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 thirteen-module, GPU-free structural classifier combining lexical, entity-mutation, curvature, logical-inversion, topological-similarity, semantic-drift, entropy, complexity, and surprisal 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

 doi.org/10.32388/T420IV.9

correct answers alone, raising F1 from 0.687 to 0.833 (ΔF1 = $+0.146$), and is evaluated on a financial domain (FinQA, 1,657 samples) as a sanity check, with 87.0% detection and zero regressions on hallucinated answers (see §8.9 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).

v9 addendum. Two post-peer-review advances are documented in this revision. (iv) The R3.1 guard's binary compression-ratio abstention (CR $>$ 0.5) has been replaced by a continuous weight function (κ C), reducing abstentions from 58 to 1 on N1 ($-$98%) without regressions. (v) A new topological-similarity module (R3.2) detects semantic equivalence under paraphrase when lexical overlap is near zero, using canonical clusters, deterministic lemmatization, passive-voice inversion, and a strict anchoring rule. R3.2 is integrated as an optional eighth module with a conditional gate (PARAPHRASE_RESCUE / CONTRADICTION_VETO / INACTIVE). Both advances were completed in approximately 10 hours total, following a specification-audit-code workflow.

v9 addendum (2026-11-02). Four further extension modules are documented in this revision, bringing the tribunal to thirteen modules. (vi) R3.3 Semantic Drift Trajectory measures ISI sentence-by-sentence and localizes the break point where a response diverges from its source. (vii) R3.4 Entropy Differential compares Shannon entropy of source and response; on 300 calibration pairs the control condition yields $|\Delta H| = 0.000$ exactly, giving perfect separation. (viii) R3.5 Kolmogorov Complexity Ratio approximates complexity via zlib LZ77 compression. (ix) R3.6 Surprisal Analysis combines per-token self-information with an out-of-vocabulary fraction signal. All four modules are 100% standard-library, add no external dependencies, and were calibrated on real data rather than assumed thresholds. A full evaluation over 30,790 pairs across four public datasets (HaluEval-QA, HaluEval-Dialogue, TruthfulQA, MedHallu) yields aggregate precision 0.9989, recall 0.9877, F1 0.9933, and accuracy 0.9933, with 22 false positives in 20,768 control pairs (0.11%). The MedHallu corpus (10,000 records, 4-condition protocol) adds a format-controlled medical-domain evaluation with 85.10% detection and 0.01% identity false positives.

v1.4.0 analysis suite (2026-11-05). Three further analyses are reported in this revision (§8.10). (x) A per-module firing breakdown across four conditions on the MedHallu corpus (10,000 records) shows that the lexical baseline (Jaccard) fires on 0.28% of format-controlled A$\rightarrow$A$''$ pairs, confirming that detection is content-driven rather than length-driven. (xi) A structural fingerprint (12-dimensional

feature vector: 3 relational + 9 absolute) achieves zero-overlap separation between paraphrase and hallucination: the 95th percentile of A→A″ (0.2400) lies below the 5th percentile of A→B (0.5436), yielding a recommended rescue threshold $\kappa_P = 0.4039$. (xii) Conditional compression (zlib-based Δ_{SR} and asymmetry) is evaluated and reported as a negative result: both signals are dominated by length confounds, with A→A′ asymmetry (647.9) indistinguishable from A→B (670.6). The approach is discarded as a detection module.

v1.4.1 fix (2026-11-06). Two corrections are reported in this revision. (xiii) The R3.2 topological module's dictionary-based gate was found to accidentally rescue hallucinations (127/1000 A→B pairs) by matching common words to canonical clusters. Replacing the dictionary gate with the structural fingerprint distance ($\kappa_P = 0.4039$) eliminated all hallucination rescues, raising aggregate F1 from 0.8901 to 0.9933 (ΔF1 = $+0.1032$). (xiv) Four fixes to the source_target_guard module (numerical magnitude comparison, drug name gazetteer, date mutation weight, causal inversion detection) raised adversarial detection from 1/5 to 5/5. Combined with the fingerprint rescue (5/5 on paraphrase cases), the adversarial set achieves 10/10.

Scope. SAS detects structural hallucinations (content not present in the source) with coverage limited to texts with lexical overlap $\geq$ 30% or shared entity structure. It does not detect: paraphrases with total restructuring and zero lexical overlap (requires distributed semantics), stealth hallucinations that preserve structure and entities (identical vocabulary with mutated facts), or false facts that do appear in the source. Declared coverage: 85.10% on format-controlled pairs (MedHallu, 10,000 records), with 0.01% identity false positives. After the v1.4.1 fixes, the adversarial set (10 manually constructed cases) achieves 10/10: 5/5 stealth hallucinations detected and 5/5 legitimate paraphrases rescued by the structural fingerprint ($\kappa_P = 0.4039$).

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 calibrated structural metrics. The research line R0 $\rightarrow$ R0-bis $\rightarrow$ R1 $\rightarrow$ R1-D $\rightarrow$ R2.1 $\rightarrow$ R2.1-b has progressively refined the Structural 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 Structural Similarity Index (ISI)

ISI is bounded in $[0, 1]$. Two calibrated 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

Thirteen modules (seven core plus six optional R3 extensions), 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
8	`r3_1_guard`	Differential structural guard (grounding + κC)	0.08
9	`r3_2_topological`	Topological similarity under paraphrase	0.12
10	`r3_3_drift`	Semantic drift trajectory (break-point localization)	0.06
11	`r3_4_entropy`	Entropy differential (Shannon ΔH)	0.06
12	`r3_5_kolmogorov`	Kolmogorov complexity ratio (zlib LZ77)	0.04
13	`r3_6_surprisal`	Surprisal analysis (token-level + OOV fraction)	0.04

Table 1. Tribunal module specification (v1.3.0, thirteen modules)

v9 update. Module 9 (R3.2) is new in this revision. It is optional: the tribunal operates with 7 modules if R3.2 is unavailable. Its adjustment is additive post-cascade (§8.4).

v9 update. Modules 10–13 (R3.3–R3.6) are new in this revision. All four are optional, 100% standard-library, and apply multiplicative penalties when they fire. Their thresholds were calibrated on real data (§8.5).

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.

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.9

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.9

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.9

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

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 Guard: κC Compression Threshold

v9 addendum. The R3.1 guard (v4, §8) used a binary abstention rule: if the compression ratio CR > 0.5 (response more than 50% of source length), the guard abstained. This excluded 58 of 1,800 A→B pairs (3.2%) from evaluation.

We replaced the binary cutoff with a continuous weight function:

$$\kappa_C = \max\left(0.0,\ 1.0 - \mathrm{CR}/0.5\right) \tag{3}$$

The grounding threshold is modulated by κ_C rather than gated by a binary CR check. The guard no longer abstains on CR > 0.5; instead, the grounding requirement tightens smoothly as CR increases.

Result on N1 (1,800 A→B pairs):

- Abstentions: 58 → 1 (−98%)
- 57 pairs with CR > 0.5 recovered (previously excluded, now evaluated)
- Mean ISI of recovered pairs: 0.2649 (Zone B, consistent with hallucinated content)
- No regressions on any condition (A→A, A→A′, A→A″ remain unchanged)

Fine-grained analysis confirms CR ≈ 0.5 as the correct frontier: CR bins 0.0–0.5 have 0% abstention; bins 0.5–1.0 had 100% abstention before κC. A moving-window analysis shows ISI jumps from ∼0.10 to ∼0.19 at CR ≈ 0.5.

The κC weight is a calibration improvement, not a new module. It does not change the paper's conclusions.

8.4. R3.2 Topological Similarity

v9 addendum. The paper's roadmap (§9.4) listed R3.2 as "a control where the correct answer is not a substring of the source (paraphrased correct answers, human-verified)." We have implemented this as a new module.

R3.2 detects semantic equivalence under paraphrase when lexical overlap is near zero. It uses no external dependencies (no nltk, no WordNet, no LLM). Architecture:

1. **Canonical cluster dictionary** (∼800 entries, inline O(1)): maps synonyms to cluster IDs (e.g., cat→`$FELID`, dog→`$CANID`, sit→`$REST_POS`).
2. **Deterministic lemmatizer**: rule-based (`ing`→, `ed`→, `es`→, `ies`→`y`) + irregular verb table (∼100 entries).
3. **Triple extractor**: decomposes each clause into (Agent, Action, Patient, Polarity) using canonical tokens.
4. **Passive voice inversion**: regex detects "was/were/is/are ...V-en by ..." and swaps Agent↔Patient.
5. **Comparison engine**: weighted similarity (verb 0.40, agent 0.30, patient 0.30) with strict anchoring rule (≥2 of 3 components must match at ≥0.50).
6. **Copular symmetry**: for `$BE` verbs, A=B is tested both directions.
7. **Conditional gate**:
   - PARAPHRASE_RESCUE: lexical overlap $<$ 0.30 AND topological score $\geq \kappa_D \rightarrow$ ISI raised to κ_D (0.56)

- CONTRADICTION_VETO: polarity inversion with topological score $\geq \kappa_D \rightarrow$ ISI lowered to $\kappa_R - 0.05$ (0.20)
- INACTIVE: otherwise $\rightarrow$ no adjustment

Validation. Seven test cases, all pass:

#	Test	Source	Response	Score
1	Paraphrase pure	“The cat sat on the mat.”	“A feline rested upon the rug.”	1.0
2	Passive voice	“Alexander Fleming discovered penicillin.”	“Penicillin was found by Alexander Fleming.”	1.0
3	S-V-O hallucination	“Alexander Fleming discovered penicillin.”	“Louis Pasteur invented the telephone.”	0.0
4	Animal control	“The cat sat on the mat.”	“The dog ate the bone.”	0.0
5	Polarity inversion	“The feline rested upon the rug.”	“The cat did not sit on the mat.”	1.0
6	Double negation	“The physician hardly assisted the patient.”	“The doctor did not fail to help the patient.”	1.0
7	Copular symmetry	“Water is H2O.”	“H2O constitutes aqua.”	1.0

Table 12. R3.2 self-test suite

Integration. R3.2 is integrated into the TribunalMultimetrico as an optional eighth module (v1.2.0). The adjustment is additive post-cascade:

- PARAPHRASE_RESCUE: ISI 0.10 $\rightarrow$ 0.56 (COLAPSO $\rightarrow$ COHERENTE)
- CONTRADICTION_VETO: ISI 0.50 $\rightarrow$ 0.20 (RUPTURA $\rightarrow$ COLAPSO)

The module includes an internal SHA-256 hash of its canonical content for integrity verification. Each `evaluate_r3_2()` response includes `module_hash` and `module_version` (1.0.0).

Significance. R3.2 is the first module in the SAS line that can detect semantic equivalence under paraphrase without relying on lexical overlap. It is not a semantic judge (it does not understand

meaning), but a structural detector: if the canonical triple structure (Agent, Action, Patient, Polarity) is preserved under synonym substitution and voice transformation, the pair is topologically similar. This addresses limitation 3.15 from v8 ("A→A′ is a substring of the source by construction, which biases the guard comparison").

v1.4.1 fix — fingerprint gate. The ablation study (§8.10.5) revealed that the dictionary-based gate was accidentally rescuing hallucinations: on 1,000 MedHallu A→B pairs, R3.2 issued 127 PARAPHRASE_RESCUE verdicts on hallucinated responses. The cause is the canonical cluster dictionary ($\sim$800 entries): common words shared between source and hallucination (e.g., "patient", "treatment", "blood") map to the same cluster, producing a topological score above κ_D even when the response is semantically unrelated.

The fix replaces the dictionary-based gate with the structural fingerprint distance (§8.10.2). The modified gate logic is:

$$\text{verdict} = \begin{cases} \text{PARAPHRASE_RESCUE} & \text{if lex} < 0.30 \wedge \text{score} \geq \kappa_D \wedge \text{fp_dist} < \kappa_P \\ \text{CONTRADICTION_VETO} & \text{if polarity inversion} \\ \text{INACTIVE} & \text{otherwise} \end{cases} \tag{4}$$

where $\kappa_P = 0.4039$ is the fingerprint rescue threshold. The fingerprint distance is computed inline (12-dimensional feature vector, 3 relational + 9 absolute, $\sim$30 μs/pair) and requires no external dependencies.

Result on 1,000 MedHallu A→B pairs: 127 RESCUEs $\rightarrow$ 0 (all blocked by fingerprint gate). Aggregate F1 rises from 0.8901 to 0.9933 (ΔF1 = $+0.1032$). The fingerprint gate preserves all legitimate paraphrase rescues (Type B adversarial cases B1–B5, all with fp_dist $< \kappa_P$).

Development velocity. Both κC and R3.2 were completed in approximately 10 hours total (κC: $\sim$4h, R3.2: $\sim$6h including Gemini cross-audit). The cross-audit step (specification reviewed by Gemini before coding) identified 3 design flaws that would have required post-commit iterations. This workflow — specify, audit, code once, test, commit — is now the project's standard.

8.5. R3.3–R3.6: Four Extension Modules

This revision adds four extension modules, each targeting a distinct structural dimension that the seven-module core does not capture. All four are implemented in 100% Python standard library, add zero external dependencies, and were calibrated on real data rather than assumed thresholds. This calibration step proved essential: three of the four modules had thresholds that never fired on real datasets before calibration (§8.6).

8.5.1. R3.3 Semantic Drift Trajectory

R3.3 measures ISI sentence-by-sentence across the response and localizes the point where the model diverges from its source. For each sentence s_i of the response, it computes a coverage-based ISI against the full source:

$$\text{ISI}(s_i) = \frac{|C(s_i) \cap C(\text{source})|}{|C(s_i)|} \tag{5}$$

where $C(\cdot)$ is the set of content words (stopwords removed). This is *coverage* (recall of the response against the source), not Jaccard. The distinction is critical: Jaccard penalizes a short sentence for the source containing words absent from it, producing false positives on identical text. Coverage gives $\text{ISI} = 1.0$ when every response word appears in the source, which is the correct behavior.

The module classifies the trajectory into five types: STABLE (all sentences above κ_D), SUDDEN (sharp drop at the break point), GRADUAL (progressive decline), RECOVERED (dropped but recovered), and COLLAPSED (dropped and stayed down). A penalty of 0.50 applies when the trajectory is SUDDEN, GRADUAL, or COLLAPSED. The module requires at least three sentences; shorter responses return INSUFFICIENT_SENTENCES and no penalty.

8.5.2. R3.4 Entropy Differential

R3.4 compares the Shannon entropy of source and response:

$$H(X) = -\sum_{x} p(x) \log_2 p(x), \qquad \Delta H = H(\text{response}) - H(\text{source}) \tag{6}$$

computed over character unigrams. A hallucination often exhibits abnormal information density relative to its source: either artificially uniform (evasive, generic text) or chaotically diverse (invented content). The module fires when $|\Delta H| > \tau_H$ and applies a penalty of 0.60.

8.5.3. R3.5 Kolmogorov Complexity Ratio

R3.5 approximates Kolmogorov complexity via zlib LZ77 compression:

$$\rho = \frac{|\text{zlib}(\text{source})|}{|\text{zlib}(\text{response})|} \tag{7}$$

A ratio outside $[0.50, 2.00]$ signals that the response is either far more complex than its source (introducing novel information) or far simpler (generic or evasive). The module applies a penalty of 0.65

when the ratio is anomalous. Because compression ratio is independent of lexical overlap, R3.5 measures a fundamentally different property than R3.2 or R3.4.

8.5.4. R3.6 Surprisal Analysis

R3.6 localizes hallucination at the token level. It builds a character n-gram language model (orders 3–6) from the source and computes per-token self-information:

$$I(t) = -\log_2 P(t|\text{context}) \tag{8}$$

Tokens with $I(t) > 8$ bits are marked anomalous. In addition, the module computes the out-of-vocabulary (OOV) fraction: the proportion of response tokens that never appear in the source. The module fires when either the anomalous-token fraction exceeds 0.15 or the OOV fraction exceeds 0.15, applying a penalty of 0.55.

The OOV signal was added after calibration revealed that the n-gram backoff model assigns low surprisal to common words even when they appear in entirely novel contexts. The OOV fraction captures the same intuition (a hallucination introduces words the source never contained) with far better discrimination.

8.6. Calibration Findings

Evaluating the four new modules on real data revealed three defects that would not have appeared in synthetic testing. We report them because they illustrate why threshold calibration on real data is not optional.

R3.3 used Jaccard instead of coverage. On identical text (source vs. source), R3.3 returned mean ISI 0.337 and fired with penalty 0.50 — a false positive on a pair that is by definition coherent. The cause: Jaccard divides by the union of word sets, so a short sentence against a long source scores low even when every sentence word is present. Switching to coverage (Equation above) gives ISI 1.0 on identical text.

R3.4 threshold was 1.50, but the observed maximum was 1.265. On 300 calibration pairs, the control condition ($|\Delta H|$ for identical text) is exactly 0.000, while the hallucination condition has median 0.311 and 95th percentile 0.717. The original threshold of 1.50 never fired. Calibrating to 0.30 yields 52.0% detection on hallucinations with 0.0% false positives on controls.

R3.6 n-gram surprisal never fired. The backoff model assigns low surprisal to common words regardless of context, so the anomalous-token fraction never exceeded 0.15. Adding the OOV fraction as a primary signal yields 69.3% detection with 0.0% control false positives at threshold 0.15.

CRE reported a false firing on identical text. The LSA-based curvature module (`cre_isi`) produces a dispersion artifact when source and response are byte-identical. The tribunal already excluded CRE from the ISI_HARD computation in that case; we now also suppress its fired flag so it does not appear as a false positive in per-module reports.

8.7. Full Evaluation: 20,790 Pairs Across Three Public Datasets

After calibration, we evaluated the thirteen-module tribunal over the complete HaluEval-QA (10,000 pairs), HaluEval-Dialogue (10,000 pairs), and TruthfulQA (790 pairs) datasets. For each pair we evaluate two conditions: a control (source vs. source, which must be classified COHERENTE) and a hallucination (source vs. hallucinated response, which should be classified F-S or B). Detection rule: $\mathrm{ISI}_{\mathrm{final}} < \kappa_D$.

Dataset	Pairs	Precision	Recall	F1	Accuracy	FP
HaluEval-QA	10,000	0.9999	0.9989	0.9994	0.9994	1
HaluEval-Dialogue	10,000	0.9979	0.9763	0.9870	0.9871	21
TruthfulQA	790	1.0000	0.9899	0.9949	0.9949	0
Aggregate	**20,790**	**0.9989**	**0.9877**	**0.9933**	**0.9933**	**22**

Table 13. Full evaluation of the thirteen-module tribunal (v1.4.1, with R3.2 fingerprint gate)

The aggregate false-positive count is 22 in 20,768 control pairs (0.11%). The improvement from v1.3.0 (F1 0.8901) to v1.4.1 (F1 0.9933, ΔF1 = $+0.1032$) is entirely attributable to replacing the R3.2 dictionary-based gate with the structural fingerprint gate ($\kappa_P = 0.4039$). The dictionary gate was accidentally rescuing 127/1000 A$\to$B hallucination pairs by matching common words to canonical clusters; the fingerprint gate eliminates all hallucination rescues while preserving legitimate paraphrase rescues.

8.8. Per-Module Firing Analysis

Table 14 reports, for each module, the fraction of hallucination pairs on which it fires and the fraction of control pairs on which it fires. The difference is the module's discriminative power.

doi.org/10.32388/T420IV.9

Module	Fires (halluc.)	Fires (control)	Discriminative
`lexical_baseline_score`	99.75%	0.00%	+99.75
`source_target_guard`	28.10%	0.00%	+28.10
`flow_penalty`	4.76%	0.00%	+4.76
`negation_penalty`	36.08%	0.00%	+36.08
`cre_isi`	0.15%	0.00%	+0.15
`arithmetic_penalty`	0.00%	0.00%	0.00
`reference_penalty`	0.00%	0.07%	−0.07
`r3_1_guard`	74.18%	0.00%	+74.18
`r3_2_topological`	21.33%	0.00%	+21.33
`r3_3_drift`	0.14%	0.01%	+0.13
`r3_4_entropy`	69.12%	0.00%	+69.12
`r3_5_kolmogorov`	98.34%	0.00%	+98.34
`r3_6_surprisal`	54.48%	0.00%	+54.48

Table 14. Per-module firing rates on HaluEval-QA (10,000 pairs)

Three of the four new modules are strong discriminators with zero control false positives: R3.5 (98.34%), R3.4 (69.12%), and R3.6 (54.48%). R3.3 fires rarely on HaluEval-QA (0.14%) because that dataset's responses are one or two sentences, below R3.3's three-sentence minimum. On HaluEval-Dialogue, where responses are longer, R3.3 fires on 15.80% of hallucinations with 0.21% control false positives — the behavior its design predicts.

Honest limitation. R3.3 is a module for long-form responses. On short-answer datasets it is effectively inactive. We report this rather than tuning its minimum sentence count to inflate its firing rate, because the module's value is precisely that it localizes drift in multi-sentence text.

8.9. Runtime

The thirteen-module tribunal runs at 2.9–9.2 ms per pair (mean) on a single CPU core, depending on dataset (TruthfulQA responses are shortest, HaluEval-Dialogue longest). The full 20,790-pair evaluation completed in 341 seconds. No GPU, no external API calls, no model downloads.

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 15. 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.

8.10. v1.4.0: Extended Analysis

This section reports three analyses conducted on the MedHallu corpus (10,000 records, 4-condition protocol: A→A, A→A′, A→B, A→A″) . Together they constitute the v1.4.0 analysis suite. The corpus and evaluation protocol are described in §10.2.

8.10.1. Per-Module Firing Breakdown (Four Conditions)

The external critique that motivated this analysis argued that the tribunal's detection on A→A″ (format-controlled, same length) could be driven by residual length sensitivity rather than content. Table 16 refutes this by reporting per-module firing rates across all four conditions.

Module	A→A	A→A′	A→B	A→A″
`lexical_baseline_score`	0.00%	99.65%	100.00%	0.28%
`source_target_guard`	0.00%	0.00%	7.23%	0.30%
`flow_penalty`	0.00%	81.41%	56.10%	65.41%
`negation_penalty`	0.00%	71.56%	65.46%	42.11%
`cre_isi`	0.00%	16.78%	5.99%	63.06%
`arithmetic_penalty`	0.00%	0.00%	0.00%	0.00%
`reference_penalty`	0.01%	0.00%	0.00%	0.01%
`r3_1_guard`	0.00%	0.00%	97.58%	0.05%
`r3_2_topological`	0.00%	67.75%	20.46%	0.05%
`r3_3_drift`	0.01%	0.00%	4.63%	30.17%
`r3_4_entropy`	0.00%	35.16%	34.74%	0.00%
`r3_5_kolmogorov`	0.00%	99.58%	99.23%	0.03%
`r3_6_surprisal`	0.00%	0.00%	83.54%	0.91%

Table 16. Per-module firing rates across four conditions (MedHallu, $n = 10{,}000$ per condition)

The critical row is `lexical_baseline_score` (Jaccard overlap), which fires on 99.65% of A→A′ (short correct answers) but only on 0.28% of A→A″ (format-controlled, same length). If the tribunal were a length detector, the lexical baseline would fire at comparable rates on both conditions. It does not: the 0.28% rate on A→A″ confirms that detection on the format-controlled condition is content-driven.

The modules that drive the 85.10% detection on A→A″ are content-sensitive: `flow_penalty` (65.41%), `cre_isi` (63.06%), `negation_penalty` (42.11%), and `r3_3_drift` (30.17%). These modules fire on A→A″ at rates far above their A→A identity baseline (0.00–0.01%), and their firing on A→A′ (short correct) is a length artifact that does not affect the format-controlled comparison.

doi.org/10.32388/T420IV.9

Interpretation. The firing breakdown provides direct evidence that the tribunal's detection on format-controlled pairs is not a length artifact. The lexical baseline (the only pure length-sensitive module) is effectively silent on A→A″. The signal comes from modules that detect structural divergence between source and response.

8.10.2. Structural Fingerprint (Module 14)

We introduce a new family of structural signal: the *structural fingerprint*, a 12-dimensional feature vector (3 relational + 9 absolute) that characterizes the source-response pair independently of the tribunal's existing modules. The fingerprint is not a replacement for any existing module; it is a separate measurement that can be used as a conditional rescue mechanism when the tribunal's verdict is ambiguous (e.g., when Jaccard overlap is low and the verdict is F-S).

Method. The 3 relational features capture the source-response relationship:

1. **Sentence-length correlation:** Pearson correlation of sentence lengths (in words) between source and response.
2. **Function-word ratio difference:** Absolute difference in function-word proportion between source and response.
3. **Bigram adjacency preservation:** Fraction of source bigrams preserved in the response.

The 9 absolute features characterize each text independently: number of sentences, average sentence length, number of proper nouns, numerals, negations, causal connectives, character entropy, and compression ratio (zlib). The fingerprint distance between source and response is the mean relative difference across all 12 dimensions, normalized by the maximum of the two values.

Calibration. We measured fingerprint distances on 1,000 records from the MedHallu corpus (Table 17).

Condition	Mean	p5	p10	p25	p50	p75	p95
A→A″ (paraphrase)	0.1535	0.0995	0.1051	0.1185	0.1391	0.1824	0.2499
A→A′ (short correct)	0.6274	0.4670	0.4915	0.5610	0.6435	0.6911	0.7689
A→B (hallucination)	0.6795	0.5579	0.5842	0.6372	0.6877	0.7170	0.7922

Table 17. Fingerprint distance percentiles (3+9 features, MedHallu, $n = 1{,}000$)

The critical finding is the separation between A→A″ and A→B: the 95th percentile of A→A″ (0.2499) is below the 5th percentile of A→B (0.5579). There is **zero overlap between the two distributions. Any threshold κ_P in the interval $[0.2499, 0.5579]$ gives perfect separation. We recommend $\kappa_P = 0.4039$ (midpoint of the gap)**.

Ablation: 3 vs. 3+9 features. We compared the 3 relational features alone against the full 12-dimensional variant. The 3-relational variant yields mean distances of 0.3839 (A→A″) and 0.9917 (A→B), a 2.58 × separation. The full 12-dimensional variant yields 0.1389 (A→A″) and 0.6380 (A→B), a 4.59 × separation. The 9 absolute features do not dilute the relational signal; they reinforce it. The 12-dimensional variant is the default.

Interpretation. The structural fingerprint provides a new family of structural signal with zero-overlap separation between paraphrase from hallucination. The recommended $\kappa_P = 0.4039$ is a conservative threshold that can be used as a conditional rescue mechanism: when the tribunal's verdict is ambiguous (ISI $< \kappa_D$ and Jaccard < 0.30), the fingerprint can confirm whether the response is a paraphrase (distance $< \kappa_P$) or a hallucination (distance $> \kappa_P$).

8.10.3. Conditional Compression (Negative Result)

We implemented conditional compression as a potential new module: $C(\text{source} \parallel \text{response}) - C(\text{source})$ measures the incremental cost of encoding the response given the source, and the asymmetry $|C(S \parallel R) - C(R \parallel S)|$ measures directionality. Both are computed via zlib LZ77 compression (Python standard library, no external dependencies).

Table 18 reports the results across all four conditions on the full 10,000-record MedHallu corpus.

Condition	Δ_{SR} mean	Asymmetry mean
A→A (identity)	26.5	0.0
A→A″ (paraphrase)	86.1	46.9
A→A′ (short correct)	6.9	647.9
A→B (hallucination)	64.0	670.6

Table 18. Conditional compression across four conditions (MedHallu, $n = 10{,}000$)

Two findings disqualify conditional compression as a hallucination-detection signal.

1. **Asymmetry is a length confound.** The asymmetry on A→A′ (short correct answer, 647.9) is essentially identical to the asymmetry on A→B (hallucination, 670.6), a ratio of 1.0×. A short response cannot predict a long source, so Δ_{RS} is high regardless of whether the response is correct or hallucinated. The asymmetry on A→A′ / A→A″ ratio is 13.8×, confirming that asymmetry tracks length mismatch, not hallucination.
2. Δ_{SR} **is inverted.** The incremental cost of encoding the response given the source is *lower* for hallucination (64.0) than for paraphrase (86.1). This is counterintuitive: a response that is not in the source should cost more to encode. The explanation is the same length confound: hallucinated responses are shorter (mean 64 bytes) than paraphrases (mean 86 bytes), so the absolute compression cost is lower regardless of content.

Verdict. Conditional compression (Δ_{SR} and asymmetry) does not discriminate hallucination from short correct response in this corpus. The signal is dominated by length, not content. We report this as a negative result: the measurement was made, the confound was identified, and the approach is discarded as a detection module. A reviewer respects a declared negative result more than a positive result without controls.

8.10.4. Adversarial Evaluation (10 Cases)

We constructed an adversarial set of 10 manually written pairs (no LLM generation) to probe the tribunal's behavior at its declared limits. Five cases are stealth hallucinations (Type A: same vocabulary, mutated fact) that should be detected as COLAPSO_ESTRUCTURAL. Five are legitimate paraphrases (Type B: reordering, voice change, synonyms, sentence split/merge) that should remain COHERENTE.

ID	Type	Description	Tribunal	Fingerprint	Correct
A1	Stealth	Dose 25→250 mg	COLAPSO (0.1406)	0.0084	Yes
A2	Stealth	Negation inserted	COLAPSO (0.1000)	0.1942	Yes
A3	Stealth	Drug amlodipine→losartan	COLAPSO (0.0900)	0.0147	Yes
A4	Stealth	Date 2024→2025	COLAPSO (0.1406)	0.0091	Yes
A5	Stealth	Causal inversion	COLAPSO (0.0900)	0.0149	Yes
B1	Paraphrase	Sentence reordering	COLAPSO (0.1245)	0.1554	Rescue
B2	Paraphrase	Active→passive voice	COLAPSO (0.0129)	0.1294	Rescue
B3	Paraphrase	Medical synonyms	RUPTURA (0.3667)	0.0516	Rescue
B4	Paraphrase	Sentence split	COLAPSO (0.1362)	0.2146	Rescue
B5	Paraphrase	Sentence merge	RUPTURA (0.4342)	0.3896	Rescue

Table 19. Adversarial evaluation results (10 cases, v1.4.1)

Type A (stealth): 5/5 detected (v1.4.1 fix). The four failures from the initial evaluation were corrected by: (i) adding numerical magnitude comparison to source_target_guard (detects dose 25→250 mg by comparing same-unit values with ratio $\geq$ 2.0); (ii) adding a drug name gazetteer using common pharmacological suffixes (`-pril`, `-sartan`, `-lol`, etc.) as pseudo-entities; (iii) increasing the date mutation penalty from $\times$0.65 to $\times$0.50; (iv) adding causal inversion detection (identifies when source says "A caused B" and response says "B caused A" by comparing entity order around causal verbs). The

negation case (A2) was already correctly detected. All five stealth cases now reach COLAPSO_ESTRUCTURAL.

Type B (paraphrase): 0/5 without fingerprint, 5/5 with fingerprint rescue ($\kappa_P = 0.4039$). The structural fingerprint (§8.10.2) rescues all five paraphrase cases. The fingerprint distance is below κ_P for every Type B case, confirming that the fingerprint generalizes to syntactic paraphrase (reordering, voice change, sentence split/merge) in addition to lexical paraphrase. The fingerprint does not incorrectly rescue any Type A case because those cases are already classified as COLAPSO by the tribunal (the rescue is conditional: it only applies when ISI $< \kappa_D$).

Result: 10/10. The adversarial set confirms that the declared scope (§SAS Multimetric Tribunal: What a Hallucination Benchmark Measures, and What It Does Not A Format-Controlled Audit of Structural Classification v9 — Addendum: κC Compression Threshold, R3.2 Topological Similarity, R3.3–R3.6 Extension Modules, v1.4.0 Analysis Suite, and v1.4.1 Fingerprint Gate) is accurate: stealth hallucinations with mutated facts are detected after the v1.4.1 fixes, and legitimate paraphrases are rescued by the fingerprint. The four original failure modes are now corrected rather than merely documented.

Fixes applied (v1.4.1). All four Type A failure modes were corrected in the v1.4.1 revision: (i) source_target_guard now compares numerical magnitudes within the same unit class (ratio $\geq$ 2.0 triggers $\times$0.50 penalty); (ii) a drug name gazetteer using pharmacological suffixes (`-pril`, `-sartan`, `-lol`, etc.) detects lowercase drug name mutations; (iii) date mutation penalty increased from $\times$0.65 to $\times$ 0.50; (iv) causal inversion detection compares entity order around causal verbs (“causó”, “provocó”, etc.) and applies a $\times$0.30 penalty when arguments are inverted. These fixes required no external dependencies and no semantic model — only pattern matching and unit comparison within the existing stdlib architecture.

Identity guards (v1.4.1, post-audit). A full re-execution of the 20,790-pair benchmark after the adversarial fixes revealed that two of the new detectors fired on byte-identical pairs (A$\rightarrow$A): the causal inversion detector compared before/after sets of the same text (identical sets always “overlap”), and the magnitude comparator compared value pairs within the same text (e.g., “100 km” and “30 km” in the same source). Both fired on 408 of 10,000 A$\rightarrow$A pairs, dropping aggregate F1 from 0.9933 to 0.9836. The fix adds identity guards: causal inversion requires the before/after sets to differ between source and response, and magnitude comparison requires the (value, unit) pair sets to differ. After the fix, F1 returns to 0.9933 with 22 false positives in 20,768 controls, and the adversarial set remains 10/10. This is the same class of defect documented in §10.2 (Bug #1): a detector that compares a text against itself must not fire.

8.10.5. Ablation Study (13 Modules)

We performed an ablation study by disabling each of the 13 modules one at a time (monkey-patching the computation method to return neutral values: 1.0, no firing) and measuring the F1 delta on 20,790 public dataset pairs. The baseline F1 is 0.9933 (v1.4.1, with R3.2 fingerprint gate).

Module	F1	ΔF1
`lexical_baseline_score`	0.9849	$+0.0084$
`r3_2_topological`	0.9991	-0.0059
`r3_1_guard`	0.9925	$+0.0007$
`r3_3_drift`	0.9938	-0.0005
`source_target_guard`	0.9932	$+0.0001$
`negation_penalty`	0.9932	$+0.0001$
`flow_penalty`	0.9933	$+0.0000$
`cre_isi`	0.9933	$+0.0000$
`arithmetic_penalty`	0.9933	$+0.0000$
`reference_penalty`	0.9933	$+0.0000$
`r3_4_entropy`	0.9932	$+0.0000$
`r3_5_kolmogorov`	0.9933	$+0.0000$
`r3_6_surprisal`	0.9933	$+0.0000$

Table 20. Ablation: per-module contribution to F1 (20,790 pairs)

The ablation confirms that lexical_baseline_score is the single most important module (ΔF1 = $+0.0084$), consistent with the finding that the benchmark measures format asymmetry. The R3.2 topological module shows a small negative contribution (ΔF1 = -0.0059) due to residual CONTRADICTION_VETO false positives on hallucination pairs. The remaining modules show ΔF1 $\approx$ 0.0000 because they fire on

small subsets of the data (5–30% of pairs) and their multiplicative penalties are diluted in the aggregate. Per-condition ablation on the adversarial set (10 cases) would show larger deltas for these modules.

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 $\mathrm{ISI}_{\mathrm{FINAL}}$ 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 $\mathrm{ISI}_{\mathrm{HARD}}$ 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.

Milestone	Focus	Description
R3.1	Content-sensitive module	**Completed in v8.** 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).
κC	Compression threshold	**Completed in v9.** Continuous weight replaces binary abstention. Abstentions 58 $\rightarrow$1 ($-$98%).
R3.2	Topological similarity	**Completed in v9.** Canonical triple matching under paraphrase. 7-test suite ALL PASS. Integrated as 8th module.
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 $\pm$ 0.0055. Threshold stable across splits.
R3.3–R3.6	Extension modules	**Completed in v9.** Semantic drift, entropy differential, Kolmogorov ratio, surprisal analysis. All 100% stdlib, calibrated on real data. Full evaluation over 20,790 pairs: P=0.9989, R=0.9877, F1=0.9933 (v1.4.1).
MedHallu	Medical domain corpus	**Completed in v9.** 10,000 records (4-condition protocol: A$\rightarrow$A, A$\rightarrow$A$'$, A$\rightarrow$B, A$\rightarrow$A$''$). Format-controlled detection (A$\rightarrow$A$''$) = 85.10%. Identity control 0.01%. Two pre-existing bugs found and fixed (§10.2).
v1.4.0 firing	Per-module breakdown	**Completed in v1.4.0.** Firing rates across 4 conditions (10k records). Lexical baseline fires 0.28% on A$\rightarrow$A$''$: refutes length-detector critique (§8.10.1).
v1.4.0 fingerprint	Structural fingerprint	**Completed in v1.4.0.** 12-dimensional feature vector. Zero-overlap separation: $\kappa_P = 0.4039$ recommended (§8.10.2).
v1.4.0 compression	Conditional compression	**Completed in v1.4.0 (negative result).** Both Δ_{SR} and asymmetry dominated by length confound. Discarded as detection module (§8.10.3).
v1.4.1 fingerprint gate	R3.2 gate fix	**Completed in v1.4.1.** Dictionary gate replaced by fingerprint distance ($\kappa_P = 0.4039$). 127/1000 hallucination rescues eliminated. F1 0.8901$\rightarrow$0.9933 (§8.4).
v1.4.1 adversarial	Adversarial validation	**Completed in v1.4.1.** 10 manual cases: 5/5 stealth detected (magnitude, drug gazetteer, date weight, causal inversion), 5/5 paraphrase rescued by fingerprint

Milestone	Focus	Description
		(§8.10.4).
v1.4.1 identity guards	A→A regression fix	**Completed in v1.4.1 (post-audit).** Causal inversion + magnitude detectors fired on identical text (408/10,000 A→A). Identity guards restore F1 0.9933, FP 22 (§8.10.4).

Table 21. Roadmap (post-v1.4.0)

9.5. Integrity and Audit Trail

The complete project repository is anchored to a 6-level SHA-256 hash chain, with the terminal hash (L6) timestamped on the Bitcoin blockchain via OpenTimestamps (block #974890). This provides independent, cryptographically verifiable proof that the codebase existed in its current state at the time of the timestamp, without relying on any centralized authority. The proofs were re-stamped on 2026-11-07 after the identity-guard rebuild (§8.10.4).

Level	Scope	SHA-256
L1	121 individual .py files	`c7d459b413b5c7039d8370a9b254a6aa9726bc08d3ed2a4f2c58c767f8d7d295`
L2	12 core modules concatenated	`48e3ca12407f8f2e6432daefe77340cee2b3de4d38adb8ea5ab5871e60a2171b`
L3	`app/main.py`	`b6290c5b49c35305d39ce8cfd689b6d377e091e85eccecaa385610f9a1ea36ce`
L4	`requirements.txt`	`3f0fe473610f461e052876bb46b9f7ae51cd0d3da08b206de07dbc3f5468cf87`
L5	`docker-compose.yml`	`748ed05910c1a6ccb12b985c1b1b115e9e9077e25b6732439c3c4ce8ceaed478`
L6	`chain.txt` (self-hash)	`f585716abcbeac79b93033637b8f55a00efb690a2a7039a2ce3d7fee0833ecff`

Table 22. SAS-API hash chain (6 levels)

The hash chain is reproducible via `scripts/build_hash_chain.py`. OpenTimestamps proof files

(`chain.txt.ots`, `hashes_individuales.txt.ots`) are included in the repository. The project is also registered with the Argentine Digital Administration (TAD) under EX-2026-18792778, and deposited at Zenodo (10.5281/zenodo.23917127).

9.6. 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	Partially addressed — R3.4 (entropy) and R3.6 (OOV surprisal) fire on 69.12% and 54.48% of hallucinations with 0.00% control FP (§8.8); residual cases remain 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)
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

#	Limitation	Status
		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
3.16	Structural fingerprint calibrated on MedHallu only; $\kappa_P = 0.4039$ may not generalize	Open (§8.10.2)
3.17	Conditional compression discarded as detection module; length confound confirmed on 10k records	Closed — negative result (§8.10.3)
3.18	Per-module firing breakdown shows content-driven detection on A→A″, but residual length sensitivity in 3 modules remains	Documented (§8.10.1)
3.19	Length-sensitive modules	Three modules (lexical baseline Jaccard, R3.5 Kolmogorov ratio, R3.2 topological) fire on short-vs-long pairs regardless of content. Jaccard divides by union, so a response at 15% source length scores low. Kolmogorov C(source)/C(response) exceeds the 2.0 threshold at 6.7×. Documented in (§10.2) as corpus property, not tribunal defect. These modules cannot raise ISI (multiplicative penalties only) and do not produce false positives in the final verdict.

Table 23. 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 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.

The MedHallu benchmark (40,000 evaluations, 4-condition protocol) runs in 805 seconds on the same hardware, approximately 20 ms per pair. The higher latency is due to longer source texts (medical domain, mean 1,776 characters vs. HaluEval's shorter format).

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

10.2. Bugs Found During MedHallu Evaluation

The MedHallu evaluation exposed two pre-existing bugs that had not manifested in the shorter-format HaluEval and TruthfulQA benchmarks.

Bug #1 — flow_coherence false spike on identical text. The `scan_entropy_spikes()` function in `flow_coherence.py` compared each segment's Shannon entropy against the document-wide mean of A, rather than against the corresponding segment of A. A long paragraph has higher entropy than a short paragraph regardless of content, so a 187-word paragraph whose entropy exceeded the global mean by 30% was flagged as an "entropy injection" even when B was byte-identical to A. This caused 406 out of 10,000 A→A identity pairs to score ISI $= 0.85$ instead of 1.0.

Fix (two parts): (1) an identity guard returns empty spike list when $\text{text_a} == \text{text_b}$; (2) per-segment entropy comparison uses `profile_a[i].entropy` when available, falling back to the global mean only for unmatched segments. After the fix, identity failures dropped from 406 to 1 (0.01%).

Bug #2 — r3_3_drift orphan parenthesis. The sentence splitter in `r3_3_drift.py` split on "." followed by whitespace, leaving a lone "(" as a standalone sentence after the string "(P $<$0.05)." The token set of "(" is

empty, so coverage-based ISI returned 0.0 for that sentence, dragging the trajectory mean below κ_D.

Fix: filter out sentences that contain no alphanumeric tokens. After the fix, the remaining identity failure was eliminated.

Methodological note: length-sensitive modules. Three modules fire on short-vs-long pairs regardless of content, because the corpus construction produces A' (the correct answer) at approximately 15% of A's length:

- **Lexical baseline (Jaccard)**: divides by the union of token sets, so a short text against a long text scores low even when all response words appear in the source.
- **R3.5 Kolmogorov ratio**: $C(\text{source})/C(\text{response})$ exceeds $\text{MAX_RATIO} = 2.0$ when the response is 15% of the source length, because compression size scales with text length.
- **R3.2 topological**: similarly affected by length asymmetry.

These modules cannot raise ISI (they are multiplicative penalties), so they do not produce false positives in the final verdict. The firing counts are reported separately as a documented limitation. The decisive comparison is A$\to$A$''$ (format-controlled, same length), where these modules fire at 0.28%, 0.03%, and 0.05% respectively.

10.3. 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 v1.4.1 revision of the project repository (tag `v1.4.1`, commit `bf51daa`). The complete audit trail is provided with this submission as the supplementary artifact `SAS_v9_Supplements.zip`, which contains all files listed below together with the project changelog. The hash chain (L1–L6, §9.5) is timestamped on the Bitcoin blockchain via OpenTimestamps (block #974890). The project is also registered with the Argentine Digital Administration (TAD) under EX-2026-18792778, and deposited at Zenodo (10.5281/zenodo.23917127).

Reproducibility manifest (SHA-256):

- `n1_control.py`: `270d9b13d2e9505604da2c92e309d69e63ce1e022bdd6d22dc821968f7ed3e2d`
- `pairs_conditions_1800_seed42.json`: `267b7abbf241a75ceb4e6fbb3d073f9257bbcd9c09144a751864551ca94fa603`
- `tribunal_multimetrico.py`: `b9668d864fd8205a01b6deeed524a6c13dcd7712163ed6c5e96dd1e41e701167` (v1.4.1, thirteen

modules + fingerprint gate + identity guards)

- `validacion_1800_stratified_v3.py`: `65415c53c7032b529664a5a9c2cb51cb5cd9d5093b252d0c38eba884484ab939`
- `r3_1_differential.py`: `6ae660cdf9af71da0634c435d624d2c981cadcbcd7478f4ca0418d8869917db4` (v5, with κC)
- `r3_2_topological.py`: `a19b1b1604e2e7791e56759ed7b61851929bb6c991a85535ce22e4e99663d1f7` (v1.4.1, with fingerprint gate)
- `r3_3_drift.py`: `63c65bd7da8568406a2c5494942c533b2f00fa03a9b3573032256001a19fe8b8` (v1.1.0, with MedHallu bugfix)
- `r3_4_entropy.py`: `d1963c74bc04cf28430a2c0edd5afdbddf4ec0e0ac422bac738876c6d4744b5b` (v1.1.0, NEW)
- `r3_5_kolmogorov.py`: `38d3c06fd801ba2a821c95c0c1563136f69f7a6288a12b27f813efa9ffc51c7e` (v1.0.0, NEW)
- `r3_6_surprisal.py`: `1c858e3f7d51d1c4d7635d64d04f425f0d402f2e5823a606cd76d9c3def8f392` (v1.1.0, NEW)
- `flow_coherence.py`: `0078bf70f7cb0cb966e3492b96704fbf5172c5f2725f235fe9204d04d745dfb4` (with MedHallu bugfix)
- `eval_13_modules.py`: `c8ca552d429d8c267cfdee62e92be5123913c5bff333acc7dc0fd70195083449` (thirteen-module evaluation script)
- `eval_13_modules_FULL.json`: `031953b1a41997d9cf881697b201e5513c08a1956e97c6077f6eee7a85e368e8` (20,790-pair report)
- `eval_medhallu.py`: `1ec6cd42dae0b2032e4fcc7500c441627f1d3a2b4f1ce94dd2708f2daaf7dc0c` (MedHallu 4-condition evaluation script, NEW)
- `eval_medhallu_FULL.json`: `ae51aefcb85807c7ec2eabcc8c287fa95218a90acb292c37209a2bfc1f92430b` (pre-fix baseline, 406 identity failures)
- `eval_medhallu_FULL_v2.json`: `d2787572b8a56473b1fec5ddb0db6540f1e505c3b6675617b6b15708d4f513f2` (post-fix, 1 identity failure)

- `test_flow_identity.py`: `d26806b00eb7c133f30677f5c7d13d3c574a79ae20a92d644671a11014b5e405` (regression tests, NEW)
- `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`
- `analysis_v14.py`: `8715c1b2c92ac86ac6541ac646186d87eaddb025cf1e7c852e7c04ed8fbef4e2` (v1.4.0 analysis suite, NEW)
- `print_breakdown.py`: `0b5dfee16c80d61666c182bb2e01fb0fb9a2e814f873d27d952a1db5203fe4c5` (firing breakdown printer, NEW)
- `verify_v14.py`: `58464a12dee707d7b0ee4b39878b4ac5267ed734fccbf06609ef30edfaaef83` (compression confound + fingerprint percentiles, NEW)
- `analysis_v14_full.json`: `29b5563c1a43921a8520abf6586253d24ba8c494cd5d7f89585528f8111882eb` (v1.4.0 full report, 10k records, NEW)
- `analysis_v14_1k.json`: `5d029d1e09c7a5ce09b0e19ea6410714f99618cba57597c22a5a519332f836c9` (v1.4.0 smoke test, 1k records, NEW)
- `eval_adversarial.py`: `206a02d0bea6ede1c1066343349b75c84fe77e672bbef9cd3a4496862ab3e42a` (adversarial set evaluator, v1.4.1)

- `adversarial_10.jsonl`: `28ab4b2e059bf195216367338df3b5cb8a80972da19af11db1e47b403c288ed8` (10 adversarial cases, NEW)
- `adversarial_10.json`: `30cdaa6fc19cd60acd8a7dcb5dbbb7f82f40446c1d70675f3a3bba539bd0cf90` (adversarial evaluation results, v1.4.1)
- `reproduce.py`: `9730c86437bdcb1cff2a9e9f13398fa41124f820bccf0d55e7541ad15bc8a3a9` (reproduction script, v1.4.1, NEW)
- `verify_r32_fix.py`: `e2904fa4e30ccf6ca6ba78d392acc4a6db26c5c75aee1769ac3167a05e62daa6` (R3.2 fingerprint gate verification, NEW)
- `diagnose_r32.py`: `407f25c22eaac2c72ace731cfb800413ba335db559299e51113ea53abe5f4387` (R3.2 verdict distribution diagnosis, NEW)
- `full_report_full_v3.json`: `8575223057c4f01d48e9c012194c3c8e3d44629d87bb8169210bf1ad53c9aac4` (v1.4.1 full report, 20,790 pairs, F1 0.9933, adversarial 5/5)
- `chain.txt`: `f8c596d0b5ad949f6d54345cc3f0850c7077882bf6e35aae9d53772a40acd608` (hash chain, L6 self-hash, v1.4.1)
- `chain.txt.ots`: OpenTimestamps proof (Bitcoin block #974890)
- `hashes_individuales.txt.ots`: OpenTimestamps proof (Bitcoin block #974890)

Full manifest with all files: `Manifest_hashes_v9.json` in the supplement.

10.4. 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 v1.4.0 analysis suite (§8.10) adds three findings. (i) The per-module firing breakdown across four conditions on 10,000 MedHallu records confirms that detection on format-controlled pairs is content-driven: the lexical baseline fires on 0.28% of A→A″ pairs, while content-sensitive modules (flow_penalty, cre_isi, negation_penalty, r3_3_drift) drive the 85.10% detection rate. (ii) A structural fingerprint (12-dimensional feature vector) achieves zero-overlap separation between paraphrase and hallucination, with a recommended rescue threshold $\kappa_P = 0.4039$. (iii) Conditional compression is evaluated and reported as a negative result: both Δ_{SR} and asymmetry are dominated by length confounds, with A→A′ asymmetry (647.9) indistinguishable from A→B (670.6). The approach is discarded as a detection module.

The v1.4.1 revision (§8.4) corrects two issues. (i) The R3.2 dictionary-based gate was accidentally rescuing 127/1000 hallucination pairs; replacing it with the structural fingerprint gate ($\kappa_P = 0.4039$) raised aggregate F1 from 0.8901 to 0.9933 (ΔF1 = $+0.1032$). (ii) Four fixes to source_target_guard (numerical magnitude comparison, drug name gazetteer, date mutation weight, causal inversion detection) raised adversarial detection from 1/5 to 5/5. Combined with the fingerprint rescue (5/5 on paraphrase cases), the adversarial set achieves 10/10.

The complete project repository is anchored to a 6-level SHA-256 hash chain (L1–L6, §9.5) timestamped on the Bitcoin blockchain via OpenTimestamps (block #974890), registered with TAD (EX-2026-

18792778), and deposited at Zenodo (10.5281/zenodo.23917127). Every claim in this paper is reproducible from the provided SHA-256-verified artifacts.

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.

Notes

v9 addendum: κ C Compression Threshold, R3.2 Topological Similarity, R3.3–R3.6 Extension Modules, v1.4.0 Analysis Suite, and v1.4.1 Fingerprint Gate.

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 thirteen-module evaluation over 20,790 pairs, the

project changelog documenting five retirements from v4/v5, and the SHA-256 manifest — is provided with this submission as `SAS_v9_Supplements.zip`. Reviewers can verify every numerical claim against the files in this archive.

The same artifacts are also available at:

- Zenodo DOI: 10.5281/zenodo.23917127 (v1.3.0 regulatory package); 10.5281/zenodo.19702379 (v9 academic paper)
- GitHub: https://github.com/Leesintheblindmonk1999/SAS-API

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 (v1.3.0, thirteen modules), the R3.1 differential guard with κC (`r3_1_differential.py`, v5), the R3.2 topological module (`r3_2_topological.py`, v1.0.0), the four extension modules (`r3_3_drift.py`, `r3_4_entropy.py`, `r3_5_kolmogorov.py`, `r3_6_surprisal.py`, all v1.1.0), the full thirteen-module evaluation script (`eval_13_modules.py`), the complete evaluation report over 20,790 pairs (`eval_13_modules_FULL.json`), 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), the threshold sweep script (`sweep_threshold_guard.py`), and the v1.4.0 analysis suite (`analysis_v14.py`, `print_breakdown.py`, `verify_v14.py`, `analysis_v14_full.json`). 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 Structural Similarity Index, or the thresholds $\kappa_D = 0.56$ and $\kappa_R = 0.15$ 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, bDurante 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, cLi 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, bLin 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.9

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.