

Soft Tokens Are Off-Manifold: Geometric Correction for Continuous-Space Reasoning in Large Language Models

Jing Gao^{1,†} Weiqiang Jin^{1,3,†} Yunqi Mi² Peigen Ye³ Hongyang Du³
Biao Zhao¹ Yimin Deng^{2,4} Chengxu Liu² Xueming Qian^{1,*} Guoshuai Zhao^{2,*}

¹School of Information and Communications Engineering,
Xi'an Jiaotong University, Xi'an 710049, China

²School of Software Engineering, Xi'an Jiaotong University, Xi'an 710049, China

³Department of Electrical and Computer Engineering, The University of Hong Kong,
Hong Kong SAR 999077, China

⁴City University of Hong Kong, Hong Kong SAR, China

gaojing0423@stu.xjtu.edu.cn weiqiangjin@stu.xjtu.edu.cn miyunqi@stu.xjtu.edu.cn

yepg@hku.hk duhy@hku.hk biao Zhao@xjtu.edu.cn

dymanne@stu.xjtu.edu.cn chengxu.liu@xjtu.edu.cn

qianxm@mail.xjtu.edu.cn guoshuai.zhao@xjtu.edu.cn

[†]Equal contribution. ^{*}Corresponding authors.

Abstract

Continuous-space reasoning methods replace discrete intermediate tokens with probability-weighted soft embeddings to preserve information discarded by greedy Chain-of-Thought decoding. We show this promise is limited by probability concentration and local representation geometry. When next-token distributions are low-entropy and concentrated, probability weighting makes soft tokens nearly collinear with the greedy anchor, a geometric manifestation of collapse toward default decoding. When distributions are high-entropy, diffuse, or stochastically perturbed, continuous states may leave the greedy path but drift beyond the local scale of natural candidate representations. We diagnose these regimes with step-level probes over representation divergence, local norm calibration, and behavior-cloud membership. To address them, we propose *Manifold Thinking*, a training-free inference framework combining structured tangent movement away from the dominant greedy direction with local norm rescaling to the candidate-token neighborhood. Across 18 matched model–benchmark configurations in the primary suite, Manifold Thinking attains the highest mean score in every row; 10 of the 17 non-degenerate paired comparisons are significant at $p < 0.05$, while seven are positive but non-significant trends. Additional LiveCodeBench results obtained with a different backbone set are reported separately in the appendix. These results suggest useful continuous reasoning requires geometric correction rather than unconstrained linear token mixing alone.

1 Introduction

Chain-of-thought (CoT) prompting has become a standard way to elicit multi-step reasoning from large language models (LLMs) (Wei et al., 2022; Wang et al., 2022; Kojima et al., 2022). Yet standard CoT remains token-discrete: at every step, a full probability distribution over possible continuations is collapsed into one sampled or greedy token. This discretization is convenient for text generation, but it retains only one continuation and may discard lower-probability alternatives that soft-token methods aim to preserve (Zhang et al., 2026; Wu et al., 2025).

Continuous Space Reasoning (CSR) attempts to reduce this bottleneck by carrying probability-weighted soft-token embeddings through intermediate reasoning steps (Zhang et al., 2026). By delaying commitment to a single token, CSR can preserve a local mixture of candidate meanings. Yet this benefit rests on a geometric assumption: linear mixtures in the ambient embedding space should remain locally consistent with the model’s representation geometry. Prior work shows that contextualized representations in pretrained language models can be strongly anisotropic (Ethayarajh, 2019). Motivated by this evidence, we test rather than assume whether linear mixtures of candidate token embeddings remain locally compatible with the representation geometry used by the model.

Recent work on Soft Thinking exposes one symptom of this issue. Wu et al. (2025) show that vanilla Soft Thinking often behaves as a single-threaded reasoner: despite being formed from Top-

K candidates, the soft token frequently inherits the greedy component. Stochastic Soft Thinking mitigates this through the Gumbel-Softmax trick (Wu et al., 2025; Jang et al., 2017), making it an important training-free stochastic baseline. We therefore treat stochasticity as a partial answer rather than a full geometric correction. Stochastic perturbation can move a continuous state away from the greedy path, but random departure does not specify whether the new state remains locally natural on the candidate-induced representation manifold. We summarize the closest lines of work in Section 5 and provide an extended discussion in Appendix C.

Motivated by this gap, we study soft-token reasoning geometrically. Our meta-experiments identify two failure regimes. In the low-entropy greedy anchor collapse regime, probability concentration contracts soft tokens toward the greedy anchor, producing the geometric manifestation observed in prior Soft Thinking analyses. In the high-entropy local-scale drift regime, soft or Gumbel-perturbed tokens can depart from the greedy path but lose calibration to the natural scale of the local Top- K candidate cloud. The key requirement is therefore not mere difference from greedy decoding: useful continuous reasoning should remain in a region that the model can interpret as a plausible intermediate state. Our reference-vector analysis probes dominant-direction responses across symbolic domains and anchor choices, showing that the greedy anchor is the most stable empirical proxy for the dominant semantic direction.

To address both regimes, we propose **Manifold Thinking**, a training-free inference framework for geometry-aware continuous reasoning. At each decoding step, Manifold Thinking first uses the Top- K candidates to define a local candidate neighborhood. *Tangent Navigation* then extracts structured orthogonal innovation from sub-dominant candidates relative to the dominant semantic direction, providing non-random movement away from greedy collapse. *Manifold Projection*, implemented as local norm rescaling, finally re-anchors the shifted state to a candidate-derived scale, keeping exploration tied to the empirical geometry of naturally occurring token representations. The method changes only inference-time soft token construction and does not require model fine-tuning. By modifying how frozen LLMs construct intermediate states during language generation, it directly targets NLP questions in reasoning, generation, and inference-time methods.

Our contributions are:

- We diagnose two distinct failure regimes: probability-driven low-entropy greedy anchor collapse, with a geometric manifestation, and genuinely geometric high-entropy local-scale drift. Section 2 introduces step-level probes for representation divergence, local norm calibration, and behavior-cloud membership.
- We introduce Manifold Thinking, combining tangent movement and local norm rescaling to correct continuous reasoning states while preserving a single inference trajectory.
- We evaluate against verified baselines, including greedy decoding, Pure CoT, vanilla Soft Thinking, and Stochastic Soft Thinking, across math and code benchmarks. Results show geometry-aware correction provides benefits beyond uncorrected soft-token mixing or stochastic diversification alone.

2 Meta-Experiments: Diagnosing Continuous-State Failure Regimes

We test when probability-weighted soft tokens form useful continuous states. The study is diagnostic; downstream evaluation appears in Section 4. We ask whether vanilla soft tokens collapse toward Top-1, whether departures stay calibrated, and whether behavior metrics separate aggressive from supported exploration. Unless noted, probes use Qwen2.5-7B (Qwen et al., 2025), Top- $K = 20$, 10 problems, and up to 100 steps per problem.

2.1 Step-wise Dual-Channel Probing

We compare input-space and hidden-response proximity. At step t , the vanilla soft token is

$$\mathbf{e}_{soft}^t = \sum_{k=1}^K p_k^t \mathbf{e}_k^t, \quad (1)$$

where $\mathbf{e}_k^t$ and p_k^t are the Top- K embeddings and normalized probabilities. Let $\mathcal{F}_t(\cdot)$ be the one-step continuation computation. We measure hidden-response divergence from Top-1 by

$$RD_t(\mathbf{v}) = 1 - \frac{\mathcal{F}_t(\mathbf{v}^t) \cdot \mathcal{F}_t(\mathbf{e}_{top1}^t)}{\|\mathcal{F}_t(\mathbf{v}^t)\| \|\mathcal{F}_t(\mathbf{e}_{top1}^t)\|}. \quad (2)$$

RD_t is relative divergence, not semantic correctness.

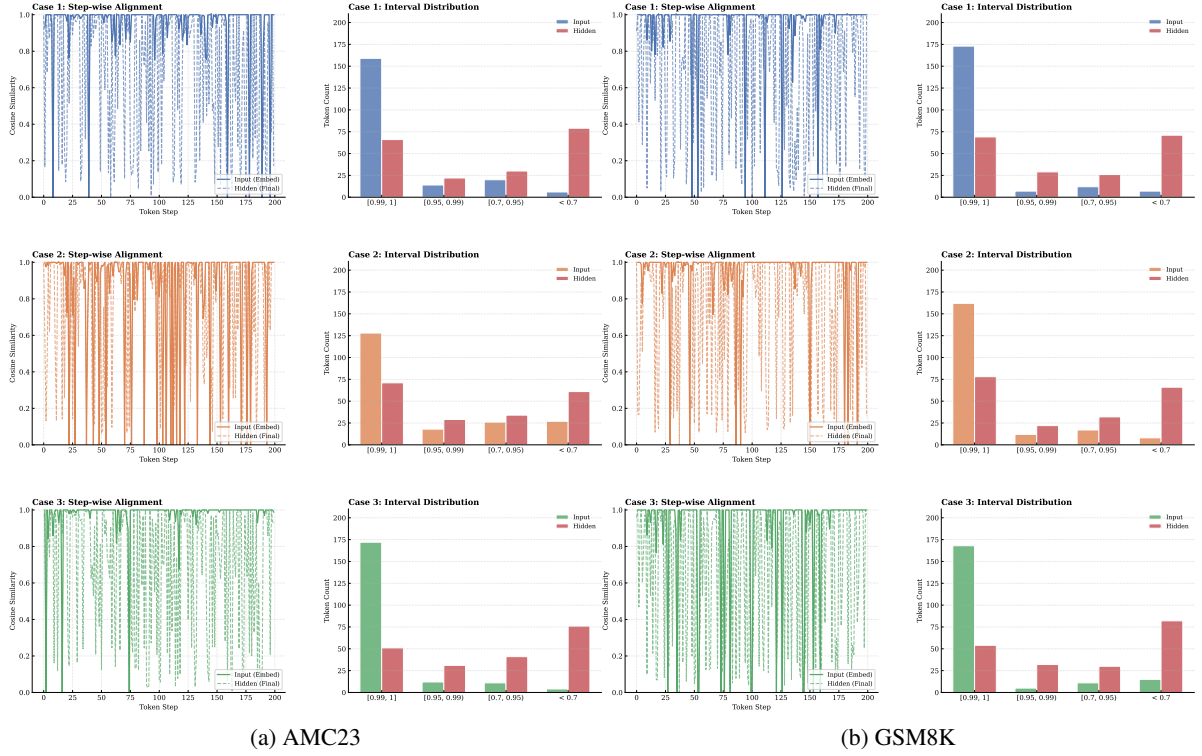


Figure 1: **Step-wise dual-channel probing.** Soft tokens can be nearly collinear with Top-1, especially at low entropy, while hidden responses need not mirror inputs. We separate input-space departure from response divergence.

Figure 1 shows vanilla Soft Thinking often remains single-threaded at low entropy, consistent with prior analyses (Wu et al., 2025). Top-1 distance diagnoses collapse but cannot certify useful exploration.

2.2 Low-Entropy Collapse and High-Entropy Drift

At low entropy, probability concentration drives Top-1 collapse: when p_{top1}^t dominates, $\mathbf{e}_{soft}^t$ becomes nearly collinear with $\mathbf{e}_{top1}^t$, which is the geometric manifestation of that concentration. At high entropy, soft or stochastic tokens may depart from Top-1, but they become poorly norm-calibrated relative to the local Top- K scale. We measure this norm-level drift by

$$Z_{norm}(\mathbf{v}^t) = \frac{||\mathbf{v}^t||_2 - \mu_t}{\sigma_t + \epsilon}, \quad (3)$$

Here μ_t is a scalar—the mean norm of the current Top- K candidate embeddings—and σ_t is the corresponding scalar standard deviation; neither is a soft-token vector. Lower Z_{norm} means better local-scale calibration.

Table 1 supports this diagnosis. At low entropy, continuous variants stay Top-1-like; at high entropy,

Data	Entropy	N	Metric	Soft	Gumbel	Proj.
AMC23	Low	735	Top-1 dist. $\uparrow$	0.000001	0.000386	0.000002
			$Z_{norm} \downarrow$	0.990	1.014	0.994
AMC23	High	99	Top-1 dist. $\uparrow$	0.117006	0.249608	0.153288
			$Z_{norm} \downarrow$	7.995	6.652	0.809
MATH500	Low	714	Top-1 dist. $\uparrow$	0.000001	0.001732	0.000002
			$Z_{norm} \downarrow$	1.011	1.020	1.018
MATH500	High	116	Top-1 dist. $\uparrow$	0.130007	0.280445	0.167099
			$Z_{norm} \downarrow$	9.507	7.976	1.015

Table 1: **Collapse and drift diagnostics.** At high entropy, soft and Gumbel tokens depart from Top-1 but drift in norm. The projected probe preserves departure while re-anchoring to the local Top- K scale.

they depart but become poorly norm-calibrated. Projection gives lower Z_{norm} than soft and Gumbel in 92.93% of high-entropy AMC23 steps and 89.66% of high-entropy MATH500 steps.

2.3 Behavior-Cloud Membership

Geometry alone is insufficient: a vector may have acceptable norm yet induce unnatural next-token behavior. At each step, we compare a synthetic state’s behavior with the local Top- K continuation cloud. We report kernel-density energy E_{KDE} , effective support N_{eff} , and Hellinger departure from Top-1 behavior H_{top1} . Lower E_{KDE} and higher N_{eff} mean stronger cloud membership; higher

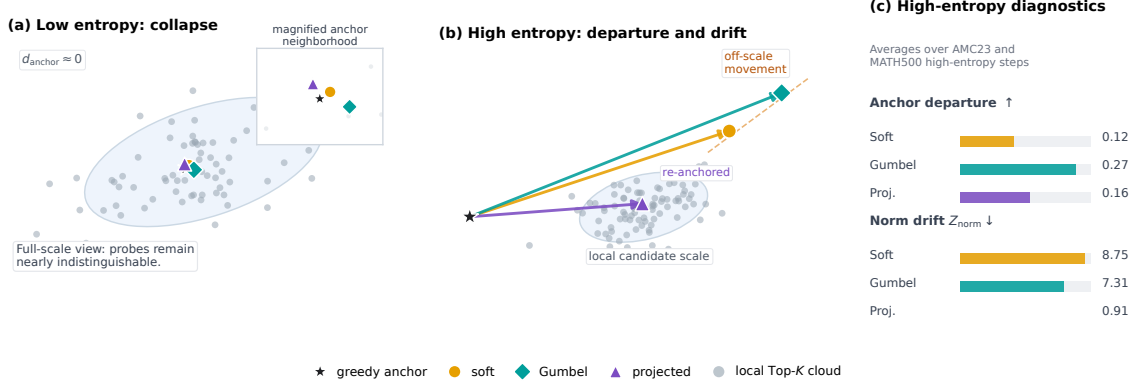


Figure 2: **Soft-token collapse and high-entropy drift.** The representative projection shows that low-entropy probes collapse near the greedy anchor, with the inset magnifying this neighborhood for legibility. At high entropy, soft and stochastic probes move farther from the anchor but become poorly norm-calibrated, while projection re-anchors the state near the local Top- K cloud. The diagnostic bars summarize the same high-entropy pattern with average anchor departure and norm drift.

Data	Method	$E_{KDE} \downarrow$	$N_{\text{eff}} \uparrow$	$H_{\text{top1}} \uparrow$
AMC23	Soft	0.366009	18.901830	0.079935
	Gumbel $\tau_g = 0.5$	0.373411	18.896671	0.336589
	Tangent Probe	0.366852	18.907352	0.152680
	Projected Probe	0.367110	18.905917	0.149506
MATH500	Soft	0.416652	18.817928	0.156263
	Gumbel $\tau_g = 0.5$	0.424191	18.648180	0.358765
	Tangent Probe	0.415682	18.781376	0.227579
	Projected Probe	0.413645	18.794482	0.228459

Table 2: **Behavior-cloud membership in high-entropy steps.** Gumbel explores more aggressively, but without better cloud membership. Projected Probe preserves departure while better supported than Gumbel under E_{KDE} and N_{eff} .

H_{top1} means stronger greedy-path departure. Full definitions are given in Appendix D.

Table 2 separates exploration strength from cloud support. Gumbel has the largest H_{top1} on both datasets, hence the strongest stochastic departure from greedy behavior. The projected probe has lower E_{KDE} and higher N_{eff} than Gumbel on both datasets, so its behavior is better locally supported. These diagnostics motivate the method. They also clarify the role of stochasticity in this paper. Random perturbation is useful as a stress test because it exposes whether leaving the Top-1 path is sufficient, but the behavior-cloud results show that departure alone is an incomplete criterion. A useful continuous state must remain close enough to the local candidate-induced continuation family that the frozen model can still process it as a plausible intermediate state. Accordingly, later experiments use stochastic soft-token reasoning as

a strong comparator rather than treating it as geometric correction. The diagnostic target is not to maximize distance from greedy decoding, but to preserve interpretable local support while allowing non-greedy semantic movement. This criterion links the probes to the later construction. Tangent Navigation is meant to preserve departures that come from actual Top- K candidates rather than arbitrary noise, while Manifold Projection keeps the resulting state within a locally observed representation scale. Thus the diagnostics are not merely descriptive: they specify what a corrected soft token should satisfy at each step. It should depart enough to keep alternatives available, yet remain close enough to the local continuation cloud to stay legible to the frozen model.

3 Methodology

This section introduces *Manifold Thinking*, a training-free framework for geometry-aware continuous-space reasoning. The meta-experiments in Section 2 separate two effects: low-entropy probability concentration manifests geometrically as collapse toward the dominant semantic direction, while high-entropy or perturbed states can genuinely drift from local candidate geometry. Manifold Thinking addresses both cases by treating each step as a local geometry-constrained update rather than unconstrained probability mixing.

Figure 3 places Manifold Thinking inside the autoregressive loop. Standard decoding repeatedly selects a token from the vocabulary distribution and feeds it back into the context. Mode A com-

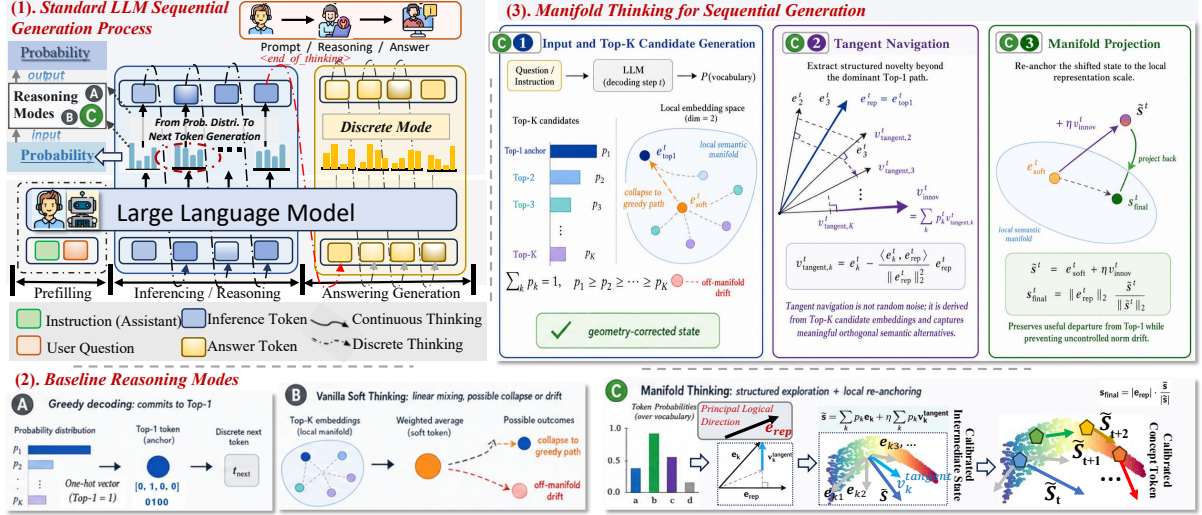


Figure 3: **Main overview of Manifold Thinking.** Panel (1), *Standard LLM Sequential Generation Process*, shows the usual prefill, reasoning/inference, and answer pipeline. Panel (2), *Baseline Reasoning Modes*, contrasts **A** greedy Top-1 decoding with **B** vanilla Soft Thinking, where linear Top- K mixing may collapse toward the greedy path or drift off the local manifold. Panel (3), *Manifold Thinking for Sequential Generation*, presents **C** as geometry-corrected reasoning: **C1** builds the Top- K semantic neighborhood, **C2** performs tangent navigation around the dominant semantic anchor, and **C3** projects the shifted state back to the local representation scale.

mits to Top-1, which is stable but discards subdominant alternatives; mode **B** averages Top- K embeddings, but the unconstrained average may stay too close to the anchor or enter an unsupported region. Manifold Thinking (**C**) keeps a single trajectory while replacing averaging with **C1**–**C3**: candidate-neighborhood construction, tangent innovation extraction, and local norm re-anchoring.

3.1 Preliminaries and Design Principles

Let V be the vocabulary and let $E \in \mathbb{R}^{|V| \times d}$ be the token embedding matrix. At step t , the model produces a distribution over V conditioned on context C_t . CSR constructs a soft token from the Top- K candidates:

$$\mathbf{e}_{\text{soft}}^t = \sum_{k=1}^K p_k^t \mathbf{e}_k^t, \quad (4)$$

where $\mathbf{e}_k^t$ is the k -th candidate embedding and p_k^t is its normalized Top- K probability. This convex combination is simple, but assumes that ambient-space interpolation remains semantically and computationally valid.

We use the *Latent Semantic Manifold* to denote the empirical high-density region occupied by natural token embeddings and hidden states under normal model computation. A synthetic state is more on-manifold when it is locally consistent in density, subspace, norm, and induced behavior, and

more off-manifold when it violates these criteria. We use this notion operationally rather than as a closed-form global manifold. The latent semantic manifold is therefore a global empirical concept, while the local reasoning manifold defined below is its step-wise operational approximation induced by the current Top- K candidate neighborhood.

Definition 3.1 (Local Reasoning Manifold).

For a frozen language model M and task family, the reasoning manifold denotes the representation states used during successful intermediate reasoning under the model’s learned dynamics. At step t , this global manifold is unobserved, so Manifold Thinking uses a local reasoning manifold induced by the current context C_t :

$$\widehat{\mathcal{M}}_t = \mathcal{N}(\{(\mathbf{e}_k^t, p_k^t, \mathcal{F}_t(\mathbf{e}_k^t))\}_{k=1}^K),$$

where $\mathcal{F}_t(\cdot)$ is the one-step continuation computation and $\mathcal{N}(\cdot)$ is the neighborhood jointly supported by Top- K embeddings, probabilities, and induced continuation behaviors. Concretely, it is operationalized through four observed quantities: normalized candidate probabilities, candidate embedding directions, the empirical scale of candidate norms, and one-step continuation behavior. All main experiments use $K = 20$. A synthetic state $\mathbf{v}^t$ is

locally on-manifold when its direction, norm, and induced behavior stay within this neighborhood’s empirical scale, and off-manifold when they leave these ranges.

A local anchor provides a stable reference for step-wise departures. It should summarize the most reliable direction in the current candidate neighborhood, so tangent components capture information not already explained by the dominant semantic direction. Manifold Thinking follows three requirements: retain structured signals from sub-dominant candidates, use actual candidate embeddings rather than arbitrary noise, and re-anchor the shifted state to a context-dependent scale. Additional methodological details on Stochastic Soft Thinking, the compact update, tensor-parallel implementation, and complexity are given in Appendix F.

3.2 Latent Manifold Navigation

At each step, the Top- K candidates define the local neighborhood, corresponding to **C1** in Figure 3. Manifold Thinking then constructs a corrected continuous reasoning state $\mathbf{s}_{final}^t$ in two stages: Tangent Navigation extracts a structured direction relative to the dominant semantic anchor, and Manifold Projection re-anchors the state to a locally valid representation scale.

3.2.1 Tangent Navigation

In the main method, the local anchor at step t is the Top-1 token embedding:

$$\mathbf{e}_{top1}^t = \mathbf{e}_1^t. \quad (5)$$

Top-1 is used as a stable reference axis, not as the destination of the correction; alternative reference vectors are compared in Appendix H. Each candidate is decomposed into a component parallel to $\mathbf{e}_{top1}^t$ and a component orthogonal to it. The orthogonal projection matrix is

$$\mathbf{P}_{\perp}^t = \mathbf{I} - \frac{\mathbf{e}_{top1}^t (\mathbf{e}_{top1}^t)^{\top}}{\|\mathbf{e}_{top1}^t\|_2^2}. \quad (6)$$

For candidate k , the tangent component is

$$\mathbf{v}_{tangent,k}^t = \mathbf{P}_{\perp}^t \mathbf{e}_k^t = \mathbf{e}_k^t - \frac{\langle \mathbf{e}_k^t, \mathbf{e}_{top1}^t \rangle}{\|\mathbf{e}_{top1}^t\|_2^2} \mathbf{e}_{top1}^t. \quad (7)$$

We aggregate tangent components with the original Top- K probabilities:

$$\mathbf{v}_{innov}^t = \sum_{k=1}^K p_k^t \mathbf{v}_{tangent,k}^t. \quad (8)$$

The intermediate state is

$$\tilde{\mathbf{s}}^t = \mathbf{e}_{soft}^t + \eta \mathbf{v}_{innov}^t, \quad (9)$$

where $\eta \geq 0$ controls navigation strength. Unlike random noise, $\mathbf{v}_{innov}^t$ is built from semantic candidates and contains only components orthogonal to the local reference direction. The update corrects the vanilla soft-token state rather than reconstructing from the anchor: using $\mathbf{e}_{top1}^t$ as the baseline would replace the probability-weighted centroid with a single embedding instead of correcting the state that CSR actually produces.

3.2.2 Local Norm Rescaling (Manifold Projection)

Tangent Navigation increases exploration, but the shifted state may have a norm inconsistent with natural representations. To prevent accumulated mismatch, we apply local norm re-anchoring:

$$\mathbf{s}_{final}^t = \|\mathbf{e}_{top1}^t\|_2 \cdot \frac{\tilde{\mathbf{s}}^t}{\|\tilde{\mathbf{s}}^t\|_2}. \quad (10)$$

The step is not an exact projection onto a known global manifold, but a lightweight local approximation: the target scale is the current anchor norm, not a global average or arbitrary scalar. The final state $\mathbf{s}_{final}^t$ is used as the current continuous reasoning token, remaining distinct from the anchor direction while locally consistent with the empirical representation manifold.

3.3 Theoretical Interpretation

Existing CSR methods implicitly assume that the model’s response to a weighted embedding average approximates the weighted average of individual responses. This amounts to a local linearity assumption in ambient embedding space. Transformer layers combine attention with nonlinear operations such as softmax and position-wise feed-forward networks (Vaswani et al., 2017), so this local linearity is not guaranteed. Together with evidence of anisotropy in contextualized language-model representations (Ethayarajh, 2019), this motivates our empirical tests of whether unconstrained aggregation becomes dominant-direction aligned or locally miscalibrated.

Manifold Thinking can be read as a first-order local approximation to constrained navigation on the empirical latent semantic manifold. Tangent Navigation supplies a direction orthogonal to the local anchor reference, introducing structured exploration. Manifold Projection calibrates

the intermediate state to the local representation scale, reducing the risk that exploration becomes off-manifold drift. The resulting inference process conditions later predictions on the sequence of geometry-corrected continuous states $s_{final}^1, \dots, s_{final}^m$ rather than on unconstrained embedding averages. Thus, stochasticity alone is insufficient: it can increase departure from the dominant path, but without a manifold constraint may not reduce local scale, density, or behavior-cloud violations.

4 Experiments and Analysis

We evaluate *Manifold Thinking* on math and programming benchmarks using the verified suite. The main baselines are Greedy CoT, Pure CoT, Vanilla Soft Thinking, and Stochastic Soft Thinking. Section 2 studies Gumbel Soft Thinking as a diagnostic probe; downstream, the corresponding baseline is Stochastic Soft Thinking with the Gumbel-Softmax trick. Section 4.4 compares this probe with the full Manifold probe in behavior space, showing why geometric correction is not just stochasticity.

4.1 Experimental Setup

Models and benchmarks. The matched primary suite uses three open-source backbones: Qwen2.5-7B (Qwen et al., 2025), Llama-3.1-8B (Grattafiori et al., 2024), and the inference-specialized DeepSeek-R1-Distill-Qwen-7B (Guo et al., 2025). It covers four math benchmarks, GSM8K (Cobbe et al., 2021), MATH500 (Hendrycks et al., 2021; Lightman et al., 2024), AMC23 (Bai et al., 2023), and OlympiadBench-Math (He et al., 2024), plus two programming benchmarks, HumanEval (Chen et al., 2021) and MBPP (Austin et al., 2021). LiveCodeBench (LCB) (Jain et al., 2025) was evaluated with a different backbone set—QwQ-32B (Team et al., 2025), Llama-3.1-8B, and DeepSeek-R1-Distill-Qwen-32B (Guo et al., 2025)—so these results are reported separately in Appendix M.1 rather than pooled into Table 3. Math uses exact-match accuracy; code uses execution-based pass@1. Appendix B gives hyperparameters, baselines, and dataset descriptions; Appendix K gives the math, multiple-choice, and code prompts.

Implementation details. All main experiments use one consolidated inference configuration. Appendix B specifies the Top- K neighborhood, anchor, temperatures, navigation factor, budget, hardware, and baseline settings; Appendix G reports

sensitivity analyses; and Appendix L analyzes inference overhead.

Main-table scope. Appendix B defines the main-table baselines: Greedy CoT, Pure CoT, Vanilla Soft Thinking, and Stochastic Soft Thinking. Table 3 therefore focuses on single-trajectory, training-free baselines. For cost control, multi-trajectory voting, structured search, and training-based soft-token grounding are treated as alternative-route baselines rather than primary main-table baselines. Appendix E explains these routes, covers Self-Consistency, Tree/Graph-of-Thoughts (Wang et al., 2022; Yao et al., 2023; Besta et al., 2024), COCONUT (Hao et al., 2024), and Soft Tokens, Hard Truths (Butt et al., 2025), and includes a cost-normalized Self-Consistency comparison. For inspection, Appendix J gives two decoding-based interpretability procedures and human judgments over mapped reasoning traces.

4.2 Quantitative Results and Analysis

Table 3 shows that Manifold Thinking achieves the highest mean score in all 18 matched model-dataset configurations. Among the 17 non-degenerate paired comparisons, 10 reach $p < 0.05$ and seven are positive but non-significant trends; the remaining row has constant paired differences and is marked with a dash. The final columns give the mean gain over the strongest non-Manifold baseline and the paired test over the same five seeds; Appendix M gives the per-seed audit and full statistical interpretation.

Within the matched suite, all six HumanEval and MBPP comparisons are significant, with gains over the strongest non-Manifold baseline ranging from +0.61 to +3.11 points. This provides consistent code-generation evidence without relying on the separately scoped LCB runs; their aggregate and per-seed results appear in Appendix M.1, Table 22.

Math benchmarks show a complementary pattern: Manifold Thinking has the highest mean in every math setting, with clear gains on Qwen2.5-7B MATH500 (+1.60), Llama-3.1-8B MATH500 (+2.20), and DeepSeek-R1-Distill-Qwen-7B OlympiadBench-Math (+1.04). On smaller or more variable suites such as AMC23 and parts of OlympiadBench-Math, gains are positive mean trends rather than significance evidence; AMC23 contains only 40 problems, so one additional correct answer changes accuracy by 2.5 points.

Stochastic Soft Thinking remains a strong high-

Model	Dataset	Greedy CoT (%)	Pure CoT (%)	Vanilla Soft Thinking (%)	Stochastic Soft Thinking (%)	Manifold Thinking (%)	Δ	p -value
Qwen2.5-7B	GSM8K	92.57 $\pm$ 0.00	92.80 $\pm$ 0.59	91.81 $\pm$ 1.02	<u>93.10 $\pm$ 1.66</u>	93.63 $\pm$ 0.55	+0.53	0.345
	MATH500	75.80 $\pm$ 0.00	<u>77.20 $\pm$ 1.50</u>	73.00 $\pm$ 2.41	76.60 $\pm$ 3.85	78.80 $\pm$ 1.38	+1.60	< 0.001
	AMC23	<u>62.50 $\pm$ 0.00</u>	45.00 $\pm$ 1.77	60.00 $\pm$ 3.95	57.50 $\pm$ 8.29	65.00 $\pm$ 2.50	+2.50	0.089
	OlympiadBench-Math	35.61 $\pm$ 0.00	36.80 $\pm$ 1.24	35.91 $\pm$ 1.89	<u>37.83 $\pm$ 2.94</u>	38.87 $\pm$ 1.12	+1.04	0.287
	HumanEval	<u>76.83 $\pm$ 0.00</u>	75.61 $\pm$ 0.43	76.22 $\pm$ 0.61	<u>76.22 $\pm$ 1.14</u>	77.44 $\pm$ 0.43	+0.61	0.034
	MBPP	71.21 $\pm$ 0.00	70.82 $\pm$ 0.28	71.98 $\pm$ 0.48	<u>73.15 $\pm$ 0.91</u>	76.26 $\pm$ 0.28	+3.11	0.002
Llama-3.1-8B	GSM8K	85.97 $\pm$ 0.00	85.60 $\pm$ 0.60	84.00 $\pm$ 1.06	85.22 $\pm$ 1.56	86.81 $\pm$ 0.54	+0.83	0.026
	MATH500	<u>48.80 $\pm$ 0.00</u>	46.80 $\pm$ 1.56	45.80 $\pm$ 2.60	47.60 $\pm$ 3.89	51.00 $\pm$ 1.38	+2.20	0.023
	AMC23	22.50 $\pm$ 0.00	27.50 $\pm$ 1.77	12.50 $\pm$ 3.95	25.00 $\pm$ 8.10	30.00 $\pm$ 1.77	+2.50	0.089
	OlympiadBench-Math	23.74 $\pm$ 0.00	<u>24.33 $\pm$ 1.09</u>	23.44 $\pm$ 1.75	<u>25.07 $\pm$ 2.57</u>	25.96 $\pm$ 0.91	+0.89	0.296
	HumanEval	53.66 $\pm$ 0.00	<u>54.27 $\pm$ 0.43</u>	51.83 $\pm$ 0.75	<u>53.05 $\pm$ 1.43</u>	54.88 $\pm$ 0.43	+0.61	0.034
	MBPP	65.37 $\pm$ 0.00	64.20 $\pm$ 0.39	63.81 $\pm$ 0.55	<u>65.76 $\pm$ 0.91</u>	68.09 $\pm$ 0.28	+2.33	0.002
DeepSeek-R1-Distill-Qwen-7B	GSM8K	90.83 $\pm$ 0.00	90.98 $\pm$ 0.57	91.43 $\pm$ 0.91	91.74 $\pm$ 1.46	92.80 $\pm$ 0.52	+1.06	0.071
	MATH500	87.40 $\pm$ 0.00	<u>91.40 $\pm$ 0.14</u>	88.00 $\pm$ 0.32	90.80 $\pm$ 0.66	91.60 $\pm$ 0.14	+0.20	–
	AMC23	70.00 $\pm$ 0.00	<u>72.50 $\pm$ 1.77</u>	<u>77.50 $\pm$ 3.95</u>	75.00 $\pm$ 8.10	80.00 $\pm$ 2.50	+2.50	0.266
	OlympiadBench-Math	39.76 $\pm$ 0.00	40.95 $\pm$ 1.15	40.36 $\pm$ 1.93	40.65 $\pm$ 2.88	41.99 $\pm$ 1.02	+1.04	< 0.001
	HumanEval	67.07 $\pm$ 0.00	70.73 $\pm$ 0.61	71.95 $\pm$ 0.86	<u>73.78 $\pm$ 1.43</u>	76.22 $\pm$ 0.43	+2.44	0.009
	MBPP	<u>75.88 $\pm$ 0.00</u>	72.76 $\pm$ 0.28	74.71 $\pm$ 0.39	<u>75.49 $\pm$ 0.83</u>	77.43 $\pm$ 0.28	+1.56	< 0.001

Table 3: Comparative performance of **Manifold Thinking** against discrete, vanilla soft-token, and stochastic soft-token baselines on the matched primary suite of six benchmarks and three backbones. Each cell reports mean $\pm$ standard deviation over 5 seeds; deterministic Greedy CoT repeats the same path, so its standard deviation is 0.00. **Bold** marks the best result in each row, and underlining marks the best non-Manifold baseline. Δ and estimated p -values compare Manifold Thinking with the strongest non-Manifold baseline using a two-sided paired t -test over matched seeds when paired differences have nonzero variance. A dash indicates constant paired differences with zero variance, for which the paired t -test is undefined; it is not a missing value.

uncertainty comparator, becoming the strongest non-Manifold baseline on several rows: Qwen2.5-7B for GSM8K, OlympiadBench-Math, and MBPP; Llama-3.1-8B for OlympiadBench-Math and MBPP; and DeepSeek-R1-Distill-Qwen-7B for GSM8K and HumanEval. Manifold Thinking’s best mean results support the central claim: useful continuous-space reasoning needs more than random diversification; it needs exploratory movement that stays locally calibrated and geometrically grounded.

4.3 Ablation Study

Table 4 ablates the two core components with Qwen2.5-7B on math and code benchmarks. GSM8K and HumanEval provide standard settings, while OlympiadBench-Math stresses harder mathematical reasoning.

The ablation shows complementarity between Tangent Navigation and Manifold Projection. Removing Tangent Navigation lowers performance on all three benchmarks, with the largest reduction on OlympiadBench-Math; removing Manifold Projection also lowers all three results. The full method performs best by coupling structured departure with local re-anchoring. The corresponding LCB ablation uses QwQ-32B and is therefore reported separately in Appendix M.1, Table 23. Appendix I gives a concrete AMC23 case study.

4.4 Gumbel vs. Full Manifold Probe in Behavior Space

Section 2.3 shows that Gumbel probes move farther in embedding space from Top-1 behavior, while the projected probe has stronger behavior-cloud support. We restate this comparison with the metrics of Table 2, directly contrasting Gumbel with the full Manifold probe. Here, *Full Manifold Probe* applies Tangent Navigation followed by Manifold Projection, the full-method counterpart of the Projected Probe in Table 2.

Table 5 shows that stronger stochastic departure is not the same as better continuous reasoning. Gumbel gives larger H_{top1} , indicating stronger movement away from the greedy behavior, but the full Manifold probe gives lower E_{KDE} and higher N_{eff} on both datasets, indicating stronger behavior-cloud membership. This supports Manifold Thinking’s core mechanism: useful exploration should move beyond Top-1 collapse while staying inside the locally supported behavior cloud. In other words, stochasticity helps only when sampled directions remain compatible with the model’s local semantic geometry. A behavior-space visualization of this comparison is provided in Appendix D.2.

5 Related Work

Reasoning and test-time computation. Chain-of-thought exposes intermediate reasoning; Self-Consistency and tree/graph search improve robustness through multiple paths at additional inference

Method	GSM8K	Olympiad Math	HumanEval
Manifold Thinking	93.63 $\pm$ 0.55	38.87 $\pm$ 1.12	77.44 $\pm$ 0.43
– w/o Tangent Nav.	91.74 $\pm$ 0.72	35.46 $\pm$ 1.28	75.00 $\pm$ 0.75
– w/o Manifold Proj.	92.19 $\pm$ 0.60	37.39 $\pm$ 1.18	76.83 $\pm$ 0.43
Stochastic Soft Thinking	93.10 $\pm$ 1.66	37.83 $\pm$ 2.94	76.22 $\pm$ 1.14

Table 4: Ablation on Qwen2.5-7B across matched math/code settings. Values report mean $\pm$ std accuracy or pass@1 across 5 seeds.

Data	Probe	$E_{KDE} \downarrow$	$N_{\text{eff}} \uparrow$	$H_{\text{top1}} \uparrow$
AMC23	Gumbel $\tau_g = 0.5$	0.392	18.51	0.343
	Full Manifold	0.352	19.11	0.215
MATH500	Gumbel $\tau_g = 0.5$	0.446	18.14	0.372
	Full Manifold	0.399	19.07	0.248

Table 5: Direct behavior-cloud comparison between Gumbel and the full Manifold probe. Consistent with Section 2.3, Gumbel has larger Top-1 departure, while the Manifold probe has lower cloud energy and higher effective support.

cost (Wei et al., 2022; Wang et al., 2022; Yao et al., 2023; Besta et al., 2024). Manifold Thinking instead enriches the continuous state within a single trajectory.

Continuous reasoning and representation geometry. Training-based latent reasoning such as COCONUT moves computation beyond discrete tokens (Hao et al., 2024). Training-free Soft Thinking forms probability-weighted mixtures (Zhang et al., 2026), and Stochastic Soft Thinking adds Gumbel diversification (Wu et al., 2025; Jang et al., 2017); neither enforces locally supported geometry. Motivated by anisotropy in contextualized language-model representations (Ethayarajh, 2019), Manifold Thinking combines tangent innovation with local-scale projection. Appendix C provides the detailed related work; Appendix A maps the remaining supplementary material.

6 Conclusion

We present *Manifold Thinking*, a training-free framework that treats soft tokens as geometric states rather than unconstrained mixtures. Vanilla soft-token reasoning exhibits two failure regimes: low-entropy states collapse toward the greedy anchor, while high-entropy or stochastic states depart from it but drift beyond natural candidate continuations. Manifold Thinking combines Tangent Navigation for structured non-greedy movement with Manifold Projection for local re-anchoring. Across

math and code benchmarks, it outperforms discrete CoT and vanilla and stochastic soft-token baselines without additional training or architectural changes. The results suggest that useful continuous reasoning requires exploration within representation regions the model treats as plausible reasoning states.

Limitations

We approximate the token manifold available to a frozen language model at each decoding step using the local Top- K candidate set, a training-free surrogate rather than a complete characterization. Appendix J provides engineering-oriented interpretability procedures and human judgments over mapped reasoning traces, but these mappings are surrogate views rather than direct access to latent computation. Our experiments cover mathematical reasoning, multiple-choice math, and code generation with open-weight models. Future work should study broader model families, multilingual and tool-augmented reasoning, longer interactions, and larger dense and mixture-of-experts models.

Ethical Considerations

We use established math/code benchmarks and open-weight models without collecting personal or sensitive data. The author team audited only model-generated traces (Appendix J); no external participants were recruited, and only aggregate judgments are reported. Stronger reasoning may enable misuse, so deployments should retain model safeguards and human oversight. The method adds inference-time computation.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62372364) and the Shaanxi Province Natural Science Foundation, China (No. 2026JC-YXQN-213).

References

- Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and 1 others. 2021. [Program synthesis with large language models](#). *arXiv preprint arXiv:2108.07732*.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.
- Loïc Barrault, Paul-Ambroise Duquenne, Maha El-bayad, Artyom Kozhevnikov, Belen Alastruey, Pierre Andrews, Mariano Coria, Guillaume Couairon, Marta R Costa-jussà, David Dale, and 1 others. 2024. [Large concept models: Language modeling in a sentence representation space](#). *arXiv preprint arXiv:2412.08821*.
- Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. [Representation learning: A review and new perspectives](#). *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.
- Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and 1 others. 2024. [Graph of thoughts: Solving elaborate problems with large language models](#). In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 17682–17690.
- Eden Biran, Daniela Gottesman, Sohee Yang, Mor Geva, and Amir Globerson. 2024. [Hopping too late: Exploring the limitations of large language models on multi-hop queries](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14113–14130.
- Natasha Butt, Ariel Kwiatkowski, Ismail Labiad, Julia Kempe, and Yann Ollivier. 2025. [Soft tokens, hard truths](#). *arXiv preprint arXiv:2509.19170*.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, and 1 others. 2021. [Evaluating large language models trained on code](#). *arXiv preprint arXiv:2107.03374*.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. [Training verifiers to solve math word problems](#). *arXiv preprint arXiv:2110.14168*.
- Romain Cosentino and Sarath Shekizhar. 2024. [Reasoning in large language models: A geometric perspective](#). *arXiv preprint arXiv:2407.02678*.
- Kawin Ethayarajh. 2019. [How contextual are contextualized word representations](#). *Comparing the geometry of BERT, ELMo, and GPT-2 Embeddings*, 2.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. [The llama 3 herd of models](#). *arXiv preprint arXiv:2407.21783*.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *arXiv preprint arXiv:2501.12948*.
- Wes Gurnee and Max Tegmark. 2024. [Language models represent space and time](#). In *International Conference on Learning Representations*, volume 2024, pages 2483–2503.
- Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. 2024. [Training large language models to reason in a continuous latent space](#). *arXiv preprint arXiv:2412.06769*.
- Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, and 1 others. 2024. [Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3828–3850.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.
- Naman Jain, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2025. [Livecodebench: Holistic and contamination free evaluation of large language models for code](#). In *International Conference on Learning Representations*, volume 2025, pages 58791–58831.
- Eric Jang, Shixiang Gu, and Ben Poole. 2017. [Categorical reparameterization with gumbel-softmax](#). *Preprint*, arXiv:1611.01144.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. [Large language models are zero-shot reasoners](#). *Advances in neural information processing systems*, 35:22199–22213.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. [Let’s verify step by step](#). In *The Twelfth International Conference on Learning Representations*.

- Samuel Marks and Max Tegmark. 2023. [The geometry of truth: Emergent linear structure in large language model representations of true/false datasets](#). *arXiv preprint arXiv:2310.06824*.
- Deepak Narayanan, Mohammad Shoeybi, Jared Casper, Patrick LeGresley, Mostofa Patwary, Vijay Korthikanti, Dmitri Vainbrand, Prethvi Kashinkunti, Julie Bernauer, Bryan Catanzaro, and 1 others. 2021. Efficient large-scale language model training on gpu clusters using megatron-lm. In *Proceedings of the international conference for high performance computing, networking, storage and analysis*, pages 1–15.
- Shuren Qi, Fei Wang, Tiejong Zeng, and Fenglei Fan. 2025. [Rethink deep learning with invariance in data representation](#). In *Companion Proceedings of the ACM on Web Conference 2025*, pages 45–48.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, and 25 others. 2025. [Qwen2.5 technical report](#).
- Abulhair Saparov and He He. 2022. [Language models are greedy reasoners: A systematic formal analysis of chain-of-thought](#). *arXiv preprint arXiv:2210.01240*.
- Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. [Scaling llm test-time compute optimally can be more effective than scaling model parameters](#). *arXiv preprint arXiv:2408.03314*.
- Qwen Team and 1 others. 2025. Qwq-32b: Embracing the power of reinforcement learning.
- Joshua B Tenenbaum, Vin de Silva, and John C Langford. 2000. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*, volume 30.
- Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. [Improving text embeddings with large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. [Self-consistency improves chain of thought reasoning in language models](#). *arXiv preprint arXiv:2203.11171*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). *Advances in neural information processing systems*, 35:24824–24837.
- Junhong Wu, Jinliang Lu, Zixuan Ren, Gangqiang Hu, Zhi Wu, Dai Dai, and Hua Wu. 2025. [Llms are single-threaded reasoners: Demystifying the working mechanism of soft thinking](#). *arXiv preprint arXiv:2508.03440*.
- Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. 2024. [Do large language models latently perform multi-hop reasoning?](#) In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10210–10229.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. [Tree of thoughts: Deliberate problem solving with large language models](#). *Advances in neural information processing systems*, 36:11809–11822.
- Zhen Zhang, Xuehai He, Weixiang Yan, Ao Shen, Chenyang Zhao, and Xin Wang. 2026. Soft thinking: Unlocking the reasoning potential of llms in continuous concept space. *Advances in Neural Information Processing Systems*, 38:168990–169012.

A Appendix Roadmap

The appendix is organized to separate experimental details, related work, diagnostic evidence, seed-level audit trails, method details, and implementation material. Table 6 gives a compact map of the remaining supplementary sections.

B Experimental, Dataset, and Baseline Details

This appendix records the experimental protocol, primary baseline definitions, and benchmark details used by the matched primary suite and supplementary evaluations. Table 7 summarizes the main hyperparameters, while Table 8 lists the benchmark roles and evaluated split sizes. The exact prompts are provided in Appendix K: general mathematical prompts for GSM8K, MATH500, and OlympiadBench-Math; multiple-choice prompts for AMC23; and code-generation prompts for HumanEval, MBPP, and LCB.

Implementation protocol. All reported main experiments use $\text{Top-}K = 20$ to define the local candidate neighborhood for soft-token construction. Section 3.2 defines the main-method anchor directly as the Top-1 token embedding $\mathbf{e}_{\text{top1}}$. This choice follows the reference-vector analysis in Appendix H, where Top-1 anchoring gives the strongest downstream accuracy and the most stable drift diagnostics compared with Soft Centroid and sub-dominant-token anchors. We set the decoding temperature to $T = 0.6$ for mathematical and multiple-choice tasks and $T = 0.2$ for code-generation tasks; the former matches the published Soft Thinking default (Zhang et al., 2026), whereas the lower code-generation temperature is our task-specific choice. The navigation factor is fixed to $\eta = 0.2$ across all reported model–dataset configurations; preliminary probing and the extended analyses in Appendix G show that this setting balances tangent innovation against excessive manifold displacement. Experiments use a maximum generation length of 24000 tokens and are run on $8 \times$ NVIDIA RTX 4090 GPUs.

Primary baselines. Greedy CoT performs standard token-by-token greedy decoding. Pure CoT is a single sampled discrete chain-of-thought baseline ($N = 1$ per seed) using the same prompt family and task temperature, without voting or answer aggregation (Wei et al., 2022). Vanilla Soft Thinking is the training-free CSR baseline that constructs

continuous soft tokens without geometric correction (Zhang et al., 2026). Stochastic Soft Thinking follows the original authors’ suggested Gumbel-Softmax trick (Jang et al., 2017; Wu et al., 2025): we add i.i.d. Gumbel noise to the Top- K logits, apply a softmax at the same task temperature, and use the resulting differentiable weights to form the soft token. This baseline covers the latest stochastic soft-token strategy considered in our single-trajectory, training-free setting. Our method is also training-free and uses a single decoding trajectory. Its extra computation is $O(Kd)$ per step, which is a constant-factor overhead under fixed Top- K and model dimension, rather than additional model calls. Measured latency/throughput overheads are provided in Appendix L.

The separately scoped LCB comparison uses QwQ-32B, Llama-3.1-8B, and DeepSeek-R1-Distill-Qwen-32B; this heterogeneous mapping is excluded from the matched three-backbone primary grid and is reported in Appendix M.1.

C Detailed Related Work

This section expands the compact overview in Section 5 with additional distinctions among multi-trajectory inference, training-based latent reasoning, training-free soft-token construction, and geometric views of representation space.

Discrete reasoning and test-time scaling. Chain-of-Thought prompting improves arithmetic and symbolic reasoning by making intermediate steps explicit (Wei et al., 2022; Kojima et al., 2022). In this paradigm, the model is encouraged to decompose a problem into natural-language reasoning steps before producing the final answer, which makes the intermediate computation inspectable and often improves accuracy on multi-step tasks. Formal analyses further show that language models can follow locally plausible but globally incomplete proof paths, motivating explicit aggregation or search (Saparov and He, 2022). Follow-up methods improve the reliability of this discrete reasoning trace in different ways. Self-Consistency samples multiple CoT trajectories and aggregates their final answers, reducing the brittleness of a single sampled path (Wang et al., 2022). Tree-of-Thoughts and Graph-of-Thoughts generalize the linear trace into search structures whose nodes are partial thoughts, enabling explicit branching, evaluation, and recombination (Yao et al., 2023; Besta et al., 2024). More broadly, test-time scaling stud-

Appendix	Section	Contents
B	Experimental, Dataset, and Baseline Details	Consolidated experimental protocol, current hyperparameter settings, primary baseline definitions, benchmark descriptions, and links to the exact prompts.
C	Detailed Related Work	Extended background on discrete reasoning, test-time scaling, continuous and stochastic soft reasoning, and representation geometry.
D	Additional Meta-Experiment Diagnostics	Behavior-cloud metric definitions used to quantify whether a continuous state remains supported by nearby candidate behaviors.
E	Alternative-Route Baselines	Scope rationale for multi-trajectory, search-based, and training-based baselines, including a representative Self-Consistency comparison.
F	Additional Methodological Details	Relation to Stochastic Soft Thinking, the compact mathematical update, tensor-parallel implementation logic, and computational complexity.
G	Hyperparameter Sensitivity and Configuration	Joint η - T grid, final hyperparameter choices, dynamic η schedules, and extended ablations over η , T , and Top- K .
H	Analysis of Principal Directions	Anchor-choice diagnostics showing why the Top-1 direction is the most stable reference for tangent extraction.
I	Case Study	A worked AMC23 example visualizing logical bifurcation and step-wise manifold steering.
J	Interpretability Analysis	Decoding-based inspection procedures and a small human audit of mapped Manifold Thinking traces.
K	Prompting Templates	Standardized math, multiple-choice, and code-generation prompts used across the evaluation suite.
L	Inference Overhead Analysis	Tokens/sec, latency, and per-step overhead measurements for the additional vector operations.
M	Additional Downstream Analysis	Supplementary heterogeneous-backbone LCB results and per-seed correct-count audits for the primary and supplementary evaluations.

Table 6: Roadmap for the appendix sections.

Setting	Value	Notes	Dataset	Task type	Size used	Metric	Evaluation role
Backbone models	Qwen2.5-7B; Llama-3.1-8B; DeepSeek-R1-Distill-Qwen-7B	Same three model families are used for all main tables.	GSM8K	Grade-school math word problems	1319	Exact-match accuracy	Tests arithmetic and multi-step natural-language reasoning.
Random seeds	5	All stochastic methods report mean $\pm$ standard deviation over matched seeds.	MATH500 (Hendrycks et al., 2021; Lightman et al., 2024)	Competition-style math	500	Exact-match accuracy	Uses the 500-problem held-out subset of MATH.
Top- K candidate set	20	Defines the local candidate neighborhood for Vanilla Soft Thinking, Stochastic Soft Thinking, and Manifold Thinking.	AMC23 (Bai et al., 2023)	2023 AMC 12A/12B math	40	Exact-match accuracy	Uses the 40-problem Qwen2.5-Math evaluation set.
Local anchor	Top-1 token embedding	Main Manifold Thinking experiments use $\mathbf{e}_{\text{top1}}$; Appendix H compares anchor alternatives.	OlympiadBench-Math	Olympiad-style math	674	Exact-match accuracy	Tests high-difficulty mathematical reasoning on a larger math-only slice.
Navigation factor	$\eta = 0.2$	Fixed across datasets after preliminary probing; Appendix G reports sensitivity analyses.	HumanEval	Python code generation	164	Execution pass@1	Measures functional program synthesis from docstring-style prompts.
Temperature	0.6 / 0.2	Math and multiple-choice tasks use $T = 0.6$; code-generation tasks use $T = 0.2$.	MBPP	Python code generation	257	Execution pass@1	Tests short natural-language programming problems with unit tests.
Greedy CoT	Deterministic	Greedy decoding is evaluated once and repeated across seed columns for paired reporting.	LCB	Code generation	279	Execution pass@1	Supplementary evaluation of recent competitive-programming-style tasks with a separately reported backbone set.
Pure CoT	One sampled trajectory per seed	Uses the task-specific temperature without soft-token interpolation.					
Stochastic Soft Thinking	Gumbel-Softmax over Top- K candidates	Uses the same local candidate set as Manifold Thinking; diagnostic Gumbel probes use $\tau_g = 0.5$.					
Maximum generation length	24000 tokens	Shared upper bound used in the current evaluation configuration.					
Hardware	8 $\times$ NVIDIA RTX 4090	Matches the throughput and overhead measurements in Appendix L.					

Table 7: Current experimental hyperparameter summary. Values are intended as the reproducibility defaults for the reported experiments; extended sensitivity results are given in Appendix G.

ies show that allocating additional inference computation can improve reasoning performance (Snell et al., 2024). The weakness of these approaches is cost: gains usually require extra trajectories, ver-

Table 8: Benchmark descriptions and evaluated split sizes for the six primary benchmarks and the supplementary LCB evaluation.

ifier calls, or broader search. Our goal is different: we ask whether a single trajectory can carry richer intermediate information through continuous states.

Continuous and stochastic soft reasoning. Analyses of standard LLMs find evidence for latent multi-hop computations in hidden states (Yang et al., 2024; Biran et al., 2024), whereas Large Concept Models perform autoregressive generation in a sentence-representation space (Barrault

et al., 2024). Instruction-tuned text-embedding models learn global text representations through contrastive fine-tuning (Wang et al., 2024); they are representation-learning methods rather than latent-reasoning methods. Training-based methods such as COCONUT (Hao et al., 2024) show that latent states can support reasoning, but require additional optimization. Vanilla Soft Thinking (Zhang et al., 2026) is closer to our setting because it constructs training-free soft tokens as probability-weighted mixtures of filtered candidate embeddings. However, vanilla soft-token mixing can become Top-1 dominated. Stochastic Soft Thinking (Wu et al., 2025) is a strong stochastic baseline because it implements the Gumbel-Softmax trick (Jang et al., 2017) over Top- K logits before soft-token construction, but it still does not explicitly constrain where the resulting vector lies in representation space.

Representation geometry. Contextualized representations in ELMo, BERT, and GPT-2 are anisotropic rather than uniformly distributed in the ambient space (Ethayarajh, 2019). More broadly, geometric deep learning emphasizes symmetry and invariance as representation priors (Qi et al., 2025), while representation-learning and manifold-learning perspectives study lower-dimensional structure, interpolation, disentanglement, and local generalization (Bengio et al., 2013; Tenenbaum et al., 2000). Recent analyses of large language models further suggest that internal activations exhibit organized geometric structure: LLMs can encode space and time as approximately linear representations (Gurnee and Tegmark, 2024), truth-related features can emerge as linear directions in activation space (Marks and Tegmark, 2023), and reasoning capacity can be studied through intrinsic geometric properties of attention-induced representations (Cosentino and Shekkizhar, 2024). This matters for CSR because a soft token is not merely a Euclidean point; it is an input to a model trained on naturally occurring token and contextual representations. Manifold Thinking follows this view by using local Top- K geometry at inference time: tangent movement supplies structured exploration, while projection re-anchors the state to the candidate cloud’s natural scale. Training-based soft-token grounding methods such as Soft Tokens, Hard Truths (Butt et al., 2025) are complementary, but they introduce extra optimization and are therefore treated as auxiliary references rather than primary training-free baselines.

D Additional Meta-Experiment Diagnostics

This section provides the metric definitions behind the diagnostic questions studied in Section 2. We keep these details in the appendix to reserve the main text for the three primary meta-experiments and the method description.

D.1 Behavior-Cloud Metric Definitions

For each step t , we construct a reference behavior set

$$\mathcal{Y}_t = \{p_{\text{next}}(\mathbf{e}_k^t)\}_{k=1}^K, \quad (12)$$

where $p_{\text{next}}(\mathbf{v})$ is the next-token distribution induced by inserting vector $\mathbf{v}$ into the continuation computation. Distributions are evaluated over the union of Top- L output tokens, with remaining mass assigned to an OTHER bucket. Let $d_H(\cdot, \cdot)$ denote Hellinger distance. We use a kernel-density cloud energy

$$E_{KDE}(\mathbf{v}^t) = -\log \frac{1}{K} \sum_{\mathbf{q} \in \mathcal{Y}_t} \exp(-\Delta_{\mathbf{q}}(\mathbf{v}^t)),$$

$$\Delta_{\mathbf{q}}(\mathbf{v}^t) = \frac{d_H(p_{\text{next}}(\mathbf{v}^t), \mathbf{q})^2}{2\sigma_{H,t}^2}, \quad (13)$$

where $\sigma_{H,t}$ is the median pairwise Hellinger distance within $\mathcal{Y}_t$. Lower E_{KDE} indicates higher density under the local behavior cloud. We also report the effective number of supporting neighbors,

$$N_{\text{eff}}(\mathbf{v}^t) = \frac{1}{\sum_{\mathbf{q} \in \mathcal{Y}_t} w_{\mathbf{q}}(\mathbf{v}^t)^2}, \quad (14)$$

where $w_{\mathbf{q}}$ is the normalized kernel weight from Eq. 13. Finally, H_{top1} is the Hellinger distance from the Top-1 behavior.

D.2 Behavior-Cloud Visualization

Figure 4 visualizes the behavior-cloud comparison used in Section 4.4. The plot is representative rather than a raw PCA dump: it places Gumbel and the full Manifold probe according to the qualitative geometry implied by the measured cloud-energy, effective-support, and anchor-departure diagnostics. Gumbel moves farther from the greedy anchor, while the full Manifold probe remains closer to the local behavior cloud.

E Alternative-Route Baselines and Scope

This section explains how we position baselines that belong to adjacent reasoning routes but differ from our target setting. Our main experiments

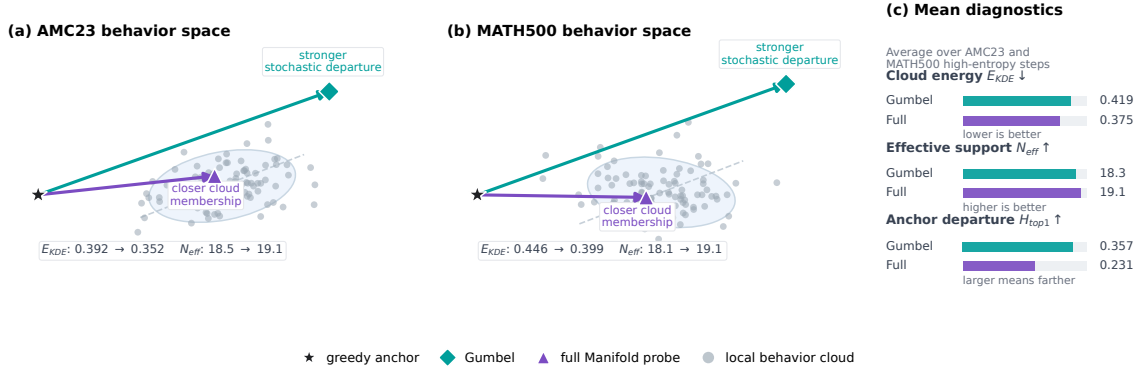


Figure 4: **Behavior-cloud support under stochastic and manifold-corrected probes.** Representative behavior-space projections show that Gumbel moves farther from the greedy anchor, while the full Manifold probe remains closer to the local behavior cloud. The diagnostic bars summarize the same pattern: the full probe lowers cloud energy and increases effective support, whereas Gumbel provides stronger anchor departure.

isolate a frozen-model, single-trajectory, training-free intervention: each method receives the same model family and operates inside one decoding trajectory without additional training, external search expansion, or multiple complete generations.

Alternative routes. Self-Consistency samples several independent CoT trajectories and aggregates final answers, usually by majority vote or answer agreement. Tree-of-Thoughts expands intermediate reasoning states as a search tree and scores or prunes candidate thoughts with heuristic or model-based evaluators. Graph-of-Thoughts generalizes this idea by allowing reasoning units to form graph-structured dependencies rather than a single tree. These search or voting methods can improve robustness, but their gains come with multiple full generations, additional scoring calls, larger memory footprints for candidate states, and extra orchestration decisions such as branching width, depth, pruning, and verifier design.

Training-based latent-reasoning methods form a second route. COCONUT (Hao et al., 2024) trains models to replace parts of explicit CoT with continuous hidden states through a staged latent-reasoning curriculum. Soft Tokens, Hard Truths (Butt et al., 2025) learns soft or fuzzy continuous CoT representations through RL-style post-training. These methods are highly relevant to continuous reasoning, but they change model parameters and training objectives. They therefore answer a different question from our paper, which asks whether a frozen LLM can be improved at inference time by correcting the geometry of the soft-token state.

Method family	Typical implementation	Setting mismatch / role in this paper
Self-Consistency (Wang et al., 2022)	Sample n independent CoT trajectories and aggregate final answers.	Multi-trajectory cost scales with n full generations; included below as a representative upper-cost comparison.
Tree-of-Thoughts (Yao et al., 2023)	Expand, score, and prune intermediate reasoning branches.	Requires branching, scoring, pruning, and repeated model/evaluator calls; related search route, not cost-matched.
Graph-of-Thoughts (Besta et al., 2024)	Organize generated reasoning units as graph nodes and operations.	Adds graph construction, scheduling, aggregation, and repeated evaluation overhead beyond single-trajectory decoding.
COCONUT (Hao et al., 2024)	Train models to use continuous hidden states through a latent-reasoning curriculum.	Requires training/fine-tuning and changes the model protocol; complementary training-based latent reasoning.
Soft Tokens, Hard Truths (Butt et al., 2025)	Use noisy soft or fuzzy tokens during RL-style post-training.	Requires post-training and parameter updates; related soft-token grounding, not a frozen-model inference baseline.

Table 9: Setting mismatch for alternative-route baselines. The main table focuses on frozen-model, single-trajectory, training-free methods.

For this reason, we do not treat all of these methods as primary main-table baselines. Doing so would mix the effect of our local geometric correction with the effect of extra trajectories, external search, or parameter updates. Nevertheless, we include Self-Consistency as a representative high-cost comparison because it is widely used, easy to interpret, and directly tests whether a single corrected continuous trajectory can approach a multi-trajectory discrete voting baseline.

E.1 Representative Comparison with Self-Consistency

Self-Consistency samples multiple independent discrete CoT trajectories and aggregates final answers. Table 10 compares this multi-trajectory baseline with a single Manifold Thinking trajectory on two math benchmarks and the MBPP code benchmark.

The comparison is intentionally not framed as a claim that Manifold Thinking universally raises the raw-accuracy ceiling above multi-trajectory

Method	Full Calls	Extra Ops	GSM8K	Olympiad-Math	MBPP
Manifold Thinking	1×	$O(Kd)$ / step	93.63%	38.87%	196/257 (76.26%)
SC@3	3×	vote over 3 paths	93.40%	38.72%	190/257 (73.93%)
SC@5	5×	vote over 5 paths	93.63%	39.02%	195/257 (75.88%)

Table 10: Cost-normalized comparison with Self-Consistency on Qwen2.5-7B. Full Calls denotes the number of complete decoding trajectories.

voting. Self-Consistency samples several complete discrete reasoning paths and selects by agreement, whereas Manifold Thinking applies a single-trajectory, training-free correction. On GSM8K, SC@5 ties Manifold Thinking; on OlympiadBench-Math, SC@5 is marginally higher. The intended comparison is therefore compute-normalized efficiency: SC@5 requires five full decoding runs, while Manifold Thinking adds only local Top- K embedding aggregation, tangent correction, and norm re-anchoring inside one trajectory.

MBPP represents the code-generation setting requested during review. SC@5 reaches 195/257 (75.88%) with five trajectories, while Manifold Thinking reaches 196/257 (76.26%) with one. For code, independently sampled programs can differ substantially in surface form, control flow, and error type; without an additional test-time oracle or a more involved program-selection rule, simple trajectory voting need not identify the functionally correct program. Manifold Thinking instead corrects the continuous state within the same generation trajectory, which is useful when syntax and long-range structural dependencies constrain valid continuations.

F Additional Methodological Details

F.1 Relation to Stochastic Soft Thinking

Stochastic Soft Thinking perturbs the Top- K logits before constructing the soft token using the Gumbel-Softmax trick (Wu et al., 2025; Jang et al., 2017). Given logits z_k^t , Gumbel noise $g_k \sim \text{Gumbel}(0, 1)$, and temperature τ_g , the stochastic weights are

$$\tilde{p}_k^t = \frac{\exp((z_k^t + g_k)/\tau_g)}{\sum_{j=1}^K \exp((z_j^t + g_j)/\tau_g)}, \quad (15)$$

and the Gumbel soft token is

$$\mathbf{e}_{\text{gumbel}}^t = \sum_{k=1}^K \tilde{p}_k^t \mathbf{e}_k^t. \quad (16)$$

This mechanism directly targets Top-1 dominance by diversifying the soft-token weights. However, it still constructs a vector through linear aggregation in the ambient embedding space and therefore does not explicitly enforce local density, norm calibra-

tion, or behavior consistency.

Manifold Thinking addresses a different aspect of the problem. Rather than perturbing the probability weights, it modifies the geometry of the resulting continuous state:

$$\mathbf{e}_{\text{soft}}^t \rightarrow \tilde{\mathbf{s}}^t \rightarrow \mathbf{s}_{\text{final}}^t, \quad (17)$$

where $\tilde{\mathbf{s}}^t$ is obtained by Tangent Navigation and $\mathbf{s}_{\text{final}}^t$ is obtained by Manifold Projection. Thus, Stochastic Soft Thinking provides stochastic diversification, while Manifold Thinking provides geometry-constrained state correction. The two mechanisms are conceptually distinct and can be combined, but our main method uses deterministic structured innovation and local re-anchoring. Training-based continuous-CoT methods, such as Soft Tokens, Hard Truths (Butt et al., 2025), address a related stability problem through additional optimization. They are therefore complementary to our frozen-model inference setting rather than primary training-free baselines.

F.2 Mathematical Update Summary

This subsection gives the model-agnostic update in compact algebraic form. At decoding step t , Manifold Thinking first constructs the usual soft token and then decomposes candidate embeddings around the local Top-1 anchor:

$$\mathbf{e}_{\text{soft}}^t = \sum_{k=1}^K p_k^t \mathbf{e}_k^t, \quad (18)$$

$$\mathbf{P}_{\perp}^t = \mathbf{I} - \frac{\mathbf{e}_{\text{top1}}^t (\mathbf{e}_{\text{top1}}^t)^{\top}}{\|\mathbf{e}_{\text{top1}}^t\|^2}, \quad (19)$$

$$\mathbf{v}_{\text{innov}}^t = \sum_{k=1}^K p_k^t \mathbf{P}_{\perp}^t \mathbf{e}_k^t, \quad (20)$$

$$\mathbf{s}_{\text{final}}^t = \|\mathbf{e}_{\text{top1}}^t\| \frac{\mathbf{e}_{\text{soft}}^t + \eta \mathbf{v}_{\text{innov}}^t}{\|\mathbf{e}_{\text{soft}}^t + \eta \mathbf{v}_{\text{innov}}^t\|}. \quad (21)$$

The update separates the two roles of the method. The orthogonal innovation term preserves subdominant semantic directions that would be suppressed by a Top-1-dominated soft token, while the final norm alignment keeps the corrected state on the local scale of the model’s native embedding space. Because all quantities are computed from the current Top- K distribution and frozen embedding table, the method remains training-free and does not introduce an additional decoding trajectory.

Tensor-parallel implementation. The same algebraic update is compatible with tensor model

parallelism, in which individual model layers are partitioned across ranks (Narayanan et al., 2021). When the embedding matrix is vocabulary-sharded by the implementation, the soft baseline, Top-1 anchor, and innovation signal must be reconstructed consistently across shards. Each rank first performs masked local embedding lookup for the Top- K candidate ids it owns, while non-local candidates are represented by zeros. The weighted soft baseline and the Top-1 anchor are then recovered by AllReduce, so the subsequent tangent update is algebraically identical to the single-device version. The projection radius is also computed from the reconstructed global Top-1 anchor rather than from a local shard, preventing rank-dependent norm drift. Thus the implementation adds only small vector reductions around the existing decoding step, while preserving the same token-level state passed to the following transformer block.

Algorithm 1 Manifold Thinking: TP-Aware Forward Pass

```

1: procedure WEIGHTEDFORWARDTP(probs, indices,  $\eta$ )
2:    $\triangleright$  Phase 1: local shard retrieval
3:   mask  $\leftarrow$  GetVocabMask(indices)
4:   embeds_local  $\leftarrow$  EmbeddingLookup(indices)
5:   embeds_local.MaskZeros(mask)
6:    $\triangleright$  Phase 2: global linear baseline
7:   v_lin_local  $\leftarrow$  Sum(probs  $\times$  embeds_local)
8:   v_linear  $\leftarrow$  AllReduce(v_lin_local)
9:    $\triangleright$  Phase 3: global Top-1 anchor
10:  e_top1_local  $\leftarrow$  embeds_local[:, 0, :]
11:  e_top1  $\leftarrow$  AllReduce(e_top1_local)
12:   $\triangleright$  Phase 4: tangent navigation
13:  R  $\leftarrow$  Norm(e_top1)
14:  u_ref  $\leftarrow$  e_top1 / (R +  $10^{-6}$ )
15:  dot  $\leftarrow$  Sum(embeds_local  $\times$  u_ref)
16:  v_tan_local  $\leftarrow$  embeds_local - (dot  $\times$  u_ref)
17:  v_innov  $\leftarrow$  AllReduce(Sum(probs  $\times$  v_tan_local))
18:   $\triangleright$  Phase 5: manifold projection
19:  s_new  $\leftarrow$  v_linear + ( $\eta \times$  v_innov)
20:  final_output  $\leftarrow$  (s_new / Norm(s_new))  $\times$  R
21:  return final_output, e_top1
22: end procedure

```

Implementation analysis. AllReduce at the baseline, anchor, and innovation stages ensures that every rank maintains a consistent global manifold center. The orthogonal residual in Phase 4 prevents η from simply scaling the redundant Top-1 direction, and the final projection preserves the local embedding scale. Thus the implementation mirrors the algebraic update while remaining compatible with tensor-parallel vocabulary sharding.

F.3 Computational Efficiency

The additional computation introduced by Manifold Thinking is lightweight. At each decoding step, Tangent Navigation computes orthogonal components for K candidate embeddings, and Manifold Projection performs a single vector nor-

(a) Mathematical Reasoning (GSM8K)				
T	$\eta = 0.1$	$\eta = 0.2$	$\eta = 0.3$	
0.4	85.22%	85.82%	85.37%	
0.6	85.97%	86.81%	85.82%	
0.8	84.84%	85.60%	84.23%	

(b) Code Generation (MBPP)				
T	$\eta = 0.1$	$\eta = 0.2$	$\eta = 0.3$	
0.1	66.93%	67.70%	65.76%	
0.2	67.32%	68.09%	66.54%	
0.4	65.76%	66.15%	63.81%	

Table 11: Joint hyperparameter grid on Llama-3.1-8B. The best setting in both domains uses the moderate navigation factor $\eta = 0.2$, while the preferred temperature differs between mathematical reasoning and code generation.

malization. The total complexity is

$$O(Kd), \quad (22)$$

where K is the Top- K width and d is the embedding dimension. In our experiments, K is small, typically $K = 20$.

This overhead is minor compared with the cost of a standard Transformer forward pass, whose self-attention and feed-forward components scale as $O(N^2d)$ and $O(Nd^2)$ with sequence length N . Manifold Thinking also avoids the multiplicative cost of multi-path sampling methods, since it operates within a single continuous reasoning trajectory. Quantitative latency and throughput measurements are reported in Appendix L.

G Hyperparameter Sensitivity and Configuration

G.1 Joint Sensitivity Analysis of η and T

To study the interaction between continuous exploratory momentum and sampling temperature, we conduct a joint hyperparameter grid on two representative domains: mathematical reasoning on GSM8K and code generation on MBPP. The grid uses Llama-3.1-8B and varies the navigation factor η and temperature T . The default setting used in the main experiments, $\eta = 0.2$, lies near the best region in both domains, while larger values begin to over-steer the latent state.

Table 11 shows that η and T are coupled. For GSM8K, the best configuration occurs at $T = 0.6$ and $\eta = 0.2$, suggesting that moderate entropy helps Tangent Navigation extract sub-dominant reasoning signals. For MBPP, the best configuration occurs at the lower temperature $T = 0.2$ and the same $\eta = 0.2$, consistent with the narrower and more syntax-sensitive manifold of code. Performance drops when either temperature or navigation

Category	Hyperparameter	Value
Navigation	η Search Space	$[0, 0.5]$
	Unified η Used in Experiments	0.2
Sampling	Top- K Search Width	20
	Top- p Threshold	0.95
Inference	Temperature T (Math / Code)	0.6 / 0.2
	Max Generation Tokens	24000

Table 12: Detailed hyperparameter settings. The unified navigation factor $\eta = 0.2$ is used across all reported model–dataset configurations after preliminary probing.

strength becomes too large, supporting the design principle that continuous reasoning benefits from exploration only when it remains geometrically anchored.

G.2 Final Hyperparameter Configuration

To ensure reproducibility, Table 12 lists the final configuration. Before the main evaluation, we performed a preliminary probe over $\eta \in [0, 0.5]$ and temperature settings. Across the joint grid and the one-dimensional η ablation, $\eta = 0.2$ gives the most balanced trade-off between exploratory tangent movement and projection stability. We therefore use this unified setting for all reported experiments.

G.3 Dynamic η Schedules

We also tested whether the navigation factor can be selected automatically at each decoding step. Since η controls the strength of the tangent correction, a natural idea is to tie it to local uncertainty or to the current amount of manifold displacement. We consider three lightweight schedules, all clipped to $[0.05, 0.35]$:

$$h_t = \frac{-\sum_{k=1}^K p_k^t \log p_k^t}{\log K}, \quad (23)$$

$$\eta_t^H = \text{clip}(0.05 + 0.30h_t), \quad (24)$$

$$\eta_t^P = \text{clip}(0.05 + 0.30(1 - p_{\text{top1}}^t)), \quad (24)$$

$$\eta_t^D = \text{clip}(0.05 + 0.30e^{-3\delta_t}), \quad (25)$$

$$\delta_t = \frac{\|\mathbf{e}_{\text{soft}}^t - \mathbf{e}_{\text{top1}}^t\|}{\|\mathbf{e}_{\text{top1}}^t\| + \epsilon}.$$

The entropy schedule increases exploration when the Top- K distribution is diffuse. The Top-1 schedule is a confidence-aware conservative rule: it reduces navigation when the model is highly confident in the dominant token. The displacement schedule increases tangent movement when the soft token is close to the local anchor, but suppresses extra movement once the state is already far from the anchor.

Table 13 shows that simple dynamic schedules do not consistently improve over the fixed $\eta = 0.2$ default. Entropy-aware scheduling slightly improves OlympiadBench-Math, and Top-1-confidence scheduling slightly improves HumanEval, but most rows degrade. This suggests that the useful navigation strength is not determined by first-order uncertainty or by a single displacement statistic alone. We suspect that η depends on richer local properties of the reasoning state, including second-order curvature of the logit surface, layer-wise Jacobian amplification of tangent perturbations, the covariance spectrum of the local Top- K behavior cloud, and the current semantic phase of reasoning. Designing a curvature-aware or phase-aware navigation controller is left for future work.

G.4 Extended Ablations over η , T , and Top- K

We further ablate the navigation factor, temperature, and Top- K width. These ablations are intended to verify that the default setting is not a narrow artifact of one task.

Table 14 confirms that $\eta = 0.2$ is a robust operating point. Performance is stable across $\eta \in [0.1, 0.3]$, suggesting that practitioners can treat 0.2 as a reliable default without extensive per-task tuning. The degradation at $\eta \geq 0.4$ is consistent with overly aggressive manifold displacement.

Table 15 shows that Manifold Thinking improves over Vanilla Soft Thinking across all temperature values. The advantage is especially visible at the default mathematical temperature, consistent with the Drift failure mode: when the distribution becomes sufficiently diffuse, norm alignment becomes more useful.

Table 16 shows that performance saturates at $K = 20$: expanding to $K = 50$ provides no additional gain but increases embedding retrieval cost, while $K = 5$ weakens the orthogonal innovation signal.

H Analysis of Principal Directions

In our primary experiments, we define the principal logical direction as the Top-1 embedding $\mathbf{e}_{\text{top1}}$. We also evaluate a probability-weighted mean (Soft Centroid) and a uniform mean of sub-dominant tokens. The goal is not to tune a new decoding strategy, but to verify whether the reference vector is consistently load-bearing across symbolic domains.

Table 17 and Figure 5 show the same ordering

Schedule	GSM8K	Olympiad-Math	HumanEval	LCB (QwQ-32B)
Fixed $\eta = 0.2$	93.63 $\pm$ 0.55	38.87 $\pm$ 1.12	77.44 $\pm$ 0.43	72.40 $\pm$ 0.25
Entropy-aware η_t^H	93.48 $\pm$ 0.72	39.02 $\pm$ 1.25	77.20 $\pm$ 0.55	71.68 $\pm$ 0.51
Top-1-conf. η_t^F	93.25 $\pm$ 0.80	38.43 $\pm$ 1.36	77.56 $\pm$ 0.51	71.33 $\pm$ 0.67
Displacement η_t^D	93.54 $\pm$ 0.62	38.72 $\pm$ 1.20	77.32 $\pm$ 0.51	72.04 $\pm$ 0.36

Table 13: Dynamic η schedules. GSM8K, OlympiadBench-Math, and HumanEval use Qwen2.5-7B; the supplementary LCB column uses QwQ-32B. Values are mean $\pm$ std across 5 independent seeds.

η	GSM8K Acc. (%)	HumanEval Pass@1 (%)
0.05	92.90 $\pm$ 0.20	76.59 $\pm$ 0.50
0.10	93.21 $\pm$ 0.20	77.07 $\pm$ 0.50
0.15	93.40 $\pm$ 0.20	77.32 $\pm$ 0.40
0.20	93.63 $\pm$ 0.55	77.44 $\pm$ 0.43
0.30	93.30 $\pm$ 0.20	76.95 $\pm$ 0.50
0.40	92.71 $\pm$ 0.30	76.22 $\pm$ 0.50
0.50	91.80 $\pm$ 0.30	75.12 $\pm$ 0.60

Table 14: Ablation over navigation factor η on Qwen2.5-7B. Values are mean $\pm$ std across 5 independent seeds.

T	GSM8K Acc. (%)		HumanEval Pass@1 (%)	
	Vanilla ST	Manifold	Vanilla ST	Manifold
0.1	91.80 $\pm$ 0.20	92.40 $\pm$ 0.20	76.46 $\pm$ 0.55	77.07 $\pm$ 0.33
0.2	92.10 $\pm$ 0.20	92.80 $\pm$ 0.20	76.22 $\pm$ 0.61	77.44 $\pm$ 0.43
0.4	92.60 $\pm$ 0.20	93.21 $\pm$ 0.20	76.10 $\pm$ 0.51	76.83 $\pm$ 0.43
0.6	91.81 $\pm$ 1.02	93.63 $\pm$ 0.55	75.49 $\pm$ 0.51	76.22 $\pm$ 0.43
0.8	92.40 $\pm$ 0.30	92.99 $\pm$ 0.30	74.76 $\pm$ 0.55	75.49 $\pm$ 0.67
1.0	91.49 $\pm$ 0.40	92.10 $\pm$ 0.40	73.66 $\pm$ 0.67	74.27 $\pm$ 0.67

Table 15: Temperature ablation on Qwen2.5-7B with $\eta = 0.2$. Values are mean $\pm$ std across 5 independent seeds. Bold entries mark the default temperatures used in the main experiments: $T = 0.6$ for math and $T = 0.2$ for code.

Top- K	GSM8K Acc. (%)	Overhead vs. $K = 20$
5	92.90 $\pm$ 0.20	-38.00%
10	93.30 $\pm$ 0.20	-19.00%
20	93.63 $\pm$ 0.55	0.00 (baseline)
50	93.50 $\pm$ 0.20	+31.00%

Table 16: Ablation over Top- K on GSM8K (Qwen2.5-7B). Values are mean $\pm$ std across 5 independent seeds.

across accuracy and diagnostics. Top-1 anchoring is best, Soft Centroid is second, and the sub-dominant-only reference is worst. As A_{ref} decreases, D_{drift} increases, indicating that references less aligned with the model’s highest-confidence token produce tangent directions that are less locally grounded. The Uniform Mean variant removes the most stable semantic anchor before orthogonalization, causing innovation components to be defined relative to low-probability directions and amplifying noise rather than meaningful sub-dominant signals.

Reference, not destination. The Top-1 embedding is not itself treated as a failure vector. At many low-entropy steps it is a reliable local direction; failure arises when probability concentration prevents a soft token from retaining alternatives at an uncertain or branching step. Manifold Thinking uses Top-1 only to define the reference axis, its orthogonal complement, and a local norm scale. Tangent Navigation adds candidate-supported components

Reference Strategy	GSM8K	MBPP	A_{ref}	D_{drift}
Top-1 Embedding (Ours)	93.63 $\pm$ 0.55	76.26 $\pm$ 0.28	0.91	0.16
Soft Centroid ($\sum p_i \mathbf{e}_i$)	92.42 $\pm$ 0.60	66.54 $\pm$ 0.62	0.82	0.26
Uniform Mean (Tokens 2-6)	88.93 $\pm$ 0.78	61.09 $\pm$ 0.73	0.60	0.53

Table 17: Principal-direction ablation on Qwen2.5-7B. Downstream columns report mean $\pm$ std across 5 independent seeds; A_{ref} is cosine alignment with the Top-1 embedding; D_{drift} is normalized distance from the local Top- K behavior cloud.

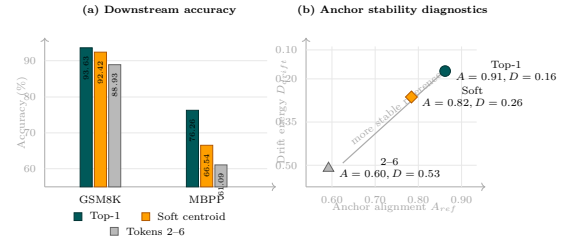


Figure 5: Principal-direction ablation and anchor diagnostics. Top-1 anchoring gives the strongest downstream accuracy and the most stable geometric diagnostics.

not explained by that axis, so when Top-1 is wrong at a high-entropy decision point the correction encourages structured departure rather than imitation. The final rescaling uses the Top-1 norm without projecting the direction back toward greedy decoding.

Why correct the soft-token state. The vanilla soft token remains the baseline in Eq. 9 because it is the probability-weighted state produced by continuous-space reasoning. Using the anchor itself as the baseline would discard that centroid and turn the update into Top-1-centered reconstruction rather than correction of the soft-token state.

I Case Study: Logical Bifurcation and Trajectory Steering

To provide a granular view of how Manifold Thinking facilitates exploratory reasoning, we visualize the decoding trajectory of a representative AMC23 instance using Qwen2.5-7B.

Semantic exploration. At Step 5, the Top-1 token probability falls below 60.00%, creating a logical hinge point where greedy decoding can prematurely collapse. Tangent Navigation generates a steering peak by using sub-dominant candidates.

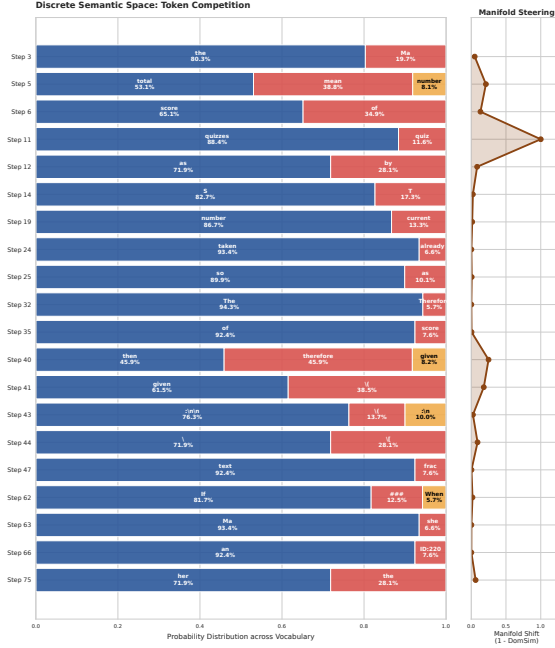


Figure 6: Visualization of Manifold Reasoning Dynamics on an AMC23 instance. The left panel displays the discrete Top- K probability distribution with token decoding; the right panel illustrates the Manifold Steering Force ($1 - \text{DomSim}$).

Latent calibration. At Steps 40 and 66, the distribution can look locally concentrated, but manifold steering remains active. The correction acts as a stabilizer rather than a replacement for the Top-1 signal.

Beyond Top-1 collapse. The discrete bars and continuous steering curve show that Manifold Thinking augments the dominant path with orthogonal pulses while preserving local semantic support.

J Interpretability Analysis

Continuous reasoning states are not directly readable as natural language tokens. Nevertheless, they can be inspected through controlled discrete mappings that expose the token-level logic most closely associated with each continuous step. We use two complementary procedures to provide engineering-oriented interpretability evidence for Manifold Thinking. The first follows the greedy-mapping strategy used in Soft Thinking analyses (Zhang et al., 2026); the second is a Manifold-specific nearest-candidate mapping designed for our geometry-corrected states.

J.1 Decoding-Based Inspection Procedures

Top-1 greedy mapping. Following prior Soft Thinking interpretability practice, we map each inspected continuous step to the ordinary Top-1

token under the current context:

$$z_{\text{top1}}^t = \arg \max_{z \in V} p(z \mid \mathcal{C}_t). \quad (26)$$

This procedure forces the latent exploration trace back onto the greedy discrete path. It is useful as a conservative readability baseline because every mapped token is a token that the frozen model would naturally choose under standard decoding. However, it can hide Manifold Thinking’s non-greedy correction, since tangent movement may be mapped back to the same dominant token even when the continuous state has shifted toward a different local semantic direction.

Manifold-nearest Top-20 mapping. To inspect the corrected Manifold token itself, we also map the final continuous state $\mathbf{s}_{\text{final}}^t$ to the nearest token within the current Top-20 candidate neighborhood:

$$k^* = \arg \max_{1 \leq k \leq 20} \frac{\langle \mathbf{s}_{\text{final}}^t, \mathbf{e}_k^t \rangle}{\|\mathbf{s}_{\text{final}}^t\|_2 \|\mathbf{e}_k^t\|_2}, \quad (27)$$

$$z_{\text{M20}}^t = z_{k^*}^t.$$

This mapping is local by construction: it does not search the full vocabulary, but restricts interpretation to the same candidate neighborhood used by Manifold Thinking. As a result, it can reveal whether tangent navigation moves the state toward a plausible sub-dominant candidate while still keeping the explanation tied to the current model distribution.

Mapping method	Engineering role	Main limitation
Top-1 greedy mapping	Provides a conservative readable trace aligned with standard decoding.	May collapse non-greedy manifold movement back to the dominant token.
Manifold-nearest Top-20 mapping	Exposes which local candidate is closest to the corrected Manifold token.	Remains a discrete surrogate and cannot fully recover the continuous state.

Table 18: Two decoding-based interpretability procedures used to inspect Manifold Thinking traces.

Statistic	Top-1 mapping	Manifold-nearest Top-20 mapping
Complete traces	10	10
Mean readability score	4.18 / 5.0	4.31 / 5.0
Mean coherence score	4.04 / 5.0	4.22 / 5.0
Mean task-relevance score	4.10 / 5.0	4.27 / 5.0
High-score rate (> 4)	80%	90%
Human-understandable rate	100%	100%

Table 19: Manual interpretability audit of mapped Manifold Thinking traces. Scores summarize human judgments over decoded surrogate traces rather than direct access to continuous states.

J.2 Manual Audit Protocol

Members of the author team conduct a small manual audit over 10 complete mapped reasoning traces sampled from the mathematical and code-generation evaluation suite; no external participants are recruited. For each trace, the evaluators inspect both the Top-1 greedy mapping and the Manifold-nearest Top-20 mapping. The evaluation uses four criteria on a 1–5 scale: semantic readability, step-to-step coherence, task relevance, and usefulness for locating the reasoning decision that leads to success or failure. We also record whether the mapped trace is human-understandable, meaning that an evaluator can follow the main reasoning intent without access to the original continuous vectors.

The audit suggests that Manifold Thinking can be made inspectable through discrete surrogate traces. Top-1 greedy mapping gives a stable conservative view of the model’s default reasoning path. The Manifold-nearest Top-20 mapping is slightly more informative for engineering analysis because it can expose sub-dominant local candidates selected by the geometry-corrected state. This makes the analysis useful not only as a readability check but also as a debugging tool: when a run succeeds or fails, the two mappings let us compare the default greedy interpretation with the candidate favored by the corrected geometry, revealing whether the method explores a plausible local alternative or merely perturbs the trace. At the same time, both mappings remain approximations: they support human inspection and failure analysis, but they should not be interpreted as exact textual reconstructions of the underlying continuous reasoning process.

K Prompting Templates

We use standardized prompt templates across benchmarks so that differences in performance reflect the reasoning mechanism rather than prompt engineering.

General Mathematics (GSM8K, MATH500, OlympiadBench-Math)

Please reason step by step, and put your final answer within `\boxed{}`.
{Question}

Multiple-Choice Reasoning (AMC23)

Please solve the following multiple-choice question. Please show your choice in the answer field with only the choice letter, e.g., "answer": "C".
{Question}

Zero-Shot Code Generation (HumanEval)

Please solve the programming task below in Python. Code should be wrapped in a markdown code block.
```python  
{Question}  
```

Functional Programming with Constraints (MBPP)

Please solve the programming task with test cases below in Python. Make sure your code satisfies the following requirements:
The function name and signature must match exactly as specified in the test cases.
Your code should be wrapped in a markdown code block without including any test cases.
Task:
{Question}
Test Cases:
```python  
{TestCases}  
```

Competitive Programming (LiveCodeBench)

You will be given a question (problem specification) and will generate a correct Python program that matches the specification and passes all tests.
Question: {Question}
If starter code exists:
You will use the following starter code to write the solution and enclose your code within delimiters.
```python  
{StarterCode}  
```  
Otherwise:
Read the inputs from stdin, solve the problem, and write the answer to stdout.
```python  
# YOUR CODE HERE  
```

L Inference Overhead Analysis

Theoretical overhead analysis. Relative to standard greedy decoding, soft-token reasoning methods construct a continuous embedding from the local Top- K candidate set. Let K denote the candidate width and d the embedding dimension. Greedy CoT feeds back only the Top-1 embedding, so the additional vector construction beyond the ordinary model forward pass is $O(d)$ and independent of K . Vanilla Soft Thinking retrieves K candidate embeddings and computes a probability-weighted

sum, giving $O(Kd)$ extra work per decoding step. Stochastic Soft Thinking, i.e., the random soft-token variant, first perturbs and normalizes the Top- K logits with Gumbel noise, an $O(K)$ operation, and then performs the same $O(Kd)$ weighted aggregation; since the embedding operation dominates, its per-step overhead is also $O(Kd)$. Manifold Thinking uses the same Top- K neighborhood and adds tangent orthogonalization, tangent aggregation, and norm re-anchoring, each linear in the K candidate embeddings, so its per-step overhead remains $O(Kd)$. Thus Vanilla Soft Thinking, Stochastic Soft Thinking, and Manifold Thinking share the same asymptotic order and differ mainly by constant factors; in practice, wall-clock overhead also depends on memory access and synchronization, especially AllReduce in tensor-parallel settings. This comparison isolates the incremental soft-token construction cost; the dominant transformer forward pass is shared by all four decoding paths. We measure end-to-end latency and throughput on $8 \times$ NVIDIA RTX 4090 GPUs using 100 randomly sampled problems from GSM8K and HumanEval with batch size 1.

Benchmark	Method	Tok./s	Latency	Overhead
GSM8K	Greedy CoT	$\sim 1.9k$	~ 0.12 s/sample	$1.00 \times$
	Vanilla ST	$\sim 1.1k$	~ 0.26 s/sample	+116.70%
	Manifold	$\sim 1.1k$	~ 0.27 s/sample	+125.00%
HumanEval	Greedy CoT	$\sim 1.3k$	~ 0.20 s/sample	$1.00 \times$
	Vanilla ST	$\sim 1.1k$	~ 0.23 s/sample	+15.00%
	Manifold	$\sim 0.94k$	~ 0.25 s/sample	+25.00%

Table 20: End-to-end inference overhead on $8 \times$ NVIDIA RTX 4090 GPUs.

Phase	Operation	$\mu\text{s}/\text{step}$
1–2	Lookup + baseline AllReduce	18.4
3	Top-1 anchor AllReduce	4.1
4	Tangent Navigation + AllReduce	9.7
5	Manifold Projection	2.3
Total	Added phases	34.5
Base	Greedy embedding lookup	313.2

Table 21: Per-step overhead breakdown for Manifold Thinking.

Vanilla Soft Thinking and Manifold Thinking have comparable overhead profiles because both require embedding retrieval and weighted aggregation. Relative to Greedy CoT, soft-token methods introduce a visible end-to-end latency cost because they perform embedding retrieval and weighted aggregation at every decoding step. The incremental gap between Vanilla Soft Thinking and Manifold

Thinking is much smaller: Manifold increases latency by about 3.80% over Vanilla ST on GSM8K and 8.70% on HumanEval. Table 20 reports the end-to-end throughput, latency, and overhead measurements used in this interpretation. Table 21 shows that Tangent Navigation is the dominant incremental component, while norm alignment is lightweight. This one-trajectory overhead remains substantially more favorable than multi-sample methods such as Self-Consistency, which incur $n \times$ full generation cost for SC@ n .

Practical interpretation. The overhead is additive rather than multiplicative. Manifold Thinking does not request additional complete model calls, construct external search trees, or train an auxiliary controller; it inserts a small number of vector operations inside the same decoding trajectory. This distinction is important for deployment because the expensive part of inference remains the backbone forward pass, while the added work is localized to the current Top- K candidate set and embedding dimension. The tensor-parallel synchronization cost in Table 21 is also bounded: the method needs global agreement on the baseline vector, the Top-1 anchor, and the tangent innovation, but it does not introduce iterative branch scoring or answer-level voting.

Optimization note. The current implementation is intentionally direct, so the reported overhead should be read as a conservative engineering estimate rather than a lower bound. Embedding lookup, tangent projection, and norm alignment can be fused with the soft-token construction kernel, and the Top- K embeddings are already available in many decoding stacks for sampling or logit filtering. Thus, the appendix evidence supports the main experimental framing: the method pays a small per-step geometric correction cost to avoid the much larger compute profile of multi-trajectory inference.

Connection to cost-normalized results. Together with Appendix E, these measurements explain why our main tables emphasize single-trajectory baselines.

M Additional Downstream Analysis

M.1 Supplementary LiveCodeBench Results

The LCB runs use QwQ-32B, Llama-3.1-8B, and DeepSeek-R1-Distill-Qwen-32B rather than the matched Qwen2.5-7B/Llama-3.1-8B/DeepSeek-R1-Distill-Qwen-7B grid used in Table 3. We

Model	Greedy CoT	Pure CoT	Vanilla Soft Thinking	Stochastic Soft Thinking	Manifold Thinking	Δ	p -value
QwQ-32B	57.35 $\pm$ 0.00	62.01 $\pm$ 0.36	62.72 $\pm$ 0.51	64.87 $\pm$ 0.84	72.40 $\pm$ 0.25	+7.53	< 0.001
Llama-3.1-8B	22.22 $\pm$ 0.00	24.37 $\pm$ 0.25	26.88 $\pm$ 0.36	<u>27.96 $\pm$ 0.76</u>	30.11 $\pm$ 0.25	+2.15	0.001
DeepSeek-R1-Distill-Qwen-32B	43.37 $\pm$ 0.00	57.35 $\pm$ 0.25	59.50 $\pm$ 0.36	<u>62.01 $\pm$ 0.67</u>	67.74 $\pm$ 0.25	+5.73	< 0.001

Table 22: Supplementary LCB pass@1 results over five matched seeds within each model. Bold marks the best result and underlining the strongest non-Manifold baseline. The three heterogeneous backbones are not pooled into the matched primary-table tally.

Method	LCB pass@1 (%)
Manifold Thinking	72.40 $\pm$ 0.25
– w/o Tangent Nav.	61.29 $\pm$ 0.72
– w/o Manifold Proj.	68.10 $\pm$ 0.57
Stochastic Soft Thinking	64.87 $\pm$ 0.84

Table 23: Supplementary LCB ablation on QwQ-32B over five seeds.

therefore report them as a transparent supplementary comparison and exclude them from the 18-configuration primary-table tally. Within each LCB model, all methods use five matched seeds and the same evaluation split.

When accuracy is close, an inference method that preserves the one-pass budget is preferable to a method that obtains similar gains by multiplying full generations. When Manifold Thinking wins outright, the advantage is stronger: it improves the continuous reasoning state while staying inside the same latency regime as Vanilla Soft Thinking. This makes the method a practical add-on for settings where throughput, memory pressure, and distributed synchronization are first-order constraints rather than secondary implementation details.

M.2 Statistical Interpretation of Mean Improvements

Manifold Thinking has the highest mean score in all 18 matched model–dataset configurations in Table 3, but this statement is distinct from claiming significance in every row. Of the 17 primary comparisons for which a paired test is defined, 10 reach $p < 0.05$ and seven show positive but non-significant mean trends. The remaining DeepSeek-R1-Distill-Qwen-7B/MATH500 row is a degenerate paired comparison: relative to the strongest non-Manifold baseline, Manifold Thinking solves exactly one additional problem under every matched seed, so each paired difference is $+1/500$ and the paired-difference variance is zero. A two-sided paired t -test cannot estimate a standard error in that case, and the dash in Table 3 denotes an undefined test rather than a missing value. All three supplementary LCB comparisons in Table 22 also reach $p < 0.05$; including them gives 13 significant results, seven positive but non-significant trends, and one degenerate comparison across all 21 reported configurations, while preserving the primary/sup-

plementary scope distinction.

AMC23 illustrates small-sample sensitivity. It contains 40 problems, so one additional correct answer changes accuracy by 2.5 percentage points. Manifold Thinking has a positive mean gain for all three models, but none of the AMC23 paired comparisons reaches the significance threshold; these rows should therefore be read as positive mean trends rather than strong significance evidence. Several GSM8K and OlympiadBench-Math rows likewise have positive but variable seed-level differences against the strongest stochastic baseline. The primary-table claim is thus a consistent ordering of mean performance, with statistically significant evidence in most non-degenerate comparisons and directional evidence in the remaining rows.

M.3 Per-Seed Result Audit

Tables 24–28 and Table 30 provide the per-seed results underlying Table 3; Table 29 underlies the supplementary LCB aggregate in Table 22. Rows for the same model–dataset pair use the same seed sequence across methods, keeping the multi-run comparison aligned. The greedy rows are deterministic, sampled baselines fluctuate across seeds, and Manifold Thinking keeps its average advantage over the strongest stochastic soft-token baseline while remaining within a single forward trajectory. Across individual seeds, several close comparisons fluctuate locally; nevertheless, the mean results remain aligned with the aggregate improvements reported in Table 3 and the separately scoped LCB table. The purpose of this audit is not to claim that every seed favors the same method on every dataset; rather, it makes the variability explicit. By exposing correct counts rather than only percentages, the table helps distinguish one-problem fluctuations on small benchmarks from robust improvements on larger or longer-generation settings.

Model	Method	S1	S2	S3	S4	S5	Mean $\pm$ Std (%)
Qwen2.5-7B	Greedy CoT	1221/1319	1221/1319	1221/1319	1221/1319	1221/1319	92.57 $\pm$ 0.00
	Pure CoT	1234/1319	1216/1319	1227/1319	1227/1319	1216/1319	92.80 $\pm$ 0.59
	Vanilla Soft Thinking	1227/1319	1199/1319	1216/1319	1218/1319	1195/1319	91.81 $\pm$ 1.02
	Stochastic Soft Thinking	1253/1319	1207/1319	1237/1319	1240/1319	1203/1319	93.10 $\pm$ 1.66
	Manifold Thinking	1244/1319	1228/1319	1238/1319	1238/1319	1227/1319	93.63 $\pm$ 0.55
Llama-3.1-8B	Greedy CoT	1134/1319	1134/1319	1134/1319	1134/1319	1134/1319	85.97 $\pm$ 0.00
	Pure CoT	1120/1319	1141/1319	1124/1319	1130/1319	1130/1319	85.60 $\pm$ 0.60
	Vanilla Soft Thinking	1090/1319	1121/1319	1101/1319	1123/1319	1105/1319	84.00 $\pm$ 1.06
	Stochastic Soft Thinking	1097/1319	1146/1319	1113/1319	1143/1319	1121/1319	85.22 $\pm$ 1.56
	Manifold Thinking	1137/1319	1156/1319	1141/1319	1145/1319	1146/1319	86.81 $\pm$ 0.54
DeepSeek-R1-Distill-Qwen-7B	Greedy CoT	1198/1319	1198/1319	1198/1319	1198/1319	1198/1319	90.83 $\pm$ 0.00
	Pure CoT	1193/1319	1211/1319	1194/1319	1198/1319	1204/1319	90.98 $\pm$ 0.57
	Vanilla Soft Thinking	1193/1319	1224/1319	1197/1319	1208/1319	1208/1319	91.43 $\pm$ 0.91
	Stochastic Soft Thinking	1188/1319	1237/1319	1197/1319	1220/1319	1208/1319	91.74 $\pm$ 1.46
	Manifold Thinking	1217/1319	1234/1319	1218/1319	1225/1319	1226/1319	92.80 $\pm$ 0.52

Table 24: Per-seed correct counts for GSM8K (1319 problems).

Model	Method	S1	S2	S3	S4	S5	Mean $\pm$ Std (%)
Qwen2.5-7B	Greedy CoT	379/500	379/500	379/500	379/500	379/500	75.80 $\pm$ 0.00
	Pure CoT	379/500	397/500	380/500	384/500	390/500	77.20 $\pm$ 1.50
	Vanilla Soft Thinking	352/500	383/500	356/500	367/500	367/500	73.00 $\pm$ 2.41
	Stochastic Soft Thinking	361/500	410/500	370/500	393/500	381/500	76.60 $\pm$ 3.85
	Manifold Thinking	387/500	404/500	388/500	395/500	396/500	78.80 $\pm$ 1.38
Llama-3.1-8B	Greedy CoT	244/500	244/500	244/500	244/500	244/500	48.80 $\pm$ 0.00
	Pure CoT	245/500	227/500	236/500	226/500	236/500	46.80 $\pm$ 1.56
	Vanilla Soft Thinking	246/500	215/500	235/500	217/500	232/500	45.80 $\pm$ 2.60
	Stochastic Soft Thinking	262/500	213/500	245/500	224/500	246/500	47.60 $\pm$ 3.89
	Manifold Thinking	265/500	249/500	257/500	248/500	256/500	51.00 $\pm$ 1.38
DeepSeek-R1-Distill-Qwen-7B	Greedy CoT	437/500	437/500	437/500	437/500	437/500	87.40 $\pm$ 0.00
	Pure CoT	457/500	458/500	456/500	457/500	457/500	91.40 $\pm$ 0.14
	Vanilla Soft Thinking	441/500	438/500	440/500	442/500	439/500	88.00 $\pm$ 0.32
	Stochastic Soft Thinking	457/500	451/500	455/500	457/500	450/500	90.80 $\pm$ 0.66
	Manifold Thinking	458/500	459/500	457/500	458/500	458/500	91.60 $\pm$ 0.14

Table 25: Per-seed correct counts for MATH500 (500 problems).

Model	Method	S1	S2	S3	S4	S5	Mean $\pm$ Std (%)
Qwen2.5-7B	Greedy CoT	25/40	25/40	25/40	25/40	25/40	62.50 $\pm$ 0.00
	Pure CoT	18/40	19/40	17/40	18/40	18/40	45.00 $\pm$ 1.77
	Vanilla Soft Thinking	25/40	22/40	24/40	26/40	23/40	60.00 $\pm$ 3.95
	Stochastic Soft Thinking	26/40	20/40	24/40	26/40	19/40	57.50 $\pm$ 8.29
	Manifold Thinking	27/40	26/40	25/40	27/40	25/40	65.00 $\pm$ 2.50
Llama-3.1-8B	Greedy CoT	9/40	9/40	9/40	9/40	9/40	22.50 $\pm$ 0.00
	Pure CoT	12/40	11/40	11/40	10/40	11/40	27.50 $\pm$ 1.77
	Vanilla Soft Thinking	3/40	6/40	7/40	4/40	5/40	12.50 $\pm$ 3.95
	Stochastic Soft Thinking	7/40	13/40	12/40	6/40	12/40	25.00 $\pm$ 8.10
	Manifold Thinking	12/40	13/40	11/40	12/40	12/40	30.00 $\pm$ 1.77
DeepSeek-R1-Distill-Qwen-7B	Greedy CoT	28/40	28/40	28/40	28/40	28/40	70.00 $\pm$ 0.00
	Pure CoT	30/40	28/40	29/40	29/40	29/40	72.50 $\pm$ 1.77
	Vanilla Soft Thinking	33/40	30/40	29/40	32/40	31/40	77.50 $\pm$ 3.95
	Stochastic Soft Thinking	32/40	26/40	33/40	27/40	32/40	75.00 $\pm$ 8.10
	Manifold Thinking	33/40	31/40	33/40	32/40	31/40	80.00 $\pm$ 2.50

Table 26: Per-seed correct counts for AMC23 (40 problems).

Reading guide. GSM8K is the lowest-variance math setting, while AMC23 is deliberately high-variance because one problem corresponds to 2.50 percentage points. The supplementary LCB comparisons use QwQ-32B, Llama-3.1-8B, and DeepSeek-R1-Distill-Qwen-32B and are all significant, but their heterogeneous backbone set means they are interpreted separately from the matched primary grid. OlympiadBench-Math supplies a harder math benchmark; scores are lower than GSM8K and MATH500, but the method ranking

remains aligned with the main evaluation.

Model	Method	S1	S2	S3	S4	S5	Mean $\pm$ Std (%)
Qwen2.5-7B	Greedy CoT	126/164	126/164	126/164	126/164	126/164	76.83 $\pm$ 0.00
	Pure CoT	125/164	124/164	124/164	123/164	124/164	75.61 $\pm$ 0.43
	Vanilla Soft Thinking	126/164	124/164	126/164	125/164	124/164	76.22 $\pm$ 0.61
	Stochastic Soft Thinking	127/164	123/164	126/164	126/164	123/164	76.22 $\pm$ 1.14
	Manifold Thinking	128/164	127/164	126/164	127/164	127/164	77.44 $\pm$ 0.43
Llama-3.1-8B	Greedy CoT	88/164	88/164	88/164	88/164	88/164	53.66 $\pm$ 0.00
	Pure CoT	89/164	90/164	88/164	89/164	89/164	54.27 $\pm$ 0.43
	Vanilla Soft Thinking	87/164	84/164	85/164	85/164	84/164	51.83 $\pm$ 0.75
	Stochastic Soft Thinking	85/164	90/164	86/164	89/164	85/164	53.05 $\pm$ 1.43
	Manifold Thinking	90/164	91/164	90/164	89/164	90/164	54.88 $\pm$ 0.43
DeepSeek-R1-Distill-Qwen-7B	Greedy CoT	110/164	110/164	110/164	110/164	110/164	67.07 $\pm$ 0.00
	Pure CoT	117/164	115/164	116/164	117/164	115/164	70.73 $\pm$ 0.61
	Vanilla Soft Thinking	117/164	119/164	120/164	117/164	117/164	71.95 $\pm$ 0.86
	Stochastic Soft Thinking	124/164	119/164	123/164	120/164	119/164	73.78 $\pm$ 1.43
	Manifold Thinking	126/164	124/164	125/164	125/164	125/164	76.22 $\pm$ 0.43

Table 27: Per-seed correct counts for HumanEval (164 problems).

Model	Method	S1	S2	S3	S4	S5	Mean $\pm$ Std (%)
Qwen2.5-7B	Greedy CoT	183/257	183/257	183/257	183/257	183/257	71.21 $\pm$ 0.00
	Pure CoT	182/257	183/257	181/257	182/257	182/257	70.82 $\pm$ 0.28
	Vanilla Soft Thinking	187/257	184/257	185/257	185/257	184/257	71.98 $\pm$ 0.48
	Stochastic Soft Thinking	186/257	191/257	187/257	190/257	186/257	73.15 $\pm$ 0.91
	Manifold Thinking	196/257	197/257	196/257	195/257	196/257	76.26 $\pm$ 0.28
Llama-3.1-8B	Greedy CoT	168/257	168/257	168/257	168/257	168/257	65.37 $\pm$ 0.00
	Pure CoT	166/257	164/257	165/257	166/257	164/257	64.20 $\pm$ 0.39
	Vanilla Soft Thinking	163/257	165/257	166/257	163/257	163/257	63.81 $\pm$ 0.55
	Stochastic Soft Thinking	172/257	167/257	171/257	168/257	167/257	65.76 $\pm$ 0.91
	Manifold Thinking	176/257	174/257	175/257	175/257	175/257	68.09 $\pm$ 0.28
DeepSeek-R1-Distill-Qwen-7B	Greedy CoT	195/257	195/257	195/257	195/257	195/257	75.88 $\pm$ 0.00
	Pure CoT	187/257	187/257	188/257	186/257	187/257	72.76 $\pm$ 0.28
	Vanilla Soft Thinking	193/257	193/257	191/257	192/257	191/257	74.71 $\pm$ 0.39
	Stochastic Soft Thinking	196/257	193/257	196/257	191/257	194/257	75.49 $\pm$ 0.83
	Manifold Thinking	199/257	199/257	200/257	198/257	199/257	77.43 $\pm$ 0.28

Table 28: Per-seed correct counts for MBPP (257 problems).

Model	Method	S1	S2	S3	S4	S5	Mean $\pm$ Std (%)
QwQ-32B	Greedy CoT	160/279	160/279	160/279	160/279	160/279	57.35 $\pm$ 0.00
	Pure CoT	174/279	172/279	173/279	174/279	172/279	62.01 $\pm$ 0.36
	Vanilla Soft Thinking	174/279	176/279	177/279	174/279	174/279	62.72 $\pm$ 0.51
	Stochastic Soft Thinking	184/279	179/279	183/279	180/279	179/279	64.87 $\pm$ 0.84
	Manifold Thinking	203/279	201/279	202/279	202/279	202/279	72.40 $\pm$ 0.25
Llama-3.1-8B	Greedy CoT	62/279	62/279	62/279	62/279	62/279	22.22 $\pm$ 0.00
	Pure CoT	68/279	68/279	69/279	67/279	68/279	24.37 $\pm$ 0.25
	Vanilla Soft Thinking	76/279	76/279	74/279	75/279	74/279	26.88 $\pm$ 0.36
	Stochastic Soft Thinking	80/279	77/279	80/279	75/279	78/279	27.96 $\pm$ 0.76
	Manifold Thinking	84/279	84/279	85/279	83/279	84/279	30.11 $\pm$ 0.25
DeepSeek-R1-Distill-Qwen-32B	Greedy CoT	121/279	121/279	121/279	121/279	121/279	43.37 $\pm$ 0.00
	Pure CoT	161/279	160/279	160/279	159/279	160/279	57.35 $\pm$ 0.25
	Vanilla Soft Thinking	167/279	165/279	167/279	166/279	165/279	59.50 $\pm$ 0.36
	Stochastic Soft Thinking	175/279	171/279	174/279	174/279	171/279	62.01 $\pm$ 0.67
	Manifold Thinking	190/279	189/279	188/279	189/279	189/279	67.74 $\pm$ 0.25

Table 29: Supplementary per-seed correct counts for LCB (279 problems) using the separately reported backbone set.

Model	Method	S1	S2	S3	S4	S5	Mean $\pm$ Std (%)
Qwen2.5-7B	Greedy CoT	240/674	240/674	240/674	240/674	240/674	35.61 $\pm$ 0.00
	Pure CoT	256/674	238/674	253/674	253/674	240/674	36.80 $\pm$ 1.24
	Vanilla Soft Thinking	256/674	225/674	250/674	246/674	233/674	35.91 $\pm$ 1.89
	Stochastic Soft Thinking	274/674	227/674	271/674	260/674	243/674	37.83 $\pm$ 2.94
	Manifold Thinking	270/674	253/674	266/674	266/674	255/674	38.87 $\pm$ 1.12
Llama-3.1-8B	Greedy CoT	160/674	160/674	160/674	160/674	160/674	23.74 $\pm$ 0.00
	Pure CoT	170/674	153/674	171/674	161/674	165/674	24.33 $\pm$ 1.09
	Vanilla Soft Thinking	170/674	142/674	167/674	150/674	161/674	23.44 $\pm$ 1.75
	Stochastic Soft Thinking	187/674	145/674	183/674	159/674	171/674	25.07 $\pm$ 2.57
	Manifold Thinking	181/674	166/674	180/674	172/674	176/674	25.96 $\pm$ 0.91
DeepSeek-R1-Distill-Qwen-7B	Greedy CoT	268/674	268/674	268/674	268/674	268/674	39.76 $\pm$ 0.00
	Pure CoT	287/674	269/674	278/674	268/674	278/674	40.95 $\pm$ 1.15
	Vanilla Soft Thinking	289/674	258/674	278/674	260/674	275/674	40.36 $\pm$ 1.93
	Stochastic Soft Thinking	298/674	249/674	281/674	260/674	282/674	40.65 $\pm$ 2.88
	Manifold Thinking	293/674	277/674	285/674	276/674	284/674	41.99 $\pm$ 1.02

Table 30: Per-seed correct counts for OlympiadBench-Math (674 problems).