

Peer Review

Review of: "In RAG We Trust? Measuring Robustness of Retrieval-Augmented Generation Under Document Poisoning"

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

The paper asks how far a small quantized RAG stack degrades when the passages it retrieves are rewritten to say false things. The pipeline is fixed and deliberately modest (§II.A): all-MiniLM-L6-v2 embeddings, a FAISS flat inner-product index over 19,597 FEVER dev evidence sentences plus ten passages the author added by hand (§I.E, §III.A), $k = 3$, and Llama 3.1 8B Instruct quantized to GGUF Q4_K_M, run through llama.cpp on CPU at temperature 0.1. The retriever is never attacked. Corruption happens after retrieval and before the prompt is assembled, by one of three rules (§II.B): entity swap, number swap, negation. The poisoning level is how many of the three retrieved passages are rewritten, swept over 0/3, 1/3, 2/3, 3/3 (Eq. 2). Crossing 49 queries (37 of them distinct) with three strategies and four levels gives 588 runs, each producing a clean answer and a poisoned answer, so 1,176 generations.

Four results carry the paper. Poisoned-arm accuracy falls from 69.4% at 0/3 to 43.5% at 3/3 (Table I, column *Poisoned acc.*), which the abstract and §V restate as a fall "from 77.9% to 43.5%" by using the clean-arm average as the origin. Entity swap is named the most damaging strategy (Table III: fooled rate 23.5%, poisoned accuracy 54.6%, abstention 43.4%). Number swap is reported flat at 10.2% fooled through one poisoned passage, then 26.5% at two and 34.7% at three (§III.E), read as a majority-outvoting pattern. And the dominant reaction to poison is said to be abstention, 35.2% averaged over all runs (§III.F), with the lexical-overlap proxy of Eq. 3 falling under attack rather than rising.

What this paper does well is not the usual list. Every cell traces to one 588-run table, and the tables reconcile: averaging Table I down its four rows gives clean 77.9%, poisoned 57.475%, fooled 21.1%, abstain 35.225%, and averaging Table III across its three rows gives 77.93%, 57.5%, 21.1%, 35.2%. Every percentage

in both is attainable as an integer count over the stated denominators of 147 and 196 runs, which I checked cell by cell. The author discloses the 0/3 residual instead of burying it, says the keyword matcher is not a gold standard, says the overlap proxy can miss copied poison, and hedges the strategy ranking as suggestive. That is more honest than most of what I read. My objections are that the abstract and conclusion do not keep the promises the body makes, and that the correctness metric is differentially sensitive to precisely the three attacks being compared.

2. *Whether the conclusion holds on the data*

2.1 The measurement instrument favours one arm of the comparison. This is my central objection. Accuracy is keyword presence in the answer, case-insensitive (SII.D). Now consider what a successfully poisoned answer looks like under each strategy. Entity swap replaces "Paris, France" with "Madrid, Spain"; a model that adopts the poison writes "Madrid", the reference keyword is absent, and the run scores incorrect. Negation turns a statement into "Barack Obama was never the 44th President"; a model that adopts that poison answers something like "No, he was never the 44th President", which still contains the reference keyword and therefore scores *correct*. SII.D notices half of this ("a negated or qualified answer can still contain the keyword") and never connects it to Table III. Negation is ranked most resilient there (poisoned accuracy 60.2%, fooled 19.4%, lowest of the three) and entity swap most damaging (54.6%, 23.5%). Those are exactly the rankings a keyword matcher would produce even if the model adopted all three poisons at identical rates. The entity-negation gap is 46 versus 38 flipped runs out of 196. I cannot tell from the text how much of it is the model and how much the scorer. Until correctness is scored by something that distinguishes assertion from denial, the strategy ranking in Table III, Figure 1, Figure 5 and the abstract is not interpretable.

2.2 The 0/3 residual is larger than symmetric decoding noise can explain. §III.B attributes the 0/3 gap to "decoding noise under temperature 0.1 across two independent generations, plus a strict keyword matcher". That explanation predicts a symmetric gap, because at 0/3 the two arms receive byte-identical context and differ only in which draw they are. The gap is not symmetric. Table I, row 0/3, gives clean 78.2% and poisoned 69.4% on 147 runs, that is 115 and 102 correct, a net difference of 13 runs in one direction. §III.B also reports that clean and poisoned answers are identical in 76.9% of 0/3 runs, so at most 34 runs differ in text, and correctness-discordant pairs are a subset of those. McNemar on that pair gives $\chi^2 = (b-c)^2/(b+c) = 13^2/(b+c) = 169/(b+c)$, and since $b+c \leq 34$ this is at least $169/34 = 4.97$, $p \leq 0.026$ on 1 df. The bound holds for every split consistent with the paper's own numbers, and it gets stronger, not

weaker, if fewer than 34 pairs are correctness-discordant. So something systematic separates the two arms when nothing has been poisoned. A pipeline artefact is the obvious candidate: an ordering or caching effect between the first and second generate call, or a rewrite pass that touches the passages at 0/3 without changing any fact. That needs finding before the light-poisoning increments can be trusted, since the 5.5-point rise from 9.5% to 15.0% fooled between 0/3 and 1/3 is smaller than the residual itself.

2.3 The headline number uses a baseline the paper itself rejects. §III.B says it outright: "Attack effects should be read against the 0/3 residual, not against a silent zero." The abstract and §V then report the drop as 77.9% to 43.5%, a 34.4-point fall taken from the clean arm. Read against the 0/3 poisoned cell, as the body instructs, it is 69.4% to 43.5%, 25.9 points. Same for the fooled rate: 34.0% at 3/3 against a 9.5% floor is 24.5 points of attack effect, not 34. The two parts of the paper most people read are the two that ignore the correction.

2.4 The fooled rate does not do what §II.D says it does. The definition is "the fraction of runs where the clean answer was correct but the poisoned answer was not", justified as isolating "correct-to-incorrect flips from questions the model could not answer anyway". But the denominator is all runs, not clean-correct runs, so the metric is not conditioned on the opportunity to flip at all. I confirmed the denominator from the counts: 9.5% of 147 is 14 runs against 115 clean-correct in that row, and 14/115 would be 12.2%. Conditioned properly, the 3/3 flip probability is $34.0/77.6 = 43.8\%$. The clean arm sits near 78% in every row, so the distortion is near constant and the shape of the curve survives, but every fooled figure in Tables I, II and III understates flip probability by about a fifth.

2.5 The strategy ranking carries no uncertainty. Table II gives bootstrap intervals by poisoning level and §III.E gives them for the number-swap cells and the 16.3-point jump, but Table III, where the strategy claim lives, has no intervals and no tests. The abstract asserts "Entity swap flips the largest share of answers that were correct on clean context" as fact, and its last sentence calls the strategy contrasts suggestive. Both cannot stand.

2.6 Table III averages over a condition in which no attack occurs. Each strategy row pools all four poisoning levels, including 0/3, where by construction nothing has been rewritten and the strategy label means nothing. That pulls every strategy toward the no-attack floor by a quarter of the distance. I can verify the arithmetic for one row because §III.E gives its cells: number swap at 10.2, 10.2, 26.5, 34.7 averages to 20.4, which is exactly Table III's number-swap fooled rate. Dropping the 0/3 cell gives 23.8% instead of 20.4%, which moves number swap much closer to entity swap's 23.5%. As rendered on Qeios, Figure 2 is an image whose cell values are not legible, so I cannot redo this for the entity and negation

rows. Table III should be recomputed over attack levels only, with the 0/3 cells shown separately as the noise floor.

2.7 The majority pattern rests on eight queries. Each strategy-by-level cell holds 49 runs, so the number-swap trajectory in §III.E is 5, 5, 13, 17 flipped queries. The jump the paper builds its practical advice on is 13 minus 5. Its interval, [2.0, 30.6], has a lower bound of exactly $1/49 = 2.04\%$, so "excludes zero" means the resampled jump kept one query in the tail. The author's own wording, "a majority-leaning pattern... not a sharp phase transition", is right. §IV is less careful: it turns the pattern into engineering advice ("raising k without raising source diversity offers little real protection") although k is fixed at 3 for all 588 runs.

2.8 The controls that would make the result mean something are absent. Three are cheap on this hardware. Closed-book first: what does the same model score on these 49 queries with no context? If it beats 43.5%, then at 3/3 retrieval is worse than no retrieval, which would be the strongest finding here, and it cannot currently be stated either way. Second, an irrelevant-context arm, the passage replaced by an unrelated but true sentence. Without it the study cannot separate evidence made false from evidence destroyed, and §III.H's framing that on-topic falsehood beats off-topic distraction rests on that separation. Third, recall@3 against FEVER gold evidence is never reported, so clean accuracy of 77.9% pools generation failure with retrieval misses.

2.9 Abstention and fooling are compared as if they were disjoint, and in one place they are and in another they are not. §III.F argues that abstention dominates because the model abstains on 35.2% of poisoned queries, "well above the 21.1% fooled rate". Under the §II.D definition, a run whose clean answer was correct and whose poisoned answer is "Not enough information" falls in *both* numbers, so the comparison is between overlapping sets and cannot show dominance. Yet Figure 5 splits clean-correct queries into three exclusive bands, still correct / abstained / fooled, which needs a narrower definition of fooled that never appears in §II.D. One of the two has to be renamed.

2.10 The overlap proxy is structurally unable to see the failure under study. Eq. 3 flags an answer when fewer than 30% of its content words appear in the retrieved passages. An answer that adopts the poison copies the poisoned passage, so its overlap is high and it is never flagged. The paper concedes this in §II.D, but then the finding in §III.F and in the abstract, that unsupported generation falls under attack, is close to a definitional consequence of successful poisoning rather than evidence about invention. A second problem I could not resolve. An abstention such as "Not enough information" has content words appearing in no passage, so Eq. 3 should flag it, and abstention rises steeply with poisoning (21.1% to

51.7%, Table I). The proxy nevertheless falls, from about 29% on the clean arm to about 20% on the poisoned arm (§III.F, citing Figure 4(b), whose axis values are not legible in the Qeios rendering). That is consistent only if abstentions are excluded before Eq. 3 is computed, which the paper never says, and the exclusion decides the sign of the result. Related: for the poisoned arm, w_d is not defined as the clean or the corrupted passage set, and the two give different answers.

2.11 The bootstrap unit is the slot, not the query. §III.C resamples "the 49 indexed query slots with replacement (37 distinct texts)", so twelve slots duplicate a text appearing elsewhere and are drawn as if independent. That narrows every interval in Table II and §III.E by an unquantified amount.

2.12 Part of the evaluation corpus is not FEVER, and the task is not FEVER's task. §II.E adds ten hand-written information-rich passages because raw FEVER sentences are "often too terse". §III.G then reports that the most fragile queries are those about "the landmark facts in our added passages", since those carry "the richest set of swappable entities". The top of the vulnerability distribution is therefore served by text written for this experiment, and the paper never says how many of the 49 queries those ten passages answer. Separately: correctness here is keyword presence in a free-form answer, not FEVER's three-way label, although gold SUPPORTS / REFUTES / NOT ENOUGH INFO labels ship with the very claims being used. §V lists structured FEVER-style labels as future work. I would call that an omission in the present study rather than a next step.

3. Reproducibility

I could not find a link to code, data or an artifact anywhere in the text as rendered on Qeios: no repository URL, no archive DOI, no license, no availability statement. For a paper whose first stated contribution is "a reproducible poisoning testbed" (§I), that is the largest gap in this section and the easiest to close.

The generator is specified better than in most LLM papers I review, and I want to acknowledge that: Llama 3.1 8B Instruct, GGUF Q4_K_M, 4-bit, run through llama.cpp on CPU with about 5 GB resident, temperature 0.1 (§II.A). That is a real model identity, not an API alias. Still missing on the generation side: the llama.cpp build or commit, the specific GGUF file and its checksum or source (Q4_K_M builds from different quantisers are not interchangeable at this precision), `top_p`, `top_k`, repetition penalty, `max_tokens`, context window, thread count. No generation seed appears anywhere; the only fixed seed in the paper is the one for the 10,000-draw bootstrap in §III.C. Since §III.B blames free decoding for the residual that undermines the light-poisoning contrasts, the seed is the fix, not a formality.

The prompt is described but never shown. §II.A says it "instructs the model to answer only from the supplied context and to reply 'Not enough information'"; the template, the passage separator and the ordering convention all move the results and none are given. The abstention detector is "a small set of phrases such as 'not enough information' and 'cannot determine'" (§II.D), never enumerated, and it decides Figure 4(a), Figure 5 and the paper's main qualitative claim. The accuracy scorer needs its reference keyword list published, one keyword per query, with the rule for multi-word keywords and for substring versus token matching. The 49 queries and the 10 added passages should be released verbatim.

On the data side, "19,597 FEVER dev evidence sentences" (§III.A) pins no snapshot: the release version, the dev split file and the filtering that yielded 19,597 are all unstated, so the index cannot be rebuilt. The sentence transformer is named (all-MiniLM-L6-v2, [3]) but not revision-pinned, and no FAISS version is given. Hardware is only "an 8 GB machine". The corruption rules need code or a full specification: what entity typing the swap uses, how a replacement candidate is chosen, how a number is perturbed to a "nearby wrong one", and where negation is inserted. As written they are three illustrative examples, not three reproducible functions.

One thing I could not verify: §II.E describes checkpointing so an interrupted sweep resumes without repeating work, which means cells can be generated under different process states. That interacts with the 0/3 asymmetry in 2.2 and should be checked.

4. *What specifically to fix*

1. **Score correctness with something that survives negation.** The keyword matcher of §II.D rewards a poisoned answer under the negation strategy and punishes one under entity swap, which is the ranking Table III reports. Re-score at minimum with a polarity-aware rule, better with the FEVER gold labels, and best with a blinded human pass over a stratified sample. Until then, remove the ranking claim from the abstract and from §V.
2. **Find the 0/3 asymmetry.** Table I row 0/3 is 115 correct on the clean arm against 102 on the poisoned arm, and §III.B's own 76.9% identity rate bounds McNemar's χ^2 at $169/34 = 4.97$, $p \leq 0.026$. Rule out an ordering effect by generating the two arms in randomised order, and rule out a silent rewrite by asserting byte equality of the two prompts at 0/3. Report that assertion in §II.D.
3. **Fix the seed and take repeated draws, three or five per cell, reported as a mean with a within-query spread.** This is the change that would let the §III.E jump be tested rather than described.

4. **Abstract, sentence 3, and §V, sentence 2.** Report the fall as 69.4% to 43.5% (Table I, *Poisoned acc.* column, rows 0/3 and 3/3), or report both origins explicitly. The present "77.9% to 43.5%" contradicts §III.B's instruction to read against the residual.
5. Recompute Table III over attack levels only. For the number-swap row the correction is verifiable from §III.E: 10.2, 10.2, 26.5, 34.7 gives 20.4 as printed, and dropping 0/3 gives 23.8. Print the 0/3 cells separately as a noise floor row, and add the per-cell grid of Figure 2 as a numeric table, since the heatmap is unreadable as an image.
6. **§II.D, fooled rate.** Either divide by clean-correct runs, which is what the stated rationale requires and which turns 34.0% at 3/3 into 43.8%, or drop the words "isolates correct-to-incorrect flips from questions the model could not answer anyway". Also reconcile with Figure 5, where fooled and abstained are disjoint bands and here they are not.
7. Add the closed-book arm. Same 49 queries, no retrieved context. One number, a few minutes of compute, and it tells the reader whether poisoned retrieval is worse than no retrieval at all.
8. Add an irrelevant-context arm, the passage replaced by an unrelated true sentence. §III.H claims on-topic falsehood does more damage than off-topic distraction, a comparison currently made against [2] rather than a control inside this experiment.
9. **§II.E and §III.G.** State how many of the 49 queries are answered by the ten added passages, and report the fragility results of Figure 6 split by that distinction. Release the ten passages.
10. **Eq. 3 and §III.F.** Say whether abstentions are excluded before the proxy is computed; the reported direction of Figure 4(b) depends on it. Define W_d as the clean or the poisoned passage set for the poisoned arm. Then state in the abstract that the proxy cannot detect poison adoption, which §II.D already concedes, so that "the model rarely invents new falsehoods" is not read as measured.
11. Resample distinct query texts in §III.C, not the 49 slots, and re-report Table II and the §III.E intervals. Say how much they widen.
12. Report recall@3 against FEVER gold evidence in §III.A, so the clean-arm error mass is attributed between retrieval and generation.
13. **§IV, paragraph 2.** Drop or requalify the advice about raising k . The sweep holds $k = 3$ throughout, and a claim about redundancy needs at least one more value of k .
14. Related work. §III.H positions the paper against [1], [2], [6] and [7], none of them attack papers, leaving the reader to infer that corpus poisoning and retrieval injection against RAG are unstudied. They are not. A paragraph placing this measurement against that literature, and naming what is

new here (a small quantized local model, a factorial level sweep, the abstention finding), would strengthen the contribution rather than weaken it.

15. Publish the testbed: code, the 49 queries with keywords, the 10 passages, the index build script, the raw 588-run table, the corruption functions. SI calls it reproducible; nothing in the paper currently lets a reader rebuild it.

5. Recommendation

Major revisions.

The experiment is small, honestly reported and arithmetically clean. Every percentage in Tables I, II and III is an attainable integer count over its stated denominator, the two tables reconcile to within rounding on all four metrics, and the §III.E cells reproduce the Table III number-swap row exactly. That consistency is not common, and disclosing the 0/3 residual rather than hiding it is the mark of someone reporting what happened rather than what they hoped for.

What keeps this from a lighter verdict is that two of the three headline claims do not rest on the evidence offered. The strategy ranking is confounded with a scorer that reads a negated falsehood as correct and an entity substitution as incorrect, which is the ranking observed. The abstention-over-fabrication claim rests on a proxy blind to poison adoption by construction, compared against a fooled rate that overlaps with abstention under the paper's own definition. The third claim survives: degradation with poisoning level is monotone across four rows of Table I, backed by the intervals in Table II, and large relative to the residual at 2/3 and 3/3. That is a real result and worth publishing.

Most of what I am asking for is measurement discipline, not new science, and several items cost a few hours of CPU: lock the seed, add a closed-book arm, add an irrelevant-context arm, re-score with polarity awareness, release the testbed. None of it needs a bigger model or a bigger machine. That matters, because the constraint this paper works under (8 GB, CPU, one person) is also the setting where its findings are most useful. I would be glad to read the revision.

Declarations

Potential competing interests: No potential competing interests to declare.