

Peer Review

Review of: "SAS Multimetric Tribunal: Structural Classification of Hallucination in AI-Generated Text"

Evgenii Arsentev¹

1. Independent researcher

This is a short follow-up to my review of the previous version (10.32388/7Z13RM). I read version 3 against my list from that review and only report what changed and what is still open. A full re-read I would leave for the next version.

What changed, and changed for the better

The biggest item is done. The real negative control now exists: 127 paraphrase pairs (100 synonym replacements, 27 back-translations), with 126 false positives, i.e., 99.2% (§2.5.2, §4.5 Table 9 row R3, §4.7.2). That is an uncomfortable number, and the author put it in the abstract instead of hiding it in an appendix. I respect that.

The sampling is now documented. Section 3.1 states a stratified sample of 600 pairs per suite with seed 42, and Table 4 adds a second run with seed 123. The forensic audit in §3.1.1 is honest work too. It shows that seed 123 reproduces the v2 figures exactly (Table 4: F1 99.02%, TP 1,765, FN 35), while the primary seed 42 gives F1 98.99% and FN 36. It also withdraws three derived quantities from v2 that did not survive recomputation: the zone distribution 78.9/19.2/1.9, the firing counts that were extrapolated from 600 pairs, and ISI separation 0.8323 (now 0.8140 in Table 2). Most authors would quietly swap the numbers. This one explains them.

Smaller points from my list are also fixed. Table 8 now lists all seven modules, including arithmetic_penalty and reference_penalty at 0 firings. The note under Table 1 says the weight column does not enter the multiplicative penalty. The STG rates now add up: 66.2% of 600 dialogue pairs is about 397 firings, and 6.5% of 600 TruthfulQA pairs is 39, both inside the 617 global firings of Table 8. And the

new Known Limitations section (§4.7, Table 10) is a good addition in itself, especially items 3.4 (legal corpus pairing) and 3.5 (unsorted glob).

What is still open

1. The abstract still leads with "Precision = 100.00%", and §5 still calls the tribunal "a high-precision structural audit layer". After R3, this reads wrong. I would put the A→C result next to Table 2 as its own row (precision on paraphrases) and drop "high-precision" from the Conclusions.
2. The two comparative claims are still untested. The abstract says "We demonstrate that cascading multiplicative penalty preserves per-module signal integrity better than weighted averaging", but §4.3 gives an argument, not a run. Same for STG: §4.2 says that without STG "these cases would be misclassified", and §5 says it "successfully closes the entity-mutation gap". Either run both ablations on the same 1,800 pairs or word them as hypotheses. For this version, the wording change is enough.
3. §2.6 still sweeps κR from 0.15 upward, so the chosen value sits on the lower edge of the grid. And 1.474 bits is still not reconciled. By my arithmetic, the Table 6 distribution (50.4/47.6/2.0) gives about 1.12 bits, and the v2 distribution that §3.1.1 now traces to an unrelated κR sweep on the earlier 600-pair set (78.9/19.2/1.9) gives about 0.84 bits. Please say which zone distribution gives 1.474, and revisit the sentence that the 1,800-pair validation "preserves the same zone distribution".
4. §3.4.3 vs Table 7. Table 7 and §3.4.1 say 100% of TruthfulQA TPs fired at least one specialized module. §3.4.3 says 348 of them (61.7%, which is 348/564) have modules = [lexical_baseline_score] only. Both can't hold unless lexical_baseline_score counts as specialized, and §3.4.1 treats it as the non-specialized baseline.
5. §4.4, third bullet: a factual-QA threshold of 0.50 would be "more aggressive", but the FNs have ISI of at least 0.5625 (§3.4.1), and pairs are detected below κD . Lowering κD to 0.50 detects fewer, not more. To catch these FNs, the boundary would have to go up.
6. §2.1 still says ISI is "designed to be invariant to paraphrase, translation, and syntactic variation". R3 shows the opposite (99.2% FP on paraphrases). §2.5.2 also keeps the old "estimated 90–95%" bullet right above the measured result. Both sentences can simply go.
7. Minor, and fine to leave for v4: intervals for recall in Tables 2 and 5 (with 36 FN out of 600 TruthfulQA pairs, they matter), runtime per pair and hardware for the "lightweight" claim, and a commit hash or tag next to the GitHub link.

Recommendation

Minor revisions. Version 3 answers the main concern of my first review: precision is now measured, and the result is reported honestly even though it is bad for the method. What is left is mostly wording that still claims more than the data show (items 1, 2, and 6) plus two number checks (items 3 and 4) and one sign error (item 5). I'll be glad to do a full re-read of the version with the semantic-equivalence guard and the two ablations.

Declarations

Potential competing interests: No potential competing interests to declare.