

Peer Review

Review of: "Semantic Grounding Index: Geometric Bounds on Context Engagement in RAG Systems"

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

The paper proposes the **Semantic Grounding Index (SGI)**, defined in Equation 3 as the ratio of geodesic distances $\theta(r,q)/\theta(r,c)$ between an L2-normalised response embedding and, respectively, the question and the retrieved context on the unit hypersphere S^{d-1} (angular distance given by Equation 2). The central empirical claim, named *semantic laziness* and formalised as the inequality of expectations in Equation 6, is that hallucinated RAG responses stay angularly close to the question instead of moving toward the context. On HaluEval QA (n=5,000, §4.2) the authors report in Table 1 that valid responses score higher than hallucinated ones across five embedding models, with Cohen's d from +0.92 (mpnet row) to +1.28 (minilm row) and ROC-AUC from 0.776 (mpnet row) to 0.824 (minilm row), mean d =+1.13 and mean AUC=0.806 (Table 1, Mean row). They further claim a theoretical result: dividing the spherical triangle inequality (Equation 4) by $\theta(r,c)$ yields the SGI bounds of Equation 5, from which they say it follows that discriminative power must grow with $\theta(q,c)$ — a “prediction” they regard as confirmed by Table 2, where d rises $0.61 \rightarrow 0.90 \rightarrow 1.27$ and AUC $0.721 \rightarrow 0.768 \rightarrow 0.832$ across $\theta(q,c)$ terciles. Supporting claims are cross-model agreement (mean Pearson r =0.85, mean Spearman ρ =0.87, §5.2), subgroup structure (Table 3: long responses d =+2.05, short questions d =+1.22, long questions d =+0.65), calibration (ECE=0.10, §5.5), a deliberate negative result on TruthfulQA (Table 4: AUC=0.478, d =-0.14), and a decomposition (Table 5) showing the signal is carried by $\theta(r,q)$ rather than $\theta(r,c)$.

2. Whether the conclusion holds on the data

The authors' own Table 5 refutes the value of the proposed index. This is the most serious problem and it is internal to the manuscript, requiring no external data. Table 5 reports the effect size of the

numerator alone, $d(\theta(r,q))$, as +1.50 (mpnet), +1.62 (minilm), +1.48 (bge), +1.39 (e5), +1.44 (gte). The corresponding effect sizes for the full ratio in Table 1 are +0.92, +1.28, +1.27, +1.03, +1.13. For every one of the five models, the bare angular distance from the response to the question separates the classes better than SGI does; for mpnet, the ratio destroys 0.58 of effect size, for e5 0.36. Dividing by $\theta(r,c)$ therefore adds variance without adding discrimination. Since the ratio is the object that motivates the hypersphere framing, the bounds of Equation 5, the $\theta(q,c)$ stratification of §5.3, and the whole notion of a “grounding index”, and since the single scalar it is built from outperforms it, the paper as written argues for a construction that its own measurements do not support. §5.7 observes the asymmetry but draws only an interpretive conclusion (“not actively avoiding the context”) and never compares $d(\theta(r,q))$ in Table 5 against $d(\text{SGI})$ in Table 1. That comparison must be made explicitly, and the paper’s claims restructured around whatever survives it.

The HaluEval contrast is constructed, and the observed geometry is exactly what construction would produce. §4.2 states that HaluEval QA supplies question–knowledge–answer triples with the knowledge field used as the retrieved context, and §6.3 concedes that “HaluEval hallucinations are adversarially generated.” The two classes therefore do not come from one generator under two conditions: the positives are the benchmark’s reference answers, and the negatives are synthesised text. Any systematic difference in how the two classes were *produced* — reference answers drawn from or lexically anchored in the knowledge field, negatives written by conditioning on the question — would manifest as precisely the signature reported: hallucinated items sitting nearer the question (Table 5, $d(\theta(r,q)) \approx +1.5$) with almost no movement relative to the context (Table 5, $d(\theta(r,c)) = +0.38$ to $+0.45$). The asymmetry the authors treat as evidence for a generation mechanism is equally well explained as a fingerprint of the negative-sampling procedure. As things stand, the paper cannot distinguish the two accounts, and the mechanistic language in §5.7 (“the generation process defaults to question-proximate completions when context integration fails”) and §2.4 is not licensed by a design in which no generation failure was ever observed. §6.1 concedes the causal link is unestablished; §5.7 should concede the same.

There is no test on naturally occurring hallucinations. Both datasets used (§4.2) are constructed: HaluEval by adversarial synthesis, TruthfulQA by pairing curated true and false answers. Notably, the introduction already cites RAGTruth (reference [9]) and FaithBench (reference [8]), corpora built specifically from hallucinations emitted by real systems, and neither is used. Until SGI is evaluated on responses generated by an actual RAG pipeline and labelled post hoc, the abstract’s closing claim that SGI

is “infrastructure for identifying responses that warrant verification in production RAG deployments” is not supported by anything in §5.

The “theoretical prediction” is a restatement of the bound, not a prediction. This is where the manuscript’s derived results and its empirical results are conflated, and the distinction must be drawn sharply. Equation 4 is the spherical triangle inequality; it holds for *every* triple q, c, r on S^{d-1} , makes no reference to embeddings, labels, or hallucination, and is not a claim about data. Equation 5 is Equation 4 divided by $\theta(r,c)$; it is likewise a derived, content-free feasibility region. What that region says is that SGI lies in an interval whose lower endpoint is $|k-1|$ and upper endpoint $k+1$ with $k=\theta(q,c)/\theta(r,c)$, so as $k \rightarrow 0$ the admissible interval collapses onto the single value 1. The consequence is purely mechanical: *any* quantity plotted against $\theta(q,c)$ — SGI for valid items, SGI for hallucinated items, SGI for randomly paired items — is compressed toward 1 in the low tercile and free to spread in the high tercile. Table 2 shows exactly this compression, with both class means moving away from 1 as $\theta(q,c)$ grows (valid $1.08 \rightarrow 1.19 \rightarrow 1.31$; hallucinated $0.94 \rightarrow 0.91 \rightarrow 0.88$). The monotone rise of d in Table 2 is therefore a range-widening artefact of the statistic’s own definition before it is anything about grounding, and it cannot be treated as independent corroboration. Note also that the bound in Equation 5 constrains SGI through the *ratio* k , not through $\theta(q,c)$ alone, whereas Table 2 stratifies on $\theta(q,c)$ alone and never reports the $\theta(r,c)$ distribution per tercile — so the stratification variable does not even cleanly index the quantity the bound governs. A falsifiable version of this argument needs a null model: the same tercile analysis run on label-permuted data, or on responses paired at random with contexts. If d rises monotonically there too, §5.3 measures geometry, not hallucination.

The tercile trend is additionally vulnerable as a selection effect. Terciles of $\theta(q,c)$ sort items on far more than angular separation. Table 3 shows that response length alone moves the effect size from $d=+0.95$ (short) to $d=+2.05$ (long), a swing of 1.10 — considerably larger than the 0.66 swing ($0.61 \rightarrow 1.27$) that §5.3 attributes to the triangle-inequality mechanism. If response length, question length, or topic composition correlates with the $\theta(q,c)$ tercile, the entire §5.3 result can be a reparametrisation of Table 3. No joint analysis, stratified regression, or covariate balance table across the three terciles of Table 2 is reported, so the reader cannot rule this out. The same applies to Table 3 itself: no evidence is offered that response length is balanced across the two labels, and if HaluEval’s synthesised negatives differ in length from the reference positives, the $d=+2.05$ long-response cell is partly a length artefact.

Cohen’s d is the wrong summary for this statistic, and it is reported without uncertainty. §4.1 (Statistical Analysis) states that group comparisons use Welch’s t -test *because variances are unequal*, yet

effect sizes use Cohen's d with *pooled standard deviation* — the two choices contradict each other; under heteroscedasticity, a pooled-SD d is not interpretable, and Glass's Δ or Hedges' g with a Welch-type correction is required. Independently, SGI is a ratio of arccos values: it is non-negative, right-skewed, and unbounded above as $\theta(r,c) \rightarrow 0$, which is why §4.1 has to add $\varepsilon=10^{-8}$ to the denominator. That ε admits SGI values of order 10^8 , and a handful of such points would dominate the pooled SD and hence every d in Tables 1, 2, 3, and 5. No distributional summary, no variance, no winsorisation or trimming policy, and no histogram of SGI is given anywhere. Not one of the effect sizes in Tables 1, 2, 3, or 5 carries a confidence interval, although at $n=5,000$ an interval on d is roughly ± 0.06 and would be trivial to report; likewise, no CI accompanies any AUC. And §5.1 states that “all p-values are below 10^{-50} ” while the p-value column of Table 1 reports only “<0.01” for all five rows — these should be reconciled. Finally, if the $n=5,000$ stratified sample of §4.1 draws the valid and hallucinated answers belonging to the same HaluEval question, observations are paired, and the independent-samples Welch test of §4.1 is mis-specified; §4.1 does not say which sampling was used. For TruthfulQA, the pairing is explicit (§4.1: “treating each question's correct and incorrect answers as separate instances”), so the $p=0.045$ in Table 4 comes from an unpaired test applied to paired data.

The cross-model correlation measures agreement between instruments, not validity. §5.2 reports mean $r=0.85$, min $r=0.80$ (mpnet–bge), max $r=0.95$ (bge–gte), $\rho=0.87$, and concludes in Table 1's caption that this demonstrates SGI “captures a property of the text rather than an embedding artifact”, escalating in §5.2 to “This does not happen unless SGI measures something intrinsic to the text.” That inference does not follow. All five models listed in §4.3 are English sentence-transformer bi-encoders trained with contrastive/InfoNCE-family objectives on heavily overlapping public corpora; bge-base-en-v1.5 and gte-base are both BERT-base-scale multi-stage contrastive models, so their $r=0.95$ is the *expected* null for two near-identical instruments, not the surprise §5.2 presents it as. Convergence among correlated measurements bounds reliability; validity requires a contrast with a measurement of a different kind. The decisive missing comparison is a trivial lexical baseline — token-F1 or ROUGE overlap of r with c versus r with q , or simple answer length. Such a baseline would also correlate highly across “models”, and if it matches the $d=+1.13$ / AUC=0.806 of Table 1's Mean row, the geometric apparatus of §2 adds nothing. No baseline of any kind appears in the paper.

There is no downstream detection evaluation. The paper reports effect sizes and AUC only (Tables 1, 2, 3, 4). There is no precision, recall, F1, or false-positive rate at any operating point; no threshold-selection procedure; and no cost analysis of review effort. §4.1 states plainly that “No train/test split is applied; we

report descriptive statistics on the full samples”, which means every fitted or threshold-dependent number in the paper is in-sample. That applies directly to the calibration result: §5.5 obtains ECE=0.10 after converting SGI to probabilities “via min-max normalization”, and the min-max constants are estimated on the same data the ECE is computed on, with the bin edges (10 equal-frequency bins, §4.1) also set in-sample. Moreover, the sample of §4.1 is stratified by label, i.e., approximately balanced, whereas production RAG hallucination rates are far lower; AUC=0.806 on a balanced sample implies a precision at any usable recall that the paper never computes and that would be substantially worse at a realistic base rate. The deployment guidance in §6.2 (“Use SGI as probability estimate”) is therefore ahead of the evidence, and it sits awkwardly with §6.3, which concedes that thresholds “should be calibrated on held-out data”.

An internal inconsistency in §5.5. §5.5 states that the lowest SGI decile has a hallucination rate of approximately 100% and the highest decile approximately 65%. If the analysis sample is stratified by hallucination label (§4.1) and hence roughly balanced, the decile rates must average to about 50%; a reported range whose *minimum* is 65% cannot average to 50%. Either the stratification, the decile figure (Figure 5, right panel), or the description is wrong, and the discrepancy needs to be resolved — it also bears on the ECE=0.10, since a score whose implied probabilities never fall below 0.65 is doing far less calibrated work than the phrase “at the boundary of well-calibrated” suggests. The sentence in §5.5 that the gradient is “proving that SGI can be used as a probability estimate” is an overclaim on any reading.

The TruthfulQA negative result is welcome but under-specified and mis-described. Reporting a failure case is good practice, and §5.6 deserves credit for it. However: (i) SGI as defined in Equation 3 requires a context c , and TruthfulQA supplies none; §4.1 and §4.2 never state what was used as c for the $n=800$ TruthfulQA instances, and Table 4 reports only $\theta(r,q)$ — not SGI, not $\theta(r,c)$ — which suggests the index may not have been computed at all on this dataset. If so, the “critical negative result (AUC=0.478)” of the abstract is not a result about SGI. (ii) §5.6 describes the difference as “slightly but non-significant closer” while Table 4 gives $p=0.045$ for that same comparison; these contradict each other. (iii) AUC=0.478 is described as “worse than random guessing”, but with $n=800$ it is within sampling error of 0.5 and should be reported with a confidence interval rather than characterised as below chance. (iv) §2.3 and §5.6 call the topical-versus-factual distinction a confirmed “theoretical prediction”, yet nothing in §2 derives it; it is a reasonable post-hoc interpretation and should be labelled as such.

Novelty relative to prior work needs re-scoping. §3.1 states that Gao et al. (reference [17]) already found that hallucinated responses show smaller deviations from prompts in hidden-state space, and calls this

“consistent with our semantic laziness characterization”. If the phenomenon is already on record, the contribution reduces to the ratio form of Equation 3 plus the bounds of Equation 5 — and, per Table 5 versus Table 1, the ratio is the weaker statistic. The paper should state this delta plainly rather than presenting semantic laziness as an identification.

3. Reproducibility

Missing artefacts. There is no data availability statement, no code repository, no notebook, and no random seed anywhere in the manuscript. §4.1 says that $n=5,000$ instances were “randomly sampled” from HaluEval QA’s 10,000 and that the TruthfulQA set is “ $n=800$ after filtering incomplete entries”, but no seed, no sampled-ID list, and no filtering criterion are given, so neither sample can be reconstructed. Since §4.1 also states that no train/test split was used, there is not even a nominal split specification to reproduce.

Model identification. §4.3 names the five checkpoints (all-mpnet-base-v2 768d, all-MiniLM-L6-v2 384d, bge-base-en-v1.5 768d, e5-base-v2 768d, gte-base 768d) and §4.1 pins sentence-transformers to v2.2.2, which is good practice as far as it goes. But no Hugging Face revision hash or snapshot date is given for any of the five checkpoints, and several of these repositories have been updated since release; angular distances are sensitive to the exact weights, so revision pinning is necessary. Also unstated: torch version, device, and dtype (fp16 versus fp32 changes arccos values near the domain limits), batch size, and pooling strategy. The last matters specifically: e5-base-v2 and bge-base-en-v1.5 are documented as requiring instruction prefixes (“query:” / “passage:” and the bge query instruction respectively), yet §4.1 states that texts were encoded “without additional prompting or instruction prefixes”. Using these models off-label is a defensible choice, but it must be flagged because it plausibly depresses the e5 row of Table 1 ($d=+1.03$, the second-lowest) and confounds the cross-model comparison of §5.2.

Truncation and segmentation. §4.1 introduces spaCy `en_core_web_sm` for “sentence boundary detection and basic tokenization when segmenting long contexts”, but then states that q , c , and r are each encoded “as separate strings”. It is never said when segmentation is actually applied, nor how per-segment embeddings are aggregated into the single $\phi(c)$ that Equation 3 requires (mean of normalised vectors? normalised mean? max-similarity segment?). All five models in §4.3 have a 512-token input limit, so long HaluEval knowledge fields are either segmented or silently truncated; which occurred is material to the Context Length rows of Table 3 ($d=+0.91$ / $+0.92$ / $+1.00$), where the “long” cell may have been computed on truncated text. Report the token-length distribution of c and the truncation rate.

Formula and normalisation. Equation 3 with the clipping rule and $\epsilon=10^{-8}$ from §4.1 is clear enough to implement, but Algorithm 4.1 is referenced in §4.1 and does not appear in the article body — it must be included. The normalisation used for calibration is described only as “min-max normalization” (§5.5) with no min, no max, no statement of whether the constants were computed on the full sample, and no handling rule for the large SGI values the ϵ guard permits. The subgroup cut points of Table 3 and the tercile boundaries of Table 2 are given as ranges for $\theta(q,c)$ ([0.42,0.89], [0.89,1.12], [1.12,1.57]) but not for question, context, or response length, where only the labels Short/Medium/Long appear; the numeric thresholds and the unit (characters, tokens, sentences) are needed. The definition of “stratified by hallucination label” in §4.1 must also be made precise, since it determines whether the samples are paired (see §2 above).

Figures. Figure 1 is a schematic, not data. Figure 2’s correlation matrix appears only as three prose numbers in §5.2 (mean 0.85, min 0.80, max 0.95) — the full 5×5 Pearson and Spearman matrices should be tabulated. Figure 5’s reliability diagram and decile plot are the sole evidence for ECE=0.10 and for the 100%/65% decile claim, so the underlying bin counts and rates belong in a table.

4. What specifically to fix

1. **Confront Table 5 against Table 1 directly.** Add a column or companion table placing $d(\theta(r,q))$ (Table 5: +1.50, +1.62, +1.48, +1.39, +1.44) next to $d(\text{SGI})$ (Table 1: +0.92, +1.28, +1.27, +1.03, +1.13) and the corresponding AUCs for $\theta(r,q)$ alone, which Table 5 does not report. If the numerator dominates the ratio on both metrics, say so in the abstract and in §7, and reframe the paper around $\theta(r,q)$ or justify the ratio on some ground other than discrimination.
2. **Add baselines to Table 1.** At minimum: response length in tokens; lexical overlap of r with c versus r with q (token-F1 or ROUGE-L); cosine-similarity difference $\cos(r,c)-\cos(r,q)$; and one NLI-based detector of the kind discussed in §3.3. Report d and AUC for each on the same $n=5,000$ sample. Without these, the claim of a useful signal cannot be assessed.
3. **Add a null-model control for §5.3.** Re-run the Table 2 tercile analysis with labels permuted and, separately, with responses paired to random contexts. Report d and AUC per tercile for both nulls. If the monotone rise survives label permutation, revise §2.3, §5.3, the abstract, and §7 to state that the trend is a property of the bound in Equation 5 rather than evidence for the hypothesis in Equation 6.
4. **Restate Equation 5 honestly in §2.3.** Make explicit that the bound is a derived, label-free feasibility region governed by the ratio $k=\theta(q,c)/\theta(r,c)$, not by $\theta(q,c)$ alone; that it constrains the range of SGI

rather than the separation between two populations; and that an effect size normalized by pooled SD is not determined by it. Replace “theoretical prediction confirmed empirically” (abstract, §5.3, §7) with language that separates the derivation from the measurement.

5. **Control the §5.3 tercile analysis for the §5.4 covariates.** Report, per tercile of Table 2, the distributions of $\theta(r,c)$, response length, question length, and context length, plus the label balance. Then report d for $\theta(q,c)$ with response length held fixed (stratified or via regression), since Table 3’s response-length swing (+0.95 to +2.05) exceeds Table 2’s tercile swing (+0.61 to +1.27).
6. **Balance response length across labels in Table 3.** Report the length distribution of valid versus hallucinated responses and re-estimate the Response Length rows on a length-matched subsample, so that the headline $d=+2.05$ (Table 3, Response Length / Long) can be attributed to grounding rather than to a length difference between the reference and the synthesised class.
7. **Fix the effect-size estimator and add uncertainty.** Reconcile the use of Welch’s t -test with pooled-SD Cohen’s d (§4.1, Statistical Analysis): either justify homoscedasticity with reported per-group variances or switch to Hedges’ g / Glass’s Δ . Add bootstrap 95% confidence intervals to every d and every AUC in Tables 1, 2, 3, 4, and 5.
8. **Characterise the SGI distribution and state the outlier policy.** Given the $\epsilon=10^{-8}$ denominator guard in §4.1, report the SGI quantiles per class, the count of instances with $\theta(r,c)$ below some small threshold, and whether any trimming or winsorisation was applied before computing the d values in Tables 1, 2, 3, and 5.
9. **Specify the sampling and use the matching test.** State in §4.1 whether the $n=5,000$ HaluEval instances include both answers to the same question. If they do, replace Welch’s t -test with a paired test or cluster-robust standard errors throughout Tables 1, 2, and 3. Do the same for Table 4, where §4.1 confirms the TruthfulQA construction is paired but $p=0.045$ comes from an unpaired test.
10. **Evaluate on naturally occurring hallucinations.** Add at least one corpus of hallucinations produced by a real RAG pipeline — RAGTruth (reference [9]) and FaithBench (reference [8]) are already cited in §1 — and report Table 1’s metrics on it. Until then, restrict the claims in the abstract and §6.2 about production deployment, and promote the dataset-specificity caveat from the middle of §6.3 to the abstract.
11. **Add a downstream detection evaluation.** Introduce a held-out split (replacing the “no train/test split” protocol of §4.1), select a threshold on the training half, and report precision, recall, F1, and

false-positive rate on the held-out half — both at the balanced base rate of the current sample and at realistic lower hallucination prevalences, to show what review load an operator would actually face.

12. **Recompute the calibration out of sample.** Fit the min-max transform of §5.5 and the 10 equal-frequency bin edges of §4.1 on the training half only and report ECE on the held-out half, with the min and max constants stated numerically. Soften “proving that SGI can be used as a probability estimate” in §5.5 to match whatever the held-out ECE supports.
13. **Resolve the decile inconsistency in §5.5.** The stated range of hallucination rates (approximately 100% in the lowest SGI decile down to approximately 65% in the highest) cannot average to the roughly 50% implied by the label-stratified sampling of §4.1. Publish the per-decile n and rate behind Figure 5 (right panel) and correct whichever statement is wrong.
14. **Specify the context for TruthfulQA, or withdraw the claim.** State in §4.2 what was used as c for the $n=800$ TruthfulQA instances, given that Equation 3 requires a context and Table 4 reports only $\theta(r,q)$. If SGI was not computed there, remove the description of $AUC=0.478$ as an SGI result from the abstract and §5.6 and present it as a $\theta(r,q)$ result.
15. **Correct the internal contradiction in §5.6.** The prose calls the difference “non-significant” while Table 4 reports $p=0.045$ for the same comparison. Also add a confidence interval to $AUC=0.478$ (Table 4) and, if it covers 0.5, stop describing the value as below chance.
16. **Retract or evidence the mechanism language in §5.7 and §2.4.** Statements such as “the generation process defaults to question-proximate completions when context integration fails” (§5.7) attribute a cause to a generation process that the HaluEval design never observed, since §6.3 states the negatives were adversarially generated. Align §5.7 with the honest hedge already present in §6.1.
17. **Delete or support the validity claims attached to $r=0.85$.** Specifically, the Table 1 caption (“demonstrating that SGI captures a property of the text rather than an embedding artifact”), the sentence in §5.2 beginning “This does not happen unless SGI measures something intrinsic to the text”, and the parallel claim in §7. Replace with a reliability claim, note in §4.3 that all five checkpoints share the contrastive bi-encoder family, and report the cross-model correlation of the lexical baseline from item 2 as a reference point.
18. **Report $e5$ and bge prefix handling.** §4.1 states that no instruction prefixes were used; note explicitly in §4.3 that $e5$ -base-v2 and bge -base-en-v1.5 are documented as expecting query/passage prefixes, and either re-run those two rows of Table 1 with the prescribed prefixes or discuss the confound this introduces into the §5.2 correlation analysis.

19. **Complete the reproducibility apparatus.** Include Algorithm 4.1, which §4.1 references but the article does not contain; add Hugging Face revision hashes or snapshot dates for all five checkpoints in §4.3; state the random seed and the sampled instance IDs for §4.1; state the “incomplete entries” filter that reduced TruthfulQA to $n=800$; give the numeric Short/Medium/Long cut points and their unit for every panel of Table 3; specify how segmented contexts are aggregated into $\phi(c)$ and report the truncation rate at the 512-token limit; and release the code and per-instance SGI scores.
20. **Tabulate what currently exists only in prose or as figures.** Give the full 5×5 Pearson and Spearman matrices behind Figure 2 (§5.2 quotes only 0.85, 0.80, 0.95, and $\rho=0.87$), and the bin table behind Figure 5. Reconcile §5.1’s “all p-values below 10^{-50} ” with the “ <0.01 ” entries in Table 1’s p-value column.
21. **Substantiate or drop “computationally efficient”.** The abstract and §7 claim efficiency, but no timing, throughput, or cost figure appears anywhere in §4 or §5. Add per-instance latency and cost for SGI and for each baseline from item 2, including the NLI detector, so the efficiency-accuracy trade-off can be judged.
22. **Re-scope the novelty statement in §3.1 and §7.** Since §3.1 reports that Gao et al. (reference [17]) already observed smaller deviations from prompts for hallucinated responses, state precisely what is new here relative to that finding.
23. **Language corrections.** §5.4: “Table 3 and Figure 4 allows” → “allow”; “These findings shows” → “show”. §5.6: “slightly but non-significant closer” → “non-significantly closer” (subject to item 15); “is methodologically important because establishes” → “because it establishes”. §5.5: “proving” → “indicating”. §6.2: “The signals are orthogonal” overstates §3.3’s “complementary” and should match it unless orthogonality is measured. Reference [21] (HALT-RAG) is listed without authors and should be completed.

5. Recommendation

Major revisions. The paper is clearly written, the geometry in §2.1 and §2.2 is correct, and the decision to publish a negative result (§5.6) and an explicit limitations section (§6.3) reflects good scientific practice. But the central construction is not currently supported: by the authors’ own Table 5, the numerator $\theta(r,q)$ outperforms the SGI ratio of Equation 3 on every one of the five models in Table 1; the “theoretical prediction” of §2.3 and §5.3 is a restatement of the derived feasibility region in Equation 5 rather than a falsifiable forecast, and is untested against a null model; the class contrast comes entirely from

constructed datasets (§4.2) whose negatives were adversarially generated (§6.3), so the geometric separation may be a property of the negative-sampling procedure; the cross-model agreement of §5.2 establishes reliability rather than validity; there is no baseline of any kind, no held-out split (§4.1), and no precision/recall evaluation at an operating threshold. These are addressable. Items 1–3, 10, and 11 of the list above are the load-bearing ones: a revision that runs the null-model control, adds lexical and NLI baselines, evaluates on a naturally occurring hallucination corpus, and reports held-out detection metrics would be a genuinely useful contribution even if the conclusion turns out to be that the ratio should be replaced by its numerator.

Declarations

Potential competing interests: No potential competing interests to declare.