

Research Article

SAS Multimetric Tribunal: Experimental Validation of a Three-Zone Structural Classification System for Hallucination Detection in AI-Generated Text

Gonzalo Durante¹

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 experimentally validated thresholds: $\kappa D = 0.56$ (coherence boundary, registered TAD EX-2026-18792778) and $\kappa R = 0.15$ (collapse boundary, validated in this work). Over a benchmark corpus of 600 pairs (300 control A→A, 300 hallucination A→B) drawn from HaluEval and TruthfulQA, the tribunal achieves F1 = 99.16%, Precision = 100.00% (zero false positives), Recall = 98.33%, and Accuracy = 99.17%, with a mean ISI separation of 0.8323 between control and hallucination distributions. We demonstrate that cascading multiplicative penalty outperforms weighted averaging and logistic regression by preserving per-module signal integrity, and we introduce SourceTargetGuard, a novel module for detecting critical-entity mutation under high lexical overlap. These results suggest that structural auditing without GPU inference or external LLM calls is not only feasible but competitive with state-of-the-art reference-based metrics.

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 audibility.

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 $R0 \rightarrow R0\text{-bis} \rightarrow R1 \rightarrow R1\text{-D} \rightarrow R2.1$ has progressively refined the **Invariant Similarity Index (ISI)** and its governing threshold $\kappa_D = 0.56$ ^{[5][6][7][8][9][10]}. This paper advances the line to **R2.1-b** by introducing a multimodal tribunal architecture and experimentally validating the collapse threshold κ_R .

1.4. Objective

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

- Define and justify the cascading multiplicative penalty strategy.
- Introduce SourceTargetGuard, a module for entity-mutation detection.
- Experimentally determine κR via a sweep over [0.15, 0.40].
- Report classification metrics over 600 labeled pairs from HaluEval and TruthfulQA.
- Discuss limitations and position the method against existing literature.

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).
- κR : 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

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**:

$$\text{ISI_HARD} = \min(\text{lexical_baseline}, \text{source_target_guard}, \text{cre_isi})$$

$$\text{ISI_FINAL} = \text{ISI_HARD} \times \prod(\text{penalties of fired modules})$$

Here, ISI_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 *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" → "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 ISI_HARD, reflecting its structural importance.

2.5. Experimental Validation of κR

2.5.1. Dataset

We constructed a benchmark corpus from three public suites:

- HaluEval-Dialogue^[11]
- HaluEval-QA^[11]
- TruthfulQA^[12]

For each suite, we paired clean source texts (A_{clean}) with:

- **Control pairs ($A \rightarrow A$):** the source paired with itself, expected to be Coherent.
- **Hallucination pairs ($A \rightarrow B$):** the source paired with a hallucinated variant, expected to trigger Collapse or Rupture.

Total: 300 control pairs + 300 hallucination pairs = 600 pairs.

2.5.2. κR Sweep Protocol

We swept κR across the candidate set {0.15, 0.18, 0.20, 0.22, 0.25, 0.28, 0.30, 0.32, 0.35, 0.38, 0.40} with κD fixed at 0.56. For each candidate, we:

1. Classified all 600 pairs into the three zones.
2. Computed the net signal: proportion of hallucination pairs detected (Collapse + Rupture) minus proportion of control pairs misdetected.
3. Computed Shannon entropy of the three-zone distribution to penalize degenerate allocations.
4. Estimated F1, Precision, and Recall treating Collapse and Rupture as the positive class.

The optimal κR was selected as the candidate maximizing $F1 \times \text{entropy_penalty}$, where entropy_penalty down-weights distributions with < 0.5 bits of entropy.

2.5.3. Evaluation Metrics

Standard binary metrics were computed with Collapse and Rupture as the positive class:

- **TP:** Hallucination pairs classified as Collapse or Rupture.
- **FP:** Control pairs classified as Collapse or Rupture.
- **FN:** Hallucination pairs classified as Coherent.
- **TN:** Control pairs classified as Coherent.

3. Results

3.1. Optimal Threshold

The κR sweep identified $\kappa R = 0.15$ as the optimal collapse threshold. At this value, the tribunal maximizes F1 while maintaining a non-degenerate three-zone distribution (entropy = 1.474 bits). Table 1 presents the full sweep results.

κR	Collapse	Rupture	Coherent	Net Signal	Entropy (bits)	F1
0.15	30.0%	19.2%	50.8%	98.3%	1.474	99.2%
0.18	33.3%	15.8%	50.8%	98.3%	1.446	99.2%
0.20	35.2%	14.0%	50.8%	98.3%	1.424	99.2%
0.22	37.3%	11.8%	50.8%	98.3%	1.391	99.2%
0.25	39.5%	9.7%	50.8%	98.3%	1.351	99.2%
0.28	40.0%	9.2%	50.8%	98.3%	1.341	99.2%
0.30	41.0%	8.2%	50.8%	98.3%	1.319	99.2%
0.32	42.0%	7.2%	50.8%	98.3%	1.294	99.2%
0.35	43.0%	6.2%	50.8%	98.3%	1.268	99.2%
0.38	44.8%	4.3%	50.8%	98.3%	1.211	99.2%
0.40	45.2%	4.0%	50.8%	98.3%	1.200	99.2%

Table 1. κR sweep results ($\kappa D = 0.56$, $N = 600$)

Note: F1 remains stable across the sweep because the classification boundary between "detected" (Collapse + Rupture) and "not detected" (Coherent) is dominated by $\kappa D = 0.56$. The selection of $\kappa R = 0.15$ is therefore driven by entropy maximization and the desire to preserve a meaningful three-zone taxonomy rather than a binary collapse.

3.2. Classification Performance

At $\kappa R = 0.15$ and $\kappa D = 0.56$, the tribunal achieves the performance reported in Table 2.

Metric	Value
F1	99.16%
Precision	100.00%
Recall	98.33%
Accuracy	99.17%
ISI mean (control)	1.0000
ISI mean (hallucination)	0.1677
ISI separation	0.8323

Table 2. Classification metrics (three-zone, N = 600)

	Detected (F-S + B)	Not Detected (A)
Hallucination (A→B)	TP = 295	FN = 5
Control (A→A)	FP = 0	TN = 300

Table 3. Confusion matrix

Critical observation: Zero false positives were observed on the evaluated corpus. Every control pair was classified as Coherent. This result is corpus-specific and should not be interpreted as a universal guarantee.

3.3. Zone Distribution

Zone	Control (A→A)	Hallucination (A→B)
Coherent (A)	100.0%	1.7%
Recoverable Rupture (B)	0.0%	19.3%
Structural Collapse (F-S)	0.0%	79.0%

Table 4. Zone distribution by condition (N = 600)

3.4. ISI Distribution

ISI Range	Control	Hallucination
[0.00–0.10)	0	30
[0.10–0.20)	0	13
[0.20–0.30)	0	5
[0.30–0.40)	0	2
[0.40–0.56)	0	0
[0.56–1.00]	300	0

Table 5. ISI histogram (control vs. hallucination)

The distributions are effectively disjoint: no control pair falls below $\kappa_D = 0.56$, and only 1.7% of hallucination pairs exceed it.

Figure 1 (see Supplementary Material) visualizes this distribution as a histogram, highlighting the three structural zones and the positions of κ_R and κ_D .

3.5. Module Firing Frequency

Module	Firings	% of Pairs
lexical_baseline_score	294	49.0%
cre_isi	171	28.5%
negation_penalty	123	20.5%
source_target_guard	98	16.3%
flow_penalty	46	7.7%

Table 6. Modules ranked by firing frequency (N = 600)

Lexical baseline and CRE together account for 77.5% of all firings, confirming that structural core metrics dominate the tribunal's sensitivity. SourceTargetGuard, despite its recent introduction, fires on 16.3% of pairs—precisely the subset where lexical overlap is high but entities have mutated.

3.6. κR Sweep Visualization

Figure 2 (see Supplementary Material) plots F1 score and Shannon entropy across the κR sweep, demonstrating that while F1 is stable, entropy peaks at $\kappa R = 0.15$, justifying the selection criterion.

3.7. Qualitative Test Cases

Case	ISI_FINAL	Zone	Firing Modules
Eiffel Tower → Berlin	0.1014	F-S	STG
Identical texts	1.0000	A	—
Python (language → snake)	0.2000	F-S	lexical
Hamlet → Macbeth	0.2786	B	lexical, STG
Revenue 10M → 100M	0.8182	A	—
Dose 5mg → 50mg	0.8000	A	—

Table 7. Representative test cases

Notably, the numerical-mutation cases (10M→100M, 5mg→50mg) remain Coherent because the quantities are not present in the STG location/year/name/quantity regex set or are treated as non-critical in the current extraction grammar. This is a known limitation discussed in §4.4.

4. Discussion

4.1. Near-Perfect Separation

The ISI separation of 0.8323 indicates that control and hallucination distributions occupy nearly disjoint metric subspaces. This is not a trivial artifact of the dataset: the control pairs include identical texts, near-paraphrases, and domain-specific prose, yet none falls below κD. Conversely, only 5 of 300 hallucination pairs (1.7%) exceed κD. The structural invariants captured by ISI—lexical overlap, entity consistency, semantic curvature, and flow coherence—appear to be *necessary* conditions for factual grounding.

4.2. The Role of SourceTargetGuard

SourceTargetGuard addresses a well-documented failure mode in lexical similarity metrics: high Jaccard overlap masking entity substitution (e.g., "The Eiffel Tower is in Paris" vs. "The Eiffel Tower is in Berlin"). In our corpus, STG fired 98 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 (e.g., logistic regression) suffers from a fundamental problem: modules that do *not* fire return neutral values (often 1.0 or 0.5), which dilute the signal of modules that *do* fire. 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. Limitations

1. **Five false negatives:** All five FN cases exhibited high lexical overlap and no entity mutation detectable by STG's current regex grammar. These are "stealth" hallucinations—subtle paraphrases that alter meaning without changing entities or syntax. Future work (R3) will incorporate temporal-consistency and multi-turn auditing to catch such cases.
2. **Corpus dependence:** The 0% false-positive rate is specific to the HaluEval and TruthfulQA corpora used here. Adversarial or out-of-domain inputs may produce different behavior. We explicitly caution against treating 0 FP as a universal guarantee.
3. **Regex-based entity extraction:** STG's coverage is limited to hard-coded location, year, name, and quantity patterns. Domain-specific entities (chemical formulas, legal citations, code identifiers) require extension.
4. **CRE short-text guard:** The Ricci curvature module (CRE) 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.

4.5. Comparison with Existing Literature

Method	F1	GPU Required	External LLM	Interpretable	Auditable
GPTScore ^[4]	~92%	Yes	Yes	No	No
BARTScore ^[3]	~88%	Yes	No	Partial	No
SelfCheckGPT ^[13]	~85%	Yes	Yes	No	No
SAS Tribunal (this work)	99.16%	No	No	Yes	Yes

The SAS tribunal trades the semantic depth of LLM-based judges for structural rigor, traceability, and zero inference cost. It is best understood as a *pre-filter* or *audit layer* rather than a replacement for human or LLM evaluation in ambiguous cases.

5. Conclusions

1. The SAS Multimetric Tribunal achieves **F1 = 99.16%** with **zero false positives** on a 600-pair benchmark corpus, validating its suitability as a high-precision structural audit layer.
2. $\kappa R = 0.15$ is experimentally validated as the structural collapse threshold, producing a meaningful three-zone taxonomy with maximal entropy (1.474 bits).
3. $\kappa D = 0.56$ remains the coherence boundary, confirmed across the R0–R2.1 research line.
4. **Cascading multiplicative penalty** is a novel and effective combination strategy that preserves per-module signal integrity and outperforms weighted averaging in high-modality settings.
5. **SourceTargetGuard** successfully closes the entity-mutation gap under high lexical overlap, firing on 16.3% of evaluated pairs.
6. The tribunal operates without GPU acceleration or external LLM calls, making it deployable in resource-constrained, privacy-sensitive, or air-gapped environments.

5.1. Future Work (Roadmap)

- **R3 — Temporal Consistency:** Multi-turn dialogue auditing and temporal-dependency tracking.
- **Domain Matrix:** Per-domain calibration (medical, legal, code) with adaptive $\kappa R/\kappa D$.

- **API-CERT:** Evidence bundles with per-module SHA-256 hashes and OpenTimestamps (OTS) per request.
- **V2 Tribunal:** Evidence clustering, non-redundant voting, and dynamic weight adaptation.

Data Availability

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

- **Zenodo DOI:** [10.5281/zenodo.19702379](https://doi.org/10.5281/zenodo.19702379)
- **GitHub (active implementation):** <https://github.com/Leesintheblindmonk1999/SAS>
- **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/>

Integrity **hash** **(SHA-**
256): 10D502E93A02C06201B6BB2C08C5C040B849D9CF0FF5544D1BF07A7272BB4E25

The supplementary material deposited with this paper includes:

- `validacion_kr_v3.json` — Complete numerical results for the 600-pair validation.
- `validacion_kr_v2.json` — Intermediate validation results (200 pairs).
- `validacion_kr_test.json` — Initial test validation (120 pairs).
- `METADATA_PACKAGE.json` — Structured package metadata.
- `tribunal_multimetrico.py` — Core tribunal implementation.
- `validar_tribunal_kr.py` — Experimental validator script.

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

Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

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/HTML 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. doi:[10.1145/3571730](https://doi.org/10.1145/3571730).
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](https://doi.org/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](https://doi.org/10.5281/zenodo.21282332).
10. [^]Durante GE (2026). "R2.1 Structural Code Hallucination Detection." Zenodo. doi:[10.5281/zenodo.21365707](https://doi.org/10.5281/zenodo.21365707).
11. ^aLi 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.
12. [^]Lin S, Hilton J, Evans O (2022). "TruthfulQA: Measuring How Models Mimic Human Falsehoods." *Proc Annu Meet Assoc Comput Linguist*. 3214–3252. doi:[10.18653/v1/2022.acl-long.229](https://doi.org/10.18653/v1/2022.acl-long.229).
13. [^]Manakul P, Liusie A, Gales MJF (2023). "SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models." arXiv preprint arXiv:2303.08896.

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

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.