SLPO: SCALING LATENT REASONING VIA A SURROGATE POLICY

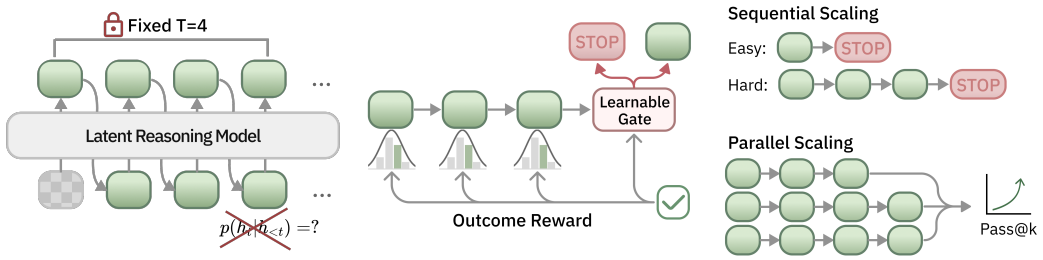
Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards has become the predominant recipe for eliciting test-time scaling in explicit Chain-of-Thought reasoners. Yet this scaling path remains computationally costly, since intermediate reasoning must be externalized as natural-language tokens. Latent reasoning instead carries intermediate computation as continuous vectors and already matches or surpasses explicit CoT at far shorter horizons. Despite this promise, latent reasoners remain largely imitation-bound, while explicit CoT has already moved past imitation via outcome-reward RL. Latent trajectories lack a tractable per-step likelihood and an adaptive stopping interface under fixed thinking budgets, so outcome rewards cannot elicit latent test-time scaling. We introduce Surrogate Latent Policy Optimization (SLPO) to bring outcome-reward RL to autoregressive latent reasoners: a differentiable surrogate policy interface over latent transitions for trajectory-level credit assignment, and a correctness-supervised stopping head that outcome-reward optimization refines into a variable-horizon policy. Across two continuous latent reasoners, two backbones, and three held-out benchmarks, SLPO improves Pass@8 and Pass@16 in all 12 backbone–dataset settings, with gains of up to 12.07 percentage points. SLPO further transfers to soft-token inference and learns difficulty-adaptive computation, allocating longer latent trajectories to harder instances.

Code is provided as the supplementary material for reproducibility.

1 INTRODUCTION



(a) Latent Reasoning (b) SLPO (c) Latent Test-Time Scaling

Figure 1: SLPO brings outcome-reward RL to latent reasoning and enables latent test-time scaling.

Large language models have demonstrated strong performance on complex reasoning tasks with chain-of-thought (CoT) reasoning (Wei et al., 2022; Wang et al., 2023). By eliciting intermediate steps in natural language, CoT decomposes difficult problems and allocates additional test-time computation before generating the final answer (Muennighoff et al., 2025; Snell et al., 2025). Recent reasoning models scale this additional test-time computation considerably further, either sequentially, by extending and revising ever longer chains (Guo et al., 2025); or in parallel, by sampling and searching among ever more trials at inference (Wang et al., 2023; Snell et al., 2025). These inference-time strategies have become known as test-time scaling (TTS), while reinforcement learning with verifiable rewards (RLVR) has emerged as the predominant training-time paradigm for eliciting and optimizing such capabilities (Wen et al., 2025; Zhang et al., 2025a).

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.

However, this test-time scaling path is computationally costly. Explicit CoT externalizes intermediate reasoning as natural-language tokens, but many of these tokens support linguistic coherence, exposition, or redundancy rather than the underlying problem-solving state (Sui et al., 2025; Wang et al., 2026a). Each token still incurs the full cost of autoregressive generation, fundamentally limiting the scalability of explicit reasoning. This motivates *continuous CoT*, or *latent reasoning*. As illustrated in Fig. 1a, latent reasoning carries intermediate computation as continuous vectors in hidden space rather than as discrete language tokens (Hao et al., 2025). Recent latent-reasoning models already match or surpass explicit CoT on standard benchmarks with far shorter intermediate computation (Chen et al., 2025; Shen et al., 2025; Tan et al., 2025; Wei et al., 2025; Wang et al., 2026b;a).

Despite the promise, latent reasoners remain imitation-bound: they are largely trained to align their intermediate states with manually-selected (Hao et al., 2025; Shen et al., 2025), compressed (Tan et al., 2025; Wei et al., 2025), or rendered (Wang et al., 2026b;a) explicit CoT representations, reproducing predefined trajectories rather than optimizing latent computation for task outcomes. Explicit CoT has already moved beyond this imitation stage: RLVR (Guo et al., 2025; Wen et al., 2025) is now a mature recipe for scaling past teacher imitation in explicit reasoning. Rollout length is inherently variable, allowing candidate trajectories to span different reasoning horizons. Because each reasoning step is a token sampled from the vocabulary distribution, policy gradients can assign trajectory-level credit through a tractable per-step likelihood, so outcome rewards can reshape both the reasoning content and test-time compute allocation. However, latent reasoning has not yet reached an equivalent outcome-optimization stage. Two obstacles prevent this extension: (1) Existing latent reasoners prescribe a fixed thinking budget throughout training and inference (Hao et al., 2025; Shen et al., 2025), freezing the compute horizon that RL would otherwise optimize. (2) As shown in Fig. 1a, intermediate steps propagate as continuous vectors that bypass the vocabulary distribution, so latent trajectories lack a tractable action likelihood through which outcome rewards can assign transition-level credit (Yue et al., 2025; Nguyen et al., 2025). In sum, absent both an adaptive stopping interface and a tractable latent-transition likelihood, outcome rewards cannot elicit latent test-time scaling.

In this work, we introduce **S**urrogate **L**atent **P**olicy **O**ptimization (SLPO), which instantiates both interfaces to bring outcome-level RL to latent reasoning, as illustrated by Fig. 1b. To make the rollout horizon learnable, we equip the reasoner with a stopping head and initialize it through a correctness-supervised cold start, providing an initial stopping policy that outcome-reward optimization subsequently refines for adaptive computation. To optimize the resulting trajectories, SLPO defines a surrogate likelihood in hidden space, allowing policy-gradient estimators to assign credit to vector-based reasoning without vocabulary-level action probabilities. Together, these components enable outcome rewards to jointly optimize latent transitions and stopping decisions in autoregressive latent reasoners, using standard algorithms such as RLOO or GRPO. The resulting policies scale latent computation with problem difficulty and improve held-out accuracy (Fig. 1c).

We summarize our main contributions as follows:

- We identify two barriers to outcome-reward optimization in latent reasoning: fixed thinking budgets constrain candidate trajectories to a prescribed horizon, while continuous latent transitions lack a tractable policy likelihood.
- We introduce SLPO, equipping latent reasoners with a learnable stopping policy for variable-length rollouts and a surrogate likelihood for credit assignment over continuous latent transitions, enabling outcome rewards to jointly optimize latent reasoning and its compute allocation.
- Across two latent reasoners, two backbones, and three held-out benchmarks, SLPO improves both Pass@8 and Pass@16 in all 12 backbone–dataset settings, with gains of up to 12.07 percentage points. The gains persist across RLOO and GRPO, transfer to soft-token inference, and accompany difficulty-adaptive latent computation.

2 PRELIMINARIES

We begin by formalizing RL over explicit CoT and reviewing latent CoT, and outline dropout-induced stochastic latent rollouts that underlie the methodology below. For each training instance $(x_i, a_i^\star)$, x_i denotes the input problem and $a_i^\star$ denotes the reference answer.

Explicit CoT and RL. In explicit CoT, a model samples a discrete reasoning trace autoregressively from the next-token distribution (Wei et al., 2022; Wang et al., 2023). At reasoning step t, $y_t \sim \pi_\theta(\cdot \mid x_i, y_{<t})$; after the reasoning trace is sampled, the model generates the final answer tokens conditioned on the trace. The resulting rollout likelihood is therefore available directly from token probabilities:

$$\log \pi_\theta(y_{1:T}, \hat{a}_i \mid x_i) = \sum_{t=1}^{T} \log \pi_\theta(y_t \mid x_i, y_{<t}) + \sum_{s=1}^{|\hat{a}_i|} \log \pi_\theta(\hat{a}_{i,s} \mid x_i, y_{1:T}, \hat{a}_{i,<s}).$$

Given a verifiable reward $R(\hat{a}_i, a_i^\star)$, policy-gradient methods assign credit to sampled rollouts by weighting this log-probability term with an advantage such as $\widehat{A}_i = R(\hat{a}_i, a_i^\star) - b_i$. Thus explicit CoT exposes the core policy interface needed by RL: the sampled reasoning trajectory already comes with token-level action probabilities.

Latent CoT. Continuous CoT replaces the discrete reasoning trace with an autoregressive trajectory in hidden space (Hao et al., 2025; Chen et al., 2025). We write this trajectory as $h_{i,1:T_{\max}} = (h_{i,1}, \ldots, h_{i,T_{\max}})$, where $h_{i,t} \in \mathbb{R}^d$ is the latent state at step t. The latent update can be written abstractly as

$$h_{i,t} = f_\theta(x_i, h_{i,<t}),$$

with the answer decoder conditioning on the completed latent trajectory. During latent reasoning, the intermediate computation bypasses the vocabulary distribution and language-head sampling, which removes the token-policy likelihood that explicit CoT exposes to RL. Prior work has shown that latent reasoning can nevertheless be made stochastic by inference under dropout (You et al., 2026). Such stochasticity supplies the exploration substrate for RL, and we denote a sampled trajectory by $\tilde{h}_{i,1:T_{\max}}$. However, how to construct a tractable log-likelihood for the realized latent states is still an open question.

3 METHODOLOGY

Existing latent reasoners typically prescribe a fixed thinking budget, so the model cannot decide when to terminate latent computation on its own. We therefore proceed in two stages. We first attach a stopping head and, for each sampled trajectory, supervise its distribution over stopping times with final-answer correctness evaluated at every candidate length (Sec. 3.1). This introduces a learnable continue-versus-stop decision before outcome-reward optimization. SLPO then makes outcome rewards usable for eliciting latent reasoning scaling: a tractable surrogate likelihood scores continuous latent transitions, and the stopping likelihood is included in the policy objective, so verifiable rewards jointly shape latent reasoning and compute allocation (Sec. 3.2). We close with the inference procedure (Sec. 3.3).

3.1 STOPPING-GATE COLD START

We augment the latent reasoner with a stopping head g_θ over latent states. Given a latent state h, the head outputs

$$s_\theta(h) = \sigma(g_\theta(h)),$$

which is the probability of terminating latent computation at that state. For an input x_i, the released backbone samples N stochastic latent trajectories up to a maximum budget:

$$h_{i,1:T_{\max}}^{(n)} \sim \pi_\theta(\cdot \mid x_i), \qquad n = 1, \ldots, N.$$

For each sampled trajectory, we enumerate candidate stopping lengths $t \in [T_{\min}, T_{\max}]$. At each candidate length, we halt latent thinking after $h_{i,t}^{(n)}$ and decode final-answers explicitly conditioned on x_i and the resulting prefix:

$$a_{i,n,t} \sim \pi_\theta(\cdot \mid x_i, h_{i,1:t}^{(n)}).$$

Using a verifiable reward, the answer-valid stopping set is

$$\mathcal{V}_i^{(n)} = \{t \in [T_{\min}, T_{\max}] \mid R(a_{i,n,t}, a_i^\star) = 1\}.$$

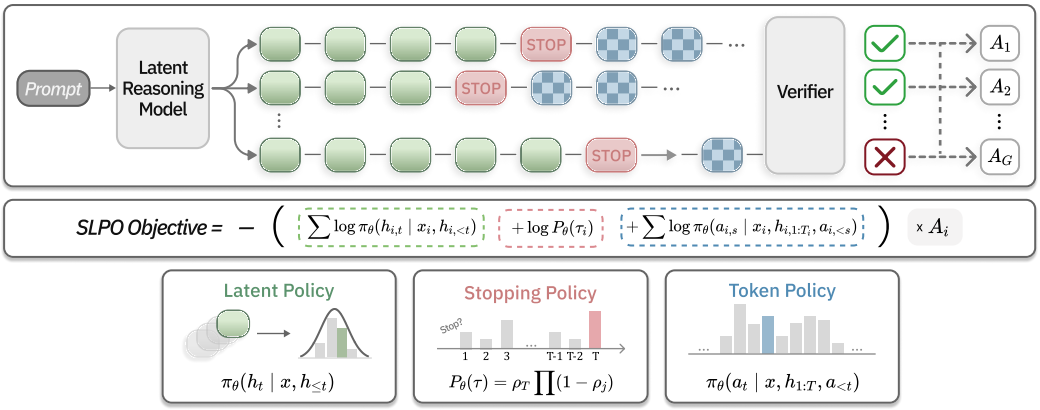


Figure 2: Overview of SLPO. Verifiable outcomes induce rollout advantages that weight the latent-surrogate, answer, and gate terms.

For a sampled trajectory, write $\rho_{i,n,t} = s_\theta(h_{i,t}^{(n)})$ as the stop probability at step t. The probability that stopping occurs at length t is then

$$P_\theta(\tau_i^{(n)} = t) = \rho_{i,n,t} \prod_{k<t} (1 - \rho_{i,n,k}).$$

The gate is trained to place probability mass on stopping times inside $\mathcal{V}_i^{(n)}$, which gives

$$\mathcal{L}_{\text{stop}}^{(i)} = -\frac{1}{N} \sum_{n=1}^{N} \log \sum_{t \in \mathcal{V}_i^{(n)}} P_\theta(\tau_i^{(n)} = t).$$

3.2 Surrogate Latent Policy Optimization

As illustrated in Fig. 2, SLPO couples latent-transition, answer-token, and stopping-time likelihoods in a single reward-weighted objective.

Surrogate transition likelihood. SLPO constructs a tractable Gaussian surrogate for stochastic latent transitions from repeated MC-dropout evaluations. For the prefix $(x_i, h_{i,<t})$, we run K stochastic forward evaluations with independent dropout masks during training:

$$z_{i,t}^{(k)} = f_\theta(x_i, h_{i,<t}; m_{i,t}^{(k)}), \qquad k = 1, \ldots, K.$$

These samples parameterize the Gaussian surrogate through an empirical mean and a floored isotropic variance:

$$\mu_{i,t} = \frac{1}{K} \sum_{k=1}^{K} z_{i,t}^{(k)}, \qquad \sigma_{i,t}^2 = \frac{1}{Kd} \sum_{k=1}^{K} \left\| z_{i,t}^{(k)} - \mu_{i,t} \right\|_2^2 + \epsilon.$$

The realized state $h_{i,t}$ is then scored under the isotropic Gaussian surrogate:

$$\log \widetilde{\pi}_\theta(h_{i,t} \mid x_i, h_{i,<t}) = -\frac{d}{2} \log(2\pi\sigma_{i,t}^2) - \frac{\|h_{i,t} - \mu_{i,t}\|_2^2}{2\sigma_{i,t}^2}.$$

MC-dropout generates the sampled trajectories, and the isotropic Gaussian supplies the per-step surrogate score used for advantage-weighted credit assignment (App. E). The sampled rollout states are treated as stop-gradient targets; gradients flow through the recomputed moments $(\mu_{i,t}, \sigma_{i,t}^2)$. Under fixed variance, a positive advantage raises $\widetilde{\pi}_\theta(h_{i,t})$ by moving $\mu_{i,t}$ toward $h_{i,t}$, while a negative advantage pushes $\mu_{i,t}$ away (Prop. 1).

Rollout policy objective. For each input x_i, we sample one or more latent rollouts, decode final answers a_i, and compute a verifiable reward

$$R_i = R(a_i, a_i^\star).$$

SLPO targets the expected rollout reward

$$J(\theta) = \mathbb{E}_{(x_i, a_i^\star)\sim\mathcal{D}} \mathbb{E}_{\xi_i \sim q_\theta(\cdot|x_i)} \big[R(a_i, a_i^\star) \big].$$

Here q_θ denotes the behavior trajectory law induced by MC-dropout, answer decoding, and the stopping gate. In practice, MC-dropout supplies behavior trajectories, while the Gaussian surrogate converts their sampled latent states into a differentiable score. The resulting update is the gradient of an explicitly defined empirical surrogate objective (Prop. 2); App. E characterizes when this update follows the expected-reward gradient and gives a sufficient condition for local improvement of $J(\theta)$. After the stopping-gate cold start, each rollout has an adaptive stopping time $\tau_i \le T_{\max}$. Write $\xi_i = (h_{i,1:\tau_i}, a_i)$. The surrogate rollout score combines the latent-transition surrogate, the answer-token likelihood, and the stopping-time likelihood term from Sec. 3.1:

$$\widetilde{\ell}_\theta(\xi_i \mid x_i) = \sum_{t=1}^{\tau_i} \log \widetilde{\pi}_\theta(h_{i,t} \mid x_i, h_{i,<t}) + \sum_{s=1}^{|a_i|} \log \pi_\theta(a_{i,s} \mid x_i, h_{i,1:\tau_i}, a_{i,<s}) + \log P_\theta(\tau_i).$$

The gate term assigns credit to stopping decisions: higher-reward stopping times reinforce continue/stop probabilities through $\rho_{i,t}$. With advantage $A_i = R_i - b_i$, where b_i may be a group-relative or leave-one-out baseline, the surrogate loss is

$$\mathcal{L}_{\mathrm{RL}} = -A_i \widetilde{\ell}_\theta(\xi_i \mid x_i).$$

Differentiating this loss yields the detached empirical update characterized in App. E (Prop. 2); RLOO or GRPO specify the stop-gradient advantage A_i. The composite score thereby extends outcome-level credit into the latent recurrence while jointly optimizing answer generation and trajectory length.

3.3 Inference

At inference, latent reasoning proceeds sequentially. After each latent state, the stopping head emits a stop probability, and the realized stopping time is the first step where this probability exceeds a threshold. The final answer is decoded only after this stopping event. Thus adaptive computation changes the number of latent steps used per instance without requiring inference-time enumeration of candidate prefixes. We leave the algorithm details to App. B.5.1.

4 Experiment Setup

Our experiments test whether SLPO (i) consistently improves outcome-level performance across latent reasoners, (ii) remains compatible with different policy-optimization algorithms, (iii) transfers across continuous and soft latent interfaces, and (iv) converts a fixed latent budget into difficulty-adaptive test-time computation.

Datasets. We evaluate SLPO on grade-school mathematical reasoning. We train on `GSM8K-Aug` (Hao et al., 2025), and evaluate on three held-out benchmarks: (1) `GSM8K-Test`, the official GSM8K test split; (2) `GSM-Hard` (You et al., 2026), a harder variant of GSM8K-Test with larger-magnitude numbers; and (3) `MultiArith` (You et al., 2026), a multi-step arithmetic set. The soft-latent generalization experiment additionally evaluates `MATH500`, `AIME2025`, and `AMC23`, following prior works (Hao et al., 2025; Shen et al., 2025; Tan et al., 2025; You et al., 2026).

Models and baselines. We apply SLPO to publicly released COCONUT (Hao et al., 2025) and CODI (Shen et al., 2025) checkpoints, using GPT-2 (124M) and Llama-3.2-1B-Instruct backbones. We compare our variants against commonly used latent-reasoning baselines: **COCONUT** (Hao et al., 2025) and **CODI** (Shen et al., 2025), **CoLaR** (Tan et al., 2025), **ReGuLaR** (Wang et al., 2026a), **DART** (Jiang et al., 2025), and **Latent-SFT** (Deng et al., 2025), and the explicit-reasoning methods **CoT-SFT** and **iCoT** (Hao et al., 2025). Detailed baseline configurations are provided in App. B.2. For soft-latent generalization experiments (Sec. 6.2), we apply SLPO to Llama-3.2-1B-Instruct and Llama-3.2-3B-Instruct, and compare against CoT with soft-token inference (Zhang et al., 2025b) and LEPO (Zhou et al., 2026); the corresponding setup appears in App. C.2.

Table 1: Comparison between latent-reasoning baselines and our variants. **Acc** is deterministic accuracy with dropout disabled; **Pass@**k for $k \in \{8, 16\}$ is the mean over three independent MC-dropout evaluation seeds ($p = 0.1$).

Method	GSM8K			GSM-Hard			MultiArith			Average		
	Acc	P@8	P@16	Acc	P@8	P@16	Acc	P@8	P@16	Acc	P@8	P@16
GPT2-124M												
COCONUT	34.12	45.79	49.13	7.66	10.70	12.06	80.86	88.79	91.38	40.88	48.43	50.86
COCONUT+SLPO	**35.63**	**49.13**	**51.55**	**7.66**	**10.85**	**12.52**	**83.10**	**91.38**	**92.59**	**42.13**	**50.45**	**52.22**
CODI	42.30	52.08	54.36	9.26	11.84	12.67	90.21	96.90	97.41	47.59	53.61	54.81
CODI+SLPO	**42.76**	**54.13**	**56.71**	**9.71**	**12.06**	**13.05**	**90.52**	**97.24**	**97.76**	**47.66**	**54.48**	**55.84**
Llama3.2-1B												
COCONUT	21.23	24.64	38.13	4.55	6.98	9.94	41.1	45.00	63.10	22.29	25.54	37.06
COCONUT+SLPO	**22.37**	**30.48**	**41.85**	**5.77**	**8.19**	**10.93**	**46.90**	**57.07**	**71.38**	**25.01**	**31.91**	**41.38**
CODI	55.22	63.91	67.48	12.82	15.02	15.63	95.52	98.28	98.79	54.59	59.07	60.63
CODI+SLPO	**55.27**	**65.13**	**70.28**	**13.20**	**15.86**	**16.77**	**96.38**	**98.79**	**99.48**	**54.95**	**59.93**	**62.18**

Training and Evaluation Protocols. Unless otherwise specified, `+SLPO` first trains the stopping gate and then optimizes SLPO with RLOO on `GSM8K-Aug` using $(K, G) = (4, 8)$. We report **Acc** from a single deterministic rollout (dropout disabled), with **Pass@**k for $k \in \{8, 16\}$ draws k independent MC-dropout rollouts ($p = 0.1$) per problem, we repeat this evaluation with three independent dropout seeds and report the mean. (see further details in App. B.5). For Tab. 2, we additionally report average reasoning length (#L): token-chain length for explicit CoT, the fixed latent budget for ungated latent baselines, and the realized stopping time under $T_{\max} = 12$ for `+SLPO`.

5 Main Results

5.1 Consistent Scaling Across Latent Reasoners

Tab. 1 shows a consistent latent test-time scaling effect: SLPO raises both Pass@8 and Pass@16 on all 12 backbone–dataset combinations. The gains span curriculum-trained COCONUT and distillation-based CODI, GPT-2 and Llama-3.2-1B, and all three held-out benchmarks without altering the underlying vector recurrence. Deterministic accuracy also improves on average for every backbone–reasoner pair. The improvement reaches 12.07 percentage points, or 26.8% relative, for Pass@8 on MultiArith with Llama-3.2-1B COCONUT. This substantial expansion in successful parallel rollouts is the clearest signature of the latent policy elicited by outcome-reward optimization.

5.2 Accuracy with Adaptive Latent Computation

Tab. 2 widens the comparison to explicit CoT-SFT, iCoT, and latent baselines. On Llama-3.2-1B, CODI+SLPO achieves the strongest average accuracy among latent methods with 5.79 latent steps on average; compared with Latent-SFT, it raises average accuracy by 1.05 percentage points while using 50.3% fewer latent steps. On GSM8K, CODI+SLPO surpasses CoT-SFT by 1.17 percentage points while using 75.2% fewer reasoning steps. The same optimization improves COCONUT, showing that SLPO strengthens both curriculum-trained and distillation-trained latent initializations. On GPT-2, CODI+SLPO likewise achieves the strongest average accuracy among the latent methods. These results position SLPO as an outcome-optimization layer that improves latent reasoning while allowing the learned gate to allocate, rather than uniformly spend, the available compute budget.

6 Analysis

We analyze SLPO from four perspectives: how well it generalizes across policy-optimization algorithms, how well it extends to soft latent inference, the geometry of latent rollouts it produces, and how the stopping gate adapts the allocation of latent computation to problem difficulty. See further analyses of rollout hyperparameters and computational cost in App. D.1 and App. D.2.

Table 2: Broader model results on grade-school mathematical reasoning benchmarks. **Acc** is deterministic accuracy. #L is average reasoning length: fixed latent budget for ungated COCONUT/CODI ($T_{\max} = 6$), realized stopping time for `+SLPO` under $T_{\max} = 12$, and the corresponding length metric reported by each baseline. DART and Latent-SFT Acc/#L are cited from Jiang et al. (2025); Deng et al. (2025).

Type	Method	GSM8K		GSM-Hard		MultiArith		Average	
		Acc	#L	Acc	#L	Acc	#L	Acc	#L
`GPT2-124M`									
Explicit	iCoT	30.12	0.00	5.73	0.00	55.52	0.00	30.41	0.00
	CoT-SFT	44.13	25.10	9.83	28.92	90.72	10.92	48.20	21.60
Latent	COCONUT	34.14	6.00	7.66	6.00	80.86	6.00	40.87	6.00
	CODI	42.30	6.00	9.26	6.00	90.21	6.00	47.59	6.00
	COCONUT+SLPO	35.63	5.73	7.66	6.08	83.10	6.28	42.13	6.03
	CODI+SLPO	**42.76**	11.83	**9.71**	11.94	**90.52**	11.44	**47.66**	11.74
`Llama3.2-1B`									
Explicit	iCoT	19.84	0.00	3.30	0.00	38.18	0.00	20.62	0.00
	CoT-SFT	54.10	25.42	15.66	34.24	99.32	13.71	56.31	24.46
Latent	COCONUT	21.23	6.00	4.55	6.00	41.12	6.00	22.29	6.00
	CODI	55.22	6.00	12.82	6.00	95.52	6.00	54.59	6.00
	CoLaR	26.57	5.56	5.87	7.01	86.39	3.23	39.60	5.27
	ReGuLaR	34.86	3.69	8.27	4.12	89.20	2.28	44.13	3.36
	DART	42.60	20.00	10.90	20.00	84.80	20.00	46.10	20.00
	Latent-SFT	52.41	12.82	12.60	15.04	96.82	7.19	53.88	11.66
	COCONUT+SLPO	22.37	7.52	5.77	7.88	46.90	7.49	25.01	7.63
	CODI+SLPO	**55.27**	6.30	**13.20**	6.41	**96.38**	4.65	**54.95**	5.79

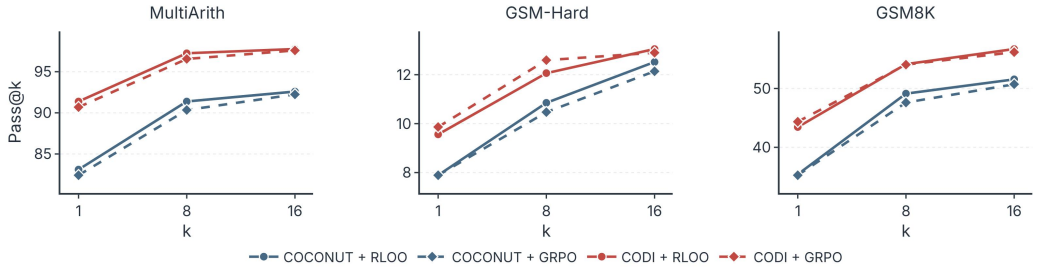


Figure 3: Pass@k for $k \in \{1, 8, 16\}$ under SLPO with RLOO versus GRPO on COCONUT and CODI. Blue curves denote COCONUT; red curves denote CODI. Solid curves use RLOO and dashed curves use GRPO. Full values appear in App. C.1.

6.1 Generalization Across Policy-Optimization Algorithms

SLPO defines a surrogate likelihood on hidden transitions that converts verifiable final-answer rewards into vector-space credit, as formalized in Sec. 3.2. Here we analyze whether SLPO adapts across outcome-reward policy-optimization algorithms by comparing RLOO and GRPO under the same surrogate, rollout budget, and latent backbones.

Fig. 3 shows closely aligned scaling curves for RLOO and GRPO across COCONUT and CODI on MultiArith, GSM-Hard, and GSM8K. Both algorithms preserve SLPO's gains over the released backbone initialization across latent reasoners and sampling budgets. The surrogate transition likelihood therefore provides a reusable outcome-reward interface across distinct advantage estimators.

SLPO targets outcome-reward optimization over complete latent trajectories in any autoregressive latent reasoner whose intermediate steps are vector states propagated by the backbone rather than vocabulary-level actions, as formalized in Sec. 3.2. Soft latent inference constitutes another vector-based latent paradigm (Deng et al., 2026; Zhou et al., 2026), recurring on a vocabulary-weighted token embedding rather than the hidden state. We thus evaluate SLPO under this setting to test whether the objective generalizes across vector-based latent inference mechanisms. Tab. 3 compares

Table 3: Soft latent inference results on `Llama3.2-1B` and `Llama3.2-3B`.

Method	Llama3.2-1B						Llama3.2-3B			
	GSM8K			MATH500			AIME 2025		AMC23	
	P@1	P@32	Len.	P@1	P@32	Len.	P@1	P@32	P@1	P@32
CoT	37.30	80.50	200.18	21.10	70.40	599.02	0.52	13.33	21.48	75.00
CoT (w/ latent inference)	35.78	79.23	188.27	20.80	69.50	605.27	0.83	16.67	21.64	77.50
GRPO	45.56	**87.57**	238.50	25.80	69.40	602.98	0.42	13.33	24.14	75.00
GRPO (w/ latent inference)	46.25	80.74	245.49	26.60	69.00	581.73	0.42	16.67	23.98	72.50
Soft Tokens	44.96	81.24	175.45	22.40	68.74	575.02	0.42	10.00	20.55	72.50
HRPO	33.32	70.33	158.27	18.20	64.22	551.27	0.42	6.67	17.50	67.50
LEPO	40.56	77.56	197.43	24.80	70.20	572.27	0.96	16.67	27.03	77.50
SLPO (ours)	**46.70**	82.03	**256.65**	**27.20**	**71.60**	**642.32**	**3.33**	**20.00**	**32.50**	**77.50**

SLPO with vocabulary-routed baselines on `Llama3.2-1B` and `Llama3.2-3B`; Fig. 4 tracks mean rollout length during soft-token training on the 1B setup. Implementation details appear in App. C.2.

6.2 Generalization to Soft Latent Inference

Tab. 3 establishes that SLPO transfers beyond hidden-state recurrence. At 1B, SLPO achieves the best deterministic accuracy on both GSM8K and MATH500 and the best MATH500 Pass@32. At 3B, it attains the strongest results on the more selective AIME 2025 and AMC23 evaluations, raising AIME 2025 Pass@1 from 0.96 to 3.33 (3.47×) and AMC23 Pass@1 from 27.03 to 32.50 (+5.47 percentage points; 20.2% relative). Fig. 4 exposes the corresponding sequential scaling behavior: SLPO steadily expands rollout length under reward optimization, whereas GRPO and LEPO remain near initialization or contract. SLPO thus elicits both stronger outcomes and increased test-time computation through a distinct soft-token latent interface.

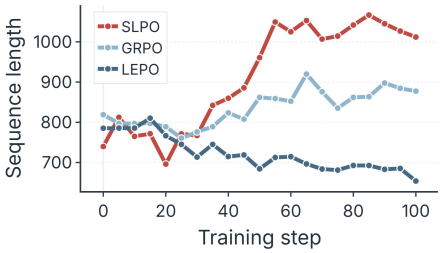


Figure 4: Mean generated sequence length during soft-token outcome-reward optimization on `Llama3.2-1B`.

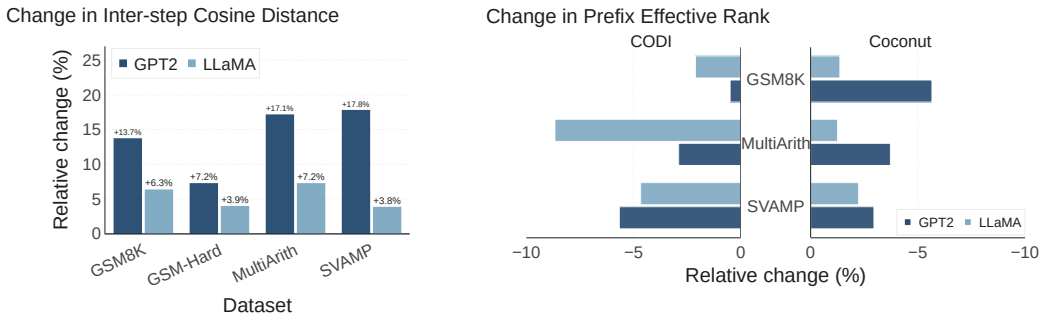
6.3 Latent Geometry Analysis

Outcome-reward RL reinforces successful latent rollouts, but the underlying latent geometry change remains unclear. We thus track two complementary signals to characterize how SLPO reshapes latent thinking geometry: inter-step separation, the mean cosine distance between consecutive latent states $h_{i,t}$ and $h_{i,t+1}$ along each rollout, and relative change in prefix effective rank. Fig. 5a reports the former and Fig. 5b the latter; see metric definitions, sampling details, and prefix-length rank curves in App. C.3.

For this geometry probe we additionally include `SVAMP` (Hao et al., 2025), a grade-school set with lexical perturbations. Fig. 5a reports the relative change in mean inter-step cosine distance after SLPO across backbones and datasets. Inter-step distance increases for every backbone–dataset pair, with the largest change on GSM-Hard. Successive latent states therefore become more differentiated after SLPO, producing a clearer stage-wise progression through the latent thinking process.

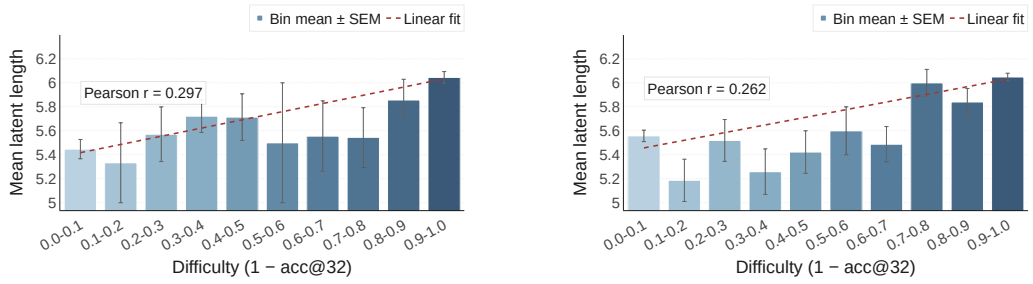
Inter-step separation captures pairwise progression along each rollout. We next track prefix effective rank, which measures how many independent directions the latent prefix spans, with definition given in App. C.3.4. Fig. 5b reports the relative change in prefix effective rank after SLPO across backbones, datasets, and latent-training methods. Effective rank decreases in every backbone–dataset–method combination. The latent prefix therefore concentrates into a lower-dimensional subspace after SLPO.

Together, the rise in inter-step separation and this rank compression show that SLPO yields stronger differentiation between successive thinking steps while steering rollouts through a sharpened set of task-relevant directions.



(a) Inter-step cosine distance (b) Relative rank change

Figure 5: Latent geometry pre- and post-SLPO. Fig. a: relative change in mean cosine distance between consecutive latent states across backbones and datasets. Fig. b: relative change in prefix effective rank across backbones, datasets; metric definitions and prefix-length rank curves appear in App. C.3.



(a) GSM8K validation (b) GSM8K test

Figure 6: Relationship between problem difficulty and latent length under the learned stopping gate. Bars show bin means with standard errors, and the dashed line shows the linear fit. Per-problem Pearson correlations are $r = 0.30$ (validation) and $r = 0.26$ (test), both with two-sided $p < 0.001$.

6.4 Difficulty-Adaptive Compute Allocation Analysis

The stopping-gate cold start trains the stop gate from the correctness of answers decoded under varying latent lengths (Sec. 3.1). We test whether the resulting gate allocates latent computation by difficulty by measuring difficulty as the empirical failure rate $d_i = 1 - \text{acc@32}_i$ and comparing gate-selected latent length across difficulty bins (Fig. 6).

As shown in Fig. 6, harder problems receive longer latent trajectories on both validation and test. Easy bins concentrate on shorter prefixes, while the hardest bins draw the longest hidden computation (Pearson $r = 0.30$ on validation and $r = 0.26$ on test; two-sided t-tests, both $p < 0.001$). The learned gate therefore provides adaptive test-time scaling in latent space: harder instances draw more computation, while easy instances terminate early rather than following a single global budget.

7 Conclusion and Future Work

We introduce SLPO for outcome-reward optimization in autoregressive latent reasoners. SLPO defines a differentiable surrogate policy interface over MC-dropout latent transitions, extending trajectory-level reward credit directly into vector-space reasoning. A correctness-supervised stopping-gate cold start establishes a prior over stopping times, allowing SLPO to optimize adaptive computation alongside latent reasoning. Across continuous and soft thinking inference, SLPO realizes latent test-time scaling through higher Pass@k under parallel sampling and longer latent trajectories on harder instances with improved deterministic accuracy. Future work will extend SLPO to open-ended reasoning and multimodal latent architectures.

Ethics Statement

This work studies reinforcement learning for latent reasoning on existing mathematical reasoning benchmarks using publicly available model checkpoints. The resulting models may inherit biases and generate incorrect outputs, as with the underlying language models. Applications beyond the evaluated reasoning tasks require additional assessment of reliability and potential misuse.

Reproducibility Statement

Section 4 describes the datasets, models, and evaluation protocols. Appendix B documents model initialization, baseline configurations, training hyperparameters, hardware, and inference procedures. Appendix C.2 provides the setup for the soft-latent experiments, and Appendix E states the assumptions and proofs underlying the surrogate optimization analysis. Code is provided as the supplementary material for reproducibility.

AI use statement

We used generative AI tools to refine manuscript wording and to help retrieve and summarize related literature. The authors reviewed the AI-assisted text and checked the literature summaries against the cited sources. We take full responsibility for the final content of this work.

References

Xinghao Chen, Anhao Zhao, Heming Xia, Xuan Lu, Hanlin Wang, Yanjun Chen, Wei Zhang, Jian Wang, Wenjie Li, and Xiaoyu Shen. Reasoning beyond language: A comprehensive survey on latent chain-of-thought reasoning, 2025. URL https://arxiv.org/abs/2505.16782.

Jingcheng Deng, Liang Pang, Zihao Wei, Shichen Xu, Zenghao Duan, Kun Xu, Yang Song, Huawei Shen, and Xueqi Cheng. LLM latent reasoning as chain of superposition, 2025. URL https://arxiv.org/abs/2510.15522.

Jingcheng Deng, Zihao Wei, Liang Pang, Junhong Wu, Shicheng Xu, Zenghao Duan, and Huawei Shen. Latent-GRPO: Group relative policy optimization for latent reasoning, 2026. URL https://arxiv.org/abs/2604.27998.

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.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason E Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=Itxz7S4Ip3.

Lianghuan Huang, Shuo Li, Sagnik Anupam, Insup Lee, and Osbert Bastani. Effective reinforcement learning for reasoning in language models. *arXiv preprint arXiv:2505.17218*, 2025.

Nan Jiang, Ziming Wu, De-Chuan Zhan, Fuming Lai, and Shaobing Lian. DART: Distilling autoregressive reasoning to silent thought. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5100–5108, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.256. URL https://aclanthology.org/2025.emnlp-main.256/.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20275–20321, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1025. URL https://aclanthology.org/2025.emnlp-main.1025/.

Phuc Minh Nguyen, Chinh D. La, Duy M. H. Nguyen, Nitesh V. Chawla, Binh T. Nguyen, and Khoa D. Doan. The reasoning boundary paradox: How reinforcement learning constrains language models. *arXiv preprint arXiv:2510.02230*, 2025.

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.

Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. CODI: Compressing chain-of-thought into continuous space via self-distillation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 677–693, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.36. URL https://aclanthology.org/2025.emnlp-main.36/.

Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4FWAwZtd2n.

Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Na Zou, Hanjie Chen, and Xia Hu. Stop overthinking: A survey on efficient reasoning for large language models, 2025. URL https://arxiv.org/abs/2503.16419.

Wenhui Tan, Jiaze Li, Jianzhong Ju, Zhenbo Luo, Ruihua Song, and Jian Luan. Think silently, think fast: Dynamic latent compression of LLM reasoning chains. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=AQsko3PPUe.

Fanmeng Wang, Haotian Liu, Guojiang Zhao, Hongteng Xu, and Zhifeng Gao. Regular: Variational latent reasoning guided by rendered chain-of-thought, 2026a. URL https://arxiv.org/abs/2601.23184.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Yifan Wang, Shiyu Li, Peiming Li, Xiaochen Yang, Yang Tang, and Zheng Wei. Render-of-thought: Rendering textual chain-of-thought as images for visual latent reasoning, 2026b. URL https://arxiv.org/abs/2601.14750.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Xilin Wei, Xiaoran Liu, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Jiaqi Wang, Xipeng Qiu, and Dahua Lin. Sim-cot: Supervised implicit chain-of-thought, 2025. URL https://arxiv.org/abs/2509.20317.

Xumeng Wen, Zihan Liu, Shun Zheng, Shengyu Ye, Zhirong Wu, Yang Wang, Zhijian Xu, Xiao Liang, Junjie Li, Ziming Miao, Jiang Bian, and Mao Yang. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms. *arXiv preprint arXiv:2506.14245*, 2025.

Runyang You, Yongqi Li, Meng Liu, Wenjie Wang, Liqiang Nie, and Wenjie Li. Parallel test-time scaling for latent reasoning models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, San Diego, USA, November 2026. Association for Computational Linguistics.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*, 2025.

Kaiyan Zhang, Yuxin Zuo, Bingxiang He, Youbang Sun, Runze Liu, Che Jiang, et al. A survey of reinforcement learning for large reasoning models. *arXiv preprint arXiv:2509.08827*, 2025a.

Zhen Zhang, Xuehai He, Weixiang Yan, Ao Shen, Chenyang Zhao, Shuohang Wang, Yelong Shen, and Xin Eric Wang. Soft thinking: Unlocking the reasoning potential of LLMs in continuous concept space. In *Advances in Neural Information Processing Systems*, 2025b. URL https://arxiv.org/abs/2505.15778.

Yuyan Zhou, Jiarui Yu, Hande Dong, Zhezheng Hao, Hong Wang, Jianqing Zhang, and Qiang Lin. LEPO: Latent reasoning policy optimization for large language models, 2026. URL https://arxiv.org/abs/2604.17892.

A Related Work

Latent reasoning. Latent reasoning instead autoregresses in hidden space, compressing redundant explicit traces into continuous thoughts (Sui et al., 2025; Chen et al., 2025; Hao et al., 2025; Shen et al., 2025; Tan et al., 2025; Wei et al., 2025; Wang et al., 2026b;a; Jiang et al., 2025; Deng et al., 2025). These approaches compress reasoning through curriculum training, self-distillation, dynamic or variational latent traces, and vocabulary-space representations (Hao et al., 2025; Shen et al., 2025; Tan et al., 2025; Wang et al., 2026a; Jiang et al., 2025; Deng et al., 2025; Wei et al., 2025; Wang et al., 2026b). These methods optimize latent states under SFT-style supervision from explicit reasoning, typically with a fixed inference budget; parallel test-time scaling instead keeps weights frozen, stochastically rolls out latent trajectories, and aggregates them with a learned reward model (You et al., 2026).

Reinforcement learning. Outcome-reward RL has become central to scaling explicit CoT (Zhang et al., 2025a; Guo et al., 2025; Wen et al., 2025). Token-policy objectives such as GRPO and RLOO remain standard (Shao et al., 2024; Huang et al., 2025), but continuous hidden transitions expose no action log-probabilities (Yue et al., 2025; Nguyen et al., 2025). Recent methods address this gap with soft-token rollouts, vocabulary-space latent tokens, or learned latent heads (Zhou et al., 2026; Deng et al., 2026; Tan et al., 2025). These designs depend on token distributions, embedding mixtures, or architecture-specific latent heads, so they do not carry over to arbitrary vector-based latent reasoners with plain hidden-state recurrence. Outcome-reward optimization in this setting remains unexplored.

B Implementation Details

This appendix supplements Sec. 4 with model, training, and evaluation details omitted from the main text for space.

B.1 Base Models and Initialization

All SLPO experiments start from publicly available latent-reasoning checkpoints, using official baseline releases when available and open third-party reimplementations otherwise. COCONUT and CODI use GPT-2 (124M) or Llama-3.2-1B-Instruct, depending on the row in the main tables; CoLaR, ReGuLaR, DART, and Latent-SFT use Llama-3.2-1B-Instruct. We do not retrain the imitation stage of these latent reasoners: The stopping-gate cold start and SLPO are applied directly on top of these initialized weights.

B.2 Baseline Configurations

We use the default inference configuration of each latent baseline whenever possible. Ungated COCONUT and CODI use a fixed budget of six latent thoughts, matching the original papers; `+SLPO` uses the same backbones with maximum budget $T_{\max} = 12$ and reports the realized stopping time

as #L (App. B.3.1–B.5.1). CoLaR uses thinking speed $2\times$ with a maximum latent budget of 64 steps. ReGuLaR follows the released evaluation protocol in Wang et al. (2026a). DART Acc and #L are taken from Jiang et al. (2025), with #L equal to their fixed Silent Thought count $C{=}20$ (non-autoregressive; all ST tokens are processed in one forward pass). Latent-SFT Acc and #L are taken from the $r{=}2$ Llama-3.2-1B setting in Deng et al. (2025) (GSM8K/GSM-Hard/MultiArith; averages exclude SVAMP). For fair comparison with explicit CoT baselines, CoT-SFT and iCoT numbers follow the training protocol reported by Hao et al. (2025); Shen et al. (2025); Tan et al. (2025).

For CoT-SFT and iCoT, #L is measured on the generated reasoning chain before answer prediction. For latent methods without the stopping gate, #L is the fixed latent budget used at inference.

B.3 Training Procedure

B.3.1 Stopping-gate Cold Start

The stopping-gate cold start is applied first on the same GSM8K-Aug split. For each training problem, the released backbone samples $N = 4$ stochastic latent trajectories with rollout dropout ($p = 0.1$) up to $T_{\max} = 12$, the same maximum budget used at SLPO training and gated inference (raised from the ungated COCONUT/CODI horizon of six). Candidate stopping lengths lie in $[T_{\min}, T_{\max}]$ with $T_{\min} = 3$. For every trajectory and candidate prefix length, answer decoding runs with dropout disabled to stabilize the online supervision signal. The stopping head is trained with the first-stop objective in Sec. 3.1. We do not add a compute-aware auxiliary term.

Optimization. This stage uses Adafactor at learning rate 10^{-4} for 15 epochs. For GPT-2 (COCONUT and CODI), we use a per-device batch size of 32 with 2 gradient-accumulation steps; for Llama-3.2-1B (COCONUT and CODI), we use a per-device batch size of 8 with 8 accumulation steps. The stop gate itself is regularized with dropout $p = 0.1$ during this cold start. We evaluate every 500 optimizer steps, restoring the checkpoint with the highest validation accuracy under early stopping with patience 20.

B.3.2 SLPO Training

Reward and advantage. Each training instance $(x_i, a_i^\star)$ receives a binary outcome reward $R_i = \mathbb{I}[\hat{a}_i = a_i^\star]$ after answer parsing. Advantages are estimated with RLOO over $G = 8$ sampled latent trajectories per problem. The surrogate transition likelihood at each step uses $K = 4$ independent MC-dropout forwards.

Stochastic rollouts. Latent trajectories are sampled with MC-dropout at rate $p = 0.1$ during rollout generation. The realized latent states are treated as stop-gradient targets; gradients flow through the recomputed moments $(\mu_{i,t}, \sigma^2_{i,t})$ in Sec. 3.2, not through the rollout sampling path itself.

Optimization. We optimize with Adafactor at learning rate 10^{-6}. Training runs for up to 40,000 optimizer steps under a constant schedule with warmup over the first 2,000 steps. We do not apply PPO-style clipping or a KL penalty: the SLPO RLOO objective uses $\beta = 0$ and omits reference-policy log-probability terms. Answer-token likelihoods and the first-stop gate term are included in the rollout objective as in Sec. 3.2.

B.4 Runtime Environment

Hardware. SLPO is launched with Accelerate over 4 processes on 4 GPUs (4×RTX 5880 Ada). Each process uses a per-device batch size of 16; with 4 GPUs, the effective prompt batch is 64.

B.5 Inference and Evaluation

B.5.1 Stopping-gate Inference

Alg. 1 gives the inference procedure after the stopping-gate cold start and SLPO. Latent reasoning and answer decoding both run with dropout disabled, and the stopping gate uses deterministic thresholding. Unless otherwise noted, $T_{\max} = 12$ matches the cold-start budget in App. B.3.1.

Algorithm 1 Adaptive latent inference with stopping gate

Require: Input x; latent reasoner f_θ with stopping head g_θ; maximum budget $T_{\max}$; stop threshold τ
Disable dropout for latent updates and answer decoding
Initialize latent prefix $\tilde{h}_{<1} \leftarrow \emptyset$
for $t = 1$ to $T_{\max}$ **do**
  $\tilde{h}_t \leftarrow f_\theta(x, \tilde{h}_{<t})$
  $\rho_t \leftarrow \sigma(g_\theta(\tilde{h}_t))$
  if $\rho_t \geq \tau$ **then**
    break
  end if
end for
Let $T^\star \leftarrow t$ be the realized stopping step
Decode final answer $\hat{a} \sim \pi_\theta(\cdot \mid x, \tilde{h}_{1:T^\star})$
return $\hat{a}$

We sweep candidate thresholds $\{0.5, 0.6, 0.7, 0.8, 0.9\}$ on the validation split and select the threshold with the highest validation accuracy. When a single default threshold is required, we use 0.5.

B.5.2 Accuracy and Pass@k Estimation

Acc uses a single deterministic forward pass per problem: latent reasoning and answer decoding both run with dropout disabled, following the stopping-gate inference path in Sec. 3.3. For $k > 1$, one Pass@k evaluation draws k independent stochastic latent rollouts per problem with MC-dropout ($p = 0.1$); a problem counts as solved if at least one trajectory yields the correct parsed answer. Because dropout seeds affect the sampled trajectories, we repeat this evaluation R=3 times with independent seeds and report the mean Pass@k over the three runs in Tab. 1.

C Supplementary Experimental Material

C.1 Policy-Optimization Algorithm Comparison

This appendix reports the full Pass@k values underlying Fig. 3. We keep the SLPO surrogate, rollout budget, and latent backbones fixed, and swap only the outcome-reward policy-optimization algorithm between RLOO and GRPO.

Table 4: Pass@k under SLPO with RLOO versus GRPO on COCONUT and CODI.

		COCONUT			CODI		
Dataset	Algorithm	Pass@1	Pass@8	Pass@16	Pass@1	Pass@8	Pass@16
MultiArith	RLOO	83.10	91.38	92.59	90.52	97.24	97.76
	GRPO	82.41	90.34	92.24	90.69	96.55	97.59
GSM-Hard	RLOO	7.66	10.85	12.52	9.71	12.06	13.05
	GRPO	7.89	10.47	12.14	9.86	12.59	12.90
GSM8K	RLOO	35.63	49.13	51.55	42.76	54.13	56.71
	GRPO	35.25	47.61	50.72	44.35	54.06	56.18

C.2 Soft Latent Inference Transfer

This appendix supplements Sec. 6.2 with the setup behind Tab. 3. We keep the soft-token inference mechanism unchanged (Zhang et al., 2025b), apply SLPO on Llama3.2–1B and Llama3.2–3B, and

compare against CoT, vocabulary-routed GRPO, Soft Tokens, HRPO, and LEPO (Zhou et al., 2026) under the same stack. This transfer setting uses SLPO without the stopping-gate cold start.

Soft-token interface. Each latent step is a probability-weighted embedding rather than a backbone hidden state. The model maps logits ℓ_t to $p_t = \mathrm{softmax}(\ell_t/\tau)$ and forms

$$z_t = p_t^\top E.$$

Answer generation after the latent block still uses the ordinary token likelihood.

SLPO objective. We apply the surrogate in Sec. 3.2 by identifying $\tilde{h}_{i,t} \equiv z_{i,t}$ and scoring transitions in embedding space. The isotropic Gaussian moments and likelihood follow Sec. 3.2, with variance floor $\epsilon = 10^{-6}$. On the optimized latent prefix, the surrogate term $-\widehat{A}_i \log \widetilde{\pi}_\theta(z_{i,t} \mid x_i, z_{i,<t})$ replaces the soft-token cross-entropy; subsequent answer tokens retain the standard token log-probability objective.

Training and evaluation. Rollouts use a latent block of length 32. We score only an early prefix with the surrogate—8 steps by default, with a 16-step variant also evaluated—using $K = 2$ MC-dropout samples per step. Stochastic forwards follow App. B.3.2 ($p = 0.1$). Outcome rewards and group-relative advantages follow the GRPO-style soft-token baseline. We evaluate GSM8K and MATH500 at 1B scale, and AIME 2025 and AMC23 at 3B scale (Tab. 3); the 3B runs keep the same interface and surrogate settings, changing only the backbone scale. Other optimization details follow the soft-token baseline recipe shared with LEPO and vocabulary-routed GRPO.

C.3 Latent Geometry Analysis

This appendix supplements Sec. 6.3 with the metric definitions, aggregation rules, and figure construction used in the latent-geometry analysis.

C.3.1 Sampling Protocol

We compare pre- and post-SLPO checkpoints on the same evaluation instances. For each problem x_i, we draw N stochastic latent trajectories with MC-dropout at rate $p = 0.1$, using the same rollout mechanism as SLPO training (Sec. 3.2). Each trajectory is

$$h_{i,1:T}^{(n)} = \big(h_{i,1}^{(n)}, \ldots, h_{i,T}^{(n)}\big), \qquad n = 1, \ldots, N,$$

with fixed latent budget $T = 6$. Unless stated otherwise, we evaluate the first 32 problems per dataset, use $N = 16$ dropout samples per problem, and report metrics on COCONUT and CODI with GPT-2 or Llama backbones.

C.3.2 Step Representations

Dropout sampling induces local stochasticity around each latent step. To obtain a representative latent vector for step t, we aggregate sampled states across trajectories:

$$c_{i,t} = \frac{1}{N} \sum_{n=1}^{N} h_{i,t}^{(n)}. \tag{1}$$

The sequence $(c_{i,1}, \ldots, c_{i,T})$ is the step-wise representation used by both metrics below.

C.3.3 Inter-Step Separation

Inter-step separation measures how far successive step representations move apart along this trajectory. We use cosine distance

$$\mathrm{dcos}(u, v) = 1 - \frac{u^\top v}{\|u\|_2 \|v\|_2}, \tag{2}$$

and define the per-problem inter-step score as the mean over consecutive step representations:

$$D_i^{\mathrm{inter}} = \frac{1}{T-1} \sum_{t=1}^{T-1} \mathrm{dcos}(c_{i,t}, c_{i,t+1}). \tag{3}$$

Larger D_i^{inter} means successive latent steps occupy more separated positions in hidden space. For a dataset $\mathcal{S}$, we average over problems,

$$\bar{D}^{\text{inter}} = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} D_i^{\text{inter}},$$

and report the relative change from pre- to post-SLPO,

$$\Delta^{\text{inter}} = 100 \cdot \frac{\bar{D}^{\text{inter}}_{\text{post}} - \bar{D}^{\text{inter}}_{\text{pre}}}{\bar{D}^{\text{inter}}_{\text{pre}}}. \tag{4}$$

Fig. a plots Δ^{inter} for COCONUT across GSM8K, GSM-Hard, MultiArith, and SVAMP.

C.3.4 Prefix Effective Rank

To characterize how the latent prefix distributes across the hidden subspace as t grows, we track how many independent directions the latent prefix $\{c_1, \ldots, c_t\}$ spans as t grows. For prefix length $t \geq 2$, stack the first t step representations into a matrix

$$C_{i,t} = \begin{bmatrix} c_{i,1}^\top \\ \vdots \\ c_{i,t}^\top \end{bmatrix} \in \mathbb{R}^{t \times d},$$

row-center across latent steps,

$$\tilde{C}_{i,t} = C_{i,t} - \mathbf{1}\bar{c}_{i,t}^\top, \qquad \bar{c}_{i,t} = \frac{1}{t} \sum_{j=1}^{t} c_{i,j},$$

and compute the effective rank

$$r_{\text{eff}}(i,t) = \frac{\|\tilde{C}_{i,t}\|_F^2}{\|\tilde{C}_{i,t}\|_2^2} = \frac{\sum_{j=1}^{\min(t,d)} \sigma_j\big(\tilde{C}_{i,t}\big)^2}{\sigma_1\big(\tilde{C}_{i,t}\big)^2}, \tag{5}$$

where $\sigma_j(\cdot)$ denotes the j-th singular value. This is the stable rank of the centered step-representation matrix: it equals 1 when the step representations lie on one line, and grows as the prefix spans more independent directions. At each t, we average $r_{\text{eff}}(i,t)$ over problems and plot the resulting before/after curves. The x-axis therefore marks prefix length rather than a single latent state; because adding step representations also raises the maximum attainable rank, the upward slope along t partly reflects accumulation. The comparison of interest is pre- versus post-SLPO at the same t.

C.3.5 Figure Construction

Fig. 5 in the main text has two panels. Fig. a reports Δ^{inter} for COCONUT (equation 3–equation 4). Fig. b reports $r_{\text{eff}}(i,t)$ from equation 5 on two representative CODI settings: GPT-2 on SVAMP and Llama on MultiArith.

Figs. 7 and 8 extend the rank analysis to all dataset–backbone combinations for COCONUT and CODI. The main-text conclusion follows the same pre- versus post-SLPO comparison at fixed t: the post-SLPO curve lies lower while the progression shape across latent steps is preserved.

D Further Analysis

D.1 Hyperparameter Sensitivity

SLPO couples surrogate transition-likelihood estimation with reward optimization through rollout hyperparameters K and G. K controls the number of stochastic forward passes used to estimate each surrogate in Sec. 3.2, whereas G is the group size controlling how many complete trajectories are sampled per problem. We vary each factor over $\{2, 4, 8\}$ while holding the other fixed (Fig. 9); full values appear in App. D.1.

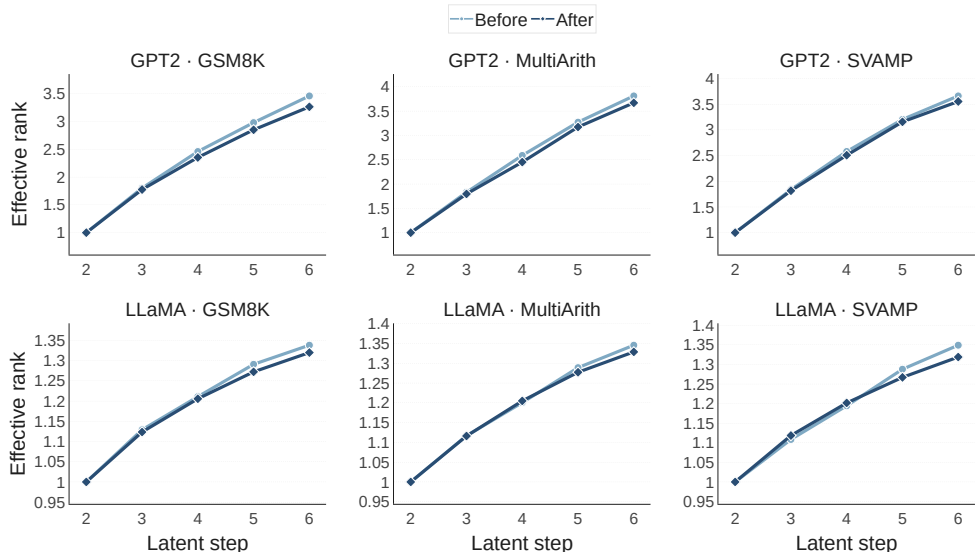


Figure 7: Full prefix effective rank curves pre- and post-SLPO for COCONUT. The plot shows equation 5 at prefix length t, averaged over the first 32 problems in each dataset. Curves are grouped by backbone and dataset.

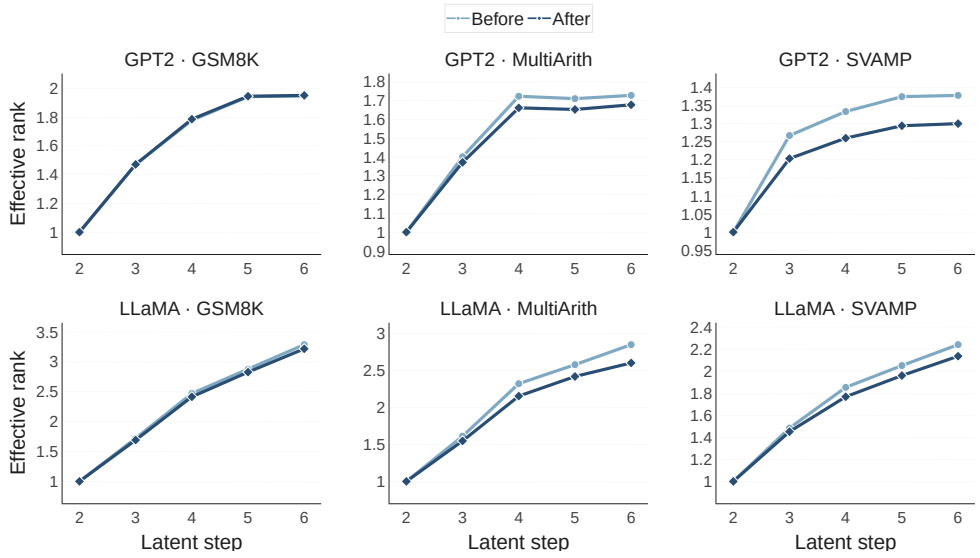
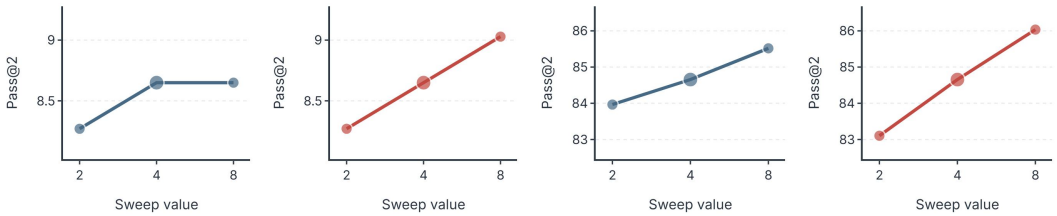


Figure 8: Full prefix effective rank curves pre- and post-SLPO for CODI. The plot shows equation 5 at prefix length t, averaged over the first 32 problems in each dataset. Curves are grouped by backbone and dataset; the main-text rank panels (Fig. 5) show two representative cases from this figure.

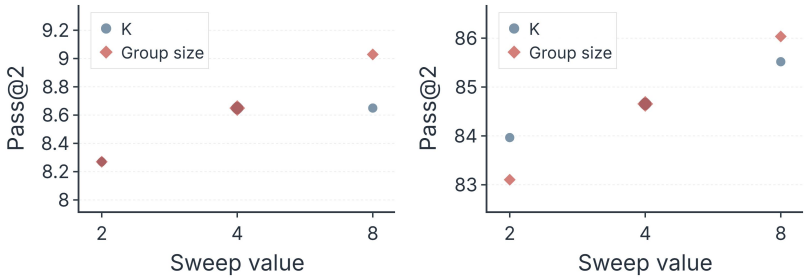
Fig. 9 identifies group size G as the primary rollout-scaling axis: increasing G consistently improves Pass@2 on both GSM-Hard and MultiArith. Performance is substantially less sensitive to K, indicating that a modest number of stochastic forward passes already provides an effective surrogate estimate. SLPO can therefore direct additional rollout compute toward exploring more complete trajectories, where it yields the clearest return.

This appendix reports the full Pass@2 values underlying Fig. 9. We vary one rollout hyperparameter over $\{2, 4, 8\}$ while holding the other fixed at $K = 4$ or $G = 4$.



(a) K sweep, GSM-Hard (b) G sweep, GSM-Hard (c) K sweep, MultiArith (d) G sweep, MultiArith

Figure 9: Controlled hyperparameter sweeps for SLPO. Each panel varies one factor over $\{2, 4, 8\}$ while holding the other fixed (K sweep at $G = 4$; G sweep at $K = 4$). Full values are in App. D.1.



(a) GSM-Hard (b) MultiArith

Figure 10: Full Pass@2 sweeps for K (with $G = 4$ fixed) and G (with $K = 4$ fixed). The larger marker at 4 marks the value used in that one-factor sweep.

D.2 Computational Profile

SLPO introduces training-time computation through G sampled trajectories and K stochastic forwards for estimating each surrogate transition likelihood. We profile this cost on the same RTX 5880 Ada GPU class using two unique prompts, with two warmup steps followed by eight measured optimizer steps. Tab. 5 reports mean wall-clock time per step and its rollout, surrogate-construction, and backward components. The profile isolates the SLPO training stage; the one-time stopping-gate cold start is described separately in App. B.3.1.

Table 5: Per-step computational profile on RTX 5880 Ada. All configurations use MC-dropout. Relative time is normalized by the paper configuration $(K, G) = (4, 8)$. **Sur./step** is the fraction of total step time spent constructing the surrogate likelihood.

K	G	Step (s)	Rel.	Rollout (s)	Surrogate (s)	Backward (s)	Sur./step
1	8	0.624	0.55×	0.321	0.108	0.177	17%
4	**8**	**1.130**	**1.00×**	0.294	0.274	0.545	24%
8	8	2.021	1.79×	0.319	0.585	1.101	29%
4	4	0.676	0.60×	0.243	0.146	0.272	22%

Under the paper configuration $(K, G) = (4, 8)$, an optimizer step takes 1.130 seconds; surrogate construction accounts for 0.274 seconds, or 24% of the total, while rollout generation and the backward pass account for 0.294 and 0.545 seconds. The profile also exposes a direct compute control: reducing the group size from $G = 8$ to $G = 4$ at $K = 4$ lowers step time to 0.676 seconds (0.60×), whereas increasing K from 4 to 8 at $G = 8$ raises it to 2.021 seconds (1.79×). Together with the hyperparameter analysis in Fig. 9, which finds larger returns from rollout diversity than from additional surrogate samples, these measurements support allocating training compute to G while keeping K moderate.

The surrogate estimator and the K stochastic forwards are used only during training. At inference, SLPO performs the original latent recurrence together with the learned stopping head; its compute is therefore governed by the realized adaptive trajectory length rather than by K or G.

E SURROGATE OPTIMIZATION GUARANTEES

We establish five properties of the surrogate in Sec. 3.2: its advantage-weighted update, the empirical objective implemented by SLPO, its relation to the expected-reward gradient, a sufficient condition for expected-reward ascent, and the regularity induced by its variance floor. Throughout, write h for a stop-gradient latent state sampled by MC-dropout, $(\mu_\theta, \sigma_\theta^2)$ for the isotropic moments recomputed from K dropout forwards, and

$$\log \widetilde{\pi}_\theta(h) = -\tfrac{d}{2}\log(2\pi\sigma_\theta^2) - \tfrac{\|h-\mu_\theta\|_2^2}{2\sigma_\theta^2}.$$

Answer-token and stopping terms are unchanged from Sec. 3.2 and are omitted from the latent-only statements below.

Proposition 1 (Mean-gradient component under fixed variance)**.** *Fix $\sigma_\theta^2 > 0$ (or treat σ_θ^2 as stop-gradient), and hold (h, A) fixed. Then*

$$-A \log \widetilde{\pi}_\theta(h) = \frac{A}{2\sigma_\theta^2}\|h - \mu_\theta\|_2^2 + C(A, \sigma_\theta^2),$$

so

$$-A\, \nabla_\theta \log \widetilde{\pi}_\theta(h) = \frac{A}{\sigma_\theta^2}\,(\mu_\theta - h)^\top \nabla_\theta \mu_\theta$$

whenever gradients are taken only through μ_θ. A positive advantage therefore pulls μ_θ toward h, and a negative advantage pushes it away. When σ_θ^2 is also trainable, this identity describes only the mean-gradient component of the latent update; the variance path is additional.

Proof sketch. The isotropic surrogate log-density expands as $\log \widetilde{\pi}_\theta(h) = -\|h-\mu_\theta\|_2^2/(2\sigma_\theta^2) + c(\sigma_\theta^2)$. Multiplying by $-A$ gives the stated loss identity. Differentiating in μ_θ and applying the chain rule yields the gradient form. □

Proposition 2 (Detached empirical surrogate loss)**.** *Draw a frozen rollout batch $\{\xi_i\}_{i=1}^G$ from MC-dropout under parameters $\bar{\theta}$, compute stop-gradient advantages $A_i = \mathrm{sg}(R_i - b_i)$, and define the empirical surrogate loss*

$$\widehat{\mathcal{L}}_{\mathrm{surr}}(\theta; \bar{\theta}) = -\frac{1}{G}\sum_{i=1}^{G} A_i \widetilde{\ell}_\theta(\xi_i \mid x).$$

Then the implemented gradient of the latent, answer, and gate terms in Sec. 3.2 coincides with $\nabla_\theta \widehat{\mathcal{L}}_{\mathrm{surr}}(\theta; \bar{\theta})$ evaluated at $\theta = \bar{\theta}$, with the sampled latent states in $\widetilde{\ell}_\theta$ treated as stop-gradient targets. Thus the implementation optimizes an explicit reward-weighted surrogate score on a detached batch, establishing the empirical optimization target of SLPO.

Proof sketch. After sampling, (ξ_i, A_i) no longer depend on the differentiation variable θ. Differentiating $\widehat{\mathcal{L}}_{\mathrm{surr}}$ therefore passes only through the scored factors $\widetilde{\ell}_\theta(\xi_i \mid x)$, which is precisely the implemented backward pass. □

Proposition 3 (Gradient approximation under score alignment)**.** *Let $q_\theta(\xi \mid x)$ denote a behavior trajectory law that admits a differentiable score, and suppose the expected-reward gradient admits the score representation*

$$g_J = \mathbb{E}_{\xi\sim q_\theta}\left[A(\xi)\nabla_\theta \log q_\theta(\xi \mid x)\right]$$

for a square-integrable advantage $A(\xi)$. Define the corresponding population surrogate direction

$$g_{\mathrm{surr}} = \mathbb{E}_{\xi\sim q_\theta}\left[A(\xi)\nabla_\theta \widetilde{\ell}_\theta(\xi \mid x)\right].$$

If

$$\mathbb{E}_{\xi\sim q_\theta}\left[\left\|\nabla_\theta \widetilde{\ell}_\theta(\xi \mid x) - \nabla_\theta \log q_\theta(\xi \mid x)\right\|_2^2\right] \le \varepsilon_{\mathrm{score}}^2,$$

then

$$\|g_{\mathrm{surr}} - g_J\|_2 \le \sqrt{\mathbb{E}_{q_\theta}[A(\xi)^2]}\,\varepsilon_{\mathrm{score}}.$$

Proof sketch. Write the gradient difference as the expectation of the advantage multiplied by the score mismatch. Cauchy–Schwarz gives the stated bound. □

Proposition 4 (Local expected-reward ascent)**.** *Suppose J has an L-Lipschitz gradient on a neighborhood of θ, and let g_{surr} be an ascent direction obtained from the population surrogate objective. If*

$$\|g_{\mathrm{surr}} - \nabla_\theta J(\theta)\|_2 \le \delta,$$

Then for any $\eta > 0$ such that the segment $\{\theta + s\eta g_{\mathrm{surr}} : s \in [0,1]\}$ lies in this neighborhood,

$$J(\theta + \eta g_{\mathrm{surr}}) \ge J(\theta) + \eta \left(\|g_{\mathrm{surr}}\|_2^2 - \delta \|g_{\mathrm{surr}}\|_2\right) - \frac{L\eta^2}{2}\|g_{\mathrm{surr}}\|_2^2.$$

Consequently, if $\|g_{\mathrm{surr}}\|_2 > \delta$ and

$$0 < \eta < \frac{2(\|g_{\mathrm{surr}}\|_2 - \delta)}{L\|g_{\mathrm{surr}}\|_2},$$

the surrogate step strictly increases the expected reward J.

Proof sketch. L-smoothness gives

$$J(\theta + \eta g_{\mathrm{surr}}) \ge J(\theta) + \eta\langle \nabla_\theta J(\theta), g_{\mathrm{surr}}\rangle - \frac{L\eta^2}{2}\|g_{\mathrm{surr}}\|_2^2.$$

Writing $\nabla_\theta J(\theta) = g_{\mathrm{surr}} + e$ with $\|e\|_2 \le \delta$ and applying Cauchy–Schwarz yields the result. □

Role of MC-dropout. MC-dropout supplies the exploratory behavior trajectories, while a finite-K moment-matched Gaussian score turns those trajectories into a differentiable optimization interface. Larger K reduces Monte Carlo error in the estimated moments, and Prop. 3 shows how score alignment controls the population-gradient approximation. Prop. 4 then converts this approximation into an explicit condition for expected-reward improvement.

Proposition 5 (Variance-floor regularity)**.** *For fixed h and a variance floor $\epsilon > 0$,*

$$\sup_{\mu,\, \sigma^2 \ge \epsilon} \log \widetilde{\pi}(h; \mu, \sigma^2) = -\tfrac{d}{2}\log(2\pi\epsilon),$$

attained at $\mu = h$ and $\sigma^2 = \epsilon$. For every fixed $\sigma^2 \ge \epsilon$, $\mu = h$ is the unique maximizing mean. The floor is essential for the finite optimum: over $\sigma^2 > 0$,

$$\sup_{\mu, \sigma^2 > 0} \log \widetilde{\pi}(h; \mu, \sigma^2) = +\infty,$$

because the supremum diverges as $\sigma^2 \to 0^+$ along $\mu = h$.

Proof sketch. Completing the square in the isotropic Gaussian density shows that, for fixed σ^2, the maximum over μ is attained uniquely at $\mu = h$ and equals $-\frac{d}{2}\log(2\pi\sigma^2)$. The variance floor maximizes this expression at $\sigma^2 = \epsilon$, giving the finite joint optimum above. Removing the floor and sending $\sigma^2 \to 0^+$ along $\mu = h$ yields the stated divergence. □

Implications. Prop. 1 identifies the mean-matching force in the latent update when variance is held fixed. Prop. 2 characterizes the implemented step as gradient descent on a detached, reward-weighted surrogate score. Prop. 3 links score alignment to population-gradient approximation, and Prop. 4 gives a sufficient local condition for expected-reward improvement. RLOO and GRPO instantiate the stop-gradient advantage A_i within this shared interface. Prop. 5 establishes the finite optimum induced by the variance floor ϵ.