

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 fifth round; I read version 6 in full against my version 5 review.

Version 6 keeps the contribution as the audit procedure (six steps, §2), not a detector (§1.4, §1.5). The headline F1 = 98.99% on the 1,800-pair benchmark is kept as reported, and the format-controlled negative N1 shows that it measures format asymmetry: 1,569 of 1,800 verbatim correct answers presented alone are flagged (87.2%) (§5.4). New in this version is a differential guard called R3.1 (§8). On N1 it corrects 1,133 of the 1,569 false positives, and F1 goes from 0.6873 to 0.8330 (§8.2, Table 11). On FinQA (1,657 samples) it detects 208 of 239 hallucinated answers, 87.0% (§8.3, Table 12). That is the claim.

2. Whether the conclusion holds on the data

What I value most is how fully my version 5 review was answered. I raised eight points: seven numbered items plus the field names. Version 6 answers all eight: six are fixed cleanly, and two stay partly open, the weighted-average row of Table 9 and cre_isi (see section 4). Field names per suite are now in Table 2 (§3.5). Table 9 (§7) adds a multiplicative row with an optimized threshold, 0.550 (F1 0.688). §5.4 calls the two rates close, 10.8 points apart, with Wilson intervals. §9.1 now measures length: Pearson $r = +0.5996$ ($A \rightarrow B$) and $+0.8696$ ($A \rightarrow A'$) between ISI_FINAL and the response-to-source length ratio. §9.5 says "No limitation above" and explains why $\kappa_D = 0.56$ survives. The licence is now CC BY 4.0, as in the Qeios header.

I rechecked the arithmetic, and it holds: both Wilson intervals in §5.4 (from 1,569/1,800 and 1,764/1,800), Table 11 ($TP = 1,800 - 204 = 1,596$, precision $1,596 / 2,032 = 0.7854$, $F1\ 3,192 / 3,832 = 0.8330$), and the optimized multiplicative $F1$ in Table 9 ($3,524 / 5,123 = 0.688$).

The new guard result is weaker than its wording. The threshold 0.95 was chosen by a sweep over the N1 audit data (§8.1), and $F1\ 0.833$ is reported on the same N1 set (Table 11), with no held-out split. Also, A' is a substring of A by construction, and the guard's grounding counts content words of the response present in the source, so A' is expected to reach grounding near 1.0, as §8.2 itself reports. The Abstract and limitation 3.14 (Table 14) say the guard "is not a constructional artifact", while limitation 3.15 (Table 14) says the substring construction "biases the guard comparison" and is Open.

The format conclusion of §11 holds. The R3.1 numbers should be read as in-sample.

3. Reproducibility

§10.2 now gives the tag v6.0.0 and a SHA-256 manifest that adds the guard, the sweep script, the FinQA and MedHallu conversions, and the FinQA evaluation. Artifacts are in SAS_v6_Supplements.zip and, per Data Availability, on Zenodo and GitHub. I did not run the code or open the archive; this is a reading of the text on Qeios.

Two small gaps. Version 5 gave the commit hash cdd2e12 with its tag, while version 6 gives only the tag v6.0.0 (§10.2), so I would add the commit hash again. The file counts also disagree: §10.2 lists 12 hashes and says "Full manifest with all 18 files", while Data Availability says the supplement has 26 files and 28 manifest entries.

4. What specifically to fix

1. Threshold 0.95 for the guard was swept on N1 (§8.1), and $F1\ 0.833$ is reported on the same N1 (Table 11). I would split N1 by ID into a tuning half and a test half, and call 0.833 an in-sample figure in the Abstract and §11.
2. In Table 10, row $A \rightarrow A$ ", Equiv 362 (20.1%) and Abstain 1,800 (100%) cannot both hold in one row of 1,800 pairs. One cell needs correcting.
3. §8.2 says the 168 regressions on $A \rightarrow B$ "are responses where the guard abstains by compression ratio". Table 10 shows only 61 abstentions on $A \rightarrow B$, fewer than 168, while 204 pairs are marked Equiv. Could the 168 be pairs the guard calls equivalent rather than abstentions? If so, I would correct that sentence, and the same wording in §11.

4. For FinQA (§8.3, Table 12), detection 87.0% is measured cross-query, a hallucination against the source of a random different sample. Since grounding is the share of response content words found in the source (§8.1), would any unrelated text not score low? On correct answers, the guard abstains on 1,328 of 1,418 (93.6%). Table 12 has no regression column, so "zero regressions" is not defined there. I would define it and present FinQA as a sanity check, not as generalization.
5. §1.3 still says "four quantitative claims" while §1.5 and §9.3 say five. §9.2 points to "Table 7" for *cre_isi* firing, which is now Table 8. §6 and §9.2 say "The tribunal has no content-sensitive module", while §9.3 retires that claim as superseded by R3.1. One sentence on whether R3.1 is part of the tribunal would settle it.
6. In Table 9, the optimized weighted-average row now reads TP 1,735 and FP 1,461 (v5: 1,759 and 1,511). With FN = 65, $F1 = 3,470 / 4,996 = 0.695$, not the printed 0.694. Which counts are current?
7. Remaining from my v5 point on *cre_isi*: §9.2 now explains the firing gate, thank you. Yet it says short responses ($A \rightarrow B$, $A \rightarrow A'$) rarely reach the dispersion threshold, so firing depends on the response too, not only on the source as §6 says. Also, MedHallu (10,000 pairs) is listed as a contribution (§1.5) without any result. I would report one, or move it to the roadmap.

5. Recommendation

Minor revisions. All eight points of my version 5 review are answered, six of them fully, and the format conclusion holds on the data. What remains is how R3.1 is described: the in-sample threshold (§8.1), two inconsistencies (Table 10, §8.2), one row of Table 9, and the FinQA framing (§8.3). All can be fixed in text plus one held-out split. I look forward to R3.2 (Table 13), the control with paraphrased correct answers.

Declarations

Potential competing interests: No potential competing interests to declare.