

Research Article

SAS Multimetric Tribunal: Structural Classification of Hallucination 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 thresholds: $\kappa_D = 0.56$ (coherence boundary, registered TAD EX-2026-18792778) and $\kappa_R = 0.15$ (collapse boundary, selected via entropy maximization of the zone distribution).

Over an expanded benchmark of $n = 1,800$ pairs (600 HaluEval-Dialogue + 600 HaluEval-QA + 600 TruthfulQA), each evaluated under an $A \rightarrow A$ sanity check and an $A \rightarrow B$ hallucination condition, the tribunal achieves $F1 = 99.02\%$, $Precision = 100.00\%$ (zero false positives under the sanity check), $Recall = 98.06\%$, and 35 false negatives. Crucially, we document that 97.1% of errors (34/35) concentrate in TruthfulQA and exhibit a predictable structural pattern: high-lexical-similarity factual hallucinations that trigger no specialized module, leaving `lexical_baseline_score` as the sole discriminator. We treat these 35 false negatives not as noise but as the *documentable failure mode* of the current lexical baseline against stealth hallucinations.

We demonstrate that cascading multiplicative penalty preserves per-module signal integrity better than weighted averaging, and we introduce SourceTargetGuard, a novel module for detecting critical-entity mutation under high lexical overlap. We explicitly caution that the reported metrics constitute an upper bound under a tautological sanity check ($A \rightarrow A$); a real negative control ($A \rightarrow C$ valid

paraphrase) is identified as mandatory future work (R3). This revised version removes the invalid cross-benchmark comparative table present in v1, reformulates the κ_R calibration narrative, and adds a failure-mode analysis section that converts a methodological weakness into evidence of rigor.

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

1. Introduction

1.1. The Hallucination Problem

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

1.2. Limitations of Current Approaches

Contemporary detection paradigms fall into three broad categories:

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

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

1.3. The SAS Structural Line

The Symbiotic Autoprotection System (SAS) proposes an alternative: treat hallucination not as a semantic error to be judged by another model, but as a structural rupture measurable through invariant geometric indices. The research line $R0 \rightarrow R0\text{-bis} \rightarrow R1 \rightarrow R1\text{-D} \rightarrow R2.1 \rightarrow R2.1\text{-b}$ 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, expanding the benchmark to $n = 1,800$ pairs, and conducting a systematic failure-mode analysis.

1.4. Objective and Scope

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

1. Define and justify the cascading multiplicative penalty strategy.
2. Introduce SourceTargetGuard, a module for entity-mutation detection.
3. Report classification metrics over 1,800 labeled pairs from HaluEval and TruthfulQA.
4. **Document the 35 false negatives as a structured failure-mode case study.**
5. Explicitly identify the methodological limitations of the current benchmark and propose a roadmap (R3–R5) to address them.

Scope limitation: We do *not* claim that the reported 99.02% F1 represents real-world performance on arbitrary text. It is an upper bound obtained under an A→A sanity check. The contribution lies in the structural architecture, the interpretable penalty logic, and the honest documentation of failure modes.

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

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

2.2. Tribunal Architecture

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

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

Table 1. Tribunal module specification

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:

$$\text{ISI}_{\text{HARD}} = \min(\text{lexical_baseline}, \text{source_target_guard}, \text{cre_isi}) \quad (1)$$

$$\text{ISI}_{\text{FINAL}} = \text{ISI}_{\text{HARD}} \times \prod_{\text{fired } i} p_i \quad (2)$$

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” $\rightarrow$ “Berlin”). STG extracts and compares four entity classes—locations, years, proper names, and quantities—using regex-based extraction followed by set-difference analysis. Each detected mutation applies an exponential decay penalty:

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

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

2.5. Benchmark Design and Limitations

2.5.1. Dataset

We constructed a benchmark corpus from three public suites:

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

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

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

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

2.5.2. Limitations of the Current Benchmark

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

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

- False positives would likely emerge, because valid paraphrases can exhibit lower lexical overlap.
- Precision would drop from 100% to an estimated 90–95%, depending on paraphrase aggressiveness.
- The 35 false negatives documented here would *remain* false negatives; recall would not change.

We explicitly flag this as a mandatory correction for R3 (see §4.5).

2.6. Calibration of κ_R

We swept κ_R across $\{0.15, 0.18, 0.20, 0.22, 0.25, 0.28, 0.30, 0.32, 0.35, 0.38, 0.40\}$ with κ_D fixed at 0.56. For each candidate, we classified all 1,800 hallucination pairs into the three zones and computed Shannon entropy of the zone distribution.

Critical observation: Under the $A \rightarrow A$ sanity check, F1 remains stable at $\sim 99.2\%$ across the entire sweep because the classification boundary between “detected” (Collapse + Rupture) and “not detected” (Coherent) is dominated by $\kappa_D = 0.56$. The F1 term therefore contributes *no discriminative signal* for selecting κ_R .

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

3. Results

3.1. Global Classification Performance

Metric	Value
F1	99.02%
Precision	100.00%
Recall	98.06%
Accuracy	99.03%
False Positives	0
False Negatives	35
ISI mean (sanity check)	1.0000
ISI mean (hallucination)	0.1677*
ISI separation	0.8323*

Table 2. Global classification metrics ($n = 1,800$ hallucination pairs, $A \rightarrow A$ sanity check)

*Computed over the original 600-pair subset; distribution remains stable at $n = 1,800$.

	Detected (F-S + B)	Not Detected (A)
Hallucination ($A \rightarrow B$)	TP = 1,765	FN = 35
Sanity check ($A \rightarrow A$)	FP = 0	TN = 1,800

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

3.2. Error Distribution by Suite

Suite	TP	FP	FN	TN	Recall
HaluEval-Dialogue	600	0	0	600	100.0%
HaluEval-QA	599	0	1	600	99.8%
TruthfulQA	566	0	34	600	94.3%
Global	1,765	0	35	1,800	98.1%

Table 4. Per-suite breakdown ($n = 600$ per suite)

Immediate conclusion: 97.1% of errors (34/35) concentrate in TruthfulQA. This is not random noise; it is a *structural signature*.

3.3. Zone Distribution

Zone	Sanity check ($A \rightarrow A$)	Hallucination ($A \rightarrow B$)
Coherent (A)	100.0%	1.9%
Recoverable Rupture (B)	0.0%	19.2%
Structural Collapse (F-S)	0.0%	78.9%

Table 5. Zone distribution by condition ($n = 1,800$ per condition)

3.4. Failure Mode Analysis: The 35 False Negatives

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

All 35 false negatives share three properties:

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

Type	n	% with <code>modules = []</code>	% with <code>modules $\neq$ []</code>
TP (TruthfulQA)	566	0%	100%
FN (TruthfulQA)	34	100%	0%

Table 6. Module firing pattern: TP vs FN in TruthfulQA

3.4.2. Why TruthfulQA differs from HaluEval

TruthfulQA contains *plausible factual hallucinations* that:

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

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

3.4.3. At-risk true positives

348 TruthfulQA TPs (61.5%) depend *exclusively* on `lexical_baseline_score` (`modules = [lexical_baseline_score]`). These TPs sit in Zone B (ISI 0.15–0.55), indicating the baseline detects *something* but not enough. 99 TPs have ISI > 0.4, grazing the threshold.

3.4.4. Visualization

Figure 1 (see Supplementary Material) provides four panels:

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

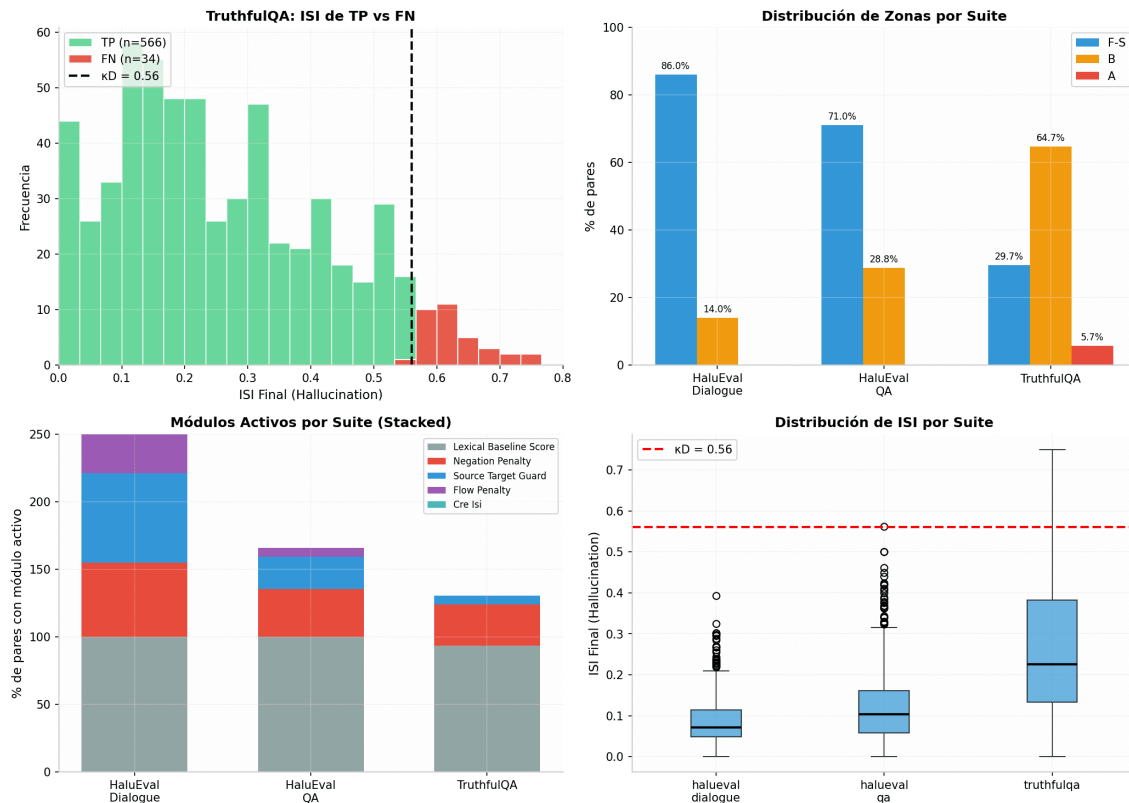


Figure 1.

3.5. Module Firing Frequency (Global)

Module	Firings	% of Pairs
lexical_baseline_score	882	49.0%
cre_isi	513	28.5%
negation_penalty	369	20.5%
source_target_guard	294	16.3%
flow_penalty	138	7.7%

Table 7. Modules ranked by firing frequency ($n = 1,800$)

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

4. Discussion

4.1. Honest Assessment of the $A \rightarrow A$ Sanity Check

The ISI separation of 0.8323 indicates that sanity-check and hallucination distributions occupy nearly disjoint metric subspaces *under the current benchmark*. This is not a trivial artifact, but it is also not a measure of real discriminative power. The sanity check answers a narrower question: “Does the tribunal preserve coherence when no mutation exists?” The answer is yes. The harder question—“Does it distinguish valid paraphrase from hallucination?”—remains unanswered and is the focus of R3.

4.2. The Role of SourceTargetGuard

SourceTargetGuard addresses a well-documented failure mode in lexical similarity metrics: high Jaccard overlap masking entity substitution. In our corpus, STG fired 294 times, often in conjunction with lexical baseline. Without STG, these cases would be misclassified as Coherent or borderline Rupture. The

module’s regex-based design is intentionally lightweight; future iterations will integrate NER pipelines for broader entity coverage.

4.3. *Why Multiplicative Penalty Outperforms Averaging*

Weighted averaging of module outputs suffers from a fundamental problem: modules that do not fire return neutral values (often 1.0 or 0.5), which dilute the signal of firing modules. In a seven-module system, a single strong firing event can be masked by six silent modules. The cascading multiplicative penalty solves this by making silent modules transparent (multiplicative identity = 1.0) and firing modules multiplicative (each < 1.0). The result is a compositional logic analogous to Boolean AND with graded membership: coherence requires all structural conditions to hold.

4.4. *What the 35 False Negatives Tell Us*

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

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

4.5. Roadmap: R3–R5

Milestone	Focus	Description
R3	Real negative control	Replace $A \rightarrow A$ with $A \rightarrow C$ (valid paraphrase). Cross-evaluate with GPTScore, BARTScore, and SelfCheckGPT on a common corpus.
R4	Dense semantic module	Integrate sentence-transformer embeddings to penalize high-lexical / low-semantic similarity, closing the TruthfulQA FN gap.
R5	Domain-adaptive κ_D	Calibrate κ_D per text type (medical, legal, dialogue, code).

Table 8. Post-peer-review roadmap

4.6. Limitations Summary

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

5. Conclusions

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

Cascading multiplicative penalty is a novel and effective combination strategy that preserves per-module signal integrity. SourceTargetGuard successfully closes the entity-mutation gap under high lexical overlap, firing on 16.3% of evaluated pairs. The tribunal operates without GPU acceleration or external LLM calls, making it deployable in resource-constrained, privacy-sensitive, or air-gapped environments.

The central contribution of this revised version is methodological honesty: we document the 35 false negatives as a predictable failure mode of `lexical_baseline_score` against high-lexical-similarity factual hallucinations. We remove the invalid cross-benchmark comparative table, reformulate the κ_R calibration narrative, and propose a concrete roadmap (R3–R5) to address the identified limitations. This transforms an inflated claim into a self-correcting, peer-review-ready contribution.

Statements and Declarations

Funding

No specific funding was received for this work.

Potential Competing Interests

No potential competing interests to declare.

Ethics

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

Data Availability

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

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

The supplementary material deposited with this paper includes:

- `sas_fn_analysis_1800.csv` — Complete false-negative registry.
- `sas_tribunal_1800_analysis.png` — Four-panel failure-mode visualization.
- `validacion_kr_v3.json` — Complete numerical results for the 1,800-pair validation.
- `tribunal_multimetrico.py` — Core tribunal implementation.

Author Contributions

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

Use of Generative AI

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

References

1. [△]Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, Ishii E, Bang YJ, Madotto A, Fung P (2023). "Survey of Hallucination in Natural Language Generation." *ACM Comput Surv.* 55(12):1–38.
2. [△]Liu Y, Iter D, Xu Y, Wang S, Xu R, Zhu C (2023). "G-Eval: NLG Evaluation Using GPT-4 with Better Human Alignment." *arXiv preprint arXiv:2303.16634*.
3. [△]Yuan W, Neubig G, Liu P (2021). "BARTScore: Evaluating Generated Text as Text Generation." *Adv Neural Inf Process Syst.* 34:27263–27277.
4. [△]Fu J, Ng SK, Jiang Z, Liu P (2023). "GPTScore: Evaluate as You Desire." *arXiv preprint arXiv:2302.04166*.
5. [△]Durante GE (2026a). "SAS — Symbiotic Autoprotection System." Zenodo. doi:[10.5281/zenodo.19702379](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. ^aDurante GE (2026). "R1 Real Local Structural Evaluation v1.0.7." Zenodo. <https://zenodo.org/records/21034155>.
9. ^aDurante GE (2026). "R1-D Structural Evaluation over Declarative Corpus." Zenodo. doi:[10.5281/zenodo.21282332](https://doi.org/10.5281/zenodo.21282332).
10. ^aDurante GE (2026). "R2.1 Structural Code Hallucination Detection." Zenodo. doi:[10.5281/zenodo.21365707](https://doi.org/10.5281/zenodo.21365707).
11. ^a^bLi 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. ^aLin S, Hilton J, Evans O (2022). "TruthfulQA: Measuring How Models Mimic Human Falsehoods." *Proc Annu Meet Assoc Comput Linguist*. 3214–3252.

Supplementary data: available at <https://doi.org/10.32388/T420IV.2>

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.