BENDING THE THOUGHT: CURVATURE-GUIDED REINFORCEMENT LEARNING FOR LLM REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards (RLVR) improves large language model reasoning using only final-answer feedback. This raises a token-level credit assignment problem: only a small subset of local computations may account for the final result, while the reward fails to indicate the locations of such computations. Existing token-level credit allocation researches, such as token entropy or update magnitude, are mostly defined on the policy side and mainly capture uncertainty or distributional change in the probability space, while neglecting the model's internal computational processes. Inspired by the cognitive neural-manifold, we view reasoning as a trajectory in the hidden representation space. Under a smooth-trajectory idealization, we derive an upper bound on the second-order variation of a smooth scalar readout that includes a curvature-dependent term. This motivates us to treat curvature as a candidate signal for token-level update credit assignment. Building on this analysis, we introduce ARC (Adaptive Reweighting by Curvature), a curvature-guided RLVR framework that converts response-normalized curvature scores into bounded, detached weights on the token-level policy surrogate. Empirical analysis reveals a correlation between high-curvature positions and reasoning markers, alongside layer-dependent differences in entropy-based token selections. Evaluations on six mathematical reasoning benchmarks show improved performance over the baseline for both Qwen3-8B-Base and Qwen3-14B-Base, including gains of 5.98 and 8.27 percentage points on AIME 2024, respectively.

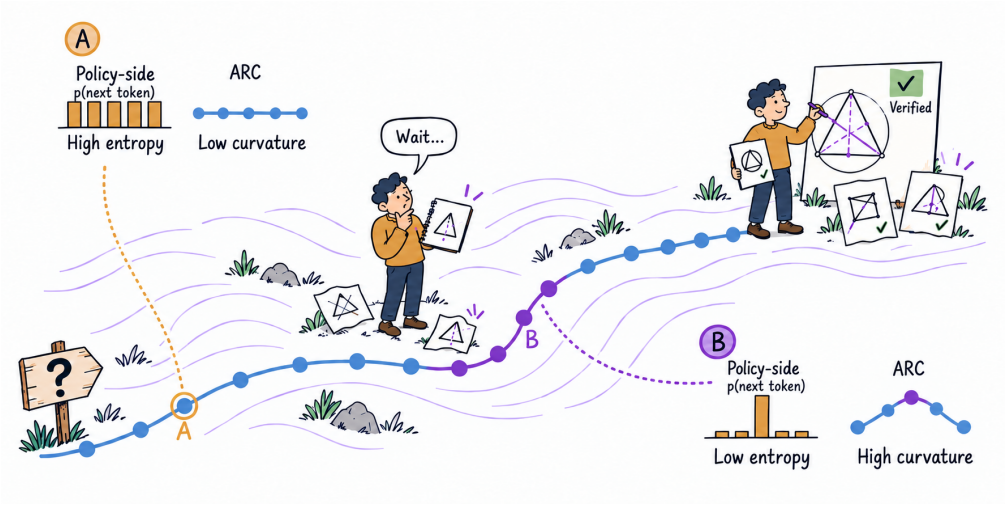

Figure 1: Reasoning corresponds to high-curvature transitions on a latent trajectory, while low-curvature regions reflect smooth progression.

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

1 INTRODUCTION

Outcome supervision has made reinforcement learning for reasoning models scalable. Instead of relying on dense annotations of intermediate reasoning steps, reinforcement learning with verifiable rewards (RLVR) trains language models from automatically checkable final outcomes, such as mathematical correctness or executable program results. This paradigm has been central to recent large reasoning models, including OpenAI o1, DeepSeek-R1, and Qwen3 (Jaech et al., 2024; Guo et al., 2025; Yang et al., 2025), and enables learning from sampled reasoning traces rather than human-written solutions (Shao et al., 2024; Yu et al., 2026a). However, outcome supervision leaves a token-level credit assignment problem: a verifier evaluates the completed response, while standard objectives apply a shared response-level advantage across its tokens (Shao et al., 2024). This does not explicitly distinguish tokens with varying contributions to the final outcome and constitutes an instance of temporal credit assignment (Sutton et al., 1998).

Recent works have begun to reveal the sparsity of useful RLVR updates. High-entropy-token methods argue that a small subset of uncertain tokens acts as reasoning forks, and show that updating only these tokens can preserve or even improve RLVR performance (Wang et al., 2026). Prefix-based optimization emphasizes that early tokens can strongly constrain subsequent reasoning trajectories (Sun et al., 2026b). Update-direction analysis further suggests that signed token-level probability shifts provide informative coordinates for identifying where RLVR changes model behavior (Huang et al., 2026). Together, these findings suggest that verifier-derived learning signals should be allocated selectively rather than uniformly across response tokens. Yet the existing criteria primarily describe output uncertainty, output changes, or token position; they do not directly reveal how the model's underlying computation evolves as it reasons. This motivates us to look beyond output-level statistics and ask whether the model's internal representations can provide a more direct signal of token importance.

The internal representation of the model offers an additional perspective. Existing criteria characterize token importance through output uncertainty, probability changes, or response position, but do not directly track how the model's internal state evolves as reasoning unfolds. At each generation step, the transformer updates its hidden representation, producing a sequence of states that traces a trajectory through representation space. Recent studies suggest that hidden representations and their trajectories carry information about belief states, semantic uncertainty, and reasoning structure that may not be fully captured by output-distribution statistics (Shai et al., 2024; Kossen et al., 2024; Sun et al., 2026a; Damirchi et al., 2026). If reasoning proceeds as an evolving internal process, then a reasoning transition should appear not only in the token distribution, but also in how this trajectory evolves over time. This raises our central question: do turning points in reasoning coincide with turning points in the model's internal trajectory?

Once the hidden states are viewed as a trajectory, changes in the model's ongoing computation can be characterized by changes in the trajectory's local direction. A smooth progression corresponds to a relatively stable direction of evolution, whereas a pause, revision, or shift in the line of reasoning may produce a local bend in the hidden-state path. We refer to these local changes in trajectory direction as hidden turns and use local curvature to quantify their strength. This trajectory-based view is also related to neuroscience, where population activity is often modeled as low-dimensional trajectories whose geometry reflects computation and behavior (Saxena & Cunningham, 2019; Vyas et al., 2020; Gallego et al., 2020; Perich et al., 2025).

As illustrated in Fig. 1, policy-side methods focus on how uncertain the model is about the next token, whereas ARC follows the evolution of the reasoning trajectory itself. A token can have high entropy while the hidden trajectory continues smoothly, so uncertainty alone need not indicate a change in reasoning direction. ARC instead assigns greater importance to high-curvature locations, where the internal trajectory bends sharply and where the model may be pausing, revising, or shifting its line of reasoning.

We quantify hidden turns using discrete trajectory curvature. Under a smooth-trajectory idealization, our analysis derives an upper bound on the variation of a semantic readout that includes a curvature-dependent term, motivating curvature as a candidate allocation signal. We further examine its alignment with reasoning markers, its distribution across response positions, and its overlap with entropy-based selections.

Building on these observations, we propose ARC (Adaptive Reweighting by Curvature), which maps response-normalized curvature scores to bounded weights on the token-level policy surrogate. Evaluations on Qwen3-8B-Base and Qwen3-14B-Base across six mathematical reasoning benchmarks show improvements over the baseline.

Our contributions are summarized as follows:

- We introduce hidden turns, quantified by discrete trajectory curvature, as a representation-based signal for RLVR update allocation, and analyze their relationships with reasoning markers and entropy-based selections.
- We derive a semantic variation bound and relate it to ambient trajectory curvature, providing a geometric motivation for the allocation signal.
- We propose ARC, which uses hidden-trajectory curvature to weight token-level policy updates, and demonstrate its effectiveness across two model scales and six mathematical reasoning benchmarks.

2 Preliminaries: Dynamic Sampling Policy Optimization (DAPO)

We first introduce the DAPO Yu et al. (2026a) objective. Given a prompt q and ground-truth answer a, DAPO samples a group of responses $\{o^i\}_{i=1}^G$ from the old policy $\pi_{\theta_{\text{old}}}(\cdot \mid q)$. We denote

$$r_t^i(\theta) = \frac{\pi_\theta(o_t^i \mid q, o_{<t}^i)}{\pi_{\theta_{\text{old}}}(o_t^i \mid q, o_{<t}^i)}$$

be the token-level probability ratio, and let R^i denote the verifier reward of the i-th sampled response o^i. DAPO uses a group-relative advantage computed within the sampled group:

$$\hat{A}_t^i = \frac{R^i - \texttt{mean}(\{R^i\}_{i=1}^G)}{\texttt{std}(\{R^i\}_{i=1}^G)}. \tag{1}$$

For outcome-only RLVR, this rollout-level advantage is assigned to all valid response tokens, i.e., $\hat{A}_t^i = \hat{A}^i$ for $t = 1, \ldots, |o^i|$, where $|o^i|$ denotes the response length. DAPO removes the KL penalty used in earlier GRPO-style objectives, adopts dynamic sampling, uses token-level policy-gradient aggregation, and introduces a decoupled clipping range with separate lower and upper coefficients ϵ_{low} and ϵ_{high}. Its maximization objective can be written as

$$\mathcal{J}_{\text{DAPO}}(\theta) = \mathbb{E}_{\mathcal{B},(q,a),\{o^i\}_{i=1}^G} \left[\frac{1}{\sum_{i=1}^G |o^i|} \sum_{i=1}^G \sum_{t=1}^{|o^i|} \omega_t^i \min\left(r_t^i(\theta)\hat{A}_t^i,\ \text{clip}\big(r_t^i(\theta), 1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}}\big)\hat{A}_t^i \right) \right], \tag{2}$$

subject to the dynamic-sampling condition

$$0 < \left| \{ o^i \mid \text{is_equivalent}(a, o^i) \} \right| < G, \tag{3}$$

where token weights $\omega_t^i \geq 0$, uniform token aggregation corresponds to $\omega_t^i = 1$. The condition in Eq. 3 filters out groups in which all sampled responses are either correct or incorrect, so that the remaining groups contain non-degenerate reward variation for estimating the advantage. In this work, we use DAPO as the baseline RLVR algorithm.

3 Curvature as a Second-Order Semantic Variation

3.1 Discrete trajectory curvature

For a response containing T valid tokens, let $h_t^l \in \mathbb{R}^m$ be its hidden representation at position t in layer l. We treat $\{h_t^l\}_{t=1}^T$ as a discrete trajectory in the ambient representation space. A raw second difference depends on the parametrization of the trajectory. We therefore follow the arc-length

parametrization principle from differential geometry (Do Carmo, 2016; Lee, 2018) and estimate curvature as the change of unit tangent direction per unit local arc length. At the layer l, define:

$$\Delta h_t^l = h_t^l - h_{t-1}^l, \qquad \ell_t^l = \|\Delta h_t^l\|_2 + \epsilon, \qquad \tau_t^l = \frac{\Delta h_t^l}{\ell_t^l}, \tag{4}$$

where Δh_t^l, ℓ_t^l and τ_t^l represents the raw tangent vector, the norm of raw tangent vector and the unit tangent vector at token t respectively.

For $2 \le t \le T-1$, the trajectory curvature is

$$\kappa_t^l = \frac{\|\tau_{t+1}^l - \tau_t^l\|_2}{(\ell_{t+1}^l + \ell_t^l)/2 + \epsilon}. \tag{5}$$

Unlike entropy, which describes uncertainty in the next-token distribution, curvature describes how the hidden trajectory bends and can therefore reveal latent turning points that need not coincide with high output uncertainty.

To make this interpretation reliable, we remove degenerate local steps that could create numerical curvature artifacts. When adjacent hidden states are nearly identical, the local arc-length denominator becomes close to zero, producing spuriously large curvature values that do not correspond to meaningful hidden turns. We therefore compute the 5th percentile δ^l length of all finite nonzero step lengths for each trajectory, and evaluate curvature at position t only when both adjacent steps exceed δ^l; otherwise the score is set to zero. Boundary values are copied from the nearest interior positions.

To connect these geometric events to semantic changes in reasoning, we analyze smooth semantic functions on the hidden manifold. The following proposition shows that curvature controls a trajectory-induced component of their second-order variation, thereby providing a theoretical basis for treating high-curvature positions as candidate reasoning turning points.

3.2 Curvature as a semantic-change signal

The reason to look at curvature is simple: if a smooth readout of the hidden state tracks a property such as solution phase, current subgoal, or answer confidence, then a sharp change in the hidden trajectory can contribute to a sharp change in that readout. The following bound formalizes this connection between hidden-trajectory geometry and local semantic variation.

Proposition 1 (Semantic variation bound)**.** *Let $M \subset \mathbb{R}^m$ be a smooth embedded submanifold equipped with the induced Riemannian metric g, and let $\alpha : [0, S] \to M$ be a C^2 curve parameterized by arc length. Let $f \in C^2(U)$, where $U \subset M$ is an open neighborhood of the trajectory. Assume that, along the trajectory,*

$$\|\operatorname{Hess}_M f\| \le L, \qquad \|\operatorname{grad}_M f\| \le G.$$

Define $u(s) = f(\alpha(s))$, $v(s) = \dot\alpha(s)$, the geodesic curvature $\kappa_g(s) = \|\nabla_v^M v\|$, and the ambient curvature $\kappa_{\mathrm{amb}}(s) = \|\ddot\alpha(s)\|$. Then

$$\|u''(s)\| \le L + G \cdot \kappa_g(s) \le L + G \cdot \kappa_{\mathrm{amb}}(s).$$

This proposition gives curvature a principled role in identifying latent turning points. It shows that, for any smooth semantic observer, the second-order variation of the observed semantic quantity along the hidden trajectory is controlled by an intrinsic smoothness term and a curvature-dependent term. Thus, high-curvature positions are natural candidates for locations where the model's latent computation changes rapidly. The proof and a short manifold interpretation are provided in Appendix A.

3.3 Curvature-soft allocation

We use token-level curvature to softly redistribute the verifier-derived policy loss over response tokens. Let $\kappa_{i,t}$ denote the curvature score of token t in response i, and let $\mathcal{R}_i$ denote the set of response-token indices. We first standardize curvature within each response by applying a per-response z-score normalization, producing $z_{i,t}$. The normalized score is then converted into a soft allocation weight:

$$\tilde w_{i,t} = 1 + \gamma \cdot \mathrm{sigmoid}(z_{i,t}), \tag{6}$$

where γ controls the strength of curvature-based reweighting. We normalize the weights to have unit mean within each response, preserving its total allocation weight while changing the relative emphasis across tokens:

$$w_{i,t} = \frac{\tilde{w}_{i,t}}{\frac{1}{|\mathcal{R}_i|}\sum_{s\in\mathcal{R}_i}\tilde{w}_{i,s}}. \tag{7}$$

Given the original token-level policy loss $\ell_{i,t}^{\text{policy}}$, the curvature-soft objective is:

$$\mathcal{L}_{\text{ARC}} = \frac{1}{\sum_i |\mathcal{R}_i|}\sum_i\sum_{t\in\mathcal{R}_i}\text{stopgrad}(w_{i,t})\,\ell_{i,t}^{\text{policy}}. \tag{8}$$

The weights are computed with stop-gradient. Thus, curvature changes where the verifier-derived update is concentrated, while preserving the response-level supervision signal.

4 Experiments

Our experiments are organized around three questions. **First**, do high-curvature positions emerge in specific layers and response regions, and do these positions align with reasoning-relevant parts? **Second**, does curvature provide information beyond policy-side uncertainty signals such as entropy? **Third**, could curvature-guided allocation improve RLVR training performance? The first two questions evaluate curvature as a diagnostic signal before it is used to modify training, while the third tests whether the same signal remains useful when incorporated into the RLVR objective.

4.1 Experiment setup

Diagnostic setup For diagnostic analysis, we generate reasoning traces with Qwen3-8B (Yang et al., 2025) on the Reasoning360 math dataset (Cheng et al., 2026). We sample 200 prompts and generate 8 responses per prompt, yielding 1,600 reasoning traces in total. For each generated trace, we compute token-level curvature scores across layers and compute token entropy scores, which serve as a representative policy-side uncertainty signal. Unless otherwise specified, high-score positions are selected independently for each response and layer by taking the top 20% of tokens under the corresponding score.

We evaluate whether these selected positions capture reasoning-relevant events from four complementary perspectives: response-position enrichment, alignment with reasoning markers, layer-wise behavior, and complementarity with policy-side signals. Marker alignment uses a ± 3 token window around predefined reasoning markers, and confidence intervals are computed by bootstrapping over responses. The marker selection criteria, full marker list, implementation details, and sensitivity analyses are provided in Appendix B. To examine whether the observed patterns are specific to Qwen3, we further repeat the diagnostic analysis on Llama-3.1-8B-Instruct; the resulting trends are highly consistent with the main analysis, with results reported in Appendix E.

RL training. We initialize from Qwen3-8/14B-Base (Yang et al., 2025) and train on the Reasoning360 math dataset (Cheng et al., 2026). We evaluate these methods on six mathematical reasoning benchmarks: AIME'24, AIME'25, AMC'23, MATH500 (Hendrycks et al., 2021), Minerva, and Olympiad (He et al., 2024). Evaluation uses 32 sampled responses per problem to compute average accuracy rate, curvature weights use hyperparameter $\gamma = 0.6$, and for Qwen3-8/14B-Base, we compute the curvature weights at Layer 21/24.

4.2 A First Glance: Curvature Peaks More Land on Reasoning Events

We first test whether high-curvature positions concentrate near reasoning-relevant regions of generated responses. We compare top-curvature tokens with top-policy-uncertainty tokens, low-curvature tokens, shuffled-curvature controls, and random selections, which respectively control for output uncertainty, generic selection effects, response-position bias, and chance. In the main text, we report two diagnostic readouts: overlap with predefined reasoning markers and enrichment over normalized response regions. Marker overlap measures whether selected tokens fall within a fixed window around visible reasoning transitions, such as checking, correction, conclusion, or subgoal shifts.

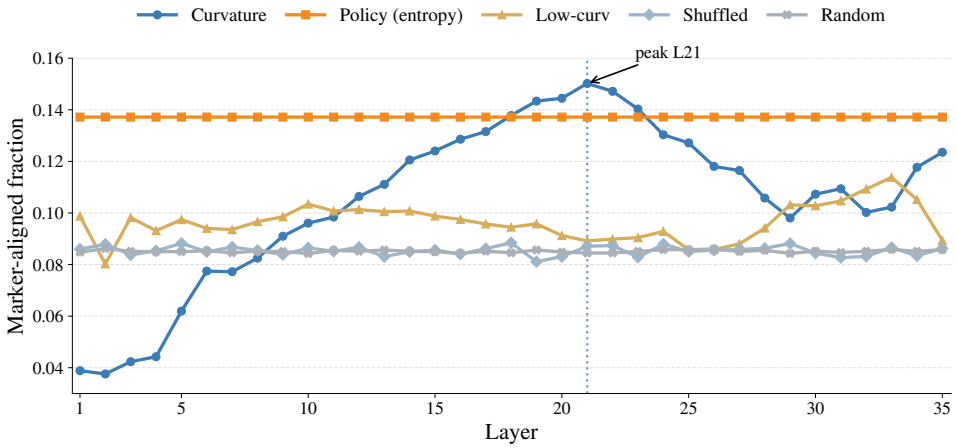

Figure 2: Layer-wise marker alignment of high-curvature positions. The marker-aligned fraction across layers for curvature, policy-side uncertainty, low-curvature, shuffled, and random controls.

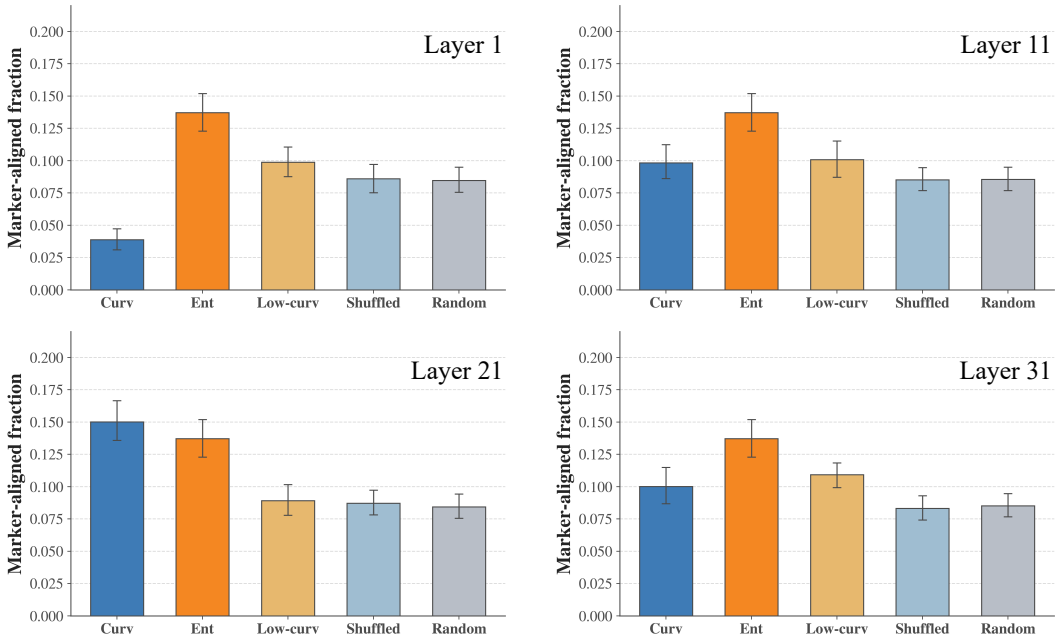

Figure 3: Marker alignment on representative layers. Y-axis shows the fraction of selected tokens near predefined reasoning markers; error bars denote bootstrap confidence intervals over responses. Curvature exhibits a layer-dependent pattern, peaking around Layer 21.

Response-region enrichment further tests whether curvature concentrates in particular parts of the trace rather than merely selecting prompt copying, answer formatting, or final delimiters.

Figure 2 shows a clear layer-dependent pattern. Curvature has weak marker alignment in early layers, increases steadily through the middle layers, and reaches its strongest alignment around Layer 21. The representative-layer view in Figure 3 confirms this trend: entropy is stronger in early and later layers, but curvature becomes the most marker-aligned signal at the peak layer. This suggests that hidden-trajectory curvature does not simply reflect output uncertainty or response-position bias; rather, it becomes informative in specific intermediate layers where latent reasoning transitions are most visible. From a geometric perspective, this layer-wise rise and fall maybe consistent with the formation and subsequent collapse of a reasoning manifold across the model depth: intermediate

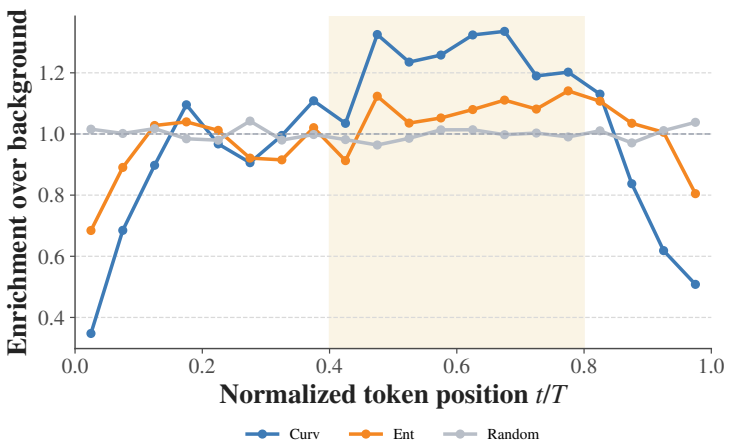

Figure 4: Response-position enrichment of selected tokens. Curves show enrichment over a uniform background along normalized token position, with the shaded area indicating the middle response region. Curvature-selected tokens are more enriched in the middle of the trace than policy-side uncertainty and random controls, suggesting localization to response-internal reasoning transitions.

layers transiently organize hidden trajectories around reasoning-relevant transitions, whereas earlier and later layers are less dominated by this structured reasoning geometry.

4.3 Reasoning Enrichment

For response-region enrichment, we divide each generated response into early, middle, and late regions using a $0.4/0.4/0.2$ split. This split is an operational coarse partition motivated by classical and empirical models of mathematical problem solving, which distinguish understanding, planning, implementation, and verification or reflection phases (Schoenfeld, 2014; Rott et al., 2021). In generated mathematical solutions, the final region is often dominated by answer consolidation and formatting, while the main derivation occupies the central part of the trace. We therefore use a shorter late region.

Figure 4 further shows that curvature is preferentially enriched in the reasoning-intensive middle region of the response. After the initial setup portion, curvature enrichment rises sharply and stays well above background throughout the main derivation interval, before dropping near the final answer region. This means they are more likely to occur where the model develops, checks, or redirects its reasoning, rather than at prompt-copying or answer-formatting positions. By contrast, the policy-side uncertainty signal varies more smoothly across the response and shows only a mild middle-region increase. Random selections remain close to the background level. Thus, curvature provides a more localized signal for response-internal reasoning transitions, whereas the policy-side signal behaves more like a broad uncertainty profile.

4.4 Curvature–Entropy Complementarity

We next test whether curvature is redundant with policy-side uncertainty. To this end, we partition top-curvature and top-entropy selections into four groups: their overlap, curvature-only tokens, entropy-only tokens, and random controls. This decomposition separates two effects: whether curvature identifies positions also regarded as uncertain by the policy, and whether it contributes additional marker-proximal positions after removing this overlap.

Figure 5 shows that the overlap group has the strongest marker alignment in middle and later layers, indicating that hidden-trajectory curvature and policy-side uncertainty can reinforce each other near reasoning-relevant transitions. Importantly, curvature-only tokens remain nontrivial and marker-proximal, especially around the layers where curvature alignment is strongest. This suggests that curvature is not merely a restatement of output uncertainty, but provides complementary information about latent reasoning transitions.

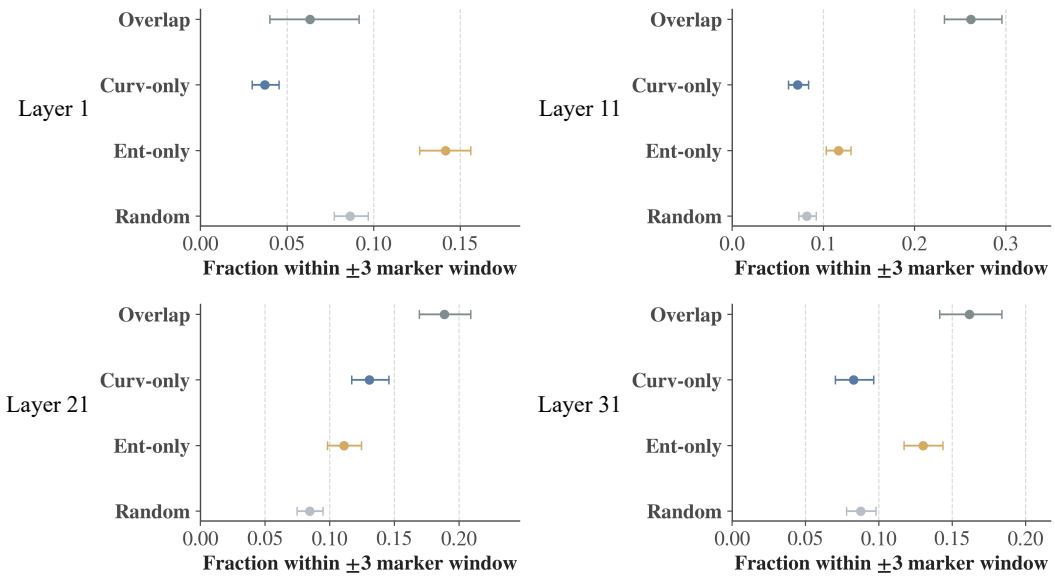

Figure 5: Curvature–entropy complementarity diagnostic. Top-curvature and top-entropy selections are partitioned into overlap, curvature-only, entropy-only, and random groups. Curvature-only positions remain marker-aligned after removing tokens also selected by entropy.

4.5 RLVR: Turning the Prior into an Update

We next examine whether hidden-state turning information remains useful when incorporated into RLVR training. We compare the DAPO baseline with entropy-based baselines and several curvature-based variants. Entropy-Hard Top-20% (Ent-Hard) restricts policy updates to tokens with the highest entropy. Our main method, Curvature-Soft (Cur), uses curvature-dependent weights to softly reweight token-level policy updates, emphasizing high-curvature positions without discarding the remaining tokens. We further consider Entropy-Soft (Ent-Soft), which replaces curvature with entropy while retaining the same soft-weighting scheme, enabling a controlled comparison of the two scoring signals. Finally, we apply curvature-based reweighting only to high-entropy tokens (Cur w/ Ent) to examine whether entropy and curvature provide complementary signals. Table 1 shows that Cur improves over DAPO across all six benchmarks for both model sizes.

We further compare ARC with a within-response shuffled-weight control that preserves the allocation-weight distribution. The results in Appendix C support the importance of aligning curvature weights with their original token positions.

In pilot experiments, we also explored Curvature-Hard Top-20%, which updates only the highest-curvature tokens. However, this strategy led to rapid response-length growth, suggesting an overthinking-like failure mode in which the policy may learn to prolong or repeatedly introduce reasoning transitions. We therefore do not include the hard-curvature variant as our main result. This observation motivates our central design choice: curvature should serve as a geometric prior for softly allocating credit, rather than as a hard local optimization target.

Why does curvature-based credit allocation help? We attribute the improvement to a better alignment between token-level updates and the internal structure of reasoning. Standard policy optimization distributes verifier-derived credit broadly across response tokens, while curvature highlights positions where the hidden trajectory changes direction and reasoning is more likely to undergo a meaningful transition. By softly assigning larger weights to these positions, our method strengthens updates at potentially more consequential reasoning steps without discarding the rest of the trajectory. This reduces noisy credit assignment while avoiding the length inflation observed with hard curvature selection. The fact that curvature-only performs best on most benchmarks, whereas entropy-conditioned variants do not consistently improve it, further suggests that curvature captures information beyond token-level uncertainty.

Table 1: Evaluation scores on six mathematical reasoning benchmarks.

Benchmark	AIME'24	AIME'25	AMC'23	MATH500	Minerva	Olympiad
RLVR from the Qwen3-14B Base Model						
Baseline	42.95	37.06	90.26	91.5	40.23	57.88
Ent-Hard	46.16	40.91	91.99	93.4	42.83	61.50
Ent-Soft	47.21	41.23	91.42	93.7	42.95	60.78
Cur	**51.22**	**42.35**	**92.89**	**94.6**	**44.69**	**63.78**
Cur w/ Ent	48.01	41.12	91.56	93.7	44.21	62.95
Improvement vs Baseline	**+8.27**	**+5.29**	**+2.63**	**+3.1**	**+4.46**	**+5.90**
RLVR from the Qwen3-8B Base Model						
Baseline	29.65	22.18	77.32	87.4	35.09	49.76
Ent-Hard	30.83	23.33	78.28	87.6	35.71	50.34
Ent-Soft	31.56	23.63	**79.59**	87.9	35.52	50.59
Cur	**35.63**	**25.63**	78.28	**88.4**	**36.63**	**50.75**
Cur w/ Ent	32.71	23.54	76.56	87.4	36.14	50.49
Improvement vs Baseline	**+5.98**	**+3.45**	**+0.96**	**+1.0**	**+1.54**	**+0.99**

5 Related Work

Credit assignment and sparse token optimization in RLVR. A growing line of work studies how to assign outcome rewards more precisely within generated reasoning traces. VinePPO uses Monte-Carlo estimates to improve intermediate credit assignment, while GRPO-λ introduce eligibility-trace-style mechanisms for token-level credit assignment Kazemnejad et al. (2024); Parthasarathi et al. (2025). Other methods construct denser supervision by resetting failed trajectories, deriving verifiable subproblems, or converting intermediate verifiers into process-level rewards Samanta et al. (2026); Jiang et al. (2026); Yuan et al. (2026). In parallel, several works find that useful RLVR updates are highly sparse. High-entropy minority token methods identify uncertain tokens as reasoning forks Wang et al. (2026); ERPO further treats transient high-entropy states as critical decision pivots Yu et al. (2026b); and update-direction analyses locate critical positions using token-level probability changes such as signed $\Delta \log p$ Huang et al. (2026). These methods differ technically, but most of them define token importance from the policy side, rather than LLM's hidden reasoning representations.

Hidden representations and neural manifold. A complementary literature studies the structure of internal representations. Transformers can represent belief-state geometry in their residual streams (Shai et al., 2024). Semantic Entropy Probes show that hidden states can approximate semantic uncertainty from a single generation (Kossen et al., 2024). Recent trajectory-based studies explicitly model reasoning as movement through representation space and find correctness-related geometric signals (Sun et al., 2026a; Damirchi et al., 2026). In neuroscience, neural-manifold analysis treats high-dimensional neural activity as low-dimensional state-space trajectories whose geometry relates to behavior and cognition (Mitchell-Heggs et al., 2022). Our focus is on using local hidden-trajectory geometry to define token-level weights in an outcome-supervised policy surrogate. The diagnostic analysis and training evaluation examine the utility of this geometric prior for RLVR.

6 Conclusion

We studied RLVR credit assignment through hidden-state geometry, treating reasoning traces as latent trajectories and using curvature to identify local directional changes. Diagnostic experiments reveal layer-dependent alignment with reasoning markers, enrichment in middle response regions, and selection patterns complementary to entropy. Used as detached soft weights, curvature guides verifier-derived updates toward hidden turns and improves RLVR performance across two model scales and six mathematical reasoning benchmarks.

More broadly, these results motivate hidden-state geometry as a tool for studying and shaping model computation. Future work could extend this perspective to world models, exploring whether curvature helps identify transition boundaries, improve latent state abstraction, or encourage more structured predictive dynamics.

REFERENCES

Jorge Zhoujun Cheng, Shibo Hao, Tianyang Liu, Fan Zhou, Yutao Xie, Feng Yao, Yuexin Bian, Nilabjo Dey, Yonghao Zhuang, Yuheng Zha, et al. Revisiting reinforcement learning for llm reasoning from a cross-domain perspective. *Advances in Neural Information Processing Systems*, 38, 2026.

Hamed Damirchi, Ignacio Meza De la Jara, Ehsan Abbasnejad, Afshar Shamsi, Zhen Zhang, and Javen Shi. Truth as a trajectory: What internal representations reveal about large language model reasoning. *arXiv preprint arXiv:2603.01326*, 2026.

Manfredo P Do Carmo. *Differential geometry of curves and surfaces: revised and updated second edition*. Courier Dover Publications, 2016.

Juan A Gallego, Matthew G Perich, Raeed H Chowdhury, Sara A Solla, and Lee E Miller. Long-term stability of cortical population dynamics underlying consistent behavior. *Nature neuroscience*, 23 (2):260–270, 2020.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. 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)*, pp. 3828–3850, 2024.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Kexin Huang, Haoming Meng, Junkang Wu, Jinda Lu, Chiyu Ma, Ziqian Chen, Xue Wang, Bolin Ding, Jiancan Wu, Xiang Wang, et al. On the direction of rlvr updates for llm reasoning: Identification and exploitation. *arXiv preprint arXiv:2603.22117*, 2026.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Xitai Jiang, Zihan Tang, Wenze Lin, Yang Yue, Shenzhi Wang, and Gao Huang. From reasoning chains to verifiable subproblems: Curriculum reinforcement learning enables credit assignment for llm reasoning. *arXiv preprint arXiv:2605.22074*, 2026.

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. Vineppo: Unlocking rl potential for llm reasoning through refined credit assignment. 2024.

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

John M Lee. *Introduction to Riemannian manifolds*, volume 2. Springer, 2018.

Rufus Mitchell-Heggs, Seigfred Prado, Giuseppe P. Gava, Mary Ann Go, and Simon R. Schultz. Neural manifold analysis of brain circuit dynamics in health and disease. *arXiv preprint arXiv:2203.11874*, 2022.

Prasanna Parthasarathi, Mathieu Reymond, Boxing Chen, Yufei Cui, and Sarath Chandar. GRPO-λ: Credit assignment improves llm reasoning. *arXiv preprint arXiv:2510.00194*, 2025.

Matthew G Perich, Devika Narain, and Juan A Gallego. A neural manifold view of the brain. *Nature Neuroscience*, 28(8):1582–1597, 2025.

Benjamin Rott, Birte Specht, and Christine Knipping. A descriptive phase model of problem-solving processes. *ZDM–Mathematics Education*, 53(4):737–752, 2021.

Ankur Samanta, Akshayaa Magesh, Ayush Jain, Youliang Yu, Daniel Jiang, Kavosh Asadi, Kaveh Hassani, Paul Sajda, Jalaj Bhandari, and Yonathan Efroni. Credit assignment with resets in language model reasoning. *arXiv preprint arXiv:2605.25507*, 2026.

Shreya Saxena and John P Cunningham. Towards the neural population doctrine. *Current opinion in neurobiology*, 55:103–111, 2019.

Alan H Schoenfeld. *Mathematical problem solving*. Elsevier, 2014.

Adam S. Shai, Sarah E. Marzen, Lucas Teixeira, Alexander Gietelink Oldenziel, and Paul M. Riechers. Transformers represent belief state geometry in their residual stream. *arXiv preprint arXiv:2405.15943*, 2024.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Lihao Sun, Hang Dong, Bo Qiao, Qingwei Lin, Dongmei Zhang, and Saravan Rajmohan. Llm reasoning as trajectories: Step-specific representation geometry and correctness signals. *arXiv preprint arXiv:2604.05655*, 2026a.

Yiliu Sun, Zicheng Zhao, Yang Wei, Yanfang Zhang, and Chen Gong. Well begun, half done: Reinforcement learning with prefix optimization for llm reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 33144–33152, 2026b.

Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Saurabh Vyas, Matthew D Golub, David Sussillo, and Krishna V Shenoy. Computation through neural population dynamics. *Annual review of neuroscience*, 43(1):249–275, 2020.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xiong-Hui Chen, Jianxin Yang, Zhenru Zhang, et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *Advances in Neural Information Processing Systems*, 38:115452–115486, 2026.

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

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026a.

Song Yu, Li Li, Wenwen Zhao, and Zhisheng Yang. Erpo: Token-level entropy-regulated policy optimization for large reasoning models. *arXiv preprint arXiv:2603.28204*, 2026b.

Huining Yuan, Zelai Xu, Huaijie Wang, Xiangmin Yi, Jiaxuan Gao, Xiao-Ping Zhang, Yu Wang, Chao Yu, and Yi Wu. Verifiable process rewards for agentic reasoning. *arXiv preprint arXiv:2605.10325*, 2026.

A THEORY OF HIDDEN-TRAJECTORY CURVATURE

For a fixed layer of an LLM, generation produces a sequence of hidden states $h_1, \ldots, h_T \in \mathbb{R}^m$. We idealize the region visited by these states as a smooth embedded manifold $M \subset \mathbb{R}^m$, equipped with the metric g induced by the ambient Euclidean space. A continuous idealization of the sequence is a C^2 curve

$$\alpha : [0, S] \longrightarrow M, \qquad \|\dot{\alpha}(s)\|_g = 1, \tag{9}$$

where s denotes arc length. This idealization provides a setting for analyzing local variation along the hidden trajectory.

Let $U \subset M$ be an open neighborhood of the trajectory and let $f : U \to \mathbb{R}$ be a C^2 scalar semantic readout. Defining f on a neighborhood within M allows its Riemannian gradient and Hessian to be evaluated along the curve. We study the second-order variation of

$$u(s) = f(\alpha(s)). \tag{10}$$

Two kinds of curvature enter the analysis. Turning within M contributes to the variation of u through the intrinsic acceleration of the curve. The curvature computed from observed hidden states, however, is measured in the ambient space and also includes bending normal to M. We first relate the corresponding connections and Hessians, then give sufficient conditions for the intrinsic regularity bounds used in Proposition 1, and finally connect its intrinsic bound to ambient trajectory curvature.

We write D for the ambient Euclidean connection and ∇^M for the Levi-Civita connection on M. The notation $\nabla^{\mathbb{R}^m} F$ denotes the Euclidean gradient of an ambient function, whereas $\operatorname{grad}_M f$ denotes the Riemannian gradient.

Lemma A.1 (Connection decomposition)**.** *Let X and Y be smooth tangent vector fields on M. Then*

$$D_X Y = \nabla^M_X Y + \mathrm{II}(X, Y), \tag{11}$$

where $\mathrm{II}(X, Y)$ is the second fundamental form and takes values in the normal bundle of M.

Proof. Decompose the ambient derivative into tangent and normal components:

$$D_X Y = (D_X Y)^\top + (D_X Y)^\perp.$$

The tangent projection defines a metric-compatible, torsion-free connection on M and therefore equals its Levi-Civita connection: $(D_X Y)^\top = \nabla^M_X Y$. By definition, $(D_X Y)^\perp = \mathrm{II}(X, Y)$, which gives equation 11. □

Lemma A.2 (Hessian decomposition)**.** *Let F be a C^2 extension of f to an ambient neighborhood of U. For tangent vector fields X, Y on U,*

$$\mathrm{Hess}_M f(X, Y) = \mathrm{Hess}_{\mathbb{R}^m} F(X, Y) + \left\langle \nabla^{\mathbb{R}^m} F, \mathrm{II}(X, Y) \right\rangle_2. \tag{12}$$

Proof. Because $F|_U = f$, the definition of the Riemannian Hessian gives

$$\begin{aligned} \mathrm{Hess}_M f(X, Y) &= X(Yf) - (\nabla^M_X Y) f \\ &= X(YF) - \left\langle \nabla^{\mathbb{R}^m} F, \nabla^M_X Y \right\rangle_2. \end{aligned} \tag{13}$$

Substituting Eq. equation 11, we obtain

$$\begin{aligned} \mathrm{Hess}_M f(X, Y) &= X(YF) - \left\langle \nabla^{\mathbb{R}^m} F, D_X Y \right\rangle_2 \\ &\quad + \left\langle \nabla^{\mathbb{R}^m} F, \mathrm{II}(X, Y) \right\rangle_2 \\ &= \mathrm{Hess}_{\mathbb{R}^m} F(X, Y) + \left\langle \nabla^{\mathbb{R}^m} F, \mathrm{II}(X, Y) \right\rangle_2. \end{aligned} \tag{14}$$

□

The Hessian in Eq. equation 12 describes intrinsic variation of the readout. The next lemma shows how ambient regularity and the embedding geometry can jointly provide the intrinsic bounds assumed in Proposition 1.

Lemma A.3 (Sufficient conditions for intrinsic regularity)**.** *Let F be as in Lemma A.2. Suppose that, at every point $x = \alpha(s)$, there are uniform constants $L, G, B \geq 0$ such that*

$$\left\|\nabla^2 F(x)\right\| \leq L, \qquad \left\|\nabla^{\mathbb{R}^m} F(x)\right\| \leq G, \tag{15}$$

and, for all $X, Y \in T_x M$,

$$\|\mathrm{II}_x(X, Y)\| \leq B\|X\| \cdot \|Y\|. \tag{16}$$

Then, along the trajectory,

$$\|\operatorname{grad}_M f\| \leq G, \qquad \|\operatorname{Hess}_M f\| \leq L + G \cdot B. \tag{17}$$

Proof. The Riemannian gradient is the tangent projection of the ambient gradient:

$$\operatorname{grad}_M f = \left(\nabla^{\mathbb{R}^m} F\right)^{\top}.$$

For any tangent vectors X, Y, Lemma A.2 and Cauchy–Schwarz give

$$\begin{aligned} \|\operatorname{Hess}_M f(X, Y)\| &\leq L\|X\|\|Y\| + G\|\mathrm{II}(X, Y)\| \\ &\leq (L + G \cdot B)\|X\|\|Y\|. \end{aligned} \tag{18}$$

Taking the supremum over unit tangent vectors yields the Hessian bound. □

Proof of Proposition 1.

Proof. Let $v(s) = \dot{\alpha}(s)$. Since α is parameterized by arc length, $\|v(s)\| = 1$. The chain rule gives

$$u'(s) = \langle \operatorname{grad}_M f, v \rangle_g.$$

Differentiating along the curve and using metric compatibility yields

$$\begin{aligned} u''(s) &= \left\langle \nabla_v^M \operatorname{grad}_M f, v \right\rangle_g + \left\langle \operatorname{grad}_M f, \nabla_v^M v \right\rangle_g \\ &= \operatorname{Hess}_M f(v, v) + \left\langle \operatorname{grad}_M f, \nabla_v^M v \right\rangle_g. \end{aligned} \tag{19}$$

The intrinsic assumptions of the proposition and Cauchy–Schwarz imply

$$\begin{aligned} \|u''(s)\| &\leq \|\operatorname{Hess}_M f\|\|v\|^2 + \|\operatorname{grad}_M f\|\|\nabla_v^M v\| \\ &\leq L + G\kappa_g(s), \end{aligned} \tag{20}$$

where $\kappa_g(s) = \|\nabla_v^M v\|_g$ is the geodesic curvature.

To relate this bound to the ambient curvature, apply Lemma A.1 along α:

$$\ddot{\alpha} = D_v v = \nabla_v^M v + \mathrm{II}(v, v).$$

The tangent and normal components are orthogonal, so

$$\kappa_{\mathrm{amb}}(s)^2 = \kappa_g(s)^2 + \|\mathrm{II}(v, v)\|_2^2, \qquad \kappa_{\mathrm{amb}}(s) = \|\ddot{\alpha}(s)\|_2. \tag{21}$$

Consequently, $\kappa_g(s) \leq \kappa_{\mathrm{amb}}(s)$ and

$$|u''(s)| \leq L + G\kappa_g(s) \leq L + G\kappa_{\mathrm{amb}}(s).$$

□

B Additional Diagnostic Analyses

B.1 Marker Selection

We use reasoning markers as fixed surface-level anchors for evaluating whether selected tokens occur near visible reasoning events. The markers are grouped into transition, action, and answer-related terms. Transition markers capture logical shifts such as consequence, contrast, or conclusion; action markers capture explicit problem-solving operations such as assuming, defining, simplifying, verifying, or proving; and answer markers capture final answer consolidation. These markers are not used for training and are not intended to exhaustively annotate reasoning. They only provide a coarse diagnostic proxy for reasoning-relevant regions. Table 2 lists the complete marker set.

Table 2: Reasoning markers used in the marker-alignment analysis.

Marker type	Terms
Transition	*therefore*, *thus*, *hence*, *however*, *alternatively*, *otherwise*, *instead*, *consequently*, *contradiction*, *finally*, *conclude*, *implies*
Action	*suppose*, *assume*, *define*, *denote*, *consider*, *solve*, *substitute*, *simplify*, *verify*, *derive*, *prove*
Answer	*answer*, *boxed*, *final*, *solution*

B.2 Sensitivity Analysis

We conduct additional analyses to verify that the diagnostic conclusions in Sections 4.2–4.4 are not artifacts of particular analysis choices. We focus on four potential confounds: the fraction of selected tokens, the marker-alignment window size, the use of a binary marker-overlap criterion, and the overlap between curvature-selected and entropy-selected positions.

Sensitivity to the top-selection ratio. The main diagnostic experiments select the top 20% positions within each response and layer. To test whether the observed curvature-marker alignment depends on this threshold, we vary the selected fraction over $\{5\%, 10\%, 20\%, 30\%\}$ while keeping the marker window fixed at ± 3 tokens. The left column of Figure 6 reports the marker-aligned fraction across representative layers. The qualitative pattern remains stable across selection ratios: curvature is weakly aligned in early layers, becomes increasingly marker-aligned in intermediate layers, and is strongest near the peak layer. This indicates that the main diagnostic result is not a consequence of the default top-20% selection rule.

Sensitivity to the marker-window size. Marker alignment in the main text uses a ± 3 token window around reasoning markers. Since this window is a diagnostic choice rather than a model assumption, we further vary it over $\{\pm 1, \pm 3, \pm 5, \pm 7\}$ tokens while keeping the top-selection ratio fixed at 20%. The right column of Figure 6 shows that the marker-aligned fraction increases as the window becomes wider, as expected. More importantly, the relative behavior of curvature, entropy, shuffled curvature, and random controls remains stable. Thus, the curvature-marker alignment is not an artifact of the particular ± 3 token window used in the main analysis.

Threshold-free distance-to-marker analysis. The previous analyses use a binary criterion: a selected token is either inside or outside a marker window. To avoid relying only on this thresholded readout, we also measure the continuous distance from each selected token to the nearest reasoning marker. Figure 7 reports the distribution of mean marker distances for curvature, entropy, low-curvature, shuffled-curvature, and random selections. In the curvature-sensitive layers, high-curvature positions remain closer to reasoning markers than shuffled and random controls. This provides a threshold-free confirmation that curvature-selected tokens are not merely selected by chance or by response-position bias.

Robustness of curvature–entropy complementarity. Finally, we test whether the complementarity result in Section 4.4 is driven only by tokens selected by both curvature and entropy. For each response and layer, we partition selected positions into overlap, curvature-only, entropy-only, and random groups. We then measure the distance from each group to the nearest reasoning marker. As shown in Figure 8, curvature-only positions remain marker-proximal after removing tokens that overlap with entropy-selected positions, especially in the intermediate layers where curvature is most informative. This supports the claim that curvature provides information beyond policy-side uncertainty, rather than simply reproducing entropy-selected reasoning forks.

Complementarity controls. We further check whether the curvature-only effect is caused by group-size imbalance or response-position bias. Figure 9 reports the selected-token count and response-stage distribution at the peak layer. Curvature-only positions form a non-negligible group and are mainly distributed in the middle response region, rather than being concentrated only near the final answer. This supports that curvature contributes complementary reasoning-relevant positions beyond entropy.

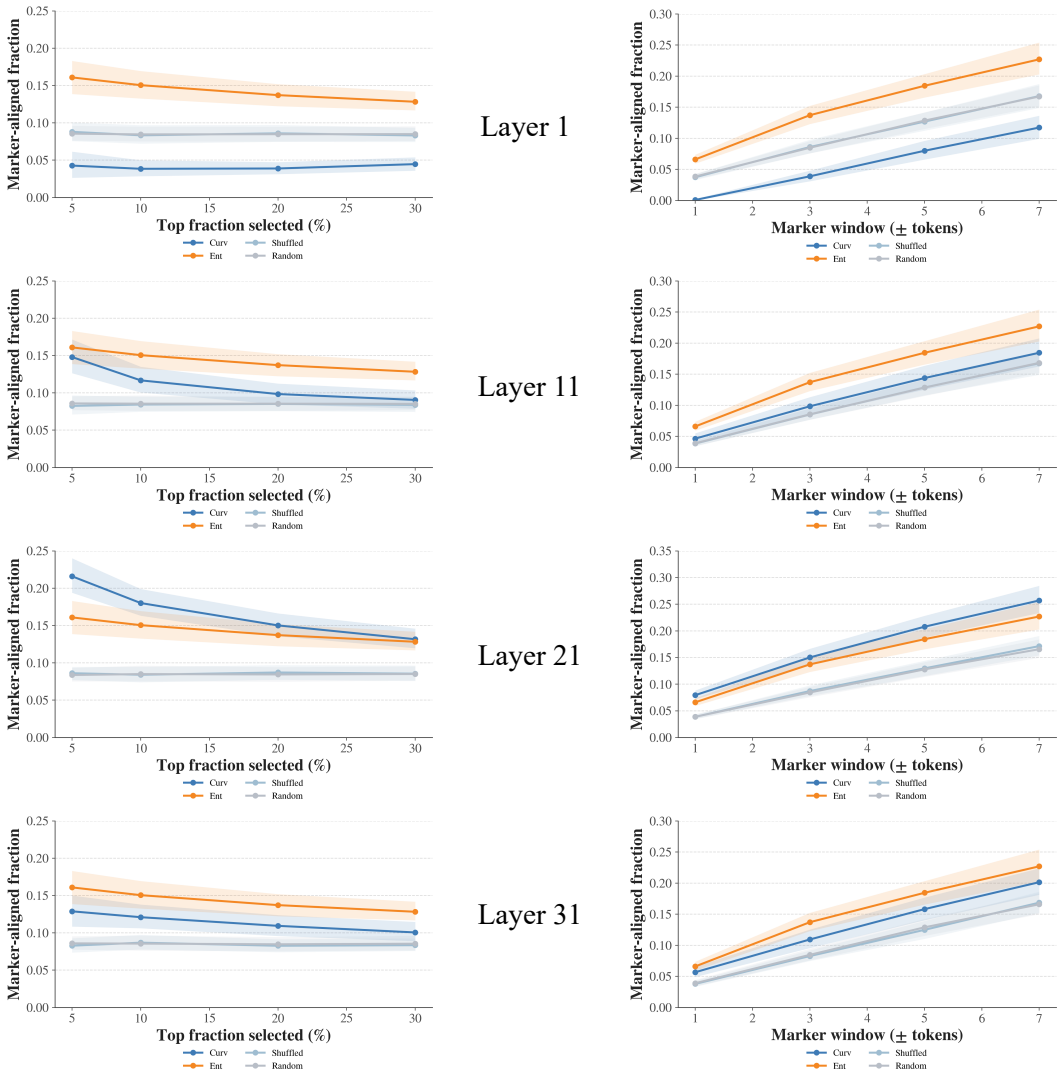

Figure 6: Sensitivity to top-selection ratio and marker-window size. Left column: marker-aligned fraction when varying the selected top fraction from 5% to 30%. Right column: marker-aligned fraction when varying the marker window from ± 1 to ± 7 tokens. Rows correspond to representative layers.

C Ablation Experiment

To test whether the correspondence between curvature weights and token positions contributes to performance, we randomly permute the final normalized weights within each response. This preserves the weight distribution and unit mean while disrupting their alignment with the original token positions. All other training and evaluation settings are kept unchanged. The ablation result is in the Table 3.

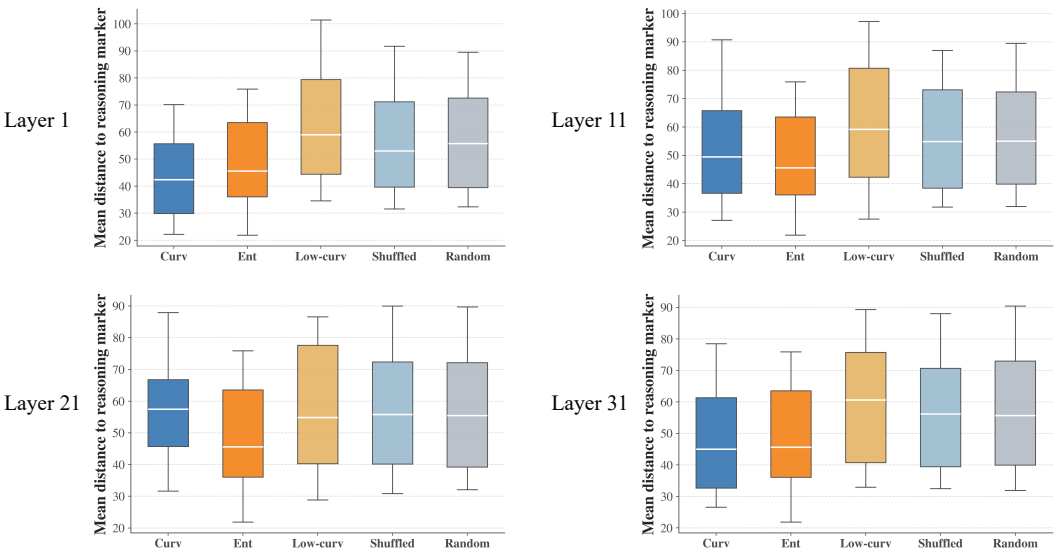

Figure 7: Threshold-free distance-to-marker diagnostic. For each selected group, we measure the mean distance from selected tokens to the nearest reasoning marker. High-curvature selections remain closer to reasoning markers than shuffled and random controls in curvature-sensitive layers.

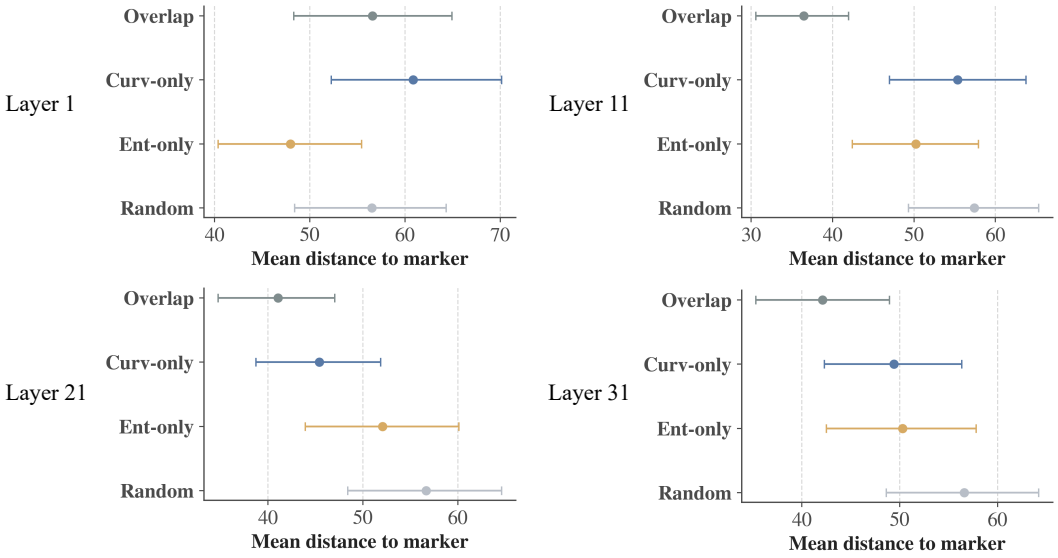

Figure 8: Complementarity distance-to-marker diagnostic. Top-curvature and top-entropy selections are partitioned into overlap, curvature-only, entropy-only, and random groups. Curvature-only positions remain marker-proximal after removing tokens that are also selected by entropy.

D Qualitative Case Study

We provide a qualitative case study to illustrate how curvature-selected tokens appear in an actual generated reasoning trace. Green indicates high-entropy positions, while red indicates high-curvature positions. The example is used only for qualitative illustration; the quantitative evidence is provided by the diagnostic analyses in the main text and Appendix B.

Question. Wang put his phone into the pocket of his coat, and the coat was hanging by the door. Later, he took the coat to the study room, but he did not take the phone out. Where is the phone most likely to be, and why?

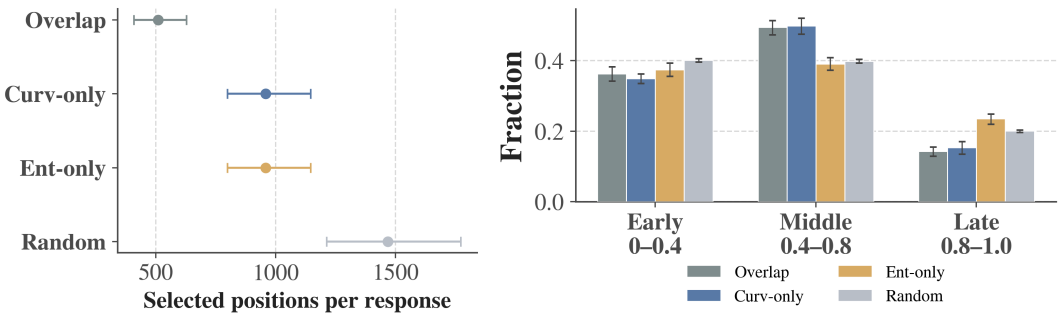

Figure 9: Complementarity controls at Layer 21. Curvature-only positions are non-negligible in count and are mainly distributed in the middle response region, indicating that the complementarity result is not explained by group-size imbalance or final-answer bias.

Table 3: Ablation evaluation scores on six mathematical reasoning benchmarks.

Benchmark	AIME'24	AIME'25	AMC'23	MATH500	Minerva	Olympiad
RLVR from the Qwen3-14B Base Model						
Baseline	42.95	37.06	90.26	91.5	40.23	57.88
Ent-Hard	46.16	40.91	91.99	93.4	42.83	61.50
Cur	**51.22**	**42.35**	**92.89**	**94.6**	**44.69**	**63.78**
Cur-Shuffled	39.36	32.91	85.32	87.8	37.64	54.59
Improvement vs Baseline	**+8.27**	**+5.29**	**+2.63**	**+3.1**	**+4.46**	**+5.90**
RLVR from the Qwen3-8B Base Model						
Baseline	29.65	22.18	77.32	87.4	35.09	49.76
Ent-Hard	30.83	23.33	78.28	87.6	35.71	50.34
Cur	**35.63**	**25.63**	78.28	**88.4**	**36.63**	**50.75**
Cur-Shuffled	27.85	19.73	74.79	86.1	33.26	46.76
Improvement vs Baseline	**+5.98**	**+3.45**	**+0.96**	**+1.0**	**+1.54**	**+0.99**

Generated reasoning trace.

> Let us reason step by step:
>
> 1. **Initial state.** Wang put the phone in the pocket of his coat, then hung the coat by the door.
> 2. **State transfer.** Later, Wang took the coat to the study room. At this point, the key question is whether Wang took the phone out of the pocket.
> 3. **Persistence check.** During the process of moving the coat, Wang did not do anything to remove the phone. There is no additional information indicating that Wang performed any phone-related action while going to the study room.
> 4. **Conclusion update.** Therefore, we can infer that the phone is still in the pocket of the coat; only the coat is now hanging in the study room.
>
> In summary, the phone is most likely still in the pocket of Wang's coat, while the coat is now somewhere in his study room. This is because the entire reasoning trace does not provide any evidence that the phone was taken out.

Interpretation. This case illustrates the qualitative difference between entropy and curvature. High-entropy tokens mainly appear around locally uncertain or content-bearing phrases, such as identifying the current point of reasoning or describing the movement event. In contrast, high-curvature tokens appear near transitions that change the logical state of the reasoning trace: the temporal update then, the evidence check indicating, and the final inference markers Therefore, only, and any. This supports our interpretation that curvature highlights latent reasoning turns rather than merely uncertain next-token choices.

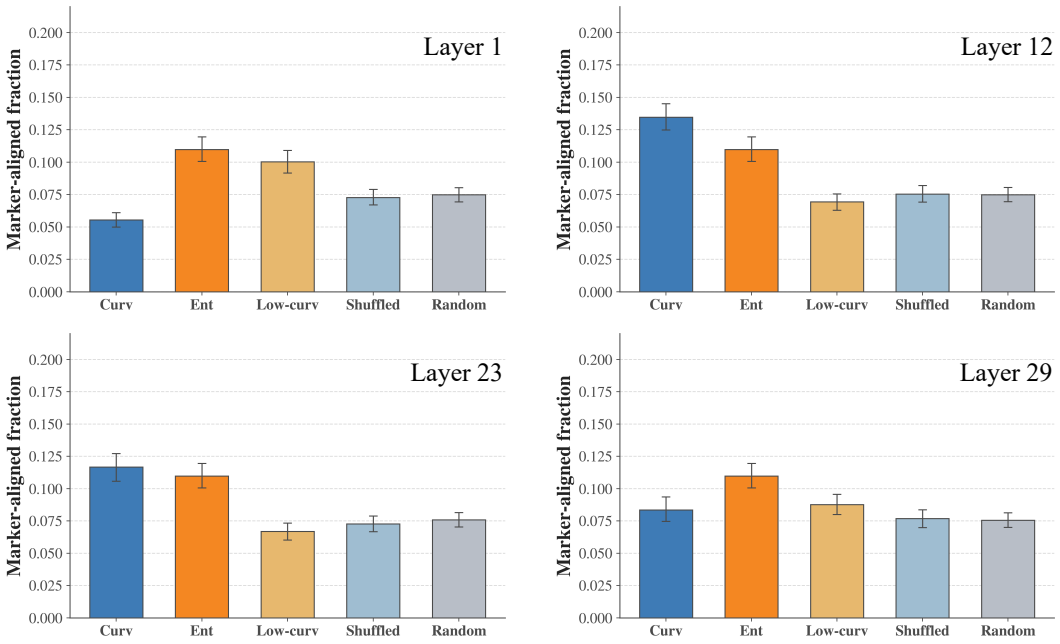

Figure 10: Marker alignment on representative layers of Llama-3.1-8B-Instruct. Curvature reaches its strongest marker alignment around Layer 12, indicating that the peak curvature-sensitive layer shifts across model families while preserving the same intermediate-layer pattern.

E Cross-Model Diagnostic Results on Llama-3.1-8B-Instruct

To examine whether the observed curvature patterns are specific to the Qwen3 model family, we repeat the diagnostic analysis on Llama-3.1-8B-Instruct. We use the same diagnostic protocol as in the main experiments: responses are generated on the Reasoning360 math dataset, token-level curvature and entropy scores are computed across layers, high-score positions are selected within each response and layer using the top-20% criterion, and marker alignment is evaluated using a ± 3 token window around the same reasoning markers.

Figure 10 reports marker alignment on representative Llama layers. The peak layer shifts to Layer 12, rather than the Qwen3 peak layer reported in the main text. This shift is expected because the two models differ in architecture, depth, and representation organization. Importantly, the qualitative pattern remains consistent: curvature is weakly aligned with reasoning markers in early layers, becomes strongest in an intermediate layer, and then decreases in later layers. This supports the interpretation that curvature identifies a transient reasoning-sensitive representation regime, rather than a fixed absolute layer index.

Figure 11 further evaluates curvature–entropy complementarity. As in the Qwen3 analysis, the overlap between curvature and entropy is strongly marker-aligned, showing that the two signals can reinforce each other near reasoning-relevant events. More importantly, curvature-only positions remain marker-proximal in the intermediate layers after removing tokens also selected by entropy. This indicates that curvature still contributes information beyond policy-side uncertainty in a different model family.

Finally, Figure 12 shows response-position enrichment on Llama-3.1-8B-Instruct. Curvature-selected tokens are enriched in the reasoning-intensive middle-to-late portion of the response, while random selections remain close to background. Together, these results show that the main diagnostic conclusions are not tied to Qwen3: hidden-trajectory curvature remains a representation-level signal for reasoning-event localization, although the exact peak layer can vary across model families.

Peak-layer sensitivity on Llama. Since the strongest curvature-marker alignment on Llama-3.1-8B-Instruct occurs around Layer 12, we additionally perform sensitivity checks at this peak layer.

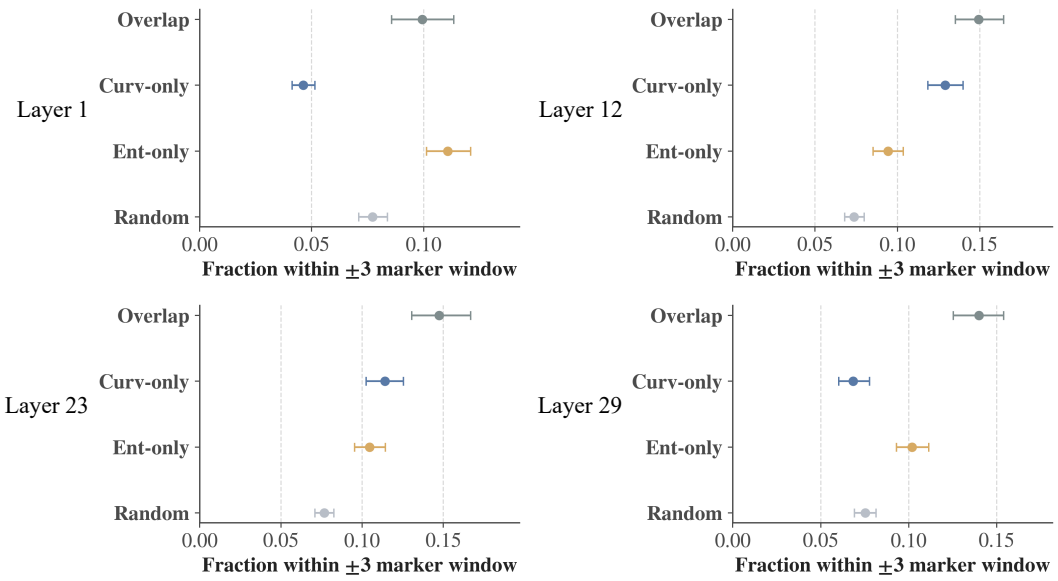

Figure 11: Curvature–entropy complementarity on Llama-3.1-8B-Instruct. Top-curvature and top-entropy selections are partitioned into overlap, curvature-only, entropy-only, and random groups. Curvature-only positions remain marker-proximal in intermediate layers after removing tokens also selected by entropy.

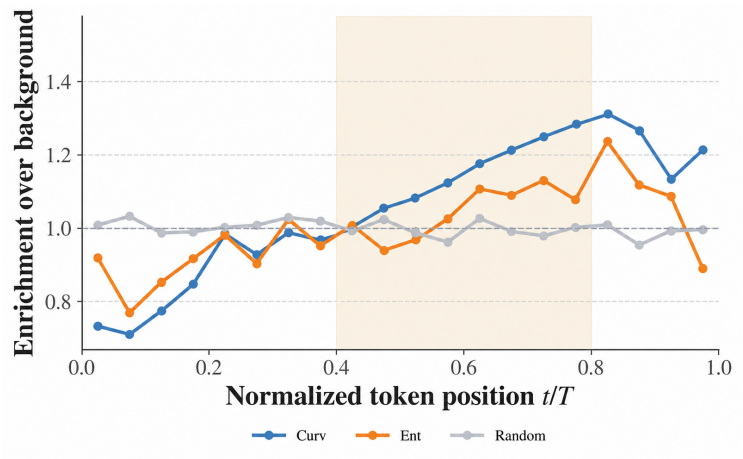

Figure 12: Response-position enrichment on Llama-3.1-8B-Instruct. Curvature-selected positions are enriched in the reasoning-intensive region of the response, showing a pattern consistent with the Qwen3 diagnostics.

As shown in Figure 13, curvature remains more marker-aligned than shuffled and random controls across different top-selection ratios and marker-window sizes. The distance-to-marker diagnostic further confirms that curvature-selected tokens are closer to reasoning markers than low-curvature, shuffled, and random selections. Finally, the complementarity distance analysis shows that curvature-only positions remain marker-proximal after removing overlap with entropy-selected tokens.

This peak-layer sensitivity analysis further suggests that the Llama result is not driven by the default top-20% selection rule, the ± 3 marker window, or curvature–entropy overlap.

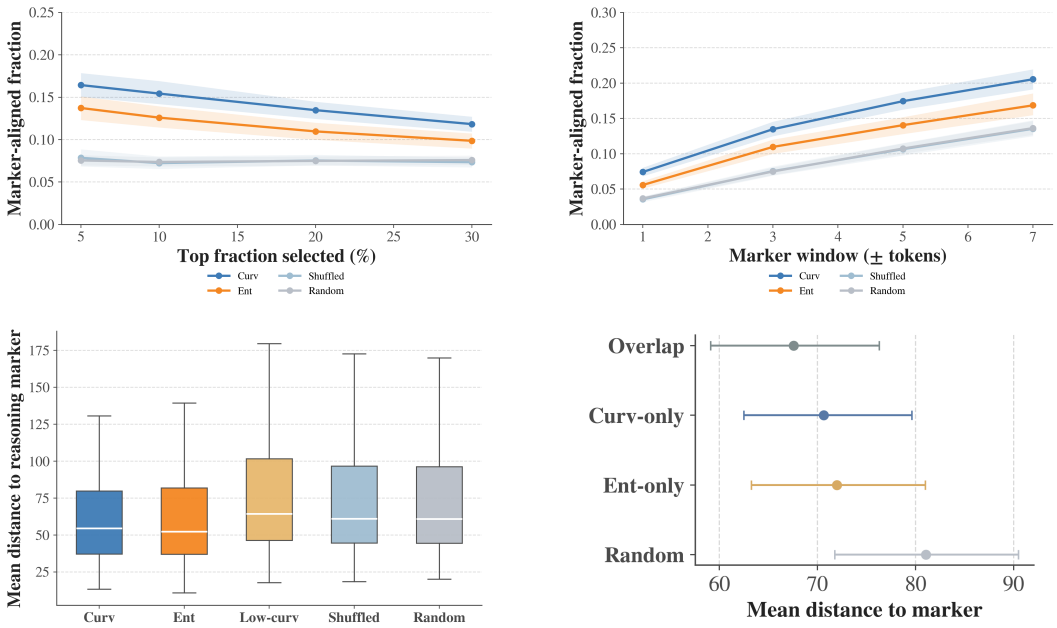

Figure 13: Peak-layer sensitivity analysis on Llama-3.1-8B-Instruct. All panels are computed at Layer 12, where curvature shows the strongest marker alignment. Top left: sensitivity to the top-selection ratio. Top right: sensitivity to the marker-window size. Bottom left: threshold-free distance-to-marker diagnostic. Bottom right: curvature–entropy complementarity distance diagnostic. The conclusions remain stable under these analysis choices.

F LIMITATIONS

Our analysis has several limitations. The marker-alignment diagnostics rely on surface markers, which provide only a coarse proxy for reasoning events and cannot capture all implicit semantic transitions. Curvature should also be interpreted as an allocation prior rather than a causal explanation of correctness: a high-curvature token indicates a hidden-state turn, but not every turn is necessarily useful for solving the problem. This motivates our use of detached, mean-preserving soft weights instead of hard curvature targets. In addition, our experiments focus mainly on mathematical RLVR, and broader validation across tasks, model scales, and domains remains future work. Finally, curvature requires access to intermediate hidden states, which introduces additional activation storage and computation compared with policy-side signals such as entropy.