Multi-Path Collaborative Reasoning via Reinforcement Learning

Jindi Lv[1,2]   Yuhao Zhou[1]   Zheng Zhu[2]   Xiaofeng Wang[2,3]   Guan Huang[2]   Jiancheng Lv[1]

[1]Sichuan University   [2]GigaAI   [3]Tsinghua University

Project page: https://multi-path-collaborative-reasoning.github.io/
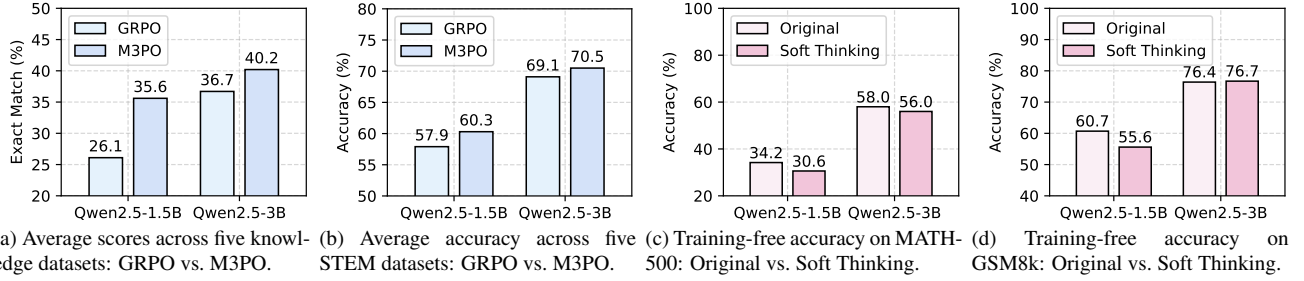
Figure 1. Comparative analysis of reasoning paradigms. Figures (a) and (b) present results on knowledge and STEM benchmarks, while Figures (c) and (d) evaluate Soft Thinking without training on MATH-500 and GSM8k. Soft Thinking shows degraded performance, indicating extraneous inference noise. In contrast, M3PO consistently improves performance on knowledge- and reasoning-intensive benchmarks (with up to a **9.5%** average gain on five knowledge datasets), demonstrating the superiority of multi-path reasoning paradigm.

Abstract

Chain-of-Thought (CoT) reasoning has significantly advanced the problem-solving capabilities of Large Language Models (LLMs), yet conventional CoT often exhibits internal determinism during decoding, limiting exploration of plausible alternatives. Recent methods attempt to address this by generating soft abstract tokens to enable reasoning in a continuous semantic space. However, we find that such approaches remain constrained by the greedy nature of autoregressive decoding, which fundamentally isolates the model from alternative reasoning possibilities. In this work, we propose Multi-Path Perception Policy Optimization (M3PO), a novel reinforcement learning framework that explicitly injects collective insights into the reasoning process. M3PO leverages parallel policy rollouts as naturally diverse reasoning sources and integrates cross-path interactions into policy updates through a lightweight collaborative mechanism. This design allows each trajectory to refine its reasoning with peer feedback, thereby cultivating more reliable multi-step reasoning patterns. Empirical results show that M3PO achieves state-of-the-art performance on both knowledge- and reasoning-intensive benchmarks. Models trained with M3PO maintain interpretability and inference efficiency, underscoring the promise of multi-path collaborative learning for robust reasoning.

1. Introduction

The advent of Chain-of-Thought (CoT) [14, 21, 51] has markedly improved the reasoning capabilities of Large Language Models (LLMs) [29, 43, 50] on complex tasks. By generating explicit intermediate steps in natural language, CoT decomposes problems and guides models step-by-step toward the answer [4]. However, the discrete token decoding underlying standard CoT induces internal determinism, which limits the exploration of plausible alternatives [57].

Recent efforts seek to transcend discrete token decoding by introducing continuity into the reasoning process. Latent reasoning [15, 39, 41] propagates hidden states instead of tokens to enable differentiable reasoning, but such representations often deviate from the pretrained embedding space, requiring costly alignment and limiting compatibility [38]. To address this, Soft Thinking [52, 61] performs soft aggregation within the input embedding space, preserving model compatibility while simulating a smooth semantic flow.

Nevertheless, we reveal that Soft Thinking inherently lacks the ability to represent diverse semantic trajectories in parallel [52]. As shown in Figure 2, during consecutive decoding, each subsequent distribution consistently follows the dominant semantic token from the previous step, resulting in a coherent but exclusive progression along a single trajectory. Although soft aggregation improves information capacity, it primarily reinforces the dominant path while introducing extraneous noise over time, as evidenced in Fig-

1

ures 1c and 1d.

This phenomenon stems from the greedy nature of LLMs in autoregressive generation, where hidden states evolve at each step toward the most confident semantic direction [52]. Such locally optimal decisions are amplified over long reasoning chains, resulting in a cumulative bias toward dominant paths and exclusion of alternatives. Thus, when a reasoning trajectory begins with a flawed premise, the lack of timely corrective feedback allows erroneous logic to propagate unimpeded through subsequent steps [53].

In this work, we propose **M3PO**, a novel **M**ulti-**P**ath **P**erception **P**olicy **O**ptimization framework for reasoning robustness. M3PO leverages the inherent independence among policy rollouts to create natural multi-path sources. Building on this perspective, we design a collaborative mechanism that explicitly incorporates cross-path interactions into the policy update. Through reward-guided policy optimization, M3PO progressively internalizes more consistent and reliable reasoning patterns.

This paradigm promotes robust policy optimization by allowing trajectories to benefit from cross-path insights. Specifically, the parameter-free gate facilitates step-level information exchange, providing trajectories with opportunities to gently refine their reasoning chain and reduce the persistence of local biases. In particular, by maintaining and refining multiple hypotheses in parallel, M3PO aligns with human-like cognition [2, 9, 10], striking an adaptive balance between exploration and exploitation.

Empirically, as shown in Figures 1a and 1b, M3PO achieves superior performance over the standard CoT approach across both knowledge- and reasoning-intensive benchmarks. The advantage is particularly pronounced on knowledge-oriented tasks, where it delivers an average improvement of 9.5%. These results validate the distinct advantage of our proposed multi-path reasoning paradigm.

We highlight the main contributions of this paper below:

- We hypothesize parallel rollouts in reinforcement learning as a natural source of reasoning diversity, eliminating the need for curated auxiliary datasets.
- We design a multi-path collaborative learning mechanism that integrates cross-path interactions into policy updates, enabling implicit exploration-exploitation balance.
- Our framework achieves state-of-the-art reasoning performance (up to +9.5% accuracy over conventional CoT) on standard benchmarks without additional parameters.

2. Related Works

2.1. Continuous Space Reasoning

Recent works have explored continuous space reasoning as an alternative to explicit CoT generation [12, 31, 45, 46, 49]. Several studies demonstrate that Transformers can perform multi-hop reasoning implicitly through their inter-

Figure 2. An illustrative case of probability distribution in vanilla Soft Thinking method. Each step aligns closely with the top-1 token from the prior step, leading to rapid concentration along a single semantic path.

nal hidden states, without generating intermediate textual steps [36, 55]. Deng et al. [7, 8] further exploit this idea by using hidden dynamics directly for reasoning, bypassing symbolic output altogether. Geiping et al. [11] extend this paradigm with a depth-recurrent architecture that reuses Transformer layers during inference, increasing computational depth per token and enabling internal processing.

Another line focuses on explicit continuous reasoning through latent representations [7, 41]. COCONUT [15] enables reasoning in the hidden state space by replacing discrete tokens with continuous vectors, eliminating the need for human-readable chains. CODI [39] learns to align recurrent state updates via self-distillation, forming structured reasoning trajectories in the latent space. Although conceptual advances, they often struggle to scale to larger models and maintain performance on complex reasoning tasks.

More recently, training-free approaches such as Soft Thinking [13, 52, 61] and Mixture-of-Inputs (MoI) [62] have emerged, leveraging output distributions to align hidden states with input embeddings, facilitating iterative reasoning in continuous space. Building on this, HRPO [59] introduces a hybrid policy that fuses discrete token embeddings with soft embeddings through a gating mechanism, enabling richer internal state evolution during reasoning.

Despite these innovations, Soft Thinking methods mainly increase representational capacity without enabling meaningful exploration across reasoning paths. We propose a multi-path collaborative policy optimization framework that leverages parallel rollouts during training to encourage interaction among trajectories. By learning from alternative reasoning paths, the model breaks self-reinforcing logical loops and develops more robust inference capabilities.
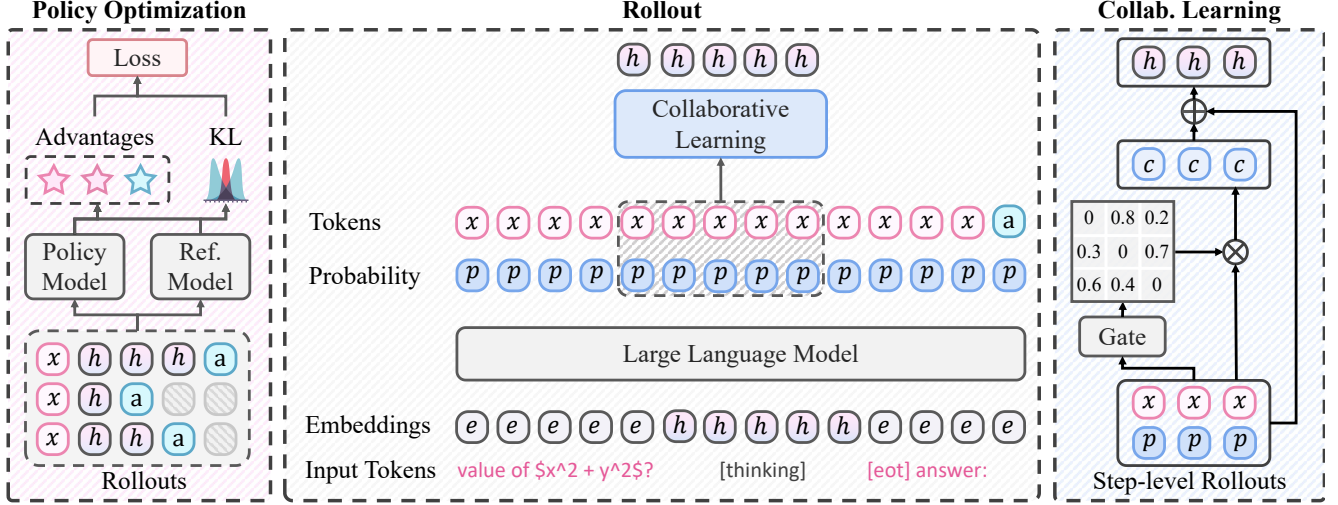
Figure 3. Overview of the M3PO framework. For a given question, multiple trajectories are first generated in parallel (Rollout). During thinking, these paths dynamically interact via a novel collaborative mechanism to refine intermediate reasoning steps (Collaborative Learning). The optimized trajectories are then used for policy updates via group-relative advantage estimation, closing the loop between multi-path exploration and policy learning (Policy Optimization).

2.2. Reinforcement Learning

Reinforcement learning (RL) has emerged as a powerful framework for enhancing language models through feedback-driven decision making. In natural language generation, RL enables models to learn from implicit human preferences by maximizing cumulative rewards [30]. This is typically achieved with policy gradient methods such as REINFORCE [40], which update the policy using sampled responses and their rewards. Actor-critic approaches like A2C [28] reduce variance with learned baselines, while PPO [34] improves stability through a clipped objective that limits policy changes and prevents performance collapse.

An alternative line of work uses preference-based optimization, where models are fine-tuned on pairwise response comparisons [3, 35, 48, 60]. Methods like DPO [33] bypass explicit reward modeling by directly optimizing for preferred outputs, often eliminating the need for a reference model. Yet despite their efficiency, offline approaches generally lag behind online RL in complex reasoning settings. Recent online methods such as RLOO [1], GRPO [37], and REINFORCE++ [20] address variance and scalability by computing baselines from multiple responses in the same batch. Leveraging group-level comparisons, they achieve more stable and memory-efficient training, making them well suited for long-horizon reasoning tasks.

Building on these advances, we propose an online RL framework that incorporates cross-path interactions during policy updates. By evaluating reasoning trajectories against each other within groups, the model learns to identify and correct flawed reasoning. This strengthens policy robustness by internalizing reliable reasoning patterns, yielding more accurate and stable text generation.

3. Method

3.1. Overall Framework

The M3PO framework establishes a closed-loop optimization process that seamlessly integrates multi-path reasoning with a step-level collaborative mechanism. As illustrated in Figure 3, the framework comprises three core components: a parallel rollout pipeline, a collaboration learning mechanism, and a policy optimization module.

Formally, let V be the vocabulary of size $|V|$ and d be the embedding dimension. Given an input question x and a token embedding matrix $E \in \mathbb{R}^{|V| \times d}$, the policy π_θ generates N parallel rollouts. Each trajectory τ_i is represented as a sequence of embeddings:

$$\tau_i = [E(x), \bar{h}_i^{(1)}, \ldots, \bar{h}_i^{(L)}, E(a_i)], \quad i = 1, 2, \ldots, N \quad (1)$$

where a_i denotes the final answer token, and $\bar{h}_i^{(l)}$ represents the hybrid thinking embedding at step l, produced by the collaborative mechanism.

This mechanism is activated only during the thinking phase. In contrast, the answer phase remains isolated to preserve the independence of final decisions. At each step l, the model outputs token distributions $p_i^{(l)} = \pi_\theta(\cdot \mid \tau_i^{(l)})$ for $i = 1, \ldots, N$. The set $\{p_1^{(l)}, \ldots, p_N^{(l)}\}$ is fed into a parameter-free gating function $\mathcal{G}$, which computes cooperation weights based on distributional consistency and value estimates. These weights are used to aggregate cross-path information and produce the hybrid thinking embedding $\bar{h}_i^{(l)}$, effectively fusing global insights into local trajectory.

3

Finally, the policy optimization module uses the collaboratively refined rollouts $\{\tau_1, \ldots, \tau_N\}$ to compute advantages and update π_θ using a group-relative objective. This closed-loop design leverages reward signals to guide cross-path collaboration, ensuring that information exchange amplifies high-reward reasoning patterns while suppressing erroneous trajectories, thereby improving thinking quality and learning stability.

3.2. Multi-Path Collaborative Mechanism

The multi-path collaborative mechanism addresses the challenge of self-reinforcing reasoning loops by strategically introducing external perspectives during the thinking phase. This framework facilitates controlled information exchange across parallel trajectories, enabling each path to potentially break away from its own reasoning constraints while maintaining training stability.

At each thinking step l, for the i-th trajectory with its sampled token embedding $e_i^{(l)}$ and the cross-path contextual embedding $c_i^{(l)}$, the mechanism constructs a hybrid thinking embedding through convex combination:

$$\overline{h}_i^{(l)} = (1 - \lambda)\, e_i^{(l)} + \lambda\, c_i^{(l)}, \quad \lambda \in [0, 1] \tag{2}$$

The hyperparameter λ balances two crucial aspects, with the $(1 - \lambda)$ term preserving the trajectory's intrinsic reasoning direction and the λ term incorporating diversified information from alternative paths. This design allows each trajectory to maintain its distinctive characteristics while gaining opportunities to refer other reasoning patterns.

The cross-path contextual embedding $c_i^{(l)}$ is generated through a distribution-similarity fusion mechanism. The process begins with computing a similarity matrix from output distributions:

$$S_{ij}^{(l)} = \frac{p_i^{(l)} \cdot p_j^{(l)}}{\|p_i^{(l)}\|\|p_j^{(l)}\|}, \quad i, j = 1, \ldots, N \tag{3}$$

Diagonal elements are masked to prevent self-reinforcement:

$$\tilde{S}_{ij}^{(l)} = \begin{cases} 0 & i = j, \\ S_{ij}^{(l)} & \text{otherwise.} \end{cases} \tag{4}$$

Cooperative weights are obtained through temperature-scaled normalization:

$$A_{ij}^{(l)} = \begin{cases} 0 & i = j, \\ \dfrac{\exp(\tilde{S}_{ij}^{(l)}/T)}{\sum_{k \neq i} \exp(\tilde{S}_{ik}^{(l)}/T)} & \text{otherwise,} \end{cases} \tag{5}$$

yielding the final contextual embedding:

$$c_i^{(l)} = \sum_{j=1}^{N} A_{ij}^{(l)} e_j^{(l)} \in R^d. \tag{6}$$

This fusion mechanism ensures moderated information exchange by fostering strong interactions between distributionally aligned paths to smooth the evolution of reasoning states, while permitting controlled diversity through weaker engagements with divergent trajectories to prevent disruptive perturbations.

The resulting hybrid embedding $\overline{h}_i^{(l)}$ is used as the input for the next reasoning step, thereby directly shaping the model's subsequent reasoning and token predictions through integrated individual reasoning and collective insights. Notably, the mechanism operates exclusively during thinking steps, while the final answer generation remains independent to preserve decision diversity. When a trajectory exits the thinking mode at some step, it no longer participates in subsequent cross-path collaboration. Before the softmax normalization, we set $\tilde{S}_{ir}^{(l)} = \tilde{S}_{ri}^{(l)} = 0$ for all i for the exited trajectory r, so that it neither contributes to nor receives information in later steps.

Through this local collaborative process, the mechanism provides refined reasoning trajectories for the subsequent global policy optimization, enabling the identification and reinforcement of more robust reasoning strategies.

3.3. Policy Optimization

M3PO optimizes the policy model π_θ via reinforcement learning, leveraging the intrinsic reasoning capabilities of LLMs without explicit supervision and curated auxiliary datasets. Building upon the refined thinking embeddings produced by a collaborative mechanism, we formulate the optimization objective to leverage both explicit reward signals and implicit guidance from cross-path interactions.

Formally, given an input question x and a set of N refined rollouts $\{\tau_1, \tau_2, \ldots, \tau_N\}$ generated through our collaborative mechanism, we compute trajectory-level rewards $R(\tau_i)$ based on final answer correctness. The advantage for each trajectory is estimated through group-relative standardization:

$$A(\tau_i) = \frac{R(\tau_i) - \mu}{\sigma} \tag{7}$$

where μ and σ represent the mean and standard deviation of rewards within the group of N rollouts. This advantage estimation provides a stable learning signal without requiring a separate value function.

The policy optimization objective combines advantage-weighted likelihood maximization with KL-divergence regularization:

$$\nabla_\theta J_{\text{M3PO}}(\theta) = \mathbb{E}\left[\frac{1}{N} \sum_{i=1}^{N} \left(\sum_{t=1}^{L} \nabla_\theta \log \pi_\theta(e_i^{(t)} \mid x, \overline{h}_i^{(<t)})\right.\right.$$
$$\left.\left. \cdot A(\tau_i)\right)\right] - \beta \nabla_\theta D_{\text{KL}}[\pi_\theta \| \pi_{\text{ref}}], \tag{8}$$

4

where β controls the strength of the KL constraint, and π_{ref} denotes the reference model for regularization. Crucially, the gradient updates are computed using the hybrid thinking embeddings $\bar{h}_i^{(<t)}$ as context, ensuring that the policy learns to leverage the collaborative reasoning patterns cultivated during the thinking phase.

Note that M3PO uses raw log probabilities in its policy gradient (Equation (8)) rather than likelihood ratios as in PPO [34] or GRPO [37], eliminating the need for ratio clipping under our conservative update schedule. Furthermore, each trajectory is used only once for gradient updates since the hybrid representations are intrinsically tied to θ, maintaining strict on-policy training. This makes M3PO both lightweight and compatible with other RL optimizations.

Inference. M3PO employs single-path decoding at test time, matching standard LLMs inference. This efficiency is enabled by the reinforcement learning objective, which trains the policy π_{θ} to maximize the reward via a multi-path collaboration mechanism. As the policy iteratively updates to capture high-reward trajectories, it effectively internalizes robust reasoning patterns and learns a more reliable and self-consistent way of thinking. Thus, even in the absence of explicit path interaction at test time, the optimized policy still exhibits enhanced reasoning accuracy and stability.

4. Experiments

We evaluate M3PO on two benchmark categories. The first is knowledge-intensive tasks, including open-domain and multi-hop question answering (QA), which assess factual utilization and evidence aggregation; the second is reasoning-intensive STEM problems that probe structured problem solving in scientific and mathematical settings.

4.1. Evaluation on Knowledge Benchmarks

Datasets and baselines. For knowledge-intensive reasoning, we use five widely adopted open-domain and multi-hop QA datasets: Natural Questions (NQ) [24], TriviaQA [23], HotpotQA [56], 2WikiMultiHopQA (2WikiMQA) [18], and Bamboogle [32]. Following HRPO [59], we employ the E5-base embedding model [47] to retrieve the top three Wikipedia documents as context for each prompt. We train on a joint corpus formed by merging the NQ and HotpotQA training sets, and we report exact match results on each dataset's official evaluation split. Detailed implementation settings are provided in the Appendix.

For the 1.5B and 3B model sizes [54], we evaluate M3PO against representative post-training and retrieval baselines, including Supervised Fine-Tuning (SFT), the RL methods PPO [34] and GRPO [37], Retrieval-Augmented Generation (RAG) [25], and Hybrid Reasoning Policy Optimization (HRPO) [59], an RL-based approach built on soft thinking [61]. We also report results with the larger

Table 1. Evaluation performance on QA benchmarks. This table reports exact match scores on five open-domain and multi-hop QA datasets using top-3 retrieved documents. The upper section shows RAG baselines with Qwen2.5-7B, while the lower sections show the performance of smaller Qwen models (1.5B and 3B) under different training strategies. M3PO delivers consistently strong performance across all datasets, with especially notable gains on the NQ dataset, demonstrating its effectiveness in enhancing reasoning robustness through internal refinement of reasoning processes.

	NQ	TriviaQA	HotpotQA	2WikiMQA	Bamboogle	Average
Qwen2.5-7B-Instruct						
QA	13.4	40.8	18.3	**25.0**	12.0	21.9
CoT	4.8	18.5	9.2	11.1	23.2	13.4
IRCoT	22.4	47.8	13.3	14.9	22.4	24.2
Search-o1	15.1	44.3	18.7	17.6	**29.6**	25.1
RAG	**34.9**	**58.5**	**29.9**	23.5	20.8	**33.5**
Qwen2.5-1.5B-Instruct						
SFT	9.4	19.3	12.9	21.0	2.4	13.0
RAG	28.8	47.7	22.8	20.3	7.2	25.4
PPO	32.7	52.7	25.6	24.2	18.4	30.7
GRPO	29.3	48.0	20.2	21.3	12.0	26.1
HRPO	36.4	55.3	27.3	27.6	21.6	33.7
M3PO	**41.4**	**56.8**	**28.7**	**27.9**	**23.2**	**35.6**
Qwen2.5-3B-Instruct						
SFT	24.9	29.2	18.6	24.8	11.2	21.7
RAG	34.8	54.4	25.5	22.6	8.0	29.1
PPO	35.6	56.3	30.4	29.3	24.0	35.1
GRPO	38.1	57.0	30.8	30.3	27.2	36.7
HRPO	37.8	59.3	31.6	**31.8**	29.6	38.0
M3PO	**44.1**	**61.0**	**33.2**	31.4	**31.2**	**40.2**

Qwen2.5-7B-Instruct [54] backbone using standard QA and retrieval pipelines, namely direct inference (QA), Chain-of-Thought (CoT) [51], interleaving retrieval with CoT (IR-CoT) [44], Search-o1 [26], and RAG [25]. The best result in each block of Table 1 is highlighted in bold.

Results analysis. Table 1 shows that M3PO consistently achieves the highest average exact match (EM) scores on knowledge-intensive benchmarks for both Qwen-1.5B and Qwen-3B. With Qwen-1.5B, M3PO attains 35.6% EM, surpassing the strongest Qwen-7B RAG baseline by 2.1%. With Qwen-3B, it reaches 40.2% EM, a gain of 6.7% over the same 7B baseline. These results demonstrate that training-time multi-path collaboration can effectively compensate for limited model capacity.

Relative to RL-based methods, M3PO yields substantial gains. In the 1.5B setting, it exceeds GRPO by 9.5%, which supports the claim that explicit cross-path collaboration improves policy learning beyond group-relative updates alone. The advantage is particularly pronounced on the NQ dataset, where M3PO reaches 41.4% EM at 1.5B and 44.1% at 3B, representing improvements of 6.5% and 9.2% over the 7B RAG baseline. Moreover, M3PO con-

sistently outperforms HRPO, another RL-based hybrid reasoning method, achieving higher average scores while remaining parameter-free. This further confirms M3PO's advantage in enabling lightweight and efficient deployment.

These results validate that by leveraging parallel rollouts as independent reasoning sources and enabling cross-path collaboration, M3PO cultivates more robust reasoning patterns without sacrificing efficiency or architectural changes.

4.2. Evaluation on STEM Benchmarks

Datasets and baselines. For reasoning-intensive STEM evaluation, we employ five benchmarks spanning mathematical and scientific domains: GSM8k [6], MATH [17], MATH500 [27], MMLU-STEM [16], and ARC-Challenge [5]. For training, we use the GSM8k training split for GSM8k, the MATH training split for MATH and MATH500, and a merged corpus that combines the auxiliary MMLU and ARC-C training sets for MMLU-ST and ARC-C. Detailed settings are provided in Appendix.

For models at the 1.5B and 3B scales [54], we compare against several strong baselines: SFT, standard RL methods like PPO [34] and GRPO [37], and the RL-based latent reasoning method HRPO [59]. To further contextualize the performance of M3PO, we also include comparisons with larger models ($\geq$ 7B parameters) using few-shot CoT reasoning, specifically DeepSeekMath-7B [37], Gemma-2-9B [42], Qwen2.5-7B [54], and MAmmoTH2-7B [58].

Results analysis. As shown in Table 2, M3PO achieves state-of-the-art performance with compact Qwen models while competing effectively with substantially larger LLMs.

In particular, SFT consistently underperforms RL-based approaches, confirming the importance of verifiable rewards for complex reasoning. Moreover, M3PO with a 3B model attains an average accuracy of 70.5%, outperforming the strongest 7B baseline by 5.3%. It also surpasses HRPO on all datasets, indicating that its collaborative learning mechanism better internalizes robust reasoning patterns. A compelling case is the MATH benchmark, where the 3B model trained with M3PO achieves 60.7% accuracy, significantly exceeding the best 7B baseline by 10.9%.

Collectively, these findings validate that M3PO's collaborative reasoning paradigm effectively unlocks large-model capabilities within compact architectures, demonstrating a promising direction for developing an efficient yet powerful reasoning framework.

4.3. Ablation Study

Latent reasoning paradigm. We integrate different latent reasoning paradigms into the same policy optimization framework and train the Qwen2.5-3B-Instruct model on the MATH dataset to ensure fair comparison. The variants are: (1) Hidden States, which feed the final-layer hidden state

Table 2. Evaluation performance on STEM benchmarks. This table reports accuracy on five reasoning-intensive datasets. The upper section shows few-shot results from LLMs with at least 7B parameters, while the lower sections present the performance of smaller Qwen models (1.5B and 3B) under various training strategies. M3PO achieves the best or competitive performance across all datasets, demonstrating that the multi-path collaborative mechanism effectively enhances complex reasoning and supports strong generalization via policy optimization.

	GSM8k	MATH	MATH500	MMLU-ST	ARC-C	Average
Larger LLMs (Size $\geq$ 7B)						
DeepSeekMath-7B	64.2	36.2	34.6	56.5	67.8	51.9
Gemma-2-9B	70.7	37.7	36.4	65.1	68.2	55.6
Qwen2.5-7B	**85.4**	**49.8**	46.4	**72.3**	63.7	63.5
MAmmoTH2-7B	68.4	36.7	39.6	62.4	81.7	57.8
MAmmoTH2-8B	70.4	35.8	**73.2**	64.2	**82.2**	**65.2**
Qwen2.5-1.5B-Instruct						
SFT	56.0	30.0	30.2	40.3	60.2	43.3
PPO	67.6	45.4	44.8	56.6	71.5	57.2
GRPO	68.2	46.0	45.2	56.2	73.7	57.9
HRPO	69.9	43.8	45.8	56.9	74.2	58.1
M3PO	**70.2**	**47.4**	**48.0**	**58.1**	**77.6**	**60.3**
Qwen2.5-3B-Instruct						
SFT	67.0	34.8	36.0	45.4	47.4	46.1
PPO	81.9	59.7	60.4	58.2	81.1	68.2
GRPO	83.4	60.2	60.4	60.1	81.4	69.1
HRPO	83.5	58.6	60.2	59.0	82.0	68.7
M3PO	**84.8**	**60.7**	**63.0**	**61.6**	**82.6**	**70.5**

back as the next input; (2) Soft Thinking, which forms the next input by a soft weighted aggregation in the input embedding space; (3) HRPO, a hybrid approach that combines the sampled next input with a soft aggregation via a gating mechanism; and (4) M3PO, our hybrid reasoning method with explicit cross-path interaction (see Equation (2)).

Figure 4 visualizes the exponential moving average (EMA) of training rewards. The hidden states approach remains zero reward, primarily due to the distributional discrepancy between hidden states and the pretrained embedding space, leading to incompatibility and performance degradation. While both Soft Thinking and HRPO utilize the soft aggregation scheme, Soft Thinking exhibits slower convergence and attains lower reward levels. This performance pattern highlights the benefits of hybrid reasoning and underscores the importance of preserving the distinctive characteristics of each reasoning trajectory.

In contrast, M3PO achieves both faster convergence and higher stabilized reward levels compared to HRPO, although both methods employ hybrid reasoning principles. This performance differential confirms the superior effectiveness of our explicit multi-path interaction mechanism over conventional soft aggregation methods, validating the efficacy of structured cross-path collaboration in leverag-
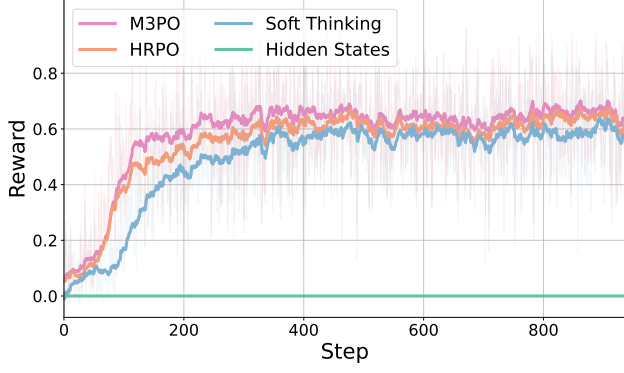
Figure 4. Training performance across latent reasoning variants. M3PO not only converges more rapidly but also attains consistently higher final rewards, underscoring the efficacy and robustness of our hybrid reasoning paradigm in complex reasoning tasks.
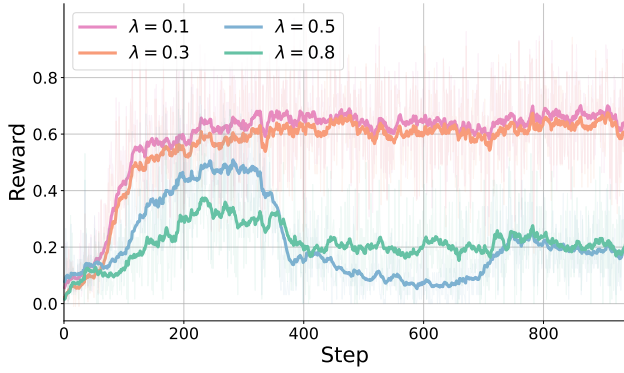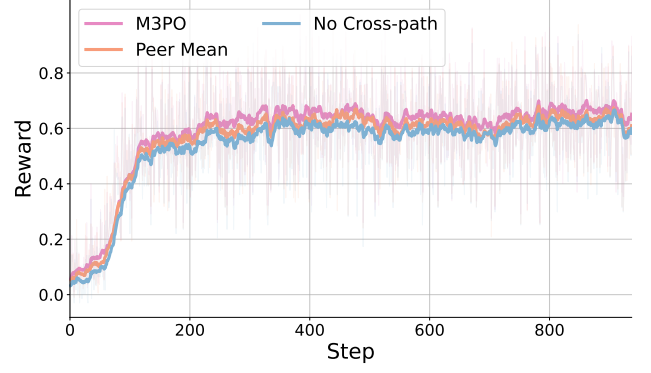

Figure 6. Ablation study on the impact of the distribution-similarity fusion mechanism. M3PO achieves the highest rewards, highlighting the effectiveness of our warm cross-path fusion strategy and the value of peer insights.
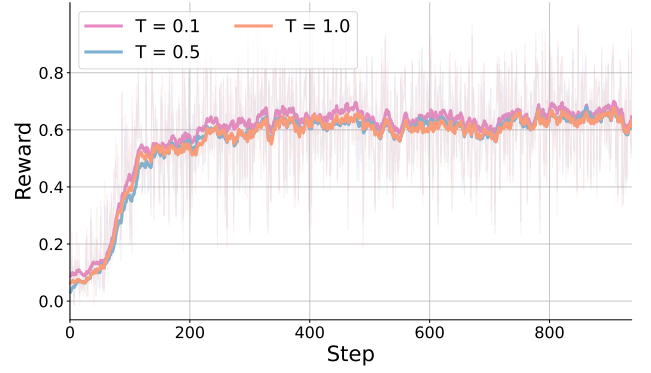

Figure 5. Sensitivity of the blending coefficient λ in Equation (2). Performance collapses when $\lambda \geq 0.5$, highlighting the critical balance between intrinsic reasoning direction and peer insights.

Figure 7. Sensitivity of the temperature T in Equation (5). Performance peaks at $T = 0.1$, where sharp attention weights enhance reasoning by selectively focusing on relevant peers.

ing independent rollouts. Importantly, unlike HRPO which requires additional parameters for its gating mechanism, M3PO maintains a fully parameter-efficient design while delivering stronger performance.

Effect of cross-path fusion mechanism. To evaluate the impact of our cross-path fusion strategy based on distributional similarity, we compare against two alternative approaches: (1) Peer Mean, which averages peer embeddings without gating; (2) No Cross-Path, which entirely removes peer insights and operates like conventional CoT reasoning.

Figure 6 shows the EMA of training rewards on MATH with Qwen2.5-3B. Replacing M3PO's similarity-guided fusion with uniform averaging, which removes selective gating and assigns equal weight to all peers, leads to measurable performance degradation. This result confirms that distributional divergences among trajectories introduce conflicting signals, hindering effective utilization of complementary reasoning paths. The performance degradation stems from a fundamental incompatibility between uniform aggregation and the structured reasoning processes inher-

ent in pretrained models. This mismatch introduces propagating noise that progressively undermines both reasoning coherence and learning stability.

Furthermore, complete removal of peer insights yields the lowest performance when the model reverts to standard CoT reasoning. This ablation result consistently underperforms both M3PO and the peer mean variant, strongly validating two crucial aspects of our approach. First, incorporating multi-path perspectives provides fundamental benefits over isolated reasoning trajectories. Second, our gated collaboration mechanism offers specific advantages over naive aggregation methods by selectively integrating the most consistent cross-path signals, thereby enhancing both reasoning robustness and training stability.

Sensitivity of λ on hybrid reasoning. M3PO is built on the hybrid reasoning scheme, which introduces a blending coefficient λ in Equation (2). To systematically examine the effect of this balancing strategy, we conduct sensitivity analysis with λ values from [0.1, 0.3, 0.5, 0.8]. Figure 5 illustrates the training rewards on the MATH

Question: What is the ones digit of $\(22^{22 \cdot (11^{11})}\)$?

Soft Thinking (653 tokens): To find the ones digit of $\(22^{22 \cdot 11^{11}} \)$, we need to focus on the ones digit of the base number $\($ which is 2 , and determine the pattern in the ones digits of powers of 2.\n Let, let's observe ... We observe that the ones digits repeat every 4 numbers: $\(2, 4, 8, 6.\n$ ... We calculate:\n $\[\n 22 \cdot 4 = 2\n\]\nThis,\n \(22^{22 \cdot 11^{11}} \equiv 2^{22 \mod 11^{11} \mod 4}\)$.\n Now, we simplify $\(22 \mod 4 \):\n\[\n22 \div 4 \ 5 \quad \text{(remainder is }2)}\n\]\n\nThus, \(22 \equiv 2 \pmod{4} \)$.\n Therefore:\n we $22 \cdot 11^{11} \equiv 2 \cdot 11^{11} \pmod{4} \)$.\n ... We simplify:\n\[\n3 \1 \equiv 3 \pmod{4}\n \quad 3^2 \equiv 9 \pmod{4}...$ Since $\(\(3^{11} \equiv 3 \pmod{4} \)$.\n ... \nFinally, we ones digit of $\(2^2 = 4 \).\n\{$, the ones digit of $\(22^{22 \cdot 11^{11}} \)$ is \{reasoning process} #### 4 ✔️

M3PO (379 tokens): To determine the ones digit of $\(22^{22 \cdot (11^{11})} \)$, we will focus on the ones digit of the base number 22, which is 2, and analyze the pattern of the ones digits of powers of 2.\n First, let's examine the ones digits of the first few powers of 2:\n\[\n\begin{align*}\n 2^1 & = 2 \quad (\text{ones digit is } 2) \\\n 2^2 & = 4 \quad (\text{ones digit is } 4)\\\n 2^3 & = 8 \quad (\text{ones digit is } 8) \\\n2^4 & = 16 \quad (\text{ones digit is } 6) \\\n 2^5 & = 32 \quad (\text{ones digit is } 2) \\\n\end{align*}\n\]\nThe ones digits repeat every 4 numbers: \(2, 4, 8, 6).\n Next, we need to determine the exponent $\(22 \cdot (11^{11})\)$ modulo 4 to find where the exponent falls within these repeating cycles. We start by simplifying $\(22 \mod 4\):\n\[\n22 \div 4 = 5 \quad \text{remainder } 2 \quad \Rightarrow \quad 22 \equiv 2 \pmod{4}\n\]\n$ The problem reduces to finding the ones digit of $\(2^2\):\n \[\n2^2 = 4 \quad (\text{ones digit is } 4)\n\]\n$ Thus, the ones digit of $\(22^{22 \cdot (11^{11})}\)$ is:\n \{reasoning process} #### 4 ✔️

Figure 8. A Comparison of training rollouts for Soft Thinking and M3PO on a mathematical problem. The complete Soft Thinking rollout is provided in the Appendix. Text highlighted in red indicates erroneous insertions, discontinuities, or formatting defects. Soft Thinking exhibits noisy and spurious text, while M3PO exhibits clean steps with consistent formatting.

dataset using Qwen2.5-3B. A clear performance collapse occurs when $\lambda \geq 0.5$, where the anchor trajectory loses its dominant role in the hybrid fusion. This outcome aligns with expectations, as substantially diminishing the contribution of the anchor trajectory disrupts the logical coherence essential for valid multi-step inference. Such disruption proves particularly critical in autoregressive models, where local inconsistencies propagate and amplify through subsequent reasoning steps.

Moreover, $\lambda = 0.1$ achieves consistently better performance than $\lambda = 0.3$, while Figure 6 confirms that $\lambda = 0.1$ (M3PO) outperforms the $\lambda = 0$ ablation (No Cross-Path). This indicates that optimal performance requires both preserving the structural integrity of the original reasoning flow and allowing controlled integration of peer-derived insights. Striking this balance allows the model to harness complementary reasoning signals without compromising coherence, maximizing its overall problem-solving capability.

Sensitivity of T on collaborative learning. To regulate information flow in multi-path collaborative learning, we introduce a temperature parameter T that governs the selectivity of attention over peer trajectories. We assess its effectiveness via sensitivity analysis with $T \in [0.1, 0.5, 1.0]$.

Figure 7 illustrates the training rewards on the MATH dataset using Qwen2.5-3B. The results show that $T = 0.1$ achieves superior performance, while both $T = 0.5$ and $T = 1.0$ exhibit comparable but lower performance levels. This confirms that sharper attention weights enhance collaboration quality by concentrating on the most relevant cross-path signals, whereas higher temperatures reduce selectivity and compromise reasoning stability through overly dispersed attention distributions.

The temperature parameter T works synergistically with the blending parameter λ to create a dual-control mechanism for robust multi-path reasoning. While λ balances the influence between the intrinsic reasoning direction and external insights, the temperature T fine-tunes the quality of incorporated external signals by filtering out less relevant contributions. This hierarchical gating strategy ensures that the model selectively perceives the most beneficial cross-path information while maintaining training stability and compatibility with the pretrained architecture.

4.4. Qualitative Results

Figure 8 compares reasoning rollouts of Soft Thinking and M3PO under identical optimization conditions. Soft Thinking produces erroneous insertions and logical discontinuities, suggesting that soft aggregation can propagate and amplify noise during autoregressive generation. In contrast, M3PO maintains a coherent, well-structured reasoning chain throughout the process, demonstrating its superior stability and noise resistance. This qualitative difference underscores the value of structured cross-path collaboration in ensuring reasoning coherence and interpretability.

5. Conclusion

In this work, we present M3PO, a multi-path collaborative RL framework that enhances reasoning robustness through structured trajectory interaction. By integrating cross-path insights into policy optimization, M3PO develops more reliable reasoning patterns while maintaining training stability. Extensive experiments demonstrate consistent improvements across various benchmarks.

Limitations. Computational resources limited our exploration to models up to 3B parameters. Future work will investigate M3PO's scalability and adaptive collaboration mechanisms. Despite this, M3PO establishes a solid foundation for multi-path collaborative reasoning that effectively balances performance and deployment feasibility.

References

[1] Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms. *arXiv preprint arXiv:2402.14740*, 2024. 3

[2] Yael Benn, Anna A Ivanova, Oliver Clark, Zachary Mineroff, Chloe Seikus, Jack Santos Silva, Rosemary Varley, and Evelina Fedorenko. The language network is not engaged in object categorization. *Cerebral Cortex*, 33(19):10380–10400, 2023. 2

[3] Shreyas Chaudhari, Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan, Ameet Deshpande, and Bruno Castro da Silva. Rlhf deciphered: A critical analysis of reinforcement learning from human feedback for llms. *ACM Computing Surveys*, 58(2):1–37, 2025. 3

[4] K Cho and A Oh. Advances in neural information processing systems 35: Annual conference on neural information processing systems 2022, 2022. 1

[5] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018. 6

[6] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. 6

[7] Yuntian Deng, Kiran Prasad, Roland Fernandez, Paul Smolensky, Vishrav Chaudhary, and Stuart Shieber. Implicit chain of thought reasoning via knowledge distillation. *arXiv preprint arXiv:2311.01460*, 2023. 2

[8] Yuntian Deng, Yejin Choi, and Stuart Shieber. From explicit cot to implicit cot: Learning to internalize cot step by step. *arXiv preprint arXiv:2405.14838*, 2024. 2

[9] Evelina Fedorenko and Rosemary Varley. Language and thought are not the same thing: evidence from neuroimaging and neurological patients. *Annals of the New York Academy of Sciences*, 1369(1):132–153, 2016. 2

[10] Evelina Fedorenko, Steven T Piantadosi, and Edward AF Gibson. Language is primarily a tool for communication rather than thought. *Nature*, 630(8017):575–586, 2024. 2

[11] Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian R Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. Scaling up test-time compute with latent reasoning: A recurrent depth approach. *arXiv preprint arXiv:2502.05171*, 2025. 2

[12] Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. Think before you speak: Training language models with pause tokens. *arXiv preprint arXiv:2310.02226*, 2023. 2

[13] Halil Alperen Gozeten, M Emrullah Ildiz, Xuechen Zhang, Hrayr Harutyunyan, Ankit Singh Rawat, and Samet Oymak. Continuous chain of thought enables parallel exploration and reasoning. *arXiv preprint arXiv:2505.23648*, 2025. 2

[14] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 1

[15] Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024. 1, 2

[16] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020. 6

[17] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021. 6

[18] Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. *arXiv preprint arXiv:2011.01060*, 2020. 5

[19] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 1

[20] Jian Hu. Reinforce++: A simple and efficient approach for aligning large language models. *arXiv preprint arXiv:2501.03262*, 2025. 3

[21] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024. 1

[22] Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025. 1

[23] Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017. 5

[24] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019. 5

[25] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020. 5

[26] Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. Search-o1: Agentic search-enhanced large reasoning models. *arXiv preprint arXiv:2501.05366*, 2025. 5

[27] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2023. 6

[28] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937. PmLR, 2016. 3

[29] R OpenAI. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2(5):1, 2023. 1

[30] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022. 3

[31] Jacob Pfau, William Merrill, and Samuel R Bowman. Let's think dot by dot: Hidden computation in transformer language models. *arXiv preprint arXiv:2404.15758*, 2024. 2

[32] Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*, 2022. 5

[33] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023. 3

[34] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 3, 5, 6

[35] Ashish Kumar Shakya, Gopinatha Pillai, and Sohom Chakrabarty. Reinforcement learning algorithms: A brief survey. *Expert Systems with Applications*, 231:120495, 2023. 3

[36] Yuval Shalev, Amir Feder, and Ariel Goldstein. Distributional reasoning in llms: Parallel reasoning processes in multi-hop reasoning. *arXiv preprint arXiv:2406.13858*, 2024. 2

[37] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. *URL https://arxiv. org/abs/2402.03300*, 2(3):5, 2024. 3, 5, 6

[38] Xuan Shen, Yizhou Wang, Xiangxi Shi, Yanzhi Wang, Pu Zhao, and Jiuxiang Gu. Efficient reasoning with hidden thinking. *arXiv preprint arXiv:2501.19201*, 2025. 1

[39] Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. Codi: Compressing chain-of-thought into continuous space via self-distillation. *arXiv preprint arXiv:2502.21074*, 2025. 1, 2

[40] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999. 3

[41] Jihoon Tack, Jack Lanchantin, Jane Yu, Andrew Cohen, Ilia Kulikov, Janice Lan, Shibo Hao, Yuandong Tian, Jason Weston, and Xian Li. Llm pretraining with continuous concepts. *arXiv preprint arXiv:2502.08524*, 2025. 1, 2

[42] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024. 6

[43] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1

[44] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 10014–10037, 2023. 5

[45] Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965, 2023. 2

[46] Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. Towards understanding chain-of-thought prompting: An empirical study of what matters. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 2717–2739, 2023. 2

[47] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pretraining. *arXiv preprint arXiv:2212.03533*, 2022. 5

[48] Xu Wang, Sen Wang, Xingxing Liang, Dawei Zhao, Jincai Huang, Xin Xu, Bin Dai, and Qiguang Miao. Deep reinforcement learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 35(4):5064–5078, 2022. 3

[49] Xinyi Wang, Lucas Caccia, Oleksiy Ostapenko, Xingdi Yuan, William Yang Wang, and Alessandro Sordoni. Guiding language model reasoning with planning tokens. *arXiv preprint arXiv:2310.05707*, 2023. 2

[50] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022. 1

[51] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 1, 5

[52] Chünhung Wu, Jinliang Lu, Zixuan Ren, Gangqiang Hu, Zhi Wu, Dai Dai, and Hua Wu. Llms are single-threaded reasoners: Demystifying the working mechanism of soft thinking. *arXiv preprint arXiv:2508.03440*, 2025. 1, 2

[53] Siheng Xiong, Ali Payani, and Faramarz Fekri. Enhancing long chain-of-thought reasoning through multi-path plan aggregation. *arXiv preprint arXiv:2510.11620*, 2025. 2

[54] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. 5, 6

[55] Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. Do large language models latently perform multi-hop reasoning? *arXiv preprint arXiv:2402.16837*, 2024. 2

[56] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*, 2018. 5

[57] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023. 1

[58] Xiang Yue, Tianyu Zheng, Ge Zhang, and Wenhu Chen. Mammoth2: Scaling instructions from the web. *Advances in Neural Information Processing Systems*, 37:90629–90660, 2024. 6, 1

[59] Zhenrui Yue, Bowen Jin, Huimin Zeng, Honglei Zhuang, Zhen Qin, Jinsung Yoon, Lanyu Shang, Jiawei Han, and Dong Wang. Hybrid latent reasoning via reinforcement learning. *arXiv preprint arXiv:2505.18454*, 2025. 2, 5, 6, 1

[60] Kaiyan Zhang, Yuxin Zuo, Bingxiang He, Youbang Sun, Runze Liu, Che Jiang, Yuchen Fan, Kai Tian, Guoli Jia, Pengfei Li, et al. A survey of reinforcement learning for large reasoning models. *arXiv preprint arXiv:2509.08827*, 2025. 3

[61] Zhen Zhang, Xuehai He, Weixiang Yan, Ao Shen, Chenyang Zhao, Shuohang Wang, Yelong Shen, and Xin Eric Wang. Soft thinking: Unlocking the reasoning potential of llms in continuous concept space. *arXiv preprint arXiv:2505.15778*, 2025. 1, 2, 5

[62] Yufan Zhuang, Liyuan Liu, Chandan Singh, Jingbo Shang, and Jianfeng Gao. Text generation beyond discrete token sampling. *arXiv preprint arXiv:2505.14827*, 2025. 2

Multi-Path Collaborative Reasoning via Reinforcement Learning

Supplementary Material

6. Implementation

M3PO is a lightweight multi-path collaborative learning framework designed to be compatible with any LLM architecture. It introduces no additional trainable parameters and preserves the original inference efficiency of the base model. To enable efficient training under this framework, we integrate optimized kernel implementations from Unsloth[1] and apply low-rank adaptation (LoRA) [19] for parameter-efficient fine-tuning. All M3PO experiments adopt the hyperparameter settings detailed in Table 3, which serve as our default unless otherwise specified.

Table 3. Training settings for M3PO.

Algorithm	M3PO
Epochs	1
Optimizer	AdamW 8bit
Optimizer Momentum	$\beta_1, \beta_2 = 0.9, 0.99$
Weight Decay	0.1
Learning Rate	5e-6
M3PO β	0.005
Max Gradient Norm	0.1
Gradient Accumulation Step	4
Group size N in M3PO	4 / 8
Total Train Batch Size	64
LR Scheduler	Cosine with Warmup
Warmup Ratio	0.1
Precision (WA)	BF16-mixed
Completion Length	1024
λ in Equation (2)	0.1
T in Equation (5)	0.1
LoRA Modules	query, key, value, dense
LoRA Rank	32
LoRA α	64

Thanks to the lightweight design of M3PO and our optimized kernels, M3PO runs efficiently on a single GPU for all tasks. Notably, all STEM benchmark datasets are used directly from their HuggingFace[2] repositories without additional preprocessing. We use a fixed group size of 4 for knowledge-intensive tasks. For complex reasoning benchmarks, including GSM8k, MATH, and MMLU-STEM, we generate 8 hybrid completions per query to enhance exploration and robustness.

The prompt construction follows a consistent template during both training and evaluation. Each input begins with

---

[1] https://github.com/unslothai/unsloth
[2] https://huggingface.co

a system message instructing the model to carry out step-by-step reasoning before generating its final answer, followed by the user query. The full prompt is then formatted according to the model's native chat template. In line with established practice [59], we use the minimal delimiter #### to separate the model's reasoning chain from its final answer. This delimiter tokenizes as a single unit, adding no sequence length overhead while providing a clear signal for transitioning from latent collaborative reasoning to autoregressive answer generation. To ensure reasoning integrity, a penalty mechanism assigns zero reward to completions containing repeated delimiters, regardless of answer correctness. This prevents early termination of the reasoning process and encourages complete reasoning chains. Full prompt examples for different task types, including system messages and representative queries, are provided in Figures 9, 10, and 11, respectively.

We use greedy decoding and the standard inference pipeline of the base LLM to ensure reproducibility. For knowledge tasks, exact match scores are reported on validation and test splits following [22]. For mathematical reasoning benchmarks including GSM8K, MATH, and MATH-500, along with multiple-choice datasets (MMLU-STEM and ARC-Challenge), we follow the post-processing and scoring procedures defined in [58].

7. Additional Results

Comparison to latent reasoning methods. Beyond the RL baselines examined in our main experiments, we further compare M3PO against representative latent reasoning approaches, including HRPO, Coconut, and CODI. All methods are evaluated on the GSM8k and MATH benchmarks using the Qwen-1.5B backbone. For Coconut, we utilize its augmented CoT training data, while for CODI we adopt the original CoT trajectories from each dataset. HRPO is implemented under the same settings as M3PO.

As summarized in Table 4, M3PO achieves the highest accuracy on both reasoning tasks, outperforming all latent reasoning baselines by a consistent margin. Coconut

Table 4. Performance comparison of M3PO against alternative latent reasoning methods. M3PO achieves consistent improvements over all baselines on both benchmarks.

	Coconut		CODI		HRPO		M3PO	
	GSM8k	MATH	GSM8k	MATH	GSM8k	MATH	GSM8k	MATH
Accuracy	31.5	-	65.8	41.9	69.9	43.8	**70.2**	**47.4**

Example Prompt for Knowledge Tasks

```
<|im_start|>system
A conversation between User and Assistant. The user asks a question, and the assistant solves it
. The assistant first thinks about the reasoning process in the mind and then provides the user
with the answer. The final answer is provided after the #### tag, i.e., {reasoning process} ####
 {answer}.<|im_end|>
<|im_start|>user
Context (which may or may not be relevant):
Anton Zingarevich::::Anton Zingarevich Anton Zingarevich is a Russian businessman ...
Anton Zingarevich::::The couple married in late 2009 and had a child ...
Nigel Howe::::Nigel Howe Nigel Howe (born 7 April 1958) is a British property developer...

Question: who wrote the first declaration of human rights?<|im_end|>
<|im_start|>assistant
```

Figure 9. Example prompt for knowledge tasks in M3PO. Context is partially omitted for brevity.

Example Prompt for GSM8k / MATH / MATH500

```
<|im_start|>system
A conversation between User and Assistant. The user asks a question, and the assistant solves it
. The assistant first thinks about the reasoning process in the mind and then provides the user
with the answer. The final answer is provided after the #### tag, i.e., {reasoning process} ####
 {answer}.<|im_end|>
<|im_start|>user
The square of an integer is 182 greater than the integer itself. What is the sum of all integers
 for which this is true?<|im_end|>
<|im_start|>assistant
```

Figure 10. Example prompt for GSM8k / MATH / MATH500 in M3PO.

lags notably on GSM8k, suggesting limitations in its token compression strategy for representing complex reasoning. Although CODI shows substantial gains from CoT fine-tuning, it still trails behind M3PO in final performance. HRPO, as a hybrid reasoning approach, shows competitive results but remains consistently behind M3PO across both benchmarks. These results collectively demonstrate that our approach maintains consistent advantages over existing latent reasoning methods, highlighting the effectiveness of explicit multi-path collaboration compared to implicit representation learning techniques.

Efficiency analysis. Figures 12, 13, 14, 15 present comprehensive comparisons of reward trajectories and reasoning chain lengths across GSM8k and MATH datasets using Qwen-1.5B and 3B model variants. On GSM8k, M3PO demonstrates exceptional training stability while achieving the highest final reward among all compared methods. This advantage is particularly pronounced with the Qwen-1.5B configuration, where M3PO maintains consistent convergence throughout the training process. In contrast, both HRPO and GRPO exhibit significant performance instability, characterized by a notable collapse around 600 training steps that substantially impacts their final performance. Beyond reward superiority, M3PO yields the most compact training completions, reducing unnecessary computational overhead while maintaining solution quality.

The MATH benchmark further validates M3PO's effectiveness, where it achieves superior performance with reasoning chain lengths comparable to other methods. This consistent pattern across datasets and model scales demonstrates M3PO's robust generalization capability and computational efficiency. These empirical results establish M3PO

Figure 11. Example prompt for MMLU-ST / ARC-C in M3PO.

as an effective plug-and-play module that optimally balances performance with computational efficiency.

Notably, the core architecture of M3PO remains parameter-efficient, introducing no additional structural parameters beyond the base model. This lightweight architecture contrasts with HRPO, which embeds trainable parameters directly into its reasoning architecture. Furthermore, M3PO preserves the same computational efficiency as standard autoregressive models during deployment, while delivering superior reasoning performance. These advantages establish M3PO as an effective and practical solution for robust reasoning.

8. Qualitative Analysis

To qualitatively demonstrate the advantages of our approach, we conduct detailed case studies comparing M3PO with alternative reasoning methods. Figure 16 presents a comparison between M3PO and Soft Thinking. The Soft Thinking approach exhibits significant formatting issues, erroneous token insertions, and logical discontinuities throughout its reasoning chain. In contrast, M3PO generates a clean, logically coherent, and compact reasoning process free from these artifacts. This comparison clearly demonstrates how continuous soft aggregation in Soft Thinking introduces noise that disrupts reasoning coherence, while highlighting M3PO's superior ability to maintain structured reasoning patterns.

Figure 17 provides a case study comparing M3PO with HRPO. While both methods employ hybrid reasoning paradigms, HRPO displays noticeable repetitive looping behavior that persists to the maximum completion length without reaching a solution. In contrast, M3PO correctly solves the problem through a well-structured and logically consistent reasoning trajectory. This qualitative difference demonstrates M3PO's superior robustness in maintaining coherent reasoning progress without degenerative patterns.

Figures 18 to 23 present reasoning examples generated by M3PO across diverse domains, including knowledge-intensive tasks, scientific reasoning, and mathematical problems. These cases collectively demonstrate M3PO's consistent ability to produce readable, coherent, and compact reasoning chains that lead to correct solutions.

In knowledge-intensive domains (Figures 18 and 19), M3PO integrates factual information with logical inference, maintaining clarity throughout multi-step reasoning processes. For scientific reasoning tasks (Figures 20 and 21), the method demonstrates precise conceptual understanding and systematic problem-solving approaches. Mathematical problems (Figures 22 and 23) further showcase M3PO's capacity for structured derivation and computational accuracy.

Across all domains, M3PO generates reasoning chains characterized by logical coherence and expressive clarity. This behavior is a direct consequence of our multi-path coordination mechanism that implicitly guides gradient updates during policy optimization, enabling the model to develop robust reasoning patterns without compromising output coherence. This qualitative evidence complements our quantitative results, demonstrating the practical utility of our approach for generating interpretable and reliable reasoning across different task types.
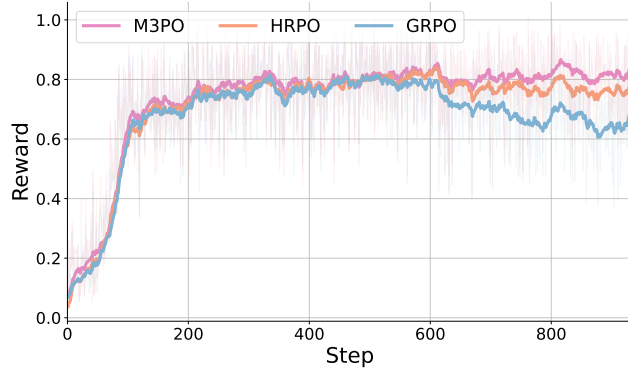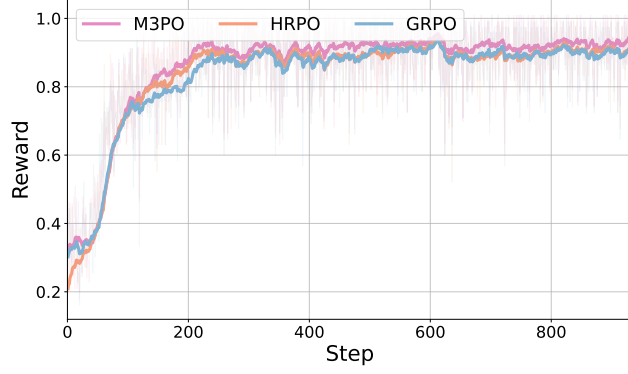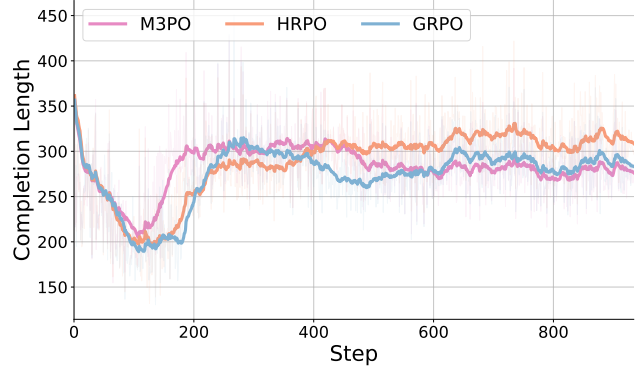
(a) Reward.
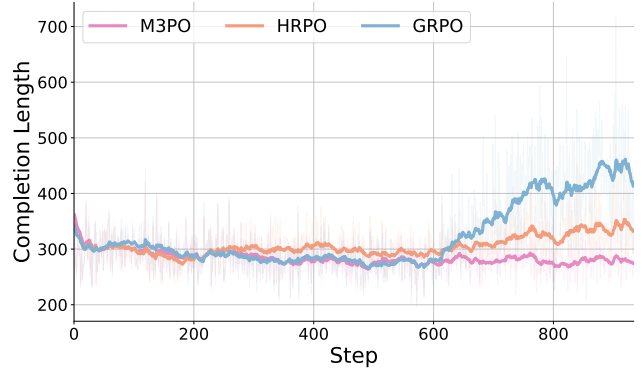
(b) Completion length.

Figure 12. Comparison of reward and completion length across training runs on GSM8k with the Qwen-1.5B backbone. M3PO exhibits the best training stability while producing the shortest completions.

(a) Reward.

(b) Completion length.

Figure 13. Comparison of reward and completion length across training runs on GSM8k using the Qwen-3B backbone. M3PO achieves the highest reward while generating the shortest completions.

(a) Reward.

(b) Completion length.

Figure 14. Comparison of reward and completion length across training runs on MATH using the Qwen-1.5B backbone. M3PO attains the highest reward while maintaining comparable completion lengths.
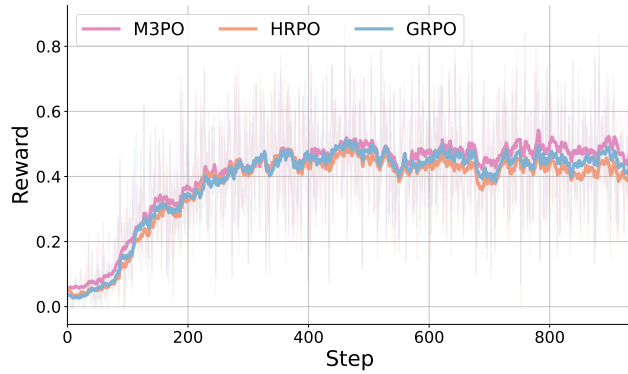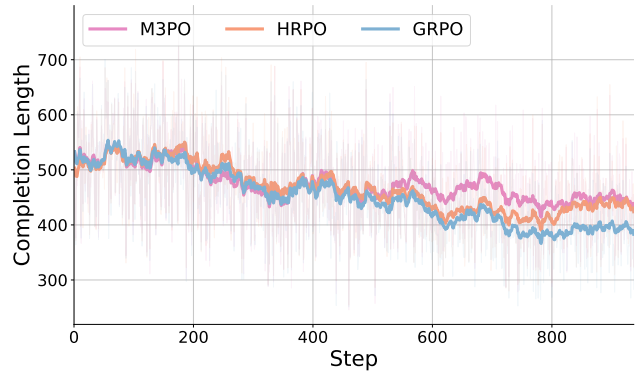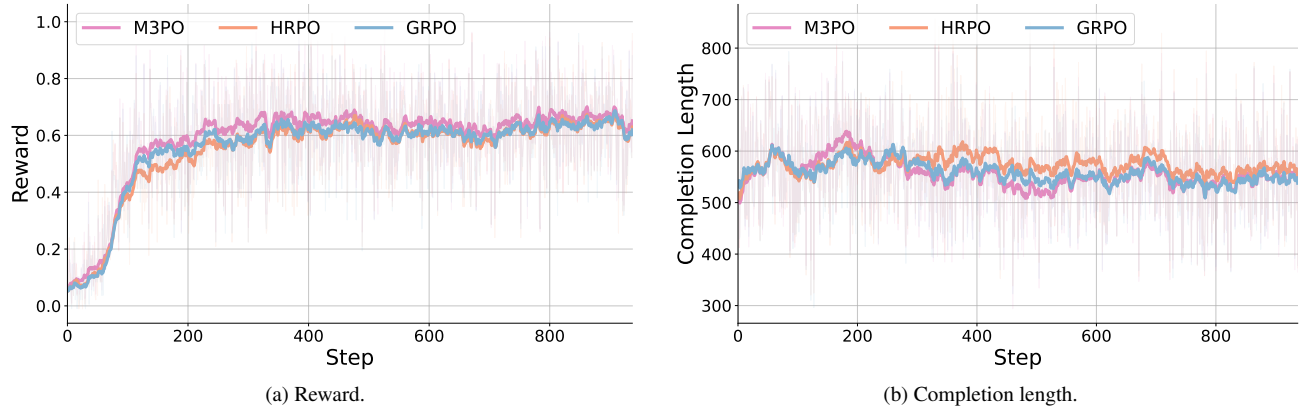
4

(a) Reward.

(b) Completion length.

Figure 15. Comparison of reward and completion length across training runs on MATH using the Qwen-3B backbone. M3PO achieves the highest reward while maintaining comparable completion length.

Question: What is the ones digit of $\(22^{22 \cdot (11^{11})}\)$?

Soft Thinking (653 tokens): To find the ones digit of $\(22^{22 \cdot 11^{11}} \)$, we need to focus on the ones digit of the base number \(which is 2 , and determine the pattern in the ones digits of powers of 2.\nLet, let's observe the ones digits of the first few powers of 2:\n\[\n\begin{align*}\n2^1 & = 2 \quad (\text{ones digit is } 2) \\\n2^2 & = 4 \quad (\text{ones digit is } 4) \\\n2^3 & = 8 \quad (\text{ones digit is } 8) \\\n2^4 & = 16 \quad (\text{ones digit is } 6) \\\n2^5 & = 32 \quad (\text{ones digit is } 2) \\\n\end{align*}\n\]\nWe observe that the ones digits repeat every 4 numbers: \(2, 4, 8, 6.\nNext, we need to determine the exponent modulo 4 to identify the position within this repeating cycle:\n We calculate:\n\[\n22 \cdot 4 = 2\n\]\nThis,\n \(22^{22 \cdot 11^{11}} \equiv 2^{22 \mod 11^{11} \mod 4}\).\nNow, we simplify \(22 \mod 4 \):\n\[\n22 \div 4 \ 5 \quad \text{(remainder is }2)}\n\]\nThus, \(22 \equiv 2 \pmod{4} \).\nTherefore:\n we 22 \cdot 11^{11} \equiv 2 \cdot 11^{11} \pmod{4} \).\nNext, we consider \(11 \mod 4 \):\n\[\n11 \equiv 3 \pmod{4}\n\]\nThus,\n\[\n11^{11} \equiv 3^{11} \pmod{4}\n\]\nWe simplify:\n\[\n3 \1 \equiv 3 \pmod{4}\n \quad 3^2 \equiv 9 \pmod{4}, \quad 3^3 \equiv 3 \pmod{4}, \quad 3^4 \equiv 1 \pmod{4}\n\]\nand so on. Since \(\(3^{11} \equiv 3 \pmod{4} \).\nThus,\n\[\n2 \cdot 11^{11} \equiv 2 \cdot 3 \equiv 6 \equiv 2 \pmod{4}\n\]\nTherefore, \(22^{22 \cdot 11^{11}} \equiv 2^2 \pmod{10} \).\nFinally, we ones digit of \(2^2 = 4 \).\n{, the ones digit of \(22^{22 \cdot 11^{11}} \) is \{reasoning process} #### 4 ✔

M3PO (379 tokens): To determine the ones digit of \(22^{22 \cdot (11^{11})} \), we will focus on the ones digit of the base number 22, which is 2, and analyze the pattern of the ones digits of powers of 2.\n First, let's examine the ones digits of the first few powers of 2:\n\[\n\begin{align*}\n 2^1 & = 2 \quad (\text{ones digit is } 2) \\\n 2^2 & = 4 \quad (\text{ones digit is } 4)\\\n 2^3 & = 8 \quad (\text{ones digit is } 8) \\\n2^4 & = 16 \quad (\text{ones digit is } 6) \\\n 2^5 & = 32 \quad (\text{ones digit is } 2) \\\n\end{align*}\n\]\nThe ones digits repeat every 4 numbers: \(2, 4, 8, 6\).\n Next, we need to determine the exponent \(22 \cdot (11^{11})\) modulo 4 to find where the exponent falls within these repeating cycles. We start by simplifying \(22 \mod 4\):\n\[\n22 \div 4 = 5 \quad \text{remainder } 2 \quad \Rightarrow \quad 22 \equiv 2 \pmod{4}\n\]\n The problem reduces to finding the ones digit of \(2^2\):\n \[\n2^2 = 4 \quad (\text{ones digit is } 4)\n\]\n Thus, the ones digit of \(22^{22 \cdot (11^{11})}\) is:\n {reasoning process} #### 4 ✔

Figure 16. Case study comparing Soft Thinking and M3PO on a mathematical problem. Text highlighted in red indicates erroneous insertions, discontinuities, or formatting defects. Soft Thinking shows noisy, spurious text, whereas M3PO presents clean, consistently formatted steps. .

5

Question: Solve for x: $$(\sqrt{12x}+12)(\sqrt{3x}-6)=4(x+3)+x-34$$?

HRPO (1024 tokens): To solve the equation \\((\\sqrt{12x} + 12)(\\sqrt{3x} - 6) = 4(x + 3) + x - 34\\), we will follow these steps:\n\n1. Simplify the right-hand side of the equation.\n2. Expand and simplify both sides of the equation.\n3. Solve for \\(x\\).\n\nFirst, simplify the right-hand side:\n\\[4(x + 3) + x - 34 = 4x + 12 + x - 34 = 5x - 22.\\]\nSo the equation becomes:\n\\[(\\sqrt{12x} + 12)(\\sqrt{3x} - 6) = 5x - 22.\\]\n\nNext, expand the left-hand side:\n\\[(\\sqrt{12x} + 12)(\\sqrt{3x} - 6) = \\sqrt{12x} \\cdot \\sqrt{3x} - 6\\sqrt{12x} + 12\\sqrt{3x} - 72.\\]\nSince \\(\\sqrt{12x} \\cdot \\sqrt{3x} = \\sqrt{(12x)(3x)} = \\sqrt{36x^2} = 6x\\), we have:\n\\[6x - 6\\sqrt{12x} + 12\\sqrt{3x} - 72.\\]\nThus, the equation is:\n\\[6x - 6\\sqrt{12x} + 12\\sqrt{3x} - 72 = 5x - 22.\\]\nRearrange terms to isolate the square root terms:\n\\[6x - 5x - 72 + 22 = 6\\sqrt{12x} - 12\\sqrt{3x},\\]\nwhich simplifies to:\n\\[x - 50 = 6\\sqrt{12x} - 12\\sqrt{3x}.\\]\nFactor out common terms on the right-hand side:\n\\[x - 50 = 6\\sqrt{12x} - 12\\sqrt{3x} = 6\\sqrt{12x} - 6\\sqrt{12x} + 12\\sqrt{3x} - 12\\sqrt{3x} = 6\\sqrt{12x} - 6\\sqrt{12x} + 12\\sqrt{3x} - 12\\sqrt{3x} = 6\\sqrt{12x} - 6\\sqrt{12x} + 12\\sqrt{3x} - 12\\sqrt{3x} = 6\\sqrt{12x} - 6\\sqrt{12x} + 12\\sqrt{3x} - 12\\sqrt{3x} = 6\\sqrt{12x} - 6\\sqrt{12x} + 12\\sqrt{3x} - 12\\sqrt{3x} = 6\\sqrt{12x} - 6\\sqrt{12x} + 12\\sqrt{3x} - 12\\sqrt{3x} = 6\\sqrt{12x} - 6\\sqrt{12x} + 12\\sqrt{3x} - 12\\sqrt{3x} = 6\\sqrt{12x} - 6\\sqrt{12x} + 12\\sqrt{3x} - 12\\sqrt{3x} = 6\\sqrt{12x} - 6\\sqrt{12x} + 12\\sqrt{3x} - 12\\sqrt{3x} = 6\\sqrt{12x} - 6\\sqrt{12x} + 12\\sqrt{3x} - 12\\sqrt{3x} ... ❌

M3PO (845 tokens): To solve the equation \\((\\sqrt{12x} + 12)(\\sqrt{3x} - 6) = 4(x + 3) + x - 34\\), we will follow these steps:\n\n1. Simplify the right-hand side of the equation.\n2. Expand and simplify both sides of the equation.\n3. Solve for \\(x\\).\n\nFirst, let's simplify the right-hand side:\n\\[4(x + 3) + x - 34 = 4x + 12 + x - 34 = 5x - 22.\\]\nSo the equation becomes:\n\\[(\\sqrt{12x} + 12)(\\sqrt{3x} - 6) = 5x - 22.\\]\n\nNext, we expand the left-hand side:\n\\[(\\sqrt{12x} + 12)(\\sqrt{3x} - 6) = \\sqrt{12x} \\cdot \\sqrt{3x} - 6\\sqrt{12x} + 12\\sqrt{3x} - 72.\\]\nSince \\(\\sqrt{12x} \\cdot \\sqrt{3x} = \\sqrt{(12x)(3x)} = \\sqrt{36x^2} = 6x\\), we can rewrite the expression as:\n\\[6x - 6\\sqrt{12x} + 12\\sqrt{3x} - 72.\\]\nWe know that \\(\\sqrt{12x} = 2\\sqrt{3x}\\), so we can substitute this into the expression:\n\\[6x - 6(2\\sqrt{3x}) + 12\\sqrt{3x} - 72 = 6x - 12\\sqrt{3x} + 12\\sqrt{3x} - 72 = 6x - 72.\\]\nThus, the equation simplifies to:\n\\[6x - 72 = 5x - 22.\\]\nNow, we solve for \\(x\\) by isolating \\(x\\):\n\\[6x - 5x = -22 + 72,\\]\n\\[x = 50.\\]\n\nFinally, we need to check if \\(x = 50\\) satisfies the original equation. Substituting \\(x = 50\\) into the original equation:\n\\[\n(\\sqrt{12 \\cdot 50} + 12)(\\sqrt{3 \\cdot 50} - 6) = 4(50 + 3) + 50 - 34,\n\\]\n\\[\n(\\sqrt{600} + 12)(\\sqrt{150} - 6) = 4 \\cdot 53 + 50 - 34,\n\\]\n\\[(10\\sqrt{6} + 12)(5\\sqrt{6} - 6) = 212 + 50 - 34,\n\\]\n\\[(10\\sqrt{6} + 12)(5\\sqrt{6} - 6) = 228.\\]\nExpanding the left-hand side:\n\\[\n(10\\sqrt{6} + 12)(5\\sqrt{6} - 6) = 10\\sqrt{6} \\cdot 5\\sqrt{6} - 10\\sqrt{6} \\cdot 6 + 12 \\cdot 5\\sqrt{6} - 12 \\cdot 6,\n\\]\n\\[= 300 - 60\\sqrt{6} + 60\\sqrt{6} - 72 = 228.\\]\nSince both sides are equal, \\(x = 50\\) is indeed a solution.\n\nTherefore, the solution is:\n{reasoning process} #### 50 ✔

Figure 17. Case study comparing HRPO and M3PO on a mathematical problem. Text in red denotes erroneous insertions, discontinuities, or formatting defects. HRPO exhibits repetitive loops that persist to the maximum completion length, whereas M3PO produces coherent, logically structured steps.

Question: who will take the throne after the queen dies?

M3PO (105 tokens): Queen Elizabeth II is the current monarch of the United Kingdom. According to the information provided, her heir apparent is her eldest son, Charles, Prince of Wales. Therefore, if Queen Elizabeth II were to die, Charles, Prince of Wales would most likely take the throne.\n\n{reasoning process} The context clearly states that Queen Elizabeth II's eldest son, Charles, is her heir apparent. This means he is the next in line to become the monarch after her. {/reasoning_process}\n#### Charles, Prince of Wales ✔

Figure 18. Reasoning example 1: M3PO on a knowledge-intensive task. The context is omitted for brevity.

Question: What year did Guns N Roses perform a promo for a movie starring Arnold Schwarzenegger as a former New York Police detective?

M3PO (258 tokens): To determine the year Guns N' Roses performed a promo for a movie starring Arnold Schwarzenegger as a former New York Police detective, let's follow these steps:\n\n1. Identify the movie: The context mentions \"End of Days (film)\", which is a 1999 American supernatural action horror film starring Arnold Schwarzenegger.\n\n2. Confirm the role: The movie indeed features Arnold Schwarzenegger as a former New York Police Department detective named Jericho Cane.\n\n3. Check for promotional activities: The context states that during \"End of Days\" editing, soundtrack songs were overlaid in scenes that are typically silent in thriller films. This implies that Guns N' Roses likely performed a promo for the movie.\n\n4. Determine the year: Given that \"End of Days\" is a 1999 film, we can conclude that Guns N' Roses performed a promo for this movie in 1999.\n\n{reasoning process} The movie featuring Arnold Schwarzenegger as a former New York Police detective is \"End of Days\", which was released in 1999. Therefore, Guns N' Roses performed a promo for this movie in 1999. {/reasoning_process}\n#### 1999 ✔

Figure 19. Reasoning example 2: M3PO on a knowledge-intensive task. The context is omitted for brevity.

Question: Find the degree for the given field extension Q(sqrt(2), sqrt(3)) over Q.

M3PO (307 tokens): The degree of the extension Q[sqrt(2), sqrt(3)] over Q is the dimension of Q[sqrt(2), sqrt(3)] as a vector space over Q.\nLet's consider the basis of Q[sqrt(2), sqrt(3)] over Q. We know that sqrt(2) and sqrt(3) are linearly independent over Q. So we can take the set {1, sqrt(2), sqrt(3), sqrt(6)} as a basis. This is because any element in Q[sqrt(2), sqrt(3)] can be written uniquely as a linear combination of these four elements with coefficients in Q.\nSo the dimension of Q[sqrt(2), sqrt(3)] over Q is 4.\n{reasoning process} Let K = Q[sqrt(2), sqrt(3)]. Then K is a field extension of Q. Since sqrt(2) and sqrt(3) are not in Q, they generate K over Q. The minimal polynomial of sqrt(2) over Q is x^2 - 2, which has degree 2. Similarly, the minimal polynomial of sqrt(3) over Q is x^2 - 3, which also has degree 2. Since sqrt(2) and sqrt(3) are algebraically independent over Q, the degree of the extension K/Q is the product of the degrees of the minimal polynomials, which is 2 * 2 = 4. {reasoning process} #### B ✔

Figure 20. Reasoning example 3: M3PO on a MMLU-STEM task. The options are omitted for brevity.

Question: When an electric current is passed through water (H_{2}O) using a specific apparatus, hydrogen (H_{2}) and oxygen (O_{2}) gases form. How can this reaction be shown in a chemical equation?",

M3PO (303 tokens): Let's think about this step by step:\n\n1. We need to write a chemical equation for the decomposition of water into hydrogen and oxygen when electricity is applied.\n\n2. Water (H\u2082O) is composed of hydrogen (H\u2082) and oxygen (O\u2082). \n\n3. When electricity is passed through water, it splits into its constituent elements - hydrogen and oxygen.\n\n4. The balanced chemical equation for this reaction is: \n \\[\n 2H_2O(l) \\rightarrow 2H_2(g) + O_2(g)\n \\]\n\n5. Let's verify each option:\n   - Option A: \\(H_2O \\rightarrow H + O + H\\) is incorrect because water does not decompose into hydrogen and oxygen atoms.\n   - Option B: \\(2H_2O(l) \\rightarrow 2H_2(g) + O_2(g)\\) correctly represents the reaction where two molecules of water produce two molecules of hydrogen gas and one molecule of oxygen gas.\n   - Option C: \\(H:O:H \\rightarrow H_2O\\) is not a balanced chemical equation; it just shows the molecular structure of water.\n   - Option D: \\(H_2O(l) \\rightarrow 2H(g) + O(g)\\) is incorrect as it suggests that water decomposes into individual hydrogen and oxygen atoms rather than molecules.\n\nTherefore, the correct chemical equation is given by option B.\n\n{reasoning process} #### B ✔

Figure 21. Reasoning example 4: M3PO on an ARC-Challenge task. The options are omitted for brevity.

Question: Simplify and write the result with a rational denominator: $$\\sqrt{\\sqrt[3]{\\sqrt{\\frac{1}{729}}}}$$

M3PO (456 tokens): To simplify the expression \\(\\sqrt{\\sqrt[3]{\\sqrt{\\frac{1}{729}}}}\\), we will break it down step by step.\n\nFirst, let's simplify the innermost part of the expression, which is \\(\\sqrt{\\frac{1}{729}}\\).\n\nThe number 729 can be written as \\(729 = 3^6\\). Therefore,\n\\[\n\\sqrt{\\frac{1}{729}} = \\sqrt{\\frac{1}{3^6}} = \\left(\\frac{1}{3^6}\\right)^{1/2} = \\frac{1^{1/2}}{(3^6)^{1/2}} = \\frac{1}{3^3} = \\frac{1}{27}.\n\\]\n\nNext, we need to simplify \\(\\sqrt[3]{\\frac{1}{27}}\\). Since \\(27 = 3^3\\),\n\\[\n\\sqrt[3]{\\frac{1}{27}} = \\sqrt[3]{\\frac{1}{3^3}} = \\left(\\frac{1}{3^3}\\right)^{1/3} = \\frac{1^{1/3}}{(3^3)^{1/3}} = \\frac{1}{3}.\n\\]\n\nFinally, we need to simplify \\(\\sqrt{\\frac{1}{3}}\\). This can be written as\n\\[\n\\sqrt{\\frac{1}{3}} = \\left(\\frac{1}{3}\\right)^{1/2} = \\frac{1^{1/2}}{3^{1/2}} = \\frac{1}{\\sqrt{3}}.\n\\]\n\nTo rationalize the denominator, we multiply the numerator and the denominator by \\(\\sqrt{3}\\):\n\\[\n\\frac{1}{\\sqrt{3}} \\cdot \\frac{\\sqrt{3}}{\\sqrt{3}} = \\frac{\\sqrt{3}}{3}.\n\\]\n\nTherefore, the simplified form of the given expression is\n{reasoning process} #### \\(\\frac{\\sqrt{3}}{3}\\) ✔

Figure 22. Reasoning example 5: M3PO on a mathematical task.

Question: Simplify $(5-3i)(-4+3i)$.

M3PO (431 tokens): To simplify the expression \\((5-3i)(-4+3i)\\), we will use the distributive property (also known as the FOIL method for binomials). The FOIL method stands for First, Outer, Inner, Last, which refers to the positions of the terms in the binomials.\n\nLet's break it down step by step:\n\n1. **First**: Multiply the first terms in each binomial.\n   \\[\n   5 \\cdot (-4) = -20\n   \\]\n\n2. **Outer**: Multiply the outer terms in the product.\n   \\[\n   5 \\cdot 3i = 15i\n   \\]\n\n3. **Inner**: Multiply the inner terms in the product.\n   \\[\n   -3i \\cdot (-4) = 12i\n   \\]\n\n4. **Last**: Multiply the last terms in each binomial.\n   \\[\n   -3i \\cdot 3i = -9i^2\n   \\]\n\nNow, combine all these results:\n\\[\n-20 + 15i + 12i - 9i^2\n\\]\n\nNext, combine the like terms (the real parts and the imaginary parts):\n\\[\n-20 + (15i + 12i) - 9i^2 = -20 + 27i - 9i^2\n\\]\n\nRecall that \\(i^2 = -1\\). Therefore, \\(-9i^2\\) can be simplified as follows:\n\\[\n-9i^2 = -9(-1) = 9\n\\]\n\nSubstitute this back into the expression:\n\\[\n-20 + 27i + 9\n\\]\n\nCombine the real parts:\n\\[\n-20 + 9 = -11\n\\]\n\nSo, the expression simplifies to:\n\\[\n-11 + 27i\n\\]\n\nThus, the final answer is:\n{reasoning process} #### $-11 + 27i$ ✔

Figure 23. Reasoning example 6: M3PO on a mathematical task.