

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 sixth round with this work. I read version 7 in full against my review of version 6 and the author's point-by-point reply under it.

Version 7 still presents the audit procedure as its contribution, not a detector (§1.4, §1.5). The headline F1 = 98.99% on the 1,800-pair benchmark is explained by format asymmetry, which the format-controlled negative N1 shows (§5.4). The content-sensitive differential guard R3.1 is now an optional add-on, outside the seven-module core (§6). New in this version is a held-out check of the guard threshold: a 50/50 split of N1 by ID, stratified by suite, repeated over 100 random splits (§8.2). The guard raises F1 on N1 from 0.6873 to 0.8330 (Table 11). FinQA is now called a domain sanity check, not a generalization claim (§8.3, Abstract). MedHallu moves to the roadmap with evaluation pending (§9.4). That is the claim as I understand it.

2. *Whether the conclusion holds on the data*

Version 7 fixed the earlier points well, and I checked the numbers. Table 9 now gives F1 0.695 for the optimized weighted average, and §7 gives $\Delta F1 = -0.007$ (0.688 vs. 0.695). With TP 1,735, FP 1,461, FN 65, I get $3,470 / 4,996 = 0.695$, so it matches. §8.2 now splits the 168 regressions into 61 by compression-ratio abstention and 107 by false equivalence, and §11 uses the same wording. Table 10, row A→A", shows Equiv 0 with a dagger footnote (362 of the 1,800 abstentions have grounding ≥ 0.95 but are overridden by the

compression-ratio abstention). §1.3 now says "five quantitative claims," §9.2 points to Table 8 for cre_isi firing, and §6 states that R3.1 is an optional add-on.

Two questions remain. In the held-out check (§8.2 and §9.4 R3.1b), the tuning half ($n = 900$ IDs) has its optimal threshold at 0.97 ($F1 = 0.882$), but the held-out half is evaluated at 0.95, not 0.97 ($F1 = 0.833$, $\sigma = 0.0055$ over 100 splits). So the held-out number tests the full-set threshold 0.95, which was chosen on all 1,800 IDs, including the held-out half (§8.1: the sweep minimizes total error at 0.95, $n = 1,800$). For the sentence "The threshold is stable across splits," I would add the distribution of the threshold chosen in each of the 100 tuning halves, and the held-out F1 at that tuned threshold.

In Table 10, row A→B, Equiv is 204 (11.3%) and Abstain is 61 (3.4%), so 265 pairs are not flagged by the guard, but Regressed is $168 = 61 + 107$. That leaves $204 - 107 = 97$ false equivalences not counted as regressions. Could you say what they are, for example, pairs the tribunal had already missed? One sentence or a footnote would help.

3. Reproducibility

In round 5, I asked for a commit hash, and §10.2 now gives it: tag v7.0.0 together with commit e45f82b. Thank you for that. The file counts also agree now. §10.2 says "Full manifest with all 25 files," and Data Availability says the supplement has 25 files with 27 manifest entries, including the manifest files themselves.

One thing I would add: The SHA-256 list printed in §10.2 has 12 entries, and none of them is a script or a results file for the new held-out split (§8.2). It would help to put the split script and its per-split output (the 100 thresholds and F1 values) into the manifest.

I did not run the code or open the archive (SAS.v7.Supplements.zip, Zenodo 10.5281/zenodo.19702379, GitHub). This is a reading of the text on Qeios.

4. What specifically to fix

1. In §8.2 and §9.4 (R3.1b), I would report the threshold chosen on each of the 100 tuning halves and the held-out F1 at that threshold, next to the held-out F1 at 0.95 (0.833). The tuning-half optimum is 0.97 ($F1 = 0.882$), so "stable across splits" is not yet shown.

2. In Table 10, row A→A", Equiv is 0 and Abstain is 1,800 (100%), yet Corrected is 9. If every pair abstains, it is not clear how 9 are corrected. One line in the footnote would settle it.

3. In Table 10, row A→B, there are 204 Equiv versus 107 false-equivalence regressions (§8.2). It would help to explain the remaining 97.

4. For FinQA (§8.3, Table 12), on correct answers (n = 1,418), the guard flags 10 (0.7%) as non-equivalent, while the text says this yields "zero regressions on the correct side as well". If "regression" is counted against the tribunal's verdict, I would say so; if not, these 10 look like false positives. Also, the Abstract, §1.5, and §11 still say the guard "transfers to a financial domain", next to "domain sanity check". I would keep only the sanity-check wording.

5. §9.3 lists "the format interpretation of F1" among the five retired claims, while the main finding of the paper is that F1 reflects format (§5.4, §9.1). I would name each retired claim together with the measurement that replaced it, so a reader does not take this as retiring the main result.

5. Recommendation

Minor revisions. Version 7 closes six of my seven points from round 5 cleanly in the text: the arithmetic in Table 9, the 168 breakdown, the Table 10 footnote, the counts and table references, R3.1 as an optional add-on, and MedHallu moved to the roadmap. What remains is small. I would run the held-out check with the tuned threshold, and I would add one line of explanation for three table cells: Table 10 A→A" Corrected 9, Table 10 A→B 97, and the 10 flagged correct answers in Table 12. I would be glad to read version 8 and the R3.2 paraphrased control when it comes.

Declarations

Potential competing interests: No potential competing interests to declare.