

Peer Review

Review of: "A Geometric Taxonomy of Hallucinations in LLMs"

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

The paper argues that the single word “hallucination” hides three different failure modes, and that the three can be separated by where the response sits on the unit hypersphere. Type I is unfaithfulness: the model ignores the context it was given and stays near the query. Type II is confabulation: the model invents entities or mechanisms, and the response displacement leaves the region where grounded answers land. Type III is a factual error inside a correct conceptual frame, which the author claims angular geometry cannot see. Two detectors follow. SGI (Eq. 1, §3.1.1) is the ratio $\theta(r,q)/\theta(r,c)$, above 1 when the response has moved toward the context. Γ (Eq. 2, §3.1.2) is the dot product of the normalized query-to-response displacement with a mean grounding direction $\hat{\mu}$ estimated from a reference set of verified grounded pairs.

The empirical case has four parts. Table 1 reports SGI on HaluEval QA ($n = 10,000$) across five embedding models, mean AUROC 0.805, mean grounded SGI 1.180 against 0.910 for hallucinated. Table 2 reports Γ on 142 human-written confabulations at AUROC 0.958, 95% CI [0.924, 0.992], against an NLI cross-encoder at 0.611. Table 3 extends to three external human-annotated benchmarks, with domain-specific Γ at 0.581 (WikiBio), 0.648 (FELM), and 0.695 (ExpertQA). Table 6 uses TruthfulQA as the Type III probe: a logistic regression on raw embeddings reaches 0.731, but cosine similarity is inverted at 0.365, and held-out Γ falls to 0.535 with $p = 0.265$, which the author reads as a stylistic annotation artifact rather than factual-error detection.

Let me say up front what is genuinely good here, because the rest of my report is critical. The three-way split is not arbitrary: it is tied to a mechanism, and the mechanism makes a falsifiable prediction about what will and will not be detectable. The TruthfulQA analysis in §4.3 is the kind of work most papers skip because it takes a number that flatters the method (0.731) and argues it away. The ExpertQA result in

Table 3 (Γ 0.695 [0.654, 0.738] against NLI 0.452 [0.426, 0.478]) has non-overlapping intervals and is the one place where the central claim is carried by data I can check. Table 1's row means all reconcile: $5.902/5 = 1.1804$ grounded, $4.550/5 = 0.910$ hallucinated, $4.023/5 = 0.8046$ AUROC, $1.89/5 = 0.378$ TPR. The arithmetic in §4.2.1 checks out too, $0.958 - 0.611 = 0.347$. My objections are that the strongest number in the abstract may not be held out, that Tables 3 and 4 contradict each other, and that the external validation is reported selectively.

2. Whether the conclusion holds on the data

2.1 The headline AUROC is measured on a dataset written by the author to match the method's definition. §4.2.1 says the 142 confabulations were "prepared by the authors to represent prototypical Type II errors: redefined terms, fabricated institutions, or invented mechanisms. These introduce semantically foreign content." But semantically foreign content is what Γ measures, by construction. The items were selected for the property being detected, so AUROC 0.958 shows that the operationalization is coherent, not that the detector works on errors arising in the wild. The honest external estimate is Table 3's 0.581 to 0.695, and the abstract leads with 0.958.

2.2 I could not determine whether Table 2 is held out, and the paper's own TruthfulQA result shows why that matters. Γ needs a reference set R from which $\hat{\mu}$ is computed (§3.1.2). For TruthfulQA, the paper is explicit about separation: Table 6 labels the row " Γ (80/20 split)" and §4.3 says "Held-out Γ calibration, which cannot propagate this style signal through $\hat{\mu}$, yields AUROC=0.535." Table 2 carries no such statement. Its row is labelled " Γ local (n=142)", the same n as the evaluation set, which reads as though $\hat{\mu}$ and the scored items came from the same 142 pairs. I am not asserting that they did. The text does not let me rule it out, and this is the first thing a reader will want settled.

2.3 The external validation is reported selectively. The abstract and §7 both state that Γ outperforms NLI "by $\Delta=0.243$ on expert-domain data." True, on ExpertQA. On the other two datasets in the same table, the comparison runs the other way or vanishes. From Table 3: WikiBio Γ 0.581 against NLI 0.712, so $\Delta = -0.131$, and the baseline wins by a wide margin. FELM Γ 0.648 against NLI 0.624 gives $\Delta = +0.024$, which against CIs of [0.510, 0.801] and [0.532, 0.704] is nothing. Two of three external benchmarks fail to show the advertised advantage, and only the third reaches the abstract. Both stray values appear in the corrupted header row of Table 3 as rendered on Qeios, so the author computed them; they simply did not survive into the summary.

2.4 Table 3 is not legible as rendered on Qeios, and Table 4 disagrees with it. Table 3’s header row on the Qeios page reads “Dataset | -0.131 dom. | 95% CI | n=81 | NLI | 95% CI | +0.024”, with row labels “WikiBio (n=900)”, “FELM (Γ)”, and “ExpertQA (Δ)”. The sample sizes and the Δ values have been transposed into headers and labels. §4.2.3 gives 102 matched pairs for WikiBio, 81 for FELM, and 900 for ExpertQA, so “WikiBio (n=900)” is wrong under any reading, and the caption compounds it with “ Δ : n=102”. Possibly a rendering fault, but it needs fixing in the version of record.

The harder problem is Table 4. Its diagonal should be the same quantity as Table 3’s domain-specific Γ , since both describe a detector calibrated and tested within one dataset. They do not match. WikiBio reads 0.537 against 0.581; ExpertQA 0.691 against 0.695, close enough to be rounding; FELM 0.822 against 0.648, a gap of 0.174 that no rounding explains. Either the two tables use different estimators and must say so, or one of them is wrong.

2.5 The “symmetry” claim reads the wrong two cells. §4.2.3 and the caption of Table 4 both assert that “the FELM $\leftrightarrow$ ExpertQA symmetry (0.822) confirms that open-ended factual confabulation shares a consistent geometric regime across subject domains.” In the matrix, train FELM $\rightarrow$ test ExpertQA is 0.653, and train ExpertQA $\rightarrow$ test FELM is 0.822. Those are the two cells a symmetry claim needs, and they differ by 0.169. The value 0.822 appears twice: once as the FELM diagonal, once as ExpertQA $\rightarrow$ FELM. So the claimed symmetry holds between an in-domain diagonal and an off-diagonal transfer, which is not a symmetry, or else a cell has been duplicated in error. The sentence cannot stand as written, and it is one of the two load-bearing generalization claims in §7.

2.6 Below-chance transfer is strong signal, not incommensurability. Table 4 gives WikiBio $\rightarrow$ FELM at 0.200 and WikiBio $\rightarrow$ ExpertQA at 0.288. §4.2.3 calls this WikiBio’s $\hat{\mu}$ “actively mislead[ing]” and “geometrically incommensurable with the confabulation regime.” That undersells it. An AUROC of 0.200 inverts to 0.800 and 0.288 to 0.712, so WikiBio’s grounding direction is close to a good detector with the sign flipped. Incommensurable directions give values near 0.50, not 0.20. The paper reasons exactly this way about the TruthfulQA cosine inversion (0.365, Table 6). The same treatment is owed here, and it may be informative: WikiBio’s criterion looks anti-correlated with the confabulation regime, not orthogonal to it.

2.7 The architecture ablation does not support the interpretation given. §4.2.4 introduces Table 5 with “Higher-dimensional embedding spaces provide finer geometric resolution for displacement measurement.” Three rows: all-MiniLM-L6-v2 at 384 dimensions, 0.800 [0.730, 0.870]; all-mpnet-base-v2 at 768, 0.880 [0.820, 0.940]; sentence-t5-large at 768, 0.958 [0.924, 0.992]. The across-dimension step is

$0.880 - 0.800 = 0.080$. The within-dimension step, between two models of identical width, is $0.958 - 0.880 = 0.078$. Same size. Dimensionality explains none of the variation that model identity does not already explain, and the 384-to-768 comparison is not established anyway, since $[0.730, 0.870]$ and $[0.820, 0.940]$ overlap across a span of 0.050. Three models on one 142-item set cannot carry a claim about dimension.

2.8 Type I is evaluated without any baseline, and the stated reason does not hold. §3.2 says the NLI baseline “applies to Type II and Type III experiments only” because SGI takes (q,c,r) triples and “no standard entailment model accepts three inputs in this configuration.” But faithfulness to a provided context is the most heavily benchmarked task in this area, and the natural baseline is a two-input one: entailment of r by c , ignoring q . §2 cites the relevant systems itself, reporting AlignScore at AUC-ROC 0.87 and HHEM at 74–77% balanced accuracy. SGI’s best single-model figure in Table 1 is 0.824 (sentence-t5-large), below the AlignScore number the paper quotes. That comparison is not decisive, since the datasets differ, but it is the one every reader will make.

2.9 The style confound is diagnosed by argument, not measured. §4.3 explains the TruthfulQA LR result this way: “truthful answers are longer and hedged; false answers are shorter and declarative. Longer responses accumulate semantic content orthogonal to the query, producing larger displacement vectors that the LR exploits across 768 dimensions.” I think this is probably right. It is also entirely unquantified. No length statistics appear anywhere: no mean or median token count per class, no correlation between length and cosine, and above all no length-only classifier. That last one is a few lines of code, and it would settle the question. If length alone reaches AUROC near 0.731 on these 817 pairs, the argument is proven; if it reaches 0.55, something else is driving the LR.

A related gap: Table 6 puts LR on Γ features at 0.786 $[0.765, 0.807]$, above LR on raw 768-dimensional embeddings at 0.731 $[0.706, 0.756]$, intervals not overlapping. Both rows are labelled “style artifact.” Why a small feature set beats the embedding it came from is not discussed, and under the style explanation, it is counterintuitive.

2.10 The confound test is applied only where it hurts, never where it helps. §4.3 asks whether a positive result could be a length artifact and answers yes. That question is never put to Table 1 or Table 2. HaluEval’s hallucinated responses are LLM-generated on a fabricate-this prompt, which §1 itself criticizes as capturing “how models write when asked to lie”; such responses plausibly differ in length and register from grounded ones. The 142 items of Table 2 were hand-written, and hand-written negatives are a classic source of phrasing regularities. Table 1’s cross-model $r = 0.85$ is offered as proof

that SGI “measures a property of the text, not an embedding artifact”; length is also a property of the text, so that correlation is equally consistent with the confound.

2.11 The Type III conclusion is a null result promoted to a theoretical constraint. The evidence that Type III errors are undetectable is one line: held-out Γ on TruthfulQA at 0.535, CI [0.473, 0.600] crossing 0.50, $p = 0.265$ (Table 6). A failure to reject, on one dataset, with one embedding model, at roughly 163 test pairs given the 80/20 split of 817. From this, the Discussion concludes that “Type III errors are geometrically invisible by construction” and §7 that the finding confirms “a theoretical constraint rather than a methodological gap.” §6 says the opposite, and says it well: “The Type II/III boundary is defined geometrically but not yet formally proven ... a formal proof on S^{d-1} for within-frame substitutions remains open.” I agree with §6. Bring the other three into line with it.

A second problem sits underneath. §4.3 argues that TruthfulQA is confounded by annotation style and unfit as a factuality probe, and §4.2.3 says WikiBio’s criterion “conflat[es] Type II and Type III”. Those are the only two Type III instruments here, so the claim that Type III is undetectable rests on datasets the paper itself calls unsuitable for measuring Type III.

3. Reproducibility

This paper does not run language models, so the usual complaint about missing decoding parameters applies only in part: the generations come from HaluEval, WikiBio GPT-3, FELM, ExpertQA, and TruthfulQA, all public. Credit where due, the encoders are named precisely (sentence-t5-large at 768 dimensions and cross-encoder/nli-deberta-v3-small, both §3.2), the resampling counts are given (1,000 bootstrap, 5,000 permutation), and the comparison test is named (Hanley–McNeil). The following are missing.

Artifacts. The Reproducibility statement says “Code and datasets released upon publication” and “Human-crafted confabulation dataset released with code.” No repository URL, no archive DOI, no anonymized link. For a reviewer, that is a promise, not an artifact, and the 142-item dataset is the one thing here that cannot be obtained anywhere else. It carries the headline number, so it should be posted now.

Model revisions. Names without revisions. HuggingFace checkpoints are mutable, so both encoders should carry a commit hash, with the sentence-transformers and PyTorch versions stated.

Splits and seeds. No seed for the bootstrap, the permutation tests, or the TruthfulQA 80/20 split. A single random split at $n = 817$ gives a test AUROC with real variance, and the conclusion of §4.3 rests on that one

split landing at 0.535. Repeated splits with a reported spread would make it much safer.

Reference set construction. §3.1.2 defines R as “verified grounded pairs” without saying, for any experiment, where those pairs came from, how many there were, or how they were verified. The only k in the paper is “ $k=15$ ” in Table 2, attached to the row labelled “ Γ global,” though k belongs to the local variant.

Aggregation of the external benchmarks. All three carry annotations finer than the response: sentence-level for WikiBio, segment-level for FELM, claim-level for ExpertQA (§4.2.3). The conversion into 102, 81, and 900 “matched pairs” is never described. Whether a response counts as hallucinated when one segment is flagged or when a majority is, and how the matched partner is chosen, will move the AUROCs. This is the most consequential undocumented preprocessing step in the paper.

Numbers with no table behind them. The cross-domain degradation of 3.8% (abstract, Table 2 caption) has no per-domain table, though the 142 items span finance, medical, and legal, so three train-on-one numbers should be printable. §4.2.2 reports in-domain HaluEval AUROC of “0.68–0.99” and “Dialogue $\rightarrow$ QA: 0.38” with no table; a range 0.31 wide tells a reader almost nothing, and 0.38 is another below-chance value of the kind discussed in 2.6. The mean cosine of -0.07 between domain grounding directions has no pairwise matrix. The one text number I could check does hold: 0.965 against 0.958 is $0.007/0.958 = 0.73\%$, matching the caption’s “0.7%”.

Notation on the headline interval. The abstract and §7 write AUROC 0.958 ± 0.034 ; Table 2 gives $[0.924, 0.992]$. Since $0.958 - 0.924 = 0.034$ and $0.992 - 0.958 = 0.034$, the $\pm$ is a 95% interval half-width, but most readers will take it for a standard error. Write the interval. Both Table 2 intervals are also exactly symmetric, with the local row’s upper bound at 0.999, which is unusual for a percentile bootstrap near the ceiling at $n = 142$; name the interval method used.

Compute. Nothing on hardware or runtime. For a method whose selling point in §3.1.2 is that global Γ costs “one embedding and one dot product: $O(d)$ ”, a measured per-item latency against the NLI baseline would make that concrete rather than asymptotic.

4. What specifically to fix

1. Table 4, and the sentence in §4.2.3 beginning “The FELM $\leftrightarrow$ ExpertQA symmetry”. Check the matrix for a duplicated cell. As printed, FELM $\rightarrow$ ExpertQA is 0.653 and ExpertQA $\rightarrow$ FELM is 0.822,

differing by 0.169, so the symmetry claim is not supported by the cells it cites. If the matrix is correct, rewrite the claim as asymmetric transfer and say which direction works.

2. **Reconcile Table 4's diagonal with Table 3.** FELM reads 0.822 in one and 0.648 in the other; WikiBio, 0.537 against 0.581. State whether these are the same estimator. If they are, one set is wrong.
3. **Abstract, sentence 4; §7, sentence 3.** Report all three external Δ values, not one: WikiBio -0.131 , FELM $+0.024$, ExpertQA $+0.243$. The current phrasing lets a reader infer that Γ beat the baseline on the external benchmarks generally. It did so on one of three.
4. **§4.2.1 and Table 2: state the train/test separation.** If $\hat{\mu}$ for the 142 confabulations came from held-out grounded pairs, say which and how many. If it did not, report the held-out number as the headline, as §4.3 does for TruthfulQA.
5. **Run a length-only classifier on TruthfulQA and put it in Table 6**, with mean and median response length by class. The style-confound argument in §4.3 is the paper's best idea, and it has no measurement behind it.
6. Apply the same length check to Table 1 and Table 2, reporting it whichever way it comes out. The cross-model $r = 0.85$ in §4.1 does not rule out a length artifact, since length is a property of the text too.
7. **Add a context-only entailment baseline to Table 1.** The reason given in §3.2 for omitting one, that no entailment model takes three inputs, does not apply to NLI(c, r), which is what the faithfulness systems cited in §2 actually do. Without it, SGI's 0.824 has no comparator.
8. **§4.2.4, opening sentence, and Table 5.** The dimension claim is not carried by these three rows: 384→768 is $+0.080$, 768→768 is $+0.078$, and the first two CIs overlap over $[0.820, 0.870]$. Drop the interpretation, or add models at 1024 and above with several per width.
9. Table 3 as rendered on Qeios has the header row and row labels transposed, with -0.131 , $n=81$, and $+0.024$ in the header and (Γ) , (Δ) , $(n=900)$ in the labels; the caption " Δ : $n=102$ " is scrambled too. Regenerate it with n per dataset matching §4.2.3 (102, 81, 900).
10. **§4.2.3 on WikiBio, and §4.2.2 on Dialogue→QA.** Treat below-chance transfer as inverted signal. 0.200 inverts to 0.800 and 0.288 to 0.712, which is not what incommensurability looks like; incommensurability looks like 0.50. The paper already reasons this way about the 0.365 cosine in Table 6, so the tools are in hand.
11. Soften the Type III claim in the abstract ("identifies a theoretical constraint"), the Discussion ("geometrically invisible by construction"), and §7 ("a theoretical constraint rather than a

methodological gap”) to match §6, which already states correctly that no proof exists. The whole support is Table 6 row “ Γ (80/20 split)”: 0.535, CI [0.473, 0.600], $p = 0.265$.

12. Build or borrow a clean Type III set: correct answers with one entity, date, or quantity replaced by a plausible wrong one, frame untouched. The Type III argument currently runs on TruthfulQA and WikiBio, which §4.3 and §4.2.3 both declare unfit for it.
13. **§3.1.1, the sentence justifying angular distance.** The triangle inequality bounds the difference of two geodesic distances, not their ratio, so it does not make the ratio “geometrically principled” as claimed. Give the argument properly, or replace it with an empirical one: report AUROC for the cosine-based analogue of Eq. 1 next to SGI. While there, say what happens as $\theta(r,c)$ approaches zero and the ratio diverges.
14. Post the code and the 142-item dataset now rather than “upon publication”, with commit hashes for both encoders and seeds for the bootstrap, the permutation tests, and the 80/20 split.
15. Document how sentence-level, segment-level, and claim-level annotations became the 102, 81, and 900 matched pairs of §4.2.3. Different aggregation rules give different AUROCs, and a reader cannot currently reproduce even the inputs.
16. Print the missing tables: per-domain results behind the 3.8% degradation, the HaluEval matrix behind “0.68–0.99” in §4.2.2, the pairwise cosines behind $\hat{\mu} = -0.07$.
17. Replace “0.958±0.034” in the abstract and §7 with [0.924, 0.992], and name the bootstrap interval method in §3.2.
18. Language. A careful pass will catch the agreement errors: the abstract’s “The term ‘hallucination’ converge different failure modes”; §2’s “Academic definitions groups around two poles” and “Our approach exploit the geometric structure”; §4.2.3’s “WikiBio domain-specific detection fail”. §2 also carries an empty citation: “HHEM () achieves 74–77% balanced accuracy”.

5. Recommendation

Major revisions.

I would like to see this published. The taxonomy is a real contribution, the ExpertQA result is solid, and §4.3 argues against the author’s own favorable number in a way I wish were more common. Nothing I found is fatal to the idea.

Three things have to be settled first. Tables 3 and 4 do not agree, and the symmetry sentence in §4.2.3 cites two cells that are not symmetric; until that is resolved, I cannot tell what the transfer experiment

showed. The 0.958 rests on a dataset built to fit the method and may not be held out, which matters because the paper's own held-out result elsewhere is so much weaker. And the abstract, Discussion, and Conclusions state the Type III claim more strongly than §6 does, on one non-significant result from a dataset the paper itself calls confounded.

Items 1 through 8 are needed for this version. Items 12 and 13 are larger, and I would accept them as future work if §6 says so plainly. The rest are corrections and documentation. If items 1 and 2 reconcile in the author's favour and item 4 confirms held-out calibration, my assessment moves up considerably.

Declarations

Potential competing interests: No potential competing interests to declare.