

Peer Review

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

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

The paper presents the SAS Multimetric Tribunal, a rule-based detector that scores a (source, response) pair with seven modules (Table 1: lexical_baseline_score, source_target_guard, flow_penalty, negation_penalty, cre_isi, arithmetic_penalty, reference_penalty) and combines them into an Invariant Similarity Index. The combination is a minimum over three core modules (Eq. 1) multiplied by the penalty factors of whichever other modules fire (Eq. 2). Two thresholds split the ISI axis into three zones: Coherent (A) at $ISI \geq \kappa_D = 0.56$, Recoverable Rupture (B) between the thresholds, and Structural Collapse (F-S) below $\kappa_R = 0.15$. The new module in this version is SourceTargetGuard (§2.4), a regex comparison of locations, years, names, and quantities with fixed decay factors (0.70, 0.65, 0.80, 0.75) and a floor of 0.30.

On 600 pairs each from HaluEval-Dialogue, HaluEval-QA, and TruthfulQA, every source is scored against itself (A→A) and against a hallucinated variant (A→B). The reported result is F1 = 99.02%, Precision = 100.00%, Recall = 98.06%, with 35 false negatives (Table 2), 34 of them in TruthfulQA (Table 4). The author says openly that these figures are an upper bound under a tautological check and presents the 35 misses as a documented failure mode of the lexical baseline (§3.4). The abstract also states that the multiplicative penalty "preserves per-module signal integrity better than weighted averaging," and the Conclusions call SourceTargetGuard a module that "successfully closes the entity-mutation gap".

2. Whether the conclusion holds on the data

I want to start with what this revision gets right because it is more than most. The author removed an invalid comparison table from v1, labels A→A as a sanity check rather than a negative control (§2.5.1),

states in §2.5.2 that Precision = 100% is an artifact of the design, and publishes a false-negative registry. The confusion matrix is also internally consistent. I checked it: Recall = $1,765/1,800 = 98.06\%$; F1 = $2 \times 1,765 / (2 \times 1,765 + 35) = 3,530/3,565 = 99.02\%$; Accuracy = $(1,765 + 1,800)/3,600 = 99.03\%$ (Table 3 against Table 2). Per-suite recall in Table 4 also reproduces ($566/600 = 94.3\%$ for TruthfulQA). That is a solid base to build on. The problems are in what the numbers can mean and in several figures that do not reconcile with each other.

The benchmark has no informative negative class, so the headline metrics cannot rank this method against anything. The paper concedes that $A \rightarrow A$ guarantees zero false positives. It does not draw the next conclusion: a detector that flags every pair where the response is not character-identical to the source would score TP = 1,800, FP = 0, F1 = 100% on this benchmark. The tribunal scores 99.02% (Table 2), which is below that trivial rule. So the only quantity that carries information here is recall on $A \rightarrow B$, and recall on its own says nothing about specificity. I would not call the reported numbers an upper bound on real performance; they are uninformative about precision altogether. The phrase in §5, "validating its suitability as a high-precision structural audit layer," is not supported by any measurement in the paper.

What exactly is the "source"? I could not determine this from §2.5.1. HaluEval-QA provides knowledge, a question, a correct answer, and a hallucinated answer; TruthfulQA provides correct and incorrect answers per question. If A is the correct answer and B the hallucinated one, then the tribunal is a reference-based comparison against gold text, which is the category §1.2 criticizes for needing gold references. If A is the knowledge passage or dialogue history, the setup is different, and the high Jaccard overlap noted in §3.4.2 has another meaning. The method section needs to say which dataset field is A and which is B for each suite, and how the 600 items were drawn (TruthfulQA has 817 questions, HaluEval-Dialogue many thousands).

The two comparative claims have no experiment behind them. The abstract says "We demonstrate that cascading multiplicative penalty preserves per-module signal integrity better than weighted averaging", and §4.3 is titled "Why Multiplicative Penalty Outperforms Averaging". I found no run in which the same modules are combined by averaging and scored on the same pairs. §4.3 is an argument, not a result. The same holds for §4.2: "Without STG, these cases would be misclassified as Coherent or borderline Rupture." There is no ablation with STG removed. Both are cheap to run with the existing code, and until they are run, the claims should be stated as hypotheses.

The κ_R selection sits at the edge of the grid, and the entropy value does not match Table 5. §2.6 sweeps κ_R over {0.15, ..., 0.40} and picks 0.15 by maximum entropy, 1.474 bits. Two things bother me. First, 0.15 is the lowest value in the sweep. Raising κ_R can only move pairs from Zone B into Zone F-S, and F-S already dominates (78.9% in Table 5, column A→B), so entropy must fall as κ_R rises. A maximum at the boundary of the grid means the true maximizer lies below 0.15 and was not searched; "entropy maximization" here is equivalent to "smallest value tried". Second, I recomputed the entropy of the Table 5 hallucination column (1.9%, 19.2%, 78.9%): $-\sum p \log_2 p = 0.109 + 0.457 + 0.270 = 0.836$ bits, not 1.474. Pooling both conditions over 3,600 pairs (A: (1,800 + 35)/3,600 = 0.510; B: 0.096; F-S: 0.395) gives 1.35 bits, also not 1.474. I cannot find a distribution in the paper that produces 1.474, so either the entropy was computed on data not shown, or one of the two numbers is wrong. The "critical observation" in the same section is also stronger than stated: F1 does not merely stay "stable" across the sweep, it cannot depend on κ_R at all, since detection is defined as $ISI < \kappa_D$. And the stable value is given as "~99.2%" there, while Table 2 reports 99.02%.

Counts in §3.4 do not reconcile with the global tables. I found three.

- §3.4.3 says 348 TruthfulQA true positives sit in Zone B (348/566 = 61.5%, which checks). But Table 5 puts 19.2% of all 1,800 hallucination pairs in Zone B, that is $0.192 \times 1,800 \approx 346$ pairs across all three suites. TruthfulQA alone cannot contribute 348 of at most about 346.
- §3.4.2 says source_target_guard fires on "6.5% vs. 66.2% in dialogue". With 600 dialogue pairs, 66.2% is about 397 firings. Table 7 gives 294 STG firings globally. The per-suite figure exceeds the total.
- §3.5 says "Lexical baseline and CRE together account for 77.5% of all firings." 77.5% is the sum of their Table 7 "% of Pairs" values (49.0 + 28.5). As a share of firings, it is $(882 + 513)/(882 + 513 + 369 + 294 + 138) = 1,395/2,196 = 63.5\%$.

Perhaps some of these come from the original 600-pair subset, like the starred ISI rows of Table 2. If so, the denominators must be stated. As printed, they contradict each other.

Table 6 is close to circular. §3.4.1 says "100% of TruthfulQA TPs fired at least one specialized module", and Table 6 shows 0% of TPs with modules = []. Yet §3.4.3 says 348 of those TPs have modules = [lexical_baseline_score] only. So Table 6 counts the lexical baseline as a firing, and §3.4.1 describes it as a specialised module, while the abstract says these cases have lexical_baseline_score as "the sole discriminator". The consistent reading seems to be: a pair is detected when the lexical baseline or something else fires, and missed when nothing fires. That separates TPs from FNs by construction. It is a

description of the decision rule, not evidence about the failure mode. The same applies to the 0.5625 vs 0.5500 gap in §3.4.1: TPs are below κ_D and FNs above it by definition, so the gap tells us only that no pair fell in [0.55, 0.5625).

The proposed fix for TruthfulQA points the wrong way. §4.4 suggests that factual QA "may require a more aggressive threshold (e.g., 0.50)". Detection is $ISI < \kappa_D$. Lowering κ_D to 0.50 would move any TP with ISI in [0.50, 0.55] into the missed set and recover none of the 35 FNs, whose ISI starts at 0.5625. Catching them needs κ_D above 0.5625, which would then have to be checked against a real negative control.

Smaller points on the evidence. §2.5.2 estimates that Precision would drop "to an estimated 90–95%" under a paraphrase control; there is no data behind that range, and I would remove it. §2.1 describes ISI as "invariant to paraphrase, translation, and syntactic variation", but ISI_{HARD} takes a minimum that includes Jaccard overlap (Eq. 1, Table 1 row 1), which is not invariant to any of the three. Table 1 lists weights summing to 1.00 (0.35 + 0.25 + 0.12 + 0.08 + 0.12 + 0.04 + 0.04), yet neither Eq. 1 nor Eq. 2 uses weights. Table 7 lists five modules; arithmetic_penalty and reference_penalty are absent, presumably with zero firings, which should be said. No confidence intervals are given anywhere. With 35 misses among 1,800 positives, an interval on recall is easy to add, and per suite it matters more (Table 4, TruthfulQA, 34 of 600). Finally, "lightweight" and "zero GPU dependency" are asserted, but no runtime, latency, or memory figure is reported.

Taken together: the confusion matrix is arithmetically correct, and the failure-mode registry is a useful artifact. But the design cannot measure precision, the two comparative claims were not tested, and at least four reported numbers (the entropy, the Zone B count, the STG dialogue rate, the 77.5% share) conflict with other numbers in the same paper.

3. Reproducibility

The tribunal itself uses no LLM, so the usual complaint about missing model snapshots and decoding parameters does not apply to the detector. It does apply indirectly: the hallucinated variants come from the benchmarks, and their provenance should be cited by dataset version. On the positive side, the paper lists a Zenodo DOI (10.5281/zenodo.19702379), a GitHub repository, a public API, and names four supplementary files, including `tribunal_multimettrico.py`, `validacion_kr_v3.json`, and `sas_fn_analysis_1800.csv`. That is more than many submissions provide.

What I could not find or verify from the text:

- **Supplementary files on Qeios.** The Qeios page, as rendered, attaches only `sas-failuremode-analysis-r2-1-b-en.md` and `sas-fn-analysis-1800.csv`. The Python file, the JSON results, and the figure listed under Data Availability are not attached here. I did not check Zenodo or GitHub, since this review is of the Qeios version.
- **Figure 1.** As rendered on Qeios, Figure 1 is not legible; only the caption line appears. Its four panels are described in §3.4.4, but I could not inspect them.
- **Code version.** No commit hash, tag, or release is given for the GitHub implementation, and the public API can change without notice.
- **Data construction.** No dataset revisions, no field mapping for A and B, no sampling procedure, or seed for the 600-item subsets.
- **Module definitions.** Only STG is specified in the paper (§2.4). `flow_penalty`, `negation_penalty`, `arithmetic_penalty`, and `reference_penalty` have one-line descriptions in Table 1. `cre_isi` ("Ricci semantic curvature", with LSA embeddings per §4.6) has no formula, no LSA dimensionality, and no fitting corpus. A reader cannot reimplement the tribunal from the text.
- **Firing criteria.** What counts as the lexical baseline "firing" (Table 7, 882 firings) is never defined, although Table 6 and §3.4 depend on it.
- **Compute.** No hardware, runtime, or throughput figures.
- **Licence.** Everything is under the "Durante Invariance License v1.0". Its terms are not summarized, so I cannot tell whether independent reproduction and redistribution of results are permitted. A standard license, or a one-line summary of what this one allows, would remove the doubt.

A last note on the threshold. $\kappa_D = 0.56$ is presented as "registered TAD EX-2026-18792778" and "confirmed across the R0–R2.1-b research line". A registration records authorship; it does not validate a threshold. The cited confirmations [5]–[10] are all by the same author and all on Zenodo. That is legitimate history, but the paper should state how κ_D was originally fitted and on which data, so readers can judge whether it was tuned on material overlapping HaluEval or TruthfulQA.

4. What specifically to fix

1. **§2.5 and all of §3: add a real negative class now, not in R3.** Correct answers from TruthfulQA and the `right_answer` field of HaluEval-QA already exist, so an $A \rightarrow C$ condition needs no new annotation. Without it, Precision in Table 2 carries no information.

2. Report trivial and standard baselines on the same pairs (Tables 2–4): exact-match inequality, Jaccard alone thresholded at 0.56, and at least one off-the-shelf metric such as ROUGE-L or an NLI model. The exact-match rule reaches $F1 = 100\%$ on the current $A \rightarrow A/A \rightarrow B$ design, and the paper should say so.
3. §2.5.1: state, per suite, which field is A and which is B, the dataset versions, and how 600 items were sampled (with seed). Then revisit §1.2: if A is a gold answer, the method is reference-based.
4. Abstract, §4.3: either run the averaging comparison (same modules, weighted mean with the Table 1 weights, same 3,600 pairs) or replace "We demonstrate" with a hypothesis.
5. §4.2 and §5: run the tribunal with `source_target_guard` disabled and report how many of the 294 STG-firing pairs change zone. Only then can the Conclusions say STG "closes the entity-mutation gap".
6. §2.6: extend the κ_R grid below 0.15, since the chosen value is its lower edge; publish the full entropy curve; and reconcile 1.474 bits with Table 5, which gives 0.836 bits by my arithmetic. Also replace "~99.2%" with the value in Table 2 or explain the difference.
7. Reconcile §3.4.3 (348 TruthfulQA TPs in Zone B) with Table 5 (about 346 Zone B pairs in total).
8. The STG rate "66.2% in dialogue" (§3.4.2) implies about 397 firings, more than the 294 global firings in Table 7. Give a per-suite firing table with explicit denominators.
9. §3.5: the 77.5% is a sum of per-pair rates. The share of all firings is 63.5%. Fix the wording or the number.
10. Define "fires" for `lexical_baseline_score`, and resolve the conflict between §3.4.1 ("fired at least one specialized module"), Table 6, and §3.4.3 (modules = [`lexical_baseline_score`]). Then acknowledge that the TP/FN split in Table 6 follows from the decision rule.
11. §4.4 and Table 8, row R5: lowering κ_D to 0.50 reduces detection. Correct the direction, or drop the example.
12. Table 1: explain what the weight column does. If the weights are unused in Eq. 1–2, remove them; if they are used somewhere, write the equation.
13. Table 7 should list all seven modules, including those with zero firings.
14. Give formulas for `cre_isi`, `flow_penalty`, and `negation_penalty` in §2, or at minimum in an appendix; point to the exact function in a tagged release.
15. Add Wilson or bootstrap intervals to recall in Tables 2 and 4.
16. §2.5.2: delete the unsupported 90–95% precision estimate.

17. §2.1: drop "invariant to paraphrase, translation," or show it empirically. Jaccard is in the minimum.
18. Report runtime per pair and hardware, since "lightweight" is a headline claim.
19. Attach the Python file, the JSON results, and Figure 1 to the Qeios record, and give a commit hash.
20. §5, final paragraph: sentences such as "This transforms an inflated claim into a self-correcting, peer-review-ready contribution" judge the paper rather than report on it. Let readers decide; the space is better used for the limitation that precision is unmeasured.

5. Recommendation

Major revisions. The author has done honest work between versions, the confusion matrix checks out, and the failure-mode registry is a useful thing to publish. But the benchmark cannot measure precision, a trivial exact-match rule outperforms the method on it, the two comparative claims (multiplicative vs averaging, STG contribution) were never tested, and several numbers in §2.6 and §3.4–3.5 conflict with Tables 5 and 7. All of this can be fixed with data the author already has. I would be glad to read a version with a paraphrase control and the two ablations.

Declarations

Potential competing interests: No potential competing interests to declare.