



# PaCoRe: Learning to Scale Test-Time Compute with Parallel Coordinated Reasoning

Jingcheng Hu<sup>1,2\*</sup>, Yinmin Zhang<sup>1</sup>, Shijie Shang<sup>1</sup>, Xiaobo Yang<sup>1,3\*</sup>, Yue Peng<sup>1</sup>, Zhewei Huang<sup>1</sup>, Hebin Zhou<sup>1</sup>, Xin Wu<sup>1</sup>, Jie Cheng<sup>1</sup>, Fanqi Wan<sup>1</sup>, Xiangwen Kong<sup>1</sup>, Chengyuan Yao<sup>1</sup>, Ailin Huang<sup>1</sup>, Hongyu Zhou<sup>1</sup>, Qi Han<sup>1</sup>, Zheng Ge<sup>1</sup>, Daxin Jiang<sup>1</sup>, Xiangyu Zhang<sup>1</sup>, Heung-Yeung Shum<sup>2</sup>

<sup>1</sup>StepFun, <sup>2</sup>Tsinghua University, <sup>3</sup>Peking University

GitHub: <https://github.com/stepfun-ai/PaCoRe>

Data: <https://huggingface.co/stepfun-ai/PaCoRe-Train-8k>

Model: <https://huggingface.co/stepfun-ai/PaCoRe-8B>

## Abstract

We introduce Parallel Coordinated Reasoning (PaCoRe), a training-and-inference framework designed to overcome a central limitation of contemporary language models: their inability to scale test-time compute (TTC) far beyond sequential reasoning under a fixed context window. PaCoRe departs from the traditional sequential paradigm by driving TTC through massive parallel exploration coordinated via a message-passing architecture in multiple rounds. Each round launches many parallel reasoning trajectories, compacts their findings into context-bounded messages, and synthesizes these messages to guide the next round and ultimately produce the final answer. Trained end-to-end with large-scale, outcome-based reinforcement learning, the model masters the synthesis abilities required by PaCoRe and scales to multi-million-token effective TTC without exceeding context limits. The approach yields strong improvements across diverse domains, and notably pushes reasoning beyond frontier systems in mathematics: an 8B model reaches 94.5% on HMMT 2025, surpassing GPT-5's 93.2% by scaling effective TTC to roughly two million tokens. We open-source model checkpoints, training data, and the full inference pipeline to accelerate follow-up work.

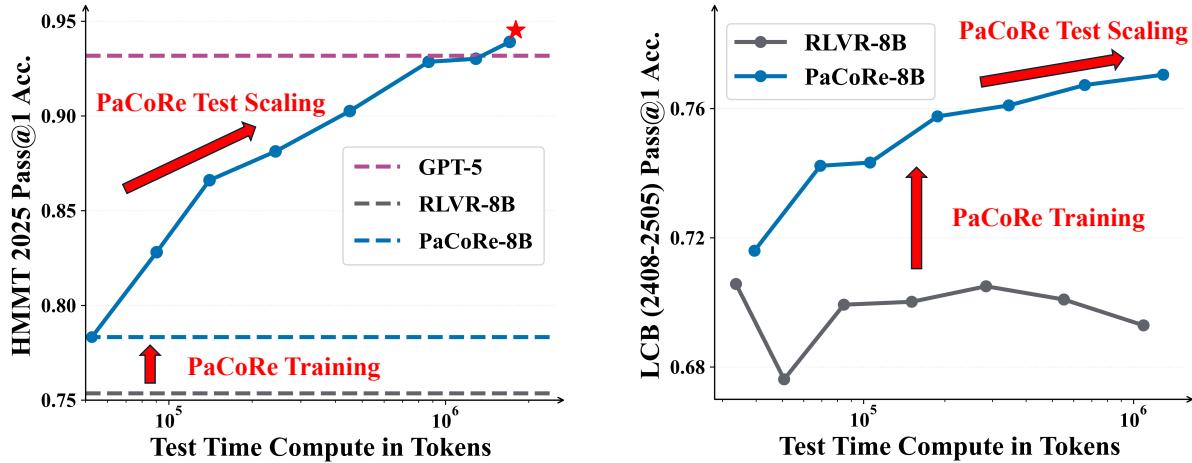


Figure 1: **Parallel Coordinated Reasoning (PaCoRe) performance.** *Left:* On HMMT 2025, PaCoRe-8B demonstrates remarkable test-time scaling by increasing both parallel trajectories and coordinated rounds, yielding steady gains and ultimately surpassing GPT-5. *Right:* On LiveCodeBench, the RLVR-8B model fails to leverage increased test-time compute, while PaCoRe-8B model effectively unlocks this synthesis capability, yielding substantial gains as the test-time compute increases.

\*Work done during internship at StepFun.

## Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
<b>2</b>	<b>Methods</b>	<b>4</b>
2.1	Inference Pipeline . . . . .	4
2.2	Training Procedure . . . . .	5
<b>3</b>	<b>Experiments</b>	<b>6</b>
3.1	Training . . . . .	7
3.2	Evaluation . . . . .	8
3.3	Analysis and Ablations . . . . .	9
<b>4</b>	<b>Related Work</b>	<b>11</b>
<b>5</b>	<b>Conclusion and Future Work</b>	<b>11</b>
<b>6</b>	<b>Acknowledgements</b>	<b>12</b>
<b>A</b>	<b>Initial Checkpoint Derivation</b>	<b>18</b>
<b>B</b>	<b>PaCoRe Training Data Preparation</b>	<b>19</b>
B.1	Math Data . . . . .	19
B.2	Competitive Code Data . . . . .	20
<b>C</b>	<b>PaCoRe Synthesis Prompt Template</b>	<b>20</b>
<b>D</b>	<b>More Ablation Studies</b>	<b>21</b>

## 1. Introduction

Long-horizon reasoning, requiring sustained exploration, cross-checking, and iterative self-correction, underpins the most demanding intellectual pursuits, from proving theorems to designing algorithms. Approaching such capabilities in AI demands substantial test-time compute (TTC)—the computation devoted to solving a single challenging instance.

A recurring lesson in the history of deep learning is that progress on hard problems often comes from scaling test-time compute through search [1, 2, 3]—expanding both in depth (sequentially) and width (in parallel)—rather than relying on a single forward pass. Yet for contemporary language models [4], a fixed context window places a hard ceiling on how much such search-driven TTC can be expressed: standard sequential reasoning [5] packs every intermediate state into a single expanding chain, strictly coupling reasoning volume to context capacity; once the window fills, the reasoning must stop.

To address this, we introduce **Parallel Coordinated Reasoning (PaCoRe)**, a framework that decouples test-time compute scaling from context limits by shifting the primary driver of inference from solely sequential depth to elevated coordinated breadth. PaCoRe iteratively coordinates parallel reasoning trajectories and compresses their insights into compact messages. The messages from the previous round are then synthesized within the model’s context window to guide subsequent exploration, and ultimately, produce the final answer.

Crucially, PaCoRe’s success hinges on shifting the model from naive aggregation and solitary problem-solving to active coordination. Vanilla reasoning models tend to rely on **simple heuristics** such as majority voting [6] on tasks with simple structure solutions, but on more diverse domains (*e.g.*, code generation and mathematical proving) they often exhibit **Reasoning Solipsism**: despite receiving rich insights from parallel branches, they ignore this context and attempt to solve the problem from scratch, wasting the test-time compute accumulated in previous rounds, as shown in Fig. 1. To overcome this, we employ large-scale, outcome-based reinforcement learning, compelling the model to master **Reasoning Synthesis**: the capacity to scrutinize parallel branches, reconcile conflicting evidence, and synthesize a unified solution that exceeds any individual trajectory.

We evaluate PaCoRe across challenging benchmarks and observe substantial gains from massively scaling TTC far beyond the limits of standard decoding. The results are especially pronounced in mathematics: on the HMMT 2025 competition benchmark, our PaCoRe-8B model achieves a score of 94.5%, surpassing the leading proprietary reasoning model, GPT-5, at 93.3%. This performance is achieved by scaling effective TTC up to nearly two million tokens per problem, orchestrated entirely within the model’s standard context window.

We summarize our primary contributions as follows:

1. We introduce **PaCoRe**, a general framework that decouples reasoning volume from model context capability by coordinating parallel reasoning, enabling multi-million token effective test-time compute.
2. We demonstrate that large-scale, outcome-based reinforcement learning is essential for inducing the synthesis capabilities required by this framework in diverse and challenging reasoning settings.
3. We release comprehensive resources including model checkpoints, training data, and the inference pipeline, to enable rigorous evaluation and foster further research within the community.

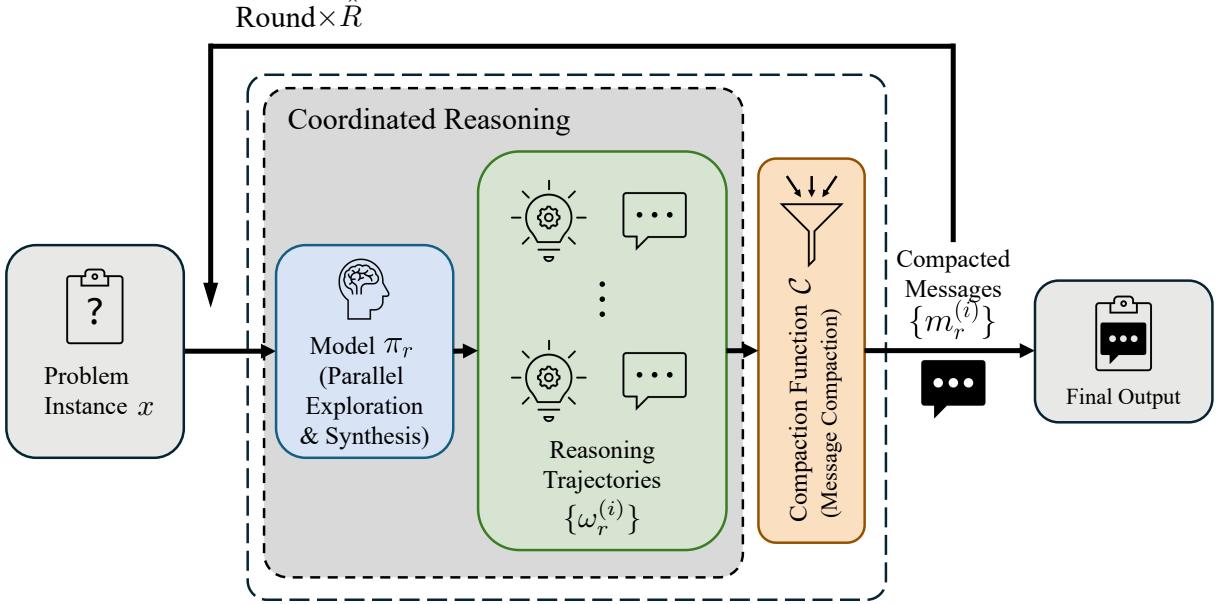


Figure 2: **Inference pipeline of PaCoRe.** Each round launches broad parallel exploration, compacts the resulting trajectories into compacted messages, and feeds these messages together with the question forward to coordinate the next round. Repeating this process  $\hat{R}$  times yields multi-million-token effective TTC while respecting fixed context limits, with the final compacted message serving as the system’s answer.

## 2. Methods

The framework comprises training and inference pipelines. The inference pipeline coordinates parallel exploration to massively scale test-time compute, growing far beyond the limits of sequential reasoning while remaining independent of context constraints. The training procedure employs large-scale, outcome-based reinforcement learning to teach the model the synthesis skills required to consolidate diverse trajectories and generate high-quality final answers.

### 2.1. Inference Pipeline

Given a problem instance  $x$ , the PaCoRe inference pipeline executes  $R$  rounds of **coordinated reasoning**. At each round  $r \in \{1, \dots, R\}$ , the system inherits a set of **compact messages**  $M_{r-1} = \{m_{r-1}^{(i)}\}_{i=1}^{K_{r-1}}$  from the previous round, and generates a new set of **reasoning trajectories**  $\Omega_r = \{\omega_r^{(i)}\}_{i=1}^{K_r}$  that explore the solution space in parallel. The final answer is simply the degenerate case where  $K_r = 1$  at the last round. For convenience, we refer to the preceding coordinated-reasoning process as consisting of  $\hat{R} = R - 1$  rounds, and refer  $\vec{K} = [K_1, \dots, K_{\hat{R}}]$  as inference trajectory configuration.

For each problem  $x$ , the model begins each round by taking the combined context  $(x, M_{r-1})$  and generating a set of parallel trajectories  $\Omega_r = \{\omega_r^{(i)}\}$ . The message set  $M_{r-1}$  provides a compacted summary of reasoning trajectories from the previous round, and each trajectory  $\omega_r^{(i)}$  contains a full chain of reasoning followed by a final conclusion. Each round then comprises two stages: (i) **Synthesis and Parallel Exploration**, where the model takes the combined context  $(x, M_{r-1})$  and produces the trajectories  $\Omega_r$ ; and (ii) **Message Compaction**, where the trajectories  $\Omega_r$  are compressed into the next-round message set  $M_r$ , allowing PaCoRe to massively scale

effective test-time compute under a fixed context window. We next detail these two stages of **coordinated reasoning**, outlining how PaCoRe constructs inputs, generates trajectories, and compacts them across rounds.

**Synthesis and Parallel Exploration.** At the beginning of round  $r$ , the system uses the prompting function  $P(x, M_{r-1})$  to serialize the problem and the compact messages, producing a structured input sequence for the model.

Once the input is constructed, the core reasoning model  $\pi_r$  is then invoked in parallel to produce  $K_r$  independent trajectories:

$$\omega_r^{(i)} \sim \pi_r(\cdot \mid P(x, M_{r-1})). \quad (1)$$

In our implementation, we employ the same model weights across all rounds for operational convenience. This parallel generation constitutes the main drive of effective TTC: even though the system can accumulate millions of tokens through parallel expansion, each round's input consumes only  $(x, M_{r-1})$ , maintaining an almost constant context cost.

**Message Compaction.** To enable multi-round coordination while respecting the fixed context window, the full trajectories in  $\Omega_r$  must be compressed into a small set of messages. We apply a compaction function:

$$M_r = C(\Omega_r), \quad (2)$$

where  $C$  transforms the set of trajectories into a new set of compact messages.

In our implementation, we leverage the structured nature of the generated trajectories. Reasoning models typically produce detailed intermediate derivations (*e.g.*, chain-of-thought or "reasoning content") followed by a final conclusion. We therefore implement  $C(\cdot)$  as a trajectory-wise extraction function  $C(\cdot)$ : it parses each  $\omega_r^{(i)} \in \Omega_r$ , retains only the final conclusion segment (forming  $m_r^{(i)} = C(\omega_r^{(i)})$ ), and discards the intermediate steps. This ensures that the input length required for synthesis remains bounded (*i.e.*,  $|x| + \sum_i |m_r^{(i)}|$  fits within the context window), even as the effective TTC—aggregate token count across all trajectories,  $\sum_r \sum_i |\omega_r^{(i)}|$ —scales vastly.

**Iterative Coordination.** The process repeats for all  $R$  rounds, with compact messages progressively refining the model's understanding of the problem. To ensure convergence, the final round uses a single trajectory ( $K_R = 1$ ), producing a final compact message

$$y = m_R^{(1)}, \quad (3)$$

which constitutes the output of the PaCoRe inference pipeline.

## 2.2. Training Procedure

The efficacy of the PaCoRe framework fundamentally depends on the core model's capacity for **synthesis**: the ability to critically evaluate diverse perspectives within the input messages  $M$ , reconcile conflicting information, and generate novel strategies that surpass the quality of any individual input.

To train this capability, we instantiate the synthesis stage of a single PaCoRe round as an episodic reinforcement learning environment, with the core model serving as the policy to be

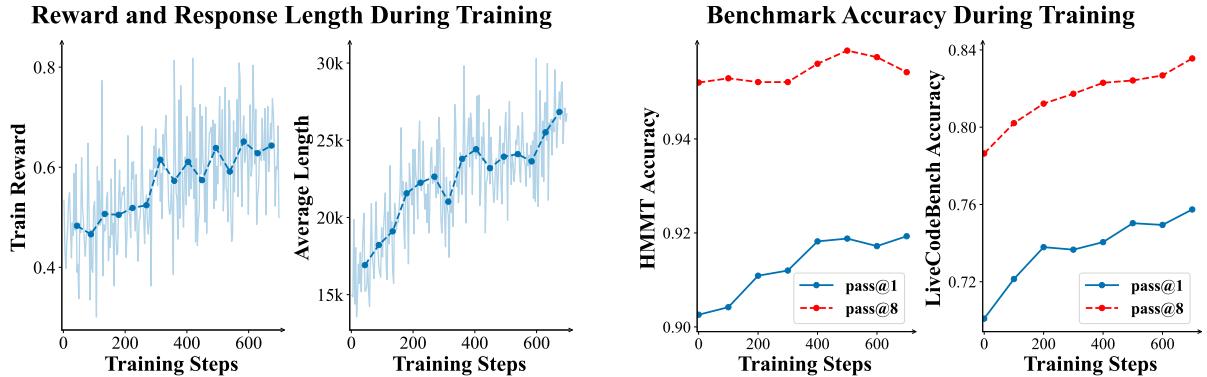


Figure 3: **PaCoRe Training dynamics.** *Left panels:* The **Training Reward** and **Response Length** steadily increase, demonstrating the training stability and effectiveness. *Right panels:* Evaluation on **HMMT 2025** and **LiveCodeBench (2408-2505)**. Performance is reported using single round coordinated reasoning in PaCoRe inference setting with  $\vec{K} = [16]$ .

optimized. In each training episode, we sample a problem  $x$  and a corresponding set of input messages  $M$  from a training distribution  $D$ :

$$(x, M) \sim D, \quad (4)$$

where  $M$  can be either generated online or drawn from a cached pool. Given the sampled pair  $(x, M)$ , the policy  $\pi_\theta$  generates a reasoning trajectory  $\omega$  from the formatted input  $P(x, M)$ , and receives a sparse terminal reward  $R(\omega) \in [0, 1]$  at the end.

We then apply large-scale, outcome-based reinforcement learning to elicit these advanced synthesis behaviors, using PPO [7] with GAE ( $\lambda = 1$ ,  $\gamma = 1$ ) [8] following the ORZ setup [9]. Training is performed on challenging domains, where each trajectory is evaluated by the correctness of its extracted final conclusion (the compacted message). In order to ensure that naive strategies like majority voting or random selection are insufficient and compel the policy to develop synthesis to succeed, we discard training instances where the average accuracy of the message set  $M$  exceeds a predefined threshold.

Compared to standard RL on static tasks with verifiable rewards [10], PaCoRe represents a significant departure. Because the input context  $M$  is composed of model-generated messages, the policy must function within an implicitly multi-agent environment rather than a fixed single-agent setting. Although the outcome-based reward aligns with conventional RL formulations for chain-of-thought optimization, PaCoRe demands higher-level synthesis and coordination across agents' outputs, introducing opportunities to examine emergent collective behaviors.

### 3. Experiments

In this section, we present comprehensive experimental results and analysis of our PaCoRe framework. We begin by detailing the training protocol used to develop our PaCoRe-8B model, followed by in-depth analysis of training results. We then present the main evaluation results, comparing PaCoRe against frontier reasoning models on challenging benchmarks. Finally, we conduct analysis and ablation studies to validate key design choices and investigate the properties of the learned synthesis capabilities.

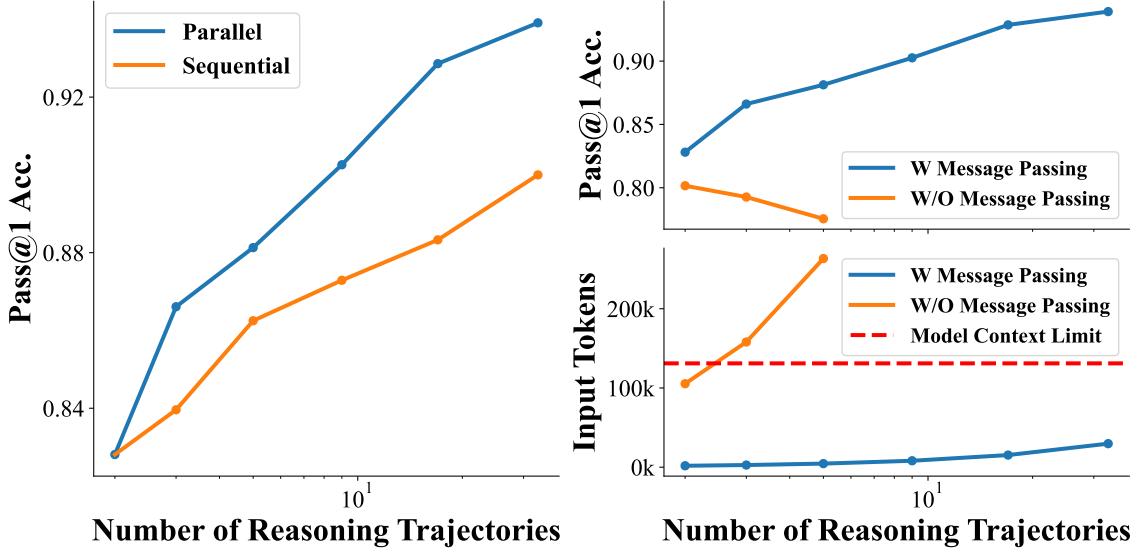


Figure 4: **Ablation of parallel reasoning and message passing.** *Left:* **Parallel** scaling ( $\vec{K} = [N, ]$ ) utilizes test-time compute more effectively than **Sequential** scaling ( $\vec{K} = [1, \dots, 1]$ ). *Right:* **Message Passing** is essential for test-time scaling. Without compaction ("W/O Message Passing"), performance degrades as test time scales and fundamentally limited by model context length, whereas standard PaCoRe ("W Message Passing") scales unboundedly and robustly. Pass@1 scores are evaluated on HMMT 2025.

### 3.1. Training

**Training Recipe.** Our PaCoRe-8B model is initialized from an internal, reasoning-oriented post-trained version of Qwen3-8B-Base [11]. We use this initial model to generate a message cache pool of size 24 for every problem in the training dataset. During training, we randomly sample the message set  $M$  from this pool, with  $|M| \sim U(16, 24)$ . We train the model using PPO with GAE ( $\lambda = 1, \gamma = 1$ ) in a strict on-policy setting. The training configuration involves sampling 16 data instances per batch, with 64 responses generated for each instance. We use a maximum sequence length of 131,072, with both generation temperature and top-p set to 1.0. We leverage competition-level mathematics and coding tasks as the primary substrate to cultivate model capabilities. To compel the model to learn synthesis, the training data distribution is refined over two stages. In **Stage 1** (250 iterations), we filter the data based on the accuracy of the message set, specifically targeting scenarios where naive strategies like random selection or majority voting are insufficient. For mathematics, we require  $0 < \text{mean(message\_acc)} < 9/24$ , and for code,  $0 < \text{mean(message\_acc)} < 15/24$ . Additionally, we apply rule-based and model-based filtering to ensure the quality of the math problems. In **Stage 2** (450 additional iterations, 700 total), we further refine the distribution. We use an intermediate Stage 1 checkpoint to evaluate synthesis accuracy on the Stage 1 dataset, retaining only instances where  $0 < \text{synthesis\_acc} < 1$ . The final checkpoint after 700 iterations is used for all subsequent evaluations.

**Training Results.** We present the training dynamics by tracking training rewards, response lengths, and performance on key benchmarks. As illustrated in Figure 3, both the training reward and response length steadily increase throughout the process. Correspondingly, the performance on both HMMT 2025 and LiveCodeBench (2408-2505) improves significantly. In these evaluations, the PaCoRe inference setting is  $\vec{K} = [16, ]$ . In order to improve experimental

Table 1: Benchmark performance and TTC spent per problem instance of different approaches. For each benchmark, we report accuracy together with total TTC (in thousands). For *Low*, *Medium*, and *High*, we apply the inference trajectory configuration as  $\vec{K} = [4]$ , [16], and [32, 4] separately.

Benchmark	HMMT 2025	LiveCodeBench	HLE <sub>text</sub>	MultiChallenge
GPT-5	93.2 (16k)	<b>83.5</b> (13k)	<b>26.0</b> (14k)	<b>71.1</b> (5.0k)
Qwen3-235B-A22B-Thinking-2507	82.3 (32k)	74.5 (21k)	18.2 (23k)	60.3 (1.6k)
GLM-4.6	88.7 (25k)	79.5 (19k)	17.2 (21k)	54.9 (2.2k)
DeepSeek-v3.1-Terminus	86.1 (20k)	74.9 (11k)	19.3 (18k)	54.4 (1.1k)
Kimi-K2-Thinking	86.5 (33k)	79.2 (25k)	23.9 (29k)	66.4 (1.7k)
RLVR-8B	75.4 (48k)	70.6 (34k)	9.3 (35k)	33.3 (1.7k)
PaCoRe-8B (low)	88.2 (243k)	75.8 (188k)	13.0 (196k)	41.8 (13k)
PaCoRe-8B (medium)	92.9 (869k)	76.7 (659k)	14.6 (694k)	45.7 (45k)
PaCoRe-8B (high)	<b>94.5</b> (1796k)	78.2 (1391k)	16.2 (1451k)	47.0 (95k)

efficiency, we utilize cached responses from the RLVR-8B model for each problem instance to directly seed the current checkpoint for the next round generation. All these results indicate stable and effective learning.

### 3.2. Evaluation

**Evaluation Setup.** We evaluate PaCoRe on HMMT 2025 [12], LiveCodeBench (2408-2505) [13], Humanity’s Last Exam (text subset) [14], and MultiChallenge [15]. These serve as representative benchmarks for mathematics, coding, science, and open-ended generation, respectively. We compare our approach against several leading reasoning models, including GPT-5 [16], Kimi-K2-Thinking [17], DeepSeek-V3.1-Terminus [18], GLM-4.6 [19], and Qwen3-235B-A22B-Thinking-2507 [20]. For GPT-5, we specifically utilize the “high” reasoning effort configuration. We report pass@1 accuracy for all approaches based on multiple independent generations per problem: 64 for HMMT, 8 for both LiveCodeBench and MultiChallenge, and 1 for Humanity’s Last Exam. We employ the final PaCoRe-8B checkpoint for all evaluations. We employ the final PaCoRe-8B checkpoint for all evaluations, maintaining a maximum generation sequence length of 131,072 tokens with a temperature and top-p of 1.0. Additionally, for both PaCoRe and RLVR evaluations, we apply the YARN [21] method with a scaling factor of 2.0 and do not modify the attention temperature during inference. We evaluate our PaCoRe inference pipeline under three levels of coordinated reasoning effort: *Low* ( $\vec{K} = [4]$ ), *Medium* ( $\vec{K} = [16]$ ), and *High* ( $\vec{K} = [32, 4]$ ). To optimize experimental efficiency without compromising validity, we utilize a caching strategy for the first round ( $r = 1$ ) for PaCoRe evaluation: we pre-generate a pool of 512 trajectories for each problem, from which the model randomly samples  $K_1$  items to seed the next round. We empirically validated that this method yields results equivalent to generating all trajectories from scratch.

**Evaluation Results.** The main results are summarized in Table 1. Across all test-time effort settings, PaCoRe-8B consistently outperforms the starting checkpoint, RLVR-8B model, on every benchmark. Even in the *Low* setting, PaCoRe-8B demonstrates robust capabilities, achieving 88.2% on HMMT2025 and 75.8% on LiveCodeBench. This configuration significantly outperforms the standard RLVR-8B baseline and surpasses considerably larger models, including Qwen3-235B-A22B-Thinking-2507 and DeepSeek-V3.1-Terminus. With further test-time scaling in the *High* setting, PaCoRe-8B reaches a remarkable 94.5% on HMMT2025, surpassing the

Table 2: **General Effectiveness of PaCoRe Data.** RLVR trained on PaCoRe data using an off-policy PPO algorithm, with only 50 iterations and 4 mini-batches per iteration, achieves significant gains on both AIME 2025 and LiveCodeBench.

Model	AIME 2025	LiveCodeBench
Ours-SFT-Qwen3-30B-A3B	81.4%	66.0%
+ RLVR with PaCoRe Data	<b>83.2%</b>	<b>74.0%</b>

Table 3: Ablation of how the message-set size  $|M|$  is determined during training.

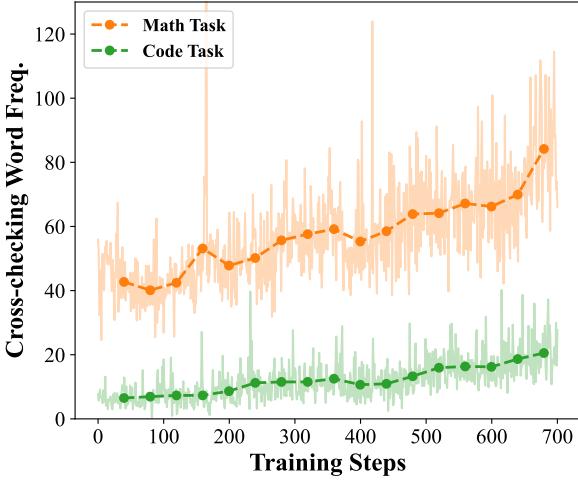
Setting	HMMT 2025	LiveCodeBench
$ M  = 4$	83.6	63.3
$ M  = 8$	83.7	64.4
$ M  = 16$	84.2	64.1
$ M  \sim U(1, 16)$	84.3	64.3
$ M  \sim U(8, 16)$	<b>85.2</b>	<b>65.1</b>

leading proprietary model, GPT-5 at 93.2%, by scaling effective TTC to approximately 2 million tokens. On LiveCodeBench (2408-2505), PaCoRe-8B achieves a strong 78.2%, remaining competitive with frontier models such as GLM-4.6 and Kimi-K2-Thinking. These results underscore the efficacy of parallel coordinated reasoning, demonstrating that scaling test-time compute enables an 8B parameter model to outperform state-of-the-art systems in some of the most complex reasoning tasks.

### 3.3. Analysis and Ablations

**Ablation of Core Design Principles.** We investigate the importance of the two central components of PaCoRe: parallel exploration and message passing. First, we compare parallel scaling (PaCoRe with  $\vec{K} = [N, ]$ ) against sequential scaling (equivalent to  $\vec{K} = [1, \dots, 1]$ ) under the same total number of generated reasoning trajectories. As shown in Figure 4 (Left), parallel coordinated reasoning utilizes TTC much more effectively than purely sequential approach. Second, we examine the role of message passing. We compare standard PaCoRe (W Message Passing) with a variant where compaction is disabled (W/O Message Passing), feeding the full trajectories into the next round context window. Figure 4 (Right) shows that without compaction, performance degrades as TTC scales and is fundamentally limited by the context length. In contrast, PaCoRe, with message passing, scales robustly and without bound.

**General Effectiveness of PaCoRe Data.** Beyond the framework itself, we discovered that the training corpus curated for PaCoRe constitutes a learning resource of exceptional density. As shown in Table 2, we empirically observe that employing our released dataset as the primary substrate for standard RLVR yields a robust performance boost. The RLVR stage uses high-quality, carefully curated training data and applies an off-policy PPO algorithm with GAE ( $\gamma = 1, \lambda = 1$ ) for only 50 iterations and 4 mini-batches per iteration on ours-SFT-Qwen3-30B-A3B. Despite this minimal compute budget, RLVR delivers substantial gains over the SFT model on both AIME25 and LiveCodeBench. This indicates that our problem set—carefully filtered to demand genuine synthesis—serves as a highly effective catalyst for training strong reasoning models in general.



**Figure 5: Evolution of synthesis-related linguistic features in model outputs across training steps.** We plot the frequency of cross-checking words (including ‘reference’, ‘参考’, ‘Ref <number>’, ‘ref <number>’) in generated solutions for **Math** and **Code** tasks during PaCoRe training. Training elicits and magnifies this synthesis capability in both domains. Notably, the model rarely performs cross-checking on Code tasks initially, corroborating the poor test time scaling in Code without PaCoRe training shown in Fig. 1.

**Table 4: Performance on SWE-Verified.** We evaluate using the Agentless framework with a 120k token context window ( $T = 1.0, p = 1.0$ ). PaCoRe-8B (low) markedly surpasses the RLVR-8B baseline, indicating generalization to agentic coding scenarios without task-specific tuning.

Model	SWE Verified
RLVR-8B	29.8%
PaCoRe-8B (low)	34.0%

**Ablation on Message Sampling Strategy in Training.** We examine the effect of the sampled message set size  $|M|$  during training. We compare training with fixed sizes versus uniformly sampled sizes. As shown in Table 3, training with diverse and relatively large message sets ( $|M| \sim U(8, 16)$ ) yields the best performance, with all evaluated under the same inference configuration of  $\vec{K} = [16, ]$ . This strategy familiarizes the model with the dynamics of massive parallel coordination and facilitates superior test-time scaling properties.

**Generalization Across Domains.** To assess cross-domain generalization beyond mathematical and logical reasoning benchmarks, we evaluate PaCoRe on realistic, repository-level software engineering tasks using the SWE-Verified [22] and MultiChallenge [15] benchmark. We note that the initialization checkpoint (RLVR-8B) underwent no specialized post-training for software engineering and multi turn dialogue tasks; thus, this study assesses the intrinsic transfer of reasoning capabilities rather than fully optimized performance. Leveraging the Agentless [23] framework, we compare a compute-efficient variant, PaCoRe-8B (low), against the RLVR-8B baseline. Inference is configured with a maximum sequence length of 120,000 tokens, temperature of 1.0 and topp of 1.0. As shown in Table 4, PaCoRe-8B (low) achieves a resolve rate of 34.0%, substantially outperforming the baseline (26.2%). Furthermore, PaCoRe again exhibits strong generalization, with PaCoRe-8B (high) improving the RLVR-8B baseline from

33.3% to 48.0% under higher TTC budgets as shown in Table 1

**Evolution of Synthesis Capabilities.** We probe the mechanism underlying PaCoRe by tracking the prevalence of "cross-checking" linguistic markers throughout training. As shown in Figure 5, outcome-based reinforcement learning drives a steady rise in these behaviors across both domains. Crucially, the Code trajectory starts with near zero cross-checking behaviors, corroborating the poor test-time scaling of untrained models observed in Figure 1. This evolution confirms that training fundamentally alters reasoning dynamics, transforming the model into a coordinated reasoner capable of rigorous synthesis.

## 4. Related Work

Classical approaches to sequentially scaling TTC, such as Chain-of-Thought (CoT) [5], pack intermediate steps into a single expanding chain. Even when boosted by large-scale reinforcement learning [24, 25], these methods ultimately saturate the context window, placing a strict ceiling on achievable TTC.

A shift towards parallel scaling has shown immense potential. Early methods [26, 3, 27] use simple or task-specific rules to integrate language models with extensive parallel search, achieving strong performance in specific domains. However, the coordination in these systems often hinges on intricate, task-specific scaffolds or domain-specific properties, hindering their general applicability.

This highlights the need for a general framework to coordinate massive parallel exploration within model context limits. Attempts at such coordination [28, 29, 30, 31, 32] synthesize parallel outputs but lacked mechanisms for message passing or context management, thus facing the same context limitations. Alternative approaches focused on context management [33, 34] use compression rules but remain rooted in the sequential paradigm, limiting TTC scaling efficiency. AggLM [6] focuses on learning aggregation strategies that surpass majority voting and reward-model ranking, while we introduce parallel coordinated reasoning with context compaction to scale test-time compute beyond context limits. Thus, scaling trajectory configurations and rounds of coordinated reasoning yields substantial gains, outperforming AggLM and even GPT-5 in the mathematics domain.

Recently, frontier proprietary systems [35, 36, 16] have demonstrated exceptional reasoning performance, reportedly utilizing massive parallel TTC under current model architectures. While these systems showcase the power of this paradigm, their undisclosed technical details hinder broader scientific progress. PaCoRe bridges this gap by providing a transparent, effective, and open-sourced framework for parallel coordinated reasoning, accelerating future research in this direction.

## 5. Conclusion and Future Work

In this work, we introduce Parallel Coordinated Reasoning (PaCoRe), a test-time scaling framework that enables massive test-time compute (TTC) without exceeding the context window constraints of contemporary language models. PaCoRe addresses the limitations of sequential reasoning by shifting the primary driver of inference to coordinated parallel breadth. It operates via a message-passing architecture that iteratively launches many parallel reasoning trajectories, compacts their findings into context-bounded messages, and synthesizes these messages to

guide subsequent exploration. Trained end-to-end with large-scale, outcome-based reinforcement learning, the model develops the synthesis capabilities required for effective coordination. Our empirical results demonstrate significant improvements across diverse domains. Notably, our PaCoRe-8B model achieves 94.5% on HMMT 2025, surpassing GPT-5 by scaling effective TTC to roughly two million tokens, all orchestrated within a standard context window.

The development of PaCoRe opens several promising avenues for future research:

- **Scaling to Extremes:** We plan to further scaling PaCoRe on model sizes, task domains and further on test time compute: applying this framework to more capable foundation models, extending the application to agentic tasks and multi-modal understanding, and expanding both the breadth (number of parallel trajectories) and depth (number of coordination rounds) to solve some extremely hard problems.
- **Boosting Token Intelligence Density:** While we currently scale by volume, we aim to maximize the utility of every unit of compute spent. This involves enabling more efficient parallel exploration through better organization, cooperation, and division of labor among trajectories.
- **Emergent Multi-Agent Intelligence:** We are interested in exploring the joint training of both the synthesis policy and the message-passing mechanism, laying minimal yet rich cooperative multi-agent learning environment, offering a valuable playground for studying emergent communication, self-organization, and collective intelligence.
- **Ouroboros for Pre- and Post-Training:** We intend to investigate the development of advanced synthetic data generation techniques with PaCoRe pipeline to improve both current pretraining and post-training processes.

## 6. Acknowledgements

We express our gratitude to Song Yuan, Wuxun Xie, Mingliang Li, and Bojun Wang for their support regarding inference, and to Xing Chen, Yuanwei Lu, Changyi Wan and Yu Zhou for their assistance with training. We also thank Shaoliang Pang, Changxin Miao, Xu Zhao, Wei Zhang, Zidong Yang, Junzhe Lin, Yuxiang Yang, Chen Xu, Xin Li and Bin Wang for their help with operational issues and platform support. We extend our thanks to Xiaoxiao Ren, Zhiguo Huang, and Kang An for their assistance with data management support. Special thanks go to Liang Zhao, Jianjian Sun, Zejia Weng, and JingJing Xie for their helpful discussions throughout this project. We acknowledge the dedicated support of the Infrastructure Team and the Data Technical Team, as well as the valuable feedback from our other colleagues at StepFun and Tsinghua University. This work was supported by computing resources and infrastructure provided by StepFun.

## References

- [1] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.
- [2] Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de Masson d’Autume, Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey Cherepanov, James Molloy, Daniel J. Mankowitz, Esme Sutherland Robson, Pushmeet Kohli, Nando de Freitas, Koray Kavukcuoglu, and Oriol Vinyals. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097, December 2022.
- [3] Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482, 2024.
- [4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.
- [5] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.
- [6] Wenting Zhao, Pranjal Aggarwal, Swarnadeep Saha, Asli Celikyilmaz, Jason Weston, and Ilia Kulikov. The majority is not always right: RL training for solution aggregation, 2025.
- [7] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017.
- [8] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation, 2018.
- [9] Jingcheng Hu, Yinmin Zhang, Qi Han, Dixin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-Reasoner-Zero: An open source approach to scaling up reinforcement learning on the base model. *arXiv preprint arXiv:2503.24290*, 2025.
- [10] Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tulu 3: Pushing frontiers in open language model post-training, 2025.
- [11] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang,

Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025.

- [12] HMMT. Hmmt 2025 feb., 2025.
- [13] Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*, 2024.
- [14] HLE Team. Humanity's last exam, 2025.
- [15] Ved Sirdeshmukh, Kaustubh Deshpande, Johannes Mols, Lifeng Jin, Ed-Yeremai Cardona, Dean Lee, Jeremy Kritz, Willow Primack, Summer Yue, and Chen Xing. Multichallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier llms, 2025.
- [16] OpenAI. Introducing gpt-5, 2025.
- [17] Moonshot AI. Introducing kimi k2 thinking, 2025.
- [18] DeepSeek. Deepseek-v3.1-terminus, 2025.
- [19] Z.ai. Glm-4.6: Advanced agentic, reasoning and coding capabilities, 2025.
- [20] Qwen Team. Qwen3-235b-a22b-thinking-2507, 2025.
- [21] Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models, 2023.
- [22] OpenAI. Introducing SWE-bench verified we're releasing a human-validated subset of swe-bench that more, 2024.
- [23] Chunqiu Steven Xia, Yinlin Deng, Soren Dunn, and Lingming Zhang. Agentless: Demystifying llm-based software engineering agents, 2024.
- [24] OpenAI. Learning to reason with llms, 2024.
- [25] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025.
- [26] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2023.
- [27] Ahmed El-Kishky, Alexander Wei, Andre Saraiva, Borys Minaiev, Daniel Selsam, David Dohan, Francis Song, Hunter Lightman, Ignasi Clavera, Jakub Pachocki, Jerry Tworek, Lorenz Kuhn, Lukasz Kaiser, Mark Chen, Max Schwarzer, Mostafa Rohaninejad, Nat McAleese, o3 contributors, Oleg Mürk, Rhythm Garg, Rui Shu, Szymon Sidor, Vineet Kosaraju, and Wenda Zhou. Competitive programming with large reasoning models, 2025.
- [28] Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. Universal self-consistency for large language model generation, 2023.

- [29] Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities, 2024.
- [30] Hao Wen, Yifan Su, Feifei Zhang, Yunxin Liu, Yunhao Liu, Ya-Qin Zhang, and Yuanchun Li. Parathinker: Native parallel thinking as a new paradigm to scale llm test-time compute, 2025.
- [31] Tong Zheng, Hongming Zhang, Wenhao Yu, Xiaoyang Wang, Runpeng Dai, Rui Liu, Huiwen Bao, Chengsong Huang, Heng Huang, and Dong Yu. Parallel-r1: Towards parallel thinking via reinforcement learning, 2025.
- [32] Siddarth Venkatraman, Vineet Jain, Sarthak Mittal, Vedant Shah, Johan Obando-Ceron, Yoshua Bengio, Brian R. Bartoldson, Bhavya Kailkhura, Guillaume Lajoie, Glen Berseth, Nikolay Malkin, and Moksh Jain. Recursive self-aggregation unlocks deep thinking in large language models, 2025.
- [33] Yuchen Yan, Yongliang Shen, Yang Liu, Jin Jiang, Mengdi Zhang, Jian Shao, and Yueting Zhuang. Infitythink: Breaking the length limits of long-context reasoning in large language models, 2025.
- [34] Milad Aghajohari, Kamran Chitsaz, Amirhossein Kazemnejad, Sarath Chandar, Alessandro Sordoni, Aaron Courville, and Siva Reddy. The markovian thinker: Architecture-agnostic linear scaling of reasoning, 2025.
- [35] xAI. Grok 4, 2025.
- [36] Google DeepMind. Gemini 2.5 deep think model card, 2025.
- [37] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.
- [38] Feng Yao, Liyuan Liu, Dinghuai Zhang, Chengyu Dong, Jingbo Shang, and Jianfeng Gao. Your efficient rl framework secretly brings you off-policy rl training, August 2025.
- [39] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Weinan Dai, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale, 2025.
- [40] Etash Guha et al. Openthoughts: Data recipes for reasoning models. *arXiv preprint arXiv:2506.04178*, 2025.
- [41] Jia Li et al. Numinamath. <https://huggingface.co/datasets/AI-MO/NuminaMath-CoT>, 2024. Technical report and dataset card.
- [42] Alon Albalak et al. Big-math: A large-scale, high-quality math dataset for reinforcement learning in language models. *arXiv preprint arXiv:2502.17387*, 2025.
- [43] Arindam Mitra, Hamed Khanpour, Corby Rosset, and Ahmed Awadallah. Orca-math: Unlocking the potential of slms in grade school math. *arXiv preprint arXiv:2402.14830*, 2024.

- [44] aslawliet. Olympiads. <https://huggingface.co/datasets/aslawliet/olympiads>, 2024. Hugging Face dataset.
- [45] aslawliet. Cn-k12. <https://huggingface.co/datasets/aslawliet/cn-k12>, 2024. Hugging Face dataset of Chinese K-12 math problems.
- [46] Open-R1 Team. Openr1-math-220k. <https://huggingface.co/datasets/open-r1/OpenR1-Math-220k>, 2025. Open-source distilled math reasoning dataset.
- [47] LIMO Authors. Less is more for reasoning: Semi-parametric math reasoners. *arXiv preprint arXiv:2502.03387*, 2025.
- [48] Niklas Muennighoff et al. s1: Simple test-time scaling. <https://arxiv.org/abs/2501.19393>, 2025.
- [49] X. He et al. Deepmath-103k: A large-scale, challenging math qa benchmark. *arXiv preprint arXiv:2504.11456*, 2025.
- [50] Aime problems and solutions. [https://artofproblemsolving.com/wiki/index.php/AIME\\_Problems\\_and\\_Solutions](https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions). Accessed: 2025-12-07.
- [51] Harvard–mit mathematics tournament (hmmt). <https://www.hmmt.org/>. Accessed: 2025-12-07.
- [52] Stanford math tournament (smt). <https://www.stanfordmathtournament.org/>. Accessed: 2025-12-07.
- [53] Carnegie mellon informatics and mathematics competition (cmimc). <https://cmimc.math.cmu.edu/>. Accessed: 2025-12-07.
- [54] Brown university math olympiad (brumo). <https://www.brumo.org/>. Accessed: 2025-12-07.
- [55] Berkeley math tournament (bmt). <https://berkeley.mt/>. Accessed: 2025-12-07.
- [56] Caltech harvey mudd math competition (chmmc). <http://chmmc.caltech.edu/problems.html>. Accessed: 2025-12-07.
- [57] Duke math meet (dmm). <https://dukemathmeet.org/>. Accessed: 2025-12-07.
- [58] Minnesota youth math outreach (mnymo). <https://www.mnyouthmathoutreach.org/>. Accessed: 2025-12-07.
- [59] Princeton university mathematics competition (pumac). <https://jason-shi-f9dm.squarespace.com/>. Accessed: 2025-12-07.
- [60] Math prize for girls. <https://mathprize.atfoundation.org/>. Accessed: 2025-12-07.
- [61] Mislav Balunović, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. MathArena: Evaluating LLMs on uncontaminated math competitions, February 2025.
- [62] Rongao Li, Jie Fu, Bo-Wen Zhang, Tao Huang, Zhihong Sun, Chen Lyu, Guang Liu, Zhi Jin, and Ge Li. Taco: Topics in algorithmic code generation dataset, 2023.
- [63] Yunjie Ji, Xiaoyu Tian, Sitong Zhao, Haotian Wang, Shuaiting Chen, Yiping Peng, Han Zhao, and Xiangang Li. Am-thinking-v1: Advancing the frontier of reasoning at 32b scale, 2025.

- [64] Michael Luo, Sijun Tan, Roy Huang, Ameen Patel, Alpay Ariyak, Qingyang Wu, Xiaoxiang Shi, Rachel Xin, Colin Cai, Maurice Weber, Ce Zhang, Li Erran Li, Raluca Ada Popa, and Ion Stoica. Deepcoder: A fully open-source 14b coder at o3-mini level. <https://pretty-radio-b75.notion.site/DeepCoder-A-Fully-Open-Source-14B-Coder-at-O3-Mini-Level-1cf81902c14680b3bee5eb349a512a51>, 2025. Notion Blog.
- [65] Zihan Wang, Siyao Liu, Yang Sun, Hongyan Li, and Kai Shen. Codecontests+: High-quality test case generation for competitive programming, 2025.

In this Appendix, we provide more elaboration on the implementation details, experiment results, and qualitative results.

## A. Initial Checkpoint Derivation

To facilitate effective test-time scaling, we first implement a reasoning-oriented post-training pipeline for Qwen3-8B-Base, comprising Supervised Fine-Tuning (SFT) and Reinforcement Learning with Verifiable Rewards (RLVR).

**Reasoning-Oriented SFT.** We collect millions of prompts from the open-source community, spanning diverse domains including mathematics, coding, science, software engineering, tool use, logical reasoning, and creative writing. Leveraging these prompts, we distill data from multiple frontier models. Following a rigorous correctness verification process, we curate a dataset comprising 10.4M samples, totaling 61.1B tokens.

To ensure data quality and integrity, we apply two pipes to further filter the dataset. First, we use predefined rules to eliminate low-quality data with degenerate patterns, such as infinite repetition. Second, we conduct comprehensive benchmark decontamination to prevent leakage. This involves both exact matching (with digit masking to catch questions featuring only numerical modifications) and  $N$ -gram matching ( $N = 64$ ). This two-pipe filtration process yields a refined dataset of 10.2M samples, totaling 59.5B tokens.

We perform SFT on Qwen3-8B-Base using the refined dataset. The training is configured with a maximum sequence length of 64k and a global batch size of 64. We employ a cosine learning rate scheduler with a 200-step warmup phase, where the learning rate peaks at  $1 \times 10^{-4}$  and anneals to a final value of  $1 \times 10^{-5}$ . We implement domain-specific sampling weights for the dataloader, which translate to varying epochs for different domains. In total, the model is trained over  $\sim 165$ B tokens.

**Reasoning-Oriented RLVR.** Similarly, we source  $\sim 500$ k prompts paired with ground truth from the open-source community, which encompass domains including mathematics, coding, science, logical reasoning, and instruction following. We implement a robust internal RL framework utilizing vLLM for inference and Megatron for training. During each step, we sample 256 prompts and generate 16 responses per prompt (max length 64k). To address long-tail inference latency, we enable the partial rollout strategy [37].

For non-coding domains, we employ LLM-as-judge (gpt-oss-120b<sup>1</sup>) to verify response consistency against the ground truth. For coding tasks, we utilize sandboxes to validate code execution against test cases with soft reward. We omit KL divergence constraints (both as reward penalties and loss terms) and entropy loss during training. We use an off-policy PPO algorithm with GAE ( $\gamma = 1, \lambda = 1$ ), while omitting importance sampling. Samples are split into 4 mini-batches per iteration. We set the actor and critic learning rates to  $2 \times 10^{-6}$  and  $5 \times 10^{-6}$ , respectively. Furthermore, we apply the Truncated Importance Sampling ratio (threshold  $C = 8$ ) proposed by Yao et al. [38] to mitigate training-inference inconsistency issues. The entire RL phase consists of 200 training iterations.

---

<sup>1</sup><https://huggingface.co/openai/gpt-oss-120b>

## B. PaCoRe Training Data Preparation

### B.1. Math Data

**Math Competition and Open Source Data.** We aggregate math problems from a mix of open-source datasets and competition archives. On the dataset side, we include open source datasets from [39, 9, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49]. On the competition side, we collect all available historical problems (including official solutions where available) from AIME, HMMT, SMT, CMIMC, BRUMO, BMT, CHMMC, DMM, MNYMO, PUMAC, and Math Prize for Girls, using their official online archives [50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60]. We exclude AIME 2024/2025, HMMT Feb 2025, BRUMO 2025, SMT 2025, CMIMC 2025, and HMMT Nov 2025 from the training pool because these contests are used as part of the MathArena benchmark [61].

**Synthetic Data.** Beyond mined data, we add a synthetic component to strengthen large-integer arithmetic, a basic skill that underlies many competition-style solutions and naturally produces diverse numeric answers. We generate 13k synthetic large-integer arithmetic problems using a small set of simple hand-written text templates. For each problem, we independently sample integers  $A$  and  $B$  uniformly between  $10^{11}$  and  $10^{13}$ , randomly choose one of four operation types—addition, subtraction, multiplication, or modular exponentiation modulo a fixed large prime—and substitute  $A$ ,  $B$ , and the chosen operation into a template to form the natural-language question. We compute the corresponding exact integer answer with standard arithmetic and record it as the ground-truth label. By restricting to integer-valued outputs, each synthetic problem has a single unambiguous label and can be checked easily.

**Data Quality Control.** We apply several stages of quality control to ensure that the math data are well-posed and rule-checkable. First, we run deterministic rule-based filters that remove problems with embedded images or external links; multi-part questions or prompts that request multiple final answers; and items whose solutions depend on vague prose or open-ended discussion rather than a single numerical or algebraic target. Next, in-house math experts spot-check samples and mark common failure types (incorrect official solutions, internally inconsistent statements, or ambiguous wording). Based on these annotations, we construct a 100-sample evaluation set with a 1:1 mix of valid and invalid QA pairs and iteratively refine a prompt for an LLM-based judge. To accommodate the strict formatting requirements of our verifiable reward environment (e.g., single-part questions with a unique final answer, as in math-verify-style checkers), we use gpt-oss-120b with this prompt in a 4-pass unanimity scheme, keeping only items that receive 4/4 positive votes on the QA-validity classification. We tune the prompt until the F1 score on the evaluation set saturates, and in practice this configuration reliably filters out most ill-posed or non-checkable problems. The filtered open-source datasets and competition problems are then combined with the synthetic arithmetic problems described above. We summarize the problem counts from each source after rule-based and model-based filtering in Table 5. As discussed in Section 3.1, we additionally apply an accuracy-based filter using a strong proposer model to select problems that are neither trivial nor degenerate; we treat this as part of curriculum design rather than basic data cleaning.

Table 5: Problem counts by source after rule-based and model-based filtering.

Source	Before Filtering	Stage 1	Stage 2
Open-source datasets	695k	0.7k	0.5k
Competition archives	8.4k	0.2k	0.3k
Synthetic arithmetic	13.4k	0.8k	1.6k
Total	717k	1.7k	2.4k

## B.2. Competitive Code Data

**Competitive Programming Contest and Open Source Data.** About 29k problems are gathered from competitive programming problem sources (for example, subsets of TACO [62], USACO<sup>2</sup>, etc) and undergo a rigorous cleaning process. The main metrics in validation are format checks in problem statements, number of test cases, and full judging of provided code submissions. We use `testlib`<sup>3</sup> library to help judge a submission code. We apply an auto-matching method on test case outputs of problems to decide a pre-defined `testlib` checker to judge. We make some modification to the checker to handling broader cases, including change ‘icmp’ function to support 64-bit integer comparision. For problems requiring a special-judge checker, we use LLMs to generate it, or simply skip this problem. There are about 14k additional problems sampled from recent open-source competitive programming datasets, including `am-thinking-v1` [63] and `deepcoder` [64]. These problems are filtered with deduplication and validation processes.

**Synthetic Data.** To address the shortage of test cases in open-source datasets, we employ a generator-validator pipeline inspired by `CodeContests+` [65]. This process utilizes LLMs to produce new test cases, which are subsequently verified against ground-truth solutions and incorrect submissions. In RL experiments, we found that a fine-grained reward function based on test case pass rates significantly outperforms a simple binary reward. Our final code data source comprises approximately 5k problems curated from `CodeForces`<sup>4</sup>.

## C. PaCoRe Synthesis Prompt Template

In the *Synthesis and Parallel Exploration* stage of the PaCoRe inference pipeline, we map the problem instance  $x$  and the set of compact messages  $M$  into a structured natural language input using a prompting function  $P(x, M)$ . The exact serialization template is provided in Table 6. By framing the compact messages as “Reference Responses,” we explicitly encourage the model to critically evaluate and synthesize the diverse perspectives accumulated from the previous round.

While the method is described in Section 3.1 in the context of a single-turn problem, this framework is naturally extensible to multi-turn, interactive environments (e.g., dialogues or agentic tool-use loops). In such settings, the pipeline is invoked at any decision point where the model must generate an action. The “Original Problem” slot in the template is populated by the most recent observation or user message, while the preceding interaction history (previous turns, tool outputs) remains intact in the context. This allows PaCoRe to maintain exactly same

<sup>2</sup><https://usaco.org/index.php?page=training>

<sup>3</sup><https://github.com/MikeMirzayanov/testlib>

<sup>4</sup><https://codeforces.com>

interface as standard chat or reasoning models, ensuring seamless compatibility with existing ecosystem.

---

You are given a problem and a list of reference responses. Your job is to analyze these references and provide your own response.

Original Problem:

`{% original_prompt %}`

Reference Responses:

```
{% for response in ref_responses %}
Reference {{ loop.index }}:
{{ response }}
{% endfor %}
```

---

Now, based on the original problem and reference responses above, please provide your own comprehensive solution.

---

**Table 6: Input serialization template for PaCoRe synthesis.** We use this template to embed the current problem  $x$  (denoted as `original_prompt`) and the compact message set  $M$  (denoted as `ref_responses`) into the model’s context. In the degenerate case where the message set is empty ( $M = \emptyset$ ), this template is bypassed, and the original problem input is passed to the model unmodified.

## D. More Ablation Studies

**Multi-round Inference Recipe.** We study the impact of the inference trajectory configuration  $\vec{K}$  in a multi-round setting. Fixing the first round width  $K_1 = 32$ , we ablate the second round width  $K_2$  (using a configuration  $\vec{K} = [32, K_2]$ ). Table 7 indicates that  $K_2 = 4$  yields the best performance on both benchmarks.

**Table 7:** Ablation over the intermediate round width  $K_2$  during inference, using a fixed multi-round configuration of  $\vec{K} = [32, K_2]$ .

$K_2$	HMMT 2025	LiveCodeBench(2408-2505)
2	94.0	77.4
4	<b>94.6</b>	<b>78.4</b>
8	94.0	77.5