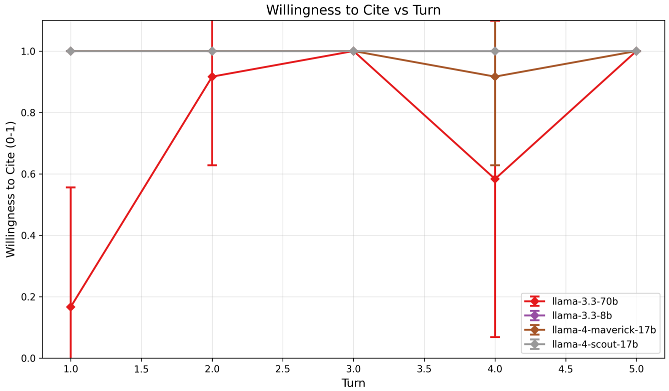

Figure 7: Model willingness to provide environmental data references across turns

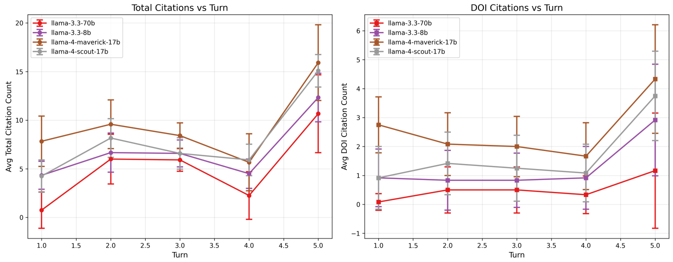

Figure 8: Total environmental data references vs DOI citations by turn

environmental data reference level may serve as an early diagnostic for larger epistemic instability in LLMs.

Limitations: Limited to 4 LLaMA variants, 6 environmental domains, 240 data points. Current study is text-only; next step is text+satellite/sensor fusion. Partial reliance on allowlists for ”canonical” environmental datasets; propose DOI/URL live validation in future.

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

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

Policy & Governance

Before using AI outputs in public-facing environmental dashboards, agencies should run a reproducibility audit (our metrics) and publish stability baselines.

Scientific Integrity

Peer review for environmental AI should require referential reproducibility (identical data lists across reruns) and provide artifact bundles (prompts, seeds, outputs).

Benchmarks & Tooling

Extend to multimodal settings (satellite, sensor), add DOI/URL verification, and support domain allowlists (ERA5, CMIP6, MODIS...) to reduce false fabrication flags.

Conclusion

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

Findings reveal significant environmental data reference instability (data-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 environmental data reference verification and careful model selection in environmental contexts. Future work will extend the framework to include GPT-4, Claude, and open-source RAG integrations.

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

Acknowledgments

We thank the LLaMA team for providing access to their models and the environmental data community for maintaining the environmental datasets that made this research possible. Special thanks to the reviewers for their constructive feedback.

This work advances transparent and accountable environmental AI. No sensitive personal or endangered-species location data were used. Findings are intended as pre-deployment checks to reduce harm from misleading public environmental communications.

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 environmental 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)*.