

Peer Review

Review of: "SAS Multimetric Tribunal: Structural Classification of Hallucination in AI-Generated Text – What a Benchmark Design Can and Cannot Show"

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

This is my third round on this paper. I read version 4 in full, against my reviews of version 2 (10.32388/7Z13RM) and version 3 (10.32388/9T9P58). The paper has been reframed, and I think rightly. The Abstract and §1.5 say the contribution is not a detection algorithm but a way to audit what a benchmark design can and cannot show, applied to the SAS Multimetric Tribunal (seven structural modules, no GPU, multiplicative penalty, $\kappa D = 0.56$, $\kappa R = 0.15$). What the author still claims (§4.2) is the per-module architecture as a place where a semantic-equivalence guard can be attached and measured.

2. Whether the conclusion holds on the data

First, what I value here. Table 4 (§2.5.2) prints four trivial baselines next to the tribunal and shows exact-match at $F1 = 100.00\%$ against the tribunal's 98.99% . Table 11 (§4.1) shows the tribunal flagging 126 of 127 valid paraphrases. And the author reports that his own SourceTargetGuard accounts for only 3 of 1,800 true positives (§2.4, Table 3). Few authors print numbers like these about their own method.

Now my seven open points from the version 3 review. Point 1, the precision headline: closed. Table 6 has a row "Specificity (R3 real negative control) 0.8%", and "high-precision" is gone from §7. Point 2, the two untested comparisons: closed, both are runs on the same 1,800 pairs. Table 2 (§2.3) gives $F1\ 98.99\%$ for the multiplicative scheme against 11.02% for the weighted average. Table 3 (§2.4) shows recall going from 98.00% to 97.83% with SourceTargetGuard disabled. Point 3, the κR grid: closed. Table 5 (§2.6) goes down

to 0.02 and gives 1.1209 bits at $\kappa R = 0.15$; the old 1.474 is traced, as its likely source, to a superseded run with two of seven modules loaded. Point 4, the stale 348: closed. Table 13 (§5.3) gives 401 of 453 TruthfulQA Zone-B true positives detected by lexical_baseline_score alone. Point 5, the sign error: closed in §6.4. Point 6: closed, §2.1 says "We do not claim invariance to paraphrase or translation". Point 7 is half done: Wilson intervals are in Tables 6 and 9; runtime, hardware, and a commit hash are not.

I rechecked the sums. Table 2: $3,528 / 3,564 = 98.99\%$ and $210 / 1,905 = 11.02\%$. Table 13: $401 / 453 = 88.5\%$. Table 14: the firings sum to 2,973, as §5.5 says. All reproduce. So the conclusions as worded in §7 hold on the data, with two reservations (items 1 and 2 below).

3. Reproducibility

The tribunal uses no LLM, so model snapshots and decoding parameters do not apply. Seeds are stated (42 and 123, Table 8). Data Availability now lists a raw result file for each new table, including per-pair ISI values for Table 5 and the 107 zone-change pair IDs for Table 3. I did not open the files; this review is of the text on Qeios.

What I still could not find in the text: a commit hash or release tag, runtime per pair, and hardware, and which dataset field is source A and which is hallucinated B in each suite (§2.5.1). §4.1 says the 127 R3 pairs are 100 synonym replacements and 27 back-translations, but not which tool made them, nor how it was checked that each pair keeps the meaning. §5.4 describes four figure panels and points to the Supplementary Material; no figure is in the article body as rendered on Qeios.

4. What specifically to fix

1. §2.3, Table 2. The weighted average is classified under the same $\kappa D = 0.56$ that the multiplicative score uses. A module that does not fire returns 1.0 (§2.2), so the average sits close to 1 by construction, and a threshold of 0.56 misses almost everything. Identical pairs score 1.0000 (Table 6), so a much higher threshold would still give zero false positives. I would fit a threshold for the averaged score or report AUROC for both schemes, and then say how much of the 8797-point gap remains. Until then, §6.3 and §7 should say "at a fixed κD ".
2. §1.4 and §4.2. The central claim is now that a single-metric baseline "has no comparable point of extension". This is an argument. The embedding guard planned as R3.1 (Table 15) could sit in front of the Jaccard < 0.56 rule just as easily. The run that tests the claim is guard plus tribunal against guard plus Jaccard, on the same pairs. Until then, it is a hypothesis.

3. §2.5.2 says "Every trivial baseline in Table 4 matches or exceeds the tribunal's F1". One row does not: ROUGE-L < 0.50 has 97.11% against 98.99%.
4. Table 15, row R3, says the tribunal's false-positive rate is "matched or exceeded by every trivial baseline tested". In Table 11, ROUGE-L < 0.50 has 123 false positives against the tribunal's 126. I would add a Wilson interval to Table 11 and call the methods indistinguishable on R3.
5. §2.6 says "Item 3.9 (§6.6) is updated accordingly". In Table 16, the κ R grid item is 3.10; 3.9 is corpus coverage.
6. §2.6 already says that Zone A is fixed at 2.0% (Table 5) and that entropy peaks when the other 98% splits evenly between B and F-S. I would add the consequence in one line: 0.15 is then simply the median ISI of the detected pairs in this sample. Otherwise, readers will take it as a property of the method.
7. §6.6 opens with "No limitation below affects the validity of κ D = 0.56", while limitation 3.2 in the same Table 16 reports specificity of 0.8% on valid paraphrases. I would soften that sentence.
8. Carried over: runtime per pair and hardware behind "GPU-free" (Abstract) and "lightweight" (§6.2), a commit hash or tag in Data Availability, the A and B field mapping in §2.5.1, and one line on what the Durante Invariance License v1.0 permits.

5. Recommendation

Minor revisions. All six main points of my version 3 review are closed, the two comparisons and the κ R grid with new runs, and the interval request from the seventh is met. What is left is small: one comparison made at a threshold set for the other scheme (Table 2), a central claim not yet measured (§4.2), two sentences that overstate Tables 4 and 11, a wrong cross-reference, and the reproducibility details. Only the threshold for the averaged score needs a new run. I look forward to the version with the semantic-equivalence guard R3.1.

Declarations

Potential competing interests: No potential competing interests to declare.