Research Article

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

Gonzalo Durante[1]

1. Independent researcher

Hallucinations in large language models (LLMs) remain a critical barrier to the safe deployment of AI-generated text in high-stakes domains. Existing detection methods typically rely on auxiliary LLMs, reference-based metrics, or human evaluation—approaches that are computationally expensive, opaque, and difficult to audit. This paper presents the SAS Multimetric Tribunal, a lightweight, fully interpretable structural classification system that evaluates (source, response) pairs across seven detection modules using a cascading multiplicative penalty strategy. The tribunal partitions the output space into three structural zones—Structural Collapse (F-S), Recoverable Rupture (B), and Coherent (A)—governed by two thresholds: $\kappa_D = 0.56$ (coherence boundary, registered TAD EX-2026-18792778) and $\kappa_R = 0.15$ (collapse boundary, selected via entropy maximization of the zone distribution).

Over an expanded benchmark of $n = 1,800$ pairs (600 HaluEval-Dialogue + 600 HaluEval-QA + 600 TruthfulQA), each evaluated under an A→A sanity check and an A→B hallucination condition, the tribunal achieves $\mathrm{F1} = 98.99\%$, Precision $= 100.00\%$ (zero false positives under the sanity check), Recall $= 98.00\%$, and 36 false negatives (range across sampling seeds: F1 98.99–99.02%, FN 35–36). Crucially, we document that 100% of errors (36/36) concentrate in TruthfulQA and exhibit a predictable structural pattern: high-lexical-similarity factual hallucinations that trigger no specialized module, leaving `lexical_baseline_score` as the sole discriminator. We treat these 36 false negatives not as noise but as the *documentable failure mode* of the current lexical baseline against stealth hallucinations.

We demonstrate that cascading multiplicative penalty preserves per-module signal integrity better than weighted averaging, and we introduce SourceTargetGuard, a novel module for detecting critical-entity mutation under high lexical overlap. We explicitly caution that the reported metrics constitute

 doi.org/10.32388/T420IV.3

an upper bound under a tautological sanity check (A→A); the real negative control (A→C valid paraphrase), executed in this revision, confirms a 99.2% false-positive rate on paraphrases, which we document as the primary open limitation. This revised version additionally reports a forensic audit of the metrics published in the previous manuscript version: we could not initially locate a validation artifact matching the previously reported numbers, re-executed the validation under a fully documented and seeded protocol, and confirm that the earlier metrics were reproducible (not fabricated) but had been insufficiently documented at the time — a distinction we consider important enough to report in full rather than silently correct. This revision integrates a systematic Known Limitations section, articulates the methodological contributions, and corrects the documentation inconsistencies identified during peer review and internal audit.

Correspondence: papers@team.qeios.com — Qeios will forward to the authors

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 paper advances the line to R2.1-b by introducing a multimodal tribunal architecture, expanding the benchmark to $n = 1,800$ pairs, and conducting a systematic failure-mode analysis.

1.4. Objective and Scope

This paper validates the SAS Multimetric Tribunal as a three-zone structural classifier. Specifically, we:

1. Define and justify the cascading multiplicative penalty strategy.
2. Introduce SourceTargetGuard, a module for entity-mutation detection.
3. Report classification metrics over 1,800 labeled pairs from HaluEval and TruthfulQA.
4. **Document the 36 false negatives as a structured failure-mode case study.**
5. **Report a forensic audit of the previously published validation metrics**, distinguishing an undocumented-but-reproducible result from a fabricated one, and correct the components that do not reproduce.
6. Explicitly identify the methodological limitations of the current benchmark and propose a roadmap (R3.1–R5), with R3 completed in this revision.

Scope limitation: We do *not* claim that the reported 98.99% F1 represents real-world performance on arbitrary text. It is an upper bound obtained under an A$\rightarrow$A sanity check. The contribution lies in the structural architecture, the interpretable penalty logic, and the honest documentation of failure modes.

1.5. Contributions

This paper does not claim a novel detection algorithm. Structural coherence scoring and threshold-based classification are established techniques in computational linguistics. The contribution is the auditing methodology:

1. **A reproducible, fully open-source pipeline** that detects hallucinations in AI-generated text without GPU hardware, without external LLM API calls, and without training data—seven structural modules in a cascading multiplicative architecture with two validated thresholds ($\kappa_D = 0.56$, $\kappa_R = 0.15$).
2. **Per-module validation with documented reversions:** every module was diagnosed by direct execution against its target domain, each attempted fix was measured against a complete firing matrix before deployment, and changes that improved one metric at the cost of zone-changing false positives were reverted. The audit trail is maintained in a SHA-256-verifiable changelog that documents every finding, fix, and reversion with its source of discovery and verification method.
3. **A corpus-integrity auditing protocol** that discovered and documented a large-scale pairing defect in the legal hallucinations suite (90% mismatched pairs), a non-determinism source in the evaluation scripts, and subtype-dependent precision variation in negation detection—each documented with text evidence rather than inferred from aggregate counts.
4. **A self-directed forensic audit of the authors' own previously published results**, applying the same verification standard to prior claims as to the tribunal's classification outputs (§3.1.1).

To the best of our knowledge, no published hallucination detection system combines structural (non-neural) detection with this level of per-module auditability, documented methodology corrections, and verifiable integrity tracking. The contribution is methodological transparency, not algorithmic novelty.

2. Methods

2.1. The Invariant Similarity Index (ISI)

The ISI is a bounded structural similarity metric in $[0, 1]$ designed to be invariant to paraphrase, translation, and syntactic variation while sensitive to factual mutation. Two thresholds partition the ISI axis:

- $\kappa_D = 0.56$: the coherence boundary. ISI $\geq \kappa_D$ indicates structural equivalence between source and response (Zone A).
- $\kappa_R = 0.15$: the collapse boundary. ISI $< \kappa_R$ indicates catastrophic structural disalignment (Zone F-S).

The interval $[\kappa_R, \kappa_D)$ defines Recoverable Rupture (Zone B): structural damage that may be salvageable through editing or re-generation.

2.2. Tribunal Architecture

The tribunal comprises seven detection modules, each producing a penalty factor in $[0, 1]$. A module that does not fire returns 1.0, leaving the signal undiluted.

#	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`	Ricci semantic curvature	0.12
6	`arithmetic_penalty`	Arithmetic error detection	0.04
7	`reference_penalty`	Reference fabrication detection	0.04

Table 1. Tribunal module specification

The Weight column reflects each module's relative importance in the ablation study (R1-D) and does not enter the cascading multiplicative penalty described in §2.3.

2.3. Cascading Multiplicative Penalty

The core innovation of the tribunal is its combination strategy. Rather than averaging module outputs—which causes silent modules (returning 1.0) to dilute the signal of firing modules—the tribunal applies a cascading multiplicative penalty:

$$\mathrm{ISI}_{\mathrm{HARD}} = \min\left(\mathrm{lexical_baseline}, \mathrm{source_target_guard}, \mathrm{cre_isi}\right) \tag{1}$$

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

Here, $\mathrm{ISI}_{\mathrm{HARD}}$ captures the structural core via the most conservative (minimum) of the three primary modules. Only modules that actually detect an anomaly multiply into the final score. This preserves the

doi.org/10.32388/T420IV.3

individual salience of each firing module and avoids the regression-to-the-mean effect inherent in weighted averaging.

2.4. SourceTargetGuard

SourceTargetGuard (STG) is introduced in this work to address a failure mode of lexical baseline: high Jaccard overlap coupled with critical-entity mutation (e.g., "Paris" $\rightarrow$ "Berlin"). STG extracts and compares four entity classes—locations, years, proper names, and quantities—using regex-based extraction followed by set-difference analysis. Each detected mutation applies an exponential decay penalty:

- Location mutation: factor 0.70 per mutation
- Year mutation: factor 0.65 per mutation
- Name mutation: factor 0.80 per mutation
- Quantity mutation: factor 0.75 per mutation

The composite STG score is the product of these factors, floored at 0.30. STG is the only module besides lexical baseline and CRE that feeds directly into $\mathrm{ISI}_{\mathrm{HARD}}$, reflecting its structural importance.

2.5. Benchmark Design and Limitations

2.5.1. Dataset

We constructed a benchmark corpus from three public suites:

- HaluEval-Dialogue [12]
- HaluEval-QA [12]
- TruthfulQA [13]

For each suite ($n = 600$ per suite), we paired clean source texts (A_{clean}) with:

- **Sanity-check pairs (A$\rightarrow$A):** the source paired with itself. This is *not* a negative control; it is a tautological sanity check that asks: "Does the tribunal self-destruct on identical text?" Expected: Zone A.
- **Hallucination pairs (A$\rightarrow$B):** the source paired with a hallucinated variant. Expected: Zone F-S or B.

Total: 1,800 sanity-check pairs + 1,800 hallucination pairs = 3,600 pairs.

2.5.2. Limitations of the Current Benchmark

The A→A design is a methodological ceiling, not a floor. Because identical texts produce Jaccard overlap = 1.0 by definition, the sanity check guarantees zero false positives for any reasonable lexical metric. The reported Precision = 100% and FP = 0 are therefore *upper-bound artifacts* of the benchmark, not evidence of universal discriminative power.

A **real negative control** would pair A_{clean} with a *valid but non-identical correct response* ($A \to C$, paraphrase). Under such a control:

- False positives would likely emerge, because valid paraphrases can exhibit lower lexical overlap.
- Precision would drop from 100% to an estimated 90–95%, depending on paraphrase aggressiveness.
- The 36 false negatives documented here would *remain* false negatives; recall would not change.

This negative control has now been executed (R3, see §4.5): over 127 paraphrase pairs (100 synonym-replacement, 27 back-translation), the tribunal produced 126 false positives (99.2% FP rate), exceeding the original 90–95% estimate and establishing the paraphrase-sensitivity limitation empirically (see §4.7, item 3.2).

2.6. Calibration of κ_R

We swept κ_R across $\{0.15, 0.18, 0.20, 0.22, 0.25, 0.28, 0.30, 0.32, 0.35, 0.38, 0.40\}$ with κ_D fixed at 0.56. For each candidate, we classified the 300 hallucination pairs of the earlier 600-pair validation into the three zones and computed Shannon entropy of the zone distribution. The sweep was run on the 600-pair validation because κ_R was fixed before the expansion to 1,800 pairs; the expanded validation preserves the same zone distribution (see footnote to Table 2).

Critical observation: Under the A→A sanity check, F1 computed over the full sweep including sanity-check pairs remains stable at ∼99.2% across the entire sweep because the classification boundary between "detected" (Collapse + Rupture) and "not detected" (Coherent) is dominated by $\kappa_D = 0.56$ (the hallucination-only F1 reported in Section 4.1 is 98.99%). The F1 term therefore contributes *no discriminative signal* for selecting κ_R.

Consequently, $\kappa_R = 0.15$ was chosen by **entropy maximization** of the three-zone distribution (entropy = 1.474 bits), not by F1 optimization. We reformulate the narrative of v1 accordingly: the selection criterion is entropy-driven taxonomy preservation, not experimental validation via F1.

3. Results

3.1. Global Classification Performance

Note on validation layers: The metrics reported in this section correspond to the expanded validation of $n = 1,800$ pairs (1,800 hallucination + 1,800 sanity-check, 600 per suite, stratified sample seed=42, current tribunal core with all seven modules verified loaded). An earlier validation of $n = 600$ pairs (300 control + 300 hallucination, across the same three suites) reported $\mathrm{F1} = 99.16\%$, Precision $= 100.00\%$, Recall $= 98.33\%$, Accuracy $= 99.17\%$, and ISI separation $= 0.8323$. Both validations are documented in the repository; the 1,800-pair expansion is the primary reference for this paper. The metrics reported here reflect the current tribunal core and a stratified sample (seed=42). §3.1.1 below reports, in full, a forensic audit of how the metrics in the previous manuscript version relate to this validation.

Metric	Value
F1	98.99%
Precision	100.00%
Recall	98.00%
Accuracy	99.00%
False Positives	0
False Negatives	36
ISI mean (sanity check)	1.0000
ISI mean (hallucination)	0.1860
ISI separation	0.8140

Table 2. Global classification metrics (metrics on the A→B hallucination condition, $n = 1,800$ pairs; sanity check on the A→A condition, $n = 1,800$ pairs)

Computed over the 1,800-pair validation (seed=42).

doi.org/10.32388/T420IV.3

	Detected (F-S + B)	Not Detected (A)
Hallucination (A→B)	TP = 1,764	FN = 36
Sanity check (A→A)	FP = 0	TN = 1,800

Table 3. Confusion matrix (global, $n = 1,800$)

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

Prior to this revision, we conducted a forensic audit of the metrics published in the previous manuscript version (Qeios, 2026-08-22: F1 = 99.02%, Recall = 98.06%, 35 false negatives, TP = 1,765). The audit was triggered because no corresponding validation artifact could initially be located in the repository despite an exhaustive search of the full commit history for the reported values and the declared 1,800+1,800 pair structure.

Re-execution with the current core and a fully documented, deterministic sampling procedure (`sorted(glob(...))` followed by `random.Random(seed).sample()`, stratified 600 pairs per suite) uncovered two defects in the interim evaluation history:

1. **Silent partial-core degradation.** A `sys.path` misconfiguration in two earlier scripts pointed at the `core/` directory itself rather than its parent, causing the internal absolute imports (`from core.X import *`) inside five of the seven modules to fail silently on import. The tribunal ran with only two modules active (`lexical_baseline_score` and `source_target_guard`), invalidating two interim results (F1 = 98.79% and F1 = 98.90%) that never reached publication but existed in the working history.
2. **Non-stratified sampling in a preliminary sweep.** An earlier sweep script selected the first 600 files in filesystem enumeration order rather than a seeded random sample, so its output was not a valid stratified measurement.

With the full seven-module core correctly loaded and a documented seed, two independent runs were executed (Table 4). **The seed = 123 run reproduces the previously published v2 metrics exactly** — F1 = 99.02%, Recall = 98.06%, TP = 1,765, FN = 35, including the identical per-suite breakdown (Dialogue 600/0, QA 599/1, TruthfulQA 566/34). This establishes that **the v2 numbers were not fabricated**: they

correspond to a real, correctly executed run under the full seven-module core. What was missing was documentation — the specific artifact and random seed used to produce them were not recorded at the time of the v2 submission.

Seed	F1	Recall	Accuracy	TP	FN	Zone (F-S/B/A)
42 (primary, this paper)	98.99%	98.00%	99.00%	1,764	36	50.4/47.6/2.0
123 (reproduces v2)	99.02%	98.06%	99.03%	1,765	35	52.0/46.1/1.9

Table 4. Seed sensitivity of the 1,800-pair validation, full seven-module core

Three further components of the v2 manuscript are, however, **superseded rather than reproduced**, and are corrected in this revision:

- The reported zone distribution (78.9%/19.2%/1.9%) matches neither seed in Table 4; it corresponds to an unrelated κ_R-sweep artifact computed on the earlier 600-pair set (§2.6), not to a measurement on the 1,800-pair validation.
- The reported per-module firing frequencies in v2 were the 600-pair counts linearly extrapolated by a factor of three, not measured directly on the 1,800-pair sample. The values reported in Table 8 of this revision are the direct 1,800-pair measurement.
- The claim that ISI separation remained stable at $n = 1,800$ (reported in v2 as 0.8323, identical to the 600-pair value) does not hold under direct recomputation: the 1,800-pair sample gives ISI separation = 0.8140 (mean ISI, hallucination condition = 0.1860, versus 0.1677 at $n = 600$). Table 2 reports the recomputed value.

We report this audit in full, rather than silently updating the numbers, because we consider the distinction between *undocumented* and *fabricated* to be methodologically important, and because the audit itself is evidence of the standard we apply throughout this research line: the same verification protocol used to validate the tribunal's classification outputs (direct re-execution, independent recomputation from raw data, documented reversions) was applied here to our own prior claims. The

complete forensic trail — affected scripts, discovered bugs, and both seed runs — is recorded in the project changelog under a SHA-256-verified entry.

3.2. Error Distribution by Suite

Suite	TP	FP	FN	TN	Recall
HaluEval-Dialogue	600	0	0	600	100.0%
HaluEval-QA	600	0	0	600	100.0%
TruthfulQA	564	0	36	600	94.0%
Global	**1,764**	**0**	**36**	**1,800**	**98.0%**

Table 5. Per-suite breakdown ($n = 600$ per suite)

Immediate conclusion: 100% of errors (36/36) concentrate in TruthfulQA. This is not random noise; it is a *structural signature*.

3.3. Zone Distribution

Zone	Sanity check (A→A)	Hallucination (A→B)
Coherent (A)	100.0%	2.0%
Recoverable Rupture (B)	0.0%	47.6%
Structural Collapse (F-S)	0.0%	50.4%

Table 6. Zone distribution by condition ($n = 1,800$ per condition)

doi.org/10.32388/T420IV.3

3.4. Failure Mode Analysis: The 36 False Negatives

3.4.1. The dominant pattern: high-lexical-similarity stealth hallucinations

All 36 false negatives share three properties:

1. **ISI just above κ_D:** The minimum FN ISI is 0.5625; the maximum TP ISI in TruthfulQA is 0.5500. The gap is only 0.0125. The threshold is calibrated at the absolute limit of what `lexical_baseline_score` can detect.
2. **Zero specialized module firings:** 100% of TruthfulQA FNs fired no specialized module ($\text{modules} = []$). By contrast, 100% of TruthfulQA TPs fired at least one specialized module.
3. **High lexical overlap with factual mutation:** TruthfulQA hallucinations maintain domain terminology and syntactic structure while altering factual content subtly (e.g., changing a historical date by one year, substituting a plausible but incorrect statistic).

Type	n	% with modules $= []$	% with modules $\neq []$
TP (TruthfulQA)	**564**	0%	100%
FN (TruthfulQA)	**36**	100%	0%

Table 7. Module firing pattern: TP vs FN in TruthfulQA

3.4.2. Why TruthfulQA differs from HaluEval

TruthfulQA contains *plausible factual hallucinations* that:

- Maintain high lexical overlap with the question/context
- Contain no explicit negations (`negation_penalty` does not fire)
- Do not obviously mutate named entities (`source_target_guard` fires only 6.5% vs. 66.2% in dialogue)
- Do not break logical flow in a way detectable by current regexes (`flow_penalty` does not fire)

By contrast, HaluEval Dialogue hallucinations are conversational, with more lexical variation and more negation/entity-change triggers.

3.4.3. At-risk true positives

348 TruthfulQA TPs (61.7%) depend *exclusively* on `lexical_baseline_score` (modules = [lexical_baseline_score]). These TPs sit in Zone B (ISI 0.15–0.55), indicating the baseline detects *something* but not enough. 99 TPs have ISI > 0.4, grazing the threshold.

3.4.4. Visualization

Four panels (see Supplementary Material) provide:

1. **Histogram:** ISI distribution of TP vs FN in TruthfulQA, with $\kappa_D = 0.56$ marked.
2. **Bar chart:** Zone distribution (F-S, B, A) per suite.
3. **Stacked bar:** Active modules per suite, showing module absence in TruthfulQA FNs.
4. **Box plot:** ISI distribution per suite, showing TruthfulQA's higher median and variance.

3.5. Module Firing Frequency (Global)

Module	Firings	% of Pairs
`lexical_baseline_score`	1758	97.7%
`source_target_guard`	617	34.3%
`flow_penalty`	272	15.1%
`negation_penalty`	194	10.8%
`cre_isi`	132	7.3%
`arithmetic_penalty`	0	0.0%
`reference_penalty`	0	0.0%

Table 8. Modules ranked by firing frequency ($n = 1,800$)

Lexical baseline has the highest pair-level firing rate, at 97.7%, confirming that structural core metrics dominate the tribunal's sensitivity. SourceTargetGuard fires on 34.3% of pairs—precisely the subset

where lexical overlap is high but entities have mutated. Flow penalty fires on 15.1%, negation on 10.8%, and CRE on 7.3%.

`arithmetic_penalty and reference_penalty fire on 0% of pairs in this benchmark.` This is reported explicitly, rather than omitted, because it is consistent with and expected from the coverage limitations documented in §4.7 (items 3.7 and 3.8): neither HaluEval nor TruthfulQA source text contains explicit arithmetic statements or citation-bearing content in the syntactic form these two modules are calibrated to detect. Their small weight in the cascading penalty (0.04 each, Table 1) reflects this scope: on the present benchmark they contribute negligibly to F1, and they are retained in the architecture for domains — legal, scientific, and citation-heavy text — that fall outside the scope of the three suites validated here (§4.7, item 3.9).

4. Discussion

4.1. Honest Assessment of the A→A Sanity Check

The ISI separation of 0.8140 indicates that sanity-check and hallucination distributions occupy nearly disjoint metric subspaces *under the current benchmark*. This is not a trivial artifact, but it is also not a measure of real discriminative power. The sanity check answers a narrower question: "Does the tribunal preserve coherence when no mutation exists?" The answer is yes. The harder question—"Does it distinguish valid paraphrase from hallucination?"—remains answered in the negative by the R3 negative control (99.2% FP on paraphrases, see §4.5), and is the focus of R3.1.

4.2. The Role of SourceTargetGuard

SourceTargetGuard addresses a well-documented failure mode in lexical similarity metrics: high Jaccard overlap masking entity substitution. In our corpus, STG fired 617 times, often in conjunction with lexical baseline. Without STG, these cases would be misclassified as Coherent or borderline Rupture. The module's regex-based design is intentionally lightweight; future iterations will integrate NER pipelines for broader entity coverage.

4.3. Why Multiplicative Penalty Outperforms Averaging

Weighted averaging of module outputs suffers from a fundamental problem: modules that do not fire return neutral values (often 1.0 or 0.5), which dilute the signal of firing modules. In a seven-module

system, a single strong firing event can be masked by six silent modules. The cascading multiplicative penalty solves this by making silent modules transparent (multiplicative identity = 1.0) and firing modules multiplicative (each < 1.0). The result is a compositional logic analogous to Boolean AND with graded membership: coherence requires all structural conditions to hold.

4.4. What the 36 False Negatives Tell Us

The false negatives are not a bug to be hidden; they are *the boundary of the method*. They reveal that:

1. `lexical_baseline_score` is the bottleneck. It needs a complement that measures semantic similarity independently of lexical overlap.
2. TruthfulQA-style hallucinations—subtle factual mutations under high lexical overlap—are the "hard case" for structural auditing.
3. A domain-adaptive κ_D may be necessary: factual QA may require a more aggressive threshold (e.g., 0.50) than conversational dialogue (0.56).

4.5. Roadmap: R3–R5

Milestone	Focus	Description
R3	Real negative control	**Completed.** A→C (valid paraphrase) executed over 127 pairs (100 synonym, 27 back-translation): 126/127 = 99.2% false-positive rate. Confirms the lexical layer's blindness to paraphrase equivalence.
R3.1	Semantic-equivalence guard	Add an embedding-based guard before lexical penalties; target FP rate $< 5\%$ on the R3 pilot.
R4	Dense semantic module	Integrate sentence-transformer embeddings to penalize high-lexical / low-semantic similarity, closing the TruthfulQA FN gap.
R5	Domain-adaptive κ_D	Calibrate κ_D per text type (medical, legal, dialogue, code).

Table 9. Post-peer-review roadmap

Note on R-series identifiers: The R-series identifiers used in this paper (R3, R4, R5) refer to the research milestones defined in this manuscript. The project changelog uses a different internal numbering for its validation cycles (e.g., R5 for the per-module pilot). See the repository for the mapping.

4.6. Limitations Summary

1. **Sanity-check ceiling:** The 0% false-positive rate is specific to the A→A design. We explicitly caution against treating it as a universal guarantee.
2. **Regex-based entity extraction:** STG's coverage is limited to hard-coded location, year, name, and quantity patterns. Domain-specific entities require extension.
3. **CRE short-text guard:** The Ricci curvature module can produce false positives on very short or identical texts due to LSA embedding dispersion. The tribunal mitigates this by excluding CRE from ISI_{HARD} when texts are identical.
4. **No semantic depth module:** The tribunal currently lacks a dense semantic similarity layer. This is the root cause of the TruthfulQA failure mode.

4.7. Known Limitations

The following limitations are documented as part of the experimental validation of the SAS Multimetric Tribunal. Each entry states the limitation, its cause, whether it has been corrected or remains open, and the scope of its impact on reported results. No limitation below affects the validity of $\kappa_D = 0.56$ or $\kappa_R = 0.15$, which remained fixed throughout the validation cycle. Table 10 indexes all nine items for quick reference; the full text of each follows.

#	Limitation	Status
3.1	Reference-penalty calibration corpus yields tautological precision	Open
3.2	Paraphrase sensitivity (structural matching)	Open — empirically confirmed (R3: 99.2% FP)
3.3	Stealth hallucinations with high lexical similarity	Open — root cause of all 36 FNs
3.4	`legal_hallucinations` corpus pairing defect	Open — recovery path identified
3.5	Non-determinism from unsorted directory listing	Corrected
3.6	`negation_penalty` recall estimate	Open — range narrows as evidence grows
3.7	`arithmetic_penalty` language and form coverage	Open — low priority
3.8	`reference_penalty`: citation additions vs. mutations	By design
3.9	Corpus coverage and domain generalization	Open

Table 10. Known limitations index

4.7.1. (3.1) Reference penalty calibration corpus

The `reference_penalty` module reports precision = 1.0 (0 false positives) over the CiteTracer calibration set (686 citation pairs). This figure is a property of the calibration corpus, not a general claim. CiteTracer partitions citations into positive cases (fabricated or manipulated citations) and negative cases (real, correctly reported citations). The negative cases contain no fabricated citations by construction, so the module has no opportunity to produce a false positive on the "clean" side. A precision estimate from this corpus measures the module's ability to avoid false alarms on real citations, not its specificity across arbitrary text. The module has not been evaluated on a corpus that includes plausible but incorrect citations in the negative set, which is the condition that would produce a meaningful false-positive rate.

Status: open. The calibration corpus is suitable for recall measurement but not for precision claims. Any future precision evaluation requires a corpus with hard-negative citations in the clean set.

4.7.2. (3.2) Paraphrase sensitivity (structural matching)

The tribunal's ISI pipeline relies on lexical and structural matching between source and response. When a response is a meaning-preserving paraphrase of the source with low lexical overlap, the lexical baseline and structural modules (`lexical_baseline_score`, `cre_isi`) can fire despite semantic equivalence. The R3 negative control study (A $\rightarrow$ C paraphrases) found a 99.2% false-positive rate under this condition (126/127 pairs): the lexical layer treats a well-formed paraphrase as a structural rupture because Jaccard overlap drops below the coherence threshold.

This is not a bug in any individual module—each module correctly measures what it is designed to measure (lexical divergence, entity change, sentence-count change). The limitation is architectural: the tribunal has no semantic-equivalence guard between the lexical layer and the zone decision. A paraphrase with high semantic fidelity but low lexical similarity is structurally indistinguishable from a hallucination under the current pipeline.

Status: open. The R3 finding is documented in the research series and the repository. A semantic-equivalence guard (R3.1) is pending work and is the highest-priority architectural addition outside the current module set.

4.7.3. (3.3) Stealth hallucinations with high lexical similarity

The 600-pair validation produced 5 false negatives (recall 98.33%); the expanded 1,800-pair validation produced 36 (recall 98.00%). All are hallucinations that maintain high lexical overlap with the source—they replace a small number of key tokens (dates, numbers, entities) within a nearly identical syntactic frame. In these cases, `lexical_baseline_score` remains above the ISI_{HARD} threshold, `source_target_guard` does not detect a critical entity mutation (the changed token is not in its entity whitelist), and no penalty module fires because the sentence structure is preserved.

This is a structural limitation of a purely lexical-structural detector: when a hallucination preserves the source's syntactic frame and only mutates tokens that fall outside the entity guard's coverage, the tribunal has no signal to distinguish it from a faithful response. These cases require either a broader entity-detection mechanism or a factual-verification layer that does not exist in the current architecture.

Status: open. The FN cases are documented in the validation report. Reducing this class requires extending `source_target_guard`'s entity coverage or adding a factual-verification module outside the current scope.

4.7.4. (3.4) Legal hallucinations corpus pairing defect

The `legal_hallucinations` suite (497k of the 596k-pair benchmark corpus) was found to have a construction defect: A_clean and B_hallucination files were numbered by independent counters, not by shared query ID. A random sample of 30 pairs showed ∼90% mismatched—the paired files represent different legal questions, not the same question answered correctly versus incorrectly. The Fase B firing matrix row for this suite (lexical 95.5%, STG 99.0%, cre 92.0%, flow 86.5%) does not measure hallucination detection; it measures how different two unrelated legal texts are. This row is excluded from module validation.

The contamination check confirmed that no reported module-level metric depends on `legal_hallucinations` for more than 0.8 points. A correctly paired subset (84,780 pairs) has since been reconstructed from the source CSV, grouping by query ID rather than by independent row counters, and verified: 20 randomly sampled reconstructed pairs (seed=42) were confirmed, by direct text inspection, to share the same underlying legal question in both files.

Status: open for full-suite module validation, with the recovery path executed for a verified subset. The reconstructed 84,780-pair corpus is available for future validation rounds; the original ∼497k unpaired files remain excluded from this paper's reported metrics.

4.7.5. (3.5) Non-determinism from unsorted directory listing

A systematic audit of the public repository (`sas-multimetric-tribunal/`) found that 10 scripts built candidate-file lists with `Path.glob("*_A_clean.txt")` without wrapping the result in `sorted()` before passing it to `random.Random(seed).sample()`. Python's `pathlib.glob` returns entries in arbitrary order by specification; the order is implementation-dependent across platforms, so the same seed can produce different samples on different operating systems. All 10 affected scripts were identified by name and line number and fixed with `sorted(glob(...))`.

The impact on the numbers reported in this paper is zero: all reported results were generated on a single platform where `glob` order was stable across the runs compared. The issue is a cross-platform portability concern for future reproduction attempts, not a source of variance in the values reported here.

Status: corrected. The fix was verified by running three consecutive calls with `sorted()` on an affected script and confirming identical output across all three runs.

4.7.6. (3.6) Negation penalty recall estimate

The recall of `negation_penalty` on `halogen/rationalization_binary` (∼21,000 pairs) is estimated at 90.6%–95.0%, based on 62 firings among 200 sampled pairs, with 1 confirmed false negative among 60 sampled non-firings ($FN_{est} \approx 2.3$). The non-firing sample was doubled from an initial 30 to 60 during the validation cycle; the second batch of 30 contained zero additional false negatives, which tightened the estimate. The range is reported instead of a fixed decimal because the estimate still rests on a single observed false negative—if the untested remainder (78 non-firings) contained one additional false negative at the same rate, recall would shift to ≈90.6%.

The precision of the same module is 71.1% (32/45 text-verified firings), with strong subtype dependence: 84.6% on primality questions, 66.7% on flight-route questions, 40.0% on senator questions. The low senator precision is a structural pattern: the "First, respond with yes or no" template in A introduces a negation marker ("or no") that is absent in B's affirmative response, producing a spurious polarity asymmetry.

Status: open. The recall estimate is the strongest available for this module but rests on a thin false-negative sample. Precision is subtype-dependent with a known, explainable failure mode.

4.7.7. (3.7) Arithmetic penalty: language and form coverage

The `arithmetic_penalty` module uses English-language regex patterns ("plus", "twice", "% of"). Spanish and other languages are not covered. Additionally, even within English, the module only detects explicit arithmetic expressions ("A plus B is C"); real numeric hallucinations in free-form text rarely take this syntactic form. The 1,600-pair firing matrix shows 0–7% firing across all suites, even though the corpus is 92–99.5% English, and the module fires on 0% of the 1,800-pair benchmark used in this paper (Table 8). The low firing rate is not primarily a language issue—it is a coverage issue with the implicit arithmetic that appears in natural text.

Status: open. The module's contribution to the global F1 (weight 0.04) is negligible on the suites validated here. Addressing implicit arithmetic detection is a separate research problem that the current regex-based approach does not attempt.

4.7.8. (3.8) Reference penalty: citation additions, not mutations

The `reference_penalty` module only penalizes citations that MODIFY a citation already present in the source (A). Citations that appear only in the response (B) with no counterpart in A—fabricated citations added without a source anchor—are excluded by design. This is not a bug: the module's detection pipeline (`detect_fabrications()`) compares citations between aligned text, and a citation with no counterpart in A has nothing to compare against. The module is calibrated for citation mutation detection, not citation addition detection, and fires on 0% of the present benchmark (Table 8), which contains no citation-bearing source text of the form the module targets.

Status: by design. Extending the module to detect added citations without a source anchor would require a different approach (external knowledge-base lookup) and is outside the current scope.

4.7.9. (3.9) Corpus coverage and domain generalization

The broader benchmark corpus comprises 596k pairs across 8 suite families. The module $\times$ domain firing matrix (Fase B) was computed on a 1,600-pair stratified sample (200 per suite, seed=42); the validation reported in this paper uses three of those suites: HaluEval-Dialogue, HaluEval-QA, and TruthfulQA. Of the seven non-empty `halogen` subfolders, only `rationalization_binary` has been validated; the remaining six have not been individually validated for A/B pairing or module performance. The `legal_hallucinations` suite is excluded from the reported metrics due to the pairing defect (item 3.4), though a verified reconstructed subset now exists for future rounds. The `truthfulqa` suite (single-sentence responses) falls below the sentence-count threshold for `cre_isi` and `flow_penalty`, producing 0% firing on both — this is by module design (both require ≥ 2 sentences or ≥ 15 content tokens), not a defect, and is precisely the structural gap that produces the 36 false negatives analyzed in §3.4. The reported global metrics (F1 = 98.99% on the 1,800-pair validation, range 98.99–99.02% across seeds; F1 = 99.16% on the earlier 600-pair validation) are limited to the three suites in the validation sets and do not generalize to domains outside these sets without per-domain validation.

Status: open. Each new domain requires independent validation of both corpus integrity and module calibration before results can be reported.

5. Conclusions

The SAS Multimetric Tribunal achieves F1 = 98.99% (range 98.99–99.02% across sampling seeds) with zero false positives under an A→A sanity check on a 1,800-pair benchmark corpus, validating its suitability as a high-precision structural audit layer. $\kappa_R = 0.15$ is selected as the structural collapse threshold by entropy maximization, producing a meaningful three-zone taxonomy. $\kappa_D = 0.56$ remains the coherence boundary, confirmed across the R0–R2.1-b research line.

Cascading multiplicative penalty is a novel and effective combination strategy that preserves per-module signal integrity. SourceTargetGuard successfully closes the entity-mutation gap under high lexical overlap, firing on 34.3% of evaluated pairs. The tribunal operates without GPU acceleration or external LLM calls, making it deployable in resource-constrained, privacy-sensitive, or air-gapped environments.

The central contribution of this revised version is methodological honesty at two levels. First, we document the 36 false negatives as a predictable failure mode of `lexical_baseline_score` against high-lexical-similarity factual hallucinations, and we report that two of the seven modules fire on 0% of the present benchmark by design rather than omitting them. Second, we extend the same verification standard to our own prior publication: a forensic audit of the previously published validation metrics (§3.1.1) found that those numbers, while reproducible under a documented seed, had not been adequately recorded at the time, and we correct the three derived quantities (zone distribution, firing frequency, ISI-separation stability) that do not survive direct recomputation on the full 1,800-pair sample. We integrate a systematic Known Limitations section covering nine documented limitations, articulate the methodological contributions (reproducible pipeline, per-module validation with documented reversions, corpus-integrity auditing, self-directed audit of prior claims), and correct the documentation inconsistencies identified during peer review. The R3 negative control, now executed, confirms the paraphrase-sensitivity limitation empirically (99.2% FP) and defines the roadmap forward (R3.1 semantic-equivalence guard, R4 dense semantic module, R5 domain-adaptive κ_D). This transforms an inflated claim into a self-correcting, peer-review-ready contribution.

Statements and Declarations

Funding

No specific funding was received for this work.

Potential Competing Interests

No potential competing interests to declare.

Ethics

This study uses publicly available benchmark datasets (HaluEval, TruthfulQA) and does not involve human subjects, animals, or identifiable personal data. No ethical approval was required.

Data Availability

All validation data, source code, and reports are available under the Durante Invariance License v1.0:

- Zenodo DOI: 10.5281/zenodo.19702379
- GitHub (active implementation): https://github.com/Leesintheblindmonk1999/sas-multimetric-tribunal
- Project Manifold 0.56 (historical archive): https://github.com/Leesintheblindmonk1999/Project_Manifold_056
- Public API: https://sas-api.onrender.com
- Landing page: https://leesintheblindmonk1999.github.io/sas-landing/

The supplementary material deposited with this paper includes:

- `sas_fn_analysis_1800.csv` — Complete false-negative registry.
- `sas_tribunal_1800_analysis.png` — Four-panel failure-mode visualization.
- `validacion_1800_stratified_v3.json` — Complete numerical results for the 1,800-pair validation (seed=42, primary).
- `validacion_1800_stratified_seed123.json` — Seed-sensitivity / v2-reproduction check (seed=123).
- `isi_1800_stratified_v3.json` — ISI mean and separation computed on the 1,800-pair validation.
- `tribunal_multimetrico.py` — Core tribunal implementation.
- `r3_negative_control/` — R3 negative control: scripts, raw results (`results_ac_pilot.jsonl`), and report.
- `CHANGELOG.md` + `CHANGELOG.md.sha256` — Complete, hash-verified audit trail, including the forensic audit described in §3.1.1.

Author Contributions

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

Use of Generative AI

The manuscript was prepared with editorial assistance from generative AI tools: Kimi Chat (Moonshot AI), DeepSeek (DeepSeek-AI), and Claude (Anthropic). These tools were used for text drafting, structural organization, LaTeX formatting, and proofreading. All claims, data, methodological decisions, threshold calibrations, and interpretive judgments remain the sole responsibility of the author.

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. ^*Durante GE (2026a). "SAS — Symbiotic Autoprotection System." Zenodo. doi:10.5281/zenodo.19702379.*
6. ^*Durante GE (2026). "R0 Infrastructure and Baseline Stability Audit." Zenodo. https://zenodo.org/records/20647532.*
7. ^*Durante GE (2026). "R0-bis Nonlinear Dependence and Redundancy Audit." Zenodo. https://zenodo.org/records/20671824.*
8. ^*Durante GE (2026). "R1 Real Local Structural Evaluation v1.0.7." Zenodo. https://zenodo.org/records/21034155.*
9. ^*Durante GE (2026). "R1-D Structural Evaluation over Declarative Corpus." Zenodo. doi:10.5281/zenodo.21282332.*
10. ^*Durante GE (2026). "R2.1 Structural Code Hallucination Detection." Zenodo. doi:10.5281/zenodo.21365707.*
11. ^*Durante GE (2026). "SAS R2.1-b, Function-Level Extension — Reference-Free Validation of Python API Calls via Sandboxed Introspection (Version v68.0)." Zenodo. doi:10.5281/zenodo.21466891.*

12. ^a, ^b*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. ^*Lin S, Hilton J, Evans O (2022). "TruthfulQA: Measuring How Models Mimic Human Falsehoods." Proc Annu Meet Assoc Comput Linguist. 3214–3252.*

Supplementary data: available at https://doi.org/10.32388/T420IV.3

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.