FROM UNCERTAINTY TO CONSEQUENCE:LEARNING FROM SEMANTIC FORKS IN LLM REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards (RLVR) typically assigns credit from terminal outcomes, making it difficult to identify which decisions matter during natural-language reasoning. Selecting high-entropy tokens makes policy updates sparse, but these uncertain positions are not equally consequential. We find that high-entropy tokens are more prevalent early in autoregressive reasoning and that replacing an early high-entropy token is more likely to change final-answer correctness. However, entropy identifies where the model faces multiple plausible continuations, but does not distinguish the semantic reasoning paths they form or provide local credit based on their terminal outcomes.

We propose SemCredit to focus policy updates on a smaller subset of early high-entropy tokens and assign them path-specific credit. SemCredit groups semantically similar reasoning segments from completed rollouts into states, organizes their sequences in a semantic prefix trie, and estimates local advantages from terminal outcomes associated with each path. These advantages modulate token-level updates without requiring step-level supervision.

Across four Qwen models and six mathematical reasoning benchmarks, SemCredit achieves the highest average performance among the methods evaluated under a controlled setting. On Qwen3-4B, it improves upon the strongest controlled baselines by 9.07 percentage points on AMC 2023 and 8.75 points on AIME 2024. On Qwen2.5-7B, it achieves the best performance on all three evaluated medical benchmarks. These results show that concentrating updates on early uncertain positions and distinguishing the semantic paths they lead to can improve reasoning with credit assigned to only a small fraction of response tokens.

1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR) improves language-model reasoning by optimizing against verifiable outcomes such as final-answer correctness (Shao et al., 2024). Beyond the 80/20 Rule shows that focusing policy updates on a small fraction of high-entropy tokens can improve RLVR (Wang et al., 2025b). ReasonMaxxer further finds that reinforcement learning updates only a small fraction of positions associated with high-entropy decision points, and that these updates can be represented in a low-dimensional parameter space (Akgül et al., 2026). Another line of work seeks intermediate credit that is finer-grained than terminal trajectory rewards. VinePPO and TreeRL estimate intermediate values by sampling continuations, while ROSE and other tree-based methods construct local credit through branch selection and feedback from sibling or descendant nodes (Kazemnejad et al., 2024; Hou et al., 2025; Zhao et al., 2026; Li et al., 2026b; Yang et al., 2025b; Ji et al., 2026; Zong et al., 2026). TEMPO, GiGPO, SALT, and T-STAR instead recover intermediate structure from completed trajectories to enable more localized credit assignment (Tran et al., 2026; Feng et al., 2025; Li et al., 2026a;c). Despite their different implementations, these methods all seek to transform terminal rewards into intermediate credit signals throughout the reasoning process.

Fine-grained credit assignment in natural-language reasoning requires knowing both which reasoning path a rollout follows and where to apply a token-level update. Token entropy identifies positions at which the model is uncertain. In autoregressive generation, changing an early uncertain token can redirect the continuation and alter final-answer correctness, resembling a butterfly effect in which a

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.

small initial difference leads to a different outcome. The token alone, however, does not reveal which reasoning path follows. Comparing reasoning segments across rollouts makes shared paths and semantic branches visible, allowing their terminal outcomes to be associated with those branches. We therefore use two levels of analysis: first locating consequential reasoning segments through their semantic paths and outcomes, then selecting high-entropy tokens within those segments for policy updates.

Based on this insight, we propose **SemCredit**, a fine-grained credit-assignment method that connects paragraph-level reasoning paths with token-level policy updates. SemCredit divides completed rollouts into reasoning segments and organizes the corresponding semantic-state sequences into a semantic prefix trie, from which local advantages are computed using terminal outcomes aggregated at the trie nodes without step-level supervision. Combined with key-token selection, SemCredit uses entropy to locate decisive tokens within each segment. The trajectory advantage of each selected token is modulated by the local advantage of its corresponding segment. Consequently, high-entropy tokens in different segments of the same rollout receive different credit rather than sharing an identical trajectory-level advantage.

Across four Qwen models of different scales and six mathematical reasoning benchmarks, SemCredit achieves the highest average performance among all methods evaluated under the controlled setting, outperforming the strongest controlled baseline for each model by 1.26–5.54 percentage points. On Qwen3-4B, SemCredit improves performance by 9.07 percentage points on AMC 2023 and 8.75 percentage points on AIME 2024. In medical reasoning experiments, SemCredit also achieves the best performance across all three evaluated benchmarks. Ablation results further show that adding local credit to early high-entropy token selection improves average performance by 1.96–8.85 percentage points across the four models.

The key takeaways of our paper are:

- In autoregressive reasoning, high-entropy tokens are more prevalent early in a rollout, and replacing an early high-entropy token is more likely to change final-answer correctness. Thus, even within an already sparse set of uncertain tokens, a smaller subset of early decisions is especially consequential for subsequent reasoning.
- SemCredit focuses updates on these early high-entropy tokens and assigns credit according to the semantic paths they belong to. By merging similar reasoning segments into a prefix trie and aggregating their terminal outcomes, it gives tokens in different segments path-specific credit without step-level supervision.
- This combination of selective updates and semantic credit improves reasoning performance across four Qwen models and six mathematical benchmarks, with further gains on three medical benchmarks. At the $\tau = 0.5$ setting, only about 11–12% of response tokens receive credit, while adding semantic-trie credit improves the six-benchmark average by 1.96–8.85 percentage points over early high-entropy selection alone.

2 Positional Effects of High-Entropy Tokens

We analyze both the positional distribution of high-entropy tokens and the outcome sensitivity of high-entropy tokens at different positions. First, we divide each reasoning trajectory into ten normalized position bins and measure the high-entropy token rate within each bin. As shown in Figure 1, high-entropy tokens are not uniformly distributed along the trajectory. Across all three models, the high-entropy rate is generally above the 20% baseline at early positions, decreases as reasoning proceeds, and falls below the baseline at later positions. This pattern indicates that model uncertainty is more concentrated during early problem interpretation and strategy formation.

We then sample high-entropy tokens from different position bins, replace each selected token with the model's highest-probability alternative, and regenerate the continuation from that position. Figure 2 reports whether the correctness of the final answer changes after this intervention. Replacing high-entropy tokens near the beginning of the trajectory is more likely to change final-answer correctness than replacing those near the end. Early positions leave a longer remaining reasoning horizon, giving a local change more opportunities to alter the subsequent reasoning process and final outcome.

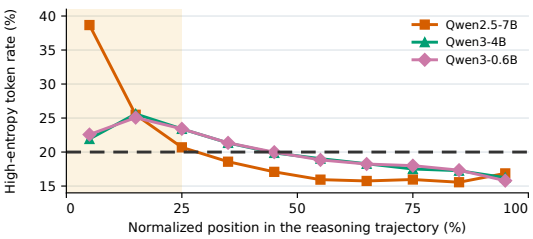

Figure 1: **High-entropy token positions.** High-entropy token rates across normalized trajectory positions for three models.

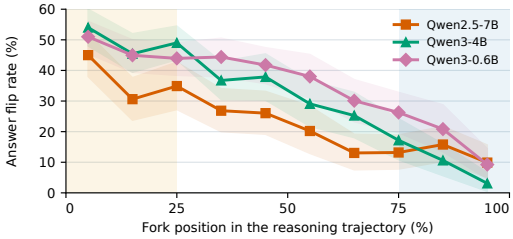

Figure 2: **Outcome sensitivity by position.** Final-answer correctness flip rates after replacing high-entropy tokens at different trajectory positions.

Together, these results show that the training value of a high-entropy token depends not only on its uncertainty but also on its position in the reasoning trajectory, motivating the use of a decision horizon to prioritize high-entropy tokens near the beginning of the response.

3 Semantic Credit Assignment from Terminal Outcomes

SemCredit converts trajectory-level terminal rewards into localized credit for reasoning segments. As illustrated in Figure 3, reasoning rollouts and their terminal outcomes are represented as sequences of semantic states and merged into a semantic prefix trie. Comparing the success rates of a node and its parent estimates whether rollouts following that semantic transition are more or less likely to end in a correct answer, relative to the preceding semantic path. This signal is then combined with the rollout-level advantage to obtain consequence-conditioned semantic credit, which guides selective policy optimization over high-entropy tokens within an early decision horizon.

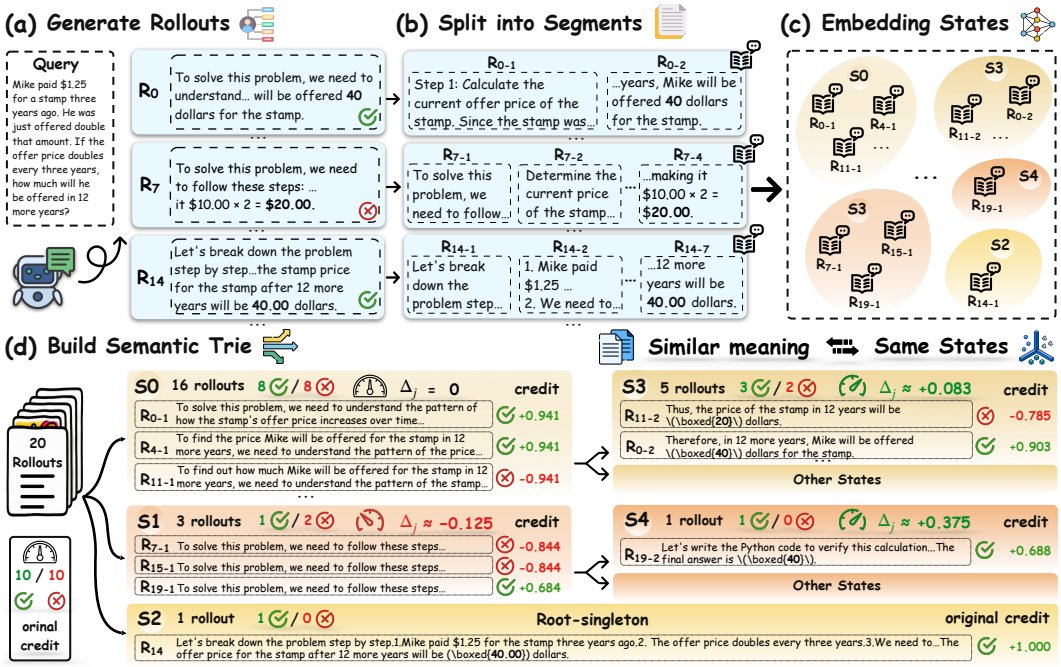

Figure 3: **From Terminal Rewards to Semantic Credit.** (a) Obtaining reasoning rollouts and their terminal outcomes. (b) Splitting each rollout into reasoning segments. (c) Embedding and clustering semantically equivalent segments into shared states. (d) The resulting state sequences are merged into a semantic prefix trie, where path-conditioned outcome statistics are used to derive semantic credit for individual reasoning segments.

3.1 SEMANTIC TRIE LOCAL ADVANTAGE

For each problem, we sample K completions from a frozen base model at a nonzero temperature and retain problems with both correct and incorrect completions. The group-relative advantage of rollout i is

$$A_i = \frac{r_i - \bar{r}}{\sigma_r + \epsilon}. \tag{1}$$

$r_i \in \{0, 1\}$ is the terminal correctness reward, $\bar{r}$ and σ_r are its mean and standard deviation within the group, and ϵ ensures numerical stability.

We split each rollout at paragraph boundaries, cluster semantically similar segments into states S_χ, and insert the resulting state sequences into a semantic prefix trie. Each node thus represents a semantic state conditioned on its preceding path. The clustering procedure is detailed in Appendix B. For node j, let n_j be the number of rollouts passing through it and c_j the number ending correctly. Its smoothed success rate is

$$\widehat{V}_j = \frac{c_j + \alpha}{n_j + 2\alpha}. \tag{2}$$

Here, α controls symmetric smoothing. We treat the complete rollout group as the parent of each initial state. The local advantage of entering node j is

$$\Delta_j = \widehat{V}_j - \widehat{V}_{\mathrm{pa}(j)}. \tag{3}$$

Thus, Δ_j measures the change in empirical downstream success rate relative to the preceding semantic path; it is an association derived from terminal outcomes, not a causal effect.

3.2 CONSEQUENCE-CONDITIONED CREDIT ASSIGNMENT

After obtaining the local advantage Δ_j of node j, we compare its direction with that of the group-relative advantage of trajectory i:

$$q_{i,j} = \mathrm{sgn}(A_i)\Delta_j, \qquad m_{i,j} = \begin{cases} 1 + q_{i,j}, & q_{i,j} \geq 0, \\ 1 + \lambda q_{i,j}, & q_{i,j} < 0, \end{cases} \qquad \lambda \in [0, 1]. \tag{4}$$

Here, $q_{i,j}$ measures directional agreement, $m_{i,j}$ is the Credit Multiplier, and λ controls attenuation when the directions conflict. Agreement strengthens the update; conflict weakens it. Since $m_{i,j}$ is nonnegative, the update direction remains determined by A_i.

We then compute a new advantage for the reasoning segment in trajectory i associated with semantic node j:

$$w_{i,j} = \frac{n_j}{n_j + \kappa}\, m_{i,j}, \qquad \widetilde{A}_{i,j} = A_i w_{i,j}. \tag{5}$$

Here, $w_{i,j}$ is the credit weight, $\widetilde{A}_{i,j}$ is the segment advantage, and κ controls support-based smoothing: $n_j/(n_j + \kappa)$ reduces the weight of nodes visited by few rollouts. Segments mapped to different nodes thus receive different advantages. For a root-singleton rollout whose initial semantic state is not shared, we use A_i directly, without the multiplier or smoothing.

3.3 SPARSE POLICY OPTIMIZATION

We select tokens using both entropy and position. The decision-token set is

$$\mathcal{D} = \left\{ (i, t) : H_{i,t} \geq \xi_{i,\rho},\ \frac{t}{T_i} \leq \tau \right\}. \tag{6}$$

Here, $H_{i,t}$ is predictive entropy under the frozen base model, $\xi_{i,\rho}$ is the threshold for the top ρ fraction of tokens in rollout i, T_i is its length, and τ is the decision-horizon cutoff. For $(i, t) \in \mathcal{D}$, $j(i, t)$ denotes the semantic node of the segment containing token t.

The training loss over decision tokens is defined as

$$\mathcal{L}_{\mathrm{decision}} = -\frac{1}{|\mathcal{D}|} \sum_{(i,t)\in\mathcal{D}} \mathrm{clip}\left(\widetilde{A}_{i,j(i,t)}, -A_{\max}, A_{\max}\right) \log \pi_{\boldsymbol{\theta}}\left(o_{i,t} \mid \boldsymbol{q}, \boldsymbol{o}_{i,<t}\right). \tag{7}$$

$\pi_{\boldsymbol{\theta}}$ is the trainable policy and $A_{\max}$ is the advantage-clipping threshold. A positive advantage increases the probability of the corresponding token, whereas a negative advantage decreases it.

For positions not selected as decision tokens, we regularize the trainable policy toward the frozen base model π_0:

$$\mathcal{L}_{\mathrm{KL}} = \frac{1}{|\mathcal{D}^c|} \sum_{(i,t)\in\mathcal{D}^c} D_{\mathrm{KL}}\left(\pi_0(\cdot \mid \boldsymbol{q}, \boldsymbol{o}_{i,<t}) \,\|\, \pi_{\boldsymbol{\theta}}(\cdot \mid \boldsymbol{q}, \boldsymbol{o}_{i,<t})\right), \tag{8}$$

where $\mathcal{D}^c$ denotes all valid response-token positions not contained in $\mathcal{D}$. The final objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{decision}} + \beta_{\mathrm{KL}}\mathcal{L}_{\mathrm{KL}}, \tag{9}$$

where β_{KL} controls the strength of KL regularization.

4 Experiments

4.1 Experimental Setup

Models and Evaluation Benchmarks. We evaluate the proposed method on Qwen2.5-7B (Yang et al., 2024), Qwen3-0.6B, Qwen3-4B, and Qwen3-8B (Yang et al., 2025a), spanning two Qwen model families and parameter scales ranging from 0.6B to 8B. Our mathematical reasoning evaluation covers six benchmarks: AIME 2024 (Maxwell-Jia, 2024), AIME 2025 (Zhang & Math-AI Team, 2025), AIME 2026 (Dekoninck et al., 2026), AMC 2023 (zwhe99, 2023), MATH-500 (Hendrycks et al., 2021b; Lightman et al., 2024), and OlympiadBench (He et al., 2024). Among these benchmarks, MATH-500 is treated as in-distribution (ID), while the remaining five are treated as out-of-distribution (OOD).

Training Data. The training data is the `simplelr_abel_level3to5` split of SimpleRL-Zoo-Data (Zeng et al., 2025), which contains mathematical problems at difficulty levels 3, 4, and 5. We sample 100 candidate problems from each difficulty level and, after filtering for informative training signals, retain at most 50 problems per model for training.

Comparison Methods. We evaluate the Base model, GRPO, DAPO with Forking Tokens, Reason-Maxxer, and SemCredit using identical training data, training frameworks, and evaluation protocols. Except for the untrained Base model, all methods are implemented through offline training with LoRA (Hu et al., 2022).

For broader comparison, we additionally report publicly available results for selected models from prior work, including SimpleRL-Zoo, Open-Reasoner-Zero, Online PasoDoble, Offline PasoDoble, Vanilla GRPO, Vanilla DAPO, DAPO w/ Forking Tokens, Archer, Entropy Env, EDGE-GRPO, and HAPO. These methods employ full-parameter training, and their results provide a broader context for assessing the performance of SemCredit relative to existing approaches.

Training Configuration. For methods trained under the unified offline setting, we use LoRA for parameter-efficient fine-tuning, with rank 32 and a scaling factor of 64. All models are trained for two epochs with a learning rate of 1×10^{-4}. We use a micro-batch size of 1 and eight gradient-accumulation steps, resulting in an effective batch size of 8. The maximum training sequence length is 8192 tokens. SemCredit uses Conan-embedding-v1 (Li et al., 2024) to encode reasoning segments and constructs consensus semantic states by combining fixed and adaptive clustering assignments. The main experiments use an entropy-selection fraction of $\rho = 0.2$, a decision-horizon fraction of $\tau = 0.5$, and a disagreement coefficient of $\lambda = 0.7$.

Evaluation Metrics. To obtain more stable evaluation results, we independently generate eight responses for each problem and report avg@8 on AIME 2024, AIME 2025, AIME 2026, and AMC 2023, which contain relatively few test problems. For the sufficiently large MATH-500 and OlympiadBench test sets, we generate one response per problem and report pass@1.

4.2 Main Results

SemCredit achieves the best overall performance under the unified offline setting. As shown in Table 1, SemCredit attains the highest Average score for all four models using the same hyperparameters without model-specific tuning. Its scores on Qwen2.5-7B, Qwen3-0.6B, Qwen3-4B, and

Table 1: Performance on six mathematical reasoning benchmarks (%). Bold and underlined values indicate the best and second-best results within each model, respectively.

Method	AIME24	AIME25	AIME26	AMC23	MATH-500	OlympiadBench	Average
Qwen2.5-7B							
Base	3.75	3.75	6.25	30.31	57.80	28.04	21.65
GRPO (Offline) (Shao et al., 2024)	7.08	5.42	4.17	35.00	64.80	31.16	24.60
DAPO w/ Forking Tokens (Offline) (Wang et al., 2025b)	10.42	<u>8.33</u>	5.83	**48.75**	69.20	32.94	<u>29.25</u>
ReasonMaxxer (Akgül et al., 2026)	8.75	7.08	<u>7.08</u>	44.69	69.00	34.12	28.45
SimpleRL-Zoo (Zeng et al., 2025)	8.80	–	–	37.50	65.60	<u>35.80</u>	–
Open-Reasoner-Zero (Hu et al., 2025)	<u>12.10</u>	–	–	46.30	**73.20**	**45.40**	–
SemCredit	**12.50**	**9.17**	**9.17**	<u>47.81</u>	<u>70.00</u>	34.42	**30.51**
Qwen3-8B							
Base	24.58	20.83	15.42	65.63	78.40	50.15	42.50
GRPO (Offline) (Shao et al., 2024)	29.58	22.92	20.00	67.50	79.60	<u>52.97</u>	45.43
DAPO w/ Forking Tokens (Offline) (Wang et al., 2025b)	26.25	20.83	14.58	66.56	80.60	48.66	42.92
ReasonMaxxer (Akgül et al., 2026)	33.75	22.50	<u>20.83</u>	**75.00**	81.00	50.74	<u>47.30</u>
Vanilla GRPO (Shao et al., 2024)	33.91	23.56	–	–	81.17	42.93	–
Vanilla DAPO (Yu et al., 2025a)	35.84	25.24	–	–	83.54	44.87	–
DAPO w/ Forking Tokens (Wang et al., 2025b)	<u>36.22</u>	25.85	–	–	82.75	46.27	–
Archer (Wang et al., 2025a)	34.76	23.33	–	–	82.57	42.69	–
Entropy Env (Cheng et al., 2026)	34.45	24.12	–	–	83.12	42.88	–
EDGE-GRPO (Zhang et al., 2025a)	35.15	24.12	–	–	83.33	43.35	–
HAPO (Liu et al., 2026)	**39.01**	<u>26.83</u>	–	–	**84.23**	45.26	–
SemCredit	35.83	**30.42**	**27.50**	<u>71.88</u>	<u>84.00</u>	**54.60**	**50.71**
Qwen3-4B							
Base	20.83	20.42	12.92	64.38	78.80	46.29	40.61
GRPO (Offline) (Shao et al., 2024)	23.33	<u>22.92</u>	<u>22.92</u>	64.38	80.80	<u>53.56</u>	44.65
DAPO w/ Forking Tokens (Offline) (Wang et al., 2025b)	<u>29.58</u>	22.08	<u>22.92</u>	<u>65.62</u>	**83.60**	<u>53.56</u>	<u>46.23</u>
ReasonMaxxer (Akgül et al., 2026)	21.67	20.42	15.42	64.69	75.40	47.18	40.80
SemCredit	**38.33**	**30.83**	**29.17**	**74.69**	<u>81.20</u>	**56.38**	**51.77**
Qwen3-0.6B							
Base	2.08	2.08	0.00	27.81	42.20	17.66	15.31
GRPO (Offline) (Shao et al., 2024)	2.50	<u>4.58</u>	0.83	<u>32.50</u>	47.60	17.21	17.54
DAPO w/ Forking Tokens (Offline) (Wang et al., 2025b)	<u>4.58</u>	<u>4.58</u>	<u>3.75</u>	**33.75**	<u>53.20</u>	<u>26.26</u>	<u>21.02</u>
ReasonMaxxer (Akgül et al., 2026)	2.92	3.33	0.83	30.63	45.20	21.66	17.43
Online PasoDoble (Zhang et al., 2025b)	1.11	1.11	–	–	38.70	11.53	–
Offline PasoDoble (Zhang et al., 2025b)	2.78	1.67	–	–	39.00	11.53	–
SemCredit	**5.00**	**8.75**	**5.00**	**33.75**	**57.00**	**27.45**	**22.83**

Qwen3-8B are 30.51%, 22.83%, 51.77%, and 50.71%, exceeding the strongest controlled baseline for each model by 1.26, 1.81, 5.54, and 3.41 percentage points, respectively. The shared configuration achieves gains across model scales under identical training data, frameworks, and evaluation protocols.

SemCredit remains competitive with full-parameter training methods despite using LoRA. The publicly reported methods included in Table 1—SimpleRL-Zoo, Open-Reasoner-Zero, PasoDoble, Vanilla GRPO, Vanilla DAPO, DAPO w/ Forking Tokens, Archer, Entropy Env, EDGE-GRPO, and HAPO—all employ full-parameter updates, whereas SemCredit updates only a small subset of parameters through LoRA. Despite its substantially smaller parameter-update budget, SemCredit matches or surpasses these methods on multiple commonly reported benchmarks. These results demonstrate that SemCredit can attain reasoning performance comparable to, or better than, existing reinforcement-learning methods while updating far fewer parameters. This comparison is intended primarily to illustrate the parameter efficiency and empirical competitiveness of SemCredit; because the publicly reported results may use different experimental configurations, it should not be interpreted as a strictly controlled comparison.

4.3 Results on Medical Reasoning

To examine whether SemCredit extends beyond mathematical reasoning to other natural-language reasoning domains, we conduct experiments on medical question-answering tasks. Like mathematical problems, medical questions require models to perform reasoning through natural language, but they differ substantially in their underlying knowledge domain and reasoning patterns. We use the MedQA training split as the training data. We evaluate Qwen2.5-7B on MedQA (Jin et al., 2021), MMLU-Med (Hendrycks et al., 2021a), and MedMCQA (Pal et al., 2022), treating MedQA as in-distribution and MMLU-Med and MedMCQA as out-of-distribution benchmarks. The results are reported in Table 2.

Table 2: Qwen2.5-7B performance on medical question-answering benchmarks (%). MedQA is in-distribution (ID), while MMLU-Med and MedMCQA are out-of-distribution (OOD). The suffix (Offline) identifies our offline adaptations. Bold denotes the best result on each benchmark.

Method	MedQA	MMLU-Med	MedMCQA
Base	44.85	65.75	50.99
GRPO (Offline) (Shao et al., 2024)	55.77	71.81	52.31
DAPO w/ Forking Tokens (Offline) (Wang et al., 2025b)	55.22	70.98	49.89
ReasonMaxxer (Akgül et al., 2026)	45.56	60.97	46.21
Ours	**56.48**	**72.91**	**53.77**

Table 3: **Component ablation of the decision-horizon constraint and credit multiplier.** Entropy trains only on the selected high-entropy tokens over the full trajectory; w/ DH additionally restricts eligible positions using $\tau = 0.5$; and w/ DH + CM further applies $m_{i,j}$ within the same window. Here, DH and CM are used only as compact configuration labels. Average is the unweighted mean over the six benchmarks, and Δ Avg reports its change, in percentage points, relative to the same-model Entropy baseline. Bold denotes the highest Average for each model.

Method	Benchmark accuracy (%)						Average	Δ Avg (pp)
	AIME24	AIME25	AIME26	AMC23	MATH-500	OlympiadBench		
Qwen2.5-7B								
Entropy	8.75	7.08	7.08	44.69	69.00	34.12	28.45	–
w/ DH	12.50	5.42	7.92	46.56	66.40	32.49	28.55	+0.10
w/ DH + CM	12.50	9.17	9.17	47.81	70.00	34.42	**30.51**	+2.06
Qwen3-8B								
Entropy	33.75	22.50	20.83	75.00	81.00	50.74	47.30	–
w/ DH	27.92	19.58	15.42	68.44	80.60	47.92	43.31	−3.99
w/ DH + CM	35.83	30.42	27.50	71.88	84.00	54.60	**50.71**	+3.41
Qwen3-4B								
Entropy	21.67	20.42	15.42	64.69	75.40	47.18	40.80	–
w/ DH	23.33	22.50	16.67	65.00	79.40	50.59	42.92	+2.12
w/ DH + CM	38.33	30.83	29.17	74.69	81.20	56.38	**51.77**	+10.97
Qwen3-0.6B								
Entropy	2.92	3.33	0.83	30.63	45.20	21.66	17.43	–
w/ DH	2.92	5.83	3.33	28.12	46.20	21.07	17.91	+0.48
w/ DH + CM	5.00	8.75	5.00	33.75	57.00	27.45	**22.83**	+5.40

As shown in Table 2, SemCredit achieves the best performance on all three medical benchmarks, with scores of 56.48%, 72.91%, and 53.77%, respectively. Relative to the strongest controlled baseline on each benchmark, these results represent improvements of 0.71, 1.10, and 1.46 percentage points. These findings indicate that semantic-aware local credit assignment is effective not only for mathematical reasoning tasks with verifiable answers, but also for medical natural-language reasoning tasks in a distinct knowledge domain.

5 Ablation and Analysis

5.1 The Role of Semantic-Trie Credit

Table 3 isolates the contribution of semantic-trie credit within the early decision horizon. Entropy identifies uncertain token positions, and the decision-horizon constraint restricts updates to the earlier part of each rollout. Neither signal, however, distinguishes reasoning segments associated with different outcomes. The semantic trie supplies this distinction: it groups similar segments across rollouts, compares the terminal outcomes associated with their branches, and uses the resulting local advantages to adjust credit within each trajectory.

Under the same decision horizon, adding semantic-trie credit improves the six-benchmark Average by 1.96, 7.40, 8.85, and 4.92 percentage points for Qwen2.5-7B, Qwen3-8B, Qwen3-4B, and Qwen3-0.6B, respectively. Across the 24 model–benchmark comparisons, 23 scores increase and one remains unchanged. DH alone yields mixed results, including a 3.99-point decrease for Qwen3-8B. The consistent gains after adding trie-based credit show why selecting early uncertain tokens is insufficient on its own: tokens in different semantic segments of the same rollout should not all inherit the same trajectory-level signal. By conditioning their update strength on path-specific outcome statistics, SemCredit gives stronger credit to segments whose local signal agrees with the final outcome and attenuates credit when the signals conflict.

5.2 Why Prioritize Early High-Entropy Tokens?

Figure 4 examines how far into a rollout high-entropy tokens should remain eligible for semantic credit. All four models achieve their highest six-benchmark Average at $\tau = 0.5$ among the tested cutoffs. Restricting updates to the first quarter performs worse, indicating that useful uncertain decisions also occur beyond the very beginning of a response. Notably, extending the update range to the entire trajectory also lowers Average performance relative to $\tau = 0.5$, by 1.08–8.53 percentage points across the four models. Thus, the observed benefit does not come from updating as many high-entropy tokens as possible. It comes from concentrating updates where an uncertain choice still has substantial opportunity to shape the subsequent reasoning path. The benchmark-level results in Appendix Figure 5 show variation in the preferred cutoff for individual benchmarks, while the macro-average peak is shared by all four models.

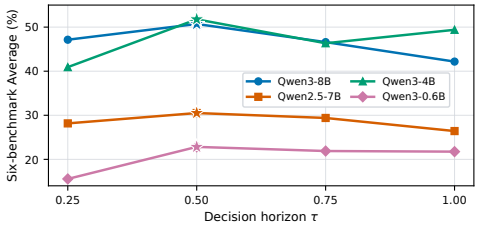

Figure 4: **Decision-horizon ablation.** Stars mark the highest score for each model among the tested settings.

Table 4 connects this positional result to the structure of the semantic trie. At $\tau = 0.5$, the eligible region already covers 97.3–100.0% of the identified, supported turning points across the four models. This concentration makes the decision horizon interpretable: the first half of the response contains nearly all observed locations where rollouts sharing a semantic prefix branch into different subsequent paths. The performance peak in Figure 4 is consistent with retaining these early branches while avoiding updates across the entire continuation.

The same table also shows how sparse the resulting supervision is. At $\tau = 0.5$, only 11.12–12.13% of completion tokens receive credit, compared with 19.05–19.47% when the horizon covers the full response. Despite using substantially fewer update positions, SemCredit at $\tau = 0.5$ achieves the highest Average for every model in the horizon sweep. Relative to the full-trajectory Entropy baseline in Table 3, its Average is higher by 2.06, 3.41, 10.97, and 5.40 percentage points for Qwen2.5-7B, Qwen3-8B, Qwen3-4B, and Qwen3-0.6B, respectively. These results show that effective credit assignment depends on selecting consequential reasoning positions and weighting their semantic paths, rather than distributing updates across a larger number of uncertain tokens.

Table 4: **Turning-point coverage and decision-token density under different values of τ.** A turning point (TP) is a supported branch node at which trajectories sharing a semantic prefix diverge. Decision-token density is the percentage of completion tokens that receive credit after selecting the top $\rho = 20\%$ high-entropy tokens per rollout and applying the cutoff τ.

	$\tau = 0.25$		$\tau = 0.50$		$\tau = 0.75$		$\tau = 1.00$	
Model	**TP (%)**	**Density (%)**	**TP (%)**	**Density (%)**	**TP (%)**	**Density (%)**	**TP (%)**	**Density (%)**
Qwen2.5-7B	92.4	7.71	97.3	12.13	97.3	15.58	100.0	19.05
Qwen3-8B	100.0	5.88	100.0	11.27	100.0	15.70	100.0	19.21
Qwen3-4B	100.0	5.51	100.0	11.12	100.0	15.78	100.0	19.20
Qwen3-0.6B	98.1	6.19	99.5	11.49	100.0	15.72	100.0	19.47

6 Related Work

Uncertainty-based sparse updates. Outcome-supervised policy optimization uses trajectory-level rewards (Schulman et al., 2017; Shao et al., 2024), while high-entropy token selection and ReasonMaxxer focus updates on uncertain decisions (Wang et al., 2025b; Akgül et al., 2026). Related work uses entropy to guide exploration, shape updates, select or compress reasoning segments, and regulate training dynamics (Cheng et al., 2026; Liu et al., 2026; Chen et al., 2026; Li et al., 2026d; Cui et al., 2025; Jin et al., 2026; Agarwal et al., 2025; Xiong et al., 2026). We retain entropy for sparse selection, but use semantic structure and descendant outcomes to determine credit strength.

Online branching and credit assignment. VinePPO and TreeRL estimate intermediate values through sampled continuations, while ROSE incorporates semantics into branch selection (Kazemnejad et al., 2024; Hou et al., 2025; Zhao et al., 2026). Other tree-based methods derive local advantages from sibling or descendant outcomes and incorporate search, pruning, or adaptive rollout allocation (Li et al., 2026b; Yang et al., 2025b; Ji et al., 2026; Zong et al., 2026; Zhang et al., 2024; Du et al., 2026). GraphPO merges semantically similar paths during generation to guide expansion (Zhan et al., 2026). These methods couple credit estimation to online sampling; our approach reconstructs structure from completed rollouts without additional continuation sampling.

Post-hoc trajectory structure. TEMPO aggregates outcomes in an exact token-prefix trie (Tran et al., 2026), whereas GiGPO, SALT, and T-STAR recover shared states, transitions, or functional structure across trajectories for localized credit (Feng et al., 2025; Li et al., 2026a;c). Representation-based studies also recover semantic states or redundant paths for reasoning analysis and test-time computation (Yu et al., 2025b; Lee et al., 2025; Sun et al., 2026). Our method induces consensus semantic states from free-form reasoning segments and organizes their sequences into a prefix trie, combining semantic sharing with path-dependent outcome estimates for offline token-level credit assignment.

7 Conclusion

We presented SemCredit, a fine-grained credit-assignment method for language-model reasoning under verifiable terminal rewards. Our analysis shows that high-entropy tokens are more prevalent early in reasoning, and that replacing an early high-entropy token is more likely to change final-answer correctness. This suggests that even among the sparse set of high-entropy tokens, a smaller subset of early decisions deserves particular attention. SemCredit focuses updates on these positions and uses a semantic prefix trie to distinguish the reasoning paths they lead to. By aggregating terminal outcomes along shared semantic paths, it assigns different credit to tokens in different reasoning segments without step-level supervision. Across mathematical and medical reasoning tasks, SemCredit improves performance while updating only a small fraction of response tokens. These findings show that effective credit assignment depends on both selecting consequential positions and evaluating the semantic paths that follow them.

8 Limitations

SemCredit estimates local credit from associations between semantic paths and terminal outcomes in a finite set of offline rollouts; these estimates do not establish that a segment causes an outcome and may be noisy for rarely visited nodes. Its states depend on paragraph segmentation and embedding-based clustering, so grouping errors can affect the resulting credit.

AI USE STATEMENT

We used generative AI tools to edit the manuscript for clarity and style. These tools were not used to generate research ideas, experimental results, analyses, or scientific claims. The authors reviewed and verified all AI-assisted edits and take full responsibility for the final manuscript.

REFERENCES

Shivam Agarwal, Zimin Zhang, Lifan Yuan, Jiawei Han, and Hao Peng. The unreasonable effectiveness of entropy minimization in llm reasoning. *Advances in Neural Information Processing Systems*, 38, 2025.

Ömer Faruk Akgül, Rajgopal Kannan, Willie Neiswanger, and Viktor Prasanna. Rethinking rl for llm reasoning: It's sparse policy selection, not capability learning. *arXiv preprint arXiv:2605.06241*, 2026.

Xinzhu Chen, Xuesheng Li, Zhongxiang Sun, and Weijie Yu. Beyond high-entropy exploration: Correctness-aware low-entropy segment-based advantage shaping for reasoning llms. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 32970–32984, 2026.

Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Xin Zhao, Zhenliang Zhang, and Furu Wei. Reasoning with exploration: An entropy perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 30377–30385, 2026. doi: 10.1609/aaai.v40i36.40290.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, et al. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025.

Jasper Dekoninck, Nikola Jovanović, Tim Gehrunger, Kári Rögnvaldsson, Ivo Petrov, Chenhao Sun, and Martin Vechev. Beyond benchmarks: MathArena as an evaluation platform for mathematics with LLMs. *arXiv preprint arXiv:2605.00674*, 2026. URL https://arxiv.org/abs/2605.00674.

Linxuan Du, Guangquan Xue, Xiaobo Liang, Qipeng Huang, Yuyang Ding, Xinyu Shi, Zhang Yijun, Ji Qi, Wenpeng Zhu, Juntao Li, et al. Escaping the echo trap: On credit assignment failure in multi-turn llm self-reflection. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 35393–35405, 2026.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for llm agent training. *Advances in Neural Information Processing Systems*, 38:46375–46408, 2025.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. 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. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.211. URL https://aclanthology.org/2024.acl-long.211/.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2021b. URL https://openreview.net/forum?id=7Bywt2mQsCe.

Zhenyu Hou, Ziniu Hu, Yujiang Li, Rui Lu, Jie Tang, and Yuxiao Dong. Treerl: Llm reinforcement learning with on-policy tree search. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12355–12369, 2025.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-Reasoner-Zero: An open source approach to scaling up reinforcement learning on the base model. *arXiv preprint arXiv:2503.24290*, 2025.

Yuxiang Ji, Ziyu Ma, Yong Wang, Guanhua Chen, Xiangxiang Chu, and Liaoni Wu. Tree search for llm agent reinforcement learning. In *International Conference on Learning Representations*, 2026.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021. doi: 10.3390/app11146421. URL https://doi.org/10.3390/app11146421.

Renren Jin, Pengzhi Gao, Yuqi Ren, Zhuowen Han, Tongxuan Zhang, Wuwei Huang, Wei Liu, Jian Luan, and Deyi Xiong. Revisiting entropy in reinforcement learning for large reasoning models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 25300–25322, 2026.

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. Vineppo: Refining credit assignment in rl training of llms. *arXiv preprint arXiv:2410.01679*, 2024.

Sungjae Lee, Hoyoung Kim, Jeongyeon Hwang, Eunhyeok Park, and Jungseul Ok. Efficient latent semantic clustering for scaling test-time computation of llms. In *EMNLP (Findings)*, pp. 24126–24144, 2025.

Jiazheng Li, Yawei Wang, Qiaojing Yan, Yijun Tian, Zhichao Xu, Huan Song, Panpan Xu, and Lin Lee Cheong. Salt: Step-level advantage assignment for long-horizon agents via trajectory graph. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 4709–4725, 2026a.

Shiyu Li, Yang Tang, Shizhe Chen, and Xi Chen. Conan-embedding: General text embedding with more and better negative samples. *arXiv preprint arXiv:2408.15710*, 2024. URL https://arxiv.org/abs/2408.15710.

Yizhi Li, Qingshui Gu, Zhoufutu Wen, Ziniu Li, Ruibin Yuan, Tianshun Xing, Shuyue Guo, Tianyu Zheng, Xin Zhou, Xingwei Qu, et al. Treepo: Enhancing policy efficacy and inference efficiency with tree modeling. In *Forty-third International Conference on Machine Learning*, 2026b.

Yu Li, Sizhe Tang, and Tian Lan. Reason in chains, learn in trees: Self-rectification and grafting for multi-turn agent policy optimization. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 4673–4686, 2026c.

Zeju Li, Jianyuan Zhong, Ziyang Zheng, Xiangyu Wen, Zhijian Xu, Yingying Cheng, Fan Zhang, and Qiang Xu. Making slow thinking faster: Compressing llm chain-of-thought via step entropy. In *International Conference on Learning Representations*, 2026d.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2305.20050.

Zheng Liu, Mengjie Liu, Siwei Wen, Mengzhang Cai, Bin Cui, Conghui He, and Wentao Zhang. Heterogeneous adaptive policy optimization: Tailoring optimization to every token's nature. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6023–6045, 2026.

Maxwell-Jia. AIME 2024 dataset, 2024. URL https://huggingface.co/datasets/Maxwell-Jia/AIME_2024.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pp. 248–260. PMLR, 2022. URL https://proceedings.mlr.press/v174/pal22a.html.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

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. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26872–26887, 2026.

Hieu Tran, Zonghai Yao, and Hong Yu. Exploiting tree structure for credit assignment in reinforcement learning with large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 10795–10810, 2026.

Jiakang Wang, Runze Liu, Fuzheng Zhang, Xiu Li, Guorui Zhou, and Ling Pan. Stabilizing knowledge, promoting reasoning: Dual-token constraints for RLVR. *arXiv preprint arXiv:2507.15778*, 2025a.

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, 2025b.

Xuan Xiong, Huan Liu, Li Gu, Zhixiang Chi, Yue Qiu, Yuanhao Yu, and Yang Wang. Etr: Entropy trend reward for efficient chain-of-thought reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17576–17594, 2026.

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, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a. URL https://arxiv.org/abs/2505.09388.

Zhicheng Yang, Zhijiang Guo, Yinya Huang, Xiaodan Liang, Yiwei Wang, and Jing Tang. Treerpo: Tree relative policy optimization. *arXiv preprint arXiv:2506.05183*, 2025b.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing

Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025a.

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

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. SimpleRL-Zoo: Investigating and taming zero reinforcement learning for open base models in the wild. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=vSMCBUgrQj.

Yuliang Zhan, Xinyu Tang, Jian Li, Dandan Zheng, Weilong Chai, Jingdong Chen, Jun Zhou, Ge Wu, Wenyue Tang, and Hao Sun. Graphpo: Graph-based policy optimization for reasoning models. *arXiv preprint arXiv:2606.18954*, 2026.

Dan Zhang, Sining Zhoubian, Ziniu Hu, Yisong Yue, Yuxiao Dong, and Jie Tang. Rest-mcts*: Llm self-training via process reward guided tree search. *Advances in Neural Information Processing Systems*, 37:64735–64772, 2024.

Xingjian Zhang, Siwei Wen, Wenjun Wu, and Lei Huang. EDGE-GRPO: Entropy-driven GRPO with guided error correction for advantage diversity. *arXiv preprint arXiv:2507.21848*, 2025a.

Yifan Zhang and Math-AI Team. American invitational mathematics examination (AIME) 2025, 2025. URL https://huggingface.co/datasets/math-ai/aime25.

Zhengxin Zhang, Chengyu Huang, Aochong Oliver Li, and Claire Cardie. Better LLM reasoning via dual-play. *arXiv preprint arXiv:2511.11881*, 2025b.

Ziqi Zhao, Zhaochun Ren, Jiahong Zou, Liu Yang, Zhiwei Xu, Xuri Ge, Zhumin Chen, Xinyu Ma, Daiting Shi, Shuaiqiang Wang, et al. Reinforced efficient reasoning via semantically diverse exploration. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 47994–48007, 2026.

Zefang Zong, Dingwei Chen, Yang Li, Qi Yi, Bo Zhou, Chengming Li, Bo Qian, Peng Chen, and Jie Jiang. At^{2}po: Agentic turn-based policy optimization via tree search. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24120–24143, 2026.

zwhe99. AMC 2023 dataset, 2023. URL https://huggingface.co/datasets/zwhe99/amc23.

A Reasoning Paths at an Early Token Fork

Two Reasoning Paths from an Early Token Fork

Problem. Two concentric circles have equal-length arcs subtending 45° on the smaller circle and 36° on the larger circle. Find the smaller-to-larger area ratio.

Model: Qwen3-0.6B **Fork:** token 3 (approximately 0.44% of the trajectory)

Shared prefix `We are given`

Path A Original rollout

Next token: `two`

`We are given two concentric circles ...`

1. Assign the same radius

Both circles initially use r, giving an inconsistent equal-arc equation:

$$r\frac{\pi}{4} = r\frac{\pi}{5}.$$

↓

2. Repair the radius assignment

The rollout introduces R for the larger circle and rewrites the equation:

$$r\frac{\pi}{4} = R\frac{\pi}{5}.$$

↓

3. Express the inverse ratio

The algebra is correct, but the explicit ratio is larger over smaller:

$$\frac{r}{4} = \frac{R}{5} \implies \frac{R}{r} = \frac{5}{4}.$$

↓

4. Square without inverting

The target requires $(r/R)^2$. The rollout instead squares $5/4$:

$$\frac{A_{\text{small}}}{A_{\text{large}}} \overset{\text{error}}{=} \left(\frac{5}{4}\right)^2.$$

Incorrect $\boxed{\frac{25}{16}}$

Path B Alternative-token sibling

Next token: `:\n\n`

`We are given:`
`- Two concentric circles ...`

1. Assign distinct radii

The rollout lists the givens and defines a separate radius for each circle:

$$\text{smaller} \mapsto r, \quad \text{larger} \mapsto R.$$

↓

2. Equate the arc lengths

The first equal-arc equation already uses the corresponding radii:

$$r\frac{\pi}{4} = R\frac{\pi}{5}.$$

↓

3. Express the target ratio

It solves for r in terms of R, giving the smaller-to-larger ratio directly:

$$r = \frac{4}{5}R \implies \frac{r}{R} = \frac{4}{5}.$$

↓

4. Square the matching ratio

Substitution preserves the requested smaller-to-larger direction:

$$\frac{A_{\text{small}}}{A_{\text{large}}} = \left(\frac{r}{R}\right)^2 = \left(\frac{4}{5}\right)^2.$$

Correct $\boxed{\frac{16}{25}}$

Where the paths diverge. The token fork changes the opening from a sentence to a list. The first mathematical divergence is the radius assignment: one shared r versus distinct r and R. Both paths then reach the same correct arc-length equation, but continue with opposite ratio directions: R/r in Path A and r/R in Path B.

During autoregressive generation, the divergent token marks where the two continuations split, but the token alone does not identify the reasoning operation that eventually fails. Reading the subsequent reasoning segments reveals two differences: the original rollout initially assigns the same radius to both circles, then repairs its arc-length equation but squares the inverse radius ratio; the alternative-token sibling assigns distinct radii from the outset and carries the requested smaller-to-larger ratio through to the answer. This motivates our paragraph-level view of reasoning paths. We group semantically similar segments into shared states and compare the terminal outcomes of rollouts that follow approximately the same sequence of states. The resulting path-conditioned outcome statistics indicate which semantic transitions are associated with correct or incorrect answers, providing local credit for the corresponding reasoning segments.

B ROLLOUT AND SEMANTIC-STATE DEFINITIONS

This appendix gives the formal definitions underlying the rollout representation, token uncertainty, and consensus semantic-state induction in Section 3.

B.1 PROBLEM FORMULATION

For each problem, we collect a group of reasoning rollouts and their terminal correctness rewards:

$$\begin{aligned} \mathcal{G}_{\boldsymbol{q}} &= ((\boldsymbol{o}_i, r_i))_{i=1}^{K}, \\ \boldsymbol{o}_i &= (o_{i,1}, \ldots, o_{i,T_i}), \qquad r_i \in \{0,1\}. \end{aligned} \tag{10}$$

$\boldsymbol{q}$ denotes the input problem, K is the number of independently sampled rollouts for that problem, $\boldsymbol{o}_i$ is the i-th rollout, T_i is its number of generated tokens, and r_i is the terminal reward indicating whether the final answer is correct. The outer parentheses define an indexed collection rather than a set, so repeated sampled rollouts remain distinct observations. All semantic structure and credit signals described below are constructed independently within each rollout group $\mathcal{G}_{\boldsymbol{q}}$.

We use the frozen base policy to measure the predictive uncertainty at every generated position:

$$\begin{aligned} H_{i,t} &= -\sum_{v=1}^{V} p_{i,t,v} \log p_{i,t,v}, \\ p_{i,t,v} &= \pi_0(v \mid \boldsymbol{q}, \boldsymbol{o}_{i,<t}). \end{aligned} \tag{11}$$

Here, $H_{i,t}$ is the predictive entropy at position t of rollout i, V is the vocabulary size, $p_{i,t,v}$ is the probability assigned to vocabulary item v, π_0 is the frozen base policy, and $\boldsymbol{o}_{i,<t}$ denotes the generated prefix preceding position t. Entropy is used only to select uncertain token positions; it does not determine whether the corresponding reasoning segment is beneficial or harmful.

B.2 CONSENSUS SEMANTIC-STATE INDUCTION

We first divide every rollout into paragraph-level reasoning segments and pool all segments generated for the same problem. A sentence encoder maps each segment into a normalized embedding:

$$\begin{aligned} \mathcal{I}_{\boldsymbol{q}} &= \{(i,\ell) : 1 \le i \le K,\ 1 \le \ell \le L_i\}, \\ \mathcal{S}_{\boldsymbol{q}} &= (s_{i,\ell})_{(i,\ell)\in\mathcal{I}_{\boldsymbol{q}}}, \\ \boldsymbol{e}_{i,\ell} &= \frac{f_{\mathrm{emb}}(s_{i,\ell})}{\|f_{\mathrm{emb}}(s_{i,\ell})\|_2}. \end{aligned} \tag{12}$$

Here, $\mathcal{I}_{\boldsymbol{q}}$ indexes all segment occurrences associated with problem $\boldsymbol{q}$, $\mathcal{S}_{\boldsymbol{q}}$ is the corresponding indexed collection, $s_{i,\ell}$ is the ℓ-th segment of rollout i, and L_i is the number of segments in that rollout. This indexed representation preserves distinct occurrences even when two segments have identical text. The function f_{emb} is the segment encoder, and $\boldsymbol{e}_{i,\ell}$ is the corresponding unit-normalized embedding; the encoder output is assumed to have nonzero norm.

A single clustering rule can either over-merge distinct reasoning operations or fragment paraphrases into unsupported states. To reduce this dependence on one geometric criterion, we construct a fixed and an adaptive partition of the same segment pool. The fixed labels z_{fix} are produced by average-linkage agglomerative clustering under a global cosine-distance threshold δ_{fix}. For the adaptive partition, let $a = (i,\ell) \in \mathcal{I}_{\boldsymbol{q}}$ index a segment occurrence, and write $s_a = s_{i,\ell}$ and $\boldsymbol{e}_a = \boldsymbol{e}_{i,\ell}$. We

construct a weighted mutual-nearest-neighbor graph:

$$\mathcal{P}_{\boldsymbol{q}} = \{\{a, b\} \subset \mathcal{I}_{\boldsymbol{q}} : a \neq b\},$$

$$\delta_{\boldsymbol{q}} = \mathrm{Quantile}_{\eta}\left((d_{\cos}(\boldsymbol{e}_a, \boldsymbol{e}_b))_{\{a,b\}\in\mathcal{P}_{\boldsymbol{q}}}\right),$$

$$\mathcal{E}_{\mathrm{adapt}} = \left\{\{a, b\} \in \mathcal{P}_{\boldsymbol{q}} \;\middle|\; \begin{array}{c} a \in \mathcal{N}_k(b),\ b \in \mathcal{N}_k(a), \\ d_{\cos}(\boldsymbol{e}_a, \boldsymbol{e}_b) \leq \delta_{\boldsymbol{q}} \end{array}\right\}, \tag{13}$$

$$\omega_{a,b} = \exp\left(-\frac{d_{\cos}(\boldsymbol{e}_a, \boldsymbol{e}_b)}{\max(\delta_{\boldsymbol{q}}, 10^{-6})}\right), \quad \{a, b\} \in \mathcal{E}_{\mathrm{adapt}}.$$

Here, $\mathcal{P}_{\boldsymbol{q}}$ is the collection of unordered segment pairs, $\mathcal{E}_{\mathrm{adapt}}$ is the edge set of the adaptive graph, and $\mathcal{N}_k(a)$ contains the $\min(k, |\mathcal{I}_{\boldsymbol{q}}| - 1)$ nearest neighbors of segment a, excluding a itself. The function $d_{\cos}$ is cosine distance, $\delta_{\boldsymbol{q}}$ is a problem-specific distance cutoff, η is the quantile used to determine that cutoff from all within-problem segment pairs, and $\omega_{a,b}$ is the graph edge weight. We apply weighted Louvain community detection with a fixed resolution parameter and random seed to obtain the adaptive labels z_{adapt}. If fewer than two segment occurrences are available, or if the resulting graph has no edges, each available segment is assigned to its own adaptive cluster.

We define a consensus semantic state by intersecting the two cluster assignments:

$$z(s_a) = \big(z_{\mathrm{fix}}(s_a), z_{\mathrm{adapt}}(s_a)\big),$$

$$s_a \sim_{\mathrm{con}} s_b \iff \begin{cases} z_{\mathrm{fix}}(s_a) = z_{\mathrm{fix}}(s_b), \\ z_{\mathrm{adapt}}(s_a) = z_{\mathrm{adapt}}(s_b), \end{cases} \tag{14}$$

$$S_\chi = \{s_a : z(s_a) = \chi\}.$$

Each rollout is then represented as an ordered sequence of consensus states:

$$\boldsymbol{S}_i = \mathrm{Collapse}\left(S_{z(s_{i,1})}, \ldots, S_{z(s_{i,L_i})}\right). \tag{15}$$

Here, $\boldsymbol{S}_i$ is the semantic-state sequence of rollout i, and $\mathrm{Collapse}(\cdot)$ removes adjacent repetitions of the same state. This prevents repeated paraphrases within one rollout from creating artificial self-transitions.

C Decision-Horizon Ablation by Benchmark

Figure 5 expands the macro-average decision-horizon ablation in Figure 4 into results for each benchmark and model.

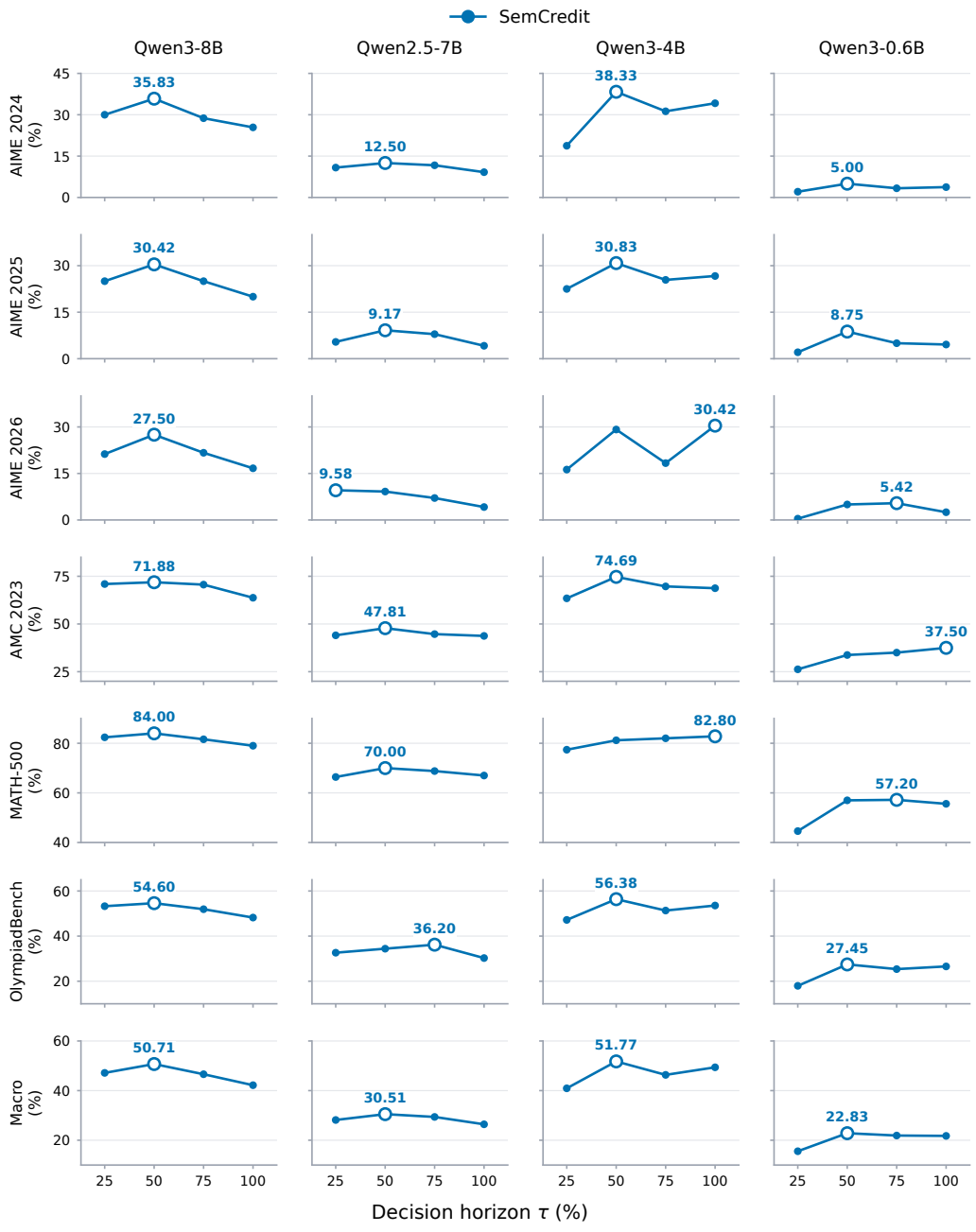

Figure 5: **Decision-horizon ablation by benchmark.** Accuracy across six mathematical reasoning benchmarks and four models as the decision-horizon cutoff τ varies from 0.25 to 1.00. Open circles mark the highest score among the tested cutoffs for each model–benchmark pair; the bottom row shows the unweighted average over the six benchmarks.