Beyond Tokens: Dynamic Latent Reasoning via Semantic Residual Refinement

Fangrui Lv[1], Lei Wang[2], Ruixin Hong[1], Yong Du[2], Xiangyu Wu[2], Tingting Gao[2], Guorui Zhou[2], Changshui Zhang[1,*]

[1]Institute for Artificial Intelligence, Tsinghua University(THUAI)
Beijing National Research Center for Information Science and Technology (BNRist)
Department of Automation, Tsinghua University Beijing, P.R.China
[2]Kuaishou Technology
lvfr23@mails.tsinghua.edu.cn, {wanglei16, duyong, wuxiangyu06, lisize, zhouguorui}@kuaishou.com,
hongrx16@tsinghua.org.cn, zcs@mail.tsinghua.edu.cn

Abstract

Chain-of-Thought prompting has remarkably advanced LLM reasoning by generating explicit step-by-step tokens, yet its discrete nature inherently limits expressiveness and efficiency, struggling with abstract, ambiguous, or semantically divergent cognition beyond linguistic tokens. Latent reasoning offers a promising alternative by operating in the model's internal continuous space for richer cognitive representations. However, existing methods typically rely on finetuning or token interpolation to bridge latent and input spaces, introducing training difficulty or semantic degradation. To this end, we propose Dynamic Latent Reasoning (DyLaR), a training-free framework that preserves semantic fidelity to latent space. DyLaR introduces a Semantic Residual Refinement module that progressively refines latent inputs by integrating semantic residuals from prior hidden states, thus capturing expressive semantic hierarchies that closely approximate continuous latent representations. To enhance flexibility, DyLaR further incorporates a dynamic switching policy that allows LLMs to alternate between discrete and latent reasoning based on model uncertainty, favoring explicit reasoning when confident and latent exploration under ambiguity. Empirical experiments across knowledge- and reasoning-intensive tasks demonstrate that DyLaR consistently outperforms strong baselines in both effectiveness and token efficiency. Qualitative analyses further illustrate its interpretability and flexibility in navigating complex reasoning scenarios.

Introduction

Large Language Models (LLMs) have demonstrated impressive reasoning capabilities (Guo et al. 2025; Yang et al. 2025; Dubey et al. 2024), particularly when guided by Chain-of-Thought (CoT) prompting, which elicits step-by-step intermediate reasoning in natural language (Wei et al. 2022; Chen et al. 2025a). However, explicit CoT reasoning operates strictly within the discrete token space, inherently limited by the expressive capacity of language and often introducing inefficiency (Qu et al. 2025; Liu et al. 2025). This contrasts with human cognition, which frequently transcends linguistic boundaries, relying on abstract insights, intuitive leaps, or divergent thoughts that are beyond precise verbalization (Pinker et al. 1994; Wittgenstein 2023).

*Corresponding author.

To address these limitations, latent reasoning has emerged as a promising alternative (Tack et al. 2025; Hao et al. 2024; Deng, Choi, and Shieber 2024). By shifting the reasoning process into the model's continuous hidden space, latent reasoning enables richer semantics, divergent explorations, and non-verbal cognitive representations, offering a more expressive and efficient medium for thought-like processing.

To enable "continuous thought" in latent space, one line of research replaces discrete tokens with the model's last hidden states as next inputs (Hao et al. 2024; Xu et al. 2025; Shen et al. 2025). While this approach improves token efficiency, it operates directly on hidden representations, which undermines interpretability and necessitates model fine-tuning to bridge the gap between hidden states and input embedding space (Hao et al. 2024), introducing training difficulty. Another line avoids additional training by softly approximating hidden states via output-probability-weighted interpolation over vocabulary embeddings (Zhang et al. 2025; Yue et al. 2025). Though training-free, these methods often suffer from semantic degradation, as hidden states typically lie outside the convex hull of token embeddings, making simple interpolation insufficient to capture the full latent semantics. Furthermore, most existing approaches rely on rigid latent reasoning patterns, i.e., fixed-length or full-sequence latent chains, which can introduce unnecessary noise in clear steps while restricting exploration in complex ones. This inflexibility significantly limits generalization across diverse reasoning scenarios (Chen et al. 2025b; Tack et al. 2025). These challenges motivate two key questions: **1) How can latent semantics be leveraged both effectively and efficiently? 2) When is it necessary to invoke latent reasoning?**

To this end, we propose Dynamic Latent Reasoning (DyLaR), a training-free framework that enables semantically faithful and flexible latent reasoning. Specifically, DyLaR comprises two core modules: **i) a Semantic Residual Refinement module (SRR)**. To leverage rich semantic information in latent representations without fine-tuning, SRR begins by computing a soft projection of hidden states onto the token embedding space. It then captures the semantic residual between hidden state and its soft projection, which reflects the unexpressed semantic information. Through iterative integration of multi-level residuals, SRR refines the la-

Method	Latent Reasoning	Training Free	Semantic fidelity	Dynamic Reasoning Mode	Output Interpretability
Zero-shot CoT (Wei et al. 2022)	-	✔	-	-	✔
Coconut (Hao et al. 2024)	✔	-	✔	-	-
SoftCoT (Xu et al. 2025)	✔	-	✔	-	-
Soft Thinking (Zhang et al. 2025)	✔	✔	-	-	✔
HRPO (Yue et al. 2025)	✔	-	-	-	✔
DyLaR (ours)	✔	✔	✔	✔	✔

Table 1: Comparisons between DyLaR and other latent reasoning methods.

tent input along a smooth, input-compatible latent trajectory, progressively correcting the initial soft projection toward full latent semantics, even beyond the convex hull of token embeddings. This facilitates semantically faithful latent reasoning without parameter updates, achieving both effectiveness and efficiency. **ii) a dynamic switch policy**. To enhance reasoning flexibility, DyLaR further incorporates a dynamic switching policy that allows the model to seamlessly alternate between explicit and latent reasoning modes based on its uncertainty. When the model exhibits high confidence, DyLaR favors explicit reasoning; when faced with abstract, ambiguous, or divergent thinking, it invokes latent reasoning, enabling a flexible and adaptive reasoning paradigm. A comprehensive comparison between DyLaR and existing latent reasoning methods is presented in Table 1.

We highlight that DyLaR makes the following contributions. First, the Semantic Residual Refinement (SRR) module enables DyLaR to capture and incrementally refine rich latent semantics, preserving semantic fidelity without additional training. Second, the dynamic switch policy empowers LLMs to adaptively alternate between reasoning modes as needed, improving both flexibility and robustness. Finally, extensive evaluations across both knowledge-intensive and STEM reasoning benchmarks with diverse LLMs consistently validate DyLaR's effectiveness and efficiency. It improves reasoning accuracy by up to 4.95 points while reducing token length by up to 17.52% compared to explicit CoT reasoning. Qualitative analyses further highlight DyLaR's ability to produce interpretable and insightful reasoning steps, offering a principled and practical path to unlocking LLMs' latent cognitive capabilities.

Related Works

Explicit LLM Reasoning. Recent advances have showcased the impressive reasoning capabilities of LLMs. A prominent paradigm, explicit reasoning via Chain-of-Thought (CoT) prompting (Wei et al. 2022; Kojima et al. 2022), guides LLMs to generate intermediate reasoning steps in natural language before producing the final answer, improving both accuracy and interpretability (Zelikman et al. 2024; Xiang et al. 2025; Tan et al. 2025; Ye et al. 2025). Building on this foundation, subsequent studies have shown that supervised finetuning (Yu et al. 2023; Yue et al. 2023) or reinforcement learning (Shao et al. 2024; Wang et al. 2024; Yu et al. 2024) can further enhance performance on complex reasoning tasks like mathematical problem-solving, often eliciting an "aha-moment" through a longer thinking process (Jaech et al. 2024; Guo et al. 2025). Despite its success, explicit CoT reasoning remains fundamentally limited by the discrete token space of natural language. The autoregressive, token-by-token generation process imposes inherent constraints on both expressiveness and efficiency (Qu et al. 2025), making it ill-suited for human-like cognition such as abstract reasoning, semantic ambiguity, and divergent thinking that transcend linguistic boundaries.

Latent LLM Reasoning. Latent reasoning has recently emerged as a promising direction for overcoming the limitations of discrete token space, which shifts CoT reasoning to a continuous hidden space within large language models. Early efforts such as Coconut (Hao et al. 2024) and CODI (Shen et al. 2025) directly replace discrete tokens with the model's last hidden states as the next input to construct continuous reasoning trajectories. SoftCoT (Xu et al. 2025) advances this idea by introducing a frozen assistant model to generate latent representations, along with a lightweight projection layer for alignment. While these methods improve token efficiency, they require model fine-tuning to align latent representations with the input embedding space, leading to increased training complexity and reduced output interpretability (Qu et al. 2025; Chen et al. 2025b). An alternative line of work avoids training by constructing latent inputs as convex combinations of token embeddings (Zhang et al. 2025; Yue et al. 2025), approximating continuous representations while maintaining compatibility with the input space. However, such interpolation-based strategies often suffer from semantic degradation, as they are confined to the convex hull of token embeddings and thus fail to capture unreachable latent semantics, compromising semantic fidelity. Moreover, most existing latent reasoning methods adopt fixed reasoning patterns, limiting their adaptability. Instead, this work proposes a training-free framework that preserves semantic integrity via residual refinement and enables uncertainty-aware dynamic switching between discrete and latent reasoning for better flexibility.

Method

In this section, we introduce Dynamic Latent Reasoning (DyLaR), a novel framework designed to leverage the rich semantics within latent representations in an effective, efficient, and flexible manner. As illustrated in Figure 1, DyLaR extends the explicit CoT paradigm by replacing discrete tokens with semantic residual-refined continuous embeddings,

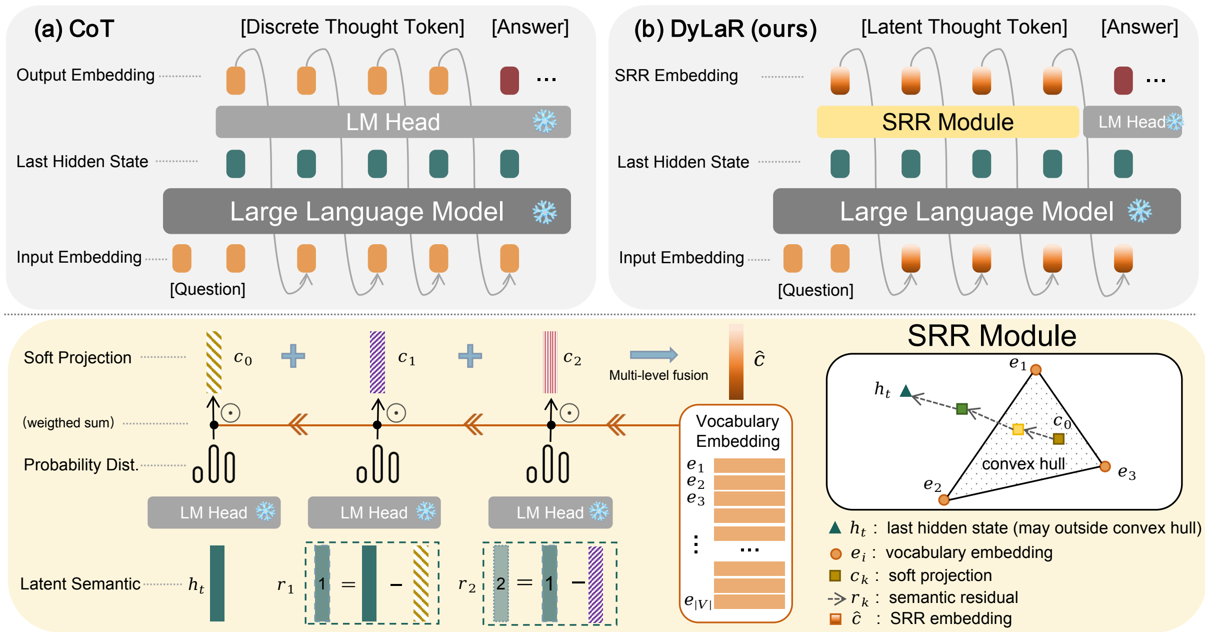


Figure 1: Framework of DyLaR. The SRR module takes the model's last hidden state and progressively refines the initial soft projection through multi-level semantic residuals, yielding a SRR embedding with enhanced semantic fidelity for the next input.

preserving semantic fidelity without finetuning. Additionally, DyLaR incorporates an entropy-aware switching policy that allows LLMs to dynamically alternate between explicit and latent reasoning modes, supporting both clear and exploratory reasoning steps on demand, beyond rigid patterns.

Preliminary: Explicit CoT Reasoning

Let V be the vocabulary of size $|V|$, $E \in \mathbb{R}^{|V|\times d}$ be the vocabulary embedding matrix. For any token index i, we denote its embedding by $e_i = E[i] \in \mathbb{R}^d$. Given input context $x = [x_1, x_2, \ldots, x_t]$ at decoding step t and its corresponding token embeddings $e = [e_{x_1}, e_{x_2}, \ldots, e_{x_t}]$, we denote the final-layer hidden states from the LLM output with $H \in \mathbb{R}^{t\times d}$:

$$H = [h_1, h_2, \ldots, h_t] = \texttt{Transformer}(e), \tag{1}$$

where `Transformer` denotes the transformer decoder. With the LM head (`Head`), the next input token x_{t+1} can be sampled discretely from the output probability distribution over the vocabulary via:

$$x_{t+1} \sim p_{t+1} = \texttt{softmax}(Wh_t). \tag{2}$$

Here, $W \in \mathbb{R}^{|V|\times d}$ is the projection matrix of `Head` (typically $W = E$ in tied embedding setups).

DyLaR: Dynamic Latent Reasoning

Semantic Residual Refinement. The discrete sampling of the next token x_{t+1} in explicit CoT paradigm commits to a single determined trajectory in language space, limiting both expressive boundary and token efficiency. In contrast, the corresponding hidden states h_t reside in a continuous latent space, encoding richer semantic information. However, it is not directly compatible with the model's input embedding space, risking semantic drift or mode collapse without finetuning. To bridge this gap, we introduce a training-free Semantic Residual Refinement (SRR) module, which first projects h_t back into the embedding space as an initial approximation and then iteratively refines it using semantic residuals derived from h_t, preserving semantic fidelity to latent representations while maintaining compatibility with input space. Specifically, SRR proceeds in three steps:

Step 1: Initial Projection. We begin by softly projecting the last hidden state h_t onto the embedding space via its output probability distribution, obtaining a coarse approximation of its latent semantics:

$$c_0 = E^\top \hat{p}_0 \in \mathbb{R}^d, \quad \hat{p}_0 = \texttt{Softmax}(Wh_t). \tag{3}$$

Although interpolating token embeddings enables richer semantics beyond discrete token expressivity, it remains confined within the convex hull of vocabulary embeddings, leaving certain latent semantic regions unreachable and causing inevitable semantic loss. To address this, SRR subsequently refines the initial soft projection using residual semantics from the hidden state, ensuring semantic fidelity.

Step 2: Residual Refinement. To uncover latent semantics not captured by the initial soft projection, we first compute the residual r_1 between the original hidden state h_t and c_0:

$$r_1 = h_t - c_0. \tag{4}$$

Semantic residuals r_k ($k \in \{1, \cdots, K\}$) are then iteratively processed by the LM head to obtain fine-grained soft projections c_k in the embedding space, from which higher-level residuals r_{k+1} are derived, enabling recursive refinement:

$$\begin{aligned} c_k &= E^\top \hat{p}_k, \quad \hat{p}_k = \text{Softmax}(Wr_k), \\ r_{k+1} &= r_k - c_k. \end{aligned} \tag{5}$$

We highlight that the soft projection of each residual uncovers previously missing semantics across different granularities, enabling progressive refinements toward a more comprehensive latent semantic representation.

Step 3: Semantic Fusion. Finally, the SRR embedding $\hat{c}$ is constructed by aggregating multi-level soft projections $c_0, \cdots, c_K$ along their latent semantic directions:

$$\hat{c} = c_0 + \sum_{k=1}^{K} \frac{\beta_k c_k}{1 + \exp\left(-\lambda \cdot \|c_k\|\right)}. \tag{6}$$

Here, $\hat{c}$ serves as the latent input for the next step. Notably, each soft projection c_k is weighted proportionally to its semantic informativeness, approximated by its L2 norm $\|c_k\|$ (Arora et al. 2017; Ethayarajh 2019). Besides, given that early projections capture core semantics while later ones primarily refine details (Zhu and Wu 2021), we impose an exponentially decaying upper bound $\beta_k = 3e^{-k}$ to each weight, preventing semantic dilution and drift.

Through iterative semantic residual refinement, the SRR module progressively integrates previously unexpressed semantic hierarchies, yielding richer representations that can theoretically extend beyond the convex hull of vocabulary embeddings. Empirically, just 1-2 refinement iterations suffice to preserve semantic fidelity that enhances latent reasoning (see Figure 2 (c)), achieving both effectiveness and efficiency. Importantly, SRR is applied only during the thinking phase (i.e., $t \in \texttt{think}$), while the final answer is still generated via standard discrete decoding.

Dynamic Switch Policy. While the SRR module facilitates effective latent reasoning, it is not always necessary. When the reasoning step is clear and deterministic—precisely mappable to a one-hot token—explicit discrete decoding is preferable, as latent representations may introduce unnecessary complexity or noise. Conversely, when reasoning involves abstraction, ambiguity, or multiple plausible paths that discrete tokens cannot faithfully capture, latent reasoning becomes essential. Therefore, we propose a dynamic switch policy that allows LLMs to adaptively alternate between latent and discrete reasoning modes as needed.

Inspired by (Gal et al. 2016; Wang et al. 2025), we employ the entropy of initial output probability $\hat{p}_0$ as a simple yet effective signal to guide each reasoning step t:

$$H(\hat{p}_0) = -\sum_{j=1}^{|V|} \hat{p}_0[j] \, \log \hat{p}_0[j]. \tag{7}$$

Low entropy corresponds to a sharp, near one-hot distribution, indicating that the current hidden state aligns closely with a specific discrete token, and the reasoning step is effectively deterministic (Rényi 1961). In contrast, high entropy reflects a uniform distribution, implying that the latent semantics exceed the expressiveness of a single token, thus necessitating latent reasoning to support abstraction or divergent thought. Hence, given an entropy threshold τ, when $H(\hat{p}_0) < \tau$, the current reasoning step t performs explicit reasoning; otherwise $H(\hat{p}_0) \geq \tau$, it switches to latent reasoning by taking the SRR embedding as next input.

Overall, this dynamic policy avoids unnecessary complexity in confident reasoning scenarios while facilitating sufficient latent exploration under uncertainty, striking a balance of exploitation and exploration. See analyses in Figure 2 (a-b) and Figure 3 for more explanation.

Theoretical Analysis

Let $\mathcal{E} = \{e_1, \ldots, e_{|V|}\} \subset \mathbb{R}^d$ denote the set of vocabulary embeddings. A probability-based soft projection over $\mathcal{E}$ yields representations in its convex hull:

$$\mathcal{C} = \mathrm{Conv}(\mathcal{E}) = \left\{ \sum_{i=1}^{|V|} \alpha_i e_i \mid \alpha_i \geq 0, \sum_i \alpha_i = 1 \right\}. \tag{8}$$

This convexity constraint limits model expressiveness, as semantic points lying outside $\mathcal{C}$ in the latent space cannot be directly represented by a single soft projection.

In DyLaR, the SRR module produces multi-level soft projections $c_k \in \mathcal{C}$, and the SRR embedding is formed as their weighted combination, which can be formulated as:

$$\hat{c} = \sum_{k=0}^{K} \gamma_k c_k = \sum_{i=1}^{|V|} \left(\sum_{k=0}^{K} \gamma_k \alpha_i^{(k)} \right) e_i, \tag{9}$$

where $\alpha_i^{(k)}$ are the probability-derived weights for the k-th soft projection, satisfying $\alpha_i^{(k)} \geq 0$ and $\sum_{i=1}^{|V|} \alpha_i^{(k)} = 1$. We analyze the expressiveness of $\hat{c}$ under two cases:

Case 1: $\gamma_k \geq 0$, but $\sum \gamma_k \neq 1$. In this case, $\hat{c}$ lies in the conic hull of $\mathcal{E}$:

$$\hat{c} \in \mathrm{Cone}(\mathcal{E}) = \left\{ \sum_{i=1}^{|V|} \delta_i e_i \mid \delta_i \geq 0 \right\} \tag{10}$$

which includes points outside the original convex hull $\mathcal{C}$, thereby expanding the representational space.

Case 2: $\gamma_k \geq 0$ and $\sum \gamma_k = 1$. Optionally, the weights of c_k can be further normalized to sum to 1. In this normalized case, since each $c_k \in \mathcal{C}$ and convex sets are closed under convex combinations, it follows that $\hat{c} \in \mathcal{C}$, which remains within the convex set of vocabulary embeddings. Nevertheless, the expressive capacity of $\hat{c}$ can exceed that of any individual projection, as it spans a richer subregion of $\mathcal{C}$.

Proposition 1 (SRR Expands Semantic Coverage) . *Let* $\mathcal{S}_K = \left\{ \sum_{k=0}^{K} \gamma_k c_k \;\middle|\; \gamma_k \geq 0,\; c_k \in \mathcal{C} \right\}$ *denote the set of SRR embeddings. Then:*

1. *For any* $K \geq 1$*, we have* $\mathcal{S}_K \supseteq \mathcal{C}$*;*
2. *Under mild assumptions on weights* γ_k*, we have*
$$\lim_{K \to \infty} \mathcal{S}_K = \mathit{Cone}(\mathcal{E}) \subseteq \mathbb{R}^d,$$
representing a broader semantic region of $\mathbb{R}^d$*.*

Theoretical analysis confirms that the SRR module systematically expands the representational space beyond single-pass discrete or soft projections, providing a principled mechanism for enhanced semantic expressiveness.

Complexity Analysis. DyLaR introduces only two lightweight additions to explicit CoT reasoning. Calculating entropy for *dynamic switch policy* incurs $O(|V|)$ time per step. For the latent reasoning steps, constructing the *SRR*

Method	CommonsenseQA		StrategyQA		TriviaQA		HotpotQA		AVG.	
	Acc ↑	#L ↓	Acc ↑	#L ↓	Acc ↑	#L ↓	Acc ↑	#L ↓	Acc ↑	#L ↓
Llama3.1-8B Series										
Zero-Shot CoT	73.87	504	78.13	450	37.87	458	11.29	512	50.29	481
Fine-Tuning										
+ Coconut	72.43	306	77.79	313	37.27	308	10.50	512	49.50 (↓ 0.79)	360 (↓ 25.21%)
+ SoftCoT	73.55	322	79.52	331	40.30	346	14.85	510	52.06 (↑ 1.77)	377 (↓ 21.57%)
Soft Thinking	73.65	475	78.43	409	37.92	418	12.68	512	50.67 (↑ 0.38)	454 (↓ 5.72%)
DyLaR (*ours*)	**74.76**	<u>354</u>	**80.64**	<u>356</u>	**44.60**	<u>367</u>	**16.43**	<u>510</u>	**54.11 (↑ 3.82)**	<u>397 (↓ 17.52%)</u>
Qwen2.5-7B Series										
Zero-shot CoT	78.13	306	77.42	357	28.35	498	8.90	512	48.20	418
Fine-Tuning										
+ Coconut	77.14	253	75.23	287	27.83	396	8.84	512	47.26 (↓ 0.94)	362 (↓ 13.45%)
+ SoftCoT	79.38	264	77.95	304	30.75	412	13.25	512	50.33 (↑ 2.13)	373 (↓ 10.82%)
Soft Thinking	79.12	299	78.30	342	28.66	460	10.40	512	49.12 (↑ 0.92)	403 (↓ 3.59%)
DyLaR (*ours*)	**79.81**	<u>276</u>	**78.64**	<u>313</u>	**36.14**	<u>428</u>	**18.01**	<u>507</u>	**53.15 (↑ 4.95)**	<u>381 (↓ 8.91%)</u>

Table 2: Main results on Knowledge QA benchmarks. All baselines are built on the instruction model, while Coconut and SoftCoT include additional fine-tuning. Accuracy on TriviaQA and HotpotQA is assessed using the exact match (EM) score, based on the top-3 retrieved documents for each query. #L denotes the count of generated tokens during reasoning. **Bold** highlights the best overall performance, and <u>underline</u> marks the best among training-free baselines.

embedding requires K LM head forward passes and dense projections, resulting in a total cost of $O(K \cdot |V| \cdot d)$. This overhead remains minimal relative to a full model forward pass, with average inference time increasing by only $\sim 5\%$ over explicit CoT reasoning, remaining on par with other latent reasoning baselines (see Appendix D for details).

Experiments

Experimental Settings

Baselines. We select two widely used open-source LLMs: the Qwen2.5-7B series (Qwen et al. 2025) and the Llama3.1-8B series (Dubey et al. 2024). To ensure a comprehensive comparison, four baseline methods are considered: an explicit CoT reasoning method, zero-shot CoT (Wei et al. 2022); two training-based latent reasoning methods, Coconut (Hao et al. 2024) and SoftCoT (Xu et al. 2025), which finetune either the full model or adapter module to align spaces; and a training-free latent reasoning method, Soft Thinking (Zhang et al. 2025), which replaces discrete tokens with probability-weighted embeddings during inference.

Evaluation Metrics and Benchmarks. We evaluate DyLaR on eight benchmarks covering both knowledge- and reasoning-intensive tasks: (1) Knowledge QA benchmarks: CommonsenseQA (Talmor et al. 2019), StrategyQA (Geva et al. 2021), TriviaQA (Joshi et al. 2017), and HotpotQA (Yang et al. 2018), spanning close-domain, open-domain, and multi-hop settings; (2) STEM reasoning benchmarks: GSM8K (Cobbe et al. 2021), MATH500 (Hendrycks et al. 2021b), MMLU-ST (Hendrycks et al. 2021a), and ARC-C (Clark et al. 2018), covering science, technology, engineering, or mathematics domains. The evaluation focuses on both effectiveness and efficiency, using accuracy (Acc) to measure effectiveness and the length of generated tokens (#L) in correct solutions to assess efficiency.

Implementation Details. For the SRR module, we set the number of refinement iterations to $K = 2$, which is empirically sufficient (see Figure 2 (c)). Following (Zhu and Wu 2021), we use $\lambda = 2$ and $\beta_k = 3e^{-k}$ in Eq. 6, as most semantic information is captured in the first pass, with diminishing returns thereafter. For the dynamic switch policy, the entropy threshold is empirically set to $\tau = 0.1$ (Figure 2 (d)). Notably, the SRR embeddings are constructed by reusing the model's existing embedding matrix. Thus, DyLaR can be plugged into any LLM's CoT pipeline with minimal effort. Additional details are provided in Appendix B.

Main Results

We present the quantitative evaluation results of DyLaR and other baseline methods on Knowledge and STEM benchmarks in Table 2 and Table 3, respectively. Experimental results demonstrate that DyLaR improves both reasoning performance and token efficiency across diverse benchmarks.

Improved Reasoning Accuracy. DyLaR consistently enhances reasoning performance across all evaluated benchmarks and models, demonstrating strong effectiveness and generalization. On Knowledge QA tasks, it boosts Qwen2.5-7B's average accuracy by 4.95% points, with substantial gains of 7.79% and 9.11% points on the challenging TriviaQA and HotpotQA datasets, respectively. For STEM benchmarks, DyLaR achieves consistent improvements of 2.08% and 1.73% points in average accuracy on LLaMA3.1-8B and Qwen2.5-7B, respectively. We attribute the larger gains on knowledge-intensive datasets to their reliance on information retrieval, integration and abstraction, which better align with DyLaR's latent reasoning, unlike the step-by-

Method	GSM8K Acc ↑	GSM8K #L ↓	MATH500 Acc ↑	MATH500 #L ↓	MMLU-ST Acc ↑	MMLU-ST #L ↓	ARC-C Acc ↑	ARC-C #L ↓	AVG. Acc ↑	AVG. #L ↓
Llama3.1-8B Series										
Zero-Shot CoT	77.48	629	45.60	1024	60.69	818	74.77	498	64.64	742
Fine-Tuning										
+ Coconut	76.29	420	44.15	1024	58.41	601	73.76	367	63.15 (↓ 1.48)	603 (↓ 18.76%)
+ SoftCoT	77.63	496	47.00	1024	61.37	643	76.54	406	65.64 (↑ 1.00)	642 (↓ 13.47%)
Soft Thinking	78.47	578	44.80	1024	60.99	770	75.51	487	64.94 (↑ 0.31)	715 (↓ 3.70%)
DyLaR (*ours*)	**79.38**	<u>537</u>	**48.12**	<u>1024</u>	**61.97**	<u>664</u>	**77.39**	<u>415</u>	**66.72 (↑ 2.08)**	<u>660 (↓ 11.08%)</u>
Qwen2.5-7B Series										
Zero-shot CoT	88.23	490	68.00	1018	73.15	691	80.83	389	77.55	647
Fine-Tuning										
+ Coconut	87.28	364	67.12	1020	71.81	483	79.67	296	76.47 (↓ 1.08)	541 (↓ 16.42%)
+ SoftCoT	89.16	407	69.00	986	73.86	515	81.83	318	78.46 (↑ 0.91)	557 (↓ 13.98%)
Soft Thinking	88.40	479	67.50	<u>832</u>	73.51	664	81.40	386	77.70 (↑ 0.15)	590 (↓ 8.77%)
DyLaR (*ours*)	**89.46**	<u>432</u>	**70.40**	871	**74.41**	<u>576</u>	**82.85**	<u>330</u>	**79.28 (↑ 1.73)**	<u>552 (↓ 14.64%)</u>

Table 3: Main results on STEM reasoning benchmarks. All baselines are based on the instruction model, while Coconut and SoftCoT include additional fine-tuning. #L denotes the count of generated tokens during reasoning. **Bold** highlights the best overall performance, and <u>underline</u> marks the best among training-free baselines.

Methods (Qwen2.5-7b)	MATH500 Acc ↑	MATH500 #L ↓	TriviaQA Acc ↑	TriviaQA #L ↓
Zero-shot CoT	68.00	1018	28.35	498
DyLaR	**70.40**	**871**	**36.14**	**428**
- w/o SRR	68.00	1018	28.35	498
w/ hidden state	0.00	1024	0.00	512
w/ interpolation	68.34	1004	29.63	461
- w/o dynamic switch	69.21	1003	32.08	473
w/ fixed latent step	68.80	1015	31.61	454

Table 4: Ablation study of DyLaR.

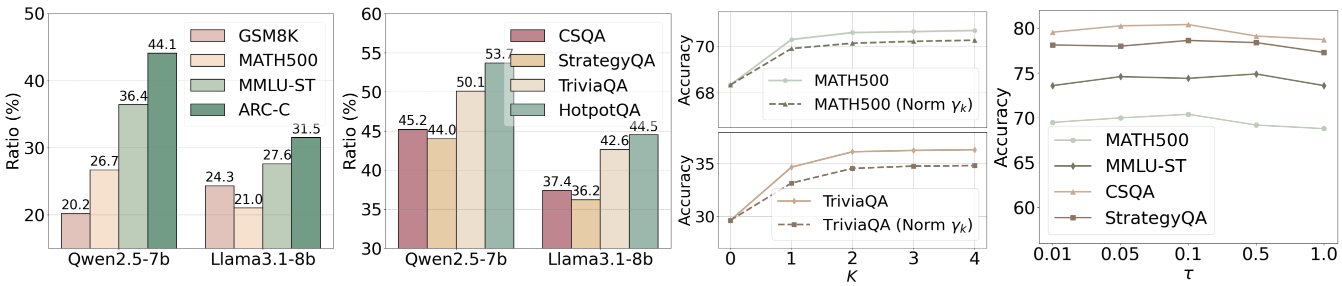
step logic required in STEM reasoning tasks. Furthermore, DyLaR outperforms both training-free and finetuning-based latent reasoning methods. Notably, Coconut even falls short of explicit CoT due to catastrophic forgetting, which aligns with findings from prior studies (Xu et al. 2025; Elita A. et al. 2025). Overall, DyLaR delivers robust accuracy gains while preserving the generalization capabilities of LLMs.

Token Efficiency. Another key advantage of DyLaR is its improved token efficiency, reducing generation length by 14.64% on Qwen and 11.08% on LLaMA for STEM reasoning benchmarks compared to standard CoT, with similar gains on knowledge tasks. This efficiency lowers inference costs and reflects more concise reasoning, crucial for real-world applications. While finetuning-based methods like Coconut achieve greater token reduction, they may sacrifice accuracy and require costly retraining. In contrast, DyLaR boosts both efficiency and accuracy without additional training. This dual benefit stems from our SRR embeddings, which maintain richer semantics at each step, enabling the model to represent abstract thoughts and explore multiple plausible paths simultaneously, thereby reaching conclusions with fewer steps. Overall, DyLaR breaks the trade-off between performance and efficiency in LLM reasoning, providing a concise yet powerful framework broadly applicable across tasks and model architectures.

Analyses of DyLaR

Impact of Semantic Residual Refinement. To assess the contribution of the SRR module, we compare two alternative strategies for incorporating latent representations, as shown in Table 4: (1) *hidden state*, which directly feed the final layer's hidden states as the next input; and (2) *interpolation*, which computes a probability-weighted sum over token embeddings. The direct leverage of hidden states fails completely with zero accuracy due to the mismatch between the latent and embedding spaces. While interpolation yields slight gains over the baseline, its effectiveness is limited, likely due to semantic loss in the approximation. In contrast, DyLaR with SRR achieves significantly superior performance, highlighting SRR's effectiveness in capturing richer semantics and its importance for latent reasoning.

Impact of Dynamic Switch Policy. We evaluate the effectiveness of the dynamic switch policy by comparing DyLaR with and without it, as shown in Table 4. In the absence of the switch, the model resorts to full latent reasoning, leading to longer outputs and decreased accuracy due to unnecessary divergent thinking and noise when the reasoning step is clear. We further examine fixed-step latent reasoning, as employed in Coconut and SoftCoT. Its inferiority indicates that rigid control hinders the exploration potential of LLMs. Differently, DyLaR's dynamic switch policy adaptively avoids unnecessary exploration along determined paths while supporting sufficient exploration when ambiguity exists. This flexibility significantly reduces generation length and achieves better performance, striking a favorable



(a) Ratio of latent reasoning steps (b) Ratio of latent reasoning steps (c) Sensitivity to SRR rounds K (d) Sensitivity to entropy threshold

Figure 2: Analysis of DyLaR. (a-b) Ratio of latent reasoning steps across all inference steps. (c) Effect of SRR refinement iterations K (with vs. without normalizing γ_k) (d) Sensitivity to entropy threshold τ for dynamic switching policy).

Question: Which planet weighs more than all the other planets of our solar system combined?

➢ ***CoT Thinking (explicit, step-by-step, logic-driven):***
To determine which planet weighs more than all the other planets of our solar system combined, I'll first consider the mass of each planet. The four terrestrial planets (Mercury, Venus, Earth, and Mars) are significantly smaller in mass compared to the four gas giants (Jupiter, Saturn, Uranus, and Neptune). \n\nJupiter has the largest planet in our solar system and is significantly more massive than the other planets. Its mass is approximately 2.5 times that of all the other planets combined. Therefore, Jupiter is the planet that weighs more than all the other planets of our solar system combined.

➢ ***DyLaR Reasoning (abstract, holistic, insight-driven):***
"Of vast majority of the system's mass is in the Sun, with most of the remaining mass contained in Jupiter" suggests that Sun and Jupiter make the two most massive bodies in the Solar System. Given the question asks for a planet that weighs more than all the other planets combined, we the Sun is not a planet, Jupiter must be the answer.

Figure 3: A comparison between explicit CoT thinking and DyLaR reasoning on a knowledge QA problem. We visualize the token with the highest probability at each step of DyLaR for readability and interpretability.

balance between exploitation and exploration.

Ratio of Latent Reasoning Steps. Figure 2(a-b) shows that DyLaR engages a higher proportion of latent reasoning in knowledge QA benchmarks than in STEM tasks. This indicates that knowledge QA often demands deeper abstraction and insight-driven inference, thereby triggering more latent reasoning. In contrast, STEM tasks generally follow more explicit and clearer logical pathways, unless ambiguity or multiple plausible paths arise. For instance, the more complex ARC-C task exhibits a higher latent reasoning ratio than the simpler GSM8K task. These findings highlight DyLaR's flexibility in adaptively switching between latent and explicit reasoning, enabling strong generalization across diverse reasoning scenarios and making it particularly effective at leveraging contextual information.

Sensitivity of Hyper-parameter. We analyze DyLaR's sensitivity to key hyperparameters using the Qwen2.5-7b model, as shown in Figure 2 (c-d). In (c), we vary the number of refinement iterations $K = \{0, 1, 2, 3, 4\}$ under both normalized and unnormalized weights γ_k (Eq. 9). Performance improves with more refinements, highlighting the benefit of incorporating semantic residuals. Typically, $1 \sim 2$ iterations are sufficient. Notably, normalized γ_k yields slightly lower but more stable results, likely due to limited expressivity within the embedding convex hull. In (d), we explore the entropy threshold τ ranging from 0.01 to 1.0, where a modest value, i.e., 0.1, performs best, providing a more precise signal for switching between reasoning modes.

Interpretability Analysis. Figure 3 illustrates a comparison between explicit CoT reasoning and DyLaR, with the highest-probability token visualized at each step. It can be seen that DyLaR exhibits high readability and interpretability. While both approaches arrive at the correct answer "Jupiter", DyLaR follows a significantly more concise reasoning path, characterized by holistic understanding, abstract thinking, and intuitive leaps. This highlights DyLaR's ability to reason effectively and efficiently with latent semantics, facilitating insightful thinking without rigid steps.

Conclusion

In this paper, we propose Dynamic Latent Reasoning (DyLaR), a novel framework introducing two key innovations: i) a Semantic Residual Refinement mechanism that incrementally integrates semantic residuals from prior hidden states, enabling semantically faithful latent inputs without finetuning, and ii) a dynamic switching policy that allows the model to adaptively alternates between discrete and latent reasoning. Notably, DyLaR achieves effective and flexible latent reasoning without additional training, making it easily integrable into any LLM's CoT pipeline with minimal effort. Extensive experiments across diverse scenarios demonstrate DyLaR's superiority in both effectiveness and efficiency. Qualitative analyses further underscore its interpretability and flexibility, offering new insights for latent reasoning.

Acknowledgements

Funded by National Science and Technology Major Project (No. 2022ZD0114903) and the Natural Science Fundation of China (NSFC. No. 62476149).

References

Arora, S.; Liang, Y.; Ma, T.; Lourie, N.; and Berant, J. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *International conference on learning representations*.

Chen, Q.; Qin, L.; Liu, J.; Peng, D.; Guan, J.; Wang, P.; Hu, M.; Zhou, Y.; Gao, T.; and Che, W. 2025a. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*.

Chen, X.; Zhao, A.; Xia, H.; Lu, X.; Wang, H.; Chen, Y.; Zhang, W.; Wang, J.; Li, W.; and Shen, X. 2025b. Reasoning Beyond Language: A Comprehensive Survey on Latent Chain-of-Thought Reasoning. *arXiv preprint arXiv:2505.16782*.

Clark, P.; Cowhey, I.; Etzioni, O.; Khot, T.; Sabharwal, A.; Schoenick, C.; and Tafjord, O. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Deng, Y.; Choi, Y.; and Shieber, S. 2024. From explicit cot to implicit cot: Learning to internalize cot step by step. *arXiv preprint arXiv:2405.14838*.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv e-prints*, arXiv–2407.

Elita A., L.; Chirag, A.; Himabindu, L.; Weld, D. S.; and Zettlemoyer, L. 2025. On the Impact of Fine-Tuning on Chain-of-Thought Reasoning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, 11679–11698. Association for Computational Linguistics.

Ethayarajh, K. 2019. How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 55–65.

Gal, Y.; Ghahramani, Z.; Weld, D. S.; and Zettlemoyer, L. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, 1050–1059. PMLR.

Geva, M.; Khashabi, D.; Segal, E.; Khot, T.; Roth, D.; and Berant, J. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9: 346–361.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Hao, S.; Sukhbaatar, S.; Su, D.; Li, X.; Hu, Z.; Weston, J.; and Tian, Y. 2024. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021a. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*.

Hendrycks, D.; Burns, C.; Kadavath, S.; Arora, A.; Basart, S.; Tang, E.; Song, D.; and Steinhardt, J. 2021b. Measuring Mathematical Problem Solving With the MATH Dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Jaech, A.; Kalai, A.; Lerer, A.; Richardson, A.; El-Kishky, A.; Low, A.; Helyar, A.; Madry, A.; Beutel, A.; Carney, A.; et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Joshi, M.; Choi, E.; Weld, D. S.; and Zettlemoyer, L. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1601–1611.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213.

Liu, Y.; Wu, J.; He, Y.; Gao, H.; Chen, H.; Bi, B.; Gong, R.; Zhang, J.; Huang, Z.; and Hooi, B. 2025. Efficient inference for large reasoning models: A survey. *arXiv preprint arXiv:2503.23077*.

Pinker, S.; Longuet-Higgins, C.; Su, D.; and Sukhbaatar, S. 1994. The language instinct: how the mind creates language. *Nature*, 368(6469): 360–360.

Qu, X.; Li, Y.; Su, Z.; Sun, W.; Yan, J.; Liu, D.; Cui, G.; Liu, D.; Liang, S.; He, J.; et al. 2025. A survey of efficient reasoning for large reasoning models: Language, multimodality, and beyond. *arXiv preprint arXiv:2503.21614*.

Qwen; :; Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; Lin, H.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Lin, J.; Dang, K.; Lu, K.; Bao, K.; Yang, K.; Yu, L.; Li, M.; Xue, M.; Zhang, P.; Zhu, Q.; Men, R.; Lin, R.; Li, T.; Tang, T.; Xia, T.; Ren, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Qiu, Z. 2025. Qwen2.5 Technical Report. arXiv:2412.15115.

Rényi, A. 1961. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics*, volume 4, 547–562. University of California Press.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Shen, Z.; Yan, H.; Zhang, L.; Hu, Z.; Du, Y.; and He, Y. 2025. Codi: Compressing chain-of-thought into continuous space via self-distillation. *arXiv preprint arXiv:2502.21074*.

Tack, J.; Lanchantin, J.; Yu, J.; Cohen, A.; Kulikov, I.; Lan, J.; Hao, S.; Tian, Y.; Weston, J.; and Li, X. 2025. Llm pretraining with continuous concepts. *arXiv preprint arXiv:2502.08524*.

Talmor, A.; Herzig, J.; Lourie, N.; and Berant, J. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4149–4158.

Tan, W.; Li, B.; Jin, C.; Huang, W.; Wang, X.; and Song, R. 2025. Think then react: Towards unconstrained action-to-reaction motion generation. In *The Thirteenth International Conference on Learning Representations*.

Wang, P.; Li, L.; Shao, Z.; Xu, R.; Dai, D.; Li, Y.; Chen, D.; Wu, Y.; and Sui, Z. 2024. Math-Shepherd: Verify and Reinforce LLMs Step-by-step without Human Annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9426–9439.

Wang, S.; Yu, L.; Gao, C.; Zheng, C.; Liu, S.; Lu, R.; Dang, K.; Chen, X.; Yang, J.; Zhang, Z.; et al. 2025. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Wittgenstein, L. 2023. Tractatus logico-philosophicus.

Xiang, V.; Snell, C.; Gandhi, K.; Albalak, A.; Singh, A.; Blagden, C.; Phung, D.; Rafailov, R.; Lile, N.; Mahan, D.; et al. 2025. Towards system 2 reasoning in llms: Learning how to think with meta chain-of-thought. *arXiv preprint arXiv:2501.04682*.

Xu, Y.; Guo, X.; Zeng, Z.; and Miao, C. 2025. Softcot: Soft chain-of-thought for efficient reasoning with llms. *arXiv preprint arXiv:2502.12134*.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W.; Salakhutdinov, R.; and Manning, C. D. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2369–2380.

Ye, Y.; Huang, Z.; Xiao, Y.; Chern, E.; Xia, S.; and Liu, P. 2025. Limo: Less is more for reasoning. *arXiv preprint arXiv:2502.03387*.

Yu, F.; Jiang, L.; Kang, H.; Hao, S.; and Qin, L. 2024. Flow of reasoning: Efficient training of llm policy with divergent thinking. *arXiv preprint arXiv:2406.05673*, 1(2): 6.

Yu, L.; Jiang, W.; Shi, H.; Yu, J.; Liu, Z.; Zhang, Y.; Kwok, J. T.; Li, Z.; Weller, A.; and Liu, W. 2023. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*.

Yue, X.; Qu, X.; Zhang, G.; Fu, Y.; Huang, W.; Sun, H.; Su, Y.; and Chen, W. 2023. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint arXiv:2309.05653*.

Yue, Z.; Jin, B.; Zeng, H.; Zhuang, H.; Qin, Z.; Yoon, J.; Shang, L.; Han, J.; and Wang, D. 2025. Hybrid Latent Reasoning via Reinforcement Learning. *arXiv preprint arXiv:2505.18454*.

Zelikman, E.; Harik, G. R.; Shao, Y.; Jayasiri, V.; Haber, N.; and Goodman, N. 2024. Quiet-STaR: Language Models Can Teach Themselves to Think Before Speaking. In *First Conference on Language Modeling*.

Zhang, Z.; He, X.; Yan, W.; Shen, A.; Zhao, C.; Wang, S.; Shen, Y.; and Wang, X. E. 2025. Soft thinking: Unlocking the reasoning potential of llms in continuous concept space. *arXiv preprint arXiv:2505.15778*.

Zhu, K.; and Wu, J. 2021. Residual attention: A simple but effective method for multi-label recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, 184–193.