

Peer Review

Review of: "100% Hallucination Elimination Using Acurai"

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

The paper presents Acurai, a pre-processing and post-processing layer placed around a commercial LLM in a retrieval-augmented setting. Its theoretical basis is what the authors call the Noun-Phrase Dominance Model: the claim (§3, paragraph after the calcium example) that "all hallucinations occur when the LLM mistakes two distinct Noun-Phrases as being the same thing", specifically when those noun phrases are semantically close, as with calcium and magnesium. From this, the method follows in three steps: split a query so that colliding noun phrases land in separate sub-queries (§3.1), rewrite the passages into short self-contained "Fully-Formatted Facts" and pair each sub-query only with the facts about its own noun phrase (§3.2), and replace names, citations, and identifiers with placeholders that are mapped back into the answer afterwards (§3.3). The text says these are example steps, "among others".

The empirical claim is that, on responses from the RAGTruth corpus that were annotated as Evident Conflict or Subtle Conflict for GPT-3.5 Turbo and GPT-4, re-sending the same query and passages through Acurai to the same model produced hallucination-free answers in every case (§4.2, Figure 2). §4.3 gives the sample as $n = 37$ with $X = 37$ successes and a 95% Wilson interval of [0.91, 1]. The title, abstract, and §6 present this as "100% hallucination elimination" and state that "no other system or method has been shown to systematically eliminate hallucinations across any sample size."

There are things here worth saying plainly. The idea that entity confusion between semantically similar referents drives a share of RAG errors is a reasonable, testable hypothesis, and the calcium/magnesium case in §3 is a vivid illustration of it. Placeholder substitution for identifiers (§3.3) is a practical trick that many deployed systems would benefit from. The authors also do something I respect: §4.1 documents concrete defects in RAGTruth with response IDs (8285, 9824, 9692), which a reader can check, and §5 lists

limitations, including cost, latency, and the three-passage setting. My concern is not with the idea. It is that the evaluation, as described, cannot support the strength of the headline.

2. Whether the conclusion holds on the data

2.1 Sample size and what 37 items can show. The whole quantitative result rests on $n = 37$ (§4.3, "n is the sample size (37)"). As rendered on Qeios, Figure 2 is not legible in the text, so I could not see how the 37 split across the four subsets listed in §4.2 (GPT-3.5 Turbo Subtle, GPT-3.5 Turbo Evident, GPT-4 Subtle, GPT-4 Evident). If the split is roughly even, each cell holds nine or ten items, and a per-model claim rests on under twenty. The paper also does not say how these 37 were drawn from RAGTruth: whether they are all conflict-labelled responses for these two models, only the QA task (Appendix A is a QA item), a random subset, or a hand-picked one. Without that, the reader cannot judge what population the interval refers to.

I checked the arithmetic in §4.3. For $X = n = 37$, the Wilson lower bound reduces to $n/(n + z^2) = 37/(37 + 3.8416) = 37/40.8416 = 0.906$, which rounds to the reported 0.91. So the number is right. The formula printed above it is not the Wilson interval, though. It takes the adjusted centre $\tilde{p} = (X + z^2/2)/(n + z^2)$ and adds $\pm z\sqrt{(\tilde{p}(1-\tilde{p}))/(n + z^2)}$, which is the Agresti–Coull construction. Applied literally, it gives $\tilde{p} = 38.9208/40.8416 = 0.953$, half-width $1.96 \times \sqrt{(0.953 \times 0.047)/40.8416} = 0.065$, so $[0.888, 1.018]$, truncated to $[0.89, 1]$. The text should print the formula it actually used. (The scrape also cuts the definition of z off after "a 95", which I assume is a rendering loss.)

The interpretation in §4.3 goes further than the interval allows: "Acurai is expected to eliminate between 91% to 100% of hallucinations in the general population under the same conditions." A confidence interval computed on 37 non-randomly selected items from one corpus, two models, and one task format is not an estimate for a general population. And even taken at face value, a lower bound of 0.91 means the data are compatible with roughly one failure in eleven. That is a respectable result. It is not "100% elimination," and the title should not say so.

2.2 No control arm, and selection on the outcome. This is the central problem. Every item in the evaluation was chosen because the original RAGTruth response hallucinated (§4.2: "Every response in each of the datasets is a hallucination"). The treatment was then applied, and the new output judged. But the comparison is between a single historical sample and a fresh sample, not between two conditions run at the same time. If the models are sampled at non-zero temperature (§4.2 says "Same Temperature As RAGTruth" and never gives the value), a plain re-run of the unmodified prompt would also turn some

fraction of these items into faithful answers. Items selected for having failed once will, on average, look better the second time. The paper needs the obvious baseline: the same 37 items re-sent to the same model snapshots without Acurai, several times, judged by the same procedure. Only the difference between that rate and 37/37 is attributable to the method.

A second baseline is also missing. Acurai sends a model a handful of short, pre-filtered fact statements per sub-query. Simply giving the model a shorter, cleaner context with an instruction to answer only from it may account for much of the effect. An ablation that removes the noun-phrase splitting while keeping the passage simplification (or the reverse) would test whether the Noun-Phrase Dominance Model is doing the work, which is the paper's theoretical claim.

2.3 No measure of the other error direction. The evaluation contains only items that originally failed. There are no items that originally succeeded, so the paper cannot show that Acurai does not introduce new hallucinations on previously correct responses, through a bad split, a wrongly remapped placeholder (§3.3), or a lossy rewrite into Fully-Formatted Facts. For a system proposed for "high-stakes applications" (abstract), that is the second half of the safety claim, and it is absent.

Faithfulness also has a trivial solution. An answer that repeats the passages verbatim, or says very little, cannot deviate from them. §6 says users "receive accurate responses, instead of being informed that a response cannot be given," but no completeness, relevance, or answer-quality measure is reported. Appendix A hints at the risk: the Acurai output ends with a "Specifics / Detail 1" block that is a near-verbatim copy of the first passage. I would ask for at least a coverage measure (what share of the query-relevant facts in the passages appear in the answer) alongside faithfulness.

2.4 Who judged, and how. §4.1 explains that the authors set aside the RAGTruth annotations because of three corpus defects, and §4.2 says: "we independently compared Acurai's response to the provided passages." As far as I can tell from the text, the judges were the authors of the method, who are presenting a named product, with no blinding, no second annotator, no written rubric beyond "deviated from the provided passages in any manner whatsoever," and no agreement statistic. With 37/37 as the claimed result, the judging procedure carries the entire burden of proof. It needs independent annotators blind to condition (original vs. Acurai), a mix of original and Acurai outputs presented in random order, and a reported agreement figure. An LLM-based faithfulness judge could serve as a secondary check, but not as a replacement.

There is an internal tension here as well. The paper rejects the RAGTruth annotations as unreliable (§4.1, "Wrong annotations", response 9692), yet uses those same annotations to decide which responses count

as hallucinations in the first place. If response 9692 was wrongly labelled an Evident Conflict, then it was never a hallucination, and a faithful Acurai answer on it shows nothing. I could not tell from the text whether 9692 and similar mislabelled items are among the 37. They should be removed or reported separately.

2.5 The showcase example, read against the paper's own criterion. I read Appendix A closely because it is the one full item the reader can check without the repository. The criterion in §4.2 is strict: any deviation from the passages is a hallucination. The passage frames the benefits with an attribution: "It has been said that this technique can provide a variety of benefits to your body". The Acurai response turns part of that into "it's been found to help reduce symptoms associated with thyroid issues and Pre-menstrual Syndrome (PMS)". "It has been said" and "it's been found" are not the same claim; the second asserts evidence the passage does not contain. The response also opens with "The technique is simple", which is nowhere in the passages, and it presents "an overall boost to your mood and mental health" without the hedge. The table at the end of Appendix A reports "Error with Acurai response using GPT-4-0613: NONE". By the RAGTruth definition of a subtle conflict, and by the paper's own "in any manner whatsoever" rule, I think at least the attribution shift would be flagged. Meanwhile, the original GPT-4 error was "base of the neck" for "base of your skull on your neck", which is arguably a smaller change. I may be stricter than the authors intend, and reasonable annotators can differ on this. That is exactly why independent, blinded annotation (2.4) is needed before a 100% figure is published.

2.6 The theoretical claims exceed their support. The statement that "all hallucinations occur" through noun-phrase confusion (§3) is universal and is not tested here. Evaluating on items where the method worked does not test it; a test would need hallucinations of other kinds (numeric errors, wrong negation, unsupported causal claims, date arithmetic) and a check of whether they reduce to noun-phrase collisions. The support offered is the authors' own blog post [13] and video [15][16][17]. §2 then says OpenAI's and Anthropic's interpretability work "precisely aligns" with the model and the authors "refer to such features as Noun-Phrases". The cited works [11][12] describe sparse-autoencoder features, many of which are not noun phrases at all. Renaming features as noun phrases does not make those studies evidence for the model, and the reference title for [13] ("OpenAI and Anthropic Confirm Acurai's Discoveries") overstates what those papers say.

2.7 The motivating 80% figure is a different quantity. §1 and §2 ("Long Context RAG") cite the Databricks study [6] to say no LLM exceeds 80% accuracy "even when provided with relevant and accurate context" (abstract). That study, as the paper itself describes it in §2, measures end-to-end

answer accuracy with varying amounts of retrieved context, which mixes retrieval quality, long-context degradation, and correctness. It is not a faithfulness score on three clean passages. Figure 1 is not legible as rendered on Qeios, so I could not check its axes. The sentence "the belief that if you send an LLM 100% factual and relevant data, you will get 100% factual results, is entirely unfounded" draws a faithfulness conclusion from a correctness-and-retrieval benchmark, and then the paper's own result is measured on a different metric. The two numbers should not be set side by side.

2.8 Claims of priority. §6 says "No other system or method has been shown to systematically eliminate hallucinations across any sample size." This is a claim about the whole literature, and no survey of RAGTruth results or faithfulness methods is given to back it. The Related Works section discusses benchmarks and interpretability but no competing mitigation methods (constrained decoding, citation-grounded generation, post-hoc verification, fine-tuned faithfulness models). Without that comparison, the sentence should go.

3. Reproducibility

The authors point to a public GitHub repository (§4.2) and say it holds the datasets and "all of the questions and responses from GPT-3.5 Turbo, GPT-4 and Acurai". That is more than many papers in this area provide, and it makes the output side checkable. I reviewed the Qeios text only and did not inspect the repository, so everything below concerns what the paper itself states.

The method is not specified. This is the main gap. §3 describes three steps "among others" by example, but gives no algorithm for any of them. How are noun-phrase collisions detected? §3.1 mentions high cosine similarity of embeddings but names no embedding model and no threshold. Who or what rewrites the passages into Fully-Formatted Facts, a human, a rule system, or an LLM, and if an LLM, which one with what prompt? How are facts assigned to sub-queries? How are sub-answers merged into the single response shown in Appendix A? A reader who wanted to re-implement Acurai from this text could not. If parts of the pipeline are proprietary, the paper should say so explicitly, and the claim should then be framed as an evaluation of a product rather than of a reproducible method.

This matters beyond replication. If any rewriting step was done by hand, or tuned while looking at the 37 items, then knowledge of where the original model failed could leak into the input, and the result would reflect the rewriter rather than the method.

Intermediate artefacts. As rendered on Qeios, the passages in §3.2 ("the following three passages were sent") and the calcium fact sets ("Acurai sent the following fact sets") do not appear in the text; both are

followed directly by the next paragraph. They may be images lost in the scrape. The rewritten queries and Fully-Formatted Facts for all 37 items should be released and described in the paper, since they are the actual inputs the models received.

Model versions and decoding. Appendix A names "GPT-4-0613". No snapshot is given for GPT-3.5 Turbo. Temperature is stated only as "Same Temperature As RAGTruth" (§4.2) with no value; top-p, max tokens, system prompt, and seed are not reported. The paper uses "GPT-4", "ChatGPT-4", and "GPT-4-0613" interchangeably (§4.2, §4.3, Appendix A), and ChatGPT is a consumer product, not an API snapshot. Dates of the API calls are not given; given that older GPT-3.5 snapshots have been retired, the exact endpoint matters for anyone trying to rerun this.

Number of runs. The paper does not say whether each item was run once or several times, or whether any outputs were discarded. With 37/37 reported, the reader needs to know that there was one pre-registered run per item and no re-tries.

Selection and judging records. The rule used to select the 37 items from RAGTruth, the per-item judging decision, and the reason for each decision are not in the paper. A table listing response ID, model, conflict type, and judged outcome would take one page.

Declarations. On the scraped page, "Affiliations" and "Declarations" appear only as collapsed headings, so I could not read them. Since the paper evaluates a named product and cites the first author's own promotional material as support [13][15][16][17], a competing-interest statement should be visible in the body of the paper.

4. What specifically to fix

1. **Title, abstract, §6.** Replace "100% Hallucination Elimination" with the measured result: 37 of 37 items judged faithful, 95% CI [0.91, 1] (§4.3), on two OpenAI models and one corpus.
2. **§4.2, add a control arm.** Re-send the same 37 unmodified prompts to the same model snapshots, ideally five or more times each at the RAGTruth temperature, and judge them with the same procedure. Report both rates and a paired test on the item level.
3. **§4.2.** Add previously non-hallucinated RAGTruth responses to the evaluation so that newly introduced hallucinations can be counted.
4. **§4.2, judging.** Use at least two annotators who are not authors, blind to condition, with a written rubric, and report Cohen's κ or a similar agreement statistic. State who performed the judging in the current version.

5. **Appendix A.** Re-examine the "NONE" verdict. The shift from "It has been said" (passage) to "it's been found" (Acurai response) and the added "The technique is simple" look like deviations under the §4.2 criterion. If the authors disagree, explain the rule that makes them acceptable, and apply that same rule to the original GPT-4 responses.
6. **§4.1 vs. §4.2.** Say whether items with wrong RAGTruth annotations (for example, response 9692) are in the 37, and exclude them from the success count.
7. **§4.2 and Figure 2.** Give the item counts per subset in a table in the text, and state the selection rule (all eligible items, random sample with seed, or manual choice, and which RAGTruth task types).
8. **§4.3.** The printed formula is Agresti–Coull, not Wilson; replace it with the Wilson formula that yields 0.906, and delete the sentence about the "general population". The z definition is truncated after "a 95".
9. **§3 as a whole.** Specify the algorithm: collision detection (embedding model, similarity threshold), the procedure that produces Fully-Formatted Facts (human or model, with prompts), fact-to-query assignment, and how sub-answers are recombined. If some of it cannot be disclosed, say so.
10. **§3.2.** Restore the three passages and the calcium fact sets, which do not appear as rendered.
11. **§4.2, "Details of methodology".** Give numbers instead of "Same Temperature As RAGTruth": temperature, top-p, max tokens, model snapshots for both models, system prompt, call dates, and how many runs per item.
12. Add an ablation that separates query splitting (§3.1) from passage rewriting (§3.2) and placeholder remapping (§3.3). Without it, the Noun-Phrase Dominance Model is not tested, only the full pipeline.
13. **§3, "all hallucinations occur when..."**. Soften to a hypothesis, or test it on hallucination types that are not entity confusion.
14. **§2, "Internal Representation of LLMs".** Remove "precisely aligns" and the equation of SAE features with noun phrases, or support it with a specific result from [11] or [12]. Replace self-published blog and video sources [13][15][16][17] with the underlying data where possible.
15. **§1, §2, and abstract.** Stop using the Databricks 80% figure [6] as a faithfulness baseline; it measures a different thing.
16. Report a coverage or completeness metric next to faithfulness.
17. **§6, last paragraph.** Remove the claim that no other method has done this at any sample size, or back it with a comparison against published mitigation methods on RAGTruth.
18. **Terminology.** Use API model names consistently ("ChatGPT-4" in §4.2 and §4.3 vs. "GPT-4-0613" in Appendix A). Also, check §4.1, "Wrong information": "No state minimum wage law" reads like a

scraped table heading rather than a factual claim, which weakens that example as a contradiction.

19. Make the competing-interest statement explicit in the body.

5. Recommendation

Major revisions. The hypothesis about noun-phrase collisions is interesting, and the practical steps in §3 may well help. But the headline result is 37 items, selected for having failed, judged by the method's authors, with no control arm and no check for newly introduced errors, and the one fully shown example (Appendix A) contains what I read as a deviation from the passages. The title and conclusions claim far more than that design can support. With a control, blinded judging, a larger and documented sample, and a specified method, this could become a useful paper.

Declarations

Potential competing interests: No potential competing interests to declare.