LEARNING WITH INTERACTION: AGENTIC DISTILLATION FOR LARGE LANGUAGE MODEL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advancements in large language models (LLMs) have demonstrated remarkable reasoning abilities to solve complex tasks, which has propelled the progress toward artificial general intelligence (AGI). However, these gains come with significant computational costs, limiting their practical deployment. A promising direction is to distill reasoning skills from larger teacher models into smaller, more efficient student models, yet existing data-centric distillation approaches suffer from passive learning, over-learning on simple tasks, and persistent knowledge gaps. To overcome these limitations, we introduce AGENTIC DISTILLATION, a novel framework for adaptive and active distillation. In AGENTIC DISTILLATION, student LLMs interact with teacher LLMs modeled as environments, receiving feedback tokens to guide their reasoning process and selectively updating their capabilities when necessary. To address the off-policy and gradient vanishing challenges introduced by feedback tokens, we devise a tailored importance sampling and clipping strategy within a unified objective that both incentivizes reasoning and injects knowledge into student LLMs. Extensive experiments show that AGENTIC DISTILLATION significantly enhances reasoning performance while improving efficiency, offering a scalable path for equipping compact LLMs with advanced reasoning abilities.

1 INTRODUCTION

In recent years, large language models (LLMs) have undergone rapid advancements, showcasing exceptional performance across various natural language processing tasks (Pu et al., 2023; Zhang & Soh, 2024; Gupta et al., 2024; Xu et al., 2024). Especially, LLMs employing long chain-of-thought (CoT) reasoning ability, which have demonstrated remarkable proficiency in solving complex tasks spanning mathematics, coding, and science, significantly advancing progress toward artificial general intelligence (AGI) (OpenAI, 2024a;b; 2025; DeepSeek-AI et al., 2025; Kimi-Team et al., 2025; Yang et al., 2025; Comanici et al., 2025; Huang & Yang, 2025; xAI, 2025).

However, the enhanced model reasoning capability introduces increased computational costs. The growth in model parameters and the extended length of CoT reasoning elevate computational demands, limiting practical applications (Chen et al., 2025). Therefore, *equipping more efficient small language models with robust reasoning capabilities via learning from stronger large models* has garnered significant attention from researchers and the broader community (DeepSeek-AI et al., 2025; Wen et al., 2025; Muennighoff et al., 2025; Guha et al., 2025; Ye et al., 2025).

A prevalent approach is *data-centric distillation*, which employs *rejection sampling* (Wang et al., 2024; Yang et al., 2024b; Shao et al., 2024; Ying et al., 2024) to generate training trajectory-level data for distilling student models. This method generates multiple reasoning trajectories from stronger LLMs for a given query, selects those with correct conclusions, and uses them to train the student

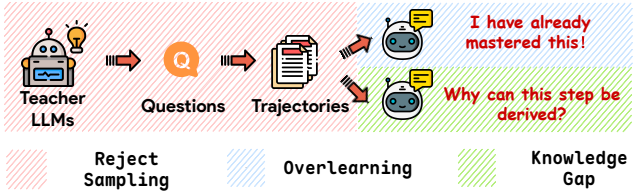

Figure 1: Illustration of *over-learning* and *knowledge grap* issues in data-centric distillation.

model via supervised finetuning (SFT). (Qin et al., 2024; DeepSeek-AI et al., 2025) or reinforcement learning (RL) (Zhang et al., 2025a). This enables student models to acquire the knowledge and reasoning capabilities of teacher LLMs. However, this approach has notable limitations due to its characteristic of **passive learning** as shown in Figure 1: 1) **Over-Learning**: Training on static and complete trajectories from teacher LLMs does not dynamically adapt to the evolving capabilities of the student model, often leading to over-learning on simpler questions (Chu et al., 2025) and wasting data and training resources on mastered questions.; 2) **Knowledge Gap**: Teacher-centered data generation may overlook the specific knowledge requirements and capability gaps of student LLMs (Liu et al., 2024a). Analogous to a classroom setting where generalized teaching overlooks students' individual knowledge deficiencies. Some reasoning steps that seem obvious to the teacher LLMs may be difficult for the student LLMs to understand and learn. This causes the student model to merely mimic the teacher's output format rather than truly acquiring reasoning ability (Chu et al., 2025; Kirk et al., 2024; Wu et al., 2025).

To address these limitations, we propose AGENTIC DISTILLATION, a novel framework for distilling knowledge and reasoning capabilities from strong LLMs into smaller student LLMs through **active** and **adaptive interaction**. Unlike traditional distillation methods that passively transfer knowledge, AGENTIC DISTILLATION empowers the student LLM to dynamically determine when to query the teacher LLM during reasoning, seeking feedback only when necessary as shown in Figure 2. This enables the student LLM to refine its reasoning process based on teacher feedback, leading to more accurate outcomes. Additionally, we design a mechanism to allow the student LLM to effectively learn essential knowledge and reasoning abilities from the teacher LLM's feedback.

Figure 2: Simple Illustration of AGENTIC DISTILLATION.

To tackle off-policy and gradient vanishing issues inherent in learning from feedback tokens, we introduce a tailored importance sampling coefficient and clipping strategy. They are seamlessly integrated into a unified objective that both incentivizes reasoning and injects knowledge into student LLMs. Notably, recent works (e.g., Search-R1) (Wang et al., 2025; Singh et al., 2025; Jin et al., 2025; Liu et al., 2025) primarily focus on enhancing LLMs' interactions with external environments (e.g., tools), often overlooking the rich information embedded in feedback, which can be utilized to improve the reasoning ability of student LLMs themselves. In contrast, AGENTIC DISTILLATION leverages teacher feedback as a direct learning signal, enabling continuous improvement of the student model. Even without interaction during inference, AGENTIC DISTILLATION-trained student LLMs can successfully reason on previously unsolvable tasks.

We conduct extensive experiments to validate the effectiveness of AGENTIC DISTILLATION. For example, AGENTIC DISTILLATION enhances the performance of Qwen2.5-7B-Instruct on mathematical reasoning benchmarks, achieving an average improvement of approximately 4 points over baseline distillation strategies. Significant gains are also observed on out-of-domain benchmarks, demonstrating AGENTIC DISTILLATION's robust generalization. Additional experiments confirm that AGENTIC DISTILLATION generalizes effectively across various student and teacher LLMs. Additionally, we investigate whether AGENTIC DISTILLATION expands the knowledge boundaries of student LLMs. Analysis of training dynamics and student LLM responses reveals that AGENTIC DISTILLATION enables student LLMs to effectively acquire new knowledge and capabilities, aligning their reasoning abilities with those of teacher LLMs.

2 METHOD

In this section, we introduce the motivation to propose the AGENTIC DISTILLATION framework for adaptive and active distillation (§ 2.1). Then, we introduce the details of proposed AGENTIC DISTILLATION (§§ 2.2 and 2.3)

2.1 PRELIMINARIES

Distillation from Strong LLMs. In a typical LLM reasoning task, given a question q from the question distribution $q \sim P(Q)$, the LLM π_θ is prompted with an instruction I to generate an answer

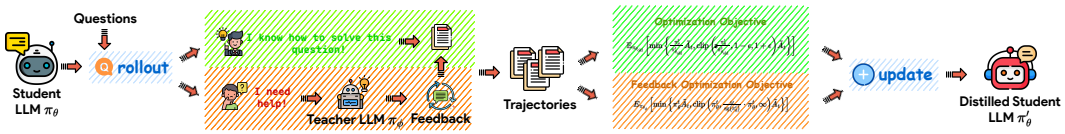

Figure 3: Illustration of AGENTIC DISTILLATION. In the AGENTIC DISTILLATION framework, during each rollout process, the student LLM initially attempts to solve a given question independently. If the student LLM fails to resolve the question, it engages in *external interaction* by querying the teacher LLM for feedback. Otherwise, the student LLM proceeds with its reasoning to derive the final answer. Subsequently, we compute the reward and optimize the student LLM using losses derived from both internal tokens and feedback tokens, respectively.

a':

$$a' \leftarrow \pi_\theta(\cdot \mid q, I). \tag{1}$$

For chain-of-thought based LLM reasoning, the reasoning process involves a step-by-step sequence, typically enclosed within tags such as `<think>` and `</think>`, represented by the token sequence $\boldsymbol{\tau}_{\text{thinking}}$. This culminates in a final conclusion $\boldsymbol{\tau}_{\text{conclusion}}$, which includes the predicted answer a':

$$[\boldsymbol{\tau}_{\text{thinking}}, \boldsymbol{\tau}_{\text{conclusion}}] \leftarrow \pi_\theta(\cdot \mid q, I). \tag{2}$$

The objective of this paper is to distill knowledge and capabilities from a strong teacher LLM π_ϕ^t to enhance a student LLM π_θ^s:

$$\pi_\theta^{s'} \leftarrow \mathtt{D}(\pi_\theta^s, \pi_\phi^t, q), \tag{3}$$

where $\mathtt{D}$ represents the distillation method, such as passive data-centric distillation (Qin et al., 2024; DeepSeek-AI et al., 2025; Wen et al., 2025) or the AGENTIC DISTILLATION proposed in this paper.

Data-Centric Distillation. The predominant distillation approach is *rejection sampling* (Yang et al., 2024b; Shao et al., 2024; Guha et al., 2025; Wen et al., 2025). Specifically, given a question set $\mathcal{Q}$, a strong teacher LLM π_t generates predictions $\mathcal{T}$ for each question $q \in \mathcal{Q}$:

$$\boldsymbol{\tau} \sim \pi_\phi^t(\cdot \mid q, I), \quad \boldsymbol{\tau} \in \mathcal{T},\, q \in \mathcal{Q}. \tag{4}$$

The prediction set $\mathcal{T}$ is then filtered based on the correctness of each prediction:

$$\mathcal{T}' = \Big\{ \boldsymbol{\tau} \mid \mathbb{I}(a, a') \Big\}, \tag{5}$$

where a denotes the ground truth answer to q and $\mathbb{I}$ is an indicator function that returns 1 only when the prediction is correct. The selected predictions $\mathcal{T}'$ are used to train the student LLM π_s:

$$\mathcal{L}(\theta) = \mathbb{E}_{q \in \mathcal{Q}, \boldsymbol{\tau} \sim \mathcal{T}'} \Big[-\log \pi_\theta^s(\boldsymbol{\tau} \mid q;\ \theta) \Big], \tag{6}$$

where θ denotes the parameters of the student LLM π_s.

Distillation from Interaction. To address the limitations of passive data-centric distillation as mentioned in § 1, we propose a novel approach that distills knowledge through active interaction with the teacher LLM. Specifically, we augment the reasoning process $\boldsymbol{\tau}_{\text{thinking}}$ to include multiple turns of interaction, comprising *queries to the teacher LLM* $\boldsymbol{\tau}_q$ and *external feedback from the teacher LLM* $\boldsymbol{\tau}_o$, formally expressed as:

$$\big[\ldots, \boldsymbol{\tau}_{q,(1)}, \boldsymbol{\tau}_{o,(1)}, \ldots, \boldsymbol{\tau}_{q,(N)}, \boldsymbol{\tau}_{o,(N)}, \ldots\big] \leftarrow \boldsymbol{\tau}_{\text{thinking}}. \tag{7}$$

The mechanisms governing interaction with the teacher LLM and the process of learning from its feedback are detailed in § 2.2 and § 2.3, respectively.

2.2 AGENTIC INTERACTION WITH TEACHER LLMS.

To distill knowledge from the teacher LLM, we design an agentic interaction mechanism that enables the student LLM to actively and flexibly interact with the teacher during the reasoning process.

When faced with a question q, the student LLM first performs basic reasoning using its internal knowledge, such as problem decomposition, solution planning, and simple arithmetic operations (Wei

et al., 2022). If the student LLM can solve the question using only its own knowledge and reasoning abilities, we argue that external knowledge distillation from an *oracle* is unnecessary. This important distinction is often overlooked by typical SFT-based methods (Qin et al., 2024; Huang et al., 2024; Muennighoff et al., 2025; Guha et al., 2025). Conversely, during reasoning, when the student recognizes that a (sub-)question exceeds the limits of its internal knowledge, it must refer to external oracle information. In such cases, we allow the student LLM to query the teacher LLM in natural language.

Specifically, we provide the student LLM with the prompt shown in Prompt 2.1 (full version is provided in Prompt A.1), which instructs it to enclose natural language queries to the teacher within `<query>` and `</query>` tags. The teacher LLM then responds with the corresponding answer (i.e., *feedback* or *observation*), appended to the student's reasoning process within `<result>` and `</result>` tags. To avoid meaningless or inefficient loops, we also impose an interaction budget limiting the number of queries the student may direct to the teacher.

Prompt 2.1: Prompt to Equip Student LLM with Agentic Interaction Capability

Reasoning Process

- **Decomposition:** Break down the user's question into a logical, step-by-step sequence of reasoning. Start from the most basic facts and build upon them.
- **External Inquiry (Optional but Encouraged):**
  - You may issue up to *max_turns* queries to an External Environment to validate hypotheses, clarify information, or advance your reasoning.
  - Each query must be a self-contained question enclosed in `<query>...</query>` tags.
  - **Wait for the `<result>...</result>` block** from the environment before continuing your reasoning.
  - **Critically analyze and integrate** the content from the `<result>...</result>` block into your reasoning chain.
  - Do not invent, assume, or hallucinate any `<result>` content. Your reasoning must be grounded in the provided results.

2.3 Learning from Agentic Interaction

This section addresses learning from agentic interaction. Given a query τ_v generated by the student LLM and feedback τ_o provided by the teacher LLM, prior RL approaches typically exclude τ_o from the loss calculation (Song et al., 2025; Liu et al., 2025), as the student LLM is not expected to generate tokens from the external environment. In contrast, our approach integrates feedback tokens $\tau_{o,(1:T_o)}$ into the RL policy loss to enable the student LLM to acquire new knowledge and capabilities.

The classical clipped surrogate objective is defined as:

$$\mathcal{J}(\theta) = \mathbb{E}_{q\sim P,\{\boldsymbol{\tau}_i\}\sim\pi_\theta}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|\boldsymbol{\tau}_i|}\sum_{t=1}^{|\boldsymbol{\tau}_i|}\Big\{\min\Big(\rho_{i,t}\tilde{A}_t^i,\ \text{clip}(\rho_{i,t},1-\epsilon,1+\epsilon)\,\tilde{A}_t^i\Big) - \beta D_{\text{KL}}\big[\pi_\theta\|\pi_{\text{ref}}\big]\Big\}\Bigg], \tag{8}$$

where the importance sampling coefficient for each token $\boldsymbol{\tau}_{i,(t)}$ at index t is given by:

$$\rho_{i,t} = \frac{\pi_\theta\left(\boldsymbol{\tau}_{i,(t)} \mid \boldsymbol{\tau}_{i,(<t)}\right)}{\pi_{\theta_{\text{old}}}\left(\boldsymbol{\tau}_{i,(t)} \mid \boldsymbol{\tau}_{i,(<t)}\right)}, \tag{9}$$

and $\pi_{\theta_{\text{old}}}$ denotes the previous policy of the student LLM. Directly applying this loss to feedback tokens $\tau_{o,(1:T_o)}$ may introduce the off-policy error due to the mismatch between feedback tokens and the student LLM's policy, which can destabilize RL training (Schulman et al., 2017; Zhang et al., 2025a).

Amending Importance Sampling Coefficient for Feedback Tokens. To mitigate the off-policy error, we introduce a modified importance sampling coefficient $\tilde{\rho}$. Within the standard clipped

surrogate loss, the off-policy error stems from sampling the trajectory τ from $\pi_{\theta_{\text{old}}}$:

$$\mathcal{J}(\theta) = \mathbb{E}_{\boldsymbol{\tau}\sim\pi_{\theta_{\text{old}}}}\left[\frac{1}{|\boldsymbol{\tau}|}\sum_{t=1}^{|\boldsymbol{\tau}|}\frac{\pi_\theta\big(\boldsymbol{\tau}_{(t)}|\boldsymbol{\tau}_{(\leq t)}\big)}{\pi_{\theta_{\text{old}}}\big(\boldsymbol{\tau}_{(t)}|\boldsymbol{\tau}_{(\leq t)}\big)}\tilde{A}_t\right], \tag{10}$$

where clipped and KL-penalty terms are omitted for simplicity. However, feedback tokens follow the distribution π_ϕ, defined by the teacher LLM, leading to the modified objective:

$$\mathcal{J}(\theta) = \mathbb{E}_{\pi_{\theta_{\text{old}}}}\left[\frac{1}{N}\sum_{\boldsymbol{\tau}_{(t)}\notin\boldsymbol{\tau}_o}\frac{\pi_\theta\big(\boldsymbol{\tau}_{(t)}|\boldsymbol{\tau}_{(\leq t)}\big)}{\pi_{\theta_{\text{old}}}\big(\boldsymbol{\tau}_{(t)}|\boldsymbol{\tau}_{(\leq t)}\big)}\tilde{A}_t\right] + \mathbb{E}_{\pi_\phi}\left[\frac{1}{M}\sum_{\boldsymbol{\tau}_{(t)}\in\boldsymbol{\tau}_o}\frac{\pi_\theta\big(\boldsymbol{\tau}_{(t)}|\boldsymbol{\tau}_{(\leq t)}\big)}{\pi_\phi\big(\boldsymbol{\tau}_{(t)}|\boldsymbol{\tau}_{(\leq t)}\big)}\tilde{A}_t\right], \tag{11}$$

where N and M are the number of non-feedback tokens and feedback tokens, respectively. Directly using the teacher LLM's distribution to compute the importance sampling coefficient is a straightforward approach but has two limitations: 1) vocabulary differences between the teacher and student LLMs may cause inconsistent distributions, and 2) computing the teacher LLM's distribution incurs additional computational overhead. To address these, we propose treating the teacher LLM's distribution as a one-hot distribution, yielding:

$$\mathcal{J}(\theta) = \mathbb{E}_{\pi_{\theta_{\text{old}}}}\left[\frac{1}{N}\sum_{\boldsymbol{\tau}_{(t)}\notin\boldsymbol{\tau}_o}\frac{\pi_\theta\big(\boldsymbol{\tau}_{(t)}|\boldsymbol{\tau}_{(\leq t)}\big)}{\pi_{\theta_{\text{old}}}\big(\boldsymbol{\tau}_{(t)}|\boldsymbol{\tau}_{(\leq t)}\big)}\tilde{A}_t\right] + \mathbb{E}_{\pi_\phi}\left[\frac{1}{M}\sum_{\boldsymbol{\tau}_{(t)}\in\boldsymbol{\tau}_o}\pi_\theta\big(\boldsymbol{\tau}_{(t)}|\boldsymbol{\tau}_{(\leq t)}\big)\tilde{A}_t\right]. \tag{12}$$

This method employs a temperature coefficient to sharpen the teacher LLM's distribution, reducing computational complexity and resolving vocabulary inconsistencies.

Gradient Vanishing for Feedback Tokens. The standard surrogate objective employs a clipping mechanism on the importance sampling coefficient to prevent excessive policy deviation from the previous policy. However, for feedback tokens, the importance sampling coefficient π_θ is inherently bounded due to the softmax activation. Consequently, we remove the standard clipping mechanism for feedback tokens. The gradient of these tokens can be computed as

$$\pi_\theta \cdot \tilde{A}_t \cdot \nabla_\theta \cdot \log \pi_\theta. \tag{13}$$

Nevertheless, when the probability of a feedback token $\tau_{o,(t)}$ in the student LLM's policy is low ($\pi_\theta(\tau_{o,(t)}) \to 0$), the gradient approaches zero ($\pi_\theta \cdot \tilde{A}_t \cdot \nabla_\theta \log \pi_\theta \to 0$). This vanishing gradient leads to suboptimal learning, particularly for off-policy feedback tokens from the teacher LLM, which are critical for the student LLM to learn effectively. These tokens often have low probabilities in the student LLM's policy, exacerbating the vanishing gradient issue and hindering knowledge transfer.

Clipping Strategy for Feedback Tokens. To address the vanishing gradient problem, we propose a clipping strategy inspired by the standard mechanism (Schulman et al., 2017):

$$\text{clip}\left(\pi_\theta,\ \frac{\omega}{\text{sg}\left(\pi_\theta\right)}\cdot\pi_\theta,\ \infty\right), \tag{14}$$

where ω is a clipping hyperparameter and $\text{sg}(\cdot)$ denotes the stop-gradient operation. This approach sets a lower bound on the importance sampling coefficient for feedback tokens, with the $\pi_\theta/\text{sg}(\pi_\theta)$ term ensuring numerical equivalence. The resulting gradients are:

$$\begin{cases}\pi_\theta \cdot \tilde{A}_t \cdot \nabla_\theta \cdot \log \pi_\theta, \text{ if } \pi_\theta \geq \omega, \\ \omega \cdot \tilde{A}_t \cdot \nabla_\theta \cdot \log \pi_\theta, \text{ if } 0 \leq \pi_\theta < \omega.\end{cases} \tag{15}$$

This ensures that feedback tokens with high advantage maintain non-vanishing gradients, mitigating the impact of policy deviation.

Final Objective. By integrating the modified importance sampling coefficient and the proposed clipping strategy for mitigating the off-policy and the vanishing gradient issues, we formulate the final objective for optimizing the student LLM:

$$\begin{aligned}\mathcal{J}(\theta) = \mathbb{E}_{\pi_{\theta_{\text{old}}}}&\left[\min\left\{\frac{\pi_\theta^t}{\pi_{\theta_{\text{old}}}^t}\tilde{A}_t, \text{clip}\left(\frac{\pi_\theta^t}{\pi_{\theta_{\text{old}}}^t}, 1-\epsilon, 1+\epsilon\right)\tilde{A}_t\right\}\right] \\ &+ \mathbb{E}_{\pi_\phi}\left[\min\left\{\pi_\theta^t\tilde{A}_t, \text{clip}\left(\pi_\theta^t, \frac{\omega}{\text{sg}\left(\pi_\theta^t\right)}\cdot\pi_\theta^t, \infty\right)\tilde{A}_t\right\}\right],\end{aligned} \tag{16}$$

where $\pi_\theta^t = \pi_\theta(\tau_{(t)}|\tau_{(\leq t)})$. Intuitively, the objective of AGENTIC DISTILLATION unifies the RLVR and SFT in a single function. For action tokens autonomously generated by student LLMs, such as problem decomposition, solving, and query formulation, we employ the standard RLVR objective function for optimization (*first part*). For feedback tokens provided by teacher LLMs, we adopt an SFT-inspired optimization objective, enhanced by advantage and clipping-controlled update (*second part*), to effectively inject new knowledge into the student LLMs.

3 EXPERIMENTS

3.1 SETUP

Baselines. We compare AGENTIC DISTILLATION against several representative LLM post-training methods: ❶ *Supervised Fine-Tuning (SFT)*: Utilizes teacher LLM-generated data through rejection sampling; ❷ *Vanilla Reinforcement Learning (RL)*: Trains the student LLM using the GRPO algorithm (Shao et al., 2024) without external environment interactions; ❸ *Reinforcement Learning with Supervised Fine-Tuning (RL+SFT)*: Combines GRPO training (Shao et al., 2024) with data generated via rejection sampling; ❹ *Reinforcement Learning with Masked Interaction (RL+MI)*: Employs the GRPO algorithm (Shao et al., 2024) with teacher LLM interactions, but excludes feedback tokens from loss computation.

Evaluation Benchmarks. We evaluated all models across four domain-specific benchmarks: ❶ *Mathematical Reasoning*: Includes AIME24, AIME25, MATH500 (Hendrycks et al., 2021), and LiveMathBench (Liu et al., 2024b); ❷ *Scientific Reasoning*: Represented by GPQA-Diamond (Rein et al., 2023); ❸ *Code Reasoning*: Comprises MBPP (Austin et al., 2021) and LiveCodeBench (Jain et al., 2025); ❹ *Puzzle Reasoning*: Includes puzzles from Reasoning-Gym (Stojanovski et al., 2025).

Implementation Details. We conducted experiments on the Qwen-2.5 series models (Yang et al., 2024a) and Llama-3.2 series models (Dubey et al., 2024), distilling from two prominent teacher LLMs from distinct families: Qwen3-30B-A3B-Instruct-2507 (Yang et al., 2025) and GPT-OSS-20B (Agarwal et al., 2025). The training corpus, sourced from DAPO (Yu et al., 2025), OpenScienceReasoning-2 [1], and Reasoning Gym (Stojanovski et al., 2025), consists of approximately 60,000 high-quality reasoning-intensive samples. Models were trained for 200 steps with a batch size of 256 and a group size of 8, selecting the best model based on validation performance. During each generation, the student LLM was allowed up to three interactions with the teacher LLM. Training was performed using the veRL (Sheng et al., 2025) and vLLM (Kwon et al., 2023) frameworks. For evaluation, we set the sampling temperature to 1.0, top-p to 1.0, top-k to -1, and the maximum generation tokens to 16384. To reduce variance, we report average performance relative to the size of each benchmark. And the prompt utilized in inference phase is shown in Prompt A.2.

3.2 MAIN RESULTS AND ANALYSIS

Table 1 illustrates the performance of AGENTIC DISTILLATION and baselines on different benchmarks, containing different student LLMs and teacher LLMs. From the experimental results, we have the following findings.

AGENTIC DISTILLATION Outperforms Baseline Methods. As illustrated in Table 1, AGENTIC DISTILLATION surpasses other training strategies, including supervised fine-tuning (SFT), vanilla reinforcement learning (RL), RL combined with SFT (RL+SFT), and RL with masked interaction (RL+MI). Notably, AGENTIC DISTILLATION achieves significant improvements on challenging reasoning benchmarks such as AIME24 and AIME25, with average accuracy gains of 4-6 points over the strongest baseline. Comparable enhancements are observed across science, code, and puzzle tasks, underscoring AGENTIC DISTILLATION's robustness in improving reasoning capabilities across diverse task settings.

AGENTIC DISTILLATION Enhances Performance Across Diverse Student LLMs. As depicted in Table 1, AGENTIC DISTILLATION consistently outperforms baseline methods across various student LLMs, including Qwen-2.5-7B-Instruct and Llama-3.2-3B-Instruct. The framework achieves

[1] https://huggingface.co/datasets/nvidia/OpenScienceReasoning-2

Table 1: Experimental results of AGENTIC DISTILLATION and baselines with Qwen2.5-7B-Instrcut as the student LLM. We report the average performance for 16 runs on AIME24 and AIME25, and 4 runs on others. We abbreviate LMB as LiveMathBench v202505, LCB as LiveCodeBench v6, RG as Reasoning Gym, MI as Masked Interaction, and AD as AGENTIC DISTILLATION. ♠ denotes the in-domain evaluation benchmark and ♣ denotes the out-of-domain benchmark. We provide performance of teacher LLMs in Table 4

Methods	Math ♠				Science ♠	Code ♣		Puzzle ♠
	AIME24 Avg@16	AIME25 Avg@16	MATH500 Avg@4	LMB Avg@4	GPQA-D Avg@4	MBPP Avg@4	LCB Avg@4	RG Avg@4
Student LLM: *Qwen2.5-7B-Instruct*, Teacher LLM: *Qwen3-30B-A3B-Instruct-2507*								
Original	9.79	7.50	73.00	10.75	33.33	58.66	15.71	9.63
+SFT	11.67	13.54	75.80	12.00	22.35	41.34	11.04	18.98
+RL	12.08	10.00	75.35	10.50	34.72	58.27	16.33	19.81
+RL+SFT	13.01	12.13	75.77	11.32	35.20	59.22	15.34	19.22
+RL+MI	11.25	6.46	72.80	10.25	34.47	58.17	14.55	9.17
+AD	**14.82**	**14.33**	**78.17**	**14.27**	**37.13**	**62.73**	**18.62**	**21.11**
Student LLM: *Qwen2.5-7B-Instruct*, Teacher LLM: *GPT-OSS-20B*								
+SFT	14.67	13.54	75.80	12.00	37.34	61.04	**22.35**	18.98
+RL	12.08	10.00	75.35	10.50	34.72	58.27	16.33	19.81
+RL+SFT	13.01	12.13	75.77	11.32	35.20	59.22	15.34	19.22
+RL+MI	12.31	9.26	71.89	11.33	31.25	56.25	13.79	10.25
+AD	**16.52**	**17.47**	**81.22**	**16.29**	**38.53**	**64.15**	20.27	**24.32**
Student LLM: *Llama-3.2-3B-Instruct*, Teacher LLM: *Qwen3-30B-A3B-Instruct-2507*								
Original	2.50	1.20	30.10	3.00	23.48	42.61	6.87	9.44
+SFT	5.50	2.67	**48.65**	5.50	22.98	43.39	3.95	13.89
+RL	8.96	1.12	44.10	7.00	25.76	52.59	10.92	13.70
+RL+SFT	6.24	3.62	41.25	8.11	26.45	53.45	10.23	12.44
+RL+MI	7.44	2.98	45.11	6.45	24.26	54.27	9.84	10.52
+AD	**10.38**	**4.42**	44.45	**9.00**	**28.54**	**58.66**	**15.70**	**16.85**

stable improvements across LLMs of different architectures and sizes, highlighting the generality of AGENTIC DISTILLATION and its potential for broad application to diverse LLM families and types.

AGENTIC DISTILLATION Improves Across Different Teacher LLMs. AGENTIC DISTILLATION consistently delivers performance improvements across different teacher models, including the short-cot based reasoning LLMs *Qwen3-30B-A3B-Instruct-2507* and the long-cot based reasoning LLMs *GPT-0SS-20B*. While baseline methods exhibit variability depending on the teacher LLM, AGENTIC DISTILLATION maintains superior results, indicating that its adaptive training mechanism is independent of the teacher model. This stability highlights AGENTIC DISTILLATION's flexibility, making it suitable for scenarios with varying teacher quality or availability.

AGENTIC DISTILLATION Generalizes to Out-of-Domain Benchmarks. AGENTIC DISTILLATION also performs generalization across in-domain and out-of-domain benchmarks. On mathematics, science, and puzzle benchmarks, which align closely with the training data, AGENTIC DISTILLATION consistently outperforms all baselines. More notably, on out-of-domain benchmarks such as code (MBPP and LiveCodeBench), AGENTIC DISTILLATION achieves substantial gains, surpassing the strongest baseline in several instances. These results demonstrate that AGENTIC DISTILLATION not only excels in task-specific settings but also enables robust generalization across domains with distinct reasoning ability.

3.3 ABLATION STUDY

Impact of Importance Sampling Coefficient in AGENTIC DISTILLATION. To assess the necessity of the modified importance sampling coefficient for feedback tokens in AGENTIC DISTILLATION, as introduced in Equation (12), we compare its performance against the importance sampling coefficient

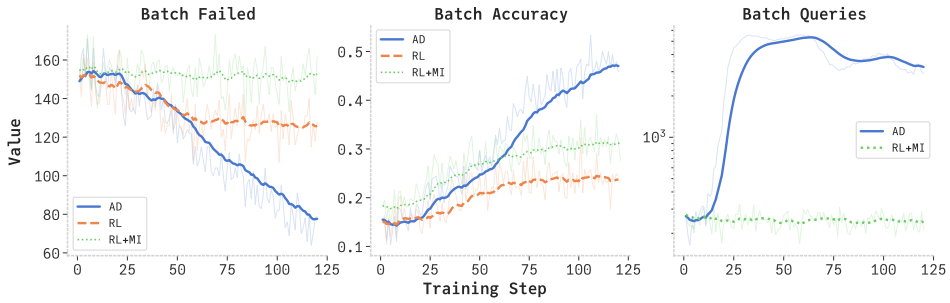

Figure 4: Training Dynamics of Reinforcement Learning, Reinforcement Learning with Masked Interaction, and AGENTIC DISTILLATION on Qwen2.5-7B-Instruct with Qwen3-30B-A3B-Instruct-2507 as the teacher LLM.

Table 2: Ablation study of AGENTIC DISTILLATION w.r.t. the modified importance sampling coefficient (abbreviated as IS) and clipping strategy (abbreviated as CS).

Methods	Math				Science	Code		Puzzle
	AIME24 Avg@16	AIME25 Avg@16	MATH500 Avg@4	LMB Avg@4	GPQA-D Avg@4	MBPP Avg@4	LCB Avg@4	RG Avg@4
Student LLM: *Qwen2.5-7B-Instruct*, Teacher LLM: *Qwen3-30B-A3B-Instruct-2507*								
AD	**14.82**	**14.33**	**78.17**	**14.27**	**37.13**	**62.73**	**18.62**	**21.11**
w/o IS	13.44	13.56	77.25	14.11	36.82	62.32	18.44	20.92
w/o CS	14.34	13.92	76.53	13.22	35.89	61.46	17.33	19.08

used in the vanilla reinforcement learning algorithm, as shown in Table 2. The results demonstrate that the proposed modified importance sampling coefficient consistently outperforms the vanilla RL approach, confirming its critical role in enhancing AGENTIC DISTILLATION's effectiveness.

Impact of Clipping Strategy in AGENTIC DISTILLATION. Similarly, we evaluate the clipping strategy proposed in Equation (14). As illustrated in Table 2, removing the clipping strategy leads to a substantial decline in model performance across all tested scenarios. This indicates that the clipping strategy effectively mitigates issues such as gradient vanishing, thereby significantly improving the performance of the student LLM.

3.4 DOES THE LLM LEARN NEW KNOWLEDGE AND CAPABILITIES THROUGH AGENTIC DISTILLATION?

To validate and elucidate the learning outcomes of AGENTIC DISTILLATION, we analyze its training dynamics and the expansion of the knowledge boundary of the LLM.

Analysis of Training Dynamics. Figure 4 illustrates the dynamics of three key metrics during the training of Qwen2.5-7B-Instruct, with Qwen3-30B-A3B-Instruct-2507 as the teacher LLM, using Reinforcement Learning (RL), Reinforcement Learning with Masked Interaction (RL+MI), and AGENTIC DISTILLATION (AD). First, we examine the proportion of problems in a batch that the student LLM fails to solve across all rollouts (*Batch Failed*). Training with AGENTIC DISTILLATION significantly reduces this proportion, indicating that AGENTIC DISTILLATION enables the student LLM to acquire new knowledge, allowing it to solve previously unsolvable problems. This improvement is mirrored in the training batch accuracy, where AGENTIC DISTILLATION-trained LLMs show markedly higher gains. Additionally, we analyze the number of queries raised by the student LLM per batch (*Batch Queries*). With AGENTIC DISTILLATION, the number of queries initially increases, then decreases, and eventually stabilizes. This trend suggests that early in training,

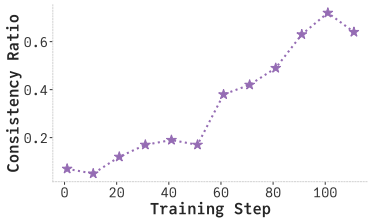

Figure 5: Knowledge Boundary Expansion of AGENTIC DISTILLATION-trained Student LLM.

the student LLM queries the teacher LLM frequently to learn new knowledge. As its knowledge boundary expands, the student LLM relies less on queries, solving problems independently. These metric dynamics demonstrate that AGENTIC DISTILLATION effectively facilitates the student LLM's acquisition of new knowledge and capabilities from the teacher LLM.

Knowledge Boundary Expansion of AGENTIC DISTILLATION-trained Student LLM. To further study the change of the knowledge boundary of the AGENTIC DISTILLATION-trained student LLM, we collect the queries raised by student LLM Qwen2.5-7B-Instruct during the training process with Qwen3-30B-A3B-Instruct-2507 as the teacher LLM. We let the student LLM at different training stages to answer these queries and report the consistency with the answers of the teacher LLM. The consistency is judged by the GPT-4o (OpenAI, 2023). As shown in Figure 5, we can observe that as training progresses, the consistency of the student LLM on these unsolvable problems that are beyond its knowledge boundary gradually aligns with that of the teacher LLM, indicating that AGENTIC DISTILLATION can effectively inject the knowledge of the teacher LLM into the student LLM.

4 RELATED WORK

Distilling from Strong LLMs. Recent advancements in LLMs have led to remarkable performance in complex reasoning tasks (OpenAI, 2024a;b; 2025; DeepSeek-AI et al., 2025; Kimi-Team et al., 2025; Yang et al., 2025; Comanici et al., 2025; Huang & Yang, 2025; xAI, 2025). However, these models are often closed-source or possess an excessively large number of parameters, limiting their practical applications. Consequently, recent research has focused on distilling the capabilities of these strong reasoning LLMs into smaller-scale LLMs. Early studies (Qin et al., 2024; Guan et al., 2025; DeepSeek-AI et al., 2025) demonstrated that a small dataset generated by strong LRMs can significantly enhance the reasoning performance of smaller LLMs. Subsequent works (Bespoke-Labs, 2025; NovaSky-Team, 2025; Ye et al., 2025; Wen et al., 2025; Guha et al., 2025; Yang et al., 2025) have further improved distillation by optimizing problem set quality, curated data, training methods, loss functions, and integration of training stages. These approaches typically rely on distilling complete reasoning trajectories, a passive learning method that often fails to address the specific capabilities and knowledge gaps of student LLMs. In contrast, our proposed method enables student LLMs to actively query strong LRMs and selectively learn knowledge beyond their current capabilities, offering a more efficient and effective distillation.

Enhancing LLM Reasoning with External Information. Despite the remarkable performance of LLMs in various reasoning tasks, their capabilities are limited by inherent knowledge constraints and the fundamental limitations of deep learning architectures, which hinder their effectiveness in certain real-world tasks (Wang et al., 2025; Yang et al., 2024b). Prior work has employed reinforcement learning algorithms to enhance LLM decision-making, equipping them with autonomous capabilities such as planning, reasoning, tool usage, memory maintenance, and self-reflection (Wang et al., 2025; Singh et al., 2025; Jin et al., 2025; Liu et al., 2025). These efforts have improved LLM performance in knowledge-intensive question answering (Jin et al., 2025; Song et al., 2025), mathematical reasoning (Li et al., 2025; Bai et al., 2025), planning (Liu et al., 2025), and real-world applications (Mialon et al., 2024; Zhang et al., 2025b). However, these methods primarily focus on enhancing LLMs' ability to utilize tools to improve task performance. In contrast, we propose a distillation approach that leverages interactions between student LLMs and an external environment, specifically, teacher LLMs, to enhance reasoning capabilities without relying on the external information during inference.

5 CONCLUSION

In this paper, we introduce AGENTIC DISTILLATION, a distillation framework that enables active and adaptive knowledge transfer from strong LLMs to smaller student models. AGENTIC DISTILLATION leverages interaction and feedback tokens from teacher LLMs, allowing student models to selectively refine their reasoning and bridge knowledge gaps. To tackle off-policy and gradient vanishing issues inherent in learning from feedback, we introduce a tailored importance sampling coefficient and clipping strategy that seamlessly integrate into the reinforcement learning objective. Extensive experiments demonstrate that AGENTIC DISTILLATION achieves consistent improvements in both in-domain and out-of-domain reasoning tasks. We believe our framework could provide a promising direction for equipping compact models with advanced reasoning abilities.

Ethics Statement

This research focuses exclusively on LLM research problems and poses no risks to safety, personal security, or privacy. No new datasets are released as part of this study. Additionally, the research does not encompass potentially harmful insights, methods, or applications, nor does it raise issues related to privacy, security, legal compliance, or research integrity. We foresee no ethical risks or conflicts of interest. We are committed to adhering to ethical guidelines throughout the research process.

Reproducibility Statement

We provide a comprehensive description of the proposed Agentic Distillation in § 2, with detailed implementation specifics provided in § A and § 3.1. All datasets utilized in this research are publicly available. Key code implementations are included in the supplementary materials for reference, and the complete code will be made publicly available upon acceptance of the paper.

References

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, Kai Chen, Mark Chen, Enoch Cheung, Aidan Clark, Dan Cook, Marat Dukhan, Casey Dvorak, Kevin Fives, Vlad Fomenko, Timur Garipov, Kristian Georgiev, Mia Glaese, Tarun Gogineni, Adam Goucher, Lukas Gross, Katia Gil Guzman, John Hallman, Jackie Hehir, Johannes Heidecke, Alec Helyar, Haitang Hu, Romain Huet, Jacob Huh, Saachi Jain, Zach Johnson, Chris Koch, Irina Kofman, Dominik Kundel, Jason Kwon, Volodymyr Kyrylov, Elaine Ya Le, Guillaume Leclerc, James Park Lennon, Scott Lessans, Mario Lezcano-Casado, Yuanzhi Li, Zhuohan Li, Ji Lin, Jordan Liss, Lily, Liu, Jiancheng Liu, Kevin Lu, Chris Lu, Zoran Martinovic, Lindsay McCallum, Josh McGrath, Scott McKinney, Aidan McLaughlin, Song Mei, Steve Mostovoy, Tong Mu, Gideon Myles, Alexander Neitz, Alex Nichol, Jakub Pachocki, Alex Paino, Dana Palmie, Ashley Pantuliano, Giambattista Parascandolo, Jongsoo Park, Leher Pathak, Carolina Paz, Ludovic Peran, Dmitry Pimenov, Michelle Pokrass, Elizabeth Proehl, Huida Qiu, Gaby Raila, Filippo Raso, Hongyu Ren, Kimmy Richardson, David Robinson, Bob Rotsted, Hadi Salman, Suvansh Sanjeev, Max Schwarzer, D. Sculley, Harshit Sikchi, Kendal Simon, Karan Singhal, Yang Song, Dane Stuckey, Zhiqing Sun, Philippe Tillet, Sam Toizer, Foivos Tsimpourlas, Nikhil Vyas, Eric Wallace, Xin Wang, Miles Wang, Olivia Watkins, Kevin Weil, Amy Wendling, Kevin Whinnery, Cedric Whitney, Hannah Wong, Lin Yang, Yu Yang, Michihiro Yasunaga, Kristen Ying, Wojciech Zaremba, Wenting Zhan, Cyril Zhang, Brian Zhang, Eddie Zhang, and Shengjia Zhao. gpt-oss-120b & gpt-oss-20b model card. *CoRR*, abs/2508.10925, 2025. 3.1

Jacob Austin, Augustus Odena, Maxwell I. Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie J. Cai, Michael Terry, Quoc V. Le, and Charles Sutton. Program synthesis with large language models. *CoRR*, abs/2108.07732, 2021. 3.1, A.5

Fei Bai, Yingqian Min, Beichen Zhang, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, Zheng Liu, Zhongyuan Wang, and Ji-Rong Wen. Towards effective code-integrated reasoning. *CoRR*, abs/2505.24480, 2025. 4

Bespoke-Labs. Bespoke-stratos: The unreasonable effectiveness of reasoning distillation. https://www.bespokelabs.ai/blog/bespoke-stratos-the-unreasonable-effectiveness-of-reasoning-distillation, 2025. Accessed: 2025-01. 4

Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *CoRR*, abs/2503.09567, 2025. 1

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. SFT memorizes, RL generalizes: A comparative study of foundation model post-training. *CoRR*, abs/2501.17161, 2025. 1

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit S. Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, Krishna Haridasan, Ahmed Omran, Nikunj Saunshi, Dara Bahri, Gaurav Mishra, Eric Chu, Toby Boyd, Brad Hekman, Aaron Parisi, Chaoyi Zhang, Kornraphop Kawintiranon, Tania Bedrax-Weiss, Oliver Wang, Ya Xu, Ollie Purkiss, Uri Mendlovic, Ilaï Deutel, Nam Nguyen, Adam Langley, Flip Korn, Lucia Rossazza, Alexandre Ramé, Sagar Waghmare, Helen Miller, Nathan Byrd, Ashrith Sheshan, Raia Hadsell Sangnie Bhardwaj, Pawel Janus, Tero Rissa, Dan Horgan, Sharon Silver, Ayzaan Wahid, Sergey Brin, Yves Raimond, Klemen Kloboves, Cindy Wang, Nitesh Bharadwaj Gundavarapu, Ilia Shumailov, Bo Wang, Mantas Pajarskas, Joe Heyward, Martin Nikoltchev, Maciej Kula, Hao Zhou, Zachary Garrett, Sushant Kafle, Sercan Arik, Ankita Goel, Mingyao Yang, Jiho Park, Koji Kojima, Parsa Mahmoudieh, Koray Kavukcuoglu, Grace Chen, Doug Fritz, Anton Bulyenov, Sudeshna Roy, Dimitris Paparas, Hadar Shemtov, Bo-Juen Chen, Robin Strudel, David Reitter, Aurko Roy, Andrey Vlasov, Changwan Ryu, Chas Leichner, Haichuan Yang, Zelda Mariet, Denis Vnukov, Tim Sohn, Amy Stuart, Wei Liang, Minmin Chen, Praynaa Rawlani, Christy Koh, JD Co-Reyes, Guangda Lai, Praseem Banzal, Dimitrios Vytiniotis, Jieru Mei, and Mu Cai. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *CoRR*, abs/2507.06261, 2025. 1, 4

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, and S. S. Li. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948, 2025. 1, 1, 2.1, 4

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024. 3.1

Xinyu Guan, Li Lyna Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. rstar-math: Small llms can master math reasoning with self-evolved deep thinking. *CoRR*, abs/2501.04519, 2025. 4

Etash Kumar Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, Ashima Suvarna, Benjamin Feuer,

Liangyu Chen, Zaid Khan, Eric Frankel, Sachin Grover, Caroline Choi, Niklas Muennighoff, Shiye Su, Wanjia Zhao, John Yang, Shreyas Pimpalgaonkar, Kartik Sharma, Charlie Cheng-Jie Ji, Yichuan Deng, Sarah M. Pratt, Vivek Ramanujan, Jon Saad-Falcon, Jeffrey Li, Achal Dave, Alon Albalak, Kushal Arora, Blake Wulfe, Chinmay Hegde, Greg Durrett, Sewoong Oh, Mohit Bansal, Saadia Gabriel, Aditya Grover, Kai-Wei Chang, Vaishaal Shankar, Aaron Gokaslan, Mike A. Merrill, Tatsunori Hashimoto, Yejin Choi, Jenia Jitsev, Reinhard Heckel, Maheswaran Sathiamoorthy, Alexandros G. Dimakis, and Ludwig Schmidt. Openthoughts: Data recipes for reasoning models. *CoRR*, abs/2506.04178, 2025. 1, 2.1, 2.2, 4

Shailja Gupta, Rajesh Ranjan, and Surya Narayan Singh. Comprehensive study on sentiment analysis: From rule-based to modern LLM based system. *CoRR*, abs/2409.09989, 2024. 1

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *NeurIPS Datasets and Benchmarks*, 2021. 3.1, A.5

Yichen Huang and Lin F. Yang. Gemini 2.5 pro capable of winning gold at IMO 2025. *CoRR*, abs/2507.15855, 2025. 1, 4

Zhen Huang, Haoyang Zou, Xuefeng Li, Yixiu Liu, Yuxiang Zheng, Ethan Chern, Shijie Xia, Yiwei Qin, Weizhe Yuan, and Pengfei Liu. O1 replication journey - part 2: Surpassing o1-preview through simple distillation, big progress or bitter lesson? *CoRR*, abs/2411.16489, 2024. 2.2

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. In *ICLR*. OpenReview.net, 2025. 3.1, A.5

Bowen Jin, Hansi Zeng, Zhenrui Yue, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *CoRR*, abs/2503.09516, 2025. 1, 4

Kimi-Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, Chuning Tang, Congcong Wang, Dehao Zhang, Enming Yuan, Enzhe Lu, Fengxiang Tang, Flood Sung, Guangda Wei, Guokun Lai, Haiqing Guo, Han Zhu, Hao Ding, Hao Hu, Hao Yang, Hao Zhang, Haotian Yao, Haotian Zhao, Haoyu Lu, Haoze Li, Haozhen Yu, Hongcheng Gao, Huabin Zheng, Huan Yuan, Jia Chen, Jianhang Guo, Jianlin Su, Jianzhou Wang, Jie Zhao, Jin Zhang, Jingyuan Liu, Junjie Yan, Junyan Wu, Lidong Shi, Ling Ye, Longhui Yu, Mengnan Dong, Neo Zhang, Ningchen Ma, Qiwei Pan, Qucheng Gong, Shaowei Liu, Shengling Ma, Shupeng Wei, Sihan Cao, Siying Huang, Tao Jiang, Weihao Gao, Weimin Xiong, Weiran He, Weixiao Huang, Wenhao Wu, Wenyang He, Xianghui Wei, Xianqing Jia, Xingzhe Wu, Xinran Xu, Xinxing Zu, Xinyu Zhou, Xuehai Pan, Y. Charles, Yang Li, Yangyang Hu, Yangyang Liu, Yanru Chen, Yejie Wang, Yibo Liu, Yidao Qin, Yifeng Liu, Ying Yang, Yiping Bao, Yulun Du, Yuxin Wu, Yuzhi Wang, Zaida Zhou, Zhaoji Wang, Zhaowei Li, Zhen Zhu, Zheng Zhang, Zhexu Wang, Zhilin Yang, Zhiqi Huang, Zihao Huang, Ziyao Xu, and Zonghan Yang. Kimi k1.5: Scaling reinforcement learning with llms. *CoRR*, abs/2501.12599, 2025. 1, 4

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of RLHF on LLM generalisation and diversity. In *ICLR*. OpenReview.net, 2024. 1

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *SOSP*, pp. 611–626. ACM, 2023. 3.1, A.1

Xuefeng Li, Haoyang Zou, and Pengfei Liu. Torl: Scaling tool-integrated RL. *CoRR*, abs/2503.23383, 2025. 4

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *ICLR*. OpenReview.net, 2024. A.5

Jiaheng Liu, Chenchen Zhang, Jinyang Guo, Yuanxing Zhang, Haoran Que, Ken Deng, Zhiqi Bai, Jie Liu, Ge Zhang, Jiakai Wang, Yanan Wu, Congnan Liu, Jiamang Wang, Lin Qu, Wenbo Su, and Bo Zheng. DDK: distilling domain knowledge for efficient large language models. In *NeurIPS*, 2024a. 1

Junnan Liu, Hongwei Liu, Linchen Xiao, Ziyi Wang, Kuikun Liu, Songyang Gao, Wenwei Zhang, Songyang Zhang, and Kai Chen. Are your llms capable of stable reasoning? *CoRR*, abs/2412.13147, 2024b. 3.1, A.5

Junnan Liu, Linhao Luo, Thuy-Trang Vu, and Gholamreza Haffari. Situatedthinker: Grounding LLM reasoning with real-world through situated thinking. *CoRR*, abs/2505.19300, 2025. 1, 2.3, 4

Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. GAIA: a benchmark for general AI assistants. In *ICLR*. OpenReview.net, 2024. 4

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel J. Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *CoRR*, abs/2501.19393, 2025. 1, 2.2

NovaSky-Team. Think less, achieve more: Cut reasoning costs by 50 https://novasky-ai.github.io/posts/reduce-overthinking, 2025. Accessed: 2025-01. 4

OpenAI. Gpt-4 is openai's most advanced system, producing safer and more useful responses. https://openai.com/index/gpt-4/, 2023. 3.4

OpenAI. Learning to reason with llms. https://openai.com/index/learning-to-reason-with-llms/, 2024a. Accessed: 2024-09. 1, 4

OpenAI. Introducing openai o3 and o4-mini. https://openai.com/index/introducing-o3-and-o4-mini/, 2024b. Accessed: 2024-12. 1, 4

OpenAI. Gpt-5 and the new era of work. https://openai.com/index/gpt-5-new-era-of-work/, 2025. Accessed: 2025-08. 1, 4

Xiao Pu, Mingqi Gao, and Xiaojun Wan. Summarization is (almost) dead. *CoRR*, abs/2309.09558, 2023. 1

Yiwei Qin, Xuefeng Li, Haoyang Zou, Yixiu Liu, Shijie Xia, Zhen Huang, Yixin Ye, Weizhe Yuan, Hector Liu, Yuanzhi Li, and Pengfei Liu. O1 replication journey: A strategic progress report - part 1. *CoRR*, abs/2410.18982, 2024. 1, 2.1, 2.2, 4

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. *CoRR*, abs/2311.12022, 2023. 3.1, A.5

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017. 2.3, 2.3

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024. 1, 2.1, 3.1

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient RLHF framework. In *EuroSys*, pp. 1279–1297. ACM, 2025. 3.1, A.1

Joykirat Singh, Raghav Magazine, Yash Pandya, and Akshay Nambi. Agentic reasoning and tool integration for llms via reinforcement learning. *CoRR*, abs/2505.01441, 2025. 1, 4

Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. R1-searcher: Incentivizing the search capability in llms via reinforcement learning. *CoRR*, abs/2503.05592, 2025. 2.3, 4

Zafir Stojanovski, Oliver Stanley, Joe Sharratt, Richard Jones, Abdulhakeem Adefioye, Jean Kaddour, and Andreas Köpf. REASONING GYM: reasoning environments for reinforcement learning with verifiable rewards. *CoRR*, abs/2505.24760, 2025. 3.1, A.4

Jun Wang, Yang Chen, Shuicheng Yan, Heng Ji, Zihu Wang, Xiaohang Yu, Yifan Zhou, Philip Torr, Hongru Wang, Zhenfei Yin, Guibin Zhang, Lei Bai, Mengyue Yang, Chen Zhang, Zaibin Zhang, Yue Liao, Xiangyuan Xue, Songtao Huang, Yijiang Li, Michael Littman, Heng Zhou, Zhongzhi Li, Yutao Fan, Hejia Geng, and Zelin Tan. The landscape of agentic reinforcement learning for llms: A survey, 2025. 1, 4

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *ACL (1)*, pp. 9426–9439. Association for Computational Linguistics, 2024. 1

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022. 2.2

Liang Wen, Yunke Cai, Fenrui Xiao, Xin He, Qi An, Zhenyu Duan, Yimin Du, Junchen Liu, Lifu Tang, Xiaowei Lv, Haosheng Zou, Yongchao Deng, Shousheng Jia, and Xiangzheng Zhang. Light-r1: Curriculum sft, DPO and RL for long COT from scratch and beyond. *CoRR*, abs/2503.10460, 2025. 1, 2.1, 4

Yongliang Wu, Yizhou Zhou, Zhou Ziheng, Yingzhe Peng, Xinyu Ye, Xinting Hu, Wenbo Zhu, Lu Qi, Ming-Hsuan Yang, and Xu Yang. On the generalization of SFT: A reinforcement learning perspective with reward rectification. *CoRR*, abs/2508.05629, 2025. 1

xAI. Grok 4. https://x.ai/news/grok-4/, 2025. Accessed: 2025-07. 1, 4

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation. In *ICML*. OpenReview.net, 2024. 1

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *CoRR*, abs/2412.15115, 2024a. 3.1, B.2

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *CoRR*, abs/2409.12122, 2024b. 1, 2.1, 4

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, Jian Yang, Jiaxi Yang, Jingren 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. *CoRR*, abs/2505.09388, 2025. 1, 3.1, 4

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. LIMO: less is more for reasoning. *CoRR*, abs/2502.03387, 2025. 1, 4

Huaiyuan Ying, Shuo Zhang, Linyang Li, Zhejian Zhou, Yunfan Shao, Zhaoye Fei, Yichuan Ma, Jiawei Hong, Kuikun Liu, Ziyi Wang, Yudong Wang, Zijian Wu, Shuaibin Li, Fengzhe Zhou, Hongwei Liu, Songyang Zhang, Wenwei Zhang, Hang Yan, Xipeng Qiu, Jiayu Wang, Kai Chen, and Dahua Lin. Internlm-math: Open math large language models toward verifiable reasoning. *CoRR*, abs/2402.06332, 2024. 1

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. *CoRR*, abs/2503.14476, 2025. 3.1, A.4

Bowen Zhang and Harold Soh. Extract, define, canonicalize: An llm-based framework for knowledge graph construction. In *EMNLP*, pp. 9820–9836. Association for Computational Linguistics, 2024. 1

Yue Zhang, Yafu Li, Ganqu Cui, Yu Cheng, Zhi Wang, Xiaoye Qu, Jianhao Yan, and Zican Hu. Learning to reason under off-policy guidance. *CoRR*, abs/2504.14945, 2025a. 1, 2.3

Zhong Zhang, Yaxi Lu, Yikun Fu, Yupeng Huo, Shenzhi Yang, Yesai Wu, Han Si, Xin Cong, Haotian Chen, Yankai Lin, Jie Xie, Wei Zhou, Wang Xu, Yuanheng Zhang, Zhou Su, Zhongwu Zhai, Xiaoming Liu, Yudong Mei, Jianming Xu, Hongyan Tian, Chongyi Wang, Chi Chen, Yuan Yao, Zhiyuan Liu, and Maosong Sun. Agentcpm-gui: Building mobile-use agents with reinforcement fine-tuning. *CoRR*, abs/2506.01391, 2025b. 4

Appendix

Table of Contents

A More Implementation Details

A.1 Training Details

Training utilized the veRL (Sheng et al., 2025) and vLLM (Kwon et al., 2023) frameworks on the clusters equipped with NVIDIA H100 GPUs. Table 3 present the detailed training parameters for Agentic Distillation.

Table 3: Training Parameters.

Parameters	**Values**
Batch Size	256
Number of Rollout Per Question	8
Rollout Temperature	1.0
Rollout Top-p	1.0
Rollout Top-k	-1
Maximum Number of Generation Tokens	16384
Learning Rate	1e-6
KL Loss Coefficient	0.001
ϵ_{min}	0.2
ϵ_{max}	0.28
Gradient Clipping	1.0
Number of Training Steps	300

A.2 Full Training Prompt

Prompt A.1 illustrates the full training prompt.

Prompt A.1: Full Training rompt

Objective:

To answer a User's question by providing a clear, verifiable reasoning process, potentially interacting with an external environment.

Interaction Protocol:

For each question you receive, you MUST follow this two-step process:

Step 1: Reasoning Process

- **Decomposition:** Break down the user's question into a logical, step-by-step sequence of reasoning. Start from the most basic facts and build upon them.
- **External Inquiry (Optional but Encouraged):**
  - You may issue up to *max_turns* queries to an External Environment to validate hypotheses, clarify information, or advance your reasoning.
  - Each query must be a self-contained question enclosed in `<query>...</query>` tags.
  - **Wait for the `<result>...</result>` block** from the environment before continuing your reasoning.
  - **Critically analyze and integrate** the content from the `<result>...</result>` block into your reasoning chain.
  - Do not invent, assume, or hallucinate any `<result>` content. Your reasoning must be grounded in the provided results.

Step 2: Final Answer

- After your reasoning is complete, state your final answer clearly.
- The final answer, and only the final answer, MUST be enclosed in "\boxed{...}".

A.3 Inference Prompt

To focus on distilling knowledge and capabilities from the teacher LLM, we prohibit the trained student LLM from interacting with the teacher LLM during the inference phase. For mathematical and puzzle reasoning benchmarks, we employ the prompt specified in Prompt A.2. For science and code reasoning benchmarks, we use the default prompts provided with the original benchmarks.

Prompt A.2: Prompt for Mathematical Reasoning Benchmarks

```
{question}
```
Please reason step by step, and put your final answer within "\boxed{...}".

A.4 Training Data

The training data of Agentic Distillation is composed of three parts:

- **DAPO-Math-17K.** DAPO-Math-17K (Yu et al., 2025) is a dataset comprising 17,000 mathematical problems with integer answers, specifically designed for large-scale reinforcement learning of LLMs. The dataset was meticulously curated to ensure accurate reward signals by collecting questions and answers from the Art of Problem Solving (AoPS) website and competition homepages, followed by manual annotation and conversion to unify answers in integer form. We utilize the English subset, consisting of 14,000 questions, for training.
- **OpenScienceReasoning-2.** OpenScienceReasoning-2 is a multi-domain synthetic dataset aimed at enhancing general-purpose reasoning in LLMs. It includes multiple-choice and open-ended question-answer pairs with detailed reasoning traces, covering diverse scientific domains such as STEM, law, economics, and humanities. We randomly sample 20,000 examples from the original dataset for training.
- **Reasoning-Gym.** Reasoning-Gym (Stojanovski et al., 2025) is a community-developed Python library featuring procedural dataset generators and algorithmically verifiable reasoning environments for training reasoning models with RL. It encompasses over 100 tasks across domains including algebra, arithmetic, computation, cognition, geometry, graph theory, logic, and various games. We generate 27,000 samples for training, with each of 27 configurations producing 1,000 samples.

```
tasks = [
    ("ab", 1.0, {
        "seed": 42,
        "length": 10,
        "size": size
    }),
    ("ab", 1.0, {
        "seed": 42,
        "length": 15,
        "size": size
    }),
    ("acre", 1.0, {
        "seed": 42,
        "size": size
    }),
    ("advanced_geometry", 1.0, {
        "seed": 42,
        "min_coord": -100,
        "max_coord": 100,
        "size": size
    }),
    ("aiw", 1.0, {
        "seed": 42,
        "max_entities": 10,
        "size": size
    }),
    ("cryptarithm", 1.0, {
```

```
            "seed": 42,
            "min_words": 5,
            "max_words": 20,
            "size": size
        }),
        ("dice", 1.0, {
            "seed": 42,
            "num_dice": 5,
            "max_dice_size": 30,
            "size": size
        }),
        ("futoshiki", 1.0, {
            "seed": 42,
            "size": size
        }),
        ("game_of_life", 1.0, {
            "seed": 42,
            "grid_size_x": 30,
            "grid_size_y": 30,
            "simulation_steps": 3,
            "size": size
        }),
        ("game_of_life", 1.0, {
            "seed": 42,
            "grid_size_x": 30,
            "grid_size_y": 30,
            "simulation_steps": 4,
            "size": size
        }),
        ("game_of_life", 1.0, {
            "seed": 42,
            "grid_size_x": 30,
            "grid_size_y": 30,
            "simulation_steps": 5,
            "size": size
        }),
        ("game_of_life_halting", 1.0, {
            "seed": 42,
            "grid_size_x": 30,
            "grid_size_y": 30,
            "difficulty": 3,
            "num_oscillators": 8,
            "max_simulation_steps": 40,
            "size": size
        }),
        ("jugs", 1.0, {
            "seed": 42,
            "difficulty": 20,
            "size": size
        }),
        ("knight_swap", 1.0, {
            "seed": 42,
            "size": size
        }),
        ("knights_knaves", 1.0, {
            "seed": 42,
            "n_people": 3,
            "depth_constraint": 3,
            "width_constraint": 3,
            "size": size
        }),
        ("knights_knaves", 1.0, {
            "seed": 42,
            "n_people": 5,
            "depth_constraint": 5,
```

```
            "width_constraint": 5,
            "size": size
        }),
        ("mahjong_puzzle", 1.0, {
            "seed": 42,
            "min_num_rounds": 30,
            "size": size
        }),
        ("needle_haystack", 1.0, {
            "seed": 42,
            "min_num_statements": 50,
            "size": size
        }),
        ("quantum_lock", 1.0, {
            "seed": 42,
            "difficulty": 10,
            "size": size
        }),
        ("quantum_lock", 1.0, {
            "seed": 42,
            "difficulty": 20,
            "size": size
        }),
        ("rush_hour", 1.0, {
            "seed": 42,
            "min_moves": 10,
            "size": size
        }),
        ("self_reference", 1.0, {
            "seed": 42,
            "difficulty": 10,
            "size": size
        }),
        ("sudoku", 1.0, {
            "seed": 42,
            "size": size
        }),
        ("zebra_puzzles", 1.0, {
            "seed": 42,
            "num_people": 4,
            "num_characteristics": 4,
            "size": size
        }),
        ("zebra_puzzles", 1.0, {
            "seed": 42,
            "num_people": 5,
            "num_characteristics": 5,
            "size": size
        }),
        ("zebra_puzzles", 1.0, {
            "seed": 42,
            "num_people": 6,
            "num_characteristics": 6,
            "size": size
        }),
        ("zebra_puzzles", 1.0, {
            "seed": 42,
            "num_people": 7,
            "num_characteristics": 7,
            "size": size
        })
    ]
```

A.5 Evaluation Benchmarks

The following details describe our evaluation benchmarks:

- **AIME24.** AIME24 comprises 30 challenging questions from the 2024 American Invitational Mathematics Examination (AIME), designed to test advanced mathematical reasoning skills.
- **AIME25.** AIME25 includes 30 challenging questions from the 2025 American Invitational Mathematics Examination (AIME), focusing on complex mathematical problem-solving.
- **MATH500.** The original MATH dataset (Hendrycks et al., 2021) contains 12,500 problems from American high school mathematics competitions. For this study, we use MATH500 (Lightman et al., 2024), a subset of the test split consisting exclusively of Level 5 questions.
- **LiveMathBench.** LiveMathBench (Liu et al., 2024b) is a continuously updated dataset of challenging mathematical problems. We utilize the December 2024 hard split, which includes 45 questions in English and Chinese.
- **GPQA.** The Graduate-Level Google-Proof Q&A Benchmark (GPQA) (Rein et al., 2023) is a challenging dataset of professional-level, multiple-choice science questions. We evaluate on its diamond subset, comprising 198 questions.
- **MBPP.** The Mostly Basic Programming Problems (MBPP) dataset (Austin et al., 2021) evaluates programming models on basic Python tasks. Constructed via crowdsourcing, the problems and solutions undergo revision and manual inspection to ensure clarity and accurate test cases.
- **LiveCodeBench.** LiveCodeBench (Jain et al., 2025) is a benchmark for comprehensive and uncontaminated evaluation of LLM code-related capabilities, incorporating questions from LeetCode, AtCoder, and Codeforces.
- **Reasoning-Gym.** Using the configurations outlined in Appendix A.4, we generate 270 samples for evaluation, with each of 27 configurations producing 10 samples.

B Additional Experimental Results and Analysis

B.1 Performance of Teacher LLMs

Table 4 shows the performance of Qwen3-30B-A3B-Instruct-2507 and GPT-OSS-20B, which are utilized as teacher LLMs in this work. The performance of teacher LLMs can be seen as the upper bound of the distillation.

Table 4: Performance of Qwen3-30B-A3B-Instruct-2507 and GPT-OSS-20B.

Methods	Math				Science	Code		Puzzle
	AIME24 Avg@16	AIME25 Avg@16	MATH500 Avg@4	LMB Avg@4	GPQA-D Avg@4	MBPP Avg@4	LCB Avg@4	RG Avg@4
Qwen3-30B-A3B-Instruct-2507								
-	76.88	63.96	96.75	44.50	55.18	84.05	44.74	19.54
GPT-OSS-20B								
-	78.62	73.75	96.45	50.50	59.22	93.68	60.53	13.98

B.2 Agentic Distillation on Larger Student LLMs

In this section, we evaluate the effectiveness of Agentic Distillation on student LLMs with larger parameter sizes, specifically training Qwen2.5-32B-Instruct (Yang et al., 2024a) with Agentic Distillation. As shown in Table 5, the evaluation results demonstrate that Agentic Distillation remains effective for larger-scale models, with Agentic Distillation-trained models outperforming baseline models across all benchmarks. Notably, the performance improvements for Qwen2.5-32B-Instruct are more pronounced compared to those for Qwen2.5-7B-Instruct. This enhanced improvement may stem from the 32B model's stronger baseline capabilities, enabling it to formulate higher-quality questions and acquire knowledge more efficiently during training with Agentic Distillation.

Table 5: Experimental results of AGENTIC DISTILLATION and baselines with Qwen2.5-32B-Instrcut as the student LLM. We report the average performance for 16 runs on AIME24 and AIME25, and 4 runs on others. We abbreviate LMB as LiveMathBench v202505, LCB as LiveCodeBench v6, RG as Reasoning Gym, MI as Masked Interaction, and AD as AGENTIC DISTILLATION.

Methods	Math				Science	Code		Puzzle
	AIME24 Avg@16	AIME25 Avg@16	MATH500 Avg@4	LMB Avg@4	GPQA-D Avg@4	MBPP Avg@4	LCB Avg@4	RG Avg@4
Student LLM: *Qwen2.5-32B-Instruct*, Teacher LLM: *Qwen3-30B-A3B-Instruct-2507*								
Original	14.38	13.12	80.85	12.25	47.10	85.12	24.71	13.06
+SFT	16.83	15.79	83.47	15.33	46.13	84.74	23.15	15.61
+RL	15.73	14.11	82.51	13.33	50.89	86.71	26.43	17.62
+AD	**17.56**	**18.19**	**86.82**	**17.37**	**51.05**	**88.74**	**26.91**	**22.48**

B.3 AGENTIC DISTILLATION ON LONG-COT STUDENT LLMS

In this section, we assess the performance of AGENTIC DISTILLATION on reasoning LLMs utilizing long CoT prompting. Given the substantial inference overhead of long CoT LLMs, we conducted experiments using DeepSeek-R1-Distill-Qwen-1.5B, with results presented in Table 6. The findings demonstrate that AGENTIC DISTILLATION achieves consistent performance improvements for student LLMs with extended reasoning chains, underscoring the generalization capability of AGENTIC DISTILLATION across such models.

Additionally, we observe a performance decline in models trained with SFT. This may be attributed to the teacher LLM, Qwen3-30B-A3B-Instruct-2507, not being optimized for long CoT reasoning. Consequently, fine-tuning based on its responses may disrupt the original reasoning patterns of the student LLM, leading to degraded performance. In contrast, AGENTIC DISTILLATION selectively injects knowledge into the student LLM via query-answer pairs, preserving its inherent reasoning patterns. This preservation represents a key advantage of AGENTIC DISTILLATION, enhancing its effectiveness without compromising the student LLM's original reasoning capabilities.

Table 6: Experimental results of AGENTIC DISTILLATION and baselines with DeepSeek-R1-Distill-Qwen-1.5B as the student LLM. We report the average performance for 16 runs on AIME24 and AIME25, and 4 runs on others. We abbreviate LMB as LiveMathBench v202505, LCB as LiveCodeBench v6, RG as Reasoning Gym, MI as Masked Interaction, and AD as AGENTIC DISTILLATION.

Methods	Math				Science	Code		Puzzle
	AIME24 Avg@16	AIME25 Avg@16	MATH500 Avg@4	LMB Avg@4	GPQA-D Avg@4	MBPP Avg@4	LCB Avg@4	RG Avg@4
Student LLM: *DeepSeek-R1-Distill-Qwen-1.5B*, Teacher LLM: *Qwen3-30B-A3B-Instruct-2507*								
Original	21.88	21.46	83.95	13.00	29.80	60.12	14.69	3.33
+SFT	18.35	19.89	77.16	14.02	26.64	55.51	15.27	10.98
+RL	28.43	25.70	86.82	17.39	34.68	65.05	14.72	13.53
+AD	**30.56**	**29.21**	**88.47**	**18.90**	**36.53**	**67.19**	**17.06**	**16.44**

C LIMITATIONS

In this paper, we propose AGENTIC DISTILLATION, a novel framework for active distillation to enhance the reasoning capabilities of LLMs. While AGENTIC DISTILLATION achieves significant performance improvements, several areas warrant further exploration. First, our framework does not explicitly guarantee the accuracy of feedback provided to the student LLM. In cases where the teacher produces suboptimal or noisy guidance, the student may inadvertently learn misleading patterns, which could diminish the effectiveness of the distillation process and lead to unstable improvements across tasks. Second, as the scale and inference complexity of teacher LLMs increase, the training time required by AGENTIC DISTILLATION may grow considerably. This not only elevates

computational costs but may also impose practical challenges when deploying the framework in resource-constrained environments or when scaling to very large datasets and extended training regimes.

D LLM USAGE

In this paper, the use of LLMs is intentionally restricted to the final stages of the research process, specifically for refining and proofreading the written content. The LLMs are employed solely to enhance the clarity, coherence, and grammatical accuracy of the text, ensuring effective and professional communication of the presented ideas. Importantly, LLMs played no role in the core components of this work, including the development of the research methodology, the design of experiments, or the analysis of results. We are aware that we will be responsible for all content in the paper.