Research Article

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

Gonzalo Durante[1]

1. Independent researcher

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

Three further findings follow from the same audit. (i) A per-module analysis shows the tribunal has no content-sensitive module in its current set: 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) Adding a semantic-equivalence guard makes the tribunal and a Jaccard baseline statistically indistinguishable (F1 0.720 vs. 0.722; AUROC 0.881 vs. 0.881), because the guard's apparent

 doi.org/10.32388/T420IV.5

success depends on a constructional property of the control (the correct answer is a substring of the source by definition) rather than on content detection.

The paper's contribution is the audit procedure that produced these retirements. Four of v4'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.

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

1. Introduction

1.1. The Hallucination Problem

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

1.2. Limitations of Current Approaches

Contemporary detection paradigms fall into three broad categories:

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

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

doi.org/10.32388/T420IV.5

1.3. The SAS Structural Line

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

Scope. We do not claim to have validated a hallucination detector, and we do not propose a fix that we have validated. We claim that the benchmark originally used to validate one does not measure hallucination detection, and we document four retirements that follow from applying the audit procedure in §2.

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. **Four retirements of v4 claims**: R3, the format interpretation of the headline F1, the 87.97-point multiplicative-vs-averaging gap, and the R3.1 semantic guard as a validated extension point. 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 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$\rightarrow$A/A$\rightarrow$B design and T is "hallucinated response".

2.3. Step 2: Construct a format-controlled negative

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

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

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

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

2.5. Step 4: Decide, and document

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

2.6. Step 5: Attribute the signal to modules

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

2.7. Step 6: Test extension claims empirically

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

3. Methods

3.1. The Invariant Similarity Index (ISI)

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

3.2. Tribunal Architecture

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

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

Table 1. Tribunal module specification

3.3. Cascading Multiplicative Penalty

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

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

Why the condition on `cre_isi` is not cosmetic. The module computes semantic curvature over LSA embeddings. On byte-identical or very short texts, embedding dispersion produces spurious firings: across the 1,800 A→A pairs, `cre_isi` fires on 971 (54%) despite the texts being byte-identical, with scores as low as 0.08. Without the conditional, the sanity check would not produce 100% precision.

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

3.4. SourceTargetGuard

SourceTargetGuard extracts and compares four entity classes (locations, years, proper names, quantities) using regex-based extraction and set-difference analysis. Each detected mutation applies an exponential

decay penalty (location 0.70, year 0.65, name 0.80, quantity 0.75), floored at 0.30.

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

3.5. Benchmark Design

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

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 2. 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 3. 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 4. 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.

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 5. $\mathrm{ISI}_{\mathrm{FINAL}}$ 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 6. 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. Detection rate is nearly identical to the hallucination detection rate (98.0%).
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 7. 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 to a property of the source (most plausibly LSA dispersion correlating with text length) and not to the source-response relationship. The tribunal has no content-sensitive module in its current set. 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', $n = 1,800$):

Scheme	Threshold	TP	FP	F1	AUROC
Multiplicative (this work)	$\kappa_D = 0.56$ (fixed)	1,764	1,569	0.687	0.514
Weighted average	0.83 (optimized)	1,759	1,511	0.695	0.632

Table 8. Combination schemes under per-scheme calibration on format-matched negatives

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 (0.687 vs. 0.695), 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.

8. The Guard Comparison

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. This claim was not tested in v4. We test it here.

8.1. Guard construction

We constructed a non-embedding guard that returns 1.0 (faithful) if the response is a substring of the source, or if all extracted entities (years, numbers, proper names) appear in the source with matching negation polarity.

8.2. Results

Method	TP	FP	F1	AUROC
Jaccard alone	1,773	1,512	0.697	0.527
Tribunal alone	1,764	1,569	0.687	0.514
Jaccard + guard	1,016	0	0.722	0.881
Tribunal + guard	1,012	0	0.720	0.881

Table 9. Four methods on A→B (positive) vs A→A' (negative), $n = 1,800$

The two guard-augmented methods are statistically indistinguishable. ΔF1 = 0.0018, ΔAUROC = 0.0000. Adding the same guard to Jaccard and to the tribunal produces the same result.

8.3. Why the guard works

The guard's apparent success is a constructional artifact. In all 1,800 A→A' pairs, the response is a substring of the source by construction (A' is extracted from A). The guard returns 1.0 for all of them, producing zero false positives. In A→B, only 430 of 1,800 responses are substrings. The guard is therefore a copy-detection rule that exploits the extraction method of the control, not a semantic-equivalence check.

8.4. Implication for the extensibility claim

The claim that a per-module architecture offers a privileged extension point over a single-scalar baseline is not supported by this experiment. A guard can be placed on top of any baseline; when placed on top of Jaccard, it produces the same result as when placed on top of the tribunal. What the per-module

architecture provides is *diagnostic visibility* (§6 was possible because module scores were exposed), not detection superiority.

9. Discussion

9.1. The format confound

The published F1 reflects response length. 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 has no content-sensitive module. `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.

9.3. The retirements

Four v4 claims do not survive this audit: R3, the format interpretation of F1, the 87.97-point gap, and the R3.1 extensibility claim. Each was retired by direct measurement, with the same protocol.

9.4. Roadmap

Milestone	Focus	Description
R3.1	Content-sensitive module	A module that measures the source-response relationship rather than a property of the source. Validated against format-matched negatives.
R3.2	Independent negative set	A control where the correct answer is *not* a substring of the source (paraphrased correct answers, human-verified). This is what R3 was intended to be.
R4	External baseline comparison	Compare against SelfCheckGPT, FActScore, G-Eval under the same format-controlled negatives.
R5	Domain-adaptive κ_D	Recalibrate per domain, on format-controlled negatives per domain.

Table 10. Roadmap (post-v5)

9.5. Known Limitations

#	Limitation	Status
3.1	Reference-penalty calibration corpus yields tautological precision	Open
3.2	Format confound in the A→B benchmark	Measured (§5)
3.3	Stealth hallucinations with high lexical similarity	Open
3.4	`legal_hallucinations` corpus pairing defect	Open
3.5	Non-determinism from unsorted directory listing	Corrected
3.6	`negation_penalty` recall estimate	Open
3.7	`arithmetic_penalty` language and form coverage	Open
3.8	`reference_penalty`: additions vs. mutations	By design
3.9	Corpus coverage and domain generalization	Open
3.10	κ_R grid extension below 0.15	Closed
3.11	No content-sensitive module in current set	Measured (§6)
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	Measured (§8)
3.15	N1 negative (A→A') is a substring of the source by construction, which biases the guard comparison	Open — a stronger control requires paraphrased correct answers, not available in this paper

Table 11. Known limitations index

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

doi.org/10.32388/T420IV.5

10. Reproducibility

10.1. Runtime

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

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

10.2. Commit and release tag

The tribunal core, the N1 control script, the validation script, the per-pair results, and the audit trail used in this paper correspond to commit hash `cdd2e12` in the project repository (tag `v5.0.0`). The complete audit trail is provided with this submission as the supplementary artifact `SAS_v5_Supplements.zip`, which contains all files listed below together with the project changelog.

Reproducibility manifest:

- `n1_control.py` (SHA-256: `270d9b13d2e9505604da2c92e309d69e63ce1e022bdd6d22dc821968f7ed3e2d`)
- `pairs_conditions_1800_seed42.json` (SHA-256: `267b7abbf241a75ceb4e6fbb3d073f9257bbcd9c09144a751864551ca94fa603`)
- `tribunal_multimetrico.py` (SHA-256: `d6a8ab5cfeff09b408ff945ff9f294bca031da2cceea7cebc0cecea4bda154d0`)
- `validacion_1800_stratified_v3.py` (SHA-256: `65415c53c7032b529664a5a9c2cb51cb5cd9d5093b252d0c38eba884484ab939`)

10.3. Determinism

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

11. Conclusions

The SAS Multimetric Tribunal reports F1 = 98.99% on a 1,800-pair benchmark. This paper shows, using a format-controlled negative on the same IDs, that the figure measures format asymmetry between source

and response rather than hallucination detection. A per-module analysis shows the tribunal has no content-sensitive module. A recalibrated comparison shows the 87.97-point advantage of multiplicative combination over weighted averaging was an artifact. A guard comparison shows that adding a semantic-equivalence guard to the tribunal and to a Jaccard baseline produces statistically indistinguishable results.

Four v4 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.

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

Appendix A. Withdrawn Control: R3

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

Statements and Declarations

Funding

No specific funding was received for this work.

Potential Competing Interests

No potential competing interests to declare.

Ethics

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

Data Availability

The complete set of artifacts supporting this paper — source code, per-pair results, the κ_R sweep, both ablation results, the per-module analysis data, the project changelog documenting four retirements from v4, and the SHA-256 manifest — is provided with this submission as `SAS_v5_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.19702379
- GitHub: https://github.com/Leesintheblindmonk1999/sas-multimetric-tribunal

Supplementary material: N1 per-pair results (`pairs_conditions_1800_seed42.json`), N1 script, the false-negative registry, both seed validation files, the κ_R sweep, both ablation results, the per-module analysis data, and the tribunal implementation. Withdrawn: `results_ac_pilot.jsonl`, `baselines_triviales_r3.json`.

License and Attribution

The manuscript is released under the **Creative Commons Attribution–NonCommercial 4.0 International License (CC BY-NC 4.0)**. The software is dual-licensed: GPL-3.0 for open-source use, a separate commercial license for proprietary integration. Inquiries: duranteg2@gmail.com. Full license texts in the repository.

Any use of the audit methodology, the Invariant Similarity Index, or the thresholds $\kappa_D = 0.56$ 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. ^Durante GE (2026a). "SAS — Symbiotic Autoprotection System." *Zenodo.* doi:10.5281/zenodo.19702379.
6. ^Durante GE (2026). "R0 Infrastructure and Baseline Stability Audit." *Zenodo.*
7. ^Durante GE (2026). "R0-bis Nonlinear Dependence and Redundancy Audit." *Zenodo.*
8. ^Durante GE (2026). "R1 Real Local Structural Evaluation v1.0.7." *Zenodo.*
9. ^Durante GE (2026). "R1-D Structural Evaluation over Declarative Corpus." *Zenodo.*
10. ^Durante GE (2026). "R2.1 Structural Code Hallucination Detection." *Zenodo.*
11. ^Durante GE (2026). "SAS R2.1-b, Function-Level Extension." *Zenodo.*
12. ^Li J, Cheng X, Zhao WX, Nie JY, Wen JR (2023). "HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models." *arXiv preprint arXiv:2305.11747.*
13. ^Lin S, Hilton J, Evans O (2022). "TruthfulQA: Measuring How Models Mimic Human Falsehoods." *Proc. ACL.* 3214–3252.

Supplementary data: available at https://doi.org/10.32388/T420IV.5

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.