Research Article

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

Gonzalo Durante[1]

1. Independent researcher

Hallucination-detection systems are usually judged by a single headline metric, computed under a benchmark design that determines what that metric can show. This paper's contribution is not a detection algorithm — structural coherence scoring is established — but a methodology for measuring exactly what a benchmark proves and what it does not, applied to the SAS Multimetric Tribunal: a seven-module, GPU-free structural classifier combining lexical, entity-mutation, curvature, and logical-inversion signals via a cascading multiplicative penalty ($\kappa_D = 0.56$, $\kappa_R = 0.15$). On a stratified 1,800-pair benchmark (HaluEval-Dialogue, HaluEval-QA, TruthfulQA), the tribunal reaches F1 = 98.99% with zero false positives under an A→A identical-text sanity check. We report this figure and then show, using a format-controlled negative (N1) on the same 1,800 IDs, that it measures format asymmetry between source and response rather than hallucination detection. The tribunal flags 1,569 of 1,800 verbatim correct answers presented alone (87.2%), and in HaluEval-QA scores the correct answer as *more* anomalous than the hallucinated response.

Three further findings follow from the same audit. (i) A per-module analysis shows the tribunal's seven-module core has no content-sensitive module (prior to the optional R3.1 guard): the one module designed to capture semantic divergence (`cre_isi`) gives the same per-ID score whether the response is byte-identical or completely different. (ii) The 87.97-point advantage of multiplicative combination over weighted averaging reported in v4 is an artifact of threshold calibration; under per-scheme thresholds on format-matched negatives, both schemes converge to F1 ≈ 0.69 and the multiplicative scheme is worse by AUROC. (iii) A new differential guard (R3.1, grounding-based with compression abstention and short-response heuristics) corrects 1,133 of the 1,569 format-driven false positives on correct answers alone, raising F1 from 0.687 to 0.833 (ΔF1 = $+0.146$), and is evaluated on a financial

 doi.org/10.32388/T420IV.8

domain (FinQA, 1,657 samples) as a sanity check, with 87.0% detection and zero regressions on hallucinated answers (see §8.3 for framing). A held-out 50/50 split over 100 random repetitions confirms that the threshold (0.95) and F1 (0.833 $\pm$ 0.0055 on held-out data) are stable across splits. The guard does not rely on substring matching and is not a constructional artifact in the sense of §8 (see §8 for the distinction between the guard's design and the N1 construction).
The paper's contribution is the audit procedure that produced these retirements. Five of v4/v5's quantitative claims do not survive it. What does survive — thresholds, the κ_R sweep, the forensic audit of v2 metrics, the per-module architecture itself as a diagnostic instrument — is reported with the same standard.

Corresponding author: Gonzalo Emir Durante, duranteg2@gmail.com

1. Introduction

1.1. The Hallucination Problem

Generative language models have achieved remarkable fluency, yet they remain prone to hallucination—the generation of text that is factually inconsistent, ungrounded, or structurally incompatible with a given source [1]. In legal, medical, and scientific applications, a single undetected hallucination can carry catastrophic consequences. The need for fast, interpretable, and auditable detection mechanisms has therefore become urgent.

1.2. Limitations of Current Approaches

Contemporary detection paradigms fall into three broad categories:

1. **LLM-as-judge:** Prompting a larger model (e.g., GPT-4) to evaluate coherence [2]. These methods are accurate but expensive, non-deterministic, and opaque.
2. **Reference-based metrics:** BARTScore [3], GPTScore [4], and similar metrics require gold-standard references or entailment models, limiting applicability to open-ended generation.
3. **Embedding-based classifiers:** Dense retrieval or fine-tuned classifiers offer speed but sacrifice interpretability and modular auditability.

None of these approaches provides a structural verdict with per-module evidence, hash-level traceability, and zero GPU dependency.

1.3. The SAS Structural Line

The Symbiotic Autoprotection System (SAS) proposes an alternative: treat hallucination not as a semantic error to be judged by another model, but as a structural rupture measurable through invariant geometric indices. The research line R0 $\rightarrow$ R0-bis $\rightarrow$ R1 $\rightarrow$ R1-D $\rightarrow$ R2.1 $\rightarrow$ R2.1-b has progressively refined the Invariant Similarity Index (ISI) and its governing threshold $\kappa_D = 0.56$ [5][6][7][8][9][10][11]. This revision advances the line to R2.1-c by applying the line's own auditing methodology to its own validation benchmark. The result is that five quantitative claims of the previous revision do not survive the audit.

1.4. Objective and Scope

This paper audits the SAS Multimetric Tribunal's own validation benchmark. Specifically:

1. We retire the R3 negative control after a construction defect was identified: in 125 of 127 pairs, the text that was paraphrased is not a response but a label string or the question text.
2. We introduce N1, a format-controlled negative on the same 1,800 IDs, constructed from verbatim correct answers already present in each dataset. N1 requires no generated text.
3. We report the N1 result: the tribunal flags 87.2% of correct answers presented alone. The published F1 measures format asymmetry, not hallucination detection.
4. We perform a per-module analysis that finds no content-sensitive module in the current core architecture (prior to the optional R3.1 guard).
5. We re-evaluate the multiplicative-vs-averaging ablation under per-scheme calibration and report that the 87.97-point gap is an artifact.
6. We test whether adding a semantic-equivalence guard creates a differential advantage for the tribunal over a Jaccard baseline, and find it does not for the v5 substring-based guard. We then develop a differential guard (R3.1) that does not rely on substrings, validate it on N1 (F1 0.833), and test it on a financial domain (FinQA, 87.0% detection, zero regressions on hallucinated answers) as a domain sanity check.

Scope. The paper's primary contribution is the audit methodology, not a production detector. The R3.1 differential guard (§8) is a validated content-sensitive module (F1 0.833 on N1, 87.0% detection on FinQA), but it is a structural signal, not a semantic judge. We claim that the original benchmark measures format,

not hallucination; that the audit procedure generalizes; and that the R3.1 guard is the first module in this line that separates content from format without relying on substring artifacts.

1.5. Contributions

This paper does not claim a novel detection algorithm. The contribution is the audit methodology, demonstrated on the paper's own design:

1. **A benchmark audit procedure** (six steps, §2) that uses format-controlled negatives constructed from verbatim dataset fields, with no generated text.
2. **Five retirements of v4/v5 claims**: R3, the format interpretation of the headline F1, the 87.97-point multiplicative-vs-averaging gap, the v5 substring-based guard as a validated extension point, and the claim that the tribunal has no content-sensitive module (superseded by R3.1). Each is measured, not argued.
3. **A per-module contribution analysis** showing that `cre_isi` responds to a property of the source rather than the source-response pair.
4. **A content-sensitive differential guard (R3.1)** that corrects 1,133 format-driven false positives on N1 (F1 0.687→0.833) and is evaluated on a financial domain (FinQA, 87.0% detection, zero regressions on hallucinated answers) as a domain sanity check.
5. **A format-controlled financial corpus**: `financiero_finqa` (1,657 samples, per-row format) with SHA-256 manifest (a second corpus, `medico_medhallu`, 10,000 pairs, is prepared and scheduled in the roadmap, §9.4).
6. **A self-directed forensic audit** of the author's own previously published results, now in its third cycle.

2. The Audit Methodology

2.1. What the procedure is for

The procedure answers a single question: *what does this benchmark actually measure, as distinct from what its authors claim it measures?* It does not ask whether a classifier is correct. It asks whether the benchmark can distinguish the class of inputs the classifier claims to detect from a class of inputs that are, by construction, not that class.

2.2. Step 1: Identify the claimed quantity

For a benchmark B with metric M and claimed target T, state M and T explicitly. In this paper, M is F1 on the A→A/A→B design and T is "hallucinated response".

2.3. Step 2: Construct a format-controlled negative

The negative must be (a) faithful by construction, not by verification; (b) matched to the positive on every dimension except the one under test; (c) same size, same sampling.

In this paper, N1 (§5) is the format-controlled negative.

2.4. Step 3: Report the confusion matrix, not the aggregate

A single accuracy or F1 cannot distinguish content detection from format detection. The audit reports the flag rate on the negative class, the 2×2 matrix, and the per-suite breakdown.

2.5. Step 4: Decide, and document

Three outcomes: the negative is clean (no revision needed); the negative is flagged at a rate near the positive (format confound; reframe); intermediate (recalibrate).

2.6. Step 5: Attribute the signal to modules

If the benchmark fails the format control, the next question is which modules are responsible. A per-module analysis across the four conditions separates modules that respond to the source-response relationship from modules that respond to a property of the source alone. A module whose per-ID score is invariant under changes to the response is not measuring the pair.

2.7. Step 6: Test extension claims empirically

If the paper claims that its architecture provides an extension point (e.g., where a guard can be added), the claim must be tested by adding the same guard to a simpler baseline and comparing. If the simpler baseline plus guard performs equivalently, the extension point is rhetorical.

3. Methods

3.1. The Invariant Similarity Index (ISI)

ISI is bounded in $[0, 1]$. Two thresholds partition the axis: $\kappa_D = 0.56$ (coherence boundary, ISI $\geq \kappa_D$ is Zone A) and $\kappa_R = 0.15$ (collapse boundary, ISI $< \kappa_R$ is Zone F-S). The interval $[\kappa_R, \kappa_D)$ defines Recoverable Rupture (Zone B). **We do not claim invariance to paraphrase, translation, or response length.**

3.2. Tribunal Architecture

Seven modules, each producing a penalty factor in $[0, 1]$. Non-firing modules return 1.0.

#	Module	Function	Weight
1	`lexical_baseline_score`	Jaccard lexical overlap	0.35
2	`source_target_guard`	Critical-entity mutation detection	0.25
3	`flow_penalty`	Semantic flow rupture	0.12
4	`negation_penalty`	Logical inversion detection	0.08
5	`cre_isi`	LSA-based semantic curvature; see §6	0.12
6	`arithmetic_penalty`	Arithmetic error detection	0.04
7	`reference_penalty`	Reference fabrication detection	0.04

Table 1. Tribunal module specification

3.3. Cascading Multiplicative Penalty

$$\mathrm{ISI}_{\mathrm{HARD}} = \begin{cases} \min(\mathrm{lb}, \mathrm{stg}, \mathrm{cre}) & \text{if cre fired and } A \neq B \\ \min(\mathrm{lb}, \mathrm{stg}) & \text{otherwise} \end{cases} \tag{1}$$

$$\mathrm{ISI}_{\mathrm{FINAL}} = \mathrm{ISI}_{\mathrm{HARD}} \times \prod_{\text{fired } i} p_i \tag{2}$$

Why the condition on `cre_isi` is not cosmetic. The module computes semantic curvature over LSA embeddings. On byte-identical or very short texts, embedding dispersion produces spurious firings:

across the 1,800 A→A pairs, `cre_isi` fires on 971 (54%) despite the texts being byte-identical, with scores as low as 0.08. Without the conditional, the sanity check would not produce 100% precision.

Direct ablation. We compare the multiplicative scheme against a weighted average of the same seven module scores. In v4, both were classified under the same $\kappa_D = 0.56$. Under that comparison, the multiplicative scheme appeared to outperform the average by 87.97 F1 points. This comparison is invalid because non-firing modules return 1.0, so the average sits near 0.9 by construction and the threshold penalizes it. We report the corrected comparison in §7.

3.4. SourceTargetGuard

SourceTargetGuard extracts and compares four entity classes (locations, years, proper names, quantities) using regex-based extraction and set-difference analysis. Each detected mutation applies an exponential decay penalty (location 0.70, year 0.65, name 0.80, quantity 0.75), floored at 0.30.

Direct ablation. Re-running the full 1,800-pair validation with STG disabled changes the zone of 107 pairs (5.94%): 104 move F-S→B, 3 move B→A. STG's measured contribution to the original benchmark is 3 true positives out of 1,800 (0.17% of global recall).

3.5. Benchmark Design

We constructed a benchmark from HaluEval-Dialogue, HaluEval-QA [12], and TruthfulQA [13]. For each suite ($n = 600$), we paired A_{clean} (context + correct answer) with a hallucinated variant B (answer alone).

Dataset fields per suite. The reviewer asked which dataset field fills each role. In all three suites, A is the dataset row that contains the context plus the correct answer, and B is the row's hallucinated counterpart. The extraction marker that separates context from answer differs per suite, and N1 uses it verbatim:

Suite	Source dataset	A construction	Marker in A
HaluEval-Dialogue	HaluEval [12]	`Knowledge:` + `Dialogue:` + `Correct response:`	`Correct response:`
HaluEval-QA	HaluEval [12]	`Knowledge:` + `Question:` + `Correct answer:`	`Correct answer:`
TruthfulQA	TruthfulQA [13]	`Knowledge:` + `Question:` + `Correct answer:`	`Correct answer:`

Table 2. Dataset fields and extraction markers per suite

A' (the correct answer alone) is the text after the marker; B is the dataset's hallucinated response field; A'' (same format as A, different content) is the prefix up to and including the marker plus B. Writing these field names down is not cosmetic: the R3 control failed exactly on a field-parsing error (§5).

Because A is context plus answer and B is answer alone, an A→B pair differs in length as well as content. Any structural metric that penalizes reduced lexical overlap will detect the pair whether the response is hallucinated or correct.

Baseline	TP	FN	FP	F1
Exact-match (A $\neq$ B)	1,800	0	0	100.00%
Jaccard < 0.56	1,773	27	0	99.24%
ROUGE-L < 0.50	1,697	103	0	97.06%
ROUGE-L < 0.65 (Youden-optimal)	1,799	1	0	99.97%
SAS Tribunal	**1,764**	**36**	**0**	**98.99%**

Table 3. Trivial baselines on the A→A/A→B benchmark, $n = 1{,}800$ each condition

3.6. Calibration of κ_R

Extended sweep on per-pair ISI values confirms $\kappa_R = 0.15$ as the entropy-maximizing value. Since Zone A is fixed at 2.0% by κ_D alone, entropy is maximized when the remaining 98% splits evenly between B

and F-S. The measured median ISI of detected pairs is 0.1424, indistinguishable from the entropy-maximizing κ_R. $\kappa_R \approx 0.15$ is a property of this sample's ISI distribution, not an independent calibration target.

4. Results (Original Benchmark)

Metric	Value	95% CI (Wilson)
F1 (A→B)	98.99%	—
Recall (A→B)	98.00%	[97.2%, 98.6%]
Precision (A→A sanity check)	100.00%	— (0 FP by construction)
Accuracy	99.00%	—
False Negatives	36	—

Table 4. Global classification metrics on the A→A/A→B design

The R3 specificity row from v4 has been withdrawn; see §5.

4.0.1. Forensic audit of the previously published (v2) metrics

Prior to v3, we audited metrics published in an earlier version (F1 = 99.02%, Recall = 98.06%, 35 FN, TP = 1,765). Re-execution under a documented protocol found that seed = 123 reproduces those figures under the full seven-module core: undocumented, not fabricated. Two defects were corrected in that cycle: a silent partial-core degradation and a non-stratified sampling bug.

doi.org/10.32388/T420IV.8

Seed	F1	Recall	Accuracy	TP	FN	Zone (F-S/B/A)
42 (primary)	98.99%	98.00%	99.00%	1,764	36	50.4/47.6/2.0
123	99.02%	98.06%	99.03%	1,765	35	52.0/46.1/1.9

Table 5. Seed sensitivity, full seven-module core

5. N1: A Format-Controlled Negative

5.1. Retirement of R3

The v4 paper reported a negative control (R3) of 127 paraphrase pairs on which every method produced FP rates above 96%. A direct audit found that in 125 of 127 pairs, the text that was paraphrased is a label string (`Dialogue:`) or the question text, due to a field-parsing error. Only 2 pairs contain a real response. The 126/127 figure, the 0.8% specificity, Table 11 of v4, the R3 row of the roadmap, limitation 3.2, and the Abstract's paraphrase claim are withdrawn.

5.2. N1 construction

N1 uses the correct answers already present in each dataset. Four conditions on the same 1,800 IDs: A→A (identity), A→A' (correct answer alone), A→B (published condition), A→A" (context + marker + hallucinated response). A full 7,200-evaluation run takes 36 seconds on CPU.

doi.org/10.32388/T420IV.8

5.3. Results

Condition	Median	Mean	Min	Max
A→A	1.0000	1.0000	1.0000	1.0000
A→A'	0.1803	0.2308	0.0099	0.8261
A→B	0.1481	0.1860	0.0000	0.7368
A→A"	0.2945	0.3575	0.0072	1.0000

Table 6. $\mathrm{ISI}_{\mathrm{FINAL}}$ by condition (N1, global, $n = 1{,}800$)

	B **flagged**	B **not flagged**
A' **flagged**	1,556 (86.4%)	13 (0.7%)
A' **not flagged**	208 (11.6%)	23 (1.3%)

Table 7. N1 matrix: A→A' vs. A→B, both under $\kappa_D = 0.56$

5.4. Interpretation

Three measurements establish that the published F1 reflects format, not content.

1. **The tribunal flags the correct answer alone.** 1,569 of 1,800 (87.2%) verbatim correct answers presented alone are flagged as collapse or rupture. The detection rate on hallucinated responses is 98.0%; the two rates are close (10.8 points apart), and their 95% Wilson intervals do not overlap ($[0.8554, 0.8863]$ for 87.2% vs. $[0.9724, 0.9855]$ for 98.0%), so the gap is statistically real but substantively small: both are high, and both reflect the format asymmetry rather than content.
2. **In HaluEval-QA, the correct answer scores worse than the hallucination.** Median ISI for A' is 0.0377, below the median for B (0.1250).
3. **Format dominates content.** A'' yields higher median ISI than B by $\Delta = +0.1480$. Presenting the same hallucination in the format of the source reduces detection.

6. Per-Module Analysis

Condition	`cre` fires	`cre` changes zone	% change
A→A	971	0 (masked)	—
A→B	132	0	0%
A→A’	68	3	4.4%
A→A”	1081	1053	97.4%

Table 8. Per-condition behavior of `cre_isi` (N1, $n = 1{,}800$ per condition)

For the same source text, `cre_isi` produces the same score whether the response is byte-identical (A→A) or completely different (A→A”). The median per-ID delta is exactly 0.0000. The module fires on 971 byte-identical pairs and 1,081 format-matched pairs, rates of 54% and 60%. If the module measured similarity between source and response, the second rate should differ substantially and the per-ID scores should change with the response. They do not.

`cre_isi` responds predominantly to a property of the source (most plausibly LSA dispersion correlating with text length) rather than to the source-response relationship. The response length also modulates firing: short responses (A→B, A→A’) rarely reach the dispersion threshold regardless of the source, while long responses (A→A, A→A”) fire at similar rates regardless of content. The tribunal has no content-sensitive module in its current core set (the R3.1 guard, which is content-sensitive, is an optional add-on; see §8). This is consistent with, and explains, the format-dominance finding.

7. The Multiplicative-vs-Averaging Ablation, Recalibrated

Under per-scheme calibration on format-matched negatives (A→B vs A→A’, $n = 1{,}800$):

doi.org/10.32388/T420IV.8

Scheme	Threshold	TP	FP	F1	AUROC
Multiplicative (this work)	0.550 (optimized)	1,762	1,561	0.688	0.514
Multiplicative (this work)	$\kappa_D = 0.56$ (fixed)	1,764	1,569	0.687	0.514
Weighted average	0.830 (optimized)	1,735	1,461	0.695	0.632
Weighted average	$\kappa_D = 0.56$ (fixed)	105	0	0.110	0.632

Table 9. Combination schemes under per-scheme calibration on format-matched negatives (both thresholds tuned on the same set)

The 87.97-point gap reported in v4 was an artifact of applying a threshold calibrated for the multiplicative scheme to the weighted average. Under per-scheme thresholds on real negatives, the two schemes converge to equivalent F1 (0.688 vs. 0.695, $\Delta\text{F1} = -0.007$), and the multiplicative scheme is *worse* by AUROC (0.514 vs. 0.632). Both are barely above chance. The multiplicative cascade does not separate hallucinated responses from correct responses better than a simple average.

For the reviewer's point 1 from the fourth round, we also report the multiplicative scheme with its threshold optimized on the same set (0.550): the F1 moves from 0.687 to 0.688, i.e. the fixed-κ_D row was already near the multiplicative optimum. The asymmetry between the two schemes' thresholds (0.550 vs. 0.830) is a property of each score's distribution, not of the calibration procedure.

8. A Content-Sensitive Guard: R3.1 Differential

The v4 paper proposed R3.1, a semantic-equivalence guard, as the extension that a per-module architecture enables and a single-scalar baseline does not. That claim was tested in v5 using a substring-based guard and was not supported: the v5 guard was a copy-detection rule that exploited the construction of the control. We then replaced it with a *differential* guard that does not rely on substrings, and we test it in this revision on the same N1 set.

doi.org/10.32388/T420IV.8

8.1. R3.1 Differential Guard construction

The R3.1 differential guard (`r3_1_differential.py`) computes semantic equivalence without substrings and without embeddings:

Grounding (primary signal). Recall of the response's content words against the source: a valid answer (A') introduces no new content words (grounding ≈ 1.0); a hallucinated response (B) introduces new content (grounding < 1.0). This is not Jaccard; it is the fraction of response content words present in the source.

Compression ratio (secondary signal). If the response exceeds 50% of the source length, the guard abstains: grounding is uninformative because the response may include the context.

Short-response heuristics ($<$ 5 words). Short answers are statistically unstable under raw grounding (a single matching token can yield grounding $= 1.0$ even with opposite meaning). Three heuristics handle this:

- *Polarity inversion:* a short response containing "no/not/none/nunca" that contradicts the source is non-equivalent regardless of grounding.
- *Isolated entity:* a bare number, year, or proper name is abstained from when grounding is high but context specificity is low (`1945`, `Paris`).
- *Stopword grounding:* all-stopword responses (`yes`, `no`) are checked against raw tokens, not content words.

Threshold calibration. The grounding threshold was swept over the N1 audit data (scripts/sweep_threshold_guard.py). The v5 default (0.50) was too permissive: a hallucination with 80% lexical overlap (`Paris is the capital of Italy` vs. `France`) scored equivalent. The sweep minimizes total error at **0.95** ($n = 1{,}800$): total error drops from 1,371 (at 0.50) to 237 (at 0.95).

8.2. R3.1 Results on N1

Condition	Equiv (guard)	Abstain	Corrected	Regressed
A→A	1,800 (100%)	0 (0%)	0	0
A→A'	1,364 (75.8%)	219 (12.2%)	1,133	0
A→B	204 (11.3%)	61 (3.4%)	0	168
A→A"	0 (0%)[†]	1,800 (100%)	9	0

Table 10. R3.1 differential guard on N1 ($n = 1{,}800$ per condition), threshold 0.95

[†] *Of the 1,800 abstentions on A→A", 362 have grounding* ≥ 0.95 *but are overridden by the compression-ratio abstention (A" is format-matched and long, so the response-to-source length ratio exceeds the 50% abstention threshold). The equiv count in the raw data is 362; the table reports 0 because the guard's verdict is "abstain" for all 1,800 pairs. The 9 corrected pairs are a subset of the 362 equiv: the guard's grounding score is* ≥ 0.95 *and the tribunal had flagged them as false positives, but the guard's verdict is "abstain" rather than "equivalent"; the correction is counted because the guard overrides the tribunal's false positive even though it does not affirmatively classify the pair.*

Method	FP	Recall	Precision	F1
Tribunal alone	1,569	0.9800	0.5293	0.6873
Tribunal + R3.1 guard	436	0.8867	0.7854	0.8330
Δ	$-1{,}133$	-0.0933	$+0.2562$	$+0.1457$

Table 11. Global metrics: tribunal alone vs. tribunal + R3.1 guard (A→B positive, A→A' negative, $n = 1{,}800$)

The guard corrects 1,133 of the 1,569 format-driven false positives on A→A' (72% of the original FP rate eliminated) while keeping A→A intact at 100% and introducing zero regressions on the negative side. On

doi.org/10.32388/T420IV.8

the positive side (A→B), recall drops from 0.9800 to 0.8867: the 168 regressions break down as 61 by compression-ratio abstention (long hallucinated responses where grounding is uninformative) and 107 by false equivalence (the guard's grounding threshold is met despite hallucinated content). The remaining 97 equiv on A→B (204 total equiv minus 107 false-equivalence regressions) are pairs where the tribunal had already returned a correct verdict (ISI below κ_D), so the guard's equiv verdict does not change the outcome. Precision rises from 0.5293 to 0.7854, and F1 from 0.6873 to 0.8330 (ΔF1 = $+0.1457$).

Why this guard is not a substring artifact. The v5 substring-based guard returned 1.0 for every A→A' pair because A' is a substring of A by construction. The differential guard does not check substrings; it checks whether the response's content words appear in the source, with abstention for long responses and strict short-response heuristics. Its success on A→A' is not an exploitation of the extraction method: A' gets grounding ≈ 1.0 because it is genuinely composed of the source's content words.

Held-out validation. Because the threshold 0.95 was chosen by sweeping over the same N1 set on which the guard is evaluated, we performed a 50/50 held-out split stratified by suite (seed = 42), repeated over 100 random splits. On the tuning half ($n = 900$ IDs), the optimal threshold is 0.97 (F1 = 0.882). On the held-out half ($n = 900$ IDs), the guard at threshold 0.95 achieves F1 = 0.833 (mean over 100 splits, $\sigma = 0.0055$), essentially identical to the in-sample figure. When each tuning half selects its own threshold, the mean held-out F1 at the tuned threshold is 0.835 ($\sigma = 0.0058$), and the threshold distribution is concentrated at 0.97 (72% of splits), with the remainder at 0.96 (18%) and 0.98 (10%). The threshold is stable across splits, and the held-out F1 confirms that the result is not an artifact of threshold overfitting. We report the in-sample figure (0.833) in Table 11 because it uses the full N1 set and is directly comparable with the tribunal-alone row.

8.3. R3.1 Generalization: Financial Domain (FinQA)

To test whether the guard generalizes outside N1, we converted financial `financiero_finqa` (1,657 samples, MIT) to the same format-controlled design (scripts/convert_finqa.py). The dataset has no paired queries (0 identical queries between correct and hallucinated samples), so we evaluate per-row: correct completions (1,418) as A/A' and hallucinated completions (239) as A"/B, and additionally test the guard cross-query (A of a random correct sample vs. B of each hallucination).

Condition	`noequiv`	`equiv`	Abstain
A→B (cross-query)	208 (87.0%)	0 (0%)	31 (13.0%)
A→A' (correct)	10 (0.7%)	80 (5.6%)	1,328 (93.6%)

Table 12. R3.1 guard on FinQA (cross-query detection, $n = 239$ hallucinated)

The guard detects 208 of 239 hallucinated financial answers (87.0%) with zero regressions on hallucinated answers (0 false positives on the 239 hallucinated queries when matched against a random correct source). On correct answers (A→A', $n = 1{,}418$), the guard abstains on 1,328 (93.6%) and flags 10 (0.7%) as non-equivalent. These 10 are counted as regressions against the tribunal's verdict (the tribunal had correctly passed them, and the guard overrides with a non-equivalent verdict), yielding zero regressions on the correct side when measured against the guard's own abstention policy. The high abstention on A→A' is a property of the domain: FinQA correct answers are bare numbers (mean 7 chars), which the isolated-entity heuristic correctly abstains from. We present FinQA as a domain sanity check rather than a generalization claim: the guard's grounding mechanism is domain-agnostic, but the 93.6% abstention rate on correct answers means the zero-regressions result is partly a property of the answer format, not solely of the guard's discrimination. During this sanity check we found and fixed a defect in the isolated-entity regex: it did not match negative numbers (`-19`, `-2.7%`), causing 5 groundings of 1.0 on hallucinated negative values whose digits appear in the financial source. The fix (negative, decimal, and percentage patterns) removed all 5 regressions without changing N1 results.

9. Discussion

9.1. The format confound

The published F1 reflects response length. This was stated as interpretation in v5 §9.1; the reviewer asked for a direct measurement. On the N1 set, we computed the correlation between $\text{ISI}_{\text{FINAL}}$ and the ratio of response length to source length:

- A→B (hallucinated response): Pearson $r = +0.5996$ ($n = 1{,}800$).

- A→A' (correct response): Pearson $r = +0.8696$ ($n = 1{,}800$).

Longer responses score higher ISI in both conditions; the correlation is stronger on correct answers than on hallucinated ones. This is consistent with the claim that ISI tracks format (length and lexical overlap) rather than content, and it makes the length explanation the measured one rather than the plausible one. Any structural metric that penalizes reduced lexical overlap produces the same result on this design. This is a property of the benchmark, not of the tribunal.

9.2. The absence of content signal

The tribunal's seven-module core has no content-sensitive module (prior to the optional R3.1 guard). `cre_isi`, the module designed to capture semantic divergence, responds to the source rather than to the pair. On format-matched negatives, the tribunal's AUROC is 0.514 — barely above chance.

The firing asymmetry in Table 8 (971 on A→A and 1,081 on A→A" vs. 132 on A→B and 68 on A→A') is gated by two conditions in the tribunal's cascade (§3.3). `cre_isi` is only included in ISI_{HARD} when it fires *and* the two texts are not byte-identical; and its score is computed from the source's LSA dispersion, so it fires on long format-matched pairs (A→A and A→A") regardless of the response, while short responses (A→B, A→A') rarely reach the dispersion threshold. The gate suppresses it on byte-identical pairs for the verdict but the firing count in Table 8 counts the module's own activation, which is why A→A still shows 971 firings without any zone change.

9.3. The retirements

Five v4/v5 claims do not survive this audit: R3 as a valid negative control (the R3 construction was found to paraphrase label strings rather than responses in 125 of 127 pairs); the interpretation of the headline F1 as measuring hallucination detection rather than format asymmetry (retired by the N1 format-controlled negative, which shows 87.2% of correct answers are flagged); the 87.97-point multiplicative-vs-averaging gap (retired by per-scheme calibration, which shows both schemes converge to F1 ≈ 0.69); the v5 substring-based guard as a validated extension point (retired by the differential guard, which does not rely on substrings); and the claim that the tribunal has no content-sensitive module (superseded by R3.1, which is content-sensitive). Each was retired or superseded by direct measurement, with the same protocol.

9.4. Roadmap

Milestone	Focus	Description
R3.1	Content-sensitive module	**Completed in this revision.** Differential guard (grounding + compression + short-response heuristics, threshold 0.95). F1 0.833 on N1 (in-sample), 87.0% detection on FinQA (domain sanity check).
R3.1b	Held-out validation	Split N1 50/50 by ID (seed=42, 100 repetitions); re-swept threshold on tuning half (optimal 0.97, F1 0.882); held-out F1 at 0.95 is 0.833 ± 0.0055. Threshold stable across splits.
MedHallu	Medical domain corpus	medico_medhallu (10,000 pairs, 4-variant) converted and manifest-verified. Evaluation under R3.1 guard pending.
R3.2	Independent negative set	A control where the correct answer is *not* a substring of the source (paraphrased correct answers, human-verified). This is what R3 was intended to be.
R4	External baseline comparison	Compare against SelfCheckGPT, FActScore, G-Eval under the same format-controlled negatives.
R5	Domain-adaptive κ_D	Recalibrate per domain, on format-controlled negatives per domain.

Table 13. Roadmap (post-v6)

doi.org/10.32388/T420IV.8

9.5. Known Limitations

#	Limitation	Status
3.1	Reference-penalty calibration corpus yields tautological precision	Open
3.2	Format confound in the A→B benchmark	Measured (§5)
3.3	Stealth hallucinations with high lexical similarity	Open
3.4	`legal_hallucinations` corpus pairing defect	Open
3.5	Non-determinism from unsorted directory listing	Corrected
3.6	`negation_penalty` recall estimate	Open
3.7	`arithmetic_penalty` language and form coverage	Open
3.8	`reference_penalty`: additions vs. mutations	By design
3.9	Corpus coverage and domain generalization	Open
3.10	κ_R grid extension below 0.15	Closed
3.11	No content-sensitive module in current core set	Measured (§6); R3.1 guard now provides one (§8)
3.12	R3 control withdrawn; N1 replaces it	Documented (§5)
3.13	87.97-point multiplicative-vs-averaging gap is an artifact	Closed (§7)

#	Limitation	Status
3.14	R3.1 semantic-guard extensibility claim not supported	Superseded — the differential guard (this revision) does not rely on substrings and is not a constructional artifact in the sense of §8 (the guard's grounding mechanism is domain-agnostic; the N1 substring construction is a property of the test, not of the guard)
3.15	N1 negative (A→A') is a substring of the source by construction, which biases the guard comparison	Open — a stronger control requires paraphrased correct answers, not available in this paper

Table 14. Known limitations index

No limitation above changes $\kappa_D = 0.56$ or $\kappa_R = 0.15$.

Why $\kappa_D = 0.56$ survives. The threshold was tuned on the original benchmark [5], and N1 shows that benchmark measures format. Two facts keep the threshold usable despite that. First, κ_D is a property of the score distribution, not of the benchmark's label semantics: it separates the two modes of $\text{ISI}_{\text{FINAL}}$ visible in N1 (format-matched pairs cluster below 0.35, identity pairs at 1.0), and it was re-confirmed on the seed-42 and seed-123 stratified runs. Second, the audit does not change the threshold; it changes what firing means. The honest formulation is: $\kappa_D = 0.56$ remains the boundary of the format distribution, and N1 shows that crossing it does not, by itself, prove hallucination. The same applies to $\kappa_R = 0.15$, which v5 established as the entropy-maximizing value of this sample's ISI distribution.

10. Reproducibility

10.1. Runtime

The N1 benchmark (7,200 evaluations) runs in 36 seconds on a single CPU, no GPU. Approximately 5 ms per pair.

Hardware: AMD Ryzen 5 5600X, 16 GB RAM, Python 3.13, single-threaded.

10.2. Commit and release tag

The tribunal core, the N1 control script, the validation script, the per-pair results, the guard, and the audit trail used in this paper correspond to the v8 revision of the project repository (tag `v8.0.0`, commit `7dbd189`). The complete audit trail is provided with this submission as the supplementary artifact `SAS_v8_Supplements.zip`, which contains all files listed below together with the project changelog.

Reproducibility manifest (SHA-256):

- `n1_control.py`: `270d9b13d2e9505604da2c92e309d69e63ce1e022bdd6d22dc821968f7ed3e2d`
- `pairs_conditions_1800_seed42.json`: `267b7abbf241a75ceb4e6fbb3d073f9257bbcd9c09144a751864551ca94fa603`
- `tribunal_multimetrico.py`: `95b106b2ced30f85b7e6f69c65ad61985b8a5c03aaf6402b55cb76bb63b0f3bb`
- `validacion_1800_stratified_v3.py`: `65415c53c7032b529664a5a9c2cb51cb5cd9d5093b252d0c38eba884484ab939`
- `r3_1_differential.py`: `3cd734b39f3376c0dee3d616fe6f4b2992557b070d474284c0e7ede870e7de1a` (v4, threshold 0.95, entity-regex fix)
- `audit_n1_guard_v4.json`: `b2ecc0b1d0df1c52cfcbabe3f776e2f906733bce2ab742e08c360decf1c499cf`
- `recalibracion_v6.json`: `553f8825dfb22d2b0070679df4c7857df9bfdb34340df5371be4c150283d8c4a`
- `convert_finqa.py`: `2561821c40e9d9b28ba840d2a0c5f26f6f4b295733c2b304c902c96817549688`
- `pairs_finqa.jsonl`: `7e5a1b8c92fa8385b6a77cc6c43158e5d42c932808bef2c72653ea6015778037`
- `evaluacion_finqa_guard.json`: `2235fc2ce412615b9e95ea075a19cbfcfa1f565b14f97f017924d9f8f41be50d`
- `convert_medhallu.py`: `74d9ae9e1aee1da3fc8d0fdd288eb7ad9df7191a74ac1471da3fbe83d65ac538`
- `sweep_threshold_guard.py`: `1cfff172d33110aba8b8e8473701aea82d623b33d673256220e13a79e8097226`

Full manifest with all 25 files: §10 of this revision and the `Manifest_hashes.txt` in the supplement.

10.3. Determinism

The tribunal uses no LLM and no trained parameters. Sampling is deterministic under `sorted(glob)` and a fixed seed. `cre_isi` uses LSA embeddings, which are deterministic given fixed vocabulary and corpus.

11. Conclusions

The SAS Multimetric Tribunal reports F1 = 98.99% on a 1,800-pair benchmark. This paper shows, using a format-controlled negative on the same IDs, that the figure measures format asymmetry between source and response rather than hallucination detection. A per-module analysis shows the tribunal's seven-module core has no content-sensitive module (prior to the optional R3.1 guard). A recalibrated comparison shows the 87.97-point advantage of multiplicative combination over weighted averaging was an artifact.

The R3.1 differential guard, developed in this revision, is the first module in the tribunal's line that separates content from format without relying on substring matching. It corrects 1,133 of the 1,569 format-driven false positives on correct answers alone (F1 0.687→0.833) and is evaluated on a financial domain (FinQA sanity check) with 87.0% detection and zero regressions on hallucinated answers. Its remaining limitations on N1 (168 regressions) break down into 61 compression-ratio abstentions and 107 false equivalences (alongside 31 abstentions on FinQA).

Five v4/v5 claims do not survive this revision. What does survive: $\kappa_D = 0.56$ and $\kappa_R = 0.15$ as thresholds; the extended κ_R sweep; the forensic audit of v2 metrics; the corrected subcount (348→453/401); SourceTargetGuard's measured contribution (3 TPs, 0.17%); the runtime measurement (36 s for 7,200 evaluations); and the per-module architecture as a diagnostic instrument — now with one content-sensitive module to start the R3 line.

The methodological contribution is the audit procedure that produced these retirements. It requires no generated text, no annotator, no model inference, and no GPU. Its first application was to a paper whose headline metric, control, ablation, and proposed extension it invalidated. Whether it generalizes to other benchmarks is the next question, and the honest answer is that this paper does not test it.

Appendix A. Withdrawn Control: R3

The v4 control (R3) is withdrawn. A direct audit found that in 125 of 127 pairs, the text that was paraphrased is a label string or the question text, due to a field-parsing error. Arithmetic verification (126/127 matching the raw count) is not content verification. The replacement control (N1) uses no generated text; every A' is a verbatim dataset field.

Statements and Declarations

Funding

No specific funding was received for this work.

Potential Competing Interests

No potential competing interests to declare.

Ethics

Publicly available benchmark datasets; no human subjects. No ethical approval required.

Data Availability

The complete set of artifacts supporting this paper — source code, per-pair results, the κ_R sweep, both ablation results, the per-module analysis data, the project changelog documenting five retirements from v4/v5, and the SHA-256 manifest — is provided with this submission as `SAS_v8_Supplements.zip`. Reviewers can verify every numerical claim against the files in this archive. The supplement contains 25 files with a SHA-256 manifest (27 entries including the manifest files themselves).

The same artifacts are also available at:

- Zenodo DOI: 10.5281/zenodo.19702379
- GitHub: https://github.com/Leesintheblindmonk1999/sas-multimetric-tribunal

Supplementary material: N1 per-pair results (`pairs_conditions_1800_seed42.json`), N1 script, the false-negative registry, both seed validation files, the κ_R sweep, both ablation results, the per-module analysis data, the tribunal implementation, the R3.1 differential guard (`r3_1_differential.py`), the N1 guard audit (`audit_n1_guard_v4.json`), the FinQA conversion (`convert_finqa.py`,

`pairs_finqa.jsonl`), the FinQA guard evaluation (`evaluacion_finqa_guard.json`), the MedHallu conversion (`convert_medhallu.py`, 10,000 pairs), and the threshold sweep script (`sweep_threshold_guard.py`). Withdrawn: `results_ac_pilot.jsonl`, `baselines_triviales_r3.json`.

License and Attribution

The manuscript is released under the **Creative Commons Attribution 4.0 International License (CC BY 4.0)**, matching the header displayed on the Qeios page. Earlier versions stated CC BY-NC 4.0; the mismatch with the platform header is resolved in favor of the displayed license. The software is dual-licensed: GPL-3.0 for open-source use, a separate commercial license for proprietary integration. Inquiries: duranteg2@gmail.com. Full license texts in the repository.

Any use of the audit methodology, the Invariant Similarity Index, or the thresholds κD=0.56 and κR=0.15 must cite this work and its author.

Author Contributions

GED conceived the study, developed the methodology, implemented the software, conducted all experiments, analyzed the results, and wrote the manuscript.

Use of Generative AI

All experiments were designed and specified by the author. Generative AI tools were used as instruments for running author-specified scripts, drafting, and cross-checking reported numbers against raw output. No tool selected an experimental design, threshold, or interpretive claim.

References

1. ^Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, Ishii E, Bang YJ, Madotto A, Fung P (2023). "Survey of Hallucination in Natural Language Generation." *ACM Comput Surv.* **55**(12):1–38.
2. ^Liu Y, Iter D, Xu Y, Wang S, Xu R, Zhu C (2023). "G-Eval: NLG Evaluation Using GPT-4 with Better Human Alignment." *arXiv preprint arXiv:2303.16634.*
3. ^Yuan W, Neubig G, Liu P (2021). "BARTScore: Evaluating Generated Text as Text Generation." *Adv Neural Inf Process Syst.* **34**:27263–27277.
4. ^Fu J, Ng SK, Jiang Z, Liu P (2023). "GPTScore: Evaluate as You Desire." *arXiv preprint arXiv:2302.04166.*

5. ^a, ^b*Durante GE (2026a). "SAS — Symbiotic Autoprotection System." Zenodo. doi:10.5281/zenodo.19702379.*
6. ^*Durante GE (2026). "R0 Infrastructure and Baseline Stability Audit." Zenodo.*
7. ^*Durante GE (2026). "R0-bis Nonlinear Dependence and Redundancy Audit." Zenodo.*
8. ^*Durante GE (2026). "R1 Real Local Structural Evaluation v1.0.7." Zenodo.*
9. ^*Durante GE (2026). "R1-D Structural Evaluation over Declarative Corpus." Zenodo.*
10. ^*Durante GE (2026). "R2.1 Structural Code Hallucination Detection." Zenodo.*
11. ^*Durante GE (2026). "SAS R2.1-b, Function-Level Extension." Zenodo.*
12. ^a, ^b, ^c*Li J, Cheng X, Zhao WX, Nie JY, Wen JR (2023). "HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models." arXiv preprint arXiv:2305.11747.*
13. ^a, ^b*Lin S, Hilton J, Evans O (2022). "TruthfulQA: Measuring How Models Mimic Human Falsehoods." Proc ACL. 3214–3252.*

Supplementary data: available at https://doi.org/10.32388/T420IV.8

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.