yAI SYSTEMS NOTE 002

Why Citation-Based RAG Still Hallucinates

Version 1.0 March 15, 2026

Yassin Hafid

Founder & CEO

LinkedIn

ABSTRACT

Citation-based RAG aims to improve groundedness by retrieving evidence and attaching citations. Yet empirical studies show that citations do not guarantee that claims are supported by retrieved evidence, nor that the model actually relied on the cited passages rather than post-hoc rationalizing. This note explains why citation-based RAG systems can still produce unsupported or misgrounded statements despite retrieving evidence.

1. The Question

If a system retrieves relevant passages and generates answers with citations, why do hallucinations and unsupported claims still occur?

2. Scope and Definitions

Citation-based RAG

A system that retrieves passages from an external corpus, generates an answer conditioned on those passages, and outputs citations intended to support the answer.

Grounding

The extent to which generated claims are supported by the retrieved evidence set.

Citation Correctness

Whether the cited source actually supports the associated claim.

Citation Faithfulness

Whether the model genuinely derived its claim from the cited source, rather than generating it from parametric memory and retroactively matching it to a retrieved passage.

---

3. Key Findings

- **Correct citations can still be unfaithful:** Citation correctness and citation faithfulness can diverge substantially, meaning citations can look valid while not reflecting the evidence actually used.[1]

- **Citations do not guarantee complete support:** Even strong systems leave portions of long-form answers uncited or unsupported; on challenging datasets like ELI5, complete citation support is absent roughly 50% of the time. Answer-level correctness masks these partial failures, motivating claim-level evaluation.[2][4]

- **Factual precision degrades with entity rarity:** Atomic-level evaluation reveals factual precision can drop dramatically, from roughly 80% for frequently-discussed entities to as low as 16% for rare ones in the same model,[4] suggesting the model may rely more heavily on parametric knowledge rather than retrieved evidence.

- **RAG reduces but does not eliminate grounding failures:** Retrieval-augmented models generate more factual outputs than parametric-only baselines;[5] however, documented failure modes persist across both the retrieval and generation stages, including selection of misaligned passages, hallucination of content unsupported by retrieved context, and incoherent integration of multiple sources.[6]

- **Prevalence of Unattributed Content:** Over 95% of answers produced by tested open-source LLMs contain at least one sentence lacking any attribution.[3]

---

4. Technical Deep Dive: Why Systems Fail

A. Post Rationalization (The "Search-to-Justify" Bias)

Faithfulness failures occur when a model generates a claim from parametric memory and then scans retrieved documents for post-hoc supporting tokens (i.e., post-rationalization) rather than causally deriving the answer from evidence. Wallat et al. find 57% of citations in a RAG-optimized model showed unfaithful behavior, though they note the causal interpretation requires further validation.[1]

B. Multi-Source Synthesis & "Hallucinated Bridges"

Hallucinations can arise during multi-source synthesis. Even if individual passages are correct, combining them can yield disjointed, incoherent, or redundant outputs; and models are readily distracted by irrelevant passages when aggregating across sources.[2][6] To maintain narrative coherence, models may confabulate connective claims between evidence fragments, what we term "hallucinated bridges" (unsupported connective statements inserted to prioritize narrative flow over evidentiary gaps), producing statements not supported by any single retrieved passage.

C. The Illusion of Groundedness

Citations should be treated as an interface for verification, not proof of correctness. Because responses often contain a mixture of supported and unsupported statements, citations can create an "illusion of groundedness" that masks partial hallucinations.

Unfaithful citations "foster misplaced trust" precisely because they are indistinguishable from faithful ones at the output level; since a post-rationalized citation looks identical to a genuine one, surface-level citation checking is an insufficient trust signal.[1] This is further complicated by the fact that citation presence does not necessarily imply factual precision. Empirically, citation presence has been shown to have near-zero correlation with factual accuracy in search-

augmented systems, with supported and unsupported statements cited at nearly identical rates.[4]

---

5. Practical Taxonomy of Failure Modes

- **Retrieval Miss:** The needed evidence is not retrieved; the model fills gaps using priors.[6]
- **Evidence Insufficiency:** Retrieved passages are related but either do not contain enough information to support the claim, or the model fails to utilize the available evidence due to multi-document synthesis limitations and context window constraints.[2][6]
- **Evidence Misinterpretation:** The model misreads the evidence (numbers, scope, negations), producing unsupported claims despite relevant context.[1]
- **Multi-source Synthesis Error:** Even when individual passages are relevant, combining them can produce incoherent, redundant, or distracted outputs; indeed, models struggle to aggregate evidence across sources without being misled by irrelevant passages.[2][6]
- **Attribution Failure:** Citations are factually disconnected from the text, incomplete, or unfaithful (post-hoc).[1][2]
- **Architectural Attribution Failure:** In Generate-then-retrieve (GTR) and post-hoc Retrieve-then-generate (RTG) pipelines, citations are assigned after generation in a step independent of the evidence used during answer generation, making faithful attribution difficult to guarantee regardless of retrieval quality.[1]

---

6. Implications for AI System Design

- **Measure claim-level grounding:** Evaluate long-form generations at the level of atomic facts to detect partial hallucinations.[4]
- **Mandate Inline Attribution:** GTR and post-hoc RTG pipelines are structurally incapable of faithful citation. Systems where citations are generated inline during answer generation, rather than assigned afterward, are a necessary (though not sufficient) condition for faithfulness.[1]
- **Evaluate citation quality explicitly:** Systems should distinguish citation correctness from citation faithfulness.[1]
- **Treat RAG as a pipeline with multiple failure points:** Improving retrieval alone cannot eliminate hallucinations as errors originate in interpretation, synthesis, and attribution.[1][6]

---

7. Open Questions

- **Faithfulness measurement:** How can we test whether a model actually relied on cited evidence rather than post-hoc rationalizing? MIRAGE[3] offers a promising direction using gradient-based saliency, but lacks scalability.
- **Evaluation unit:** What is the right unit for grounding evaluation: answer, sentence, or atomic claim, and how should results be aggregated for decision-critical settings?
- **Actionable diagnosis:** How can systems reliably distinguish retrieval failures from evidence-integration failures in a way that guides engineering fixes?
- **Benchmark realism:** What evaluation frameworks best capture multi-hop reasoning over noisy, heterogeneous document collections with partial evidence?
- **The Narrative Bias Trade-off:** How can we balance the "hallucinated bridges" required for human-readable coherence with the strict grounding required for factual reliability?

---

8. References

[1] Wallat et al., Correctness is not Faithfulness in RAG Attributions, arXiv:2412.18004, 2024.

[2] Gao et al., Enabling Large Language Models to Generate Text with Citations (ALCE), EMNLP, 2023.

[3] Qi et al., Model Internals-based Answer Attribution for Trustworthy RAG (MIRAGE), EMNLP, 2024.

[4] Min et al., FActScore: Fine-grained Atomic Evaluation of Factual Precision, EMNLP, 2023.

[5] Lewis et al., Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, NeurIPS, 2020.

[6] Gao et al., Retrieval-Augmented Generation for Large Language Models: A Survey, arXiv:2312.10997, 2023.

This note synthesizes findings from recent research on citation-based RAG systems and hallucination. The interpretations presented reflect the author's reading of the current literature.