

Peer Review

Review of: "SAS Multimetric Tribunal: What a Hallucination Benchmark Measures, and What It Does Not A Format-Controlled Audit of Structural Classification"

Evgenii Arsentev¹

1. Independent researcher

1. *What the work claims*

This is my fourth round. I read version 5 in full against my reviews of versions 2, 3, and 4 (DOIs 10.32388/7Z13RM, 10.32388/9T9P58, 10.32388/6HK3TA).

Version 5 recasts the paper as an audit of its benchmark; the contribution is the audit procedure (six steps, §2), not a detector (Abstract; §1.4). The new control N1 takes the same 1,800 IDs with verbatim correct answers and no generated text (§5.2). The tribunal flags 1,569 of 1,800 of them presented alone (87.2%), so the headline F1 = 98.99% reflects format asymmetry, not hallucination detection (§5.4). Under per-scheme thresholds, the 8797-point gap between multiplicative and averaged combination disappears: F1 0.687 vs 0.695, AUROC 0.514 vs 0.632 (§7, Table 8). With the same guard, the tribunal and Jaccard baseline reach F1 0.720 vs 0.722 (§8, Table 9). Control R3 from v4 is withdrawn: in 125 of 127 pairs, the paraphrase was a label string or the question text (§5.1; Appendix A). Four v4 claims are retired in total (§9.3). That is the claim as I understand it.

2. *Whether the conclusion holds on the data*

I value that the author retires four of his own v4 claims by direct measurement, with the same protocol (§9.3). That is rare, and I want to say it first. I also like N1 because, as §11 says, it needs no generated text, no annotator, and no GPU.

I went through my v4 points one by one, and version 5 answers each. For the averaged score, Table 8 (§7) gives AUROC 0.514 for the multiplicative scheme and 0.632 for the averaged one, so the gap reverses instead of narrowing. My point that the §4.2 extension claim was a hypothesis is answered: it is now tested in §8 (Table 9) and retired. For $\kappa_R = 0.15$, §3.6 now says the median ISI of detected pairs is 0.1424 and that 0.15 is a property of this sample. My question about R3 and Table 11 of v4 is moot since R3 is withdrawn (§5.1).

I also rechecked the arithmetic. Table 6 adds up, and all four cells sum to 1,800. Table 9 reproduces for both Jaccard + guard and Tribunal + guard. I found one small slip: in Table 8, weighted average row (TP 1,759, FP 1,511, so FN 41), $F1 = 3,518 / 5,070 = 0.694$, but the table prints 0.695. I would correct it, though it changes nothing in the argument. So the conclusions in §11 hold on the data as reported, with the reservations in section 4.

3. Reproducibility

In this round, a lot of what I asked for is now in place. §10.2 gives commit cdd2e12 and tag v5.0.0, plus a SHA-256 manifest for four files (n1_control.py, pairs_conditions_1800_seed42.json, tribunal_multimetrico.py, validacion_1800_stratified_v3.py). §10.1 states the hardware (AMD Ryzen 5 5600X, 16 GB RAM, Python 3.13, single-threaded) and the runtime, 36 seconds for 7,200 evaluations, about 5 ms per pair. §10.3 makes clear that the tribunal uses no LLM and no trained parameters, and that sampling is deterministic under sorted(glob) and a fixed seed. The artifacts are listed in Data Availability: SAS_v5_Supplements.zip, Zenodo 10.5281/zenodo.19702379, and a GitHub repository. I did not run the code or open the archive, so this assessment is of the text on Qeios only.

One thing I still could not find. §3.5 says that A is context plus the correct answer and B is the hallucinated answer alone, but it does not say which dataset field is used for each in HaluEval-Dialogue, HaluEval-QA, and TruthfulQA. Since R3 failed exactly on a field-parsing error (§5.1), I would list the field names per suite.

4. What specifically to fix

1. Table 8 (§7) is titled "per-scheme calibration," but the multiplicative row stays at the fixed $\kappa_D = 0.56$ while the weighted average gets an optimized 0.83. AUROC is not affected, but for the F1 column, I would also optimize the multiplicative threshold or label the rows "fixed" vs "optimized."
2. In Table 8, the weighted average F1 0.695 should read 0.694 (3,518 / 5,070).

3. In §5.4, the 87.2% flag rate on correct answers is called "nearly identical" to the 98.0% detection rate. They are 10.8 points apart, so "close," or both Wilson intervals, would be more exact.
4. §9.1 says "The published F1 reflects response length." Length is not measured directly anywhere in the paper, and §6 itself says "most plausibly LSA dispersion correlating with text length." I would add one measurement, for example, ISI against the length ratio of response to source, or phrase §9.1 as the likely explanation.
5. In Table 7, if `cre_isi` responds only to the source, a reader will ask why it fires on 971 pairs (A→A) and 1,081 (A→A") but on 132 (A→B) and 68 (A→A'). One sentence on what gates the firing (masking by other modules?) would help.
6. In §9.5, as rendered on Qeios, the sentence "No limitation below changes $\kappa D = 0.56$ or $\kappa R = 0.15$ " comes after Table 11, so "below" should be "above." I would also add one line on what now supports $\kappa D = 0.56$, since §11 keeps it while the benchmark it was tuned on is shown to measure format.
7. On the license, the Qeios page header shows CC BY, while "License and Attribution" says the manuscript is CC BY-NC 4.0. It would help to align one of the two with the other.

5. Recommendation

Minor revisions. I am glad to see that every point of my version 4 review is answered, and two of them by new runs (Table 8, Table 9) that went against the author's own earlier claims. What remains is small: one asymmetric threshold in Table 8, a length explanation stated as fact in §9.1, and the license mismatch. I would also tidy the field names in §3.5. I look forward to R3.2 (§9.4, Table 10).

Declarations

Potential competing interests: No potential competing interests to declare.