Geometric Deviation as an Unsupervised Pre-Generation Reliability Signal: Probing LLM Representations for Answerability

Yucheng Du
University of Southern California
yuchengd@usc.edu

Abstract

A reliable language model should be able to signal, prior to generation, when a query falls outside its knowledge. We investigate whether representation geometry can provide such a *pre-generation* signal by measuring the deviation of hidden states from an answerable reference set—requiring no labeled failure data and no access to model outputs.

Across three instruction-tuned models (Llama 3.1-8B, Qwen 2.5-7B, and Mistral-7B-Instruct) and three prompt forms (MATH, FACT, CODE), we find that geometry primarily encodes *task form*. Within mathematical prompts, unanswerable inputs consistently deviate from the answerable centroid, yielding strong separation (ROC-AUC 0.78–0.84). This single-pass pre-generation signal outperforms a simple refusal baseline and compares favorably to self-consistency. It also captures cases where models do not explicitly refuse.

In contrast, no reliable geometric signal emerges for factual prompts, indicating that the effect is form-conditional rather than universal. Code prompts show large effect sizes with higher variance, suggesting partial generalization beyond mathematical form.

A layer-wise analysis reveals that the signal arises in early layers and gradually attenuates toward the output. These results suggest that answerability-related geometry is established before the final stages of generation. Together, these findings indicate that geometric deviation can serve as a lightweight *pre-generation* signal that is reliable in structured domains with formal answerability constraints, with clear boundaries on where it generalizes.

1 Introduction

Hallucination—the generation of confident but incorrect responses—remains a central reliability challenge for deployed language models (Ji et al., 2023). Detecting likely failures *before* generation is particularly valuable: a pre-generation signal can trigger abstention or human review without adding latency to the decoding process. Prior work has approached reliability estimation through uncertainty calibration (Kadavath et al., 2022), internal probing classifiers (Slobodkin et al., 2023), representation steering (Li et al., 2023), and supervised internal-state analysis (Zhang et al., 2025a,b). However, these methods either rely on labeled training data, require access to model outputs, or are sensitive to model-specific characteristics. Whether *unsupervised* representation geometry alone can function as a practical pre-generation reliability signal—without labeled failure data or access to model outputs—remains underexplored.

We investigate a minimal approach: measuring each prompt's cosine distance from the centroid of answerable-class representations, requiring no labeled failure data, no fine-tuning, and no output sampling. Our design isolates answerability from confounding surface variation using **matched pairs**, where each unanswerable prompt shares the domain, length, and syntactic form of a corresponding answerable prompt, differing only in the property causing unanswerability. We validate across **three architecturally distinct instruction-tuned models** at the same scale (7–8B parameters): Llama 3.1-8B-Instruct (Dubey et al., 2024), Qwen 2.5-7B-Instruct (Qwen Team, 2025), and Mistral-7B-Instruct-v0.3 (Jiang et al., 2023). Holding scale constant while varying architecture and alignment recipe lets us distinguish input-driven geometric signals from model-specific artefacts.

Our main findings are: (1) Within mathematical form, geometric deviation yields strong separation (ROC-AUC 0.78–0.84 across all three models), outperforming a simple refusal baseline and a multi-sample self-consistency baseline requiring $5\times$ the inference cost—including cases that refusal-based detection may not capture; (2) The effect is *form-conditional*: within factual form, no signifi-

Proceedings of the 6th Workshop on Trustworthy NLP (TrustNLP 2026), pages 353–363

cant signal emerges across any of the three models, establishing a principled boundary; (3) Within code form, large effect sizes appear across all three models, though statistical significance is mixed at the sample size studied, suggesting the phenomenon may extend beyond mathematical form; (4) A layer-wise analysis reveals the signal *peaks at early layers* and generally decreases toward the output layer, consistent across all three models, suggesting answerability-related geometry is established early in the network; (5) Strong cross-model geometric consensus on a subset of MATH-U prompts suggests the signal reflects input structure rather than model-specific geometry; (6) Behavioral responses to geometric outliers diverge across models— Qwen refuses, Llama does not explicitly refuse on the same prompts— indicating that alignment training shapes how models *act on* geometric information, not the information itself.

2 Background

Representation geometry in LLMs. LLM hidden states exhibit strong *anisotropy*: representations cluster near a dominant direction, inflating pairwise cosine similarity even for unrelated inputs (Ethayarajh, 2019). Mean-centering removes this dominant direction and restores discriminability (Godey et al., 2024). Prior work shows that structural linguistic information is geometrically encoded in Transformer representations (Hewitt and Manning, 2019), and that task-specific function vectors emerge in later layers of instruction-tuned models (Todd et al., 2023), motivating the view that geometry can reflect semantic properties beyond surface form.

Reliability signals in LLM representations. Kadavath et al. (2022) show that LLMs are well-calibrated on multiple-choice tasks. Slobodkin et al. (2023) probe for answerability in reading comprehension via supervised classifiers on context-dependent questions. Burns et al. (2022) extract truth directions via contrastive activation differences; Li et al. (2023) show that steering attention heads can elicit truthful answers. A recent survey by Xia et al. (2025) organises uncertainty estimation approaches across four paradigms. The most closely related work is PRISM (Zhang et al., 2025a) and MHAD (Zhang et al., 2025b), which use supervised probing on internal states for hallucination detection across multiple layers. PRISM trains a prompt-guided classifier on labeled hallucination examples to identify factual errors at inference time; MHAD performs deep multi-layer representation analysis using supervised training signals derived from factuality annotations. Our work differs in three respects: it requires no labels on failure or unanswerable instances (only a reference set of answerable prompts, available by construction in structured query domains), it operates strictly before generation (no output tokens needed), and it uses matched-pair construction to explicitly disentangle surface form from answerability—enabling a controlled characterisation of *when* and *where* geometric reliability signals arise, rather than learning to discriminate post-hoc from labeled failures.

Layer-wise signal in Transformers. Probing studies have found that different linguistic properties peak at different layers: syntactic information tends to emerge in middle layers, while semantic and task-level information concentrates in later layers (Hewitt and Manning, 2019). Our layer-wise analysis adds to this literature by showing that answerability geometry—a reliability-relevant property—peaks *unusually early* (layers 2–5), suggesting that the network encodes input-level structural violations before committing to a generation strategy in deeper layers.

Hallucination and output-level baselines. Semantic entropy-based methods detect hallucinations from model outputs without accessing internal states (Farquhar et al., 2024). We compare against a lightweight output-level refusal baseline, representing the information available from generation alone. Instruction-tuned models differ in their tendency to refuse versus hallucinate on unanswerable inputs (Bai et al., 2022), a distinction we investigate empirically across three models.

3 Experimental Setup

Models. We use three instruction-tuned models at the same scale (7–8B parameters): Llama 3.1-8B-Instruct (Dubey et al., 2024), Qwen 2.5-7B-Instruct (Qwen Team, 2025), and Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), all loaded via HuggingFace Transformers (Wolf et al., 2020) in `float16` precision on Apple Silicon MPS. Holding scale constant isolates architectural and alignment recipe differences. Mistral's training recipe differs from both Llama and Qwen, providing a third alignment data point.

Representation extraction. For each prompt, we extract last-layer hidden states, apply mean pooling over all input tokens, and subtract the global mean vector computed over all prompts in a given run (Godey et al., 2024). All distances are cosine distances ($1 - \cos\theta$). For the layer-wise analysis (Section 4.3), we extract mean-pooled hidden states at every layer (including the embedding layer), yielding a matrix of shape $(n_{\text{prompts}}, n_{\text{layers}}, d)$.

Prompt forms and matched-pair construction. We study three prompt forms.

MATH ($n = 50$ pairs): well-defined arithmetic, algebra, or combinatorics questions (MATH-A) paired with structurally identical variants in which a defined quantity is replaced by an undefined one (MATH-U). Unanswerability sources include: mathematically undefined operations (e.g., $\sqrt{-169}$ in the reals; $\log_1 10$; 0^0), extremal impossibilities (e.g., "the largest prime"; "the last Fibonacci number"; "the product of all positive integers"), and unknown-quantity substitutions (e.g., "the current number of active volcanoes"). Each pair preserves domain, syntactic structure, and approximate length; the sole change is the introduction of the undefined element.

FACT ($n = 10$ pairs): verifiable factual questions paired with variants referencing unknowable future events, non-existent entities, or counterfactual premises. Examples: "capital of France" / "capital of France in 2050"; "currency of Japan" / "currency of Atlantis."

CODE ($n = 30$ pairs): Python expression questions with deterministic return values (CODE-A) paired with structurally identical variants (CODE-U) whose evaluation is undefined, raises a well-typed exception, or requires unbounded computation. Examples: `max([3,1,4])` / `max([])` (well-defined / raises `ValueError`); `sum([1,2,3])` / `sum(itertools.count())` (finite / non-terminating); `hash(42)` / `hash([1,2,3])` (hashable / `TypeError`). The CODE form tests whether the geometric signal generalizes beyond the mathematical domain to a domain where unanswerability arises from type violations and semantic ill-definedness in a programming language.

In all three forms, construction rules are applied consistently: one element is changed per pair; surface structure is preserved. This design rules out length, domain, and surface form as confounds. All prompts and analysis code are released at https://github.com/yucheng-du/geom-reliability.

Analysis. For all controlled experiments, we compute each prompt's *own_dist*— cosine distance to its form's A-only centroid—as the reliability score. Centroids are computed from the A-labeled prompts only, so no U-label information enters the score construction. In a deployment setting this reference set corresponds to a small collection of prompts known to be answerable (e.g., standard queries in a domain), requiring no annotation of failures or unanswerable instances. We report one-sided permutation tests ($n_{\text{perm}} = 5000$) on the mean gap $\overline{\text{dist}}_U - \overline{\text{dist}}_A$, recomputing the centroid at each permutation to avoid null-hypothesis violations, together with Cohen's d for effect size. Mean-centering is performed jointly over all prompts within a run. For the MATH/FACT experiments, FACT and MATH prompts are mean-centered together and share the same representational reference frame. The CODE experiments were run separately and use their own mean-centering context; own_dist values for CODE are therefore not directly comparable on an absolute scale to MATH/FACT values.

For reliability prediction, we threshold own_dist at the midpoint of the mean A and mean U distances to produce a binary classifier and report ROC-AUC and F1. The refusal-keyword baseline classifies a prompt as unanswerable if the model's generated output contains any of a curated list of refusal-indicative surface tokens: *undefined*, *cannot*, *doesn't exist*, *no such*, *not defined*, *infinite*, *ValueError*, *TypeError*, *ZeroDivisionError*, and related forms. This baseline represents the information extractable from the model's output alone—requiring a completed generation pass—and serves as a practical upper bound for lightweight output-level detection. Its recall is structurally bounded: it can only fire when the model explicitly names its uncertainty, and cannot detect hallucinations where the model generates confidently incorrect responses without refusal markers.

We additionally evaluate a **self-consistency** (SC) baseline: for each prompt, we generate $k = 5$ samples at temperature 0.7 and compute a disagreement score. For MATH and CODE, we extract the final answer token from each sample and set the score to $1 - (\text{majority count}/k)$ (*an-*

swer_disagree); for FACT, where answers are free-form, we compute the mean pairwise ROUGE-1 F1 over the last-line excerpts of all $\binom{k}{2}$ sample pairs and set the score to $1-\overline{\text{ROUGE-1}}$ (*rouge_disagree*). SC requires five generation passes per prompt and accesses model outputs. We note that this disagreement-based SC is a lightweight proxy, not full semantic entropy (Farquhar et al., 2024): it relies on surface string matching of extracted answer tokens rather than semantic clustering across outputs, and thus constitutes a lower bound on what output-level uncertainty estimation can achieve; full semantic entropy remains future work. We include SC to characterise how a post-generation multi-sample baseline compares to the single-pass pre-generation geometry signal.

4 Results

4.1 Geometry Encodes Task Form

Llama and Qwen produce well-separated clusters for the three prompt categories in the uncontrolled task-structure experiment (all $p < 0.01$, permutation test on within- vs. between-group cosine distances). MATH forms the tightest cluster (within-class distance: Llama 0.332, Qwen 0.415), reflecting the high surface uniformity of arithmetic questions. Centroid analysis reveals an asymmetry: the FACT–MATH centroid cosine (≈ -0.84 to -0.85) indicates near-orthogonality after mean-centering, while FACT–UNKNOWN is positive and moderately close ($+0.41$ Llama, $+0.58$ Qwen). The UNKNOWN cluster therefore aligns with FACT, not MATH— a pure form effect: math-form unanswerable prompts are pulled toward the MATH centroid, while fact-form unanswerable prompts align with FACT.

In the controlled experiments, the CODE form occupies a distinct cluster well-separated from both MATH and FACT, suggesting that programming language structure is encoded geometrically in instruction-tuned representations distinctly from natural-language forms. CODE within-class distances are highest of the three forms (Llama ≈ 0.889, Qwen ≈ 0.815, Mistral ≈ 0.875), reflecting greater surface heterogeneity in Python expressions relative to arithmetic questions. The CODE–MATH centroid distance is large (both forms produce tight but geometrically distant clusters), whereas CODE and FACT exhibit intermediate separation. This structure is consistent across all three models, as shown in Figure 1, suggesting that the task-form encoding is not an artifact of a specific architecture.

4.2 Answerability Signal Within Form

Table 1 reports the full controlled answerability results for all three models and all three forms.

Form	Model	n	dist_A	dist_U	Δ	p
MATH	Llama	50	0.676	1.055	+0.379	**<.0001**
	Qwen	50	0.652	1.042	+0.390	**<.0001**
	Mistral	50	0.668	1.038	+0.370	**<.0001**
FACT	Llama	10	0.326	0.406	+0.080	0.498
	Qwen	10	0.303	0.361	+0.058	0.566
	Mistral	10	0.305	0.402	+0.097	0.360
CODE	Llama	30	0.889	1.195	+0.306	0.113
	Qwen	30	0.815	1.229	+0.414	**0.008**
	Mistral	30	0.875	1.168	+0.294	0.155

Table 1: Cosine distance to answerable centroid (own_dist) for matched pairs. $\Delta = \text{dist}_U - \text{dist}_A$; p-values from one-sided permutation test ($n_{\text{perm}} = 5000$). MATH/FACT: joint mean-centering; CODE: separate mean-centering context (see §3).

MATH ($n = 50$ pairs). All three models show highly significant separation between MATH-A and MATH-U at the expanded sample size ($p < 0.0001$, Cohen's d ranging from $+1.12$ to $+1.41$). The MATH-U centroid distance (≈ 1.04–1.06) is substantially higher than MATH-A (≈ 0.65–0.68), a gap of $\approx +0.37$–$+0.39$ that is consistent across architectures. We attribute this separation to the structural-contradiction hypothesis: mathematical unanswerability forces a representation toward an undefined region of the tight MATH attractor, producing systematic centroid deviation. Figure 2 shows the full own_dist distributions: MATH-U exhibits substantially elevated values with heavy tails, while FACT-A and FACT-U distributions overlap completely—directly visualising the null FACT result.

FACT ($n = 10$ pairs). No significant separation emerges for any model ($p = 0.36$–0.57; Cohen's $d = +0.44$–$+0.76$). The null result is not underpowered: at $n = 10$ pairs, the MATH effect was already $p < 0.01$ in the original experiments. Factual unanswerability (future events, non-existent entities) is syntactically indistinguishable from ordinary factual questions and does not disrupt the FACT cluster geometry.

CODE ($n = 30$ pairs). Effect sizes are large and consistent across models ($d = +1.01$, $+1.31$, $+1.14$ for Llama, Qwen, Mistral), but statistical significance is mixed: Qwen reaches $p = 0.008$;

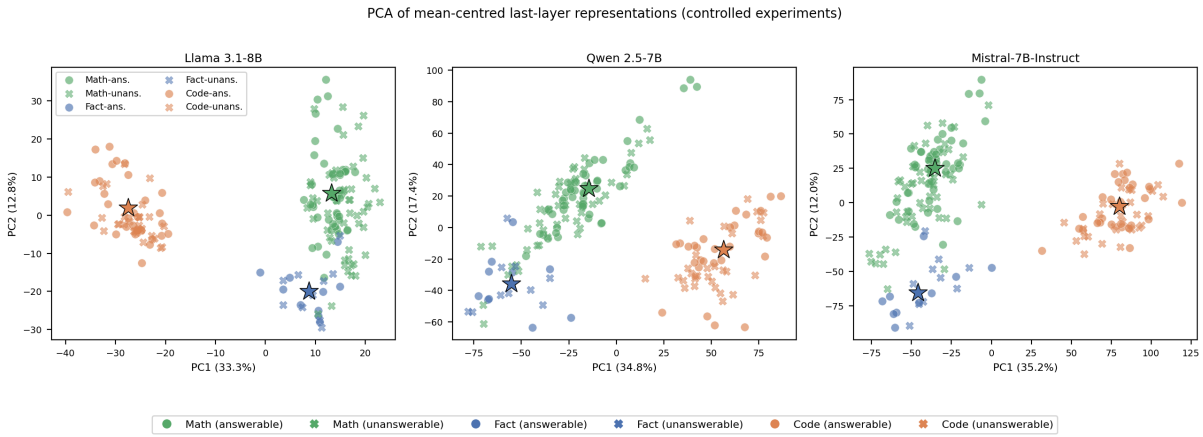

Figure 1: PCA of mean-centred last-layer representations (controlled experiments) for all three models. MATH (green), FACT (blue), and CODE (orange) occupy distinct geometric regions. Filled circles: answerable; crosses: unanswerable. Stars mark answerable-class centroids. MATH forms the tightest cluster; CODE is geometrically distant from MATH and shows higher within-class spread, consistent with greater surface heterogeneity in Python expressions. The three-form separation is consistent across all three architectures.

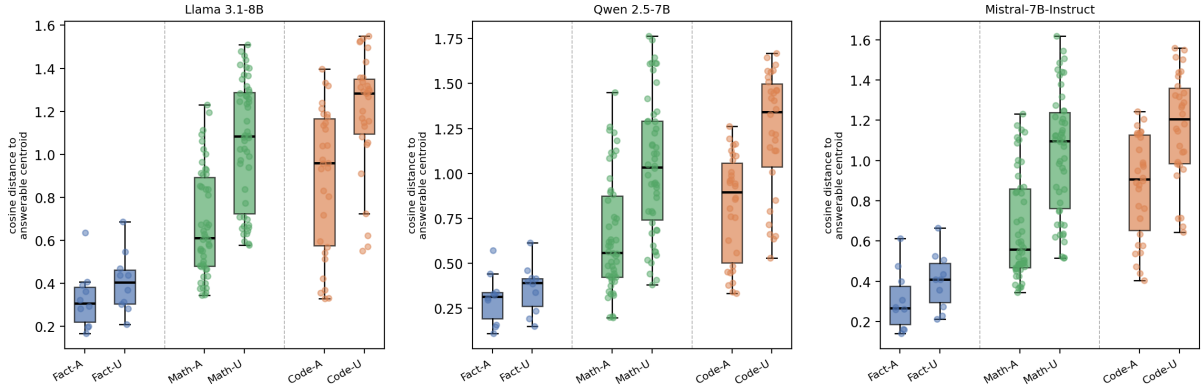

Figure 2: Distribution of own_dist (cosine distance to answerable centroid) for all three models and all three prompt forms. MATH-U and CODE-U distributions are substantially elevated relative to their answerable counterparts, consistent across all three models. FACT-A and FACT-U distributions largely overlap, consistent with the non-significant permutation tests. CODE distances are not directly comparable to MATH/FACT values as they are computed under separate mean-centering contexts. Individual prompts shown as jittered points.

Llama and Mistral are $p = 0.11$–0.16. The large d values alongside marginal p-values indicate higher within-group variance in the CODE domain: some CODE-A prompts already occupy high-deviation positions (e.g., those with unusual expression structure), widening the baseline variance and reducing power relative to MATH. We interpret this as evidence that a similar phenomenon exists in the CODE domain but requires larger n to reach conventional significance.

4.3 Layer-wise Signal Profile

To understand *where* in the Transformer stack the answerability signal arises, we extract mean-pooled hidden states at every layer for all 20 MATH matched pairs and compute the per-layer gap $\delta_l = \overline{\text{dist}}_U^{(l)} - \overline{\text{dist}}_A^{(l)}$ for all three models.

Figure 3 shows δ_l as a function of layer index. The pattern is consistent across all three models: the gap rises sharply from the embedding layer, **peaks at an early layer** (layer 2 for Llama, layer 5 for Qwen, layer 4 for Mistral; peak $\delta \approx 0.98$–1.09), and then **generally decreases** through subsequent layers to the final layer (last-layer $\delta \approx 0.44$–0.48). The last layer retains a large, practically useful gap, but is the *minimum* among all middle layers—not the maximum.

Figure 3 (bottom) clarifies the mechanism underlying the gap profile: MATH-U representations diverge from the answerable centroid from the earliest attention layers and sustain that distance

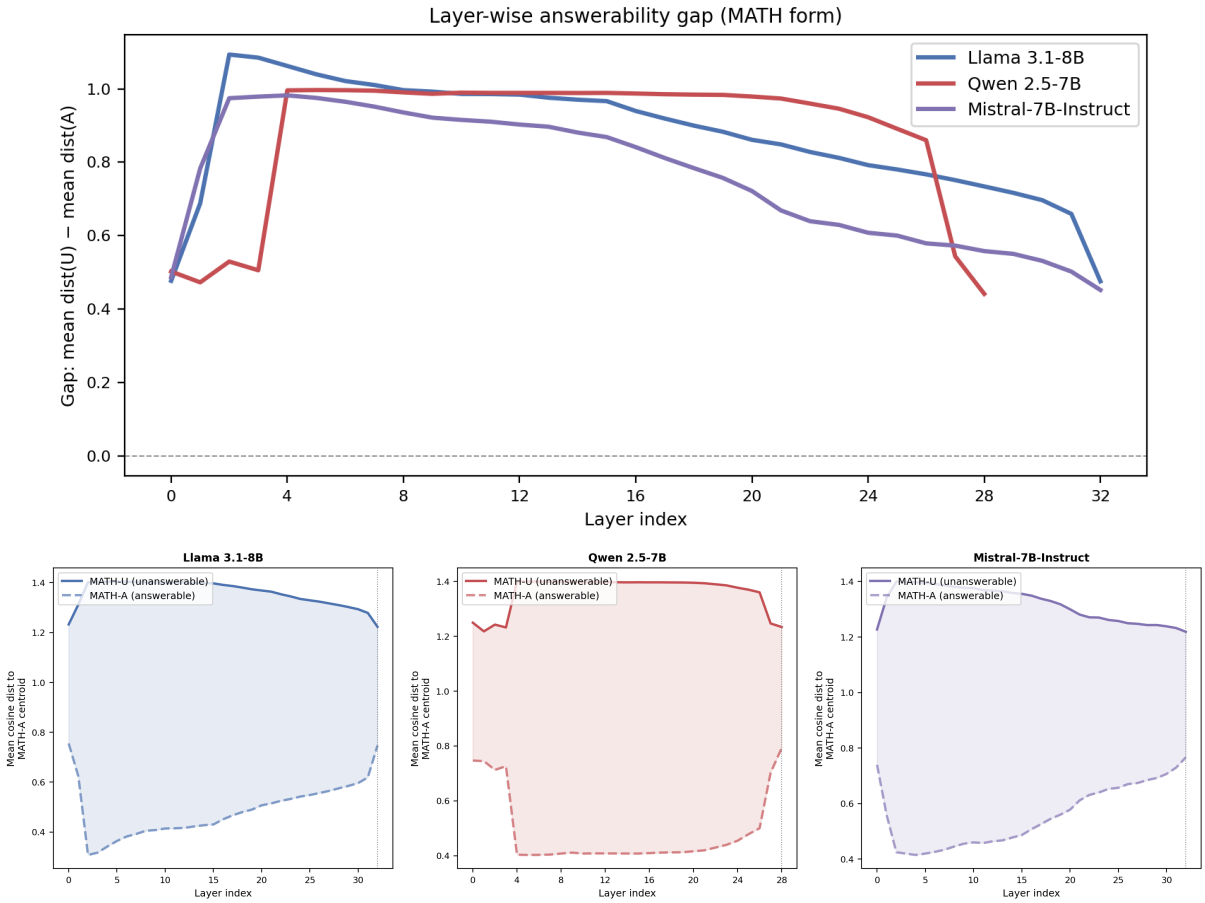

Figure 3: *Top*: Layer-wise answerability gap δ_l for MATH matched pairs ($n = 20$). All three models peak at layers 2–5 and generally decrease toward the last layer. *Bottom*: Absolute own_dist traces for MATH-U (solid) and MATH-A (dashed). The gap narrows because MATH-A rises in deeper layers, not because the MATH-U signal decays.

throughout the network. The overall decrease in δ_l is driven by the answerable class drifting *toward* the unanswerable class as depth increases, not by the unanswerable signal weakening. This suggests the network progressively adapts MATH-A representations toward a generation-ready state that incidentally reduces their distance to the answerable centroid, while MATH-U representations remain anchored in a structurally anomalous region.

This pattern indicates that answerability geometry is an emergent property of the earliest attention layers, attenuating as the network adapts representations toward generation readiness. Using the last layer in our main experiments follows prior work convention; the layer-wise profile suggests earlier layers could yield stronger classifiers if layer selection were optimized (Section 5).

4.4 Geometry–Behavior Alignment

We annotate model outputs for the original 20 MATH-U prompts (Table 2).

Model	REFUSE	PARTIAL	HALLUC
Llama 3.1-8B	5/20	5/20	10/20
Qwen 2.5-7B	7/20	3/20	10/20

Table 2: Behavioral annotation for MATH-U prompts ($n = 20$). Annotation based on qualitative reading of generated output.

Both Llama and Qwen hallucinate on 10/20 prompts. Qwen refuses more often (7/20 vs. 5/20 for Llama); Llama partially answers more (5/20 vs. 3/20).

Geometric deviation predicts behavior within each model: all four Llama prompts with own_dist > 1.2 (drifted to the FACT centroid) produce hallucination or partial answers, with zero refusals. For Qwen, the five REFUSE cases cluster within the

MATH cluster (non-drifted), while the two highest-deviation prompts (m07u: "next prime after the largest prime"; m10u: "π for a square") are correctly identified as undefined and refused. The **critical divergence**: these same prompts show the highest own_dist in *both* models, yet Llama hallucinates while Qwen refuses.

At the expanded scale ($n = 50$ MATH-U prompts), **19 of 50 prompts are misassigned to the FACT centroid in all three models simultaneously**. These 19 prompts share a common property: they involve extremal or infinite mathematical objects ("the largest prime," "the last Fibonacci number," "the average of all positive reals," "the product of all positive integers") or operations on unknown future/unbounded quantities. The near-perfect cross-model geometric consensus on these prompts— across three different architectures and training recipes—provides strong evidence that the signal reflects input-level structural properties rather than any model-specific geometry.

The 19 consensus-drift prompts fall into three structural categories: (i) extremal or infinite objects ("the largest prime," "the last Fibonacci number"), (ii) unbounded aggregates ("the exact sum of all natural numbers to infinity"), and (iii) unknown-quantity substitutions ("17 multiplied by the current moons of Jupiter"). Categories (i) and (ii) account for most cases, consistent with representations detecting *formal* impossibility rather than epistemic difficulty.

The same-geometry, different-behavior pattern between Llama and Qwen on shared outliers further supports the view that alignment training modulates how models *act on* geometric information, not the information itself: the representation-level signal of "this is anomalous" is present in all three models; the behavioral choice of whether to refuse, partially answer, or hallucinate is model-specific.

4.5 Reliability Prediction Evaluation

Table 3 evaluates own_dist as an unsupervised binary classifier distinguishing answerable from unanswerable matched pairs, compared against the refusal-keyword baseline.

MATH. Geometry substantially outperforms both baselines across all three models. Against the refusal baseline (AUC 0.63–0.73), geometry achieves AUC 0.78–0.84: refusal suffers from low recall, as only a fraction of MATH-U prompts trigger explicit refusal keywords. Against the disagreement-based SC baseline (AUC 0.30–0.62), the margin is even larger: instruction-tuned models tend to hallucinate *consistently*, producing the same incorrect answer across all five samples and yielding near-zero disagreement despite high geometric deviation. Geometry captures this pattern pre-generation, whereas SC—which detects output variance—cannot distinguish confident hallucination from correct answers in this regime; a stronger semantic entropy baseline could narrow this gap.

Form	Model	Geometry		SC	Refusal	
		AUC	F1	AUC	AUC	F1
MATH	Llama	**0.841**	**0.714**	0.624	0.630	0.413
	Qwen	**0.782**	**0.694**	0.296	0.710	0.592
	Mistral	**0.826**	**0.714**	0.524	0.730	0.630
FACT	Llama	0.690	0.632	0.460	0.550	0.182
	Qwen	0.660	0.700	0.000	**0.750**	**0.667**
	Mistral	0.710	0.700	0.470	0.550	0.308
CODE	Llama	**0.774**	**0.758**	0.441	0.633	0.421
	Qwen	0.818	0.733	0.369	**0.850**	**0.830**
	Mistral	**0.796**	**0.689**	0.497	0.733	0.636

Table 3: ROC-AUC and F1 for answerability prediction. **Geometry**: own_dist (pre-generation, zero samples); **SC**: self-consistency disagreement score (post-generation, 5 samples); **Refusal**: keyword classifier on single generated output (post-generation). MATH: $n = 50$ pairs; FACT: $n = 10$ pairs; CODE: $n = 30$ pairs. F1 threshold: midpoint of mean A and mean U own_dist (Geometry); SC F1 omitted—oracle-threshold F1 is 0.667 across all conditions, indicating SC is a near-constant classifier on this task.

FACT. Geometry yields modest AUC (0.66–0.71), consistent with the non-significant permutation tests. Qwen's refusal baseline (0.75) outperforms geometry here, reflecting Qwen's tendency to explicitly refuse future-event questions. Llama and Mistral refusal baselines are near-chance (0.55), reflecting their tendency to answer rather than refuse. SC is weakest on FACT: Qwen SC AUC $= 0.000$, because Qwen gives consistent refusal-style responses even to answerable factual questions, eliminating any disagreement signal. At $n = 10$ pairs, FACT AUC estimates carry high variance and should be treated as exploratory; the permutation tests ($p > 0.34$ across all three models) provide the more reliable evidence that no systematic geometric signal exists for factual unanswerability at the sample sizes studied.

CODE. For Llama and Mistral, geometry (0.77–0.80) outperforms both SC (0.44–0.50) and refusal (0.63–0.73). Qwen is the exception: its refusal

baseline (0.85) exceeds geometry (0.82), because Qwen explicitly names exception types ("this raises a `TypeError`") in its outputs for ill-typed CODE-U prompts, making keyword detection highly informative. This Qwen-specific behavior mirrors its elevated refusal rate in MATH and is consistent with Qwen's alignment recipe producing more explicit uncertainty acknowledgment. SC remains the weakest signal across all three models on CODE (AUC 0.37–0.50), confirming that output variance alone cannot reliably distinguish well-typed from ill-typed expressions when the model generates plausible-sounding but incorrect responses.

5 Discussion

Form dominates; answerability disrupts only when structural. Task form appears to be the primary organiser of last-layer geometry. Answerability appears as a secondary, *conditional* signal: it is most visible when unanswerability creates a structural inconsistency within the form—applying arithmetic to an undefined quantity may push a representation toward an unfamiliar region within the tight MATH cluster, whereas factual unanswerability (future events, counterfactuals) is syntactically indistinguishable from ordinary factual questions and leaves the geometric cluster intact. The CODE results are consistent with this view: ill-defined Python expressions (type errors, non-terminating operations) show a large geometric effect, though more data are needed to confirm it reaches conventional significance. The MATH/FACT asymmetry is consistent with this form-attractor account.

Answerability signal is an early-layer phenomenon. The layer-wise profile—peaking at layers 2–5 and generally decreasing thereafter—contrasts with prior probing work that finds semantic properties strongest in later layers (Hewitt and Manning, 2019). The attenuation is asymmetric (Figure 3, bottom): MATH-U distances remain elevated throughout; MATH-A distances rise in deeper layers, narrowing the gap from below rather than above. Deeper layers do not erase the anomaly signal but "normalise" answerable inputs toward a generation manifold, trading geometric separability for generation readiness. For early-warning systems, layer-2–5 activations may yield stronger answerability classifiers than the final layer at the cost of monitoring intermediate activations.

Geometry as a form-conditional reliability indicator. The MATH AUC results (0.78–0.84 across three models) establish that unsupervised geometric deviation is a viable pre-generation reliability signal in settings where unanswerability disrupts form structure. The refusal baseline can only detect failure *after* the model has decided to refuse; geometry operates over the full distribution of unanswerable inputs, including those the model does not explicitly refuse. Cross-model geometric consensus on a subset of MATH-U prompts further suggests that the signal reflects input-level structure rather than model-specific geometry, widening potential applicability. The self-consistency comparison sharpens this point: even with $5\times$ the inference cost and post-generation access, SC achieves AUC of only 0.30–0.62 on MATH—substantially below geometry (0.78–0.84). The failure mode is systematic: instruction-tuned models hallucinate *consistently*, producing the same wrong answer across all five samples and thus yielding near-zero disagreement. Geometry captures structural anomaly *before generation* on these consistently-answered prompts, where output variance is uninformative. Outside structurally disruptive settings, output-level signals such as semantic entropy (Farquhar et al., 2024) or calibrated confidence (Kadavath et al., 2022) may be more informative.

Alignment shapes the geometry–behavior link. The Llama/Qwen divergence on geometrically anomalous prompts—identical geometry, divergent behavior—suggests that alignment training shapes how models *respond to* geometric anomaly signals rather than altering the signals themselves (Bai et al., 2022). All three models agree geometrically on the 19-prompt consensus set even while differing in behavioral response. Directly testing this base-versus-instruction-tuned hypothesis remains future work.

6 Conclusion

We show that geometric deviation from an answerable reference set can serve as a pre-generation signal for answerability, particularly when unanswerability introduces structural inconsistencies within a prompt's form. Across three instruction-tuned models, the signal yields strong separation on MATH (ROC-AUC 0.78–0.84), no reliable signal on FACT, and large but variable effects on CODE—a form-conditional pattern. A layer-wise analysis localises the signal to early layers (2–5),

suggesting answerability-related geometry is established before the final stages of generation, with strong cross-model consensus on which prompts are anomalous.

The approach requires no labeled failure data, operates prior to generation, and is consistent across architecturally distinct models. The form-dependence—strongest for structurally disruptive unanswerability, weaker for open-domain factual queries—motivates combining geometric and output-based reliability signals in future work.

7 Limitations

Scale. Sample sizes are modest by benchmark standards ($n = 50$ matched pairs for MATH, $n = 10$ for FACT, $n = 30$ for CODE). FACT and CODE AUC estimates carry high variance. The CODE results in particular—large effect sizes but mixed significance—suggest the phenomenon exists but requires larger n (we estimate $n \gtrsim 80$–100 pairs) to reach conventional significance. Future work should validate on larger matched-pair sets and held-out benchmarks with independently verified answerability labels.

Baselines. We compare against a refusal-keyword proxy and a self-consistency disagreement baseline (5 samples per prompt). SC is substantially weaker than geometry on MATH (AUC 0.30–0.62 vs. 0.78–0.84), confirming that output variance is insufficient for this failure mode. However, stronger baselines remain untested: full semantic entropy (Farquhar et al., 2024), which clusters semantically equivalent outputs rather than surface-identical answers, and supervised probing methods such as PRISM (Zhang et al., 2025a) and MHAD (Zhang et al., 2025b) require labeled training data but may outperform our unsupervised signal on some form-condition combinations.

Layer selection. Our main results use the last layer for comparability with prior work. The layer-wise analysis (Section 4.3) shows that earlier layers (2–5) carry a stronger signal. Systematic comparison of pooling strategies (last layer, early layer, CLS token, last-token) across tasks is needed to identify optimal representation extraction.

Annotation. Behavioral outputs (Section 4.4) are labeled by a single annotator; inter-annotator agreement was not measured, and Mistral behavioral annotation was not performed. A follow-up study with multiple annotators would strengthen the geometry–behavior analysis.

Alternative explanations. Two confounds cannot be fully excluded: (i) MATH-U and CODE-U prompts may be lexically unusual independently of answerability, producing out-of-distribution representations; (ii) higher own_dist may partly reflect greater intra-class variance rather than a systematic centroid shift. Controlling for perplexity is a concrete next step.

Model and alignment scope. All three models are in the 7–8B range; scaling behavior and the impact of different alignment recipes (comparing base and RLHF-tuned variants of the same model family) remain open questions.

Probe choice. Our negative result on FACT relies on a single unsupervised probe: cosine distance to the answerable-class centroid on last-layer mean-pooled representations. We do not evaluate whether more sophisticated probes—e.g. PCA-projected directions, learned hyperplanes, or activations pooled from earlier layers—recover a signal on factual unanswerability at the sample size studied. The form-conditional pattern we report may therefore partly reflect this probe choice rather than an intrinsic property of the representations. Characterising how probe complexity trades off against form-domain coverage is left to future work.

8 Broader Impact

This work investigates whether internal representation geometry of LLMs can serve as an unsupervised reliability signal, with potential applications in hallucination detection and deployment safety monitoring. A pre-generation signal that fires before output is produced could complement generation-based uncertainty methods, particularly in latency-sensitive or safety-critical settings.

Our findings are encouraging but bounded: the geometric signal is reliable for structurally disruptive answerability failures (mathematical undefined operations, ill-typed code expressions) but not for general factual unanswerability. Practitioners should not deploy geometric deviation as a universal hallucination detector based on these results alone.

The layer-wise finding—that early layers carry stronger answerability signals than the last layer—suggests that lightweight online monitoring of intermediate activations could serve as a more efficient pre-generation filter than full forward-pass representation extraction.

We note that representation probing methods,

including ours, could in principle be used adversarially—for example, to construct prompts that bypass geometric detection while still causing hallucinations. However, the specificity of the signal to structural form disruption limits this concern in practice. No personal data was used in this study; all prompts are researcher-constructed.

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, and Tom Henighan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2022. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, and Aiesha Letman. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 55–65. Association for Computational Linguistics.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630.

Nathan Godey, Éric de la Clergerie, and Benoît Sagot. 2024. Anisotropy is inherent to self-attention in transformers. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, pages 35–48. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, and Lucile Saulnier. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Deep Ganguli, Jackson Kernion, Liane Lovitt, Andy Chen, Tom Brown, Jared Kaplan, Jack Clark, and Dario Amodei. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. *arXiv preprint arXiv:2306.03341*.

Qwen Team. 2025. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Aviv Slobodkin, Omer Goldman, Avi Caciularu, Ido Dagan, and Shauli Ravfogel. 2023. The curious case of hallucinatory (un)answerability: Finding truths in the hidden states of over-confident large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3607–3625. Association for Computational Linguistics.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. 2023. Function vectors in large language models. *arXiv preprint arXiv:2310.15213*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.

Zhiqiu Xia, Jinxuan Xu, Yuqian Zhang, and Hang Liu. 2025. A survey of uncertainty estimation methods on large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21381–21396. Association for Computational Linguistics.

Fujie Zhang, Peiqi Yu, Biao Yi, Baolei Zhang, Tong Li, and Zheli Liu. 2025a. Prompt-guided internal states for hallucination detection of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21806–21818. Association for Computational Linguistics.

Luan Zhang, Dandan Song, Zhijing Wu, Yuhang Tian, Changzhi Zhou, Jing Xu, Ziyi Yang, and Shuhao Zhang. 2025b. Detecting hallucination in large language models through deep internal representation analysis. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 8357–8365.

A Example Matched Prompt Pairs

Tables 4 and 5 show example matched pairs for MATH and CODE respectively.

Type	Prompt
Math-A	What is 17 multiplied by 19?
Math-U	What is 17 multiplied by the current number of moons of Jupiter?
Math-A	What is the next prime number after 29?
Math-U	What is the next prime number after the largest prime number?
Math-A	What is the value of π to 2 decimal places?
Math-U	What is the value of π for a square?
Math-A	What is the sum of the first 10 natural numbers?
Math-U	What is the exact sum of all natural numbers from 1 to infinity?
Math-A	What is the 6th Fibonacci number?
Math-U	What is the last Fibonacci number?
Math-A	What is the GCD of 24 and 36?
Math-U	What is the GCD of π and $\sqrt{2}$?

Table 4: Example MATH matched pairs. Math-U prompts encode ill-defined operations or extremal impossibilities; the sole change per pair is the undefined element.

Type	Prompt
Code-A	What does `len([1, 2, 3])` return in Python?
Code-U	What does `len(itertools.count())` return in Python?
Code-A	What does `max([3, 1, 4, 1, 5])` return in Python?
Code-U	What does `max([])` return in Python?
Code-A	What does `sum([1, 2, 3, 4, 5])` return in Python?
Code-U	What does `sum(itertools.count())` return in Python?
Code-A	What does `hash(42)` return in Python?
Code-U	What does `hash([1, 2, 3])` return in Python?
Code-A	What does `bin(10)` return in Python?
Code-U	What does `bin(3.14)` return in Python?
Code-A	What does `divmod(10, 3)` return in Python?
Code-U	What does `divmod(10, 0)` return in Python?

Table 5: Example CODE matched pairs. Code-U prompts involve non-terminating computation, type errors, runtime exceptions, or operations on undefined objects. The sole change per pair is the introduction of the ill-defined element.