Qeios

Peer Review

Review of: "100% Hallucination Elimination Using Acurai"

Iliano Fasolino[1]

1. Department of Computer Science, University of Milan, Italy

The calcium / magnesium example in §3 is a clean illustration of a real RAG failure mode: the model is given clear facts about one entity and, under a query that names a close neighbor, swaps the two. Splitting colliding noun phrases into separate sub-queries, pairing each with a short fact set, and placeholder-mapping identifiers are practical moves. The GitHub dump of questions and answers is more than many product-adjacent papers offer. The title, abstract, and §6 still announce “100% hallucination elimination,” and that claim is the problem.

Four reviews are already on the page. Arsentev (LLM844) already pressed the n = 37 design, the missing control re-runs, author-only judging, the Appendix A “NONE” verdict against the paper’s own “any deviation” rule, the unspecified collision-detection / rewrite pipeline (§3’s “among others,” unnamed embedding model and threshold, proprietary-vs-product framing, unpublished rewrite artefacts), and a coverage or answer-usefulness score beside binary faithfulness. Magee (3JME5G) noted that the title sells a general cure while the method is RAG faithfulness under supplied context, and that faithfulness to a false context is not correctness. Gandolfi (FXO089) asked for a scoped title and a clearer deployment story. Anghelescu (ZWR8V5) largely accepted the headline. The authors have not replied to any of them. I defer to Arsentev on those points; they remain unmet. What I add is a retrieval-robustness ask those reviews did not press: break the factual-chunk assumption.

§5 says the result “solely applies to RAG-based implementations that only send factual chunks,” and that RAGTruth passages are “almost exclusively relevant.” That is not a small caveat. In a poisoned or tampered corpus, or after post-retrieval rewriting by an adversary, “faithful” generation means repeating the poison. Acurai, as evaluated, never sees that regime. There is no PoisonedRAG-style injection, no contradictory neighbor planted next to a clean passage, and no test where the Fully-Formatted Facts themselves were altered after retrieval—yet the pipeline’s middle step is exactly a rewrite of passages into

 doi.org/10.32388/MLTGOM

those fact sets. If an adversary, or an ordinary upstream RAG bug, contaminates the chunk that later becomes an FFF, the method as scored has no defense beyond repeating it cleanly. For a system pitched at "enterprise and other high-stakes applications" in the abstract, honesty about faithfulness-under-attack matters more than another clean three-passage cell. Prior reviews asked for controls on the same 37 items; they did not ask the authors to break their own factual-chunk assumption. I would.

Coverage, as Arsentev asked, remains unmet after his review (author replies still 0). Faithfulness alone rewards thin or near-copy answers; Appendix A's "Specifics / Detail 1" block that tracks the first passage closely can look perfect under a deviation rubric while carrying little synthesis. That ask stands; I will not re-brief it.

I also read the Noun-Phrase Dominance Model as a hypothesis that outruns the experiment. "All hallucinations occur when the LLM mistakes two distinct Noun-Phrases as being the same thing" (§3) is universal. The evaluation only re-tests items already labelled as conflicts for GPT-3.5 Turbo and GPT-4. Numeric slips, negation flips, unsupported causal bridges, and date arithmetic are not shown to reduce to noun-phrase collisions. Treating OpenAI / Anthropic sparse-autoencoder features as "Noun-Phrases," and leaning on the authors' own videos and the blog titled as if those labs "confirm" Acurai ([13], [15]–[17]), does not turn the hypothesis into a result. Soften the theory language to match the calcium-style error class you actually intervene on. The "100%" headline compounds the same overreach: it converts a faithfulness score on a conflict-labelled slice into a claim about elimination as such.

On the statistics already debated: §4.3's interval on 37/37 successes is fine as a small-sample bound and is not a warrant for "100%" in the title or for talk of the "general population." Selection on items that once failed, without paired unmodified re-runs at the same temperature, still leaves open how much of the gain is regression to a better sample. That point stands whether or not one accepts Arsentev's reading of Appendix A.

I would retitle to the measured setting—faithfulness on conflict-labelled RAGTruth items for two OpenAI models after query/passage rewriting—drop the universal elimination language, add a corrupted-context and a previously-correct arm, and (with Arsentev) report coverage next to faithfulness and open the rewrite path enough that a reader can tell method from hand-editing. The entity-confusion idea is worth keeping. The present headline is not.

Wishing you the best.

Declarations

Potential competing interests: No potential competing interests to declare.