

---

# Confucius3-Math: A Lightweight High-Performance Reasoning LLM for Chinese K-12 Mathematics Learning

---

Lixin Wu   Na Cai   Qiao Cheng   Jiachen Wang   Yitao Duan  
NetEase Youdao, Beijing, China  
{wulixin, caina, chengqiao, wangjiachen, duan}@rd.netease.com

## Abstract

We introduce Confucius3-Math, an open-source large language model with 14B parameters that (1) runs efficiently on a single consumer-grade GPU; (2) achieves SOTA performances on a range of mathematical reasoning tasks, outperforming many models with significantly larger sizes. In particular, as part of our mission to enhancing education and knowledge dissemination with AI, Confucius3-Math is specifically committed to mathematics learning for Chinese K-12 students and educators. Built via post-training with large-scale reinforcement learning (RL), Confucius3-Math aligns with national curriculum and excels at solving mainstream Chinese K-12 mathematical problems with low cost. In this report we share our development recipe, the challenges we encounter and the techniques we develop to overcome them. In particular, we introduce three technical innovations: Targeted Entropy Regularization, Recent Sample Recovery and Policy-Specific Hardness Weighting. These innovations encompass a new entropy regularization, a novel data scheduling policy, and an improved group-relative advantage estimator. Collectively, they significantly stabilize the RL training, improve data efficiency, and boost performance. Our work demonstrates the feasibility of building strong reasoning models in a particular domain at low cost. We open-source our model and code at <https://github.com/netease-youdao/Confucius3-Math>.

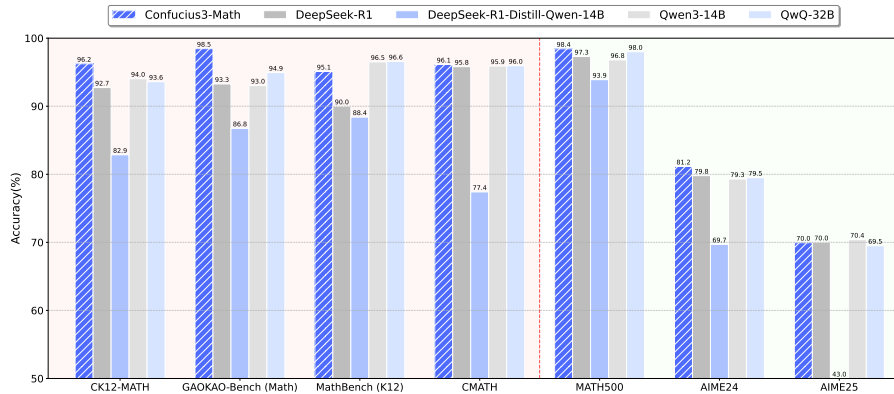


Figure 1: Benchmark performance of Confucius3-Math

# 1 Introduction

Reasoning capabilities have been the focus of the latest Large Language Model (LLM) developments. OpenAI’s o1 model (OpenAI, 2024) represents a major breakthrough in advancing LLM’s abilities in solving complex, multi-step tasks such as math and coding, igniting mainstream awareness of reasoning-specific LLMs. o1 is known to acquire its performance via Test Time Compute (TTC). However, neither the model weights nor the technical details of o1, or its sequels OpenAI (2025), are released. TTC is a well-known technique to boost model’s performance by allowing the model to “think” or compute more at inference time and has been applied successfully in many reasoning tasks, mostly in games such as Backgammon (Tesauro, 1994), poker (Brown & Sandholm, 2017), and Go (Silver et al., 2016) etc. It takes on many forms from simple consensus/majority vote to beam search and Monte-Carlo Tree Search (MCTS) (Metropolis & Ulam, 1949), depending on the application. Jones (2021) studies scaling laws between train time and test time, clarifying the benefit of TTC, also in the game setting.

With LLMs that generate results auto-regressively, TTC is more complicated. Brown et al. (2024) shows that inference-time scaling laws still exists with LLMs and the fraction of problems that are solved by any generated sample scales log-linearly with the number of samples<sup>1</sup>. However, this only suggests the *possibility* of TTC’s benefits, how to implement a feasible strategy remains challenging. The successes from game AI may not readily transfer to LLM. One major challenge is that the game models are each solving a *single* task (e.g., Go), while the reasoning LLMs are expected to solve potentially infinite number of problem: even the same math problem framed in different words represents a different state for an LLM, resulting in a huge state-action space. Another challenge is that it is much more difficult to verify the correctness of the results.

After the release of o1, there have been many works attempting to replicate it. For example, Qin et al. (2024) tries to dissect o1’s mechanisms by detailed analysis of its outputs and focused experiments. Methods exploited include SFT, DPO, tree search etc. Zhao et al. (2024) uses explicit MCTS to boost reasoning performance. These works obtain mild progress and the results are mostly speculations.

DeepSeek-R1 (DeepSeek-AI et al., 2025), released in January 2025, represents a transformative leap in reasoning capabilities and training methodologies. It is the first model to demonstrate that pure reinforcement learning (RL)—without SFT—can effectively cultivate advanced reasoning skills, enabling organic emergent reasoning behaviors like chain-of-thought and self-verification. Technical achievements aside, R1’s openness also democratizes access, enriches and strengthens the open LLM ecosystem, paving the way to further advancing reasoning capabilities and enabling practical applications.

**Reasoning LLM and Education.** We join the endeavor to push the limit of LLM’s reasoning capabilities by focusing on a particular domain, namely, K-12 education which has long been regarded as one of the main sectors in which artificial intelligence can bring about disruptive innovations. Education places a much more demanding requirement on model’s accuracy, as research has shown that incorrect answers generated by LLMs could have a negative impact on student learning (Ali et al., 2024). Surprisingly, the most powerful reasoning LLMs that achieve high scores in some complex benchmarks do not perform well on K-12 tasks (see our evaluation in 7 and studies such as Zhu et al. (2024)), calling for model adaptation to this particular domain.

Another major challenge is the inequality caused by the economic disparity of the students. The academic performance of K-12 students has been shown to be highly correlated with their socioeconomic status (Cuéllar et al., 2024). AI is expected to mitigate the inequality by making educational resources more accessible to all. In particular, LLMs, with their ability to host rich knowledge and support for natural language interaction, represent a promising solution, as they have the potential to provide, to all students with low cost at large-scale, one key element in learning: highly qualified teachers. LLMs have been shown to be effective in improving learning (e.g., Wang & Fan (2025); Kestin et al. (2024)). However, high-performance LLMs are still quite expensive to build and deploy, making their access not affordable for many students from low-income regions. Meanwhile, students who can pay for expensive models have shown a clear advantage in learning (Zhu et al., 2024). Ironically, at the current high cost, the more powerful the LLMs are, the more severe the digital divide becomes.

---

<sup>1</sup>But the performance plateaus around 100 samples - not a very high ceiling.

We believe that a strong low-cost reasoning LLM goes a long way towards equalizing education. In this work, we release Confucius3-Math and our experiences in building it. As shown in 1, Confucius3-Math is a strong reasoning model achieving SOTA performances on a range of mathematical reasoning tasks and outperforming many models with significantly larger sizes, especially in our target domain, namely Chinese K-12 mathematics learning. The contributions of our work include:

1. Confucius3-Math demonstrates the feasibility of building strong and practical reasoning models in a particular domain at low cost. Specifically, (1) Confucius3-Math achieves SOTA performance on a number of K-12 math benchmarks; (2) The training of Confucius3-Math costs only \$26K; and (3) its inference performance is about  $15\times$  that of DeepSeek R1.
2. Technically, we show that, given the right base model, a moderate amount of high-quality data, and a good training recipe, post-training via RL is capable of eliciting strong reasoning capabilities in lightweight models.
3. We introduce Targeted Entropy Regularization, Recent Sample Recovery and Policy-Specific Hardness Weighting. These innovations encompass a new entropy regularization, a novel data scheduling policy, and an improved group-relative advantage estimator. They significantly stabilize the RL training, improve data efficiency, and boost performance. Recent Sample Recovery and Policy-Specific Hardness Weighting, one novel data scheduling policy and one improved group-relative advantage estimator, respectively, that significantly improves data efficiency, stabilizes the RL training process, and boost performance.
4. We open-source Confucius3-Math and its technical details, inviting the community to utilize it in real-world applications and collaborate on developing practical reasoning LLMs.

## 2 Approach

Our goal is to build practical models that enable real-world AI learning applications. To this end, the development of Confucius3-Math is guided by the following principles:

1. Focused Domain: We target the education sector. As a first step towards building LLM models that can effectively assist learning, we especially focus on Chinese K-12 mathematics education.
2. Low Cost: The cost of both developing and deploying the model should be low. We believe that this is crucial for supporting sustainable real-world applications and providing access to all.
3. High Performance: Research has shown that incorrect answers generated by LLMs could have a negative impact on student learning (Ali et al., 2024). Therefore, the model must have adequate accuracy to be useful for realistic learning scenarios.
4. Learning Support: The model should be capable of producing sufficient auxiliary information that are essential part of learning.
5. Openness: Education is one of the foundations of our social structure with profound and long-term impacts on many people’s lives. To ensure its proper development, we believe that its key elements and mechanisms should be open, allowing people from different sectors to scrutinize, collaborate, and contribute.

We make the following design choices. First, we choose to build our model upon strong open-source models that already possess powerful capabilities through pre-training and fine-tuning. Recently, post-training has emerged as an effective way to enhance reasoning capabilities and the booming open ecosystem provides great resources for building strong models and applications with low cost.

Second, we choose model size to be 14B which is large enough to have the expressivity for complex tasks, yet not too large so that, with good engineering, can still run efficiently on consumer-grade GPU.

For post-training, we contemplate a number of choices, including RL, SFT, distillation, and their combinations. For enhancing small models, past works have shown their effectiveness in different situations. For example, DeepSeek R1 (DeepSeek-AI et al., 2025) shows that distillation obtains much larger performance gains than post-training the same model via RL. The latest Qwen3 series

also use strong-to-weak distillation to optimize their lightweight models (Qwen3-0.6B, 1.7B, 4B, 8B, and 14B) and they show that a combination of off-policy/on-policy distillation achieves significantly better performance than RL (Yang et al., 2025).

However, our objective is *not* simply a good model, but a model on par or exceeding even the strongest models with larger sizes in our domain. We believe RL, which encourages emergent behavior, holds greater potential, as indicated by its superhuman performance in tasks such as Go (Silver et al., 2016) and protein structure prediction (Jumper et al., 2021). So we select RL as the main ingredient of our solution. As our experiments turn out, this is quite effective.

### 3 Data Curation

#### 3.1 Data Sources

The data used for training the model comes from two major sources: open-source and proprietary.

**Open-Source Data.** To enhance the model’s mathematical capabilities, we collect a large number of open-source English mathematics datasets, including:

1. Datasets focused on foundational and advanced mathematics skills, such as GSM8K (Cobbe et al., 2021), applied\_math<sup>2</sup>, and Advanced-Math<sup>3</sup>.
2. Collections of competition-level mathematics problems for various levels, including MATH (Hendrycks et al., 2021), NuminaMath-1.5 (LI et al., 2024).
3. High-quality reasoning datasets from universities, forums, and research, covering multiple fields such as mathematics, science, puzzles, and coding, including OpenThoughts-114k (Guha et al., 2025), orz\_math\_57k\_collection (Hu et al., 2025), s1K-1.1 (Muennighoff et al., 2025), LIMR (Li et al., 2025), and LIMO (Ye et al., 2025), among others.

The sizes of the datasets are summerarized in table 1.

Dataset	Number of Samples
GSM8K Training Set	7,473
applied_math	10,000
Advanced-Math	1,032
MATH Training Set	7,497
NuminaMath-1.5	896,215
OpenThoughts-114k	113,957
orz_math_57k_collection	56879
s1K-1.1	1,000
LIMR	1,389
LIMO	817
<b>Total</b>	<b>1,096,241</b>

Table 1: Overview of Open Source Training Datasets

**Proprietary Data.** Since we optimize for K-12 mathematics learning scenario, in addition to relevant open-source data, we also collect math questions, and their solutions, accumulated during the operation of our business. They cover various mathematics problems for the domestic K-12 stages (primary, middle, and high school), including a rich variety of types, such as single-choice, multiple-choice, true/false, fill-in-the-blank, calculation, proof, and mixtures of multiple question types etc.

<sup>2</sup>[https://www.modelscope.cn/datasets/gavinluo/applied\\_math.git](https://www.modelscope.cn/datasets/gavinluo/applied_math.git).

<sup>3</sup><https://www.modelscope.cn/datasets/Haijian/Advanced-Math.git>.

### 3.2 Data Preparation

Prior works (Ye et al., 2025; Team, 2025) have shown that the quality and diversity of training data play a crucial role in reinforcement learning. Therefore, we have implemented a rigorous data preprocessing procedure utilizing the NeMo-Curator framework (NVIDIA).

First, we remove test splits from all open-source datasets, as well as all commonly used benchmarks in research. Additionally, since RL requires high-quality data with standard answers, we exclude all synthetic data and data without ground truth.

Next, we execute the following data processing workflow in sequence on open-source data:

1. Exact Deduplication: We retain only one instance of identical questions in the dataset.
2. Fuzzy Deduplication: We deduplicate questions with high Jaccard similarity scores with others.
3. Semantic Deduplication: We use  $k$ -means to cluster the embeddings of the questions. Within each cluster, we compute cosine similarities between each pair and remove one question if it is too similar to the other.
4. Question Type Selection: To get accurate rewards during the RL process, we remove multiple-choice, true/false, and proof questions. The first two types can be guessed and there is no easy way to verify the third.

For our proprietary data, we apply a cleaning stage, as the data originates from mass-scale automated entry with manual correction, it inherently contains significant noise. We do not perform other complex processing since the entering process has avoided duplication and the question types are all friendly to our reward model.

As a result, we retain approximately 540,000 samples of data for actual training, including 210,000 from open-source data and 330,000 from proprietary data. Note that this is the amount of training data we consume up to this release of Confucius3-Math. Our data sources, especially proprietary, are producing more. We leave the rest for future improvement.

## 4 Base Model Selection

A good base model with the right initial capabilities could avoid extra cold start, save computation and reach higher performance ceiling for the post-training. Prior to making selection, we conduct extensive explorations on different base models, including Qwen2.5-14B-Base, Qwen2.5-14B-Instruct, DeepSeek-R1-Distill-Qwen-14B, and a previous Confucius reasoning model Confucius-o1-14B. The explorations are conducted via 150 steps of RL. We monitor both the reward scores and policy entropy, which turns out to be a good indicator of the model’s exploration ability which is crucial for RL (see discussion in section 6.1).

We discover that all display certain level of trainability, with DeepSeek-R1-Distill-Qwen-14B performing the best. And indeed, the model’s outcome is correlated with its initial entropy. We postulate that the policy entropy could be a critical metric when determining whether a model can serve as an initial model for RL optimization. Interestingly, among the models, Confucius-o1-14B, which is trained from Qwen2.5-14B-Instruct, shows a lower entropy than its base, which appears to be caused by its training method: DPO without entropy regularization.

In addition, we also assess the richness of models’ output, as in an education setting, we expect the model to output the problem-solving process in its answer. The reason for emphasizing this soft metric is that in current RL training, the reward module predominantly focuses on the correctness of the response while often overlooking its compliance with the prompt, which largely hinges on the compliance capabilities of its initial model.

We ultimately select DeepSeek-R1-Distill-Qwen-14B model as the base model. Distilled from Qwen2.5-14B using DeepSeek-R1’s data, the model exhibits robust CoT and holds a greater initial edge in the field of mathematics compared to other models of comparable scale. Moreover, its answers align with our expectations for including problem-solving steps.

## 5 Reward Modeling

We adopt a two-stage, hybrid method combining both rule-based and model-based approaches for reward modeling.

The first stage acts as a filter, filtering out any output that either does not conform to the right format or contains repetitive text, a problematic phenomenon frequently observed with LLM generation (Wang et al., 2024; Xue et al., 2023). Only the response results that pass these two types of filtering can obtain the subsequent correctness reward score. We adopt this filtering mechanism rather than a reward weighting scheme mainly because the format requirement can be quickly satisfied in the early stage of training. Incorporating it as part of the reward weighting does not have a practical effect in the later stage. In addition, these two types of constraints are also hard requirements. In real usage, we cannot accept a response that does not meet these two types of constraints but is correct in result.

The second stage rewards the model for its correctness. We use either rule-based or model-based approaches based on different data sources. For open-source data, since the result of each question consists of a single expression, we continue to use conventional rule-based reward modeling. Specifically, we utilize Math-Verify (HuggingFace) for answer extraction and verification. Our proprietary data, on the other hand, contains math problems in forms typically found in Chinese K-12 schools. For example, some problems may contain multiple subquestions, and answers may depend on long textual descriptions far away. Conventional extraction rules often do not work. Therefore, we adopt a model-based approach and use LLM-as-a-Judge to determine the correctness of the model’s response. In addition to checking the correctness of the final answer, the reward model also verifies the solving process, as we need it to produce learning materials for students.

Regarding the text repetition filtering, we also find that it is not enough to only scrutinize the answer. Both the thinking and answer parts need to be examined. Otherwise repetition will appear in the thinking part in the early stage of the training and gradually spread to answer region while training progresses, resulting in an irreversible collapse.

## 6 Training

We follow a three-stage, pure-RL training pipeline. The staging is mainly for expanding context window, which progressively grows from 4K to 8K, and eventually 16K, while preserving training efficiency. We also adjust training objectives and other configurations as we address some issues we encounter, and with the availability of new tools. We use the same training template as that of R1 (DeepSeek-AI et al., 2025) and main-stream optimization frameworks such as Group Relative Policy Optimization (GRPO) (Shao et al., 2024) and Dynamic sAmpling Policy Optimization (DAPO) (Yu et al., 2025), with our own improvements which will be elaborated in sections 6.1, 6.2, and 6.3. A brief summarization of the characteristics of the stages is presented in table 2. Figure 2 shows the changes of average response lengths as the training progresses through the three stages, demonstrating the effectiveness of context expansion and the growth of model’s Chain-of-Thoughts (CoT).

Stage	Core Algorithms	Context Length
Stage 1	GRPO + Targeted Entropy Regularization	4K
Stage 2	DAPO + RSR + PSHW	8K
Stage 3	DAPO + RSR + PSHW	16K

Table 2: Training Stage Characteristics

In retrospect, we find that our approach, although independently developed, bears two major similarities with concurrent work of Skywork-OR1 (He et al., 2025), namely staged context window expansion and targeted entropy regulation. We elaborate our ideas in subsequent sections.

### 6.1 Targeted Entropy Regularization

Entropy regularization is a commonly used technique in RL policy optimization. It is believed to benefit exploration by encouraging the selection of more stochastic policies and preventing the

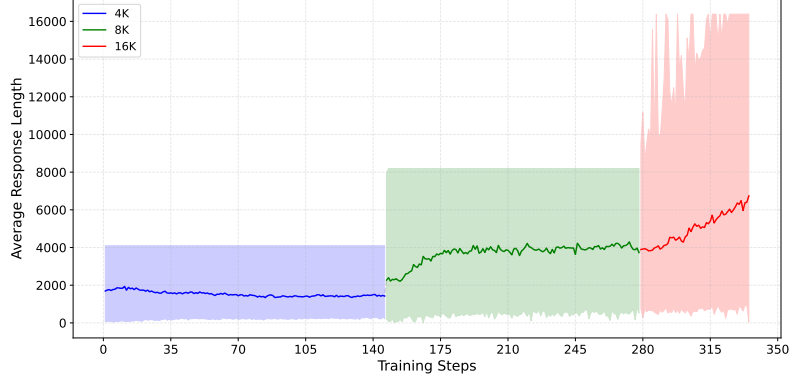


Figure 2: Multi-Stage Length Expansion

training from prematurely converging on deterministic policies (Ahmed et al., 2019; Snoswell et al., 2020; Schulman et al., 2017). It is now commonly used in LLM training (e.g., Yu et al. (2025)).

However, from our experience, entropy seems to be a delicate matter. LLMs are complex machines with output distribution over a large space of multilingual tokens. The maximum entropy principle appears to have double effects. One major issue we encounter during the training process is the “mixed language problem”, a phenomenon that also arises with DeepSeek R1 (DeepSeek-AI et al., 2025). This phenomenon appears to be correlated with high entropy, which may encourage the model to substitute tokens with the same meaning from different languages. We tried adding linguistic constraints to the reward, but it was ineffective.

Therefore we introduce “Targeted Entropy Regularization” by adding the following term to the loss function

$$|\text{entropy\_loss} - \text{entropy\_target}| * \text{entropy\_coeff}$$

where `entropy_target` is desired entropy level set empirically to 0.55. This is different from the Adaptive Entropy Control in Skywork-OR1 (He et al., 2025): their strategy aims to constrain the lower limit of entropy to prevent entropy collapse, while we constrain both the upper and lower limits. `entropy_coeff` is set to 0.001, similar to conventional entropy regularization.

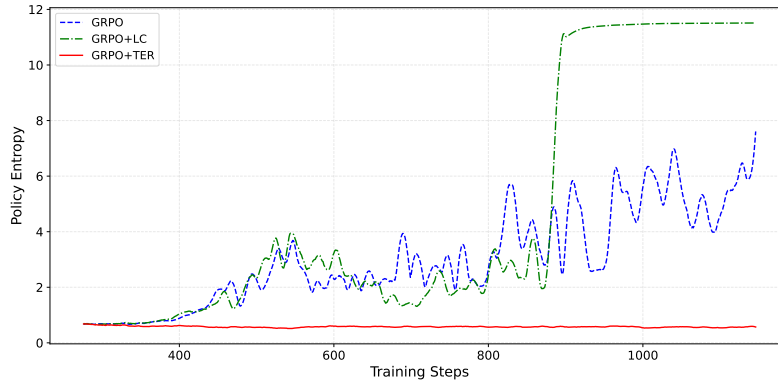


Figure 3: Effects of Target Entropy Regularization on Policy Entropy

Figure 3 plots the model output entropy using different mechanisms. Clearly Target Entropy Regularization effectively constrains the entropy to a desired level and stabilizes the training. The mixed language problem also goes away.

Our understanding is that many of the issues are rooted in the stability of training. Once we switch to a more stable training technique, namely our improved DAPO, described in the following sections, after the first stage, those nuances such as KL divergence or entropy regulation are not so useful.

## 6.2 Policy-Specific Hardness Weighting

A crucial factor influencing RL fine-tuning is the ability-difficulty gap: the gap between the current policy model’s ability to solve the problems and their difficulty levels in the given batch. The gap being both too wide or too narrow could impact the learning efficacy. Traditionally this issue is addressed with curriculum learning: ordering the training data according to their difficulty levels. Curriculum learning has also been applied in LLM RL training. For example, Wang et al. (2025) treats sample selection as an exploration/exploitation problem and models a distribution-level learnability among samples. Chen et al. (2025) adds a dedicated curriculum policy trained together with the LLM policy. Both adjust sampling probabilities throughout training to get a better sample scheduling.

We propose a policy specific approach, denoted Policy-Specific Hardness Weighting (PSHW). The key insight is, because the policy model is evolving during training, so is the learnability of samples. In other words, learnability should be relative to the current policy. Unlike curriculum learning which schedules training samples, PSHW incorporates relative difficulty into advantage estimation which is much simpler and more efficient.

PSHW can be used with any compatible RL algorithm including GRPO. We use it with DAPO. Using conventional notation as that of DAPO (Yu et al., 2025), the algorithm samples a group of outputs  $\{o_i\}_{i=1}^G$  for each question  $q$  paired with the answer  $a$ , and optimizes the policy via the following objective:

$$\begin{aligned} \mathcal{J}(\theta) = & \mathbb{E}_{(q,a) \sim \mathcal{D}, \{o_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot|q)} \left[ \sum_{i=1}^G \sum_{t=1}^{|o_i|} \min \left( r_{i,t}(\theta) \hat{A}_{i,t}, \text{clip} \left( r_{i,t}(\theta), 1 - \varepsilon_{\text{low}}, 1 + \varepsilon_{\text{high}} \right) \hat{A}_{i,t} \right) \right] \\ \text{s.t. } & 0 < \left| \{o_i \mid \text{is\_equivalent}(a, o_i)\} \right| < G, \end{aligned} \quad (1)$$

where

$$r_{i,t}(\theta) = \frac{\pi_{\theta}(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})}, \quad \hat{A}_{i,t} = [R_i - \mu] D(q), \quad \mu = \text{mean}(\{R_i\}_{i=1}^G) \quad (2)$$

and

$$R_i = \begin{cases} 1, & \text{is\_equivalent}(a, o_i) \\ -1, & \text{otherwise} \end{cases}, \quad D(q) = \alpha \times \mu + 1.256. \quad (3)$$

We set the coefficient  $\alpha = -0.256$  empirically. In essence, the mean score  $\mu$  represents how well the policy does on question  $q$ , which we believe is a good measure of the difficulty of  $q$  relative to the current policy. PSHW implements a kind of adaptive difficulty-aware advantage estimator which encourages the policy to optimize and explore problems it perceives as “hard”.

We also introduce another modification to DAPO, as obvious from equation 1. Namely, as observed by Liu et al. (2025), length normalization used in GRPO, and DAPO as well, tends to produce a response length bias. For correct responses, the algorithm favors brevity in correct answers; conversely, the normalization reduces the penalty for longer incorrect responses, which leads the policy to prefer longer incorrect responses. Therefore, similar to Liu et al. (2025), we remove the length normalization.

## 6.3 Recent Sample Recovery

One of the innovations introduced by DAPO (Yu et al., 2025) that prove to be quite effective in boosting training efficiency is Dynamic Sampling: for each batch, DAPO over-samples and filters out prompts with accuracy equal to 1 or 0. This ensures that the policy is updated with effective samples.

However, in order to fill each batch, over-sampling results in generating much more samples that are wasted. We observe up to 3 to 9 times more samples generated for each batch. Among them, some are filtered out. Still, a significant portion of the the remaining legitimate samples are also discarded because of the batch size constraint.



Ostensibly, it does not make sense to recover these samples for future use because they are generated from an old policy. Also the situation is different from experience replay (Lin, 1992) which reuses old experiences in an *off-policy* setting. Here we are doing *on-policy* learning and the samples are never used. Nevertheless, we experiment saving those legitimate overflow samples and using them in the next batch, a method we denote Recent Sample Recovery (RSR).

An orthodox way should have incorporated importance sampling in the process to account for the disparity in distributions. However, our experiments do not produce much difference so for simplicity we do not use it. Detailed algorithm is presented in algorithm 1:

---

**Algorithm 1** DAPO+RSR: DAPO with Recent Sample Recovery

---

**Input** initial policy model  $\pi_\theta$ ; reward model  $R$ ; task prompts  $\mathcal{D}$ ; **batch size**  $N$ ; hyperparameters

$\epsilon_{\text{low}}, \epsilon_{\text{high}}$

```

1: Initialize recovery buffer  $B_r = \emptyset$ 
2: for step = 1,...,M do
3:   if  $|B_r| > 0$ :
4:     Add  $B_r$  to dynamic sampling buffer  $B_c$ 
5:     Set  $B_r = \emptyset$ 
6:   Sample a batch  $\mathcal{D}_b$  from  $\mathcal{D}$ 
7:   Update the old policy model  $\pi_{\theta_{\text{old}}} \leftarrow \pi_\theta$ 
8:   Sample  $G$  outputs  $\{o_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot|q)$  for each question  $q \in \mathcal{D}_b$ 
9:   Compute rewards  $\{r_i\}_{i=1}^G$  for each sampled output  $o_i$  by running  $R$ 
10:  Filter out  $o_i$  and add the remaining to the dynamic sampling buffer  $B_c$ 
11:  if  $|B_c| < N$ :
12:    continue
13:  else:
14:    Save remaining legitimate samples for subsequent use as  $B_r = B_c[N:]$ 
15:    Select the currently used data as  $B_c = B_c[:N]$ 
16:    For each  $o_i$  in the  $B_c$ , compute  $\hat{A}_{i,t}$  for the  $t$ -th token of  $o_i$ 
17:    for iteration = 1, ...,  $\mu$  do
18:      Update the policy model  $\pi_\theta$  by maximizing the DAPO objective

```

**Output**  $\pi_\theta$

---

Figures 4 show the effects of RSR on training quality and data efficiency. Figure 4a shows the optimal model quality on evaluation set (solid line, left vertical axis) and the actual amount of raw training data consumed (dashed line, right vertical axis) during the training process. The  $x$ -axis is the number of training steps, which is proportional to the actual effective samples consumed since each batch consists of a constant number of 256 samples. The solid purple line is always above the solid green line, indicating RSR+DAPO’s superior training performance. The dashed purple line, on the other hand, grows with a slower rate than the dashed green line. At the same training steps, the RSR strategy consumes less training data, indicating that it can improve the utilization of training data. Figure 4b on the right figure is a more intuitive demonstration of RSR’s data efficiency. With the same amount of raw training data, RSR always outperforms the non-RSR strategy, with a quality difference reaching up to a 3 percent points at around the 80k data point.

Interestingly, RSR not only improves data efficiency, but also produces better results, even with the same amount of *effective* data, as indicated by the two solid lines in figure 4a. We speculate that this may be due to certain smoothing effect, since RSR blends two distributions that are not too far away. We will give this issue a more rigorous treatment in another paper.

## 7 Evaluation

**Benchmarks.** We evaluate Confucius3-Math on several benchmarks relevant to our target domain. CK12-MATH (Chinese K-12 Math) is our internal benchmark sampled from *actual* user queries. It contains 500 math problems used in daily learning, such as homeworks and quizzes, in typical Chinese schools. The performance on this dataset represents the actual experiences of a typical Chinese K-12 student interacting with the models. We also select several commonly used public benchmarks for quality assessment. These include: the math subset of GAOKAO-Bench (Zhang et al., 2023),

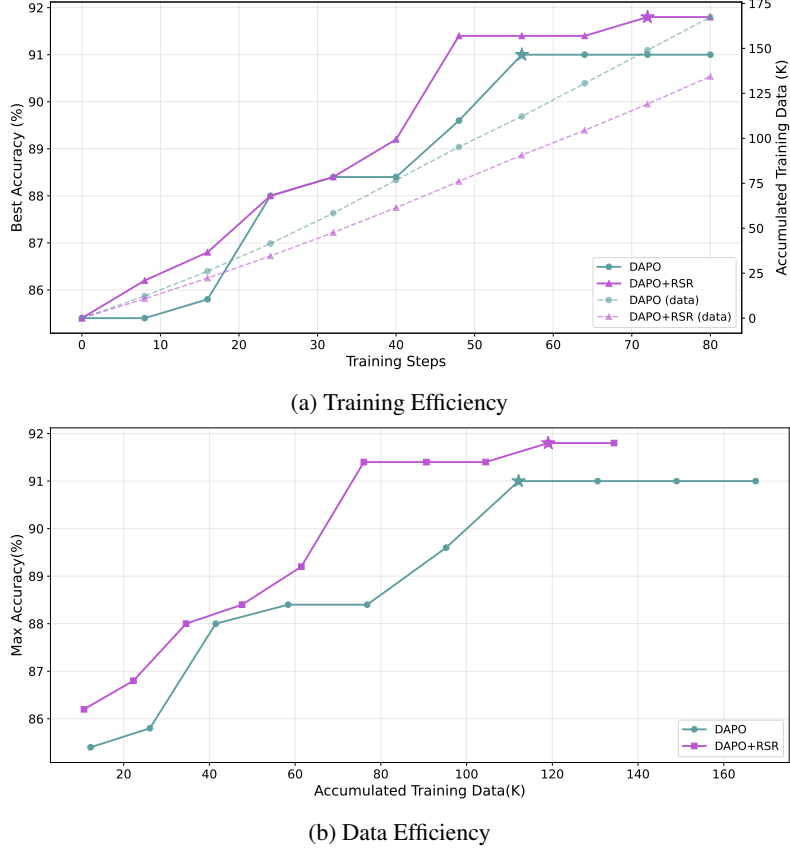


Figure 4: RSR+DAPO vs. DAPO

which contains problems from China’s National College Entrance Exam (GAOKAO), the K12 subset in MathBench (Liu et al., 2024), and CMATH (Wei et al., 2023), a Chinese Elementary School Math Word Problems dataset. In addition, we also include some widely used competition-level math datasets: MATH500, AIME24, and AIME25.

**Baselines.** We compare against DeepSeek-R1, the SOTA open-source reasoning model with 671B parameters, our base model DeepSeek-R1-Distill-Qwen-14B, as well as two other strong open-source models: Qwen3-14B (Yang et al., 2025) and QwQ-32B (QwenTeam). Qwen3-14B is one of the latest Qwen3 series with size comparable to Confucius3-Math. Pre-trained using approximately 36 trillion tokens, nearly twice the amount of Qwen2 (and our base model), and then fine-tuned with a combination of off-policy/on-policy distillation using data from Qwen3 flagship models (Yang et al., 2025), Qwen3-14B stands for the current SOTA model at its scale. QwQ-32B (QwenTeam) is also a strong reasoning model but with over twice the size of Confucius3-Math. These baselines represent different design choices in the landscape of building practical domain models.

**Evaluation Setup.** We extract corresponding performance data from their technical report if available, and use the following inference setting if not. For each model, we use the official system prompt, except for R1, for which we find that not using the system prompt leads to higher quality. The maximum generated response tokens is set to 32,768 for all models. To avoid the text repetition problem with greedy decoding found in later RL training stage of DeepSeek and Qwen3 (DeepSeek-AI et al., 2025; Yang et al., 2025), we use the sampling strategy to generate  $k$  response results and reported the pass@1 results. Specifically, for our model, we use a sampling temperature of 1.0 and a top- $p$  value of 0.7, while for other models, we use the sampling parameters recommended by the official documentation. The value of  $k$  is set differently for different benchmarks for comparability and stability. For MATH500, AIME24, and AIME25, we follow DeepSeek’s practice of setting  $k = 64$ , and for other test sets,  $k$  is set to approximately  $2,000/N$  where  $N$  is the number of samples

in the set. Pass@1 is then calculated as

$$pass@1 = \frac{1}{k} \sum_{i=1}^k p_i$$

where  $p_i$  denotes the correctness of the  $i$ -th response.

Additionally, it is worth mentioning that our internal CK12-MATH benchmark also contains the intermediate steps in a problem’s solution so we use a more stringent standard for verifying Confucius3-Math since it is trained to generate these steps in the answer block. Namely, we also evaluate the correctness of these steps using LLM-as-a-judge. A problem is deemed solved by Confucius3-Math only when both the final answer and the intermediate steps are correct.

**Results.** The main results are presented in table 3. It shows the accuracy rates of the models on a number of benchmarks. The same data is also shown visually in figure 1 in the beginning of this paper. The first 4 rows in table 3, as well as the left part of figure 1 with pink background, are the benchmarks pertinent to general K-12 learning, while the rest are competition-level benchmarks. Confucius3-Math outperforms all other models on all but two benchmarks, on which Confucius3-Math’s performance is comparable to the strongest models. In particular, on the CK12-MATH benchmark, which is the most important focus of our application scenario, Confucius3-Math leads DeepSeek R1 by 3.5 points. Qwen3-14B, as a 14B model, also performs quite well on some tasks. However, Qwen3-14B starts with a much stronger initial model, consumes more data, and relies on more powerful teacher models. We believe Confucius3-Math’s RL solution points to an alternative, and maybe more effective, technical route. As mentioned earlier, we still have a large amount of unused data relevant to this scenario and continued improvement is almost certain. This high accuracy, plus its low inference cost (see section 8), makes Confucius3-Math a viable LLM solution in our actual business operations.

We also list, in parentheses, the score improvements of Confucius3-Math against its base, DeepSeek-R1-Distill-Qwen-14B, in the last column of table 3. Confucius3-Math obtains large gains across all benchmarks, with the largest reaching 26.98 points. They demonstrate the effectiveness of our RL optimization process.

	Benchmark (subset)	DeepSeek-R1	Qwen3-14B	QwQ-32B	DeepSeek-R1-Distill-Qwen-14B	Confucius3-Math
K-12	CK12-MATH	92.74	94.04	93.60	82.86	<b>96.24 (+13.38)</b>
	GAOKAO-Bench (Math)	93.27	93.04	94.93	86.75	<b>98.46 (+11.71)</b>
	MathBench (K12)	89.99	96.51	<b>96.57</b>	88.40	95.10 (+6.70)
	CMATH	95.81	95.90	95.95	77.41	<b>96.13 (+18.72)</b>
Comp	MATH-500	97.30*	96.80*	98.00*	93.90*	<b>98.44 (+4.54)</b>
	AIME 2024	79.80*	79.30*	79.50*	69.70*	<b>81.15 (+11.45)</b>
	AIME 2025	70.00*	<b>70.40*</b>	69.50*	42.97	69.95 (+26.98)

Table 3: Comparison between Confucius3-Math and other representative models. Items with asterisk are taken from their own publications. Bold items are the best performing among all models tested. The numbers in parentheses in the last column are score lifts of Confucius3-Math against its base, DeepSeek-R1-Distill-Qwen-14B.

## 8 Inference Performance

The Confucius3-Math model outperforms DeepSeek R1 671B on K-12 math reasoning tasks. With only 14B parameters, it can be deployed with much lower cost, a crucial factor affecting the sustainability of a service. We conduct tests to verify Confucius3-Math’s inference efficiency under two configurations: (1) general throughput measurements with abundant resources, and (2) low-resource deployment feasibility test. For (1), we run both R1 and Confucius3-Math on the same high-performance hardware and measure their throughput, in terms of tokens per second. Specifically, we use a machine with the following hardware components:

- 8 × NVIDIA H800 SXM5 GPUs with 80 GB HBM3 memory

- 112 Intel(R) Xeon(R) Gold 6330 CPU @ 2.00GHz

The models are deployed using the vLLM 0.9.0 framework, and CUDA (version 12.8). Both prefix caching and chunk filling are enabled during these tests. We choose FP8 as the model quantization level. For DeepSeek R1, a parallel strategy with  $TP = 8$  is employed. Confucius3-Math inference is run on a single GPU while DeepSeek R1 671B on 8, to make sure R1 is deployed with adequate resources. We scale the results of Confucius3-Math by a factor of 8 for the two to be comparable. All other parameters are configured to the default settings of the vLLM framework.

A model with low resource requirement is advantageous, as it has more application opportunities. For this test we deploy Confucius3-Math with FP8 precision on a single GeForce RTX 4090D GPU with 24GB memory. Again we scale the numbers by 8 for comparability. Note that it is not possible to run DeepSeek R1 in this setting at all.

The test data are sampled from user queries with an average input length of 115 tokens. We measure throughput for both input and output under different request rates (QPS).

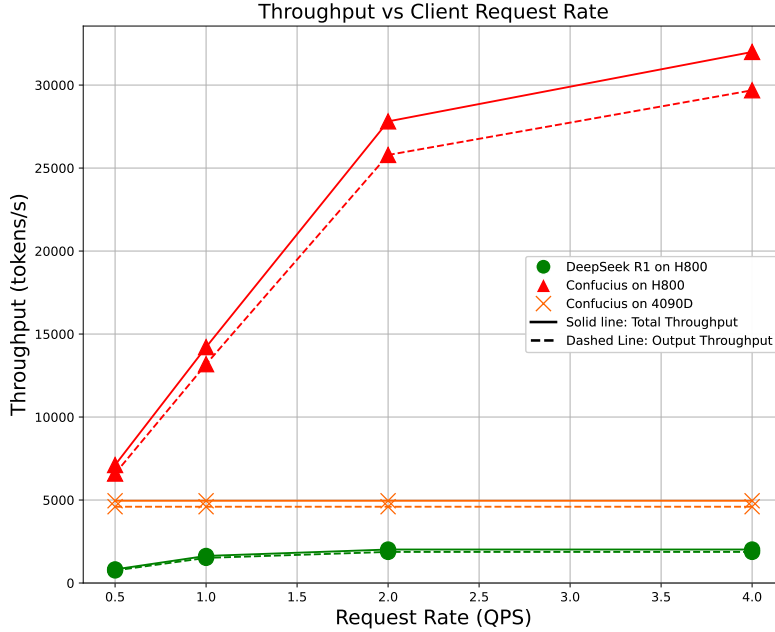


Figure 5: Deployment Performance Result

Results are shown in figure 5. For each run, solid line represents total throughput, in tokens per second, while dashed line is output throughput, also in tokens per second. DeepSeek R1 and Confucius3-Math’s performances on H800 are shown as green dots and red triangles, respectively. Confucius3-Math’s throughputs on 4090D are represented as red  $\times$ .

A few patterns are obvious. Firstly, Confucius3-Math outperforms R1 by large margins at all request rates. Secondly, R1 saturates quickly at 1,513 and 1,631 tokens per second on output and total throughput, respectively, at  $QPS = 1$ . This means that the system reaches its processing capacity at very low workload. Confucius3-Math, on the other hand, starts with 13,182 and 14,211 tokens per second at  $QPS = 1$ , scales more gracefully as QPS grows. It reaches its peak throughput at QPS 4, with a total throughput of 31,994. Compared with DeepSeek R1, this is a  $15.8\times$  speedup.

Thirdly, even running on low-resource hardware, Confucius3-Math surpasses DeepSeek R1’s performance on H800. The model achieves impressive inference input and output throughputs of up to 4,596 and 4,956 tokens per second, respectively, more than double those of R1.

## 9 Feasibility and Cost

An alternative solution, strong-to-weak distillation, is a commonly used method to enhance small models. It is actually more complex than it appears and may involve hidden costs. Acquiring outputs from teacher models can be done via API for proprietary models or hosting inference services for open models. Neither is free. And the more powerful the teacher, the more expensive. In particular, on-policy distillation, such as that used in training Qwen3-14B (Yang et al., 2025), requires more frequent and real-time access to the teacher model. This is only possible when one has access to the model weights, further narrowing down one’s choices. Again, powerful models are usually large and hosting them is not a negligible cost. For example, on-policy distillation using DeepSeek R1 as the teacher model needs, in addition to training GPUs, at least one extra H800-like server (to host the teacher) to run, let alone efficiently.

In contrast, our post-training pipeline is simple, consisting of only RL stages. We can train our model using a *single* H800 server (8 GPUs). This is a much lower investment requirement.

Training Costs	Stage 1	Stage 2	Stage 3	Total
in H800 GPU Hours	4,234	5,470	3,405	13,109
in USD	\$8.5K	\$10.9K	\$6.8K	\$26K

Table 4: Training costs of Confucius3-Math, assuming the rental price of H800 is \$2 per GPU hour.

The training cost of Confucius3-Math, divided into stages, is presented in table 4. The entire training process takes 13,109 GPU hours and costs \$26K.

The inference cost is also very low. Under full-load operation, a single H800 server can process 106.9 million tokens per hour. This translates into \$0.15 per million tokens, far more affordable than most general-purpose LLMs.

## 10 Conclusions and Future Work

Confucius3-Math demonstrates the feasibility of building strong reasoning models in a particular domain at low cost. In particular, it shows that, given the right base model, a moderate amount of high-quality data, and a good training recipe, RL is capable of producing strong reasoning capabilities in small models.

We believe that the potential of RL has not been fully exploited. As we mentioned earlier, Confucius3-Math only consumes a small portion of our data. We expect the model performance to continue to improve as we carry on with the training.

There are also many other promising directions that we are exploring. LLM post-training provides great application opportunities for RL, as well as many challenges. For example, the targeted entropy regulation method used in this work is effective. However, it lacks a principled treatment. Fundamentally, the issue it addresses is the exploration–exploitation dilemma, a problem extensively studied in conventional RL. We believe it is worth investigating the issue using mature frameworks such as Bayesian multi-armed bandit (Scott, 2010).

Application-wise, we have been focusing on mathematical capabilities so far and the model is good at solving math problems. Real-world education and learning scenarios call for more comprehensive functions such as supporting more subjects (e.g., language learning), academic evaluation, homework correction, personalized learning etc. Confucius3 series models will incorporate them in the near future. We believe that conditions are ripe for LLMs to make a real impact on education and we will continue to push their capacity to support our mission to equalizing education with AI.

## Acknowledgments

We thank the open-source community, especially Qwen, DeepSeek, and verl, among others, for their generous sharing of models, technical expertise, and tools that have empowered us in our endeavors. We are inspired by their selfless acts of collaboration and are committed to sharing our own findings and resources to support others in their journeys and advance AI boundaries.

## References

- Ahmed, Z., Le Roux, N., Norouzi, M., and Schuurmans, D. Understanding the impact of entropy on policy optimization. In *International conference on machine learning*, pp. 151–160. PMLR, 2019.
- Ali, D., Fatemi, Y., Boskabadi, E., Nikfar, M., Ugwuoke, J., and Ali, H. Chatgpt in teaching and learning: A systematic review. *Education sciences*, 14(6):643, 2024.
- Brown, B., Juravsky, J., Ehrlich, R., Clark, R., Le, Q. V., Ré, C., and Mirhoseini, A. Large language monkeys: Scaling inference compute with repeated sampling, 2024. URL <https://arxiv.org/abs/2407.21787>.
- Brown, N. and Sandholm, T. Safe and nested subgame solving for imperfect-information games. *Advances in neural information processing systems*, 30, 2017.
- Chen, X., Lu, J., Kim, M., Zhang, D., Tang, J., Piché, A., Gontier, N., Bengio, Y., and Kamalloo, E. Self-evolving curriculum for llm reasoning, 2025. URL <https://arxiv.org/abs/2505.14970>.
- Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- Cuéllar, M.-F., Dean, J., Doshi-Velez, F., Hennessy, J., Konwinski, A., Koyejo, S., Moilola, P., Pierson, E., and Patterson, D. Shaping ai’s impact on billions of lives, 2024. URL <https://arxiv.org/abs/2412.02730>.
- DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., Xue, B., Wang, B., Wu, B., Feng, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Bao, H., Xu, H., Wang, H., Ding, H., Xin, H., Gao, H., Qu, H., Li, H., Guo, J., Li, J., Wang, J., Chen, J., Yuan, J., Qiu, J., Li, J., Cai, J. L., Ni, J., Liang, J., Chen, J., Dong, K., Hu, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Zhao, L., Wang, L., Zhang, L., Xu, L., Xia, L., Zhang, M., Zhang, M., Tang, M., Li, M., Wang, M., Li, M., Tian, N., Huang, P., Zhang, P., Wang, Q., Chen, Q., Du, Q., Ge, R., Zhang, R., Pan, R., Wang, R., Chen, R. J., Jin, R. L., Chen, R., Lu, S., Zhou, S., Chen, S., Ye, S., Wang, S., Yu, S., Zhou, S., Pan, S., Li, S. S., Zhou, S., Wu, S., Ye, S., Yun, T., Pei, T., Sun, T., Wang, T., Zeng, W., Zhao, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Xiao, W. L., An, W., Liu, X., Wang, X., Chen, X., Nie, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yang, X., Li, X., Su, X., Lin, X., Li, X. Q., Jin, X., Shen, X., Chen, X., Sun, X., Wang, X., Song, X., Zhou, X., Wang, X., Shan, X., Li, Y. K., Wang, Y. Q., Wei, Y. X., Zhang, Y., Xu, Y., Li, Y., Zhao, Y., Sun, Y., Wang, Y., Yu, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Ou, Y., Wang, Y., Gong, Y., Zou, Y., He, Y., Xiong, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Zhu, Y. X., Xu, Y., Huang, Y., Li, Y., Zheng, Y., Zhu, Y., Ma, Y., Tang, Y., Zha, Y., Yan, Y., Ren, Z. Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Xie, Z., Zhang, Z., Hao, Z., Ma, Z., Yan, Z., Wu, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Pan, Z., Huang, Z., Xu, Z., Zhang, Z., and Zhang, Z. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL <https://arxiv.org/abs/2501.12948>.
- Guha, E., Marten, R., Keh, S., Raoof, N., Smyrnis, G., Bansal, H., Nezhurina, M., Mercat, J., Vu, T., Sprague, Z., Suvarna, A., Feuer, B., Chen, L., Khan, Z., Frankel, E., Grover, S., Choi, C., Muennighoff, N., Su, S., Zhao, W., Yang, J., Pimpalgaonkar, S., Sharma, K., Ji, C. C.-J., Deng, Y., Pratt, S., Ramanujan, V., Saad-Falcon, J., Li, J., Dave, A., Albalak, A., Arora, K., Wulfe, B., Hegde, C., Durrett, G., Oh, S., Bansal, M., Gabriel, S., Grover, A., Chang, K.-W., Shankar, V., Gokaslan, A., Merrill, M. A., Hashimoto, T., Choi, Y., Jitsev, J., Heckel, R., Sathiamoorthy, M., Dimakis, A. G., and Schmidt, L. Openthoughts: Data recipes for reasoning models, 2025. URL <https://arxiv.org/abs/2506.04178>.
- He, J., Liu, J., Liu, C. Y., Yan, R., Wang, C., Cheng, P., Zhang, X., Zhang, F., Xu, J., Shen, W., Li, S., Zeng, L., Wei, T., Cheng, C., An, B., Liu, Y., and Zhou, Y. Skywork open reasoner 1 technical report. *arXiv preprint arXiv:2505.22312*, 2025.
- Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. Measuring mathematical problem solving with the math dataset. *NeurIPS*, 2021.

- Hu, J., Zhang, Y., Han, Q., Jiang, D., Zhang, X., and Shum, H.-Y. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model, 2025. URL <https://arxiv.org/abs/2503.24290>.
- HuggingFace. Math-verify. <https://github.com/huggingface/Math-Verify>.
- Jones, A. L. Scaling scaling laws with board games, 2021. URL <https://arxiv.org/abs/2104.03113>.
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2. URL <https://doi.org/10.1038/s41586-021-03819-2>.
- Kestin, G., Miller, K., Klaes, A., Milbourne, T., and Ponti, G. Ai tutoring outperforms active learning. May 2024. doi: 10.21203/rs.3.rs-4243877/v1. URL <http://dx.doi.org/10.21203/rs.3.rs-4243877/v1>.
- LI, J., Beeching, E., Tunstall, L., Lipkin, B., Soletskyi, R., Huang, S. C., Rasul, K., Yu, L., Jiang, A., Shen, Z., Qin, Z., Dong, B., Zhou, L., Fleureau, Y., Lample, G., and Polu, S. NuminaMath. [<https://huggingface.co/AI-MO/NuminaMath-1.5>] ([https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina\\_dataset.pdf](https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina_dataset.pdf)), 2024.
- Li, X., Zou, H., and Liu, P. Limr: Less is more for rl scaling. <https://github.com/GAIR-NLP/LIMR>, 2025.
- Lin, L.-J. Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine Learning*, 8(3):293–321, 1992. doi: 10.1007/BF00992699. URL <https://doi.org/10.1007/BF00992699>.
- Liu, H., Zheng, Z., Qiao, Y., Duan, H., Fei, Z., Zhou, F., Zhang, W., Zhang, S., Lin, D., and Chen, K. Mathbench: Evaluating the theory and application proficiency of llms with a hierarchical mathematics benchmark. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 6884–6915, 2024.
- Liu, Z., Chen, C., Li, W., Qi, P., Pang, T., Du, C., Lee, W. S., and Lin, M. Understanding rl-zero-like training: A critical perspective, 2025. URL <https://arxiv.org/abs/2503.20783>.
- Metropolis, N. and Ulam, S. The Monte Carlo method. *J-J-AM-STAT-ASSOC*, 44(247):335–341, September 1949. ISSN 0162-1459 (print), 1537-274X (electronic). doi: <https://doi.org/10.2307/2280232>. URL <http://links.jstor.org/sici?sici=0162-1459%28194909%2944%3A247%3C335%3ATMCM%3E2.O.CO%3B2-3;http://www.jstor.org/stable/2280232>.
- Muennighoff, N., Yang, Z., Shi, W., Li, X. L., Fei-Fei, L., Hajishirzi, H., Zettlemoyer, L., Liang, P., Candès, E., and Hashimoto, T. sl: Simple test-time scaling, 2025. URL <https://arxiv.org/abs/2501.19393>.
- NVIDIA. Nemo curator: The gpu-accelerated open source framework for efficient generative ai model data curation. <https://github.com/NVIDIA/NeMo-Curator>.
- OpenAI. Learning to reason with llms. Open AI blog, 2024. URL <https://openai.com/index/learning-to-reason-with-llms/>.
- OpenAI. Introducing openai o3 and o4-mini. Open AI blog, 2025. URL <https://openai.com/index/introducing-o3-and-o4-mini/>.
- Qin, Y., Li, X., Zou, H., Liu, Y., Xia, S., Huang, Z., Ye, Y., Yuan, W., Liu, Z., Li, Y., and Liu, P. O1 replication journey: A strategic progress report – part 1. *arXiv preprint arXiv:2410.18982*, 2024. URL <https://arxiv.org/abs/2410.18982>.

- QwenTeam. Qwq-32b: Embracing the power of reinforcement learning. <https://qwenlm.github.io/blog/qwq-32b/>.
- Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- Scott, S. L. A modern bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26(6):639–658, 2010. doi: <https://doi.org/10.1002/asmb.874>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/asmb.874>.
- Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y. K., Wu, Y., and Guo, D. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL <https://arxiv.org/abs/2402.03300>.
- Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–503, 2016. URL <http://www.nature.com/nature/journal/v529/n7587/full/nature16961.html>.
- Snoswell, A. J., Singh, S. P., and Ye, N. Revisiting maximum entropy inverse reinforcement learning: New perspectives and algorithms. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 241–249. IEEE, 2020.
- Team, K. Kimi k1.5: Scaling reinforcement learning with llms, 2025. URL <https://arxiv.org/abs/2501.12599>.
- Tesauro, G. Td-gammon, a self-teaching backgammon program, achieves master-level play. *Neural Comput.*, 6(2):215–219, March 1994. ISSN 0899-7667. doi: 10.1162/neco.1994.6.2.215. URL <https://doi.org/10.1162/neco.1994.6.2.215>.
- Wang, J. and Fan, W. The effect of chatgpt on students’ learning performance, learning perception, and higher-order thinking: insights from a meta-analysis. *Humanities and Social Sciences Communications*, 12(1):621, 2025. doi: 10.1057/s41599-025-04787-y. URL <https://doi.org/10.1057/s41599-025-04787-y>.
- Wang, W., Li, Z., Lian, D., Ma, C., Song, L., and Wei, Y. Mitigating the language mismatch and repetition issues in llm-based machine translation via model editing, 2024. URL <https://arxiv.org/abs/2410.07054>.
- Wang, Z., Cui, G., Wan, K., and Zhao, W. Dump: Automated distribution-level curriculum learning for rl-based llm post-training, 2025. URL <https://arxiv.org/abs/2504.09710>.
- Wei, T., Luan, J., Liu, W., Dong, S., and Wang, B. Cmath: Can your language model pass chinese elementary school math test?, 2023.
- Xue, F., Fu, Y., Zhou, W., Zheng, Z., and You, Y. To repeat or not to repeat: insights from scaling llm under token-crisis. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS ’23*, Red Hook, NY, USA, 2023. Curran Associates Inc.
- Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.
- Ye, Y., Huang, Z., Xiao, Y., Chern, E., Xia, S., and Liu, P. Limo: Less is more for reasoning, 2025. URL <https://arxiv.org/abs/2502.03387>.
- Yu, Q., Zhang, Z., Zhu, R., Yuan, Y., Zuo, X., Yue, Y., Dai, W., Fan, T., Liu, G., Liu, L., Liu, X., Lin, H., Lin, Z., Ma, B., Sheng, G., Tong, Y., Zhang, C., Zhang, M., Zhang, W., Zhu, H., Zhu, J., Chen, J., Chen, J., Wang, C., Yu, H., Song, Y., Wei, X., Zhou, H., Liu, J., Ma, W.-Y., Zhang, Y.-Q., Yan, L., Qiao, M., Wu, Y., and Wang, M. Dapo: An open-source llm reinforcement learning system at scale, 2025. URL <https://arxiv.org/abs/2503.14476>.
- Zhang, X., Li, C., Zong, Y., Ying, Z., He, L., and Qiu, X. Evaluating the performance of large language models on gaokao benchmark. *arXiv preprint arXiv:2305.12474*, 2023.



Zhao, Y., Yin, H., Zeng, B., Wang, H., Shi, T., Lyu, C., Wang, L., Luo, W., and Zhang, K. Marco-ol: Towards open reasoning models for open-ended solutions, 2024. URL <https://arxiv.org/abs/2411.14405>.

Zhu, T., Zhang, K., and Wang, W. Y. Embracing ai in education: Understanding the surge in large language model use by secondary students, 2024. URL <https://arxiv.org/abs/2411.18708>.