SeLaR: Selective Latent Reasoning in Large Language Models

Renyu Fu **Guibo Luo**†
Guangdong Provincial Key Laboratory of Ultra High Definition Immersive Media Technology
Shenzhen Graduate School, Peking University
luogb@pku.edu.cn
github.com/Parker-rfu/SeLaReasoning

Abstract

Chain-of-Thought (CoT) has become a cornerstone of reasoning in large language models, yet its effectiveness is constrained by the limited expressiveness of discrete token sampling. Recent latent reasoning approaches attempt to alleviate this limitation by replacing discrete tokens with soft embeddings (probability-weighted mixtures of token embeddings) or hidden states, but they commonly suffer from two issues: (1) global activation injects perturbations into high-confidence steps, impairing reasoning stability; and (2) soft embeddings quickly collapse toward the highest-probability token, limiting exploration of alternative trajectories. To address these challenges, we propose ***SeLaR*** (**Se**lective **La**tent **R**easoning), a lightweight and training-free framework. SeLaR introduces an entropy-gated mechanism that activates soft embeddings only at low-confidence steps, while preserving discrete decoding at high-confidence steps. Additionally, we propose an entropy-aware contrastive regularization that pushes soft embeddings away from the highest-probability token's direction, encouraging sustained exploration of multiple latent reasoning paths. Experiments on five reasoning benchmarks demonstrate that SeLaR consistently outperforms standard CoT and state-of-the-art training-free methods.

1 Introduction

Chain-of-Thought (CoT) (Wei et al., 2023; Goyal et al., 2024; Pfau et al., 2024) has become a prevailing paradigm for enabling multi-step reasoning in large language models (Brown et al., 2020; Chowdhery et al., 2022; Du et al., 2022; Touvron et al., 2023; OpenAI et al., 2024b; Singh et al., 2025). By explicitly generating intermediate reasoning steps, CoT significantly improves performance on complex tasks such as mathematical and logical reasoning (DeepSeek-AI et al., 2025; OpenAI et al., 2024a, 2025; Abdin et al., 2025; Qwen et al., 2025; Team et al., 2025). However, CoT relies on hard token commitments at each step: the model must discretize its internal distribution into a single sampled token, potentially discarding valuable information about alternative reasoning paths. This commitment may hinder the effective propagation of uncertainty across reasoning steps, ultimately leading to suboptimal final predictions (Li et al., 2025).

Inspired by human reasoning, which often considers multiple plausible hypotheses simultaneously, recent work has explored latent reasoning paradigms that replace discrete token sampling with soft embeddings or hidden states as carriers of reasoning information (Hao et al., 2025; Cheng and Durme, 2024; Xu et al., 2025; Zhang et al., 2025b; Tan et al., 2025; Shi et al., 2025). These approaches enable richer representations and implicit branching over multiple candidate tokens during reasoning.

Existing latent reasoning methods can be categorized into fine-tuning–based and training-free approaches. Fine-tuning methods such as Coconut (Hao et al., 2025) propagate hidden states as reasoning signals, but often suffer from catastrophic forgetting (Lobo et al., 2025) due to the domain gap between hidden-state and input embedding spaces. Training-free methods such as Soft Thinking (Zhang et al., 2025b) employ soft embeddings to explore multiple reasoning trajectories, but activate them uniformly across all steps regardless of model confidence.

Our work is motivated by a key empirical observation: during CoT decoding, the entropy of the model's output distribution exhibits a clear ***long-tail*** pattern across reasoning steps. As illustrated in Figure 1, most reasoning steps cluster in a low-entropy region, reflecting confident token commitments, while a small but consistent tail extends to

† Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7073–7087

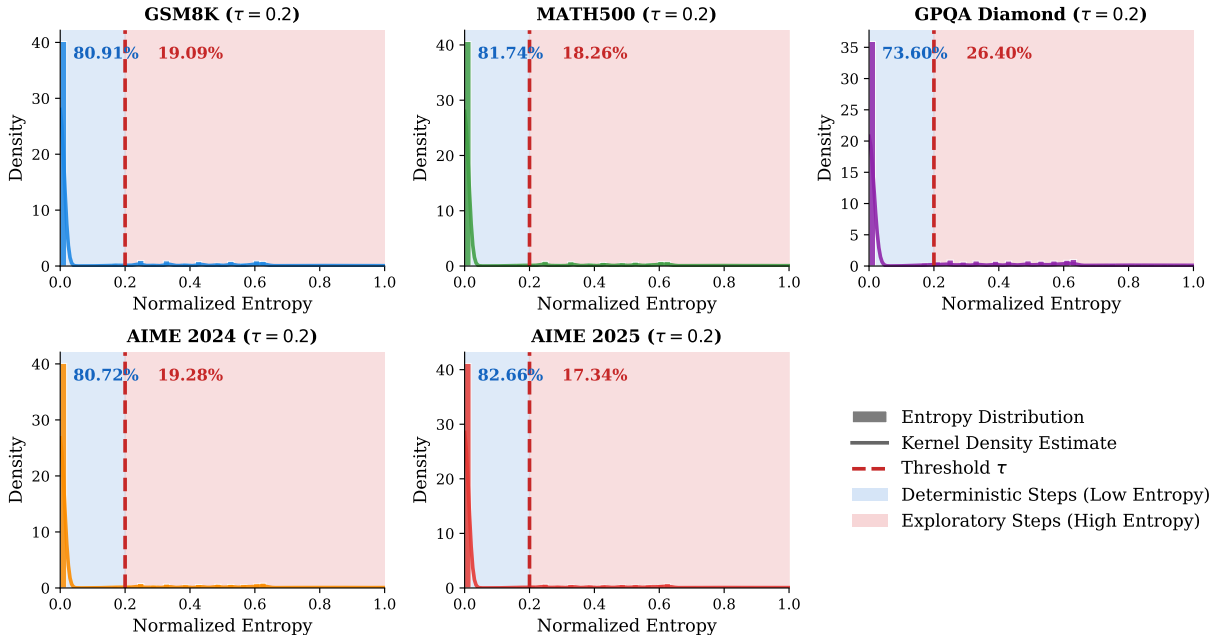

Figure 1: Normalized entropy distributions of decoding steps across five reasoning benchmarks using Qwen3-8B. Each subplot shows the density of step-wise entropy values, revealing a clear long-tail structure: the majority of steps concentrate in a low-entropy region (deterministic steps), while a smaller tail extends toward higher entropy values (exploratory steps).

higher entropy values, corresponding to moments of increased ambiguity. We refer to the former as ***deterministic steps***, where the model decisively commits to a single token, and the latter as ***exploratory steps***, where multiple candidates compete and broader exploration may be beneficial.

This entropy-based view reveals a key limitation of existing latent reasoning methods: global activation ignores the long-tail structure of model confidence, applying soft embeddings indiscriminately to both deterministic and exploratory steps. At deterministic steps where the model is already confident, this introduces unnecessary perturbations that undermine reasoning stability. Furthermore, even at exploratory steps, prior work (Wu et al., 2025) shows that soft embeddings collapse prematurely toward the dominant token, concentrating reasoning on a single trajectory and suppressing alternatives.

To address these limitations, we propose ***SeLaR***, a selective and training-free latent reasoning framework. This paper centers on two key questions: ***(i) When should latent reasoning be activated?*** SeLaR introduces an entropy-gated mechanism that activates soft embeddings only at exploratory steps, while preserving discrete decoding at deterministic steps. ***(ii) How can premature collapse toward the dominant token be mitigated?*** SeLaR incorporates an entropy-aware contrastive regularization that pushes soft embeddings away from the dominant token's direction in proportion to entropy magnitude, sustaining exploration across alternative reasoning paths. Our contributions are summarized as follows:

- We empirically show that only a small fraction of reasoning steps exhibit high uncertainty, and that activating latent reasoning exclusively at exploratory steps significantly outperforms globally applied latent reasoning.
- We propose SeLaR, a lightweight and training-free framework that selectively activates latent reasoning via entropy gating at exploratory steps, while preserving standard discrete decoding at deterministic steps. To further prevent premature collapse toward the dominant token, SeLaR introduces an entropy-aware contrastive regularization that sustains multiple latent reasoning alternatives.
- Extensive experiments on five reasoning benchmarks across multiple model scales demonstrate that SeLaR consistently outperforms standard CoT decoding and state-of-the-art training-free reasoning methods.

2 Related Work

Chain-of-Thought Reasoning

Chain-of-Thought (CoT) reasoning enhances the problem-solving capabilities of large language models by explicitly generating intermediate reasoning steps, and has become a central paradigm for improving multi-step reasoning (Zhou et al., 2023; Shinn et al., 2023; Madaan et al., 2023; Zheng et al., 2024; Wang et al., 2024a; Havrilla et al., 2024; Shao et al., 2024; Chu et al., 2024; Wang et al., 2024b; Saunshi et al., 2024; Jin et al., 2025; Wei et al., 2025; Yu et al., 2025; Lee et al., 2025). Subsequent studies have primarily focused on improving CoT through decoding and search strategies. For example, self-consistency (Wang et al., 2023) mitigates the instability of single reasoning paths by sampling multiple trajectories and aggregating their predictions, while tree- (Yao et al., 2023) or graph-based (Besta et al., 2024) search methods explicitly explore multiple discrete reasoning paths to improve robustness. Despite its strong empirical performance, CoT operates by committing to a single sequence of discrete tokens at each step, which can obscure or eliminate information about other plausible reasoning trajectories.

Latent Reasoning

Latent reasoning differs from explicit CoT by leveraging hidden states or soft embeddings (Deng et al., 2023; Geiping et al., 2025; Yang et al., 2025b; Shalev et al., 2024; Mohtashami et al., 2024; Wu et al., 2026; Wang et al., 2025; Su et al., 2025; Zhang et al., 2025a) to convey intermediate reasoning signals. Prior work in this area generally falls into two categories. Fine-tuning-based methods (Hao et al., 2025; Cheng and Durme, 2024; Xu et al., 2025) propagate hidden states across reasoning steps via full or partial fine-tuning, enabling implicit multi-step reasoning that goes beyond discrete token generation. In contrast, training-free methods (Zhang et al., 2025b; Wu et al., 2025; Shi et al., 2025) replace discrete token inputs with probability-weighted soft embeddings, allowing models to operate in continuous space without parameter updates. Our approach belongs to the latter category, but diverges from existing paradigms that apply latent reasoning globally. Specifically, SeLaR employs an entropy-gated mechanism to selectively activate latent reasoning only at exploratory steps, while incorporating a contrastive regularization strategy to prevent premature collapse toward the dominant token's trajectory during the reasoning process.

3 Method

3.1 Overview

We propose SeLaR, a selective and training-free latent reasoning framework that dynamically activates latent reasoning only when necessary. The core idea is to avoid globally propagating soft embeddings throughout the entire decoding process. Instead, SeLaR leverages token-level entropy as a confidence signal to identify high-uncertainty exploratory steps, at which latent reasoning is selectively enabled. For deterministic steps where the model exhibits high confidence, standard discrete decoding is preserved to maintain stability and efficiency.

As shown in Figure 2, SeLaR comprises two components: (1) an ***entropy-gated selective mechanism*** that determines when latent reasoning should be activated during decoding, and (2) an ***entropy-aware contrastive regularization*** that mitigates the tendency of soft embeddings to overemphasize the highest-probability token, which increasingly dominates subsequent predictions and suppresses alternative reasoning paths.

3.2 Background

Standard Chain-of-Thought Reasoning Given an input query q, a language model $\mathcal{L}$ generates a reasoning sequence $r_{1:T} = (x_1, \ldots, x_T)$ followed by a final answer a. At each decoding step t, the model produces a conditional distribution over the vocabulary $\mathcal{V}$:

$$p_t(v) = p(v \mid q, x_{<t}), \quad v \in \mathcal{V}. \tag{1}$$

Standard Chain-of-Thought (CoT) decoding commits to a single discrete token x_t at each step:

$$x_t = \begin{cases} \arg\max_{v \in \mathcal{V}} p_t(v), & \textit{Greedy} \\ v \sim \tilde{p}_t(v), & \textit{Sampling} \end{cases} \tag{2}$$

where $\tilde{p}_t$ is the filtered distribution obtained by applying temperature scaling and truncation strategies (e.g., top-k, top-p) to the original distribution p_t. The embedding e_{x_t} is then used as input for the next decoding step.

Latent Reasoning with Soft Embeddings Latent reasoning methods replace discrete token inputs with soft embeddings to preserve distributional

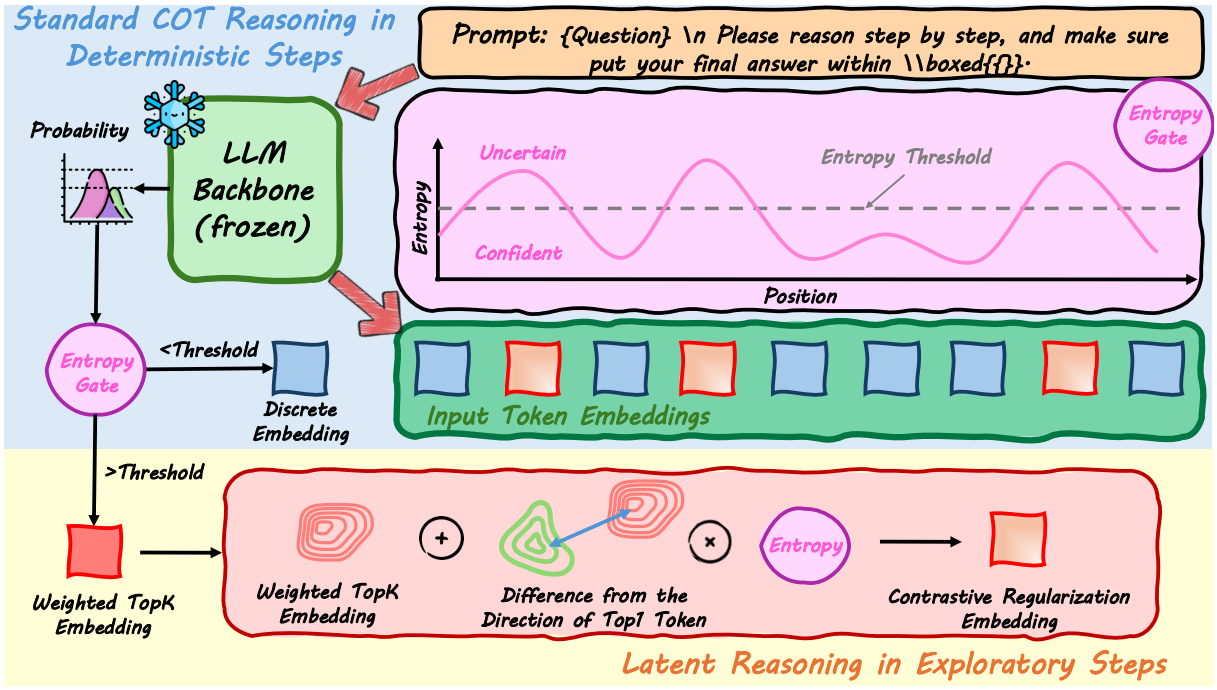

Figure 2: Overview of SeLaR. At each decoding step, we compute the normalized entropy over top-k tokens. If entropy falls below threshold τ (deterministic step), standard discrete decoding is applied. Otherwise (exploratory step), we construct a soft embedding from top-k candidates and apply contrastive regularization to push it away from the dominant token, encouraging exploration of alternative reasoning paths.

information. Let $E \in \mathbb{R}^{|\mathcal{V}| \times d}$ denote the embedding matrix. At step t, instead of committing to a sampled token, a soft embedding is computed as:

$$e_t = \sum_{v \in \mathcal{V}} p_t(v) \cdot e_v = \sum_{v \in \mathcal{V}} p_t(v) \cdot E_v. \quad (3)$$

This soft embedding is fed to the model as the next-step input, enabling implicit exploration of multiple candidate tokens within a single forward pass.

3.3 Entropy-gated Selective Mechanism

Motivation. Existing training-free latent reasoning methods propagate soft embeddings at every decoding step. While enabling richer representations, this global activation injects unnecessary perturbation into confident steps, undermining reasoning stability. Our key insight is that ***latent reasoning is only necessary when the model is uncertain***, motivating a selective mechanism that activates it only at critical exploratory steps.

Entropy as a Measure of Uncertainty At decoding step t, the model produces a predictive distribution $p_t(\cdot)$ over the vocabulary $\mathcal{V}$. Rather than computing entropy over the full vocabulary as in prior work (Shi et al., 2025), we estimate uncertainty using the top-k most probable tokens, which dominate the model's predictive mass and are most relevant for decision making. Specifically, let $\mathcal{V}_k \subset \mathcal{V}$ denote the set of top-k tokens under p_t. We first renormalize the distribution over $\mathcal{V}_k$:

$$\hat{p}_t(v) = \frac{p_t(v)}{\sum_{u \in \mathcal{V}_k} p_t(u)}, \quad v \in \mathcal{V}_k, \quad (4)$$

and define the truncated entropy as:

$$H_t = -\sum_{v \in \mathcal{V}_k} \hat{p}_t(v) \log \hat{p}_t(v), \quad (5)$$

$$\bar{H}_t = \text{clamp}\left(\frac{H_t}{\log k},\ 0,\ 1\right). \quad (6)$$

This top-k entropy captures the model's uncertainty among its most plausible candidates while avoiding perturbation from the low-probability tokens. Low entropy indicates confident predictions dominated by a small number of candidates, whereas high entropy reflects ambiguity among multiple competing tokens.

Threshold Selection The entropy threshold τ determines when latent reasoning is activated. Across models and benchmarks, $\bar{H}_t$ exhibits a clear long-tail pattern during CoT decoding: a dominant low-entropy region where a single token commands the predictive mass (deterministic steps), and a sparse

long-tail high-entropy region where multiple tokens compete (exploratory steps). The low-density transition between these two regions marks a qualitative shift from single-token dominance to multi-token competition, providing a principled and natural boundary for selecting τ. Consequently, τ is positioned within this transition band, serving as a separator that demarcates high-confidence deterministic steps from low-confidence exploratory ones. As shown in Appendix B, SeLaR is robust to the exact choice of τ, with stable performance across $\tau \in [0.3, 0.7]$.

Entropy-Gated Selective Activation Given the entropy threshold τ, the input for the next step is then computed as:

$$e_t = \begin{cases} E_{x_t}, & \text{if } \bar{H}_t \leq \tau, \\ \sum_{v \in \mathcal{V}_k} \hat{p}_t(v) \cdot e_v, & \text{if } \bar{H}_t > \tau, \end{cases} \tag{7}$$

where e_v denotes the embedding of token v. At deterministic steps, the model follows standard discrete decoding by committing to a single sampled token. At exploratory steps, latent reasoning is activated by replacing the discrete token with a soft embedding. This entropy-gated mechanism enables latent reasoning only when it is most beneficial, while maintaining the stability of standard decoding elsewhere.

3.4 Entropy-aware Contrastive Regularization

Motivation Selective activation addresses ***when*** to apply latent reasoning, but does not address ***how*** to maintain effective exploration once activated. We now turn to a complementary challenge: preventing soft embeddings from prematurely collapsing back to a single token during the reasoning process.

Premature Collapse in Latent Reasoning Although soft embeddings enable implicit exploration of multiple candidate tokens, prior work (Wu et al., 2025) has identified a *premature collapse* phenomenon: during latent reasoning, soft embeddings quickly become dominated by the highest-probability token, effectively degenerating to greedy decoding. Formally, let $v_t^* = \arg\max_{v \in \mathcal{V}_k} \hat{p}_t(v)$ denote the dominant token at step t. The soft embedding e_t tends to align increasingly with $e_{v_t^*}$ as decoding proceeds. This alignment accelerates convergence toward a single trajectory, undermining the multi-path exploration that soft embeddings are designed to enable.

Entropy-aware Contrastive Regularization To counteract premature collapse, we introduce a contrastive regularization that explicitly pushes the soft embedding away from the dominant token direction. At each exploratory step, we compute the difference between the soft embedding and the dominant token embedding:

$$\Delta_t = e_t - e_{v_t^*}, \quad \hat{\Delta}_t = \frac{\Delta_t}{\|\Delta_t\| + \epsilon}, \tag{8}$$

where $\hat{\Delta}_t$ is the unit direction pointing from $e_{v_t^*}$ toward e_t. The regularized soft embedding is then computed as:

$$\tilde{e}_t = e_t + \bar{H}_t \cdot \hat{\Delta}_t \cdot \|\Delta_t\|. \tag{9}$$

This formulation scales the repulsion from the dominant token according to the model's uncertainty: when entropy is high, the regularization effect is strong, encouraging broader exploration; as the model becomes confident, the effect diminishes naturally.

4 Experiments

4.1 Settings

Datasets We conduct experiments on five reasoning datasets, including GSM8K (Cheng and Durme, 2024), MATH500 (Hendrycks et al., 2021), AIME2024 (HuggingFaceH4, 2024), AIME2025 (Yentinglin, 2025) in the mathematical domain, and GPQA-Diamond (Rein et al., 2024) for STEM reasoning. For more details, please refer to Appendix A.2.

Baselines We compare SeLaR against four baselines: (1) standard CoT reasoning with sampling, (2) standard CoT reasoning with greedy decoding, (3) Soft Thinking (Zhang et al., 2025b), a training-free latent reasoning method that globally applies soft embeddings, and (4) SwiReasoning (SwiR) (Shi et al., 2025), which switches between explicit and latent reasoning modes based on the relative entropy increase between adjacent decoding steps. However, relying on between-step entropy deltas makes the trigger prone to spurious firing, forcing SwiR to resort to window-based smoothing heuristics that introduce additional hyperparameters.

Table 1: Detailed results on reasoning benchmarks. Results highlighted in green indicate performance comparable to or better than CoT (Sampling). Results highlighted in red indicate a performance drop relative to CoT (Sampling).

Method	GSM8K	MATH500	GPQA	AIME24	AIME25	Avg
	Qwen3-1.7B (Yang et al., 2025a)					
CoT (Sampling)	90.07	92.00	**39.39**	50.00	33.33	60.96
CoT (Greedy)	88.32	90.60	31.31	40.00	30.00	56.05
Soft Thinking	89.46	91.00	33.83	36.67	**36.67**	57.53
SwiR	89.84	92.00	37.88	46.67	23.33	57.94
SeLaR	**90.60**	**92.60**	35.35	**53.33**	**36.67**	**61.71**
	Qwen3-8B (Yang et al., 2025a)					
CoT (Sampling)	95.45	**98.00**	61.62	76.67	66.67	79.68
CoT (Greedy)	95.22	96.20	55.05	70.00	63.33	75.96
Soft Thinking	94.92	95.80	57.58	70.00	66.67	76.99
SwiR	95.68	97.00	**62.63**	60.00	66.67	76.40
SeLaR	**95.83**	97.00	61.62	**83.33**	**80.00**	**83.56**
	Qwen3-32B (Yang et al., 2025a)					
CoT (Sampling)	95.83	97.40	66.16	80.42	72.08	82.38
CoT (Greedy)	95.91	97.20	69.70	80.00	73.33	83.23
Soft Thinking	95.75	97.40	67.17	74.58	66.25	80.23
SwiR	**96.21**	**98.40**	**70.20**	82.92	73.75	84.30
SeLaR	96.06	97.60	67.17	**83.33**	**80.00**	**84.83**

Implementation Details. We evaluate SeLaR on three reasoning-oriented LLMs: Qwen3-1.7B, Qwen3-8B, and Qwen3-32B (Yang et al., 2025a). All experiments are implemented using the Hugging Face Transformers framework (Wolf et al., 2020). For fair comparison, all baselines are reproduced under identical hardware conditions (4× NVIDIA RTX PRO 6000 GPUs) using the official implementation and reported hyperparameters. Since the SwiR hyperparameters for Qwen3-32B are not publicly available, we directly adopt the baseline results reported in (Shi et al., 2025). All methods use the same decoding settings: temperature 0.6, top-p 0.95, top-k 20, and min-p 0.0. Dataset-specific entropy thresholds for SeLaR are provided in Appendix B. To further assess generality across model families, we additionally report results on DeepSeek-R1-Distill-Llama-8B (DeepSeek-AI et al., 2025) in Appendix C.

4.2 Results

Table 1 presents the main results across five reasoning benchmarks and three model scales.

Finding 1: SeLaR consistently outperforms baselines on average.

Across all model scales, SeLaR achieves the highest average accuracy, improving upon CoT (Sampling) by +0.75%, +3.88%, and +2.45% on Qwen3-1.7B, Qwen3-8B, and Qwen3-32B, respectively. Notably, SeLaR is the only method that consistently surpasses CoT across all model sizes. In contrast, Soft Thinking and SwiR exhibit inconsistent behavior: while occasionally matching or exceeding CoT on individual benchmarks, their average performance frequently falls below the CoT baseline.

Finding 2: Significant gains on challenging benchmarks.

SeLaR's advantage is most pronounced on the hardest benchmarks, AIME 2024 and AIME 2025, which demand deep multi-step reasoning and precise numerical computation. On Qwen3-8B, SeLaR improves AIME 2024 from 76.67% to 83.33% (+6.66%) and AIME 2025 from 66.67% to 80.00% (+13.33%). Similar trends hold on Qwen3-1.7B and Qwen3-32B. We attribute these gains to the complementary effect of the two components: entropy gating concentrates latent reasoning at the most uncertain and consequential steps, while contrastive regularization prevents premature collapse

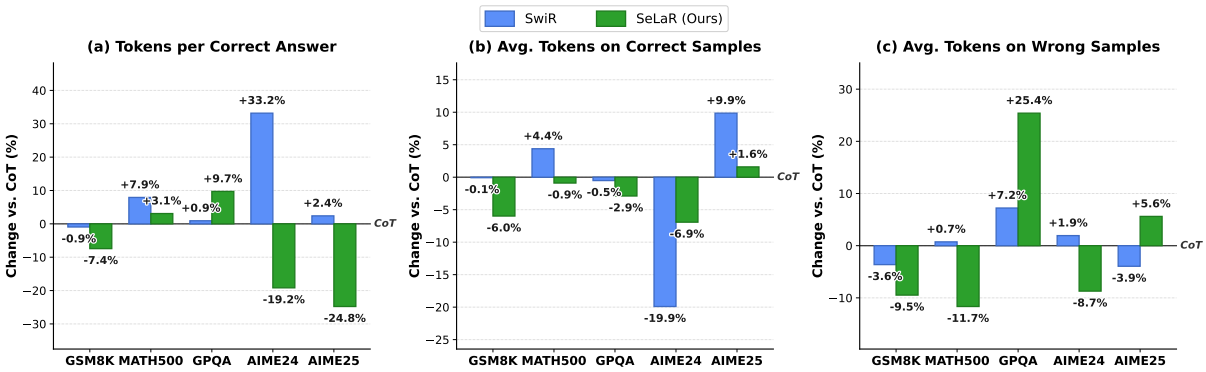

Figure 3: Computational overhead of SeLaR and SwiR relative to CoT (Sampling) on Qwen3-8B. **(a)** Tokens per correct answer: SeLaR beats SwiR on 4 of 5 benchmarks, most notably on AIME 2024/2025. **(b)** Average tokens on correctly-answered samples. **(c)** Average tokens on wrongly-answered samples.

at precisely those steps where a wrong commitment would cascade into irreversible reasoning errors. Together, they provide the greatest benefit on problems where a single misstep is most costly.

4.3 Computational Overhead

The cost-effectiveness of a reasoning method is jointly determined by the tokens it spends per problem and the accuracy it achieves. Reporting either in isolation is misleading: a method that reduces token usage while sacrificing accuracy is not truly more efficient, and one that improves accuracy at disproportionate token cost is not truly faster. We therefore report three complementary metrics that together characterize cost-effectiveness in Figure 3, all as percentage changes relative to the CoT (Sampling) baseline on Qwen3-8B.

We denote the average tokens on correctly- and wrongly-answered samples as T_c and T_w, respectively, and write accuracy as α. Our headline metric is ***Tokens per Correct Answer***:

$$TPCA = \frac{\alpha \cdot T_c + (1-\alpha) \cdot T_w}{\alpha}, \qquad (10)$$

Finding 3: On average, SeLaR is more cost-effective than SwiR, with the advantage widening on the hardest reasoning tasks.

On TPCA (Figure 3a), SeLaR outperforms SwiR by 6.5, 4.8, 52.4, and 27.2 percentage points on GSM8K, MATH500, AIME 2024, and AIME 2025, respectively. This advantage is most pronounced on AIME 2024, where SeLaR reduces TPCA by 19.2% relative to CoT while SwiR inflates it by 33.2%. The main reason is that SwiR's accuracy on AIME 2024 drops to 60%, only marginally compensated by its shorter reasoning trajectories. The sole exception is GPQA, where SeLaR's TPCA is 9.7% higher than CoT. We attribute this to GPQA's knowledge-intensive nature, where latent exploration at uncertain decoding steps provides little benefit when correct answers depend on domain recall rather than multi-step reasoning.

Finding 4: SwiR's apparent efficiency advantage on correct samples is a survivorship bias.

SeLaR's gains arise from both shorter correct trajectories and higher accuracy. On correctly-answered samples (Figure 3b), SeLaR's T_c falls within -6.0% to $+1.6\%$ of CoT across all benchmarks, confirming that selective activation and contrastive regularization introduce no runtime overhead. However, SwiR's apparent -19.9% reduction on AIME 2024 is an artifact of survivorship bias, as SwiR answers only the easier 60% of problems correctly, and those naturally require fewer tokens. TPCA corrects for this survivorship bias, revealing SwiR's true $+33.2\%$ cost inflation on AIME 2024.

4.4 Ablation Studies

We ablate each component of SeLaR on Qwen3-8B to quantify its individual contribution. Results are presented in Table 2.

Effect of Selective Activation Removing selective activation (i.e., applying soft embeddings globally at every step) leads to an average performance drop of 5.19% (from 83.56% to 78.37%), even falling below the CoT baseline. This confirms that

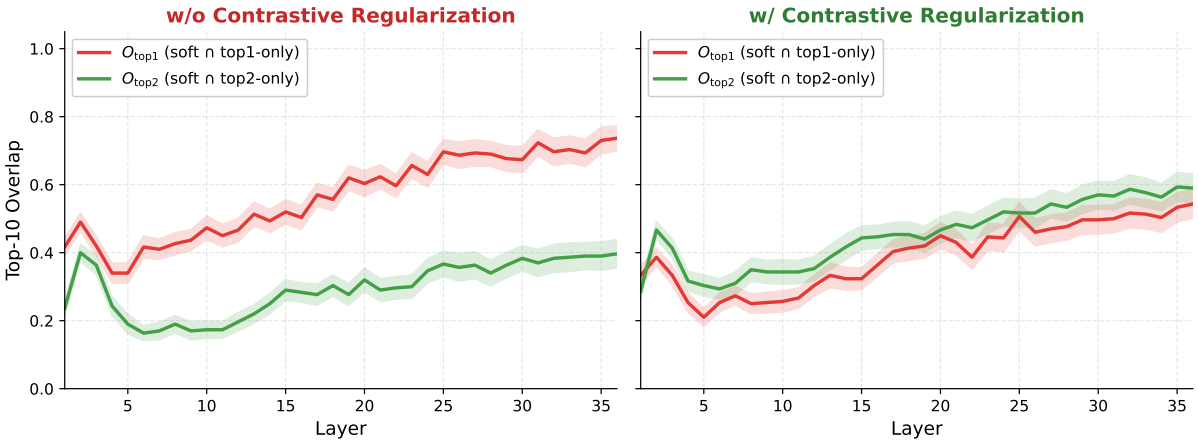

Figure 4: Layer-wise top-k overlap (k=10) between the soft-embedding forward pass and single-token reference passes, aggregated over N=200 branching steps from AIME 2025 on Qwen3-8B. Shaded bands denote standard error. **Left:** without contrastive regularization, O_{top1} dominates O_{top2} in deep layers, reproducing the collapse behavior of (Wu et al., 2025). **Right:** with contrastive regularization, *both* overlaps remain non-zero and comparable, indicating multiple reasoning trajectories coexist rather than a single one being swapped.

indiscriminate activation perturbs high-confidence steps where the model is already committed, destabilizing reasoning chains that would otherwise succeed.

Effect of Contrastive Regularization Removing contrastive regularization results in an average performance drop of 7.82% (from 83.56% to 75.74%). The degradation is particularly severe on challenging benchmarks: AIME 2024 drops from 83.33% to 70.00% and AIME 2025 drops from 80.00% to 60.00%. While this behavioral evidence confirms that contrastive regularization is indispensable, it does not reveal *how* the component intervenes inside the forward pass. We investigate this mechanism in the following subsection.

4.5 How Contrastive Regularization Works

To understand *how* contrastive regularization works, we must answer two questions: (i) does it produce consistent effects across exploratory steps rather than isolated cases, and (ii) does it truly preserve multiple reasoning trajectories, or does it merely swap the dominant top-1 token for a different top-k candidate while remaining single-threaded? We address both via a logit lens analysis inspired by the approach of (Nostalgebraist, 2020).

Logit Lens Setting. We collect $N = 200$ branching steps from 10 random AIME 2025 problems under SeLaR on Qwen3-8B, where a branching step is an exploratory step (entropy above τ) with top-1/top-2 probability ratio below 2.0. For each step t, we cache the KV state and run four independent forward passes at step $t{+}1$, with input embedding set to $e_{v_t^*}$, $e_{v_t^{**}}$, e_t, and $\tilde{e}_t$ respectively, where v_t^*, v_t^{**} are the top-1/top-2 tokens and $e_t, \tilde{e}_t$ are the soft embeddings without and with contrastive regularization. For each pass, we apply the logit lens at every layer ℓ and take the top-k projected tokens (k=10), denoted $\mathcal{T}_\ell^{\cdot}$. We then measure how much each soft-embedding pass shares with the two references:

$$O_{\text{top1}}(\ell) = \frac{|\mathcal{T}_\ell^{\text{soft}} \cap \mathcal{T}_\ell^{\text{top1}}|}{k}, \tag{11}$$

$$O_{\text{top2}}(\ell) = \frac{|\mathcal{T}_\ell^{\text{soft}} \cap \mathcal{T}_\ell^{\text{top2}}|}{k}, \tag{12}$$

quantifying how much of each candidate's reasoning content the soft-embedding forward pass still carries at layer ℓ. We average both across all N steps.

Logit Lens Results. Figure 4 shows the aggregated curves. Without contrastive regularization (left), O_{top1} rises from $\sim$0.45 to $\sim$0.73 while O_{top2} stagnates around $\sim$0.40, reproducing the collapse behavior of (Wu et al., 2025): the forward pass progressively collapses onto the top-1 trajectory and suppresses the top-2 alternative. Unlike (Wu et al., 2025), O_{top1} does not saturate to 1 because our soft embedding mixes k top candidates computed from the model's actual output distribution, rather than a manually balanced two-token mixture. The key observation is thus the relative gap between the two curves, not their absolute values.

Table 2: Ablation study on Qwen3-8B. We evaluate the contribution of each component in SeLaR.

Method	GSM8K	MATH500	GPQA	AIME24	AIME25	Avg
CoT (Sampling)	95.45	**98.00**	61.62	76.67	66.67	79.68
SeLaR (Full)	**95.83**	97.00	**61.62**	**83.33**	**80.00**	**83.56**
Component Ablation						
w/o Selective Activation	95.14	95.80	57.58	76.67	66.67	78.37
w/o Contrastive Reg.	94.92	96.20	57.58	70.00	60.00	75.74

Finding 5: Contrastive regularization prevents the collapse into single-threaded behavior, keeping multiple candidate trajectories alive in deep layers.

A natural concern is whether contrastive regularization merely shifts the collapse from the top-1 to another top-k candidate, leaving the forward pass still single-threaded. Note that if this were the case, we would expect O_{top1} to decrease and O_{top2} to increase, mirroring the left panel with the two curves swapped. Instead, both overlaps remain substantially non-zero and comparable in deep layers: O_{top2} climbs to $\sim$0.60 while O_{top1} settles at $\sim$0.55. This confirms that contrastive regularization implements genuine ***sustained exploration of latent reasoning paths***.

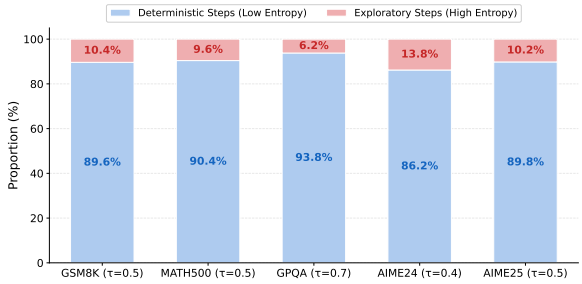

Figure 5: Activation frequency analysis on Qwen3-8B. Exploratory steps (high entropy) account for 6.2%–13.8% of total reasoning tokens.

4.6 Detailed Analysis

Sensitivity Analysis Appendix B examines the sensitivity of SeLaR to the entropy threshold τ and top-k value. We find that performance remains stable across $\tau \in [0.3, 0.7]$ with $k = 3$, achieving the best average accuracy (80.86%) at $\tau = 0.5$. For the top-k value, $k = 3$ consistently outperforms larger values, as excessive candidates dilute the soft embedding with low-probability tokens.

Activation Frequency Analysis Figure 5 shows the proportion of exploratory steps (high entropy) versus deterministic steps (low entropy) across benchmarks. Exploratory steps account for 6.2%–13.8% of total tokens, averaging 10.0%. This variation reflects dataset-specific thresholds: AIME 2024 ($\tau = 0.4$) exhibits the highest activation frequency, while GPQA-Diamond ($\tau = 0.7$) shows the lowest. This confirms that selective activation targets approximately one in ten tokens where exploration is most beneficial. Activation frequency analysis on DeepSeek-R1-Distill-Llama-8B is provided in Appendix C.

5 Conclusion

We present **SeLaR**, a training-free latent reasoning framework that selectively activates soft embeddings based on entropy. SeLaR preserves discrete token commitments at deterministic steps for stability, and activates soft embeddings at exploratory steps to enable alternative reasoning trajectories. An entropy-aware contrastive regularization further mitigates premature collapse toward the dominant token. More broadly, SeLaR addresses both *when* and *how* latent reasoning should be applied: token-level entropy signals when to activate, while contrastive regularization governs how exploration is sustained. This perspective offers new insights into designing adaptive reasoning mechanisms in large language models.

Acknowledgments

This work is supported by National Key Research and Development Program of China (2024YFE0203100), Guangdong Provincial Key Laboratory of Ultra High Definition Immersive Media Technology (Grant No. 2024B1212010006), and Shenzhen Science and Technology Program (JCYJ20230807120800001).

Limitations

While SeLaR demonstrates consistent improvements across multiple benchmarks, some limitations warrant discussion.

Constraints of Token Embedding Space Like other training-free latent reasoning methods, SeLaR operates in the token embedding space at the input level. Although contrastive regularization effectively mitigates premature collapse toward the dominant token, this approach is inherently limited in expressiveness compared to manipulating hidden states directly. Future work on latent reasoning should explore the hidden state space, which serves as the primary information carrier for reasoning in LLMs.

Sensitivity to Base Model Confidence SeLaR yields larger improvements on base models with higher confidence (e.g., Qwen3-8B) than on those with lower confidence (e.g., DeepSeek-R1-Distill-Llama-8B). Our analysis indicates that less confident models exhibit higher entropy more frequently, triggering excessive exploratory steps. Future work should investigate confidence-aware activation mechanisms or explore signals beyond entropy to better adapt to varying base model characteristics.

References

Marah Abdin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat Behl, Lingjiao Chen, Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, Piero Kauffmann, Yash Lara, Caio César Teodoro Mendes, Arindam Mitra, Besmira Nushi, Dimitris Papailiopoulos, Olli Saarikivi, Shital Shah, Vaishnavi Shrivastava, and 4 others. 2025. Phi-4-reasoning technical report. *Preprint*, arXiv:2504.21318.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2024. Graph of thoughts: Solving elaborate problems with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17682–17690.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Jeffrey Cheng and Benjamin Van Durme. 2024. Compressed chain of thought: Efficient reasoning through dense representations. *Preprint*, arXiv:2412.13171.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, and 48 others. 2022. Palm: Scaling language modeling with pathways. *Preprint*, arXiv:2204.02311.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2024. Navigate through enigmatic labyrinth a survey of chain of thought reasoning: Advances, frontiers and future. *Preprint*, arXiv:2309.15402.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Yuntian Deng, Kiran Prasad, Roland Fernandez, Paul Smolensky, Vishrav Chaudhary, and Stuart Shieber. 2023. Implicit chain of thought reasoning via knowledge distillation. *Preprint*, arXiv:2311.01460.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. *Preprint*, arXiv:2103.10360.

Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian R. Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. 2025. Scaling up test-time compute with latent reasoning: A recurrent depth approach. *Preprint*, arXiv:2502.05171.

Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. 2024. Think before you speak: Training language models with pause tokens. *Preprint*, arXiv:2310.02226.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. 2025. Training large language models to reason in a continuous latent space. *Preprint*, arXiv:2412.06769.

Alex Havrilla, Yuqing Du, Sharath Chandra Raparthy, Christoforos Nalmpantis, Jane Dwivedi-Yu, Maksym Zhuravinskyi, Eric Hambro, Sainbayar Sukhbaatar, and Roberta Raileanu. 2024. Teaching large language models to reason with reinforcement learning. *Preprint*, arXiv:2403.04642.

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

HuggingFaceH4. 2024. AIME 2024: American Invitational Mathematics Examination. Hugging Face Dataset.

Mingyu Jin, Weidi Luo, Sitao Cheng, Xinyi Wang, Wenyue Hua, Ruixiang Tang, William Yang Wang, and Yongfeng Zhang. 2025. Disentangling memory and reasoning ability in large language models. *Preprint*, arXiv:2411.13504.

Kuang-Huei Lee, Ian Fischer, Yueh-Hua Wu, Dave Marwood, Shumeet Baluja, Dale Schuurmans, and Xinyun Chen. 2025. Evolving deeper llm thinking. *Preprint*, arXiv:2501.09891.

Jindong Li, Yali Fu, Li Fan, Jiahong Liu, Yao Shu, Chengwei Qin, Menglin Yang, Irwin King, and Rex Ying. 2025. Implicit reasoning in large language models: A comprehensive survey. *Preprint*, arXiv:2509.02350.

Elita Lobo, Chirag Agarwal, and Himabindu Lakkaraju. 2025. On the impact of fine-tuning on chain-of-thought reasoning. *Preprint*, arXiv:2411.15382.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. *Preprint*, arXiv:2303.17651.

Amirkeivan Mohtashami, Matteo Pagliardini, and Martin Jaggi. 2024. Cotformer: A chain-of-thought driven architecture with budget-adaptive computation cost at inference. *Preprint*, arXiv:2310.10845.

Nostalgebraist. 2020. Interpreting gpt: The logit lens. Blog post on LessWrong.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, and 244 others. 2024a. Openai o1 system card. *Preprint*, arXiv:2412.16720.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024b. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Jacob Pfau, William Merrill, and Samuel R. Bowman. 2024. Let's think dot by dot: Hidden computation in transformer language models. *Preprint*, arXiv:2404.15758.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Nikunj Saunshi, Stefani Karp, Shankar Krishnan, Sobhan Miryoosefi, Sashank J. Reddi, and Sanjiv Kumar. 2024. On the inductive bias of stacking towards improving reasoning. *Preprint*, arXiv:2409.19044.

Yuval Shalev, Amir Feder, and Ariel Goldstein. 2024. Distributional reasoning in llms: Parallel reasoning processes in multi-hop reasoning. *Preprint*, arXiv:2406.13858.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Dachuan Shi, Abedelkadir Asi, Keying Li, Xiangchi Yuan, Leyan Pan, Wenke Lee, and Wen Xiao. 2025. Swireasoning: Switch-thinking in latent and explicit for pareto-superior reasoning llms. *Preprint*, arXiv:2510.05069.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Preprint*, arXiv:2303.11366.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 465 others. 2025. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

DiJia Su, Hanlin Zhu, Yingchen Xu, Jiantao Jiao, Yuandong Tian, and Qinqing Zheng. 2025. Token assorted: Mixing latent and text tokens for improved language model reasoning. *Preprint*, arXiv:2502.03275.

Wenhui Tan, Jiaze Li, Jianzhong Ju, Zhenbo Luo, Jian Luan, and Ruihua Song. 2025. Think silently, think fast: Dynamic latent compression of llm reasoning chains. *Preprint*, arXiv:2505.16552.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, Chuning Tang, Congcong Wang, Dehao Zhang, Enming Yuan, Enzhe Lu, Fengxiang Tang, Flood Sung, Guangda Wei, Guokun Lai, and 77 others. 2025. Kimi k1.5: Scaling reinforcement learning with llms. *Preprint*, arXiv:2501.12599.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Peiyi Wang, Lei Li, Zhihong Shao, R. X. Xu, Damai Dai, Yifei Li, Deli Chen, Y. Wu, and Zhifang Sui. 2024a. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. *Preprint*, arXiv:2312.08935.

Xiaoqiang Wang, Suyuchen Wang, Yun Zhu, and Bang Liu. 2025. System-1.5 reasoning: Traversal in language and latent spaces with dynamic shortcuts. *Preprint*, arXiv:2505.18962.

Xinyi Wang, Lucas Caccia, Oleksiy Ostapenko, Xingdi Yuan, William Yang Wang, and Alessandro Sordoni. 2024b. Guiding language model reasoning with planning tokens. *Preprint*, arXiv:2310.05707.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Ting-Ruen Wei, Haowei Liu, Xuyang Wu, and Yi Fang. 2025. A survey on feedback-based multi-step reasoning for large language models on mathematics. *Preprint*, arXiv:2502.14333.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Huggingface's transformers: State-of-the-art natural language processing. *Preprint*, arXiv:1910.03771.

Haoyi Wu, Zhihao Teng, and Kewei Tu. 2026. Parallel continuous chain-of-thought with jacobi iteration. *Preprint*, arXiv:2506.18582.

Junhong Wu, Jinliang Lu, Zixuan Ren, Gangqiang Hu, Zhi Wu, Dai Dai, and Hua Wu. 2025. Llms are single-threaded reasoners: Demystifying the working mechanism of soft thinking. *Preprint*, arXiv:2508.03440.

Yige Xu, Xu Guo, Zhiwei Zeng, and Chunyan Miao. 2025. Softcot: Soft chain-of-thought for efficient reasoning with llms. *Preprint*, arXiv:2502.12134.

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, and 41 others. 2025a. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. 2025b. Do large language models latently perform multi-hop reasoning? *Preprint*, arXiv:2402.16837.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. *Preprint*, arXiv:2210.03629.

Yentinglin. 2025. AIME 2025: American Invitational Mathematics Examination. Hugging Face Dataset.

Fangxu Yu, Lai Jiang, Haoqiang Kang, Shibo Hao, and Lianhui Qin. 2025. Flow of reasoning: Training llms for divergent reasoning with minimal examples. *Preprint*, arXiv:2406.05673.

Jintian Zhang, Yuqi Zhu, Mengshu Sun, Yujie Luo, Shuofei Qiao, Lun Du, Da Zheng, Huajun Chen, and Ningyu Zhang. 2025a. Lightthinker: Thinking step-by-step compression. *Preprint*, arXiv:2502.15589.

Zhen Zhang, Xuehai He, Weixiang Yan, Ao Shen, Chenyang Zhao, Shuohang Wang, Yelong Shen, and Xin Eric Wang. 2025b. Soft thinking: Unlocking the reasoning potential of llms in continuous concept space. *Preprint*, arXiv:2505.15778.

Chuanyang Zheng, Zhengying Liu, Enze Xie, Zhenguo Li, and Yu Li. 2024. Progressive-hint prompting improves reasoning in large language models. *Preprint*, arXiv:2304.09797.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. *Preprint*, arXiv:2205.10625.

Algorithm 1 SELAR (SELECTIVE LATENT REASONING)

Require: Question $x_{1:n}$, model $\mathcal{M}$, max steps $S_{\max}$, top-k, entropy threshold τ
Ensure: Answer $y_{1:m}$

Init: Embedding matrix E, max entropy $H_{\max} = \log k$
for $t = 1$ to $S_{\max}$ **do**
  $\ell_t \leftarrow \mathcal{M}(x_{1:t-1})$; $p_t \leftarrow \text{softmax}(\ell_t)$ ▷ Forward pass
  $\mathcal{V}_k \leftarrow \text{top-k}(p_t)$ ▷ Select top-k tokens
  $\hat{p}_t[v] \leftarrow p_t[v] / \sum_{v' \in \mathcal{V}_k} p_t[v']$ for $v \in \mathcal{V}_k$ ▷ Normalize over top-k
  $H_t \leftarrow -\sum_{v \in \mathcal{V}_k} \hat{p}_t[v] \log \hat{p}_t[v]$ ▷ Compute entropy
  $\bar{H}_t \leftarrow H_t / H_{\max}$ ▷ Normalize entropy to $[0, 1]$
  $x_t \leftarrow \text{Sample}(p_t)$ ▷ Sample discrete token for readable output

  if $\bar{H}_t < \tau$ **then** ▷ **Deterministic Step:** Low entropy
    $e_t \leftarrow E[x_t]$ ▷ Use discrete embedding
  else ▷ **Exploratory Step:** High entropy
    $e_t \leftarrow \sum_{v \in \mathcal{V}_k} \hat{p}_t[v] \cdot E[v]$ ▷ Soft embedding
    $v^* \leftarrow \arg\max_{v \in \mathcal{V}_k} p_t[v]$ ▷ Dominant token
    $\Delta_t \leftarrow e_t - E[v^*]$ ▷ Direction from dominant
    $\hat{\Delta}_t \leftarrow \Delta_t / (\|\Delta_t\| + \epsilon)$ ▷ Unit direction
    $\tilde{e}_t \leftarrow e_t + \bar{H}_t \cdot \hat{\Delta}_t \cdot \|\Delta_t\|$ ▷ Contrastive regularization
  end if
  Feed e_t as input embedding for next step
  if $x_t =$ <EOS> **then**
    break
  end if
end for
Extract answer y from $x_{n+1:t}$
return y

Appendix

A Supplementary Details

A.1 SeLaR Implementation

Alg 1 provides the detailed implementation of SeLaR. The core selective activation mechanism is shown in black: at each step, we compute the normalized entropy over top-k tokens and compare it against threshold τ to determine whether to use discrete embeddings (deterministic steps) or soft embeddings (exploratory steps). The contrastive regularization component is outlined in blue, which pushes the soft embedding away from the dominant token proportionally to the entropy, preventing premature collapse.

A.2 Dataset Details

We evaluate our method on five reasoning benchmarks spanning mathematical problem-solving and knowledge-intensive question answering.

GSM8K is a benchmark for evaluating multi-step mathematical reasoning in natural language. Following standard practice, we evaluate on the official test set, which contains 1,319 grade-school level math word problems requiring explicit step-by-step reasoning. 🤗: https://huggingface.co/datasets/openai/gsm8k.

MATH500 is a challenging subset of the MATH dataset, consisting of 500 high school competition-level problems spanning algebra, geometry, number theory, and calculus. The problems require non-trivial symbolic manipulation and multi-step deductive reasoning. 🤗: https://huggingface.co/datasets/HuggingFaceH4/MATH-500.

AIME 2024 is a benchmark of 30 problems from the 2024 American Invitational Mathematics Examination. Each problem demands deep multi-step reasoning and precise numerical computation, with answers constrained to integers within a fixed range. 🤗: https://huggingface.co/datasets/HuggingFaceH4/aime_2024.

Table 3: Sensitivity analysis on Qwen3-8B. We vary the entropy threshold τ and top-k value while keeping other settings fixed.

k	τ	GSM8K	MATH500	GPQA	AIME24	AIME25	Avg
Varying τ (fixed $k = 3$)							
3	0.3	95.00	96.40	53.03	76.67	80.00	80.22
3	0.4	95.22	96.60	54.55	**83.33**	70.00	79.94
3	0.5	95.60	**97.00**	55.05	76.67	**80.00**	**80.86**
3	0.6	**95.83**	96.00	**60.10**	76.67	70.00	79.72
3	0.7	95.53	96.40	**60.10**	76.67	56.67	77.07
Varying k (fixed $\tau = 0.5$)							
3	0.5	**95.60**	**97.00**	55.05	**76.67**	**80.00**	**80.86**
5	0.5	95.30	96.40	**61.62**	**76.67**	53.33	76.66
7	0.5	95.00	96.60	55.56	73.33	63.33	76.76

AIME 2025 is a benchmark of 30 problems from the 2025 AIME examination, featuring newly released competition problems with similar formats but increased novelty, providing a stringent test of generalization and reasoning robustness. 🤗: https://huggingface.co/datasets/yentinglin/aime_2025.

GPQA Diamond is the most difficult split of the GPQA benchmark, containing 198 expert-curated questions across mathematics, physics, chemistry, biology, and computer science. The questions are designed to resist superficial pattern matching and require advanced domain knowledge and rigorous reasoning. 🤗: https://huggingface.co/datasets/hendrydong/gpqa_diamond_mc.

B Sensitivity Analysis Details

Table 3 presents the sensitivity analysis for SeLaR on Qwen3-8B.

Effect of Entropy Threshold τ We vary τ from 0.3 to 0.7 with $k = 3$. Lower thresholds activate latent reasoning too frequently, introducing perturbation at high-confidence steps, while higher thresholds activate it too conservatively, limiting exploration at true exploratory steps. The optimal τ varies across datasets (e.g., $\tau = 0.4$–0.7), reflecting their inherent entropy characteristics: harder tasks benefit from reserving latent reasoning for highly uncertain steps, whereas tasks with more frequent exploratory steps favor earlier activation. Importantly, SeLaR remains stable across a wide range of thresholds ($\tau \in [0.3, 0.7]$), indicating that τ serves as a **coarse uncertainty gate** derived from the entropy distribution rather than a finely tuned hyperparameter.

Effect of Top-k Value We vary k from 3 to 7 while fixing $\tau = 0.5$. Smaller k values yield better average performance, with $k = 3$ achieving 80.86% compared to 76.66% for $k = 5$ and 76.76% for $k = 7$. This suggests that restricting soft embeddings to fewer high-probability candidates preserves semantic coherence, while larger k values dilute the representation with low-probability tokens that introduce perturbation.

Final Selection Based on the above analysis, we fix $k = 3$ across all experiments and select dataset-specific thresholds that maximize individual benchmark performance: $\tau = 0.6$ for GSM8K, $\tau = 0.5$ for MATH500, $\tau = 0.7$ for GPQA-Diamond, $\tau = 0.4$ for AIME 2024, and $\tau = 0.5$ for AIME 2025. These settings are used for all main results reported in Table 1.

C Experimental Results on Other Model Families

Table 4 presents results on DeepSeek-R1-Distill-Llama-8B, a model from a different family from Qwen3. SeLaR achieves the highest average accuracy (60.53%), outperforming CoT (Sampling) by 2.77% and SwiR by 1.25%. However, the improvements are less pronounced compared to Qwen3 models. As shown in Figure 7, DeepSeek-R1-Distill-Llama-8B exhibits higher activation frequencies (8.8%–15.3%) compared to Qwen3-8B (6.2%–13.8%), indicating lower confidence during reasoning. This triggers more frequent exploratory steps and introduces additional perturbation, limiting the effectiveness of selective activation. This

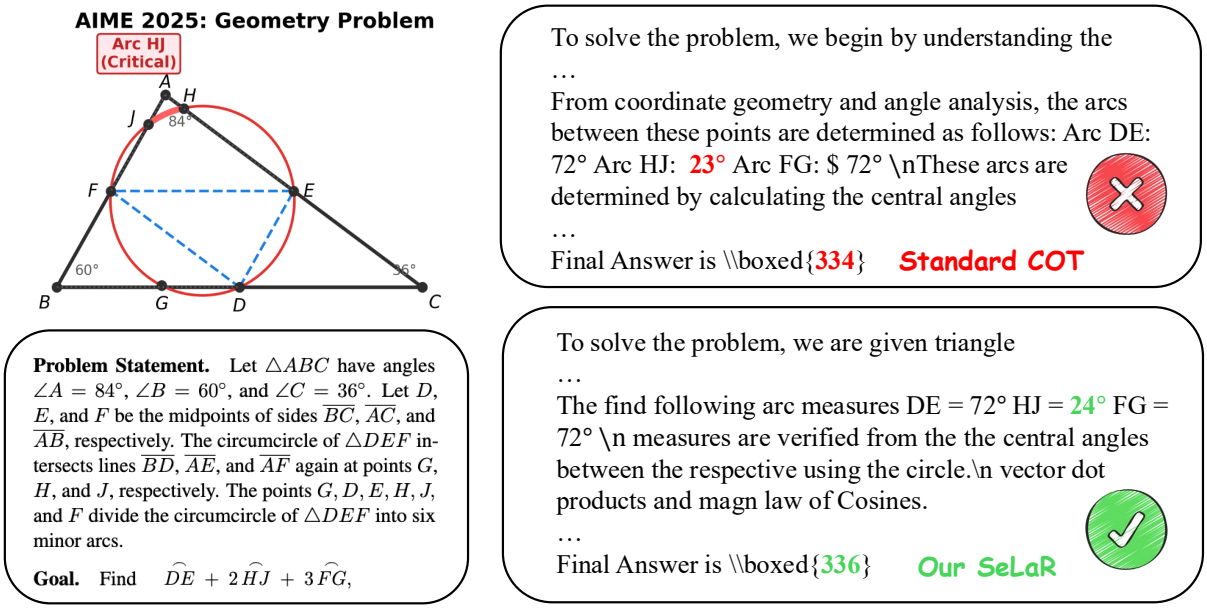

Figure 6: Case study on an AIME 2025 geometry problem. Standard CoT computes Arc HJ = 23° at the critical exploratory step, leading to an incorrect final answer of 334. SeLaR activates selective latent reasoning at this exploratory step, correctly computing Arc HJ = 24° and yielding the correct answer 336.

Table 4: Results of DeepSeek-R1-Distill-Llama-8B. Results highlighted in green indicate performance comparable to or better than CoT (Sampling). Results highlighted in red indicate a performance drop relative to CoT (Sampling).

Method	GSM8K	MATH500	GPQA	AIME24	AIME25	Avg
	DeepSeek-R1-Distill-Llama-8B (DeepSeek-AI et al., 2025)					
CoT (Sampling)	89.01	90.20	42.93	36.67	30.00	57.76
CoT (Greedy)	85.29	83.60	30.30	30.00	26.67	51.17
Soft Thinking	85.67	82.20	32.83	30.00	23.33	50.81
SwiR	89.31	87.80	**45.96**	**50.00**	23.33	59.28
SeLaR	**90.22**	**88.20**	40.91	46.67	**36.67**	**60.53**

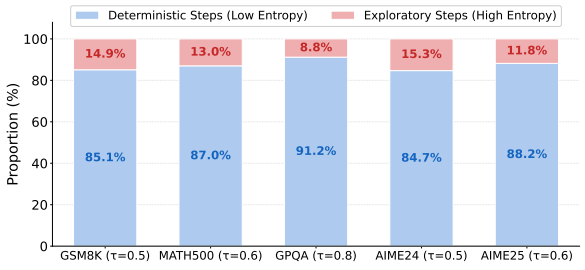

Figure 7: Activation frequency analysis on DeepSeek-R1-Distill-Llama-8B. Exploratory steps account for 8.8%–15.3% of total reasoning tokens, higher than Qwen3-8B (6.2%–13.8%).

sensitivity is a natural trade-off of the training-free design: SeLaR operates at the embedding level without modifying hidden states and therefore inevitably depends on the base model's intrinsic reasoning capability in hidden-state space.

D Case Study

Figure 6 presents a qualitative comparison on an AIME 2025 geometry problem. Both Standard CoT and SeLaR follow identical reasoning paths initially, but diverge at a critical exploratory point: computing Arc HJ. Standard CoT commits to 23° and arrives at an incorrect answer of 334, while SeLaR correctly computes 24° and yields the correct answer 336.