Geometric Instability of Hidden-State Trajectories Predicts Reasoning Failures in Large Language Models

Hoda Fakharzadehjahromy[1*], **Andreas Bueff**[1], **Fredrik Heintz**[1], **Mattias Tiger**[1]
Qing Li[2], **Jiaqi Li**[3], **Jiahui Geng**[1*]
[1]Linköping University [2]University of Groningen [3]The University of Hong Kong

Abstract

Large language models often produce fluent reasoning chains that arrive at incorrect answers, and these failures are hard to detect from output probabilities alone. We show that a strong signal of correctness is reflected in the *geometry of internal computation*: hidden-state trajectories across transformer layers differ between correct and incorrect reasoning, with correct solutions following smooth, efficient paths and failures exhibiting elevated curvature, abrupt directional changes, and increased geodesic deviation. We capture this signal through VANE, five parameter-free geometric features (velocity, acceleration/jerk, curvature, geodesic deviation, and token coherence) computed from a single forward pass over layer-wise hidden states. Across six models (1.5B to 72B parameters) and five benchmarks spanning mathematics, code, and verbal reasoning, VANE achieves 70.7 to 96.6 AUROC and exceeds token log-probability on all 30 model–benchmark pairs (mean $+25$ percentage points, up to $+50$ percentage points). The geometric signal is near-independent of output confidence ($r{=}-0.26$), and a base-model control shows it emerges with instruction tuning. As a downstream application, filtering geometrically unstable outputs raises accuracy to 96.1% at 50% coverage on a 4-bit quantized 72B model. These results show trajectory geometry is a reliable single-pass indicator of reasoning correctness, even when the model is confidently wrong. Code is available at `https://github.com/hodfa840/geometric-instability-of-hidden-state-trajectories`.

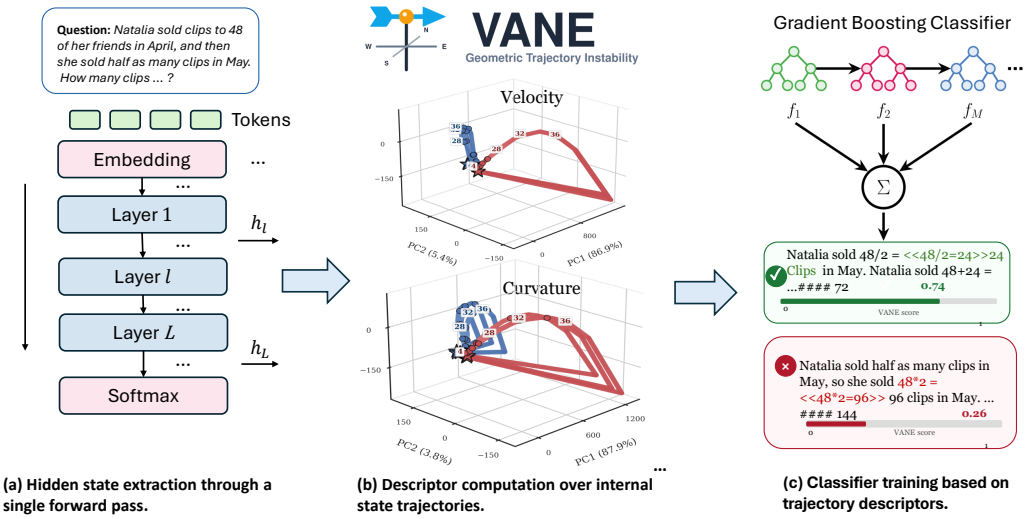
1 Introduction

Large language models (LLMs) often produce fluent reasoning that lands on the wrong answer [Song et al., 2026]. For real-world deployment, a key risk is that models do not know when they are wrong. In our experiments, self-assessment via P(True) [Kadavath et al., 2022] is barely better than random guessing, with AUROC between 41.1 and 68.4. Token-level confidence performs slightly better, but it saturates under instruction tuning and tends to reward fluent-but-wrong generations [Xiong et al., 2024, Belrose et al., 2023]. While leveraging internal activations and trajectory patterns has shown promise in detecting factual hallucinations, its applicability to complex reasoning failures remains underexplored [Song et al., 2026].

To bridge this gap, we view transformer computation through its layer-wise trajectory. Each transformer layer applies a residual update, $h_{\ell+1} = h_\ell + f_\ell(h_\ell)$, so the sequence $\{h_0, h_1, \ldots, h_L\}$ can be seen as an orbit of an iterated map through representation space. Recent studies support this idea: representational trajectory curvature has been used to study whether LLM trajectories straighten with depth [Hosseini and Fedorenko, 2023] and has been linked to prediction difficulty

*Corresponding authors: `hoda.fakhar@liu.se`, `jiahui.geng@liu.se`

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) Hidden state extraction through a single forward pass.

(b) Descriptor computation over internal state trajectories.

(c) Classifier training based on trajectory descriptors.

Figure 1: **Overview of the VANE framework for reasoning failure detection.** For a given input, (a) hidden states are extracted across all transformer layers in a single forward pass; (b) geometric trajectory descriptors (e.g., velocity, curvature) are computed over the per-layer hidden-state sequence to characterize the model's internal dynamics. The 3D visualizations in (b) show per-layer mean-pooled hidden states jointly projected to three components via PCA, with correct (blue) and incorrect (red) traces in a shared coordinate frame. (c) A gradient boosting classifier is fit once per model on these descriptors to distinguish correct from failed traces, producing the corresponding VANE score at inference time. VANE requires only one forward pass and no oracle layer selection; the PCA projection is used for visualization only.

and uncertainty [King et al., 2026]. In parallel, velocity-based analyses of hidden-state jumps show that computation can be unevenly distributed across layers, and that such layer-wise contribution imbalance may affect model capabilities [Shibata et al., 2026]. This motivates our central question: *can geometric instability along the hidden-state trajectory reveal reasoning failures that are not apparent from output confidence alone?*

We instantiate this idea with **VANE**, a reasoning-failure detection framework that reads the model's computational body language from hidden-state trajectory geometry. At its core, VANE uses five compact trajectory descriptors that capture complementary forms of geometric instability. For each layer ℓ and generated token t, we define the displacement $\Delta_\ell^{(t)} = h_{\ell+1}^{(t)} - h_\ell^{(t)}$ and the unit tangent $T_\ell^{(t)}$. From these we compute five scalars: (i) **velocity** $V_\ell = \|\Delta_\ell\|$; (ii) **representational jerk** $J_\ell = \|\Delta_{\ell+1} - \Delta_\ell\|$, the second-order difference of displacements; (iii) **Frenet–Serret curvature** $K_\ell = \|T_{\ell+1} - T_\ell\|$, capturing local directional change; (iv) **geodesic deviation** G_ℓ, measuring global departure from the input-to-output chord; and (v) **token coherence** C_ℓ, capturing cross-token directional agreement. VANE has no learned parameters at the feature stage and is agnostic to tokenizer and unembedding choice. A lightweight single-pass classifier maps this feature vector to a prediction score. VANE is not a feature search: all five metric definitions are fixed, parameter-free, and derived from the residual-update structure, with no learned representation or architecture-specific tuning.

We evaluate VANE on six models, ranging from 1.5B to 72B parameters, and five benchmarks covering math, code, and logical reasoning. VANE obtains AUROC scores between 70.7% and 96.6%. It outperforms log-probability on all 30 model–benchmark pairs, with an average gain of 25 percentage points and a maximum gain of 50 points. It also outperforms Logit-Lens [nostalgebraist, 2020], P(True) [Kadavath et al., 2022], static-representation probes [Alain and Bengio, 2017], and INSIDE [Chen et al., 2024] in most settings. Compared with the HS-Probe oracle, VANE matches or performs better on 23 of 30 settings. We also test cross-task transfer. A classifier trained only on GSM8K transfers to MATH-500, HumanEval, and CRUXEval without retraining. It remains competitive across tasks, and often outperforms other baselines, with the largest gain on Qwen3-14B. VANE also enables selective prediction: when the model abstains on geometrically unstable outputs,

retained accuracy rises by up to 19.8 percentage points at 70% coverage (Qwen3-14B) and reaches 96.1% at 50% coverage (Qwen2.5-72B, 4-bit). In contrast, log-probability-based abstention *decreases* accuracy on most models.

Contributions. Overall, our contributions are as follows: (i) We identify a consistent *geometric instability* signal in layer-wise hidden-state trajectories, where reasoning failures exhibit erratic directional shifts. We further propose VANE, a framework that extracts geometric metrics from the forward pass to quantify trajectory geometry and enable efficient failure prediction. (ii) Across 30 model–benchmark combinations, VANE consistently outperforms established baselines. It also exhibits partial cross-task transfer, with the largest gain on Qwen3-14B: a classifier trained exclusively on a single math task remains competitive on code and language reasoning without retraining. (iii) We show that the geometric signal is near-independent of token-level confidence, resides in the layer-wise trajectory rather than in any single layer, and emerges primarily from instruction tuning, independent of model scale or architecture. (iv) We show that VANE enables robust selective prediction as a lightweight abstention mechanism. Retaining only geometrically stable outputs raises accuracy by up to 19.8 percentage points at 70% coverage and achieves 96.1% accuracy at 50% coverage on a 4-bit quantized 72B model.

2 Related Work

Geometry and dynamics of transformer representations. A growing body of work interprets transformer computation through the geometry of internal representations and their evolution across depth. Feed-forward and residual blocks have been studied as mechanisms that progressively transform representations and shape predictions in vocabulary space [Geva et al., 2021, 2022, Dar et al., 2023]. In parallel, residual architectures admit a dynamical-systems view, where the layer-wise update $h_{\ell+1} = h_\ell + f_\ell(h_\ell)$ can be read as an iterated map over representation space [Chen et al., 2018, Lu et al., 2019]. Other work shows that hidden-state geometry changes systematically with depth, including shifts in intrinsic dimensionality, neighborhood structure, and local representation geometry [Ethayarajh, 2019, Valeriani et al., 2023, Yin et al., 2024]. Most closely related to our motivation are studies of representational trajectories: Hosseini and Fedorenko [2023] show that LLM sentence trajectories tend to straighten across layers, with curvature linked to prediction difficulty, while King et al. [2026] further connect representational curvature to behavioral uncertainty. Complementarily, Shibata et al. [2026] study hidden-state jumps and show that layer-wise displacement can reveal imbalanced computational contributions across depth. Together, these works show that transformer trajectories have rich and meaningful geometry. VANE builds on this view, but shifts the target from representation analysis and token-level uncertainty to per-example reasoning correctness.

Detecting failures from internal states, and the gap for reasoning. Methods that detect failures from model internals can be organized by *where* the signal is read. A first group scores the output distribution or sampled completions: token log-probability and predictive entropy, Self-Consistency [Wang et al., 2023], Semantic Entropy [Kuhn et al., 2023, Farquhar et al., 2024], P(True) [Kadavath et al., 2022], and INSIDE/EigenScore [Chen et al., 2024], with multi-pass variants paying a $K\times$ inference cost; Semantic Entropy Probes [Kossen et al., 2024] partially close this gap by distilling the multi-pass signal into a single-pass probe over hidden states. A second group trains probes on a single layer's hidden state or attention pattern, including SAPLMA [Azaria and Mitchell, 2023], CCS [Burns et al., 2023], the Lookback Lens [Chuang et al., 2024], and LLM-Check [Sriramanan et al., 2024]. Two observations motivate VANE's position in this landscape. First, almost all of this work targets open-ended factual hallucination with entailment-judged labels, and is evaluated on TriviaQA-, NQ-, or HaluEval-style benchmarks; systematic study of *reasoning failures* with verifiable correctness on GSM8K, MATH, or code benchmarks has received far less attention, with Song et al. [2026] providing a taxonomy but no operational detector. Second, activation-based detectors so far read the signal from the *values* of activations at one or more individual layers; VANE tests the complementary hypothesis that the signal lives in the *shape* of the layer-wise trajectory rather than learned high-dimensional combinations.

Table 1: AUROC (%) for failure prediction across models and benchmarks. Cells with shaded background indicate the best result per row (bold), and underlined values indicate the second-best.

Model	Benchmark	Log-Prob	Static Rep	P(True)	Logit-Lens	HS-Probe oracle	INSIDE	VANE
Qwen2.5-1.5B	GSM8K	62.0 ± 3.1	53.3 ± 4.1	51.2 ± 5.4	62.6 ± 3.0	63.5 ± 3.4	63.2 ± 7.1	$\mathbf{75.7 \pm 2.8}$
	MATH-500	56.7 ± 6.5	72.4 ± 6.3	48.2 ± 4.2	57.1 ± 4.3	75.8 ± 5.4	63.9 ± 5.1	$\mathbf{77.5 \pm 6.0}$
	HumanEval	52.6 ± 9.1	85.4 ± 7.0	53.8 ± 3.8	51.7 ± 6.8	$\mathbf{89.4 \pm 6.3}$	64.9 ± 4.9	88.4 ± 4.0
	CRUXEval	55.2 ± 5.2	67.4 ± 2.9	49.1 ± 4.1	50.3 ± 1.8	$\mathbf{73.4 \pm 2.3}$	66.6 ± 5.3	70.7 ± 2.4
	BBH	54.7 ± 4.2	66.2 ± 6.0	51.0 ± 3.9	51.6 ± 2.9	71.5 ± 4.6	46.6 ± 4.8	$\mathbf{73.8 \pm 2.1}$
Llama-3-8B	GSM8K	64.3 ± 3.3	59.0 ± 2.1	51.1 ± 2.2	61.5 ± 4.3	61.1 ± 2.2	65.9 ± 3.5	$\mathbf{78.9 \pm 1.8}$
	MATH-500	55.1 ± 3.0	78.9 ± 4.4	54.5 ± 3.5	51.9 ± 1.5	$\mathbf{84.9 \pm 4.4}$	75.2 ± 4.2	80.9 ± 2.6
	HumanEval	51.8 ± 5.6	69.7 ± 7.6	48.3 ± 3.2	46.6 ± 4.3	78.8 ± 6.1	62.5 ± 4.5	$\mathbf{80.5 \pm 6.8}$
	CRUXEval	59.7 ± 4.9	74.5 ± 2.2	50.9 ± 3.8	51.1 ± 2.1	75.4 ± 2.1	62.8 ± 4.8	$\mathbf{75.8 \pm 5.3}$
	BBH	55.8 ± 4.2	70.2 ± 2.5	46.8 ± 3.4	60.2 ± 2.6	$\mathbf{75.3 \pm 3.1}$	50.3 ± 4.1	71.0 ± 1.7
Ministral-8B	GSM8K	58.5 ± 4.7	60.1 ± 4.4	50.7 ± 3.2	60.5 ± 1.8	60.1 ± 2.9	60.7 ± 3.6	$\mathbf{72.5 \pm 3.0}$
	MATH-500	57.2 ± 5.1	79.0 ± 2.6	60.8 ± 4.5	57.4 ± 3.5	83.2 ± 2.6	58.0 ± 4.3	$\mathbf{83.6 \pm 2.1}$
	HumanEval	42.3 ± 9.4	63.9 ± 11.5	48.3 ± 3.6	42.6 ± 7.3	67.7 ± 9.2	52.5 ± 4.7	$\mathbf{76.2 \pm 7.2}$
	CRUXEval	61.4 ± 5.8	75.8 ± 2.0	50.3 ± 4.0	56.4 ± 3.2	$\mathbf{78.1 \pm 2.7}$	47.9 ± 4.2	75.4 ± 3.1
	BBH	67.0 ± 8.4	86.7 ± 2.1	51.8 ± 3.7	65.4 ± 3.3	$\mathbf{92.2 \pm 1.4}$	50.6 ± 4.4	90.3 ± 1.3
Gemma-3-12B	GSM8K	54.1 ± 2.3	55.9 ± 2.8	53.7 ± 3.0	56.1 ± 3.8	73.8 ± 3.0	50.0 ± 2.5	$\mathbf{79.3 \pm 3.3}$
	MATH-500	73.0 ± 6.9	81.4 ± 4.7	51.1 ± 4.3	66.3 ± 4.0	83.6 ± 3.5	68.0 ± 5.2	$\mathbf{92.7 \pm 3.5}$
	HumanEval	42.3 ± 12.3	74.0 ± 7.4	54.6 ± 3.9	49.7 ± 0.7	75.1 ± 10.2	46.3 ± 4.6	$\mathbf{90.2 \pm 9.9}$
	CRUXEval	59.2 ± 2.0	79.0 ± 2.3	58.1 ± 4.1	48.7 ± 0.9	$\mathbf{80.2 \pm 1.6}$	63.6 ± 5.0	73.6 ± 4.3
	BBH	55.1 ± 4.6	82.5 ± 2.5	41.1 ± 3.5	51.3 ± 3.3	82.6 ± 2.5	63.0 ± 4.8	$\mathbf{91.1 \pm 3.1}$
Qwen3-14B	GSM8K	74.5 ± 2.4	69.2 ± 2.8	52.0 ± 5.1	69.2 ± 1.0	82.4 ± 0.7	62.1 ± 5.8	$\mathbf{91.0 \pm 2.2}$
	MATH-500	47.2 ± 6.7	78.6 ± 3.9	51.0 ± 4.4	55.3 ± 3.8	84.7 ± 2.4	58.0 ± 4.9	$\mathbf{85.1 \pm 2.5}$
	HumanEval	61.0 ± 15.3	51.1 ± 12.4	50.7 ± 3.7	43.7 ± 5.9	55.1 ± 11.6	58.4 ± 5.1	$\mathbf{83.9 \pm 9.0}$
	CRUXEval	59.1 ± 4.5	77.0 ± 3.9	60.5 ± 4.2	57.1 ± 3.0	80.1 ± 2.8	58.6 ± 4.7	$\mathbf{96.6 \pm 2.0}$
	BBH	44.5 ± 10.2	88.3 ± 3.3	68.4 ± 4.6	55.8 ± 4.3	91.1 ± 1.3	49.5 ± 4.3	$\mathbf{95.0 \pm 1.1}$
Qwen2.5-72B	GSM8K	56.2 ± 6.8	74.8 ± 4.5	59.3 ± 5.4	53.5 ± 5.8	70.8 ± 3.8	56.1 ± 7.1	$\mathbf{85.5 \pm 3.1}$
	MATH-500	54.8 ± 4.2	72.7 ± 5.5	60.3 ± 4.8	53.4 ± 4.2	73.5 ± 4.4	69.0 ± 5.3	$\mathbf{88.6 \pm 2.1}$
	HumanEval	63.6 ± 5.8	76.9 ± 6.2	50.5 ± 3.9	50.7 ± 9.3	71.8 ± 5.5	63.8 ± 4.5	$\mathbf{84.9 \pm 3.8}$
	CRUXEval	63.5 ± 3.9	73.8 ± 3.1	59.1 ± 4.3	55.3 ± 1.9	77.7 ± 6.0	66.3 ± 5.1	$\mathbf{84.4 \pm 1.8}$
	BBH	64.9 ± 4.5	83.6 ± 2.8	52.4 ± 4.0	51.7 ± 2.7	84.3 ± 3.6	52.6 ± 4.6	$\mathbf{92.2 \pm 2.1}$

3 Methodology

3.1 Dynamical Motivation for Trajectory Geometry

During the forward pass of LLMs, information is propagated and progressively updated across transformer hidden layers through residual transformations of the hidden states, where each layer applies $h_{\ell+1} = h_\ell + f_\ell(h_\ell)$. Under this view, the sequence of hidden states $\{h_0, h_1, \ldots, h_L\}$ forms a trajectory through representation space, making geometric quantities such as velocity, curvature, and jerk natural descriptors of how computation unfolds across depth. When reasoning proceeds coherently, successive hidden-layer updates should remain locally aligned, yielding a smooth trajectory with gradually changing direction. By contrast, when attention and MLP sublayers drive the residual stream in inconsistent directions, the hidden-state trajectory may exhibit abrupt turns or oscillatory behavior, signaling instability in the model's internal computation. Such geometric instability can therefore reveal latent anomalies and potential reasoning failures. As shown in Figure 1, we present several representative cases where failed reasoning is accompanied by less smooth and less coherent hidden-state trajectories.

3.2 VANE: Five Geometric Metrics

We operationalize this intuition through VANE, named for the weather vane whose deflection reveals the flow it sits in: five parameter-free finite-difference geometric estimates computed over the layer-wise trajectory of a single forward pass. These discrete descriptors are analogues of classical

geometric quantities, rather than exact continuous-curve measurements. Let $h_\ell^{(t)} \in \mathbb{R}^d$ denote the hidden state at layer ℓ for generated token t, and define the layer-wise displacement $\Delta_\ell^{(t)} = h_{\ell+1}^{(t)} - h_\ell^{(t)}$ and unit tangent $T_\ell^{(t)} = \Delta_\ell^{(t)} / \|\Delta_\ell^{(t)}\|$. We compute five scalars per layer per token.

- **Velocity.** $V_\ell = \|\Delta_\ell^{(t)}\|$, the step-size magnitude (speed at which the hidden state moves per layer).
- **Representational Jerk.**[2] $J_\ell = \|\Delta_{\ell+1} - \Delta_\ell\|$, the second-order difference of displacements across layers, capturing abrupt shifts in both direction and magnitude.
- **Frenet–Serret Curvature (Local Nonlinearity).** $\kappa_\ell = \|T_{\ell+1}^{(t)} - T_\ell^{(t)}\|$, the rate of change of the unit tangent, an arc-length-normalized measure of local directional nonlinearity [do Carmo, 1976].
- **Geodesic Deviation (Global Nonlinearity).**

$$G_\ell = \frac{\left\| h_\ell - \left(h_0 + \frac{\ell}{L-1}(h_L - h_0) \right) \right\|}{\|h_L - h_0\|} \tag{1}$$

  measures the distance from the straight chord between first and last layer: a global (chord-based) measure of how far the trajectory bends away from linearity, complementary to the local curvature κ_ℓ. Values near 0 indicate a near-straight, directed path. For degenerate trajectories with $h_L = h_0$, we define $G_\ell := 0$.
- **Token Coherence (Cross-Token Efficiency).** $C_\ell = 1 - \cos(T_\ell^{(t)}, \ \bar{T}_\ell)$, where $\bar{T}_\ell = \frac{1}{|\mathcal{T}|} \sum_t T_\ell^{(t)}$, measuring per-layer directional incoherence across generated tokens. Low values indicate that successive tokens follow a shared, efficient representational direction.

3.3 Feature Aggregation and Classifier

Let $\mathcal{M} = \{V, J, \kappa, G, C\}$ denote the five metrics defined in Section 3.2, and let $\mathcal{T}$ be the set of token positions in a generated response. Each metric $m \in \mathcal{M}$ yields a per-layer, per-token scalar $m_\ell^{(t)}$ for $\ell \in \{0, 1, \ldots, L-1\}$ and $t \in \mathcal{T}$.

Token aggregation. To summarize token-level variation, we aggregate each metric across $\mathcal{T}$ under three operators: the pointwise *max* and *mean* over all $t \in \mathcal{T}$, and *ans*, the maximum over a window $\mathcal{T}_{\text{ans}} \subseteq \mathcal{T}$ of the final w tokens of the generated answer (we set $w = 20$ in all experiments). The *max* and *ans* aggregations capture worst-case behavior, while *mean* captures average smoothness.

Feature vector. For each metric and each aggregation, we collect the resulting per-layer profile and concatenate across layers, metrics, and aggregations to form a single feature vector $x \in \mathbb{R}^D$. The dimension D scales linearly with depth, is independent of the hidden-state dimension d, and does not depend on tokenizer or unembedding choice.

Classifier. Given a labeled dataset $\{(x_i, y_i)\}_{i=1}^N$ where $y_i \in \{0, 1\}$ indicates whether the generated response is correct, we train a Gradient Boosting classifier [Pedregosa et al., 2011] to estimate $\Pr(y = 1 \mid x)$. Hyperparameters are tuned with Optuna [Akiba et al., 2019] using 5-fold stratified cross-validation. Feature extraction requires a single forward pass and Algorithm 1 in Appendix A.2 summarizes the procedure.

Key properties. First, the feature map is parameter-free and deterministic: the only learned parameters in VANE are those of the downstream classifier. Second, it is single-pass: all metrics are computed from hidden states produced by one forward pass. Third, it depends only on layer-wise differences of hidden states, and therefore transfers across model families without retuning the metric definitions. All five metrics are closed-form functions of the hidden states from a single forward pass, involve no learned parameters, and capture distinct geometric aspects of the trajectory. No architecture-specific tuning is required.

[2] We use "jerk" to emphasize abrupt directional changes, though this is technically a second-order difference (analogous to acceleration in physics).

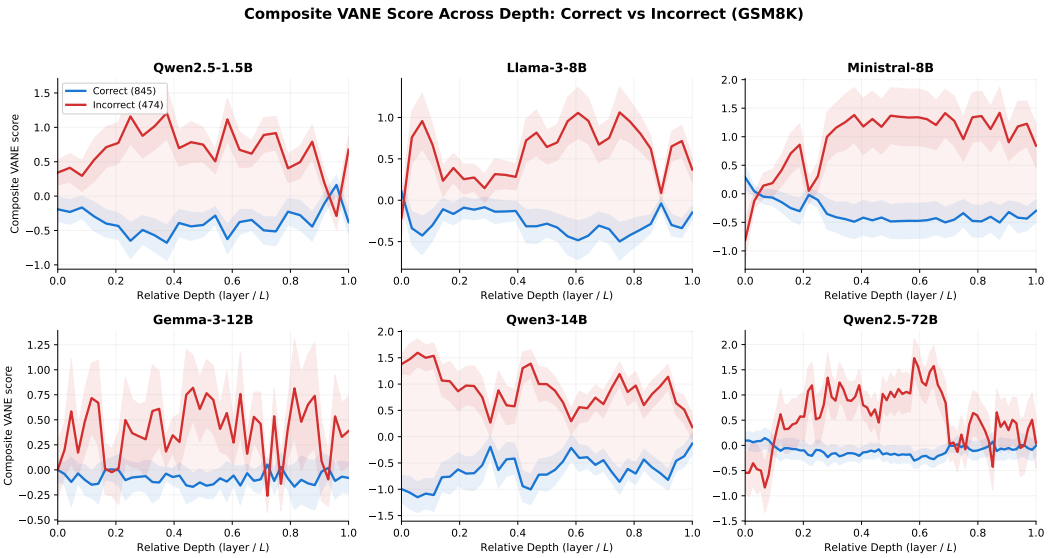


Figure 2: Composite VANE score across transformer depth (GSM8K). Each panel shows the mean composite score for correct (blue) and incorrect (red) samples as a function of relative depth (layer/L). The composite is the sum of five per-layer z-scored VANE metrics, each oriented so that higher values indicate greater geometric instability. Shaded regions denote ± 2 standard errors. Incorrect samples consistently score higher across all six models, confirming that geometric turbulence is distributed across the full trajectory rather than confined to a single layer. The composite is unbounded since it sums five z-scores.

4 Experiments

We evaluate VANE along three research questions that progress from descriptive to mechanistic. **RQ1** (section 4.2) asks whether trajectory geometry is a strong, general, and robust predictor of reasoning correctness across models, tasks, and deployment conditions. **RQ2** (section 4.3) asks where in the model's computation this signal resides: which axis, which layers, and which geometric quantities carry it. **RQ3** (section 4.4) asks where the signal comes from: whether it is attributable to architecture, scale, or training.

4.1 Experimental Setups

Models and Benchmarks. We evaluate VANE across five benchmarks spanning three categories of reasoning tasks: mathematical reasoning (GSM8K [Cobbe et al., 2021], MATH-500 [Hendrycks et al., 2021]), code reasoning (HumanEval [Chen et al., 2021], CRUXEval [Gu et al., 2024]), and natural language reasoning (BBH [Suzgun et al., 2023]). The evaluation is conducted on six model families with sizes ranging from 1.5B to 72B parameters: (i) *Qwen2.5-1.5B-Instruct* [Yang et al., 2024], (ii) *Qwen3-14B* [Qwen Team, 2025], (iii) *Llama-3-8B-Instruct* [Meta AI, 2024], (iv) *Ministral-8B-Instruct* [Mistral AI, 2024], (v) *Gemma-3-12B-IT* [Gemma Team, 2025], and (vi) *Qwen2.5-72B-Instruct* [Yang et al., 2024]. All generations use greedy decoding with default temperature and a maximum output length of 1024 tokens. AUROC is computed via 5-fold stratified cross-validation.

Baselines. We compare VANE against single-pass and multi-pass baselines, where single-pass methods compute their score from a single greedy generation, while multi-pass methods require multiple sampled generations (each involving a full autoregressive decoding pass) per query. *Log-Probability*: mean token log-probability of the generated sequence. *Static Representation*: linear probe on the final-layer hidden state. *P(True)* [Kadavath et al., 2022]: self-assessment via a follow-up prompt (Appendix G). *Logit-Lens* [nostalgebraist, 2020]: projects each layer's hidden state through the unembedding matrix to obtain the probability of the final token; layer selected via cross-validation. *HS-Probe (oracle)*: logistic regression on hidden states with the best layer chosen using test labels—a non-deployable upper bound that reveals how much signal single-layer representations contain. *INSIDE (EigenScore)* [Chen et al., 2024]: measures output uncertainty via the eigenstructure

of embedding covariance across $K=10$ temperature-sampled completions ($T=0.5$), serving as a representative uncertainty-based baseline. All methods above except INSIDE operate on a single greedy generation

4.2 RQ1: Does trajectory geometry predict reasoning correctness?

Table 1 reports AUROC across six models and five benchmarks. VANE ranks first on 23 of these 30 combinations, second on 5, and third on 2. It also beats INSIDE on every model; INSIDE samples ten completions per example, yet still underperforms our single-pass approach. On Qwen3-14B, VANE reaches 91.0 and logistic regression alone gets 91.4, confirming the signal is nearly linearly separable. Qwen2.5-72B under 4-bit quantization still achieves 85.5.

Output-space EigenScore (INSIDE) lands in a narrow 61 to 66 AUROC band on four models, weaker than expected given strong open-domain QA results in the original paper [Chen et al., 2024]. A likely factor is task shift: INSIDE was developed for factual hallucination under diverse phrasings, but GSM8K chains are highly templated, so embedding covariance carries less signal. Two outcomes were more surprising. **Gemma-3-12B** collapses to chance (50.0) because all $K=10$ sampled completions are bitwise identical despite $T=0.5$, making EigenScore constant. **Qwen2.5-72B** shows weak discrimination (56.1) with almost no cross-sample diversity. These degeneracies highlight that multi-pass diversity detectors can fail silently when a model's sampler returns near-duplicates.

Partial cross-task transfer. We train the VANE classifier on GSM8K and evaluate directly on MATH-500, HumanEval, and CRUXEval without retraining. As baselines, we include log-probability (LP) and Logit-Lens (LL), where LL selects its best layer via cross-validation on the target benchmark, giving it an advantage VANE does not have. Results appear in table 2. VANE exceeds LP on 15 of 18 settings and surpasses LL on all 18. Transfer is not uniformly dominant, and the largest gain is on Qwen3-14B, which reaches 91.6% AUROC on CRUXEval after training only on GSM8K, beating LP by 32.5 points. In an additional frozen math-to-BBH evaluation, VANE also exceeds LP on Qwen3-14B (69.0 vs. 44.5) and Llama-3-8B (60.8 vs. 55.8), but not on Ministral-8B (62.5 vs. 67.0). These results support partial cross-task transfer rather than universal generalization.

Table 2: Cross-task transfer (AUROC, %). LP = Log-Prob; LL = Logit-Lens.

	MATH-500			HumanEval			CRUXEval		
Model	VANE	LP	LL	VANE	LP	LL	VANE	LP	LL
Qwen2.5-1.5B	**72.2**	56.7	57.1	**64.8**	52.6	51.7	**61.8**	55.2	50.3
Llama-3-8B	**66.3**	55.1	51.9	**59.3**	51.8	46.6	**66.3**	59.7	51.1
Ministral-8B	**70.4**	57.2	57.4	**50.2**	42.3	42.6	60.3	**61.4**	56.4
Gemma-3-12B	**78.7**	73.0	66.3	**70.4**	42.3	49.7	56.3	**59.2**	48.7
Qwen3-14B	**63.3**	47.2	55.3	**64.8**	61.0	43.7	**91.6**	59.1	57.1
Qwen2.5-72B	**77.8**	54.8	53.4	57.8	**63.6**	50.7	**63.7**	63.5	55.3

Quantization robustness. Table 3 evaluates VANE under 4-bit NF4 quantization on MATH-500. On Gemma-3-12B, 4-bit reduces VANE by 5.6 points (92.7% to 87.1%), but VANE still leads baselines by over 10 points. Qwen3-14B shows VANE slightly improves under quantization (85.1% to 85.6%). At 72B scale, 4-bit Qwen2.5-72B reaches 88.6% AUROC, outperforming all baselines by over 15 points. Trajectory geometry is robust under quantization: log-probability depends on final softmax sharpness which quantization perturbs, whereas VANE aggregates differences across layers, allowing noise to partially cancel.

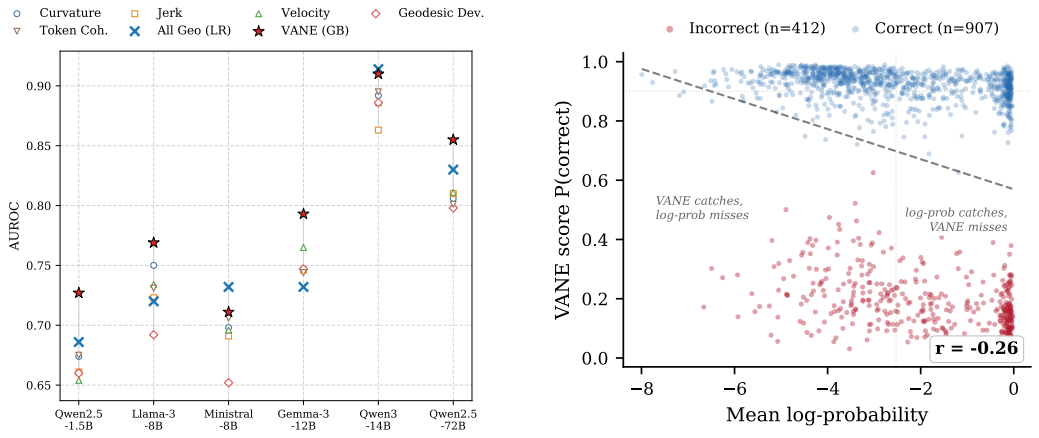
Table 3: Quantization robustness on MATH-500 (AUROC, %). LP = Log-Prob; SR = Static Rep; LL = Logit-Lens.

Model	Prec	VANE	LP	SR	LL
Gemma-3-12B	BF16	**92.7**	73.0	81.4	66.3
Gemma-3-12B	4-bit	**87.1**	68.8	76.7	56.5
Qwen3-14B	BF16	**85.1**	47.2	78.6	55.3
Qwen3-14B	4-bit	**85.6**	46.7	79.4	53.6
Qwen2.5-72B	4-bit	**88.6**	54.8	72.7	53.4

4.3 RQ2: What are the characteristics of the trajectory geometric features?

We further characterize the structural form of the signal and summarize our main observations in four findings below.

VANE (GB) surpasses individual metrics and their linear combination. We investigate each metric's contribution to understand whether the signal concentrates in a single geometric quantity or emerges from their interaction. Fig. 3a reports AUROC for each VANE metric individually, their



(a) (b)

Figure 3: (a) VANE (GB) surpasses individual metrics and their linear combination. (b) VANE score vs. mean log-probability on Llama-3-8B, showing that the geometric signal is more separatable.

linear combination (All Geo. LR), and the GradientBoosting classifier. Every metric individually beats every single-pass baseline (65.2% to 89.5%), but no metric wins consistently, with gaps below 2% across models. A linear combination recovers most of the signal, matching or exceeding Gradient Boosting on two models, while Gradient Boosting retains a 2.5% to 6.1% edge on the remaining four. These results indicate that no single geometric quantity carries the signal alone; instead, VANE benefits from non-linear interactions across complementary metrics that GradientBoosting captures.

Better separability than token confidence and the endpoint representation. Exploring the source of this advantage, we find that VANE provides better separability than output-distribution signals. For example, Fig. 3b shows that on Llama-3-8B (GSM8K), correct samples cluster at VANE scores above 80% while incorrect samples fall below 50%, whereas both classes span the full log-probability range with no separation at high confidence. The Pearson correlation between the two signals is $r = -0.26$, and concatenating log-probability as an additional VANE feature decreases AUROC on five of six models ($\Delta = -0.2\%$ to -0.7%). VANE is likewise weakly correlated with the static final-layer representation ($r = -0.18$; see Appendix G.1). These results indicate that cross-layer trajectory geometry provides stronger separability than either token confidence or the endpoint representation.

The signal mainly lives along the layer axis, not the sequence axis. VANE computes geometry along the layer axis (hidden states across depth), not the sequence axis (token positions at a fixed layer). The layer axis tracks the iterated map $h_{\ell+1} = h_\ell + f_\ell(h_\ell)$, under which curvature and jerk have a direct dynamical reading; cross-token differences at a fixed layer are dominated by token-identity content rather than computational variation. On Llama-3-8B GSM8K, layer-axis VANE reaches 76.7% AUROC while sequence-axis reaches only 68.6%; concatenating both yields just 77.1%, confirming sequence-axis features add little beyond the layer axis.

Geometry is a compact summary of all-layer information. To test whether VANE's advantage comes from its geometric descriptors or simply from reading every layer, we compare it with a matched, non-oracle all-layer probe: hidden states pooled or concatenated across all layers and classified by logistic regression, gradient boosting, or a small MLP under the same 5-fold protocol. In a separate matched run (Table 12, Appendix Q), VANE achieves a higher mean AUROC (79.0% vs. 75.0%) and wins 16 of 27 cells while using a few hundred features instead of tens of thousands; the probe's advantage is concentrated on Llama-3-8B and CRUXEval.

4.4 RQ3: Where does the signal come from?

Having located the signal in the layer-wise trajectory, we ask whether it reflects intrinsic properties like scale and architecture, or instruction tuning.

Instruction tuning amplifies geometric instability signals. To isolate the origin of the trajectory-based correctness signal, we evaluate four matched base/instruction-tuned model pairs on GSM8K (Table 4). Instruction tuning consistently strengthens the geometric signal: all four families exhibit positive gains (ΔVANE from $+3.5\%$ to $+15.1\%$). As a control, token-level log-probability (LP) shifts are inconsistent across families, and even degrade on Mistral-7B-v0.1 ($\Delta\text{LP} = -4.3\%$). This suggests instruction tuning reshapes internal dynamics, amplifying geometric signals of correctness.

Table 4: Base vs. instruction-tuned models on GSM8K (AUROC, %).

Family	Params	Base VANE	Instruct VANE	ΔVANE	Base LP	Instruct LP
OLMo-7B-0724	7B	63.4±2.6	**70.1±2.0**	+6.7	47.2±5.4	64.0±2.6
Mistral-7B-v0.1	7B	64.6±2.0	**73.2±1.4**	+8.6	62.8±2.6	58.5±3.1
Llama-3-8B	8B	61.8±4.5	**76.9±1.8**	+15.1	52.3±2.7	62.7±1.6
Qwen2.5-7B	7B	67.5±3.5	**71.0±5.1**	+3.5	53.7±4.5	54.2±7.8

Parameter scale does not drive the geometric signal. To rule out model size as a confounder, we compare a 1.5B instruction-tuned model against larger base models (Table 5). Despite fewer parameters, the 1.5B instruct checkpoint reaches 75.7 AUROC, surpassing 7B/8B base models (e.g., 64.6 for Mistral-7B, 61.8 for Llama-3-8B). Since these larger models fail to show stronger separation, scale cannot be the primary driver: combined with the base/instruct comparison above, this confirms the trajectory pattern is induced by instruction tuning rather than by model size.

Table 5: Impact of model scales (AUROC, %).

Model	Type	Params	VANE	Log-Prob
Qwen2.5-1.5B-Instruct	Instruct	1.5B	**75.7±2.8**	62.0±3.1
Qwen2.5-7B	Base	7B	**67.5±3.5**	53.7±4.5
OLMo-7B-0724	Base	7B	**63.4±2.6**	47.2±5.4
Mistral-7B-v0.1	Base	7B	**64.6±2.0**	62.8±2.6
Llama-3-8B	Base	8B	**61.8±4.5**	52.3±2.7
Gemma-2-9B	Base	9B	**62.8±4.8**	56.0±2.2

4.5 Matched controls and robustness

Confounder controls. A surface-feature bundle (length, word/line counts, code markers, digit ratio, and equations) averages 68.0 AUROC versus 78.2 for VANE, with VANE winning 23/26 cells; length alone averages 59.0, and removing length retains 89% of VANE's ranking signal. On 300 MATH-500 problems with official difficulty labels, within-level evaluation holds difficulty fixed: VANE remains predictive on four of five models (76.0–80.9 AUROC), while the smallest model is close to the difficulty-only baseline. Detailed results are in Appendix Q.

Prompting, decoding, and operating points. Under top-p=0.95 sampling, VANE remains predictive in four checks (60.4–90.6 AUROC), although performance drops most on the 1.5B model. In a sample-matched n=300 comparison, single-pass VANE exceeds Self-Consistency with its best K in all four cells; these subset-tuned values are not directly comparable with Table 1. Across 28 matched cells, mean failure-detection AUPRC is 63.1 (random baseline 33.0) and ECE is 8.5. At 50% coverage across 20 cells, VANE retains 78.7% accuracy versus 68.5% for log-probability. Appendix Q gives the protocols and values.

5 Downstream Applications

5.1 Selective prediction

A natural deployment scenario for VANE is selective prediction [Geifman and El-Yaniv, 2017]: the model answers only when its instability score falls below a threshold, and abstains otherwise. Figure 4 shows accuracy on the retained subset of GSM8K as a function of the abstention rate (the fraction of samples refused), for three representative mid-scale models (Llama-3-8B, Ministral-8B, Gemma-3-12B). Solid lines use VANE as the abstention signal; dashed lines use mean token log-probability. VANE-based selection rises monotonically on every model: accuracy climbs from 68.8% to 84.1% on Llama-3-8B, 73.8% to 86.0% on Ministral-8B, and 82.9% to 94.4% on Gemma-3-12B at 50% abstention. The gain over log-probability grows as coverage decreases (Figure 8).

Log-probability-based selection behaves very differently: on five of six models it *decreases* retained-set accuracy with higher abstention. On Llama-3-8B, filtering by log-probability drops accuracy to 60.4%, below the unconditional baseline; on Qwen3-14B, it collapses from 58.1% to 38.4%. This inversion reflects the near-independence finding: log-probability is dominated by formulaic filler tokens, while VANE reads the geometry of internal computation.

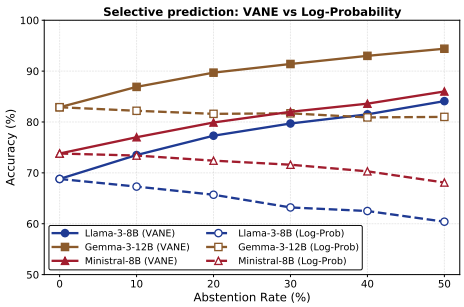


Figure 4: Selective prediction on GSM8K for three representative models; all six models are in Table 7.

5.2 Localizing failures at the token level.

Beyond aggregate trajectory statistics, geometric jerk can be localized to individual tokens, revealing *where* in the reasoning chain the model's computation destabilizes. Appendix J shows that high-jerk tokens concentrate at the exact arithmetic error site (Cohen's $d = 1.83$), suggesting VANE's signal can be traced to semantically meaningful tokens for targeted re-sampling or self-correction.

6 Conclusion

In this work, we introduce VANE, a framework that predicts LLM reasoning failures from finite-difference geometric estimates of hidden-state trajectories. Across 30 model–benchmark configurations (1.5B–72B parameters), VANE substantially outperforms output-confidence baselines. We do not claim causality; rather, VANE provides an observable proxy of internal computation whose predictive power is consistent with residual-update dynamics. Together, these results indicate that trajectory geometry exposes failure signals that the output distribution does not capture: correctness depends not only on where the model ends, but on how it computes.

Limitations. VANE requires white-box hidden-state access and supervised calibration labels, limiting use with closed APIs and label-scarce tasks. Transfer is partial and non-uniform, and performance can shift under prompting, sampling, quantization, and distribution change. Although controls reduce obvious explanations from length, formatting, and problem difficulty, residual confounding and untested error types remain possible; the descriptors are correlational finite-difference estimates, not causal mechanisms. Internal activations may also expose sensitive information and create privacy or security risks if retained. Finally, open-ended, long-form, multi-turn, and agentic settings remain outside our evaluation scope.

Acknowledgements and Disclosures

Acknowledgements and funding. This work was partially supported by the EU project TrustLLM (Grant Agreement 101135671), the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation, and the Excellence Center at Linköping-Lund in Information Technology (ELLIIT). Computational resources were provided on the Berzelius system funded by the Knut and Alice Wallenberg Foundation and operated by NAISS. The work was done at the AI4x Center of Excellence at Linköping University. **Competing interests.** The authors declare no competing interests.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631, 2019.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2017.

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, 2023.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *International Conference on Learning Representations*, 2023.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. INSIDE: LLMs' internal states retain the power of hallucination detection. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=Zj12nzlQbz`.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James Glass. Lookback Lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. *arXiv preprint arXiv:2407.07071*, 2024.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Guy Dar, Mor Geva, Ankit Gupta, and Jonathan Berant. Analyzing transformers in embedding space. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16124–16170, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.893. URL `https://aclanthology.org/2023.acl-long.893/`.

Manfredo P. do Carmo. *Differential Geometry of Curves and Surfaces*. Prentice-Hall, Englewood Cliffs, New Jersey, 1976.

Kawin Ethayarajh. 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 and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 55–65, 2019.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, pages 4878–4887, 2017.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, 2021.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.3. URL https://aclanthology.org/2022.emnlp-main.3/.

Alex Gu, Baptiste Roziere, Hugh Leather, Armando Solar-Lezama, Gabriel Synnaeve, and Sida I. Wang. CRUXEval: A benchmark for code reasoning, understanding and execution. *arXiv preprint arXiv:2401.03065*, 2024.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

Eghbal A. Hosseini and Evelina Fedorenko. Large language models implicitly learn to straighten neural sentence trajectories to construct a predictive representation of natural language. In *Advances in Neural Information Processing Systems*, 2023.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Jack King, Evelina Fedorenko, and Eghbal A. Hosseini. Representational curvature modulates behavioral uncertainty in large language models. In *International Conference on Machine Learning*, 2026.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in LLMs. *arXiv preprint arXiv:2406.15927*, 2024.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023.

Yiping Lu, Zhuohan Li, Di He, Zhiqing Sun, Bin Dong, Tao Qin, Liwei Wang, and Tie-Yan Liu. Understanding and improving transformer from a multi-particle dynamic system point of view. *arXiv preprint arXiv:1906.02762*, 2019.

Meta AI. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Mistral AI. Ministral 8B. https://mistral.ai/news/ministraux, 2024.

nostalgebraist. Interpreting GPT: The logit lens. LessWrong, 2020.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Keigo Shibata, Kazuki Yano, Ryosuke Takahashi, Jaesung Lee, Wataru Ikeda, and Jun Suzuki. Suppressing final layer hidden state jumps in transformer pretraining. *arXiv preprint arXiv:2601.18302*, 2026.

Peiyang Song, Pengrui Han, and Noah D. Goodman. Large language model reasoning failures. *arXiv preprint arXiv:2602.06176*, 2026.

Table 6: Notation used in this work.

Symbol	Meaning	Symbol	Meaning
L	number of transformer layers ($L \geq 2$)	$\mathcal{T}$	set of token positions in the generated answer
d	hidden-state dimension ($d \geq 1$)	$\mathcal{T}_{\text{ans}}$	window of the final w generated tokens ($w = 20$)
$\ell \in \{0, \ldots, L\}$	layer index	$h_\ell^{(t)}$	hidden state at layer ℓ, token t
$h_\ell \in \mathbb{R}^d$	hidden state at layer ℓ	$\Delta_\ell^{(t)}$	layer-wise displacement $h_{\ell+1}^{(t)} - h_\ell^{(t)}$
f_ℓ	residual update: $h_{\ell+1} = h_\ell + f_\ell(h_\ell)$	$T_\ell^{(t)}$	unit tangent $\Delta_\ell^{(t)} / \|\Delta_\ell^{(t)}\|$
$\mathcal{H} = (\mathbb{R}^d)^{L+1}$	space of trajectories $H = (h_0, \ldots, h_L)$	the five VANE metrics (Sec. 3.2)	$V_\ell, J_\ell, \kappa_\ell, G_\ell, C_\ell$
$\pi : \mathcal{H} \to \mathbb{R}^d$	endpoint projection $\pi(H) = h_L$	$x \in \mathbb{R}^D$	per-example feature vector
$t \in \mathcal{T}$	generated-token index	$g,\ y \in \{0, 1\}$	gradient-boosting classifier; correctness label

Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. LLM-Check: Investigating detection of hallucinations in large language models. In *Advances in Neural Information Processing Systems*, 2024.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging BIG-Bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, 2023.

Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio Ansuini, and Alberto Cazzaniga. The geometry of hidden representations of large transformer models. *Advances in Neural Information Processing Systems*, 36:51234–51252, 2023.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=gjeQKFxFpZ`.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Fan Yin, Jayanth Srinivasa, and Kai-Wei Chang. Characterizing truthfulness in large language model generations with local intrinsic dimension. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

A VANE Formal Setup and Algorithm

This appendix fixes notation used in Propositions and Lemma, and gives the end-to-end algorithm for computing the VANE feature vector. The closed-form definitions of the five metrics and the token-aggregation operators are given in Sections 3 and are not repeated here.

A.1 Notation

Throughout, $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^d$. Reading the residual update as an iterated map, the sequence H is a discrete trajectory in $\mathbb{R}^d$, and $\Delta_\ell = f_\ell(h_\ell)$. Every quantity the LM head reads—the next-token distribution, output log-probabilities, $P(\text{True})$—is a function of $\pi(H)$ alone; VANE, by contrast, reads the entire trajectory H.

A.2 Algorithm

Algorithm 1 summarises the full VANE pipeline. Given a question q and a target LLM M, the procedure performs a *single* forward pass to obtain the per-layer hidden states $\{h_\ell^{(t)}\}$, derives all five geometric metrics in closed form from finite differences of these states (lines 3–4), aggregates each metric over the generated tokens under three operators (lines 5–6), concatenates the resulting per-layer profiles into a fixed-length feature vector $x \in \mathbb{R}^D$, and finally applies the trained gradient-boosting classifier g to obtain the correctness score p_{corr}. Only g has learnable parameters; the entire feature map $H \mapsto x$ is parameter-free, deterministic, and independent of the tokenizer and unembedding matrix.

B Proposition 1: Structural Independence

Proposition 1 (Structural independence)**.** *The endpoint projection π is not injective: for every $h_L \in \mathbb{R}^d$ and every $L \geq 2$, the fiber $\pi^{-1}(h_L)$ is an (Ld)-dimensional affine subspace of $\mathcal{H}$. Moreover, there exist $H, H' \in \pi^{-1}(h_L)$ with all displacements nonzero such that $\kappa(H) \neq \kappa(H')$, $J(H) \neq J(H')$, and $G(H) \neq G(H')$ simultaneously. Consequently, for any measurable $g : \mathbb{R}^d \to \mathbb{R}$, $g(\pi(H)) = g(\pi(H'))$ while the trajectory features differ.*

Proof. **Step 1 (Fiber dimension).** Since $\mathcal{H} \cong \mathbb{R}^{(L+1)d}$, the map π is a linear projection onto the last d coordinates. It is surjective, so by rank–nullity $\ker \pi \cong \mathbb{R}^{Ld}$. The fiber $\pi^{-1}(h_L) = \ker \pi + (0, \ldots, 0, h_L)$ is an Ld-dimensional affine subspace (it does not contain the origin unless $h_L = 0$, so it is affine rather than linear in general).

Step 2 (Explicit witness construction). We exhibit two trajectories in the same fiber whose κ, J, and G values all differ. Fix $h_L = 0$ and let e_1 be the first standard basis vector (for $d = 1$, simply $e_1 = 1$).

Trajectory H (uniform). Set $h_\ell = (L-\ell)\, e_1$ for $\ell = 0, \ldots, L$. Then every displacement is $\Delta_\ell = -e_1$, all unit tangents equal $T_\ell = -e_1$, and we have:

$$J_\ell = 0 \qquad \text{for all } \ell = 0, \ldots, L-2 \quad \text{(constant displacements)}, \tag{2}$$

$$\kappa_\ell = 0 \qquad \text{for all } \ell = 0, \ldots, L-2 \quad \text{(constant unit tangents)}. \tag{3}$$

Since $h_0 = L\, e_1 \neq 0 = h_L$, the geodesic deviation is well-defined. The chord from h_0 to h_L has interpolation points $h_0 + \frac{\ell}{L-1}(h_L - h_0) = \big(L - \frac{\ell L}{L-1}\big)\, e_1$, while $h_\ell = (L-\ell)\, e_1$. The difference is $\ell\big(1 - \frac{L}{L-1}\big)e_1 = -\frac{\ell}{L-1}\, e_1$, so

$$G_\ell(H) \;=\; \frac{\ell/(L-1)}{L} \;=\; \frac{\ell}{L(L-1)} \qquad \text{for } \ell = 1, \ldots, L-1. \tag{4}$$

In particular $G_{L-1}(H) = 1/L$.

Trajectory H' (kinked). Identical to H except that h'_{L-1} is modified. We distinguish two cases by dimension.

CASE $d \geq 2$. Set $h'_{L-1} = e_1 + e_2$. Then:

$$\Delta'_{L-2} = h'_{L-1} - h_{L-2} = (e_1 + e_2) - 2\, e_1 = -e_1 + e_2,$$
$$\Delta'_{L-1} = h_L - h'_{L-1} = 0 - (e_1 + e_2) = -e_1 - e_2.$$

Both are nonzero. All other displacements ($\ell \leq L-3$ and $\ell = L$, if applicable) are unchanged from H.

Algorithm 1 VANE: geometric failure prediction (single forward pass).

Require: LLM M, question q, answer-window length w, trained classifier g
Ensure: Correctness score $p_{\text{corr}} \in [0, 1]$

```
Generate response â ~ M(q); extract h_ℓ^(t) for all ℓ ∈ {0, ..., L}, t ∈ T
Δ_ℓ^(t) ← h_{ℓ+1}^(t) − h_ℓ^(t),   T_ℓ^(t) ← Δ_ℓ^(t) / ‖Δ_ℓ^(t)‖
for m ∈ {V, J, κ, G, C} do
    Compute m_ℓ^(t) {closed-form, Sec. 3}
    for w' ∈ {max, mean, ans} do
        Aggregate over t to obtain per-layer profile m_ℓ^(w')
    end for
end for
Concatenate all profiles into x ∈ R^D
return p_corr ← g(x)
```

- *Jerk:* $J_{L-2}(H') = \|\Delta'_{L-1} - \Delta'_{L-2}\| = \|-2e_2\| = 2 \neq 0 = J_{L-2}(H)$.
- *Curvature:* The unit tangents are $T'_{L-2} = (-e_1 + e_2)/\sqrt{2}$ and $T'_{L-1} = (-e_1 - e_2)/\sqrt{2}$, so $\kappa_{L-2}(H') = \|T'_{L-1} - T'_{L-2}\| = \|-2e_2/\sqrt{2}\| = \sqrt{2} \neq 0 = \kappa_{L-2}(H)$.
- *Geodesic deviation:* The chord point at $\ell = L - 1$ is $h_0 + \frac{L-1}{L-1}(h_L - h_0) = h_L = 0$. So $G_{L-1}(H') = \|h'_{L-1} - 0\|/\|h_L - h_0\| = \sqrt{2}/L \neq 1/L = G_{L-1}(H)$.

CASE $d = 1$. Set $h'_{L-1} = -2$ (replacing $h_{L-1} = 1$ from the uniform trajectory with $L = 2$; for general L the construction is analogous). Then for $L = 2$: $h_0 = 2$, $h'_1 = -2$, $h_2 = 0$.

$$\Delta'_0 = h'_1 - h_0 = -4, \qquad \Delta'_1 = h_2 - h'_1 = 2.$$

Both are nonzero.

- *Jerk:* $J_0(H') = |\Delta'_1 - \Delta'_0| = |2 - (-4)| = 6 \neq 0 = J_0(H)$.
- *Curvature:* $T'_0 = \text{sign}(-4) = -1$ and $T'_1 = \text{sign}(2) = +1$, so $\kappa_0(H') = |+1 - (-1)| = 2 \neq 0 = \kappa_0(H)$.
- *Geodesic deviation:* $G_1(H') = |-2 - 0|/|0 - 2| = 1 \neq 1/2 = G_1(H)$.

Step 3 (Conclusion). In both cases we have $\pi(H) = \pi(H') = h_L$ but $\kappa(H) \neq \kappa(H')$, $J(H) \neq J(H')$, and $G(H) \neq G(H')$. Since any $g \circ \pi$ is constant on $\pi^{-1}(h_L)$, we have $g(\pi(H)) = g(\pi(H'))$ while the trajectory features differ. □

Remark (Scalar summaries and aggregation). Proposition 1 is stated for the sequence-valued features $\kappa, J, G \in \mathbb{R}^{L-1}$. Non-constancy of scalar summaries (e.g., mean or max over layers) follows from the same construction: the uniform trajectory H has $J_\ell = \kappa_\ell = 0$ for all ℓ, while H' has positive values at layer $L-2$, so both the mean and max of J and κ differ between H and H'. In the experiments, VANE further aggregates each metric over tokens (max, mean, or answer-window max) before concatenating the per-layer profiles; the propositions address the per-layer features directly.

Remark (Nondegeneracy of curvature). The curvature κ_ℓ requires $\Delta_\ell \neq 0$ and $\Delta_{\ell+1} \neq 0$ for the unit tangents to be well-defined. The set of trajectories in $\pi^{-1}(h_L)$ with any zero displacement is a proper algebraic subvariety of the Ld-dimensional fiber (codimension $\geq d \geq 1$), hence has Lebesgue measure zero. The construction above produces trajectories with all displacements nonzero by design.

Remark (On the G_ℓ interpolation). The chord from h_0 to h_L is parameterised as $h_0 + \frac{\ell}{L-1}(h_L - h_0)$, so the interpolation spans L points (at $\ell = 0, \ldots, L-1$) over the $L + 1$ trajectory states. At $\ell = L-1$ the chord point equals h_L, giving $G_{L-1} = \|h_{L-1} - h_L\|/\|h_L - h_0\|$. A trajectory with $G_\ell = 0$ for all $\ell = 1, \ldots, L-1$ would require $h_{L-1} = h_L$, creating $\Delta_{L-1} = 0$ and making κ_{L-2} undefined. The construction avoids this by comparing two trajectories with different but nonzero G values.

C Proposition 2: Trajectory-Gain Bound

Let $(\Omega, \mathcal{F}, P)$ be a probability space carrying random elements $H : \Omega \to \mathcal{H}$ and $Y : \Omega \to \{0, 1\}$, where Y denotes correctness on the task. Let $\phi : \mathcal{H} \to \mathbb{R}^k$ be any measurable feature map (the VANE

features, or any other). Since Y is bounded, $Y \in L^2(P)$. Consider the L^2 subspaces

$$\mathcal{G}_{\text{end}} := L^2\big(\sigma(h_L)\big), \qquad \mathcal{G}_{\text{traj}} := L^2\big(\sigma(h_L, \phi(H))\big),$$

where $\sigma(\cdot)$ denotes the σ-algebra generated by the given random variables. Because $\sigma(h_L) \subseteq \sigma(h_L, \phi(H))$, we have $\mathcal{G}_{\text{end}} \subseteq \mathcal{G}_{\text{traj}}$. The Bayes mean-squared risk of a subspace $\mathcal{G}$ is $R^*(\mathcal{G}) := \inf_{g \in \mathcal{G}} \mathbb{E}[(Y-g)^2]$, attained (by the L^2-projection theorem) at the conditional expectation of Y onto the corresponding σ-algebra.

Proposition 2 (Trajectory-gain bound)**.** *Under the setup above,*

$$R^*(\mathcal{G}_{end}) - R^*(\mathcal{G}_{traj}) \;=\; \mathbb{E}\Big[\mathrm{Var}\big(\,\mathbb{E}[Y \mid h_L, \phi(H)]\,\big|\,h_L\,\big)\Big] \;\geq\; 0, \tag{5}$$

with equality iff $\mathbb{E}[Y \mid h_L, \phi(H)] = \mathbb{E}[Y \mid h_L]$ P-almost surely. Strict inequality holds iff $\mathbb{E}[Y \mid h_L, \phi(H)]$ differs from $\mathbb{E}[Y \mid h_L]$ on a set of positive P-measure, i.e., $\phi(H)$ carries additional conditional-mean information about Y beyond h_L.

Proof. Define $\mathcal{F}_0 = \sigma(h_L)$ and $\mathcal{F}_1 = \sigma(h_L, \phi(H))$, so $\mathcal{F}_0 \subseteq \mathcal{F}_1$.

Step 1 (Bayes predictors). By the L^2-projection theorem, the infimum in each Bayes risk is attained at the conditional expectation:

$$g_0^* = \mathbb{E}[Y \mid \mathcal{F}_0], \qquad g_1^* = \mathbb{E}[Y \mid \mathcal{F}_1].$$

Step 2 (Risk as conditional variance). Therefore

$$R^*(\mathcal{G}_{\text{end}}) = \mathbb{E}\big[(Y - g_0^*)^2\big] = \mathbb{E}\big[\mathrm{Var}(Y \mid \mathcal{F}_0)\big], \qquad R^*(\mathcal{G}_{\text{traj}}) = \mathbb{E}\big[\mathrm{Var}(Y \mid \mathcal{F}_1)\big].$$

Step 3 (Conditional law of total variance). For the nested pair $\mathcal{F}_0 \subseteq \mathcal{F}_1$:

$$\mathrm{Var}(Y \mid \mathcal{F}_0) \;=\; \mathbb{E}\big[\mathrm{Var}(Y \mid \mathcal{F}_1)\,\big|\,\mathcal{F}_0\big] \;+\; \mathrm{Var}\big(\mathbb{E}[Y \mid \mathcal{F}_1]\,\big|\,\mathcal{F}_0\big) \quad P\text{-a.s.} \tag{6}$$

Step 4 (Take expectations and rearrange). Taking $\mathbb{E}[\cdot]$ of both sides of (6):

$$\mathbb{E}\big[\mathrm{Var}(Y \mid \mathcal{F}_0)\big] \;=\; \mathbb{E}\big[\mathrm{Var}(Y \mid \mathcal{F}_1)\big] \;+\; \mathbb{E}\big[\mathrm{Var}\big(\mathbb{E}[Y \mid \mathcal{F}_1]\,\big|\,\mathcal{F}_0\big)\big].$$

Substituting the expressions from Step 2:

$$R^*(\mathcal{G}_{\text{end}}) - R^*(\mathcal{G}_{\text{traj}}) \;=\; \mathbb{E}\big[\mathrm{Var}\big(\mathbb{E}[Y \mid \mathcal{F}_1]\,\big|\,\mathcal{F}_0\big)\big] \;\geq\; 0,$$

which is (5).

Step 5 (Equality condition). The right-hand side is ≥ 0 (conditional variance is non-negative). It equals zero P-a.s. iff $\mathrm{Var}(\mathbb{E}[Y \mid \mathcal{F}_1] \mid \mathcal{F}_0) = 0$ P-a.s., which holds iff $\mathbb{E}[Y \mid \mathcal{F}_1]$ is $\mathcal{F}_0$-measurable, i.e., $\mathbb{E}[Y \mid h_L, \phi(H)] = \mathbb{E}[Y \mid h_L]$ P-a.s.

The proof extends verbatim to any $Y \in L^2(P)$; binary Y is assumed here only for interpretability. □

Remark (Mean independence vs. conditional independence). The equality condition $\mathbb{E}[Y \mid h_L, \phi(H)] = \mathbb{E}[Y \mid h_L]$ is *mean independence* of Y from $\phi(H)$ given h_L: $\phi(H)$ does not improve the conditional mean predictor. This is weaker in general than full conditional independence $Y \perp\!\!\!\perp \phi(H) \mid h_L$. However, for binary $Y \in \{0,1\}$, $Y \mid h_L$ is Bernoulli with parameter $\mathbb{E}[Y \mid h_L]$, so its distribution is determined by the conditional mean. In this case, mean independence *is* equivalent to full conditional independence: the distinction is moot for binary classification.

Remark (Scope of the bound). The trajectory-gain bound (5) characterises the Bayes-optimal MSE gap, not AUROC or any particular learned classifier. Empirical AUROC improvements are *consistent with* $R^*(\mathcal{G}_{\text{end}}) - R^*(\mathcal{G}_{\text{traj}}) > 0$ but do not directly estimate it, since AUROC is a rank metric and the learned Gradient Boosting classifier is not Bayes-optimal. In particular, the bound does not rule out the existence of a better endpoint-only classifier; it establishes that existing endpoint methods leave exploitable signal on the table.

D Lemma: Local Jerk Decomposition

Lemma 1 (Local jerk decomposition under residual dynamics)**.** *Under the residual update $h_{\ell+1} = h_\ell + f_\ell(h_\ell)$, suppose f_ℓ is Λ_ℓ-Lipschitz on the segment $[h_\ell, h_{\ell+1}]$, i.e.,*

$$\|f_\ell(x) - f_\ell(y)\| \;\le\; \Lambda_\ell \,\|x - y\| \qquad \textit{for all } x, y \in \{h_\ell + t\,\Delta_\ell : t \in [0,1]\}.$$

Then

$$J_\ell \;\le\; \underbrace{\big\|f_{\ell+1}(h_{\ell+1}) - f_\ell(h_{\ell+1})\big\|}_{\textit{inter-layer disagreement}} \;+\; \underbrace{\Lambda_\ell \cdot V_\ell}_{\textit{local drift}}, \tag{7}$$

where $V_\ell = \|\Delta_\ell\| = \|f_\ell(h_\ell)\|$.

Proof. Since $\Delta_\ell = f_\ell(h_\ell)$, we have

$$J_\ell = \|\Delta_{\ell+1} - \Delta_\ell\| = \|f_{\ell+1}(h_{\ell+1}) - f_\ell(h_\ell)\|.$$

Adding and subtracting $f_\ell(h_{\ell+1})$:

$$\begin{aligned}
J_\ell &= \big\|f_{\ell+1}(h_{\ell+1}) - f_\ell(h_{\ell+1}) + f_\ell(h_{\ell+1}) - f_\ell(h_\ell)\big\| \\
&\le \big\|f_{\ell+1}(h_{\ell+1}) - f_\ell(h_{\ell+1})\big\| + \big\|f_\ell(h_{\ell+1}) - f_\ell(h_\ell)\big\| && \text{(triangle inequality)} \\
&\le \big\|f_{\ell+1}(h_{\ell+1}) - f_\ell(h_{\ell+1})\big\| + \Lambda_\ell \,\|h_{\ell+1} - h_\ell\| && \text{(local Lipschitz on } [h_\ell, h_{\ell+1}]) \\
&= \big\|f_{\ell+1}(h_{\ell+1}) - f_\ell(h_{\ell+1})\big\| + \Lambda_\ell \, V_\ell. && \square
\end{aligned}$$

Remark (Local Lipschitz constant). The constant Λ_ℓ is the Lipschitz constant of f_ℓ restricted to the segment $[h_\ell, h_{\ell+1}]$. When f_ℓ is differentiable, this can be expressed as $\Lambda_\ell = \sup_{t\in[0,1]} \|Df_\ell(h_\ell + t\,\Delta_\ell)\|_{\mathrm{op}}$. This is a local, input-dependent condition—not a global smoothness claim about the transformer block.

Remark (One-sidedness). The bound (7) is one-sided: low inter-layer disagreement *and* low velocity together imply low jerk, but the converse need not hold (low jerk can result from cancellation between the two terms). In particular, one cannot infer layer redundancy or functional equivalence from low jerk alone.

Remark (Curvature analogue). Deriving an analogous bound for curvature κ_ℓ would require bounding the Lipschitz constant of the unit-tangent map $U_\ell(x) = f_\ell(x)/\|f_\ell(x)\|$. The normalisation map $x \mapsto x/\|x\|$ is Lipschitz only away from the origin, with constant depending on a lower bound for $\|f_\ell\|$. Such a bound would need both (i) $\|f_\ell(h)\|$ bounded away from zero along the trajectory (nondegeneracy) and (ii) a contraction-type condition $\Lambda_\ell < 1$. These are significantly stronger than what the jerk decomposition requires, and we do not pursue them here.

E Connecting the Two Propositions

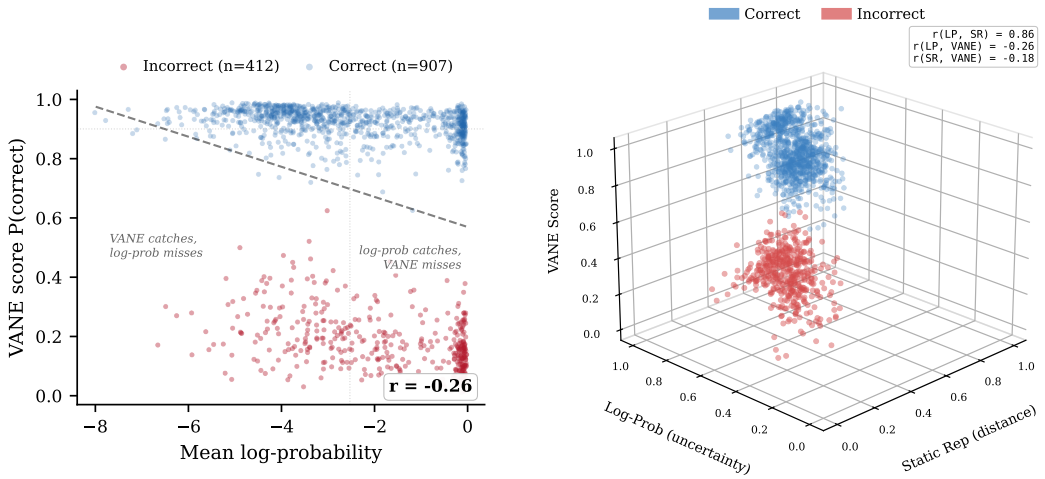
Proposition 1 is structural: it shows that trajectory features can vary within a fiber, so no endpoint-only function can reproduce them. This is a deterministic, worst-case statement; it does not by itself imply statistical improvement.

Proposition 2 is statistical: it quantifies the Bayes-optimal MSE gap from using trajectory features alongside the endpoint. The gap is strictly positive iff trajectory features carry conditional-mean information about Y beyond h_L.

Together, they establish that (a) there is room for trajectory features to carry information that endpoint-only features cannot, and (b) if they do, the improvement is quantified exactly by the within-fiber conditional variance of the trajectory-augmented Bayes predictor.

F Computational cost.

VANE adds negligible overhead: computing all five metrics for 1,319 samples takes under 2 minutes on a single A100; the classifier adds milliseconds per sample.



(a) VANE score vs. log-probability (Pearson $r = -0.26$). The upper-left quadrant contains failures VANE detects but log-prob misses.

(b) 3D scatter: VANE is independent of both log-prob ($r = -0.26$) and static representation ($r = -0.18$).

Figure 5: **Near-independence of geometric and token-level signals** (Llama-3-8B, GSM8K, N=1,319). VANE and log-probability are weakly correlated ($r = -0.26$), and combining them hurts performance.

G P(True) Baseline Details

Method. P(True) [Kadavath et al., 2022] asks the model to evaluate whether its own generated answer is correct. After generating an answer $\hat{a}$ to question q, we construct:

> *Question:* q
> *Proposed answer:* $\hat{a}$
> *Is this answer correct? Reply with only Yes or No.*

The confidence score is P(True) $= P(\text{“Yes”})/(P(\text{“Yes”}) + P(\text{“No”}))$, aggregated over multiple surface forms.

Results. P(True) achieves near-chance AUROC on all six models: 50.7 (Ministral-8B), 51.1 (Llama-3-8B), 51.2 (Qwen2.5-1.5B), 52.0 (Qwen3-14B), 53.7 (Gemma-3-12B), and 59.3 (Qwen2.5-72B). Even the highest value (59.3) is far below VANE's 85.5 on the same model.

G.1 Near-Independence from Log-Probability

A key property of VANE is that its geometric signals are only weakly correlated with token-level log-probabilities. Adding log-probability as an additional feature to VANE *decreases* AUROC on five of six models ($\Delta = -0.002$ to -0.007) and produces negligible change on the sixth. The Pearson correlation between VANE scores and log-probabilities is $r = -0.26$ (Figure 5b), indicating near-independence rather than strict orthogonality. VANE reveals a failure signal largely absent from output probabilities: cases where the model generates fluent, high-probability text while its internal computation is unstable.

H VANE Score Distributions

Figure 6 shows the distribution of VANE classifier outputs for correct and incorrect samples on Llama-3-8B (GSM8K). The distributions are well separated (KS $= 0.41$, $p < 0.001$).

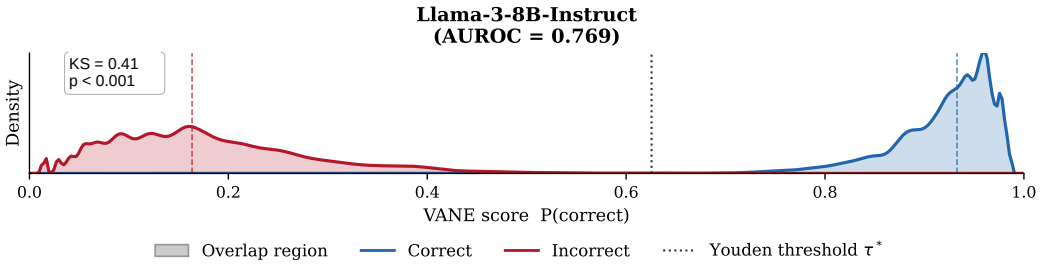
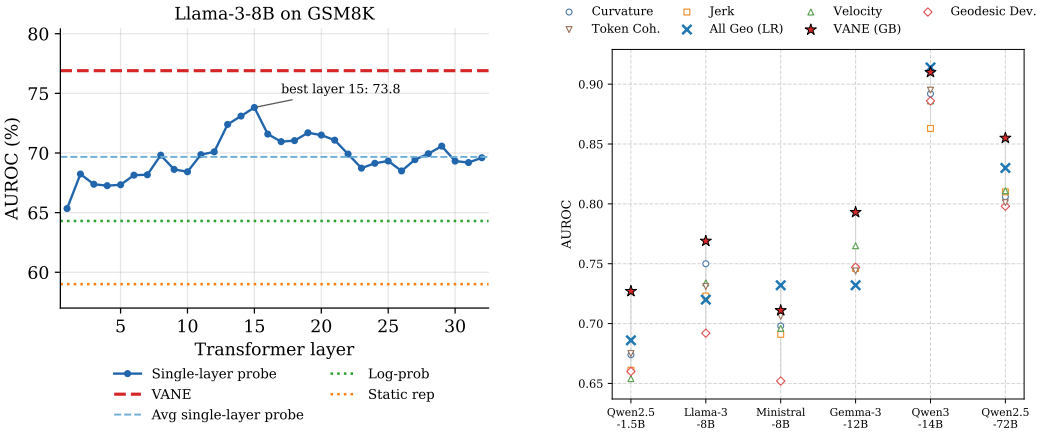


Figure 6: VANE classifier output distribution on Llama-3-8B (GSM8K). Correct samples (blue) cluster near $P(\text{correct})=1.0$; incorrect samples (red) cluster near 0 (KS $= 0.41$, $p < 0.001$).



(a) Per-layer probe AUROC.

(b) Per-metric AUROC.

Figure 7: (a) VANE exceeds the oracle-selected best layer on Llama-3-8B (GSM8K). (b) Every individual geometric metric outperforms all baselines, and linear combination nearly matches Gradient Boosting.

I Coverage Gain

Figure 8 isolates the accuracy gap between VANE and log-probability as a function of coverage. The advantage grows monotonically with increasing selectivity. Table 7 reports retained-set accuracy for all six models: on the 4-bit quantized Qwen2.5-72B, VANE reaches 96.1% at 50% coverage, compared with 87.1% for log-probability.

Table 7: Selective prediction accuracy (%) on GSM8K (out-of-fold) for all six models. V = VANE, LP = Log-Prob.

	Qwen2.5-1.5B		Llama-3-8B		Ministral-8B		Gemma-3-12B		Qwen3-14B		Qwen2.5-72B (4-bit)	
Coverage	V	LP	V	LP	V	LP	V	LP	V	LP	V	LP
100%	64.1	64.1	68.8	68.8	73.8	73.8	82.9	82.9	58.1	58.1	85.1	85.1
90%	67.9	62.8	73.5	67.3	77.0	73.4	86.9	82.2	63.4	54.4	91.0	86.5
80%	70.7	61.5	77.3	65.7	79.9	72.4	89.7	81.6	70.0	49.6	94.2	86.4
70%	73.7	59.4	79.7	63.2	82.0	71.6	91.4	81.7	77.9	43.8	94.7	86.6
60%	76.7	58.0	81.5	62.5	83.6	70.3	93.0	80.9	86.7	38.3	95.7	86.7
50%	79.2	54.9	84.1	60.4	86.0	68.1	94.4	81.0	93.2	38.4	96.1	87.1

Distributed across the trajectory, not concentrated at one layer. We fit logistic regression probes on the mean-pooled hidden state at every layer of Llama-3-8B. Even the best single-layer probe (layer 15, AUROC 0.738) falls 3.1 percentage points short of VANE (0.769), and the average single-layer probe reaches only 0.696. Figure 7a confirms this gap holds consistently: VANE exceeds every single-layer probe on Llama-3-8B (GSM8K).

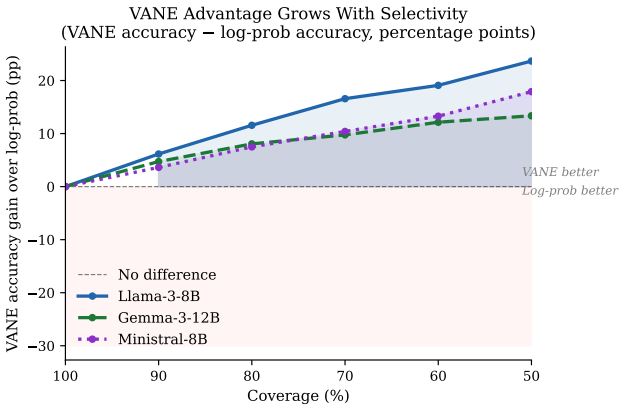


Figure 8: VANE accuracy gain over log-probability as coverage decreases. The gap grows monotonically, reaching +23.7% on Llama-3-8B at 50% coverage.

Cumulative layer ablation (Appendix K) explains why. When we include only layer 1, AUROC is 0.625; adding layers steadily improves performance until all layers yield 0.769. Later layers contribute more than earlier ones (second half: 0.741 vs. first half: 0.680), but neither half alone matches the full trajectory. In short, the failure signal builds up across the entire forward pass—no shortcut suffices.

J Token-Level Jerk Localization

Beyond global trajectory statistics, geometric jerk can be localized to individual tokens, revealing *where* in the reasoning chain the model's internal computation destabilizes. Figure 9 illustrates this on GSM8K #426, for which Llama-3-8B produces both a correct and an incorrect answer under temperature-0.8 sampling. The two generations share identical prompts and surface structure, yet their token-level jerk profiles (mid-layer mean) diverge sharply at the point of arithmetic failure, yielding Cohen's $d = 1.83$ between the two distributions. Colored spans in the figure encode low, medium, and high per-token jerk.

J.1 Classifier Sensitivity

To rule out classifier-specific overfitting, we replace Gradient Boosting with logistic regression and random forest (Appendix M). Logistic regression achieves 0.720 to 0.732 AUROC, competitive with Gradient Boosting and occasionally superior (Ministral: 0.732 vs. 0.703). On Qwen3-14B, logistic regression matches Gradient Boosting (0.914 vs. 0.910). That a linear model recovers most of the signal confirms it is a genuine geometric property, not a complex pattern exploited by tree ensembles.

Held-out split and sample complexity. As a second check against overfitting, we evaluate Llama-3-8B-Instruct on GSM8K with a fixed 60 / 20 / 20 train/val/test partition (averaged over 5 seeds): the held-out test AUROC is 0.735 ± 0.022, within one standard deviation of the 5-fold CV number (0.754). The agreement shows the 5-fold number is not inflated by the multiple re-use of samples across folds. Additionally, AUROC as a function of training-set size rises monotonically from 0.610 at 50 samples to 0.732 at 1,000 samples and then saturates near full-set performance (Appendix Figure 11), indicating that our n=1319 split is past the steep part of the learning curve.

K Why Trajectory Shape Matters: Ablation Experiments

A natural question is whether VANE's success is due to the trajectory *shape*, the ordered sequence of geometric changes across layers, or merely the aggregate statistics of individual layer representations. We design four ablation experiments to test this: single-layer probes, partial trajectories, layer-axis vs. sequence-axis geometry, and classifier sensitivity.

Question (GSM8K #426, ground truth = 11)

Aleksandra went to a restaurant for dinner. She ordered some soup, a bagel, and a piece of cake. The bagel cost $4, and the soup 25% more. The cake is only half of the price of the bagel. How much did Aleksandra need to pay for the dinner she ordered?

Correct Answer (pred = 11, gt = 11): jerk predominantly low

The bagel cost $4. The soup is 25% more than the bagel, so 25% of 4 is 1. So the cost of the soup is $4 + 1 = \$5$. The cake is half the price of the bagel, so the cost of the cake is $4 / 2 = \$2$. To find the total cost: $4 + 5 + 2 = 11$. So Aleksandra needs to pay \$11.

Incorrect Answer (pred = 12, gt = 11): jerk spikes at error

The bagel costs 4 dollars. 25% more than 4 dollars is $4 \times 0.25 = 1$ dollar. $4 + 1$ is 5. The cake is half the price of the bagel, which is 2 dollars. So the total cost of the meal is $5 + 5 + 2 = 12$. The answer is 12.

Figure 9: Token-level geometric jerk for the same GSM8K question (Llama-3-8B, T=0.8, mid-layer mean). High-jerk tokens (dark red) coincide with the arithmetic error ($5 + 5 + 2$=12 instead of $4 + 5 + 2$=11). Cohen's $d = 1.83$.

Table 8: Per-metric AUROC (%, 5-fold CV, mean ± std) on GSM8K. Individual metrics use logistic regression; VANE uses Gradient Boosting. Each geometric metric individually outperforms all baselines across all models. On Qwen3-14B, logistic regression over all geometric features (91.4%) matches Gradient Boosting (91.0%), confirming the signal is linearly separable.

Metric	Qwen2.5-1.5B	Llama-3-8B	Ministral-8B	Gemma-3-12B	Qwen3-14B	Qwen2.5-72B
Curvature	67.4±4.1	75.0±3.4	69.8±1.7	74.7±5.2	89.2±1.7	80.6±1.0
Jerk	66.1±2.7	72.3±2.2	69.1±2.1	74.6±4.3	86.3±1.4	81.0±3.6
Velocity	65.4±2.6	73.4±1.5	69.6±3.9	76.5±6.8	88.6±1.5	81.1±1.7
Geodesic Dev.	66.0±2.6	69.2±2.1	65.2±3.2	74.7±6.2	88.6±1.3	79.8±3.3
Token Coh.	67.5±2.3	73.1±1.4	70.6±4.1	74.4±5.6	89.5±1.2	80.1±3.4
All Geo. (LR)	68.6±2.7	72.0±2.1	73.2±1.6	73.2±5.4	91.4±1.3	83.0±1.6
VANE (GB)	**72.7±2.8**	**76.9±1.8**	**71.1±3.0**	**79.3±3.3**	**91.0±2.2**	**85.5±3.1**

K.1 Single-Layer Probe Analysis

We train independent logistic regression probes ($\mathbb{R}^{4096}$, L-BFGS, $C = 1$, 5-fold CV) on the mean-pooled hidden state at each transformer layer of Llama-3-8B. The average single-layer probe achieves 69.6% AUROC (+10.6 over static rep), and even oracle selection of the best layer (layer 15, 73.8%) falls 3.1 points short of VANE (76.9%). Figure 7a shows that VANE consistently exceeds all single-layer probes: the trajectory contains information that no individual layer snapshot can recover.

K.2 Partial Trajectory Analysis

To test whether the signal is concentrated in specific layers or distributed across depth, we evaluate VANE using cumulative layer windows on Llama-3-8B. Figure 10 shows AUROC as a function of the number of layers included (layers 1 through k, for $k = 1, \ldots, 31$).

The full trajectory (0.769) outperforms every partial window and the best single-layer oracle (0.738), confirming that the signal is genuinely trajectory-level. The second half carries more weight (0.741

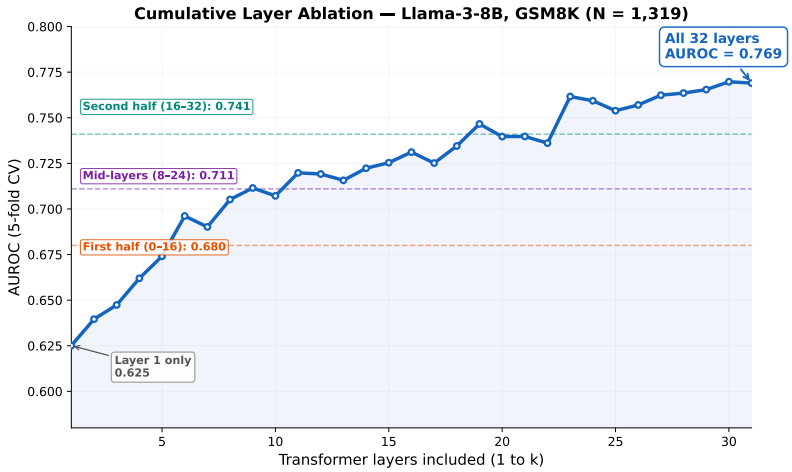


Figure 10: Cumulative layer ablation (Llama-3-8B, GSM8K, 5-fold CV). AUROC rises steadily from 0.625 (layer 1 only) to 0.769 (all layers). Dashed lines mark partial-window baselines: first half (0.680), mid-layers (0.711), and second half (0.741). The full trajectory outperforms every partial window.

vs. 0.680), consistent with reasoning errors accumulating in deeper layers, but the steady climb from layer 1 shows that early layers also contribute meaningfully.

L Noise Robustness (Full Results)

Table 9 reports AUROC under increasing Gaussian noise injected into the geometric feature vector before classification. Optuna tuning is disabled to isolate noise effects, so clean baselines are slightly lower than Table 1.

Table 9: Noise robustness: full results (AUROC, %, GSM8K, default hyperparameters). Gaussian noise (σ relative to per-feature std) added before classification.

Noise σ	Llama-3-8B	Gemma-3-12B	Ministral-8B
0 (clean)	75.3	79.1	70.3
0.01	74.8	77.7	69.3
0.05	74.7	76.3	69.0
0.10	75.1	75.9	68.7
0.25	73.2	74.9	66.7
0.50	71.9	69.7	64.5
1.00	69.2	63.3	63.0

M Classifier Sensitivity (Full Results)

Table 10 compares three classifier families on identical geometric features. Logistic regression recovers most of the signal, confirming that the geometric separation is approximately linear.

Table 10: Classifier type sensitivity: full results (AUROC %, GSM8K).

Classifier	Llama-3-8B	Gemma-3-12B	Ministral-8B
Gradient Boosting	**75.3**	**79.1**	70.3
LogisticRegression	72.0	73.2	**73.2**
RandomForest	72.5	71.7	68.0

Sample complexity. Figure 11 shows AUROC as a function of training-set size for the Gradient Boosting classifier on Llama-3-8B-Instruct GSM8K, averaged over 5 random subset seeds at each size. The curve rises monotonically from 0.610 ± 0.013 at 50 training samples to 0.732 ± 0.022 at 1,000 samples, then saturates near the full-data CV number of 0.754. Combined with the held-out

60 / 20 / 20 test AUROC of 0.735 ± 0.022 (main paper), this indicates that the classifier learns a stable signal well before the full 1,319 samples are used.

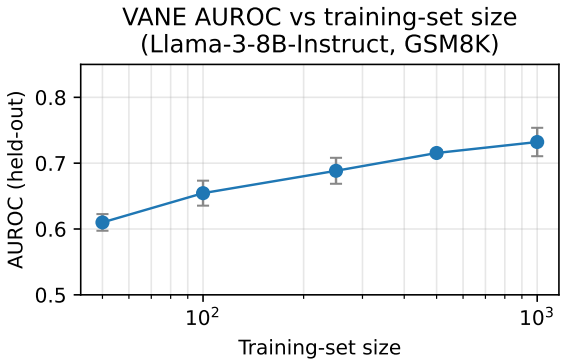


Figure 11: AUROC vs. training-set size for VANE features on Llama-3-8B-Instruct (GSM8K). Bars show standard deviation over 5 seeds per training-set size. Remaining samples used as held-out test set at each size.

N P(True) Cross-Benchmark Results

Table 11 extends the P(True) evaluation to MATH-500 and HumanEval. Verbalized self-assessment remains near chance across all three benchmarks and models.

Table 11: P(True) AUROC (%) across benchmarks. Self-assessment is consistently near-chance.

Model	GSM8K	MATH-500	HumanEval
Llama-3-8B	51.1	54.5	48.3
Gemma-3-12B	53.7	51.1	54.6
Ministral-8B	50.7	60.8	48.3

O Per-Layer Curvature Profiles

Figure 12 shows per-layer Frenet–Serret curvature for correct and incorrect samples. The divergence concentrates in mid-to-late layers, consistent with reasoning errors arising during intermediate computation.

P Per-Metric Discriminative Strength

Figure 13 shows the per-layer discriminative strength of each VANE metric, measured as |Cohen's d| between correct and incorrect samples, for two representative models: Gemma-3-12B-IT and Qwen3-14B. For each metric and each depth bin, Cohen's d is computed on the bin-averaged metric values for correct vs. incorrect samples; the absolute value captures separation strength regardless of sign. Qwen3-14B exhibits substantially stronger separation ($|d| > 1.0$) than Gemma-3-12B ($|d| \leq 0.45$), consistent with their respective AUROC gap (0.910 vs. 0.793). On both models, discriminative power concentrates in mid-to-late layers and no single metric dominates uniformly: curvature is strongest at depth $\sim$0.85 on Qwen3, while token coherence peaks at depth $\sim$0.5 on Gemma. This supports the composite view of trajectory instability underlying VANE.

Q Camera-ready Additional Controls

These experiments were reported during author response and are included here to make the camera-ready claims auditable. They use 5-fold out-of-fold AUROC unless noted otherwise and are separate matched runs, not subsets of Table 1.

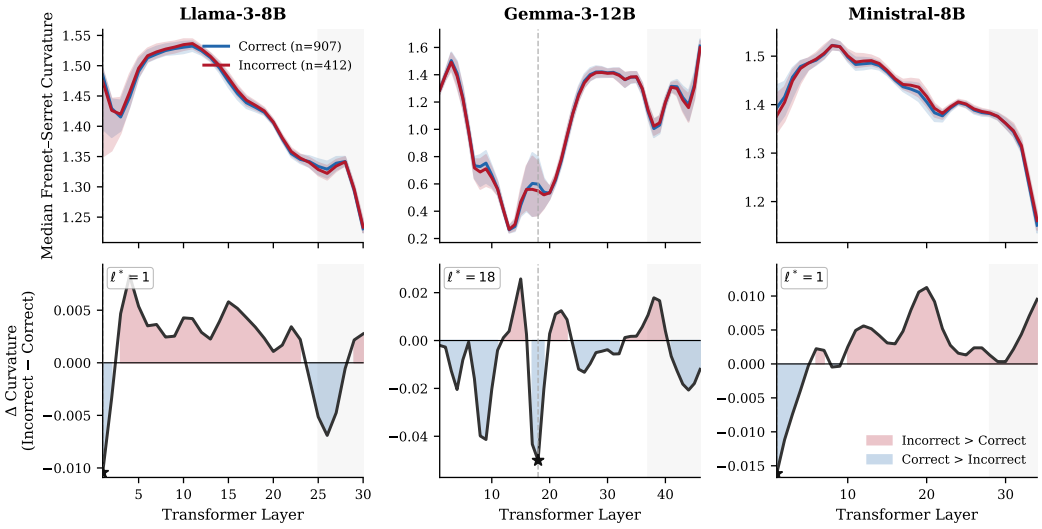


Figure 12: Per-layer Frenet–Serret curvature profiles (GSM8K). **Top:** median curvature for correct (blue) and incorrect (red) samples. **Bottom:** difference curves. ℓ^* marks the most discriminative layer.

Table 12: Matched non-oracle all-layer probe. The strongest pooled/concatenated logistic-regression, gradient-boosting, or small-MLP probe is reported per cell. "Wins" counts VANE wins across evaluated benchmarks.

Model	VANE	All-layer probe	Wins
Qwen2.5-1.5B	.748	.741	3/4
Llama-3-8B	.753	.765	0/4
Ministral-8B	.761	.748	2/4
Gemma-3-12B	.807	.781	3/4
Qwen3-14B	.810	.740	3/4
Qwen3-14B (4-bit)	.830	.727	3/4
Qwen2.5-72B (4-bit)	.833	.751	2/3
Mean / total	.790	.750	16/27

Table 13: MATH-500 difficulty control (n=300, T=0.7). "Within" scores VANE separately within each fixed official difficulty level; "+Diff." adds difficulty as an input.

Model	Difficulty only	VANE	Within	VANE + Diff.
Qwen2.5-1.5B	.681	.723	.685	.731
Ministral-8B	.731	.810	.760	.830
Gemma-3-12B	.675	.832	.809	.831
Qwen3-14B	.690	.817	.798	.816
Qwen2.5-72B (4-bit)	.681	.800	.765	.800

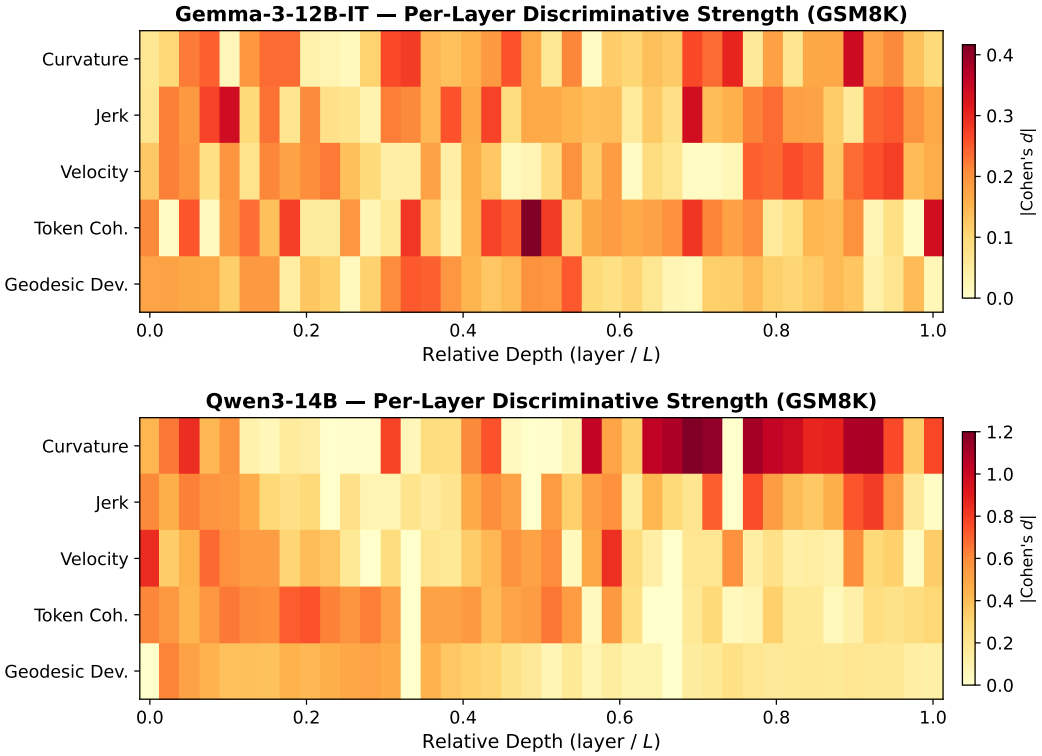


Figure 13: Per-layer discriminative strength (|Cohen's d|) of each VANE metric for Gemma-3-12B-IT (top) and Qwen3-14B (bottom) on GSM8K. Warmer colors indicate stronger separation between correct and incorrect samples. No single metric dominates across all layers; discriminative power concentrates in mid-to-late layers on both models.

Table 14: Surface-feature controls. The bundle contains response length, word/line counts, code markers, digit ratio, and equations.

Model	VANE	Surface bundle	Length only	Wins
Qwen2.5-1.5B	.748	.623	.529	3/4
Llama-3-8B	.757	.630	.567	5/5
Ministral-8B	.761	.623	.546	4/4
Gemma-3-12B	.807	.769	.639	3/4
Qwen3-14B	.797	.747	.641	4/5
Qwen2.5-72B (4-bit)	.824	.681	.608	4/4
Mean / total	.782	.680	.590	23/26

Table 15: Decoding robustness and matched Self-Consistency (SC). Sampled VANE uses top-p=0.95. SC comparisons use the same first 300 problems per cell; VANE is subset-tuned with 5-fold out-of-fold scoring, so these values carry small-sample variance and tuning optimism.

Model	Benchmark	Greedy	Sampled	SC (best K)	VANE (n=300)
Gemma-3-12B	BBH	.911	.746	–	–
Gemma-3-12B	MATH-500	.927	.906	–	–
Qwen2.5-1.5B	BBH	.738	.616	.513 (K=5)	.797
Qwen2.5-1.5B	MATH-500	.775	.604	.776 (K=9)	.814
Qwen3-14B	BBH	–	–	.618 (K=9)	.730
Qwen3-14B	MATH-500	–	–	.672 (K=9)	.747

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction claim (i) a trajectory-geometry signal that separates correct from incorrect reasoning, (ii) AUROC gains over log-probability across six instruction-tuned models and five benchmarks, (iii) near-independence from log-probability, (iv) attenuation on base models, and (v) selective-prediction gains. These are supported in §4, §G.1, §4.4, and §5.1. The abstract flags the 4-bit quantization of the 72B result, and the conclusion explicitly scopes the claims and states that they are correlational.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The dedicated "Limitations" paragraph in §6 discusses white-box access, supervised labels, partial transfer, prompt/decoding/distribution shift, residual confounding, privacy and security implications, and the exclusion of open-ended, long-form, multi-turn, and agentic settings. VANE is framed as an observable proxy, not a causal explanation.

3. **Theory, Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Appendices B and C state the assumptions and provide full proofs for the structural endpoint/trajectory result and trajectory-gain bound. They are explicitly framed as theoretical characterizations rather than empirical guarantees.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: §3 gives finite-difference definitions of all five VANE metrics, the three token aggregations (`max`, `mean`, `ans`), and the feature-vector construction; Algorithm 1 in Appendix A.2 summarises the full pipeline. §4 specifies the six open-weight models, five benchmarks, decoding, 5-fold stratified cross-validation, and classifier. All models and benchmarks are publicly available.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code is available at `https://github.com/hodfa840/geometric-instability-of-hidden-state-trajectories`, with instructions to run generation, feature extraction and evaluation for all models and benchmarks. All datasets (GSM8K, MATH-500, HumanEval, CRUXEval, BBH) and models (Qwen2.5, Qwen3, Llama-3, Ministral, Gemma-3, Mistral) are publicly available, and the full pipeline is specified in §3 and Algorithm 1.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

Justification: §3.3 specifies the classifier (GradientBoosting), hyperparameter search (Optuna, 80 trials, 5-fold stratified CV), feature dimensionality, decoding (greedy, except Fig. 9 which uses T=0.8 and is flagged in the caption), and evaluation protocol (5-fold stratified cross-validation with reported mean $\pm$ std). The 72B model is explicitly noted as 4-bit quantized; Qwen3-14B is run in default (non-thinking) mode.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Tables 1 and 8 report mean $\pm$ standard deviation over 5 folds of stratified cross-validation. Figure 2 reports ± 2 standard errors. P(True) numbers are single deterministic evaluations. The Kolmogorov–Smirnov test for the VANE score distribution (Figure 6) reports $p < 0.001$.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix F reports that VANE metric computation for 1,319 samples takes under 2 minutes on a single A100, classifier training under 60 seconds on CPU, and per-sample inference in milliseconds. Hidden-state extraction for each of the six models on each benchmark corresponds to a single forward pass per sample on the same hardware tier; the 72B model was run on a single A100 with 4-bit quantization.

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses publicly released models and benchmarks, involves no human subjects or new data collection, and does not release any new model. We have reviewed the NeurIPS Code of Ethics and the research conforms to it.

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The primary intended impact is positive: reliable failure detection enables safer deployment of LLM reasoning systems via abstention and selective prediction. Potential negative impacts are that (i) white-box detectors may give a false sense of reliability if deployed outside the evaluated setting, and (ii) the same signal could in principle be adversarially suppressed by models trained to appear geometrically smooth. Both are noted in the discussion.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release any new data, model, or high-risk artifact. It evaluates publicly released models on publicly released benchmarks.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All models (Qwen2.5/Qwen3, Llama-3, Ministral, Gemma-3, Mistral-7B-v0.1) and benchmarks (GSM8K, MATH-500, HumanEval, CRUXEval, BBH) are cited to their original papers. All are used in accordance with their respective licenses; research use is permitted for each. Specific licenses: GSM8K (MIT), MATH (MIT), HumanEval (MIT), CRUXEval (MIT), BBH (MIT); model weights under their respective open-weight licenses (Llama Community License, Gemma Terms of Use, Apache 2.0 for Qwen and Mistral variants).

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new models or datasets. Code for reproducing the VANE pipeline will be released upon acceptance.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human subjects.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human subjects and therefore does not require IRB approval.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research?

    Answer: [N/A]

    Justification: LLMs are the subject of study (we analyse their hidden-state trajectories), not a tool used for core methodology development. LLMs were used only for incidental writing and editing assistance, which does not require declaration under NeurIPS 2026 guidelines.