LDIR: A LOCAL DIRECTION-INFORMED RIEMANNIAN METRIC FOR LLM REASONING TRAJECTORIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models generate reasoning as trajectories through high-dimensional hidden state spaces, yet these trajectories are commonly analyzed using fixed Euclidean geometry. Such geometry may fail to capture the nonlinear structure of learned representations and the directional organization of reasoning dynamics. We propose a Local Direction Informed Riemannian metric (LDIR) for analyzing chain-of-thought trajectories under a geometry adapted to both representation structure and local transition directions. We first map hidden states to a latent space using a geometry-regularized variational autoencoder and use its decoder to induce a pullback metric. We then refine this metric using local directional statistics of adjacent token transitions, yielding a post pullback anisotropic geometry. We evaluate the resulting geometric features on controlled matched valid–invalid reasoning pairs from MATH500 across multiple language models, with a separate external evaluation on MGSM. Trajectory validity is defined by benchmark final answer correctness. LDIR achieves the highest mean AUROC among the compared geometries in all evaluated settings, with the clearest improvements in several Qwen settings; its small numerical gain on MGSM is statistically inconclusive. Speed related features show the clearest improvement among restricted feature subsets, and LDIR retains the highest mean AUROC after each geometric feature is residualized against log-transformed token transition count. Exploratory matched pair comparisons associate trajectories ending in incorrect answers with shorter metric paths, smaller adjacent token metric displacements, and less cumulative turning. These findings support the diagnostic value of locally conditioned trajectory geometry for final answer correctness. Our code will be made publicly available.

1 INTRODUCTION

Chain-of-thought (CoT) prompting elicits intermediate reasoning steps and improves performance on a range of reasoning tasks (Wei et al., 2022; Kojima et al., 2022). Generated explanations, however, need not faithfully reflect the computations behind an answer (Turpin et al., 2023; Lanham et al., 2023). Beyond the text, each token position has an internal representation, so a reasoning trace can be viewed as an ordered token-indexed trajectory. Recent studies have linked representation geometry to factual truth (Marks & Tegmark, 2023) and reasoning correctness (Sun et al., 2026; Masoomi et al., 2026). Several approaches characterize these structures using linear directions, displacements, and trajectory level position, velocity, and curvature (Marks & Tegmark, 2023; Zhou et al., 2026; Masoomi et al., 2026).

Neural representations can have an intrinsic dimension well below their ambient dimension (Ansuini et al., 2019), and recent work examines low-dimensional constraints in LLM inference trajectories (Ma et al., 2026). Although the ambient hidden space is Euclidean, directly applying the identity metric after mapping a trajectory into learned latent coordinates does not generally preserve the magnitude of the corresponding decoded changes (Arvanitidis et al., 2017; Shao et al., 2018). The interpretation of representation geometry can also depend on dominant dimensions and the chosen inner product (Timkey & Van Schijndel, 2021; Park et al., 2023). A decoder induced pullback metric accounts for local decoder scaling (Shao et al., 2018), but does not by itself incorporate the directions connecting successive reasoning states.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

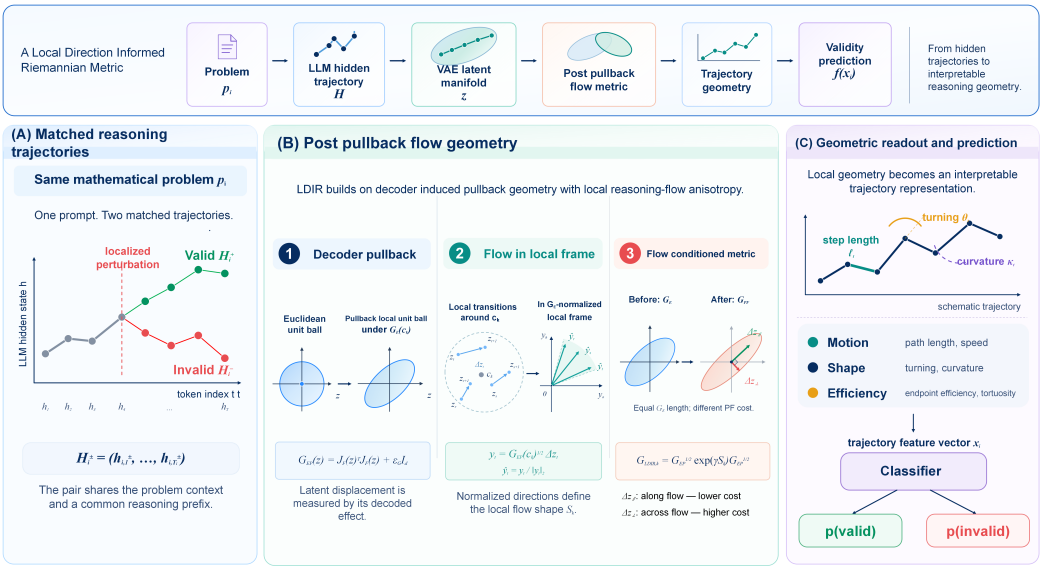

Figure 1: Overview of LDIR. (A) A localized perturbation produces matched valid and invalid token level hidden state trajectories for the same problem. (B) GeoVAE maps the trajectories to a latent manifold, where local transition axes in a pullback-orthonormal frame condition the decoder induced geometry to yield LDIR. (C) Motion, shape, and endpoint efficiency summaries form a trajectory feature vector for validity prediction.

To address this limitation, we propose the Local Direction Informed Riemannian metric (LDIR) for reasoning trajectories on a learned latent manifold. LDIR refines the decoder induced pullback geometry through post pullback flow conditioning based on local transition directions. We first use a variational autoencoder (VAE) (Kingma & Welling, 2013) to obtain compact representations of high dimensional hidden states. The decoder Jacobian induces a Riemannian pullback metric on the latent space (Arvanitidis et al., 2017; Shao et al., 2018). We then express local trajectory displacements in pullback-orthonormal coordinates, estimate their directional structure, and use it to deform the pullback metric locally. This construction uses decoder induced manifold geometry as its reference while capturing anisotropy in reasoning dynamics. It is estimated without validity labels, remains positive definite, varies smoothly over the latent space, and reduces to the shared interpolated pullback field when flow conditioning is disabled.

Figure 1 summarizes the complete pipeline. We use LDIR to construct interpretable trajectory level features for reasoning analysis and validity prediction. These features summarize metric path length, speed, turning, curvature, endpoint efficiency, and tortuosity, and are supplied to a downstream classifier for distinguishing valid from invalid reasoning. We further analyze matched valid–invalid pairs constructed from the same mathematical problem, allowing geometric changes associated with controlled reasoning errors to be examined while holding the problem context fixed. Across the evaluated model families and reasoning-generation settings, LDIR achieves the strongest mean discrimination among the compared geometries, although the magnitude of the improvement varies across settings. In the feature subset analysis, speed related measurements provide the clearest evidence of improvement. Exploratory matched-pair comparisons associate incorrect answer trajectories with shorter paths, smaller adjacent-token metric displacements, and less total turning. LDIR retains the highest mean AUROC after each feature is residualized against log-transformed token transition count. Together, these results support the diagnostic value of locally conditioned, token indexed measurements after this specific length control.

Our main contributions are as follows:

- **A local direction informed Riemannian geometry.** We introduce a Riemannian metric, estimated without validity labels after pair construction, that combines decoder induced manifold geometry with the local directional organization of reasoning trajectories. The

construction is smooth, positive definite, and reduces exactly to the shared interpolated EP field when flow conditioning is disabled.

- **Geometric validity prediction and interpretation.** We use the proposed metric to construct interpretable trajectory features and show that they provide stronger mean discrimination between valid and invalid reasoning than the corresponding baseline geometries. Exploratory matched pair analyses further identify geometric signatures associated with invalid reasoning.
- **Cross model and confound controlled evaluation.** We evaluate LDIR across multiple model families and reasoning-generation settings and analyze whether its advantage can be explained by token count.

2 PRELIMINARIES

2.1 REASONING TRAJECTORY REPRESENTATION

For a reasoning example i with T_i generated token positions, we extract an ordered sequence of final layer hidden representations $H_i = (h_{i,1}, \ldots, h_{i,T_i})$, where $h_{i,t} \in \mathbb{R}^D$. Thus, t indexes generated tokens and each transition connects adjacent token states; no semantic step pooling is used. We standardize the hidden states using training fold statistics and denote the resulting states by $\tilde{h}_{i,t}$. Their low-dimensional encoding forms a latent trajectory $Z_i = (z_{i,1}, \ldots, z_{i,T_i})$, with $z_{i,t} \in \mathbb{R}^d$ and $d \ll D$. The next two subsections specify this encoding and the geometry used to measure latent transitions.

2.2 VARIATIONAL LATENT REPRESENTATION

We use GeoVAE, a geometry regularized variational autoencoder based on the VAE framework of Kingma & Welling (2013), to obtain compact latent coordinates while retaining information about the structure of the hidden representations. Its encoder $q_\phi(z \mid \tilde{h})$ parameterizes a diagonal Gaussian with mean $\mu_\phi(\tilde{h})$ and variance $\sigma^2_\phi(\tilde{h})$, while the decoder $F_\theta : \mathbb{R}^d \to \mathbb{R}^D$ maps latent states back to standardized hidden space. Training combines reconstruction, prior regularization, and relative-distance preservation:

$$\mathcal{L}_{\text{GeoVAE}} = \mathcal{L}_{\text{rec}} + \beta_e \mathcal{L}_{\text{KL}} + \lambda_{\text{geo}} \mathcal{L}_{\text{geo}}. \tag{1}$$

Here $\mathcal{L}_{\text{rec}}$ is the reconstruction mean squared error, $\mathcal{L}_{\text{KL}}$ regularizes the posterior toward a standard normal prior, and $\mathcal{L}_{\text{geo}}$ penalizes discrepancies between log-transformed, median normalized pairwise distances in hidden and latent space. The epoch dependent weight β_e controls prior regularization, while λ_{geo} controls distance preservation. The distance term encourages relative geometric structure to survive compression, without imposing exact distance preservation. Although training uses posterior samples, trajectory analysis uses the deterministic mean $z_{i,t} = \mu_\phi(\tilde{h}_{i,t})$ so that sampling noise does not alter the measured trajectory. The complete loss definitions, architecture, and training settings are provided in Appendix A.

2.3 DECODER INDUCED PULLBACK GEOMETRY

Relative distance preservation does not make the learned coordinates isometric to hidden space. A latent displacement can produce different amounts of decoded change depending on its location and direction. For a differentiable decoder with Jacobian $J_F(z) = \partial F_\theta(z)/\partial z$, the first order relation $F_\theta(z + \delta z) - F_\theta(z) \approx J_F(z)\delta z$ connects these two quantities. Measuring the decoded change with the Euclidean inner product follows the decoder induced geometry studied by Arvanitidis et al. (2017); Shao et al. (2018). With a numerical ridge, the local reference matrix is

$$G^\star_{\text{EP}}(z) = J_F(z)^\top J_F(z) + \epsilon_{\text{proto}} I_d, \qquad \epsilon_{\text{proto}} > 0. \tag{2}$$

The ridge ensures positive definiteness even when the Jacobian is rank deficient. For a positive definite matrix M, we write $\|v\|_M = (v^\top M v)^{1/2}$; thus $\|\delta z\|_{G^\star_{\text{EP}}(z)}$ measures a latent displacement relative to the local decoder-induced geometry. The implementation evaluates this matrix at training derived prototypes and interpolates the prototype matrices to form the final EP field. We refer to this geometry as the Euclidean Pullback (EP) baseline. EP accounts for local representational scaling

but does not explicitly use the directions connecting consecutive reasoning states, which motivates the flow conditioning below.

3 METHOD

LDIR uses local adjacent-token transitions to refine the pullback reference geometry. We first account for decoder induced scaling, then estimate how transition directions are distributed locally and use this structure to adjust their relative metric costs. The resulting Local Direction-Informed Riemannian metric (LDIR) uses post pullback flow conditioning and supports a common set of trajectory summaries for validity prediction. After pair construction, neither the latent representation nor the metric is fitted using validity labels; supervision enters only at the downstream classification stage. We omit the example index i when describing individual transitions.

3.1 LOCAL DIRECTION-INFORMED RIEMANNIAN METRIC

Local directional structure. Transition directions may vary across latent space, so a single global estimate would discard their local organization. We instead collect transitions around representative locations. For consecutive states, let $\Delta z_t = z_{t+1} - z_t$ and $m_t = (z_t + z_{t+1})/2$ denote the displacement and its midpoint. Clustering the training-transition midpoints with K-means gives prototypes $\{c_k\}_{k=1}^{K}$, where $\mathcal{N}_k$ contains the transitions assigned to c_k. Each prototype is a geometric anchor, not a class representative. At each prototype, the decoder-induced reference matrix is $G_{\mathrm{EP},k} = G^{\star}_{\mathrm{EP}}(c_k)$.

Local directions should be compared after accounting for decoder induced scaling, rather than directly in latent coordinates. At prototype c_k, we therefore transform each displacement to $y_t = G^{1/2}_{\mathrm{EP},k}\Delta z_t$, using the symmetric positive definite square root. Since $\|y_t\|_2 = \|\Delta z_t\|_{G_{\mathrm{EP},k}}$, these are pullback-orthonormal coordinates for the local displacements. We then set $\hat{y}_t = y_t / \max\{\|y_t\|_2, \epsilon_y\}$ with a positive numerical floor ϵ_y, preventing large steps from dominating the directional estimate. For prototypes with $|\mathcal{N}_k| \geq 5$, these normalized directions are summarized by their uncentered second moment, with isotropic shrinkage and a ridge to stabilize estimation:

$$C_k = \frac{1}{|\mathcal{N}_k|}\sum_{t\in\mathcal{N}_k} \hat{y}_t\hat{y}_t^{\top}, \qquad C'_k = (1-\lambda_{\mathrm{sh}})C_k + \left(\frac{\lambda_{\mathrm{sh}}}{d} + \epsilon_C\right) I_d. \tag{3}$$

Here the prime in C'_k denotes regularization; $\lambda_{\mathrm{sh}} \in [0,1]$ controls shrinkage toward I_d/d, and $\epsilon_C > 0$ ensures invertibility. Because $\hat{y}_t\hat{y}_t^{\top}$ is unchanged by $\hat{y}_t \mapsto -\hat{y}_t$, this statistic estimates local axes rather than oriented flow vectors.

Flow-conditioned deformation. We use the inverse directional statistic to assign lower relative costs to directions with larger second moments. Specifically, the scale normalized precision is $\widetilde{P}_k = d(C'_k)^{-1}/\operatorname{tr}((C'_k)^{-1})$. To separate directional reweighting from a uniform change of scale, we remove the isotropic component of its matrix logarithm. Exponentiating the remaining symmetric adjustment maintains positive definiteness, while mapping it back through the pullback square root anchors the deformation to the reference geometry:

$$\begin{aligned} S_k &= \log \widetilde{P}_k - \frac{\operatorname{tr}(\log \widetilde{P}_k)}{d} I_d, \\ B_k(\gamma) &= \mathrm{clip}^{\mathrm{spec}}_{10}(\gamma S_k), \\ G_{\mathrm{LDIR},k} &= G^{1/2}_{\mathrm{EP},k} \exp(B_k(\gamma)) G^{1/2}_{\mathrm{EP},k}. \end{aligned} \tag{4}$$

Here $\mathrm{clip}^{\mathrm{spec}}_{10}$ clips the eigenvalues of its symmetric argument to $[-10, 10]$, as in the implementation. The parameter $\gamma \geq 0$ controls the conditioning strength. Since $\operatorname{tr}(S_k) = 0$, the unclipped adjustment $\exp(\gamma S_k)$ has unit determinant. Spectral clipping need not preserve that property; it also does not affect the following $\gamma = 0$ reduction. At $\gamma = 0$, each prototype metric reduces to $G_{\mathrm{EP},k}$; positive values introduce anisotropy according to the local directional statistics. When C'_k is isotropic, $\widetilde{P}_k = I_d$ and $S_k = 0$, so $G_{\mathrm{LDIR},k} = G_{\mathrm{EP},k}$ even for $\gamma > 0$. Thus the deformation requires local directional anisotropy. This weighting reflects the organization of observed transitions, not an assumption that prevalent directions correspond to correct reasoning.

A continuous metric field. Trajectory measurements require a metric between prototypes as well as at them. Using hard cluster assignments would introduce jumps at neighborhood boundaries, so the Euclidean Pullback (EP), Hidden-Space Anisotropic Pullback (HAP), and LDIR geometries use the same smooth radial-basis interpolation:

$$G_A(z) = \sum_{k=1}^{K} w_k(z) G_{A,k} + \epsilon_{\text{field}} I_d, \qquad A \in \{\text{EP}, \text{HAP}, \text{LDIR}\}. \tag{5}$$

The weights satisfy $w_k(z) \propto \exp(-\|z - c_k\|_2^2/(2\sigma^2))$ and are normalized to sum to one, with bandwidth σ estimated from within-cluster distances. Smooth, nonnegative interpolation of positive definite prototype matrices yields a smooth positive definite field. The prototype ridge $\epsilon_{\text{proto}} I_d$ and field ridge $\epsilon_{\text{field}} I_d$ are distinct, and the shared construction gives $G_{\text{LDIR}}(z; 0) = G_{\text{EP}}(z)$ exactly. For neighborhoods containing fewer than five transitions, we use $G_{\text{EP},k}$ directly. Bandwidth estimation, spectral clipping, and numerical settings are given in Appendix B.

3.2 Metric-Based Trajectory Geometry

A single path length cannot reveal whether movement is distributed evenly across token transitions or accompanied by frequent directional changes. We therefore summarize movement, turning, and endpoint relative path structure separately before combining them for diagnosis. The definitions below apply to a latent trajectory $Z = (z_1, \ldots, z_T)$ under any of the compared metric fields G; the feature set is unchanged across geometries.

Displacement and speed. To distinguish accumulated movement from its distribution across token transitions, we retain both total length and local displacement statistics. Using the midpoint approximation, each transition has length $\ell_t^G = \|\Delta z_t\|_{G(m_t)}$, and the total path length is $L^G = \sum_{t=1}^{T-1} \ell_t^G$. Because token index supplies the discrete parameter, ℓ_t^G is reported as token indexed metric speed. It is an adjacent-token metric displacement, not wall clock speed or semantic step speed. We retain L^G and the mean, standard deviation, and maximum of ℓ_t^G. Since $L^G = (T-1)\,\text{mean}_t(\ell_t^G)$, accumulated path length depends on token count even when average local movement is unchanged. This identity motivates examining token count separately in Section 4.4.

Turning and curvature. Displacement magnitude alone does not capture changes of direction. We measure the angle between consecutive transitions in a common local metric and normalize it by their mean length to distinguish a sharp turn over a short segment from the same angular change over a longer one. For $t = 2, \ldots, T-1$, let $u_t = z_t - z_{t-1}$ and $v_t = z_{t+1} - z_t$. Then

$$\theta_t^G = \arccos\left(\frac{u_t^\top G(z_t) v_t}{\|u_t\|_{G(z_t)} \|v_t\|_{G(z_t)}}\right), \qquad \kappa_t^G = \frac{2\theta_t^G}{\|u_t\|_{G(z_t)} + \|v_t\|_{G(z_t)}}. \tag{6}$$

The cosine argument is clipped to $[-1, 1]$ before applying $\arccos$. We interpret κ_t^G as a discrete turning-per-length proxy in the learned coordinates and retain the mean and maximum of both θ_t^G and κ_t^G.

Endpoint efficiency and tortuosity. Local movement and turning do not indicate how much path length is accumulated relative to the separation of the endpoints. We therefore compare the observed trajectory with the straight coordinate path $\gamma_{\text{str}}(s) = z_1 + s(z_T - z_1)$, $s \in [0, 1]$, measured under the same metric. Its length is $L_{\text{str}}^G = \int_0^1 \|z_T - z_1\|_{G(\gamma_{\text{str}}(s))}\, ds$, approximated from 32 grid points forming 31 equal segments. We retain this length, the endpoint-efficiency ratio $\eta^G = L_{\text{str}}^G / L^G$, and the reciprocal tortuosity ratio $\tau^G = L^G / L_{\text{str}}^G$.

Together, these summaries form an 11-dimensional feature vector x_i^G. We standardize the features using training fold statistics and fit a separate logistic-regression validity classifier for each geometry. Problem grouped cross validation keeps matched valid–invalid trajectories in the same fold, with all learned components fitted on training data only. The complete feature specification and evaluation protocol are provided in Appendices B and C, respectively.

4 EXPERIMENTS

Our experiments examine whether incorporating local transition structure into the metric makes reasoning trajectories more informative for validity diagnosis. We organize the evaluation around three research questions:

RQ1. Does LDIR improve trajectory level validity prediction across model and trajectory-source settings?

RQ2. Which trajectory measurements benefit from flow conditioning, and how do matched valid and invalid trajectories differ geometrically?

RQ3. To what extent can the observed advantage be explained by token count?

4.1 EXPERIMENTAL SETUP

Data and models. The primary dataset contains 121 problems drawn from MATH500, a 500 problem subset of MATH (Hendrycks et al., 2021),[1] and 242 matched valid–invalid trajectories. Invalid trajectories are produced by localized reasoning perturbations while keeping the underlying problem fixed. We label a trajectory valid when its final answer matches the benchmark answer; intermediate inferences are not individually verified. Qwen2.5-7B-Instruct (Qwen et al., 2025) is the primary backbone. We additionally evaluate Qwen2.5 models at multiple scales, Qwen3-14B (Yang et al., 2025), Phi-3.5-mini-Instruct (Abdin et al., 2024), and InternLM3-8B-Instruct (Cai et al., 2024).[2] *QwenSameText* and *QwenGenerated* use Qwen2.5-7B-Instruct-generated traces re-encoded by the target model, whereas *SelfGenerated* uses InternLM3-generated traces. Identical reasoning text need not produce identical token sequences when replayed through different model families. We therefore extract a separate token indexed trajectory for each target checkpoint and compare EP, HAP, and LDIR within that checkpoint, without assuming cross-model token alignment. Further dataset and trajectory-collection details are given in Appendix C.

Geometries and prediction. Token-level final layer hidden states are mapped to the 64 dimensional GeoVAE latent space. We compare the Euclidean Pullback baseline (EP), the Hidden Space Anisotropic Pullback baseline (HAP), and the proposed Local Direction Informed Riemannian metric (LDIR). EP pulls back the Euclidean metric of standardized hidden space, HAP introduces directional anisotropy in hidden space before pullback, and LDIR applies local direction-informed conditioning after the Euclidean pullback. The same 11 trajectory features and downstream logistic-regression classifier are used for all three geometries. Implementation details are provided in Appendices A and B.

Evaluation. We use five fold cross validation grouped by problem and report results over five seeds. All learned components are fitted on training trajectories only. Main LDIR versus EP comparisons use paired bootstrap inference, with multiple-comparison correction where reported. Full evaluation and statistical details are provided in Appendix C.

4.2 VALIDITY PREDICTION ACROSS MODEL SETTINGS (RQ1)

We compare EP, HAP, and LDIR under identical trajectories, partitions, features, and classifier settings. The evaluation concerns trajectory-level validity discrimination, not changes in the language models' answer accuracy.

LDIR attains the highest mean AUROC in all ten settings (Table 1), with gains over EP ranging from $+0.49$ to $+2.16$ AUROC points. After Holm correction, six of the ten settings remain significant, while HAP stays close to EP throughout. Overall, LDIR shows a consistent numerical advantage, although the gain varies across settings.

[1] The MATH500 dataset card attributes this subset to *Let's Verify Step by Step*: `https://huggingface.co/datasets/HuggingFaceH4/MATH-500`.

[2] The InternLM3 model card recommends the InternLM2 report as its citation. The evaluated checkpoint is `https://huggingface.co/internlm/internlm3-8b-instruct`.

Table 1: Validity prediction across model and trajectory source settings. Values are five-seed mean AUROC ($\times 100$), with seed SD in parentheses. Δ denotes LDIR minus EP and p_{H} the Holm-adjusted value. $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

Setting	EP	HAP	LDIR	Δ	p_{H}
Qwen2.5 base					
Qwen2.5-1.5B	72.33 (1.30)	72.51 (1.28)	**73.74** (1.22)	+1.41 (0.17)	0.0080^{**}
Qwen2.5-3B	71.98 (0.35)	72.07 (0.41)	**73.94** (0.62)	+1.96 (0.30)	0.0014^{**}
Qwen2.5-7B	70.64 (1.02)	70.76 (1.02)	**72.80** (1.11)	+2.16 (0.59)	0.0024^{**}
Qwen2.5 Instruct					
Qwen2.5-1.5B-Instruct	71.38 (1.05)	71.51 (1.08)	**72.46** (1.34)	+1.08 (0.44)	0.0588
Qwen2.5-3B-Instruct	71.58 (0.54)	71.67 (0.51)	**73.08** (0.87)	+1.50 (0.96)	0.0588
Qwen2.5-7B-Instruct	72.72 (0.59)	72.83 (0.61)	**74.33** (0.54)	+1.61 (0.34)	0.0030^{**}
Additional settings					
Qwen3-14B QwenSameText	72.08 (0.87)	72.23 (0.81)	**73.39** (0.68)	+1.31 (0.68)	0.0020^{**}
Phi-3.5-mini-Instruct	73.93 (0.91)	73.92 (0.91)	**74.43** (1.14)	+0.49 (0.35)	0.4774
InternLM3-8B-Instruct QwenGenerated	70.62 (1.07)	70.73 (1.15)	**71.20** (1.25)	+0.58 (0.84)	0.2112
InternLM3-8B-Instruct SelfGenerated	70.13 (1.28)	70.25 (1.24)	**71.91** (0.81)	+1.79 (0.68)	0.0004^{***}

4.3 Understanding the Diagnostic Signal (RQ2)

How does LDIR deform the reference geometry? At prototype k, we define the relative deformation $R_k = G_{\mathrm{EP},k}^{-1/2} G_{\mathrm{LDIR},k} G_{\mathrm{EP},k}^{-1/2}$. For a query location z, the analogous quantity is $R(z) = G_{\mathrm{EP}}(z)^{-1/2} G_{\mathrm{LDIR}}(z) G_{\mathrm{EP}}(z)^{-1/2}$, computed from the interpolated fields. For a unit vector q in an EP-orthonormal frame ($\|q\|_2 = 1$), $q^\top R q$ is the ratio of squared LDIR and EP costs, while $\sqrt{q^\top R q}$ is the corresponding length ratio. For a general nonzero q, the squared-cost ratio is $q^\top R q / (q^\top q)$. Ratios below one indicate contraction relative to EP, and ratios above one indicate expansion.

Figure 2 examines R_k before interpolation in the 64 dimensional latent space. Its eigenvalues μ give LDIR/EP squared cost multipliers: $\log \mu < 0$ indicates contraction and $\log \mu > 0$ expansion. The near zero mean log eigenvalue is an empirical observation for these prototypes. Exact determinant neutrality follows for the unclipped adjustment $\exp(\gamma S_k)$; clipping, interpolation, and the final field ridge do not generally preserve it. Figure 3 in Appendix B.4 evaluates $R(z)$ on held out transitions.

Which measurements benefit from the metric change? We repeat the EP versus LDIR comparison on Qwen2.5-7B-Instruct using restricted feature sets. *Speed* contains the mean, standard deviation, and maximum adjacent-token metric displacement; *Length + Speed* additionally includes total path length; *Turning + Curvature* contains their mean and maximum; *Curvature + Path* combines curvature with straight-path distance, efficiency, and tortuosity; *Path* contains only these three endpoint features; and *Full* uses all 11 features. Because these sets overlap, the analysis compares representations rather than independent feature effects.

How do matched trajectories differ? For feature f, we define the paired difference as $\Delta f_i = f(H_i^-) - f(H_i^+)$. We evaluate the 11 classifier features together with Total Turning and report the four features whose differences have the same sign in both model settings. Total Turning is not included in the classifier input. Qwen2.5 results are averaged over 121 matched pairs. For InternLM3, differences are averaged over questions within each seed and then across five seeds; uncertainty is estimated using 10,000 two-stage bootstrap replicates over seeds and matched questions. Path length and adjacent-token displacement are measured in model-specific LDIR units, so their magnitudes are not compared across models, whereas turning is measured in radians. The reported p_{hier} values are unadjusted for the 12 evaluated measurements.

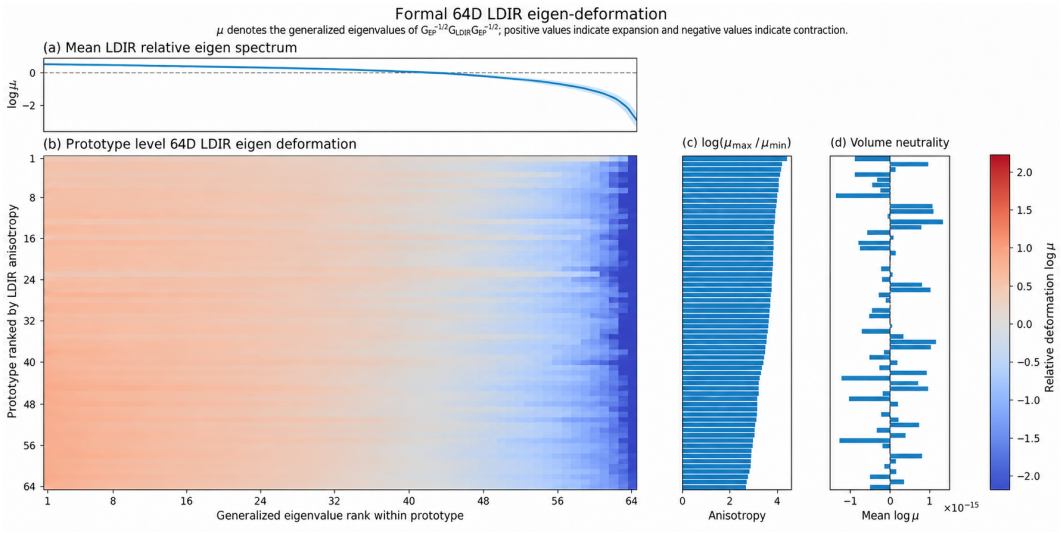

Figure 2: Prototype level spectral diagnostics of the 64 dimensional LDIR deformation relative to EP. Panels show the mean ranked spectrum, spectra across 64 prototypes, anisotropy $\log(\mu_{\max}/\mu_{\min})$, and the empirical mean log eigenvalue. Here μ is an eigenvalue of R_k, equivalently a generalized eigenvalue of $(G_{\mathrm{LDIR},k}, G_{\mathrm{EP},k})$, and hence a squared cost multiplier: $\log \mu > 0$ indicates relative expansion and $\log \mu < 0$ indicates relative contraction; the corresponding log length multiplier is $\frac{1}{2} \log \mu$. The unclipped prototype adjustment has unit determinant; spectral clipping and field construction need not preserve it.

Table 2: Feature-subset analysis on Qwen2.5-7B-Instruct. Values are AUROC ($\times 100$). Len., Spd., Turn., and Cur. denote Length, Speed, Turning, and Curvature, respectively. $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

Subset	EP	LDIR	Δ	p_{H}
Spd.	64.29	65.70	+1.41	0.0276^{*}
Len.+Spd.	66.43	67.42	+1.00	0.2160
Turn. +Cur.	69.33	70.29	+0.96	0.2160
Cur.+Path	73.29	74.23	+0.94	0.2160
Path	68.65	68.69	+0.05	0.9427
Full	72.85	**74.41**	+1.56	0.0500

Table 3: Exploratory LDIR differences (Δ: invalid minus valid; four of 12 measurements). SD and unadjusted p refer to InternLM3 hierarchical bootstrap results. $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

Feature	Q2.5 Δ	IM3 Δ	SD	p
Length	-6235.56	-3196.83	49.50	0.0002^{***}
Mean Spd.	-3.03	-2.69	0.05	0.0002^{***}
Max Spd.	-3.48	-3.55	0.42	0.0002^{***}
Tot. Turn.	-129.65	-70.20	0.21	0.0040^{**}

LDIR improves every listed subset, but Speed is the only restricted subset with an adjusted value below 0.05 (Table 2), while Path changes little. The four displayed measurements are lower for invalid traces in both settings (Table 3), with exploratory InternLM3 values $p_{\mathrm{hier}} \leq 0.0040$. Lower Total Turning does not necessarily imply a smoother path, and these matched associations are neither specific to LDIR nor evidence of causality.

4.4 Token-Count Control (RQ3)

We examine whether a fitted linear association with log-transformed token-transition count accounts for the observed advantage. After residualizing geometric features against this transformed count, LDIR retains the highest mean AUROC on both Qwen2.5-7B-Instruct (68.68 ± 0.92 vs. 66.89 ± 1.14 for EP) and InternLM3 SelfGenerated (71.46 ± 0.86 vs. 69.85 ± 1.21). Detailed procedures are provided in Appendix C.7.

Table 4: Token-count control, reported as AUROC ($\times 100$). Values are five-seed means with sample SD across seeds in parentheses.

Representation	EP	HAP	LDIR
Qwen2.5-7B-Instruct			
Token count only	52.63 (1.16)	52.63 (1.16)	52.63 (1.16)
Full geometry	72.72 (0.59)	72.83 (0.61)	**74.33** (0.54)
Geometry residualized by token count	66.89 (1.14)	66.99 (1.16)	**68.68** (0.92)
InternLM3-8B-Instruct SelfGenerated			
Token count only	55.94 (1.18)	55.94 (1.18)	55.94 (1.18)
Full geometry	70.13 (1.28)	70.25 (1.24)	**71.91** (0.81)
Geometry residualized by token count	69.85 (1.21)	69.92 (1.21)	**71.46** (0.86)

5 Related Work

Reasoning trajectories. Studies of LLM hidden states measure reasoning dynamics, correctness-related subspaces, and spectral or kinematic signatures (Zhou et al., 2026; Sun et al., 2026; Masoomi et al., 2026); related work examines low-dimensional constraints on inference (Ma et al., 2026). LDIR uses local transition directions to define the metric under which trajectory features are measured. Since geometric comparisons can depend on generation length (Gjølbye et al., 2026), we also control for token count.

Latent-space geometry. Deep generative models induce latent geometry through their decoders (Arvanitidis et al., 2017; Shao et al., 2018); pulling back Fisher–Rao geometry gives a distinct reference metric (Arvanitidis et al., 2021). GeoFaith uses VAE-based latent geometry to study CoT faithfulness (Lv et al., 2026). LDIR conditions a deterministic Euclidean decoder pullback on local transition axes, and its labels measure final-answer correctness rather than explanation faithfulness.

Validity signals and controls. Representation directions and hidden-state probes can reveal information about truth or reasoning correctness (Marks & Tegmark, 2023; Zhang et al., 2025). Interpreting such predictive signals requires controls (Hewitt & Liang, 2019; Sahoo et al., 2026). Our classifier uses geometric trajectory summaries; its performance alone does not establish that the model uses these summaries while reasoning. Additional connections to representations, interventions, discrete dynamics, and verification appear in Appendix D.

6 Conclusion

We study how local transition structure can improve the geometric diagnosis of LLM reasoning through LDIR. Rather than using reasoning dynamics only as quantities to be measured, LDIR incorporates their directional organization through post-pullback flow conditioning, refining decoder induced geometry without using validity labels during metric estimation. On controlled matched reasoning pairs, LDIR achieves the highest mean AUROC among the compared geometries across the evaluated models and generation settings. The clearest feature subset improvement appears in speed-related measurements, and LDIR retains the highest mean AUROC after residualizing features against log-transformed token transition count. An exploratory matched analysis identifies shorter metric paths, smaller adjacent token displacements, and less total turning for invalid trajectories in both examined settings. Together, these findings suggest that local transition structure provides a useful reference for measuring reasoning trajectories and revealing differences associated with benchmark final answer correctness. Future work will examine how these findings extend to naturally occurring reasoning errors, broader reasoning domains, and alternative latent representations.

AI USE STATEMENT

In this work, we used generative AI tools for polishing the paper and preliminary coding. We have not used generative AI tools for generating the research idea and designing the algorithm. We have reviewed all AI assisted work. For example, the polished text. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

We confirm that all authors of this submission have read and agree to abide by the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

The code is provided in the supplementary material to enable the reproduction of our results.

REFERENCES

Marah I Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, HS Behl, et al. Phi-3 technical report: a highly capable language model locally on your phone. corr abs/2404.14219 (2024), 2024.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. Latent space oddity: on the curvature of deep generative models. *arXiv preprint arXiv:1710.11379*, 2017.

Georgios Arvanitidis, Miguel González-Duque, Alison Pouplin, Dimitris Kalatzis, and Søren Hauberg. Pulling back information geometry. *arXiv preprint arXiv:2106.05367*, 2021.

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*, pp. 967–976, 2023.

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

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*, 2024.

Iván Vicente Moreno Cencerrado, Arnau Padrés Masdemont, Anton Gonzalvez Hawthorne, David Demitri Africa, and Lorenzo Pacchiardi. No answer needed: Predicting llm answer accuracy from question-only linear probes. *arXiv preprint arXiv:2509.10625*, 2025.

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 *International Conference on Learning Representations*, volume 2024, pp. 3056–3076, 2024.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems, 2021. *URL https://arxiv. org/abs/2110.14168*, 9, 2021.

Anders Gjølbye, Lars Kai Hansen, and Sanmi Koyejo. Reasoning models don't just think longer, they move differently. *arXiv preprint arXiv:2605.15454*, 2026.

Jinwen He, Yujia Gong, Zijin Lin, Cheng'an Wei, Yue Zhao, and Kai Chen. Llm factoscope: Uncovering llms' factual discernment through measuring inner states. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 10218–10230, 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. *arXiv preprint arXiv:2103.03874*, 2021.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. 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)*, pp. 2733–2743, 2019.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning, july 2023. *URL http://arxiv. org/abs/2307.13702*, 2023.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in neural information processing systems*, 36:41451–41530, 2023.

Jie Liang, Zhengxin Yu, Hamid Nasiri, and Peter Garraghan. Geometry of divergence: Tracking hidden-state trajectories for adaptive multi-turn reasoning. *arXiv preprint arXiv:2608.30650*, 2026.

Zhan Ling, Yunhao Fang, Xuanlin Li, Zhiao Huang, Mingu Lee, Roland Memisevic, and Hao Su. Deductive verification of chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 36:36407–36433, 2023.

Weijiang Lv, Wentong Zhao, Jiayu Wang, Yuhao Wu, Jiaheng Wei, and Xiaobo Xia. Geofaith: A spatio-temporal dual view of faithful chain-of-thought. *arXiv preprint arXiv:2605.26893*, 2026.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. Faithful chain-of-thought reasoning. In *Proceedings of the 13th international joint conference on natural language processing and the 3rd conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long papers)*, pp. 305–329, 2023.

Yanbiao Ma, Fei Luo, Linfeng Zhang, Chuangxin Zhao, Mingxuan Wang, Yinan Wu, Zhe Qian, Yang Lu, Long Chen, Zhao Cao, et al. Reasoning emerges from constrained inference manifolds in large language models. *arXiv preprint arXiv:2605.08142*, 2026.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2023.

Aria Masoomi, Mahsa Bazzaz, Adel Javanmard, and Vahab Mirrokni. Geometric signatures of reasoning: A spectral perspective on task hardness. *arXiv preprint arXiv:2607.01571*, 2026.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

A Yang Qwen, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengpeng Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, pp. 2412–15115, 2025.

Subramanyam Sahoo, Vinija Jain, Aman Chadha, and Divya Chaudhary. Linear probes detect task format, not reasoning mode in language model hidden states. In *Proceedings of the 6th Workshop on Trustworthy NLP (TrustNLP 2026)*, pp. 227–239, 2026.

Hang Shao, Abhishek Kumar, and P Thomas Fletcher. The riemannian geometry of deep generative models. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 428–4288. IEEE, 2018.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*, 2022.

Niklas Stoehr, Kevin Du, Vésteinn Snæbjarnarson, Robert West, Ryan Cotterell, and Aaron Schein. Activation scaling for steering and interpreting language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 8189–8200, 2024.

Lihao Sun, Hang Dong, Bo Qiao, Qingwei Lin, Dongmei Zhang, and Saravan Rajmohan. Llm reasoning as trajectories: Step-specific representation geometry and correctness signals. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26872–26887, 2026.

William Timkey and Marten Van Schijndel. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4527–4546, 2021.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965, 2023.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

Xuezhi Wang and Denny Zhou. Chain-of-thought reasoning without prompting. *Advances in Neural Information Processing Systems*, 37:66383–66409, 2024.

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. *arXiv preprint arXiv:2203.11171*, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Sheldon Yu, Yuxin Xiong, Junda Wu, Xintong Li, Tong Yu, Xiang Chen, Ritwik Sinha, Jingbo Shang, and Julian J McAuley. Explainable chain-of-thought reasoning: An empirical analysis on state-aware reasoning dynamics. In *EMNLP (Findings)*, pp. 16660–16667, 2025.

Anqi Zhang, Yulin Chen, Jane Pan, Chen Zhao, Aurojit Panda, Jinyang Li, and He He. Reasoning models know when they're right: Probing hidden states for self-verification. *arXiv preprint arXiv:2504.05419*, 2025.

Jason Zhang and Scott Viteri. Uncovering latent chain of thought vectors in language models. *arXiv preprint arXiv:2409.14026*, 2024.

Yufa Zhou, Yixiao Wang, Xunjian Yin, Shuyan Zhou, and Anru Zhang. The geometry of reasoning: Flowing logics in representation space. In *International Conference on Learning Representations*, volume 2026, pp. 72711–72739, 2026.

Derui Zhu, Dingfan Chen, Qing Li, Zongxiong Chen, Lei Ma, Jens Grossklags, and Mario Fritz. Pollmgraph: Unraveling hallucinations in large language models via state transition dynamics. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 4737–4751, 2024.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Additional Methodological Details

Appendix A details representation learning and latent-dimension selection. Appendix B specifies the metric construction and trajectory features, and Appendix C gives the data and evaluation protocol. Appendix D expands the related-work discussion.

A.1 GeoVAE Architecture and Trajectory Encoding

Let $h \in \mathbb{R}^D$ denote a hidden representation and $\widetilde{h}$ its standardized form. We standardize each hidden dimension using statistics computed from the training fold and apply the same transformation to the test fold. Each token-level hidden state is encoded independently, and temporal structure is captured later through differences between adjacent latent states.

The encoder represents a standardized hidden state by a diagonal Gaussian in latent space. Its mean gives the latent location, while its scale controls the stochastic variation used during training; distributional regularization is imposed by the KL term below:

$$q_\phi(z \mid \widetilde{h}) = \mathcal{N}\Big(\mu_\phi(\widetilde{h}), \operatorname{diag}\Big(\sigma_\phi^2(\widetilde{h})\Big)\Big), \tag{7}$$

and latent samples are obtained during training through

$$z = \mu_\phi(\widetilde{h}) + \sigma_\phi(\widetilde{h}) \odot \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I_d). \tag{8}$$

The reparameterization separates the random draw ε from the encoder outputs and permits gradient based optimization.

The encoder compresses the hidden state in two stages, and the decoder mirrors this map back to standardized hidden space:

$$D \longrightarrow 512 \longrightarrow 128 \longrightarrow d \longrightarrow 128 \longrightarrow 512 \longrightarrow D. \tag{9}$$

The main experiments use $d = 64$. Because stochastic samples would assign slightly different paths to the same trace, geometric analysis uses the deterministic posterior mean

$$z_t = \mu_\phi(\widetilde{h}_t) \tag{10}$$

rather than a sampled latent state.

A.2 GeoVAE Training Objective

The GeoVAE objective balances reconstruction fidelity, latent regularity, and preservation of relative hidden-state distances:

$$\mathcal{L}_{\text{GeoVAE}} = \mathcal{L}_{\text{rec}} + \beta_e \mathcal{L}_{\text{KL}} + \lambda_{\text{geo}} \mathcal{L}_{\text{geo}}. \tag{11}$$

For a mini-batch of size B, the reconstruction loss is the mean squared error in standardized hidden space,

$$\mathcal{L}_{\text{rec}} = \frac{1}{BD} \sum_{i=1}^{B} \left\| \widetilde{h}_i - F_\theta(z_i) \right\|_2^2, \tag{12}$$

where F_θ is the decoder. The KL regularization term is

$$\mathcal{L}_{\text{KL}} = -\frac{1}{2Bd} \sum_{i=1}^{B} \sum_{j=1}^{d} \left(1 + \log \sigma_{ij}^2 - \mu_{ij}^2 - \sigma_{ij}^2\right). \tag{13}$$

The KL term is averaged over both the batch and latent dimensions, matching the elementwise mean used in the implementation.

Reconstruction alone does not require the latent coordinates to retain the relative arrangement of hidden states. We therefore sample P pairs (i_p, j_p) within each mini-batch and compare their distances before and after encoding:

$$d_h^{(p)} = \left\| \widetilde{h}_{i_p} - \widetilde{h}_{j_p} \right\|_2, \tag{14}$$

$$d_z^{(p)} = \left\| \mu_\phi(\widetilde{h}_{i_p}) - \mu_\phi(\widetilde{h}_{j_p}) \right\|_2. \tag{15}$$

Table 5: Geometry preservation of GeoVAE under different latent dimensions. Spearman correlations measure geometric fidelity, while PR and effective rank summarize latent-dimension usage.

Dim.	Recon. MSE↓	Pairwise ρ↑	Transition ρ↑	Turning ρ↑	PR	Eff. Rank	d_{95}	d_{99}
8	0.4435	0.6063	0.5495	0.6172	7.97	7.98	8	8
16	0.4091	0.7128	0.6837	0.7208	15.53	15.76	15	16
20	0.3984	0.7537	0.7104	0.7415	18.65	19.26	19	20
32	0.3804	0.8059	0.7702	0.8012	27.88	29.62	29	31
48	0.3643	0.8559	0.8316	0.8488	38.33	42.64	42	47
64	**0.3521**	**0.8817**	**0.8493**	**0.8643**	47.42	54.88	56	62

Let m_h and m_z be the respective median distances over the sampled pairs. We then use

$$\mathcal{L}_{\text{geo}} = \frac{1}{P}\sum_{p=1}^{P} \text{SmoothL1}\left(\log \frac{d_h^{(p)} + \epsilon_d}{m_h + \epsilon_d}, \log \frac{d_z^{(p)} + \epsilon_d}{m_z + \epsilon_d}\right), \tag{16}$$

where $\epsilon_d > 0$ prevents numerical instability. Median normalization removes global scale differences, while the logarithm reduces the influence of unusually large pairwise distances. The term therefore constrains relative geometry without forcing the two spaces to share an absolute distance scale.

We set $\lambda_{\text{geo}} = 0.5$ and linearly increase β_e to $\beta_{\max} = 10^{-3}$ during the first 20 epochs. Training uses AdamW for 100 epochs with a learning rate of 2×10^{-4}, batch size 256, weight decay 10^{-5}. The default run-level base seed is 42; token-state subsampling uses the base seed plus the fold index. At most 120,000 training token states are used within each fold, and up to 512 randomly selected pairs are used to evaluate $\mathcal{L}_{\text{geo}}$ for each mini-batch. Gradient norms are clipped at 5, and the checkpoint with the lowest epoch average training objective is retained. Neither validity labels nor the LDIR metric are used to train the GeoVAE.

A.3 Latent-Dimension Analysis

We select the latent dimension by measuring geometric fidelity without using validity labels. Pairwise ρ is the Spearman correlation between hidden space and latent space pairwise distances. Transition ρ and Turning ρ are the corresponding correlations for adjacent-token displacement lengths and turning angles. PR and effective rank summarize the latent covariance spectrum. If λ_j are its eigenvalues and $p_j = \lambda_j / \sum_r \lambda_r$, then

$$\text{PR} = \frac{(\sum_j \lambda_j)^2}{\sum_j \lambda_j^2}, \qquad r_{\text{eff}} = \exp\left(-\sum_j p_j \log p_j\right). \tag{17}$$

These quantities measure how broadly variance is distributed across latent coordinates. Finally, d_{95} and d_{99} are the minimum numbers of principal components needed to explain 95% and 99% of the latent variance.

Reconstruction and all three preservation correlations improve over the evaluated range. We therefore use $d = 64$, which gives the strongest measured preservation while reducing the primary hidden state dimension from 3584 to 64. This choice is made before fitting the validity classifier.

B Additional Geometry Details

The metric is constructed in four stages. The decoder first defines a reference geometry; training transitions then determine local prototypes; directional statistics deform the reference metric at each prototype; and radial basis interpolation produces a continuous field. The flow statistics are computed relative to this decoder induced reference geometry.

B.1 Decoder-Induced Pullback Metric

Equal coordinate displacements in latent space need not produce equal changes after decoding. Let $F_\theta : \mathbb{R}^d \to \mathbb{R}^D$ be the trained decoder and

$$J_F(z) = \frac{\partial F_\theta(z)}{\partial z} \in \mathbb{R}^{D\times d} \tag{18}$$

its Jacobian. Write $\widetilde{h}_{\mathrm{dec}}(z) = F_\theta(z)$ for the decoded state; it need not equal an observed standardized hidden state. Its differential for an infinitesimal latent displacement is

$$\mathrm{d}\widetilde{h}_{\mathrm{dec}} = J_F(z)\,\mathrm{d}z. \tag{19}$$

Consequently, its squared Euclidean length in standardized hidden space is

$$\begin{aligned} \|\mathrm{d}\widetilde{h}_{\mathrm{dec}}\|_2^2 &= \mathrm{d}z^\top J_F(z)^\top J_F(z)\,\mathrm{d}z \\ &= \mathrm{d}z^\top G_{\mathrm{EP}}^{\mathrm{pb}}(z)\,\mathrm{d}z. \end{aligned} \tag{20}$$

where $G_{\mathrm{EP}}^{\mathrm{pb}}(z) = J_F(z)^\top J_F(z)$ is the unregularized pointwise pullback. It corrects for local coordinate distortion but does not use observed trajectory directions. In the implementation, decoder Jacobians are evaluated at the prototypes introduced below. The reference matrix at prototype c_k is

$$G_{\mathrm{EP},k} = G_{\mathrm{EP}}^{\mathrm{pb}}(c_k) + \epsilon_{\mathrm{proto}} I_d = J_F(c_k)^\top J_F(c_k) + \epsilon_{\mathrm{proto}} I_d, \qquad \epsilon_{\mathrm{proto}} = 10^{-4}. \tag{21}$$

The prototype level ridge ensures positive definiteness when the decoder Jacobian is locally rank deficient.

B.2 Prototype Construction and Local Neighborhoods

Each adjacent token transition is represented by its displacement and midpoint,

$$\Delta z_t = z_{t+1} - z_t, \qquad m_t = \frac{z_t + z_{t+1}}{2}. \tag{22}$$

For each trajectory, we compute adjacent latent differences and their midpoints independently, so transitions never cross trajectory boundaries. The resulting midpoints are pooled over the training fold and clustered with K-means. We use the cluster centers $\{c_k\}_{k=1}^K$ as local anchors in latent space. We set $K = 64$ with 10 initializations, using the run-level seed plus the fold index as the K-means random state.

Let $a(t)$ denote the nearest prototype assigned to transition t. The local neighborhood of c_k is

$$\mathcal{N}_k = \{t : a(t) = k\}. \tag{23}$$

Hard assignments are used only to estimate one local directional statistic per prototype. Evaluation uses soft radial-basis weights, avoiding jumps at cluster boundaries. Their bandwidth is set by the typical assignment radius,

$$\sigma^2 = \max\Big\{ \big(\mathrm{median}_t \, \|m_t - c_{a(t)}\|_2\big)^2, 10^{-4} \Big\}. \tag{24}$$

The floor prevents a degenerate bandwidth when assignment distances are very small.

At an arbitrary location z, normalized radial basis weights are

$$w_k(z) = \frac{\exp\big(-\|z - c_k\|_2^2/(2\sigma^2)\big)}{\sum_{j=1}^K \exp(-\|z - c_j\|_2^2/(2\sigma^2))}. \tag{25}$$

These weights satisfy $w_k(z) \geq 0$ and $\sum_k w_k(z) = 1$, so nearby prototypes contribute most while every query receives a normalized metric.

B.3 Estimation of the LDIR Metric

At prototype c_k, each local displacement is first expressed in coordinates whose Euclidean norm equals its EP length:

$$y_t = G_{\mathrm{EP},k}^{1/2}\Delta z_t, \qquad \widehat{y}_t = \frac{y_t}{\max\{\|y_t\|_2, \epsilon_y\}}, \qquad \epsilon_y > 0. \tag{26}$$

Indeed, $\|y_t\|_2^2 = \Delta z_t^\top G_{\mathrm{EP},k}\Delta z_t$. Normalizing y_t removes transition magnitude, so the following estimate describes directional occupancy rather than token-to-token speed. The numerical floor defines the expression even when $y_t = 0$; directions with $\|y_t\|_2 \geq \epsilon_y$ have unit norm.

For $|\mathcal{N}_k| \geq 5$, the uncentered directional second moment and its regularized version are

$$\begin{aligned} C_k &= \frac{1}{|\mathcal{N}_k|}\sum_{t\in\mathcal{N}_k}\widehat{y}_t\widehat{y}_t^\top, \\ C_k' &= (1-\lambda_{\mathrm{sh}})C_k + \lambda_{\mathrm{sh}}\frac{I_d}{d} + \epsilon_C I_d, \end{aligned} \tag{27}$$

where $\lambda_{\mathrm{sh}} = 0.15$ and $\epsilon_C = 10^{-4}$. The prime in C_k' denotes shrinkage and ridge regularization. The eigenvectors of C_k identify local axes of motion, and a larger eigenvalue indicates that more transitions align with that axis. Because $\widehat{y}\widehat{y}^\top = (-\widehat{y})(-\widehat{y})^\top$, this statistic is sign-invariant: LDIR learns an unoriented local axis, not a distinction between forward and reverse motion. Shrinkage limits unstable estimates in weakly supported directions.

We invert the second moment so that frequently occupied axes receive lower relative cost, and normalize the precision by its mean eigenvalue:

$$\widetilde{P}_k = \frac{(C_k')^{-1}}{\mathrm{tr}((C_k')^{-1})/d}. \tag{28}$$

We then remove its isotropic log-scale component,

$$S_k = \log(\widetilde{P}_k) - \frac{\mathrm{tr}(\log(\widetilde{P}_k))}{d} I_d. \tag{29}$$

Since $\mathrm{tr}(S_k) = 0$, the *unclipped* adjustment satisfies $\det(\exp(\gamma S_k)) = 1$. The implemented adjustment spectrally clips the eigenvalues of γS_k to $[-10, 10]$. More precisely, if $S_k = U_k\,\mathrm{diag}(s_{k,1},\ldots,s_{k,d})U_k^\top$, let

$$B_k(\gamma) = U_k\,\mathrm{diag}\big(\mathrm{clip}(\gamma s_{k,j}, -10, 10)\big)_{j=1}^{d} U_k^\top, \tag{30}$$

where $\mathrm{clip}(x, -10, 10) = \min(10, \max(-10, x))$. Mapping this implemented adjustment back from the EP-orthonormal frame gives

$$G_{\mathrm{LDIR},k} = G_{\mathrm{EP},k}^{1/2}\exp(B_k(\gamma))G_{\mathrm{EP},k}^{1/2}, \qquad \gamma = 0.75. \tag{31}$$

If q_j is an eigenvector of S_k with eigenvalue $s_{k,j}$, then in the EP-orthonormal frame its implemented LDIR/EP squared-cost multiplier is $\exp(\mathrm{clip}(\gamma s_{k,j}, -10, 10))$ and its length multiplier is its square root. When clipping is inactive, these reduce to $\exp(\gamma s_{k,j})$ and $\exp(\gamma s_{k,j}/2)$, respectively. In general $\mathrm{tr}(B_k(\gamma))$ need not vanish, so the implemented prototype determinant ratio is $\det G_{\mathrm{LDIR},k}/\det G_{\mathrm{EP},k} = \exp(\mathrm{tr}(B_k(\gamma)))$, which need not equal one. Neither interpolation nor the final field ridge preserves this determinant ratio in general. If $|\mathcal{N}_k| < 5$, the local statistic is not estimated and we set $G_{\mathrm{LDIR},k} = G_{\mathrm{EP},k}$.

All three geometries use the same interpolation rule. For $M \in \{\mathrm{EP}, \mathrm{HAP}, \mathrm{LDIR}\}$, the continuous field is

$$G_M(z) = \sum_{k=1}^{K} w_k(z)G_{M,k} + \epsilon_{\mathrm{field}} I_d, \qquad \epsilon_{\mathrm{field}} = 10^{-4}. \tag{32}$$

The field-level ridge is added after interpolation and is distinct from the prototype-level ridge in Eq. (21). Matrix exponentiation makes each LDIR prototype matrix positive definite; the nonnegative weighted sum and final ridge preserve that property. When $\gamma = 0$, $G_{\mathrm{LDIR},k} = G_{\mathrm{EP},k}$ at every prototype, so the shared interpolation recovers exactly the implemented EP field. All matrix operations use symmetrized inputs; eigenvalues used for square roots and logarithms are floored at 10^{-8}.

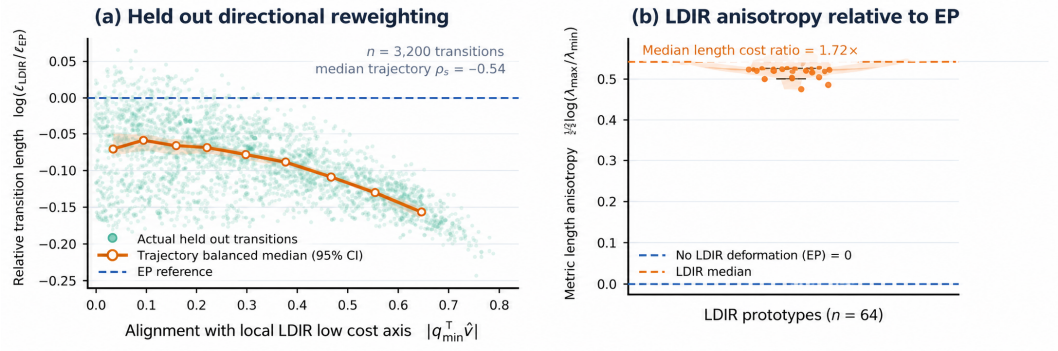

Figure 3: LDIR directional reweighting on held out Fold 1. (a) Greater alignment with the local LDIR low cost axis is associated with a smaller LDIR to EP transition length ratio. (b) Across the 64 prototype locations, the median ratio between the largest and smallest LDIR directional length costs is 1.72×. All quantities are computed in the original 64 dimensional latent space without two dimensional projection.

B.4 Empirical Diagnostics of the LDIR Deformation

Using $R(z)$ defined in Section 4.3, Figure 3 evaluates LDIR directional reweighting on held out transitions.

B.5 Compared Geometries

We compare three metric constructions using identical trajectories, training and test partitions, feature definitions, interpolation weights, and downstream classifiers. EP denotes the Euclidean Pullback baseline, HAP denotes the Hidden-Space Anisotropic Pullback baseline, and LDIR denotes the proposed Local Direction Informed Riemannian metric. EP uses the Euclidean metric in standardized hidden space before pullback, HAP adds a rank-one hidden space anisotropy before pullback, and LDIR estimates a full directional spectrum after the Euclidean pullback.

Euclidean Pullback baseline (EP). EP uses the identity metric in standardized hidden space. Its prototype matrices are the decoder pullbacks $G_{\mathrm{EP},k}$ from Eq. (21), combined by the shared field in Eq. (32). EP therefore accounts for decoder distortion but adds no trajectory-dependent directional reweighting.

Hidden Space Anisotropic Pullback baseline (HAP). Let $\Delta\widetilde{h}_t$ be an adjacent-token transition in standardized hidden space. With a positive numerical floor ϵ_{dir}, define

$$\widehat{d}_t = \frac{\Delta\widetilde{h}_t}{\max\{\|\Delta\widetilde{h}_t\|_2, \epsilon_{\mathrm{dir}}\}}. \tag{33}$$

For prototypes with $|\mathcal{N}_k| \geq 5$, let m_k be the mean of the normalized assigned transitions, and define u_k when that mean is nonzero above the numerical threshold:

$$m_k = \frac{1}{|\mathcal{N}_k|}\sum_{t\in\mathcal{N}_k}\widehat{d}_t, \qquad u_k = \frac{m_k}{\|m_k\|_2} \quad \text{if } \|m_k\|_2 > \epsilon_u, \tag{34}$$

where $\epsilon_u > 0$ is the numerical threshold for a negligible mean direction. Under this condition, the hidden space anisotropic metric is

$$M_{\mathrm{HAP},k} = bI_D + (a-b)u_k u_k^\top, \qquad a = 0.35, \quad b = 1. \tag{35}$$

Because $a < b$, motion parallel to u_k has a lower metric cost than motion orthogonal to it. Pulling this metric back through the decoder gives

$$G_{\mathrm{HAP},k} = J_F(c_k)^\top M_{\mathrm{HAP},k} J_F(c_k) + \epsilon_{\mathrm{proto}} I_d. \tag{36}$$

If $|\mathcal{N}_k| < 5$, or if the defined m_k satisfies $\|m_k\|_2 \le \epsilon_u$, set $G_{\mathrm{HAP},k} = G_{\mathrm{EP},k}$ without defining u_k. Otherwise, the rank-one form assigns cost a along u_k and cost b in every orthogonal hidden-space direction. The implementation evaluates this expression without explicitly constructing the $D \times D$ matrix:

$$J_F^\top M_{\mathrm{HAP},k} J_F = bJ_F^\top J_F + (a-b)(J_F^\top u_k)(J_F^\top u_k)^\top. \tag{37}$$

Local Direction-Informed Riemannian metric (LDIR). LDIR first establishes G_{EP} and only then estimates directional structure in pullback-orthonormal coordinates, as specified in Eqs. (26)–(32). Unlike HAP's single mean-flow axis, LDIR uses a sign-invariant second moment and can reweight multiple local axes. This comparison asks whether post pullback multi-axis conditioning provides information beyond decoder only geometry and the simpler pre-pullback rank-one control.

B.6 Complete Trajectory Feature Definitions

Here t indexes token positions, not semantic reasoning steps. For a latent trajectory $(z_1, \ldots, z_T)$, each feature is computed from adjacent-token transitions under a chosen field G. The same definitions are used for EP, HAP, and LDIR.

Metric displacement. Path length measures accumulated movement, whereas local displacement statistics describe how that movement is distributed across token transitions. The length of transition t is evaluated at its midpoint:

$$\ell_t^G = \sqrt{\Delta z_t^\top G(m_t) \Delta z_t}. \tag{38}$$

The implementation floors each quadratic form at 10^{-8} before taking its square root. The total path length is

$$L_G = \sum_{t=1}^{T-1} \ell_t^G. \tag{39}$$

Treating one token transition as one discrete time unit, ℓ_t^G is also the token indexed metric speed. It is not wall-clock generation speed. We retain the mean, standard deviation, and maximum of $\{\ell_t^G\}$ in addition to L_G.

Turning and curvature. Turning measures directional change independently of displacement magnitude. For two consecutive transitions

$$u_t = z_t - z_{t-1}, \qquad v_t = z_{t+1} - z_t, \tag{40}$$

the metric angle at their shared state is

$$\theta_t^G = \arccos\left(\frac{u_t^\top G(z_t) v_t}{\sqrt{\left(u_t^\top G(z_t) u_t\right)\left(v_t^\top G(z_t) v_t\right)}} \right). \tag{41}$$

An angle near zero indicates locally aligned motion, whereas a larger angle indicates a sharper change of direction. Curvature additionally divides this change by the local distance traveled. Both adjacent norms are evaluated in the same state metric $G(z_t)$:

$$s_t^G = \frac{1}{2}\left[\sqrt{u_t^\top G(z_t) u_t} + \sqrt{v_t^\top G(z_t) v_t} \right], \qquad \kappa_t^G = \frac{\theta_t^G}{s_t^G}. \tag{42}$$

The cosine argument is clipped to $[-1, 1]$, and norm denominators use a small positive floor. We retain the mean and maximum of both θ_t^G and κ_t^G.

Endpoint-relative structure. The preceding quantities are local. To compare the observed path with a direct connection between the same endpoints, let

$$\gamma_{\mathrm{str}}(s) = z_1 + s(z_T - z_1), \qquad s \in [0, 1], \tag{43}$$

be the straight coordinate path joining the endpoints. Its metric length is

$$L_{\mathrm{str}}^G = \int_0^1 \sqrt{(z_T - z_1)^\top G(\gamma_{\mathrm{str}}(s))(z_T - z_1)}\, \mathrm{d}s. \tag{44}$$

We use 32 equally spaced points on the coordinate path and evaluate the metric at the midpoints of the resulting 31 intervals. The endpoint efficiency and tortuosity proxies are

$$\eta_G = \frac{L^G_{\mathrm{str}}}{L_G}, \qquad \tau_G = \frac{L_G}{L^G_{\mathrm{str}}}. \tag{45}$$

Because γ_{str} is straight only in latent coordinates, L^G_{str} is not assumed to equal the true geodesic distance. Accordingly, η_G and τ_G are interpreted as endpoint-relative proxies rather than exact geodesic quantities. Ratio denominators use the same positive numerical floor as the local norms.

The resulting 11 features have four roles: accumulated and local movement (L_G and three speed summaries), directional change (two turning summaries), turning per unit distance (two curvature summaries), and endpoint-relative structure (L^G_{str}, η_G, and τ_G). In fixed order, the feature vector is

$$\begin{aligned} x_G = [&L_G, \mathrm{mean}(\ell^G), \mathrm{std}(\ell^G), \max(\ell^G), \\ &\mathrm{mean}(\theta^G), \max(\theta^G), \mathrm{mean}(\kappa^G), \max(\kappa^G), \\ &L^G_{\mathrm{str}}, \eta_G, \tau_G]. \end{aligned} \tag{46}$$

C Additional Experimental Details

C.1 Construction of Matched Invalid Reasoning

The primary dataset contains 121 MATH500 problems and 242 trajectories. For each trajectory with a benchmark correct final answer, we introduce one localized reasoning error and continue generation with the same model. The interventions cover (1) affirming the consequent; (2) incorrect reference or variable use; (3) skipped intermediate inference; (4) quantifier or scope errors; and (5) invalid inference chains.

A pair is retained only when the intervention changes the final answer from correct to incorrect. The mathematical problem, checkpoint, prompt, and decoding configuration are fixed within each pair, while the intervention position and error type are recorded. Thus, validity is defined by final-answer correctness relative to the benchmark answer rather than by formal verification of every intermediate inference.

This procedure produces matched pairs

$$q_i \longrightarrow H_i^+, \qquad q_i \longrightarrow H_i^-, \tag{47}$$

where the problem q_i is shared and the two trajectories have the same reasoning history up to the localized intervention.

C.2 Trajectory Collection

For each reasoning trace, we extract the ordered sequence

$$H = (h_1, h_2, \ldots, h_T), \qquad h_t \in \mathbb{R}^D, \tag{48}$$

where t indexes token position and D depends on the language model ($D = 3584$ for the primary Qwen2.5-7B setting). Each stored reasoning trace is tokenized and replayed through the representation model, and h_t is the final-layer hidden state at token position t. We do not segment the text into semantic reasoning steps or pool multiple tokens into one state. Token order is retained, so each example is treated as a trajectory rather than an unordered set. The same layer and extraction procedure are used for valid and invalid traces, without using validity labels.

C.3 Validity Classifier

For each geometry $G \in \{\mathrm{EP}, \mathrm{HAP}, \mathrm{LDIR}\}$, the corresponding trajectory vector x_G is standardized using statistics from the training fold and supplied to a logistic regression classifier. The classifier uses an L2 penalty with $C = 1$, balanced class weights, the L-BFGS optimizer, and a maximum of 5000 iterations. No classifier hyperparameter search is performed. Its predicted probability for the positive class is used as the trajectory-level validity score. A separate classifier is fitted for each geometry so that the comparison reflects only the information contained in the corresponding geometric features.

C.4 Cross-Validation and Leakage Prevention

We use five-fold grouped cross-validation with the problem identifier as the grouping variable. Consequently, the valid and invalid trajectories derived from the same problem always remain in the same fold. Within each split, the hidden state standardizer, GeoVAE, prototype bank, local metrics, feature standardizer, and classifier are fitted using the training fold only. The held-out fold is used solely for feature computation and prediction. Validity labels enter only at the classifier training and evaluation stages; they are not used to train the GeoVAE or estimate EP, HAP, or LDIR.

C.5 Model and Trajectory-Source Settings

Qwen2.5-7B-Instruct is the primary representation model. We additionally evaluate Qwen2.5 base and instruction-tuned models at 1.5B and 3B, Qwen2.5-7B, Qwen3-14B, Phi-3.5-mini-Instruct, and InternLM3-8B-Instruct.

For *QwenSameText* and *QwenGenerated*, reasoning traces generated by Qwen2.5-7B-Instruct are replayed through the named target representation model. *SelfGenerated* denotes trajectories generated and represented by InternLM3. Each setting fits its GeoVAE, metric field, feature standardizer, and classifier separately.

C.6 Statistical Reporting

The main comparison is repeated over seeds $\{42, 142, 242, 342, 442\}$. Fold assignments and all training-dependent components are regenerated for each seed. For each seed, AUROC is computed from the concatenated held-out predictions over the five grouped folds.

LDIR versus EP comparisons use a two-sided, 10,000-replicate paired bootstrap. For repeated-seed experiments, the bootstrap resamples seeds and matched problems within each sampled seed. Holm correction is applied to the prespecified comparison family reported in the corresponding table. Auxiliary single-run Qwen analyses use paired resampling over matched problems and are reported separately from the repeated-seed main results.

Exploratory matched-pair differences. For a measurement f, the paired difference is $\Delta f_i = f(H_i^-) - f(H_i^+)$. We examine the 11 classifier features and Total Turning, for 12 measurements in total. Table 3 displays the four whose mean paired differences have the same sign in both model settings. A negative difference means that the measurement is smaller for invalid trajectories. Total Turning is not a classifier feature.

The Qwen2.5 differences are averaged over 121 matched pairs. For InternLM3, differences are averaged over matched questions within each seed and then across five seeds. Seed SD is the standard deviation of the five seed-level differences. The InternLM3 p_{hier} values are based on 10,000 paired hierarchical bootstrap replicates that resample seeds and matched questions. They are exploratory and are not adjusted for the 12 comparisons or the selection of the four displayed measurements.

C.7 Token Count Control Details

This subsection gives the procedure underlying Table 4 in Section 4.4.

Removing linear dependence on token count. Motivated by the sensitivity of trajectory statistics to generation length (Gjølbye et al., 2026), we compare a token count only probe, the full geometry, and token count residualized geometry. For trajectory i, let $n_i = T_i - 1$ be its observed number of token transitions and $r_i = \log(1 + n_i)$. Within each training fold and geometry, we fit

$$x_{ij}^G = \alpha_j^G + \beta_j^G r_i + \varepsilon_{ij}^G$$

separately for each feature j. The training coefficients are used to form residuals for both training and held out trajectories, after which the residuals are standardized using training fold statistics. The count only probe uses r_i. We repeat the analysis on InternLM3 SelfGenerated.

Token count alone has limited predictive value in both settings. After residualization, LDIR retains the highest mean AUROC: 68.68 ± 0.92 versus 66.89 ± 1.14 for EP on Qwen2.5-7B-Instruct, and

Table 6: External validation on the English split of MGSM using 161 matched problem pairs (322 trajectories). AUROC values are five-seed means with sample standard deviation across seeds in parentheses. Δ denotes LDIR minus EP, and p is the two-sided paired hierarchical bootstrap p-value for LDIR versus EP.

Setting	EP	HAP	LDIR	Δ	p
MGSM	96.20 (1.38)	96.21 (1.41)	96.49 (1.39)	+0.29	0.2628

71.46 ± 0.86 versus 69.85 ± 1.21 on InternLM3 SelfGenerated. Residualization removes only the fitted linear association with token count; other length related effects may remain.

C.8 External Validation on MGSM

We additionally evaluate LDIR on the English split of MGSM (Shi et al., 2022). Matched valid–invalid trajectories are constructed independently of the MATH500 data using the same localized perturbation and final-answer filtering procedure described in Appendix C.1. Of 207 candidate problems, 161 yield valid matched pairs, giving 322 trajectories in total. All trajectories are generated and represented by Qwen2.5-7B-Instruct, and all learned components are refitted independently on MGSM.

We use the same five-seed grouped cross-validation and paired hierarchical bootstrap protocol as in Appendix C.6.

As shown in Table 6, LDIR improves mean AUROC from 96.20 for EP to 96.49, a difference of $+0.29$ points ($p = 0.2628$). Relative to HAP, the difference is $+0.28$ points ($p = 0.2384$). Thus, LDIR retains the highest mean AUROC on MGSM, although the improvement is not statistically conclusive.

C.9 Matched-Pair Case Study of Local LDIR Geometry

We further examine one matched valid–invalid pair to illustrate how the LDIR measurements differ after a localized reasoning error. The example is `pair_000031` from MATH500, for the problem

$$(u+4)(u-1)-(u-3)(u+6).$$

The valid continuation subtracts the second expanded expression correctly,

$$(u^2+3u-4)-(u^2+3u-18),$$

and reaches the correct answer 14. In the invalid continuation, the constant term in the second expression is changed from -18 to $+18$,

$$(u^2+3u-4)-(u^2+3u+18),$$

which yields the incorrect answer -22.

For this case, we analyze only the continuation from the first divergence to the first final answer. The shared prefix is excluded, and the later repeated solution in the valid trace is not included. Figure 4 shows the empirical distributions of LDIR token indexed speed and LDIR curvature over this continuation.

The invalid continuation has smaller local movement under LDIR: mean token indexed speed decreases from 59.6605 to 55.1542, a change of -7.55%. The maximum separation between the two empirical cumulative distributions is 0.211. At the same time, mean LDIR curvature increases from 0.0329 to 0.0357, corresponding to a change of $+8.61\%$, with a maximum ECDF gap of 0.252. Thus, in this matched continuation, the localized error is associated with smaller token-to-token movement and larger turning per unit LDIR length.

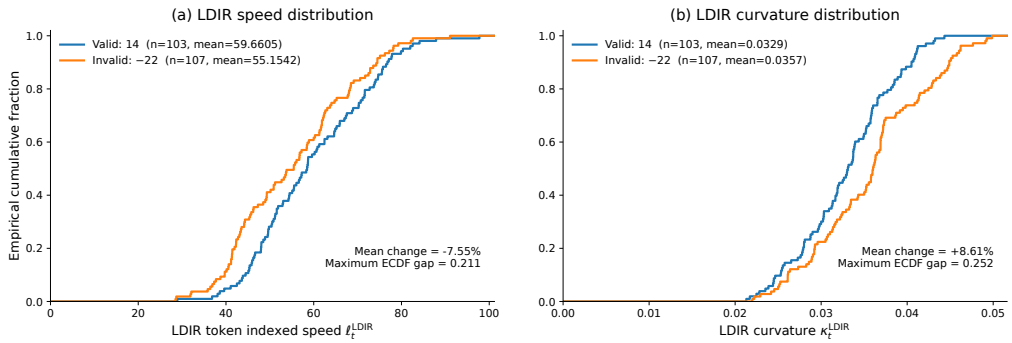

Figure 4: Local LDIR geometry for the matched reasoning pair `pair_000031`, measured from the first divergence to the first final answer. (a) The invalid continuation has a lower distribution of LDIR token indexed speed, with the mean decreasing from 59.6605 to 55.1542 (-7.55%). (b) The invalid continuation has a higher LDIR curvature distribution, with the mean increasing from 0.0329 to 0.0357 ($+8.61\%$). The maximum ECDF gaps are 0.211 for speed and 0.252 for curvature.

D Additional Related Work

Truth representations and interventions. Unsupervised consistency can recover truth-related information (Burns et al., 2022), while predictors trained on internal states assess factuality (Azaria & Mitchell, 2023; He et al., 2024). INSIDE uses the covariance spectrum of representations across multiple generated responses (Chen et al., 2024); its statistic differs from our second moment of adjacent transition directions. Counterfactuals and a causal inner product provide another view of linear representations (Park et al., 2023), and dominant dimensions can distort representation similarities (Timkey & Van Schijndel, 2021). Representation engineering studies activation-space monitoring and control (Zou et al., 2023). Activation shifts (Li et al., 2023; Turner et al., 2023), multiplicative scaling (Stoehr et al., 2024), and CoT steering vectors (Zhang & Viteri, 2024) modify or investigate model behavior. LDIR measures fixed trajectories and does not intervene in model activations.

Other trajectory dynamics. Discrete approaches cluster embeddings of reasoning steps and model their transitions with a Markov chain (Yu et al., 2025), or abstract hidden-state transitions for hallucination detection (Zhu et al., 2024). Temporal curvature and variance slope have also been studied in multi-turn reasoning (Liang et al., 2026). These approaches complement continuous, token-level measurement under LDIR.

Reasoning verification and faithfulness. Self-consistency aggregates answers across sampled reasoning paths (Wang et al., 2022), while alternative decoding can reveal CoT paths without explicit CoT prompting (Wang & Zhou, 2024). Verifiers rank mathematical solutions (Cobbe et al., 2021); process- and outcome-based supervision provide feedback at different stages (Uesato et al., 2022). Other work checks deductions under a restricted context (Ling et al., 2023) or executes symbolic reasoning chains (Lyu et al., 2023). Prompt-bias experiments and interventions show that a generated explanation need not faithfully reflect its influence on the answer (Turpin et al., 2023; Lanham et al., 2023). LDIR predicts benchmark answer correctness from trajectory features without verifying each textual inference or establishing explanation faithfulness.

Additional probing considerations. Linear probes expose information accessible to a separate classifier (Alain & Bengio, 2016). Question-only activations can also predict forthcoming answer correctness, with weaker transfer to mathematical reasoning (Cencerrado et al., 2025). These results reinforce the need to distinguish a predictive diagnostic from a claim about the model's internal reasoning mechanism.