

000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 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. 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 distillation performance, 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). In particular, LLMs employing long chain-of-thought (CoT) reasoning have demonstrated remarkable proficiency in solving intricate problems spanning mathematics, coding, and science, significantly pushing the boundaries of complex capabilities in foundation models (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 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

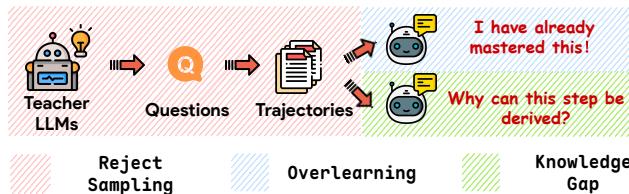


Figure 1: Illustration of *over-learning* and *knowledge gap* issues in data-centric distillation.

054 reasoning capabilities of teacher LLMs. However, this approach has notable limitations due to its
 055 characteristic of **passive learning** as shown in Figure 1: 1) **Over-Learning**: Training on static and
 056 complete trajectories from teacher LLMs does not dynamically adapt to the evolving capabilities
 057 of the student model, often leading to over-learning on simpler questions (Chu et al., 2025) and
 058 wasting data and training resources on mastered questions.; 2) **Knowledge Gap**: Teacher-centered
 059 data generation approaches may neglect the specific knowledge requirements and competency gaps
 060 of student LLMs (Liu et al., 2024a). Much like a standardized teaching approach that overlooks
 061 individual student deficiencies, reasoning processes that appear intuitive to advanced teacher models
 062 can be inaccessible to student models. As a result, student models tend to merely mimic the stylistic
 063 features of the teacher’s output rather than acquiring robust reasoning capabilities (Chu et al., 2025;
 064 Kirk et al., 2024; Wu et al., 2025).

065 To address these limitations, we propose AGENTIC DISTILLATION, a novel framework for distilling
 066 knowledge and reasoning capabilities from strong LLMs into smaller student LLMs through **active**
 067 and **adaptive interaction**. Unlike traditional distillation methods that passively transfer knowledge,
 068 AGENTIC DISTILLATION empowers the student LLM to dynamically determine when to query
 069 the teacher LLM during reasoning, seeking feedback only when necessary as shown in Figure 2.
 070 This enables the student LLM to refine its reasoning process based on teacher feedback, leading
 071 to more accurate outcomes. Additionally, we design a mechanism to allow the student LLM to
 072 effectively learn essential knowledge and reasoning abilities from the teacher LLM’s feedback.
 073



074 Figure 2: Simple Illustration of AGENTIC DISTILLATION.
 075

076 and injects knowledge into student LLMs. Notably, recent works (e.g., Search-R1) (Wang et al.,
 077 2025; Singh et al., 2025; Jin et al., 2025; Liu et al., 2025) primarily focus on enhancing LLMs’
 078 interactions with external environments (e.g., tools), often overlooking the rich information embedded
 079 in feedback, which can be utilized to improve the reasoning ability of student LLMs themselves. In
 080 contrast, AGENTIC DISTILLATION leverages teacher feedback as a direct learning signal, enabling
 081 continuous improvement of the student model. Even without interaction during inference, AGENTIC
 082 DISTILLATION-trained student LLMs can successfully reason on previously unsolvable tasks.
 083

084 We conduct extensive experiments to validate the effectiveness of AGENTIC DISTILLATION. For
 085 example, AGENTIC DISTILLATION enhances the performance of Qwen2.5-7B-Instruct on mathe-
 086 matical reasoning benchmarks, achieving an average improvement of approximately 4 points over
 087 baseline distillation strategies. Significant gains are also observed on out-of-domain benchmarks,
 088 demonstrating AGENTIC DISTILLATION’s robust generalization. Additional experiments confirm
 089 that AGENTIC DISTILLATION generalizes effectively across various student and teacher LLMs.
 090 Additionally, we investigate whether AGENTIC DISTILLATION expands the knowledge boundaries
 091 of student LLMs. Analysis of training dynamics and student LLM responses reveals that AGENTIC
 092 DISTILLATION enables student LLMs to effectively acquire new knowledge and capabilities, aligning
 093 their reasoning abilities with those of teacher LLMs.
 094

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

096 2 METHOD

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

100 2.1 PRELIMINARIES

101 **Distillation from Strong LLMs.** In a typical LLM reasoning task, given a question q from the
 102 question distribution $q \sim P(Q)$, the LLM π_θ is prompted with an instruction I to generate an answer:
 103

$$104 \quad a' \leftarrow \pi_\theta(\cdot \mid q, I). \quad (1)$$

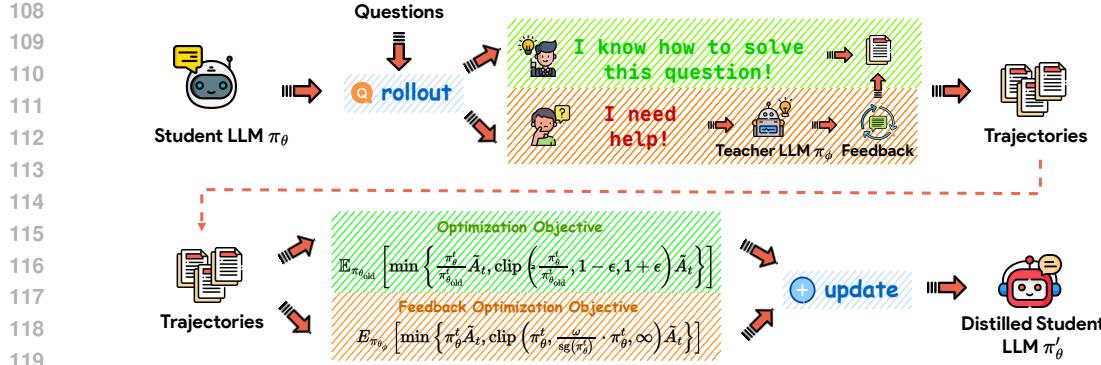


Figure 3: Illustration of AGENTIC DISTILLATION. In the AGENTIC DISTILLATION framework, during each rollout process, the student LLM initially attempts to solve a given question independently. If the student LLM fails to resolve the question, it engages in *external interaction* by querying the teacher 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.

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 τ_{thinking} . This culminates in a final conclusion $\tau_{\text{conclusion}}$, which includes the predicted answer a' :

$$[\tau_{\text{thinking}}, \tau_{\text{conclusion}}] \leftarrow \pi_\theta(\cdot | q, I). \quad (2)$$

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 :

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

where \mathbb{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 $q \in \mathcal{Q}$:

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

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

$$\mathcal{T}' = \left\{ \tau \mid \mathbb{I}(a, a') \right\}, \quad (5)$$

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 :

$$\mathcal{L}(\theta) = \mathbb{E}_{q \in \mathcal{Q}, \tau \sim \mathcal{T}'} \left[-\log \pi_\theta^s(\tau | q; \theta) \right], \quad (6)$$

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 τ_{thinking} to include multiple turns of interaction, comprising *queries to the teacher LLM τ_q* and *external feedback from the teacher LLM τ_o* , formally expressed as:

$$[\dots, \tau_{q,(1)}, \tau_{o,(1)}, \dots, \tau_{q,(N)}, \tau_{o,(N)}, \dots] \leftarrow \tau_{\text{thinking}}. \quad (7)$$

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.

162 2.2 AGENTIC INTERACTION WITH TEACHER LLMs.
163164 To distill knowledge from the teacher LLM, we design an agentic interaction mechanism that enables
165 the student LLM to actively and flexibly interact with the teacher during the reasoning process.166 When faced with a question q , the student LLM first performs basic reasoning using its internal
167 knowledge, such as problem decomposition, solution planning, and simple arithmetic operations (Wei
168 et al., 2022). If the student LLM can solve the question using only its own knowledge and reasoning
169 abilities, we argue that external knowledge distillation from an *oracle* is unnecessary. This important
170 distinction is often overlooked by typical SFT-based methods (Qin et al., 2024; Huang et al., 2024;
171 Muenninghoff et al., 2025; Guha et al., 2025). Conversely, during reasoning, when the student
172 recognizes that a (sub-)question exceeds the limits of its internal knowledge, it must refer to external
173 oracle information. In such cases, we allow the student LLM to query the teacher LLM in natural
174 language.175 Specifically, we provide the student LLM with the prompt shown in Prompt 2.1 (full version is
176 provided in Prompt A.1), which instructs it to enclose natural language queries to the teacher within
177 `<query>` and `</query>` tags. The teacher LLM then responds with the corresponding answer (i.e.,
178 *feedback* or *observation*), appended to the student’s reasoning process within `<result>` and
179 `</result>` tags. To avoid meaningless or inefficient loops, we also impose an interaction budget
180 limiting the number of queries the student may direct to the teacher.181 **Prompt 2.1: Prompt to Equip Student LLM with Agentic Interaction Capability**
182183 **Reasoning Process**

- 184 •
- Decomposition:**
- Break down the user’s question into a logical, step-by-step sequence of reasoning. Start from the
-
- 185 most basic facts and build upon them.
-
- 186 •
- External Inquiry (Optional but Encouraged):**
-
- 187 – You may issue up to
- `max_turns`
- queries to an External Environment to validate hypotheses, clarify information,
-
- 188 or advance your reasoning.
-
- 189 – Each query must be a self-contained question enclosed in
- `<query>...</query>`
- tags.
-
- 190 –
- Wait for the `<result>...</result>` block**
- from the environment before continuing your reasoning.
-
- 191 –
- Critically analyze and integrate**
- the content from the
- `<result>...</result>`
- block into your reasoning
-
- 192 chain.
-
- 193 – Do not invent, assume, or hallucinate any
- `<result>`
- content. Your reasoning must be grounded in the
-
- 194 provided results.

195 2.3 LEARNING FROM AGENTIC INTERACTION
196197 This section addresses learning from agentic interaction. Given a query τ_v generated by the student
198 LLM and feedback τ_o provided by the teacher LLM, prior RL approaches typically exclude τ_o from
199 the loss calculation (Song et al., 2025; Liu et al., 2025), as the student LLM is not expected to generate
200 tokens from the external environment. In contrast, our approach integrates feedback tokens $\tau_{o,(1:T_o)}$
201 into the RL policy loss to enable the student LLM to acquire new knowledge and capabilities.202 The classical clipped surrogate objective is defined as:
203

204
$$\mathcal{J}(\theta) = \mathbb{E}_{q \sim P, \{\tau_i\} \sim \pi_\theta} \left[\frac{1}{G} \sum_{i=1}^G \frac{1}{|\tau_i|} \sum_{t=1}^{|\tau_i|} \left\{ \min \left(\rho_{i,t} \tilde{A}_t^i, \text{clip}(\rho_{i,t}, 1-\epsilon, 1+\epsilon) \tilde{A}_t^i \right) \right. \right. \\ 205 \left. \left. - \beta D_{\text{KL}}[\pi_\theta \parallel \pi_{\text{ref}}] \right\} \right], \quad (8)$$

206
207
208
209

210 where the importance sampling coefficient for each token $\tau_{i,(t)}$ at index t is given by:
211

212
$$\rho_{i,t} = \frac{\pi_\theta(\tau_{i,(t)} \mid \tau_{i,(\leq t)})}{\pi_{\theta_{\text{old}}}(\tau_{i,(t)} \mid \tau_{i,(\leq t)})}, \quad (9)$$

213
214

215 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

216 the student LLM’s policy, which can destabilize RL training (Schulman et al., 2017; Zhang et al.,
 217 2025a).

219 **Amending Importance Sampling Coefficient for Feedback Tokens.** To mitigate the off-policy
 220 error, we introduce a modified importance sampling coefficient $\tilde{\rho}$. Within the standard clipped
 221 surrogate loss, the off-policy error stems from sampling the trajectory τ from $\pi_{\theta_{\text{old}}}$:

$$222 \quad \mathcal{J}(\theta) = \mathbb{E}_{\tau \sim \pi_{\theta_{\text{old}}}} \left[\frac{1}{|\tau|} \sum_{t=1}^{|\tau|} \frac{\pi_{\theta}(\tau_{(t)} | \tau_{(\leq t)})}{\pi_{\theta_{\text{old}}}(\tau_{(t)} | \tau_{(\leq t)})} \tilde{A}_t \right], \quad (10)$$

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

$$229 \quad \mathcal{J}(\theta) = \mathbb{E}_{\pi_{\theta_{\text{old}}}} \left[\frac{1}{N} \sum_{\tau_{(t)} \notin \tau_o} \frac{\pi_{\theta}(\tau_{(t)} | \tau_{(\leq t)})}{\pi_{\theta_{\text{old}}}(\tau_{(t)} | \tau_{(\leq t)})} \tilde{A}_t \right] + \mathbb{E}_{\pi_{\phi}} \left[\frac{1}{M} \sum_{\tau_{(t)} \in \tau_o} \frac{\pi_{\theta}(\tau_{(t)} | \tau_{(\leq t)})}{\pi_{\phi}(\tau_{(t)} | \tau_{(\leq t)})} \tilde{A}_t \right], \quad (11)$$

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

$$238 \quad \mathcal{J}(\theta) = \mathbb{E}_{\pi_{\theta_{\text{old}}}} \left[\frac{1}{N} \sum_{\tau_{(t)} \notin \tau_o} \frac{\pi_{\theta}(\tau_{(t)} | \tau_{(\leq t)})}{\pi_{\theta_{\text{old}}}(\tau_{(t)} | \tau_{(\leq t)})} \tilde{A}_t \right] + \mathbb{E}_{\pi_{\phi}} \left[\frac{1}{M} \sum_{\tau_{(t)} \in \tau_o} \pi_{\theta}(\tau_{(t)} | \tau_{(\leq t)}) \tilde{A}_t \right]. \quad (12)$$

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

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

$$250 \quad \pi_{\theta} \cdot \tilde{A}_t \cdot \nabla_{\theta} \cdot \log \pi_{\theta}. \quad (13)$$

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

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

$$260 \quad \text{clip} \left(\pi_{\theta}, \frac{\omega}{\text{sg}(\pi_{\theta})} \cdot \pi_{\theta}, \infty \right), \quad (14)$$

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

$$266 \quad \begin{cases} \pi_{\theta} \cdot \tilde{A}_t \cdot \nabla_{\theta} \cdot \log \pi_{\theta}, & \text{if } \pi_{\theta} \geq \omega, \\ \omega \cdot \tilde{A}_t \cdot \nabla_{\theta} \cdot \log \pi_{\theta}, & \text{if } 0 \leq \pi_{\theta} < \omega. \end{cases} \quad (15)$$

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

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

$$\begin{aligned} 274 \quad \mathcal{J}(\theta) = & \mathbb{E}_{\pi_{\theta_{\text{old}}}} \left[\min \left\{ \frac{\pi_{\theta}^t}{\pi_{\theta_{\text{old}}}^t} \tilde{A}_t, \text{clip} \left(\frac{\pi_{\theta}^t}{\pi_{\theta_{\text{old}}}^t}, 1 - \epsilon, 1 + \epsilon \right) \tilde{A}_t \right\} \right] \\ 275 & + \mathbb{E}_{\pi_{\phi}} \left[\min \left\{ \pi_{\theta}^t \tilde{A}_t, \text{clip} \left(\pi_{\theta}^t, \frac{\omega}{\text{sg}(\pi_{\theta}^t)} \cdot \pi_{\theta}^t, \infty \right) \tilde{A}_t \right\} \right], \end{aligned} \quad (16)$$

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

286 3 EXPERIMENTS

288 3.1 SETUP

290 **Baselines.** We compare AGENTIC DISTILLATION against several representative LLM post-training
 291 methods: ① *Supervised Fine-Tuning (SFT)*: Utilizes teacher LLM-generated data through rejection
 292 sampling; ② *Vanilla Reinforcement Learning (RL)*: Trains the student LLM using the GRPO algo-
 293 rithm (Shao et al., 2024) without external environment interactions; ③ *Reinforcement Learning with
 294 Supervised Fine-Tuning (RL+SFT)*: Combines GRPO training (Shao et al., 2024) with data generated
 295 via rejection sampling; ④ *Reinforcement Learning with Masked Interaction (RL+MI)*: Employs the
 296 GRPO algorithm (Shao et al., 2024) with teacher LLM interactions, but excludes feedback tokens
 297 from loss computation.

298 **Evaluation Benchmarks.** We evaluated all models across four domain-specific benchmarks: ①
 299 *Mathematical Reasoning*: Includes AIME24, AIME25, MATH500 (Hendrycks et al., 2021), and
 300 LiveMathBench (Liu et al., 2024b); ② *Scientific Reasoning*: Represented by GPQA-Diamond (Rein
 301 et al., 2023); ③ *Code Reasoning*: Comprises MBPP (Austin et al., 2021) and LiveCodeBench (Jain
 302 et al., 2025); ④ *Puzzle Reasoning*: Includes puzzles from Reasoning-Gym (Stojanovski et al., 2025).

303 **Implementation Details.** We conducted experiments on the Qwen-2.5 series models (Yang et al.,
 304 2024a) and Llama-3.2 series models (Dubey et al., 2024), distilling from two prominent teacher LLMs
 305 from distinct families: Qwen3-30B-A3B-Instruct-2507 (Yang et al., 2025) and GPT-OSS-20B (Agar-
 306 wal et al., 2025). The training corpus, sourced from DAPO (Yu et al., 2025), OpenScienceReasoning-
 307 2¹, and Reasoning Gym (Stojanovski et al., 2025), consists of approximately 60,000 high-quality
 308 reasoning-intensive samples. Models were trained for 200 steps with a batch size of 256 and a group
 309 size of 8, selecting the best model based on validation performance. During each generation, the
 310 student LLM was allowed up to three interactions with the teacher LLM. Training was performed
 311 using the veRL (Sheng et al., 2025) and vLLM (Kwon et al., 2023) frameworks. For evaluation, we
 312 set the sampling temperature to 1.0, top- p to 1.0, top- k to -1, and the maximum generation tokens to
 313 16384. To reduce variance, we report average performance relative to the size of each benchmark.
 314 And the prompt utilized in inference phase is shown in Prompt A.2.

316 3.2 MAIN RESULTS AND ANALYSIS

317 Table 1 illustrates the performance of AGENTIC DISTILLATION and baselines on different bench-
 318 marks, containing different student LLMs and teacher LLMs. From the experimental results, we have
 319 the following findings.

321 **AGENTIC DISTILLATION Outperforms Baseline Methods.** As illustrated in Table 1, AGENTIC
 322 DISTILLATION surpasses other training strategies, including supervised fine-tuning (SFT), vanilla

323 ¹<https://huggingface.co/datasets/nvidia/OpenScienceReasoning-2>

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

331 Methods	332 Math ♠				333 Science ♠		334 Code ♣		Puzzle ♠
	AIME24 335 Avg@16	AIME25 Avg@16	MATH500 Avg@4	LMB Avg@4	GPQA-D Avg@4	MBPP Avg@4	LCB Avg@4	RG Avg@4	
336 Student LLM: <i>Qwen2.5-7B-Instruct</i> , Teacher LLM: <i>Qwen3-30B-A3B-Instruct-2507</i>									
Original	9.79	7.50	73.00	10.75	33.33	58.66	15.71	9.63	
+SFT	11.67	<u>13.54</u>	75.80	<u>12.00</u>	22.35	41.34	11.04	18.98	
+RL	12.08	10.00	75.35	10.50	34.72	58.27	<u>16.33</u>	<u>19.81</u>	
+RL+SFT	<u>13.01</u>	12.13	<u>75.77</u>	11.32	<u>35.20</u>	<u>59.22</u>	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	
341 Student LLM: <i>Qwen2.5-7B-Instruct</i> , Teacher LLM: <i>GPT-OSS-20B</i>									
+SFT	<u>14.67</u>	<u>13.54</u>	<u>75.80</u>	<u>12.00</u>	<u>37.34</u>	<u>61.04</u>	22.35	18.98	
+RL	12.08	10.00	75.35	10.50	34.72	58.27	16.33	<u>19.81</u>	
+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	<u>20.27</u>	24.32	
347 Student LLM: <i>Llama-3.2-3B-Instruct</i> , Teacher LLM: <i>Qwen3-30B-A3B-Instruct-2507</i>									
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	<u>13.89</u>	
+RL	<u>8.96</u>	1.12	44.10	7.00	25.76	52.59	<u>10.92</u>	13.70	
+RL+SFT	6.24	<u>3.62</u>	41.25	<u>8.11</u>	<u>26.45</u>	53.45	10.23	12.44	
+RL+MI	7.44	2.98	<u>45.11</u>	6.45	24.26	<u>54.27</u>	9.84	10.52	
+AD	10.38	4.42	44.45	9.00	28.54	58.66	15.70	16.85	

355 reinforcement learning (RL), RL combined with SFT (RL+SFT), and RL with masked interaction
 356 (RL+MI). Notably, AGENTIC DISTILLATION achieves significant improvements on challenging
 357 reasoning benchmarks such as AIME24 and AIME25, with average accuracy gains of 4-6 points
 358 over the strongest baseline. Comparable enhancements are observed across science, code, and puzzle
 359 tasks, underscoring AGENTIC DISTILLATION’s robustness in improving reasoning capabilities across
 360 diverse task settings.

361 **AGENTIC DISTILLATION Enhances Performance Across Diverse Student LLMs.** As depicted
 362 in Table 1, AGENTIC DISTILLATION consistently outperforms baseline methods across various
 363 student LLMs, including Qwen-2.5-7B-Instruct and Llama-3.2-3B-Instruct. The framework achieves
 364 stable improvements across LLMs of different architectures and sizes, highlighting the generality of
 365 AGENTIC DISTILLATION and its potential for broad application to diverse LLM families and types.

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

373 **AGENTIC DISTILLATION Generalizes to Out-of-Domain Benchmarks.** AGENTIC DISTILLATION
 374 also performs generalization across in-domain and out-of-domain benchmarks. On mathematics,
 375 science, and puzzle benchmarks, which align closely with the training data, AGENTIC DISTILLATION
 376 consistently outperforms all baselines. More notably, on out-of-domain benchmarks such as code
 377 (MBPP and LiveCodeBench), AGENTIC DISTILLATION achieves substantial gains, surpassing the

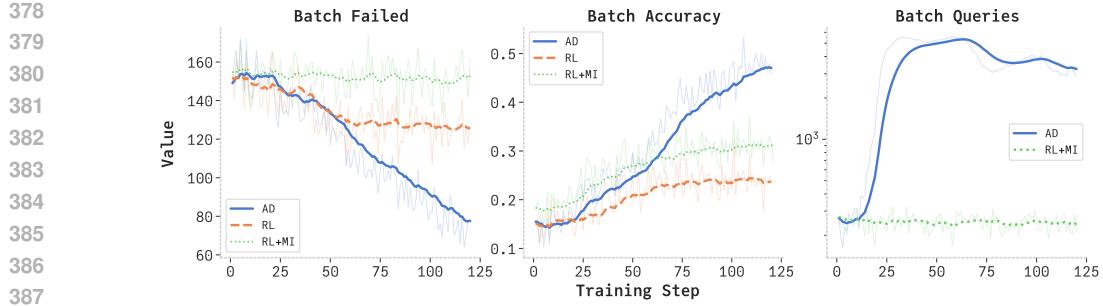


Figure 4: Training Dynamics of Reinforcement Learning, Reinforcement Learning with Masked Interaction, and AGENTIC DISTILLATION on Qwen2.5-7B-Instruct with Qwen3-30B-A3B-Instruct-2507 as the teacher LLM.

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: <i>Qwen2.5-7B-Instruct</i> , Teacher LLM: <i>Qwen3-30B-A3B-Instruct-2507</i>									
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	

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 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,

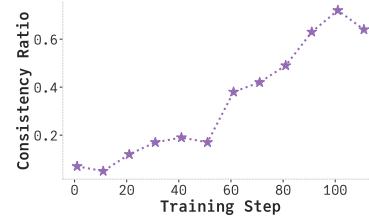


Figure 5: Knowledge Boundary Expansion of AGENTIC DISTILLATION-trained Student LLM.

432 using Reinforcement Learning (RL), Reinforcement Learning with Masked Interaction (RL+MI),
 433 and AGENTIC DISTILLATION (AD). First, we examine the proportion of problems in a batch
 434 that the student LLM fails to solve across all rollouts (*Batch Failed*). Training with AGENTIC
 435 DISTILLATION significantly reduces this proportion, indicating that AGENTIC DISTILLATION enables
 436 the student LLM to acquire new knowledge, allowing it to solve previously unsolvable problems.
 437 This improvement is mirrored in the training batch accuracy, where AGENTIC DISTILLATION-trained
 438 LLMs show markedly higher gains. Additionally, we analyze the number of queries raised by the
 439 student LLM per batch (*Batch Queries*). With AGENTIC DISTILLATION, the number of queries
 440 initially increases, then decreases, and eventually stabilizes. This trend suggests that early in training,
 441 the student LLM queries the teacher LLM frequently to learn new knowledge. As its knowledge
 442 boundary expands, the student LLM relies less on queries, solving problems independently. These
 443 metric dynamics demonstrate that AGENTIC DISTILLATION effectively facilitates the student LLM’s
 444 acquisition of new knowledge and capabilities from the teacher LLM.

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

4 RELATED WORK

456 **Distilling from Strong LLMs.** Recent advancements in LLMs have led to remarkable performance
 457 in complex reasoning tasks (OpenAI, 2024a;b; 2025; DeepSeek-AI et al., 2025; Kimi-Team et al.,
 458 2025; Yang et al., 2025; Comanici et al., 2025; Huang & Yang, 2025; xAI, 2025). However, these
 459 models are often closed-source or possess an excessively large number of parameters, limiting their
 460 practical applications. Consequently, recent research has focused on distilling the capabilities of
 461 these strong reasoning LLMs into smaller-scale LLMs. Early studies (Qin et al., 2024; Guan et al.,
 462 2025; DeepSeek-AI et al., 2025) demonstrated that a small dataset generated by strong LLMs can
 463 significantly enhance the reasoning performance of smaller LLMs. Subsequent works (Bespoke-
 464 Labs, 2025; NovaSky-Team, 2025; Ye et al., 2025; Wen et al., 2025; Guha et al., 2025; Yang et al.,
 465 2025) have further improved distillation by optimizing problem set quality, curated data, training
 466 methods, loss functions, and integration of training stages. These approaches typically rely on
 467 distilling complete reasoning trajectories, a passive learning method that often fails to address the
 468 specific capabilities and knowledge gaps of student LLMs. In contrast, our proposed method enables
 469 student LLMs to actively query strong LLMs and selectively learn knowledge beyond their current
 470 capabilities, offering a more efficient and effective distillation.

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

486 5 CONCLUSION
487488 In this paper, we introduce AGENTIC DISTILLATION, a distillation framework that enables active and
489 adaptive knowledge transfer from strong LLMs to smaller student models. AGENTIC DISTILLATION
490 leverages interaction and feedback tokens from teacher LLMs, allowing student models to selectively
491 refine their reasoning and bridge knowledge gaps. To tackle off-policy and gradient vanishing issues
492 inherent in learning from feedback, we introduce a tailored importance sampling coefficient and
493 clipping strategy that seamlessly integrate into the reinforcement learning objective. Extensive
494 experiments demonstrate that AGENTIC DISTILLATION achieves consistent improvements in both
495 in-domain and out-of-domain reasoning tasks. We believe our framework could provide a promising
496 direction for equipping compact models with advanced reasoning abilities.
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539

540 ETHICS STATEMENT
541

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

548 REPRODUCIBILITY STATEMENT
549

550 We provide a comprehensive description of the proposed AGENTIC DISTILLATION in § 2, with
551 detailed implementation specifics provided in § A and § 3.1. All datasets utilized in this research
552 are publicly available. Key code implementations are included in the supplementary materials
553 for reference, and the complete code will be made publicly available upon acceptance of the pa-
554 per.

556 REFERENCES
557

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

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

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

585 Bespoke-Labs. Bespoke-stratos: The unreasonable effectiveness of reasoning distil-
586 lation. <https://www.bespokelabs.ai/blog/bespoke-stratos-the-unreasonable-effectiveness-of->
587 reasoning-distillation, 2025. Accessed: 2025-01. 4

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

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

- 594 Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit S.
 595 Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marrs, Sam Petulla, Colin
 596 Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szektor, Nan-
 597 Jiang Jiang, Krishna Haridasan, Ahmed Omran, Nikunj Saunshi, Dara Bahri, Gaurav Mishra, Eric
 598 Chu, Toby Boyd, Brad Hekman, Aaron Parisi, Chaoyi Zhang, Kornraphop Kawintiranon, Tania
 599 Bedrax-Weiss, Oliver Wang, Ya Xu, Ollie Purkiss, Uri Mendlovic, Ilai Deutel, Nam Nguyen,
 600 Adam Langley, Flip Korn, Lucia Rossazza, Alexandre Ramé, Sagar Waghmare, Helen Miller,
 601 Nathan Byrd, Ashrith Sheshan, Raia Hadsell Sangnie Bhardwaj, Pawel Janus, Tero Rissa, Dan
 602 Horgan, Sharon Silver, Ayzaan Wahid, Sergey Brin, Yves Raimond, Klemen Kloboves, Cindy
 603 Wang, Nitesh Bharadwaj Gundavarapu, Ilia Shumailov, Bo Wang, Mantas Pajarskas, Joe Heyward,
 604 Martin Nikoltchev, Maciej Kula, Hao Zhou, Zachary Garrett, Sushant Kafle, Sercan Arik, Ankita
 605 Goel, Mingyao Yang, Jiho Park, Koji Kojima, Parsa Mahmoudieh, Koray Kavukcuoglu, Grace
 606 Chen, Doug Fritz, Anton Bulyenov, Sudeshna Roy, Dimitris Paparas, Hadar Shemtov, Bo-Juen
 607 Chen, Robin Strudel, David Reitter, Aurko Roy, Andrey Vlasov, Changwan Ryu, Chas Leichner,
 608 Haichuan Yang, Zelda Mariet, Denis Vnukov, Tim Sohn, Amy Stuart, Wei Liang, Minmin Chen,
 609 Praynaa Rawlani, Christy Koh, JD Co-Reyes, Guangda Lai, Praseem Banzal, Dimitrios Vytiniotis,
 610 Jieru Mei, and Mu Cai. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality,
 611 long context, and next generation agentic capabilities. *CoRR*, abs/2507.06261, 2025. 1, 4
- 612 DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu,
 613 Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu,
 614 Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao
 615 Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan,
 616 Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao,
 617 Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding,
 618 Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang
 619 Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong,
 620 Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao,
 621 Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang,
 622 Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang,
 623 Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L.
 624 Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang,
 625 Shuiping Yu, Shunfeng Zhou, Shuting Pan, and S. S. Li. Deepseek-r1: Incentivizing reasoning
 626 capability in llms via reinforcement learning. *CoRR*, abs/2501.12948, 2025. 1, 1, 2.1, 4
- 627 Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha
 628 Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn,
 629 Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston
 630 Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron,
 631 Bin Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris
 632 McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton
 633 Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David
 634 Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes,
 635 Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip
 636 Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme
 637 Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu,
 638 Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan
 639 Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet
 640 Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Janya Lee, Jeremy Fu, Jianfeng Chi,
 641 Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph
 642 Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasudevan Alwala, Kartikeya Upasani,
 643 Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. The llama 3 herd of models. *CoRR*,
 644 abs/2407.21783, 2024. 3.1
- 645 Xinyu Guan, Li Lyra Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao
 646 Yang. rstar-math: Small llms can master math reasoning with self-evolved deep thinking. *CoRR*,
 647 abs/2501.04519, 2025. 4
- 648 Etash Kumar Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal,
 649 Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, Ashima Suvarna, Benjamin Feuer,

- 648 Liangyu Chen, Zaid Khan, Eric Frankel, Sachin Grover, Caroline Choi, Niklas Muennighoff,
 649 Shiye Su, Wanjia Zhao, John Yang, Shreyas Pimpalgaonkar, Kartik Sharma, Charlie Cheng-Jie
 650 Ji, Yichuan Deng, Sarah M. Pratt, Vivek Ramanujan, Jon Saad-Falcon, Jeffrey Li, Achal Dave,
 651 Alon Albalak, Kushal Arora, Blake Wulfe, Chinmay Hegde, Greg Durrett, Sewoong Oh, Mohit
 652 Bansal, Saadia Gabriel, Aditya Grover, Kai-Wei Chang, Vaishaal Shankar, Aaron Gokaslan,
 653 Mike A. Merrill, Tatsunori Hashimoto, Yejin Choi, Jenia Jitsev, Reinhard Heckel, Maheswaran
 654 Sathiamoorthy, Alexandros G. Dimakis, and Ludwig Schmidt. Openthoughts: Data recipes for
 655 reasoning models. *CoRR*, abs/2506.04178, 2025. 1, 2.1, 2.2, 4
- 656 Shailja Gupta, Rajesh Ranjan, and Surya Narayan Singh. Comprehensive study on sentiment analysis:
 657 From rule-based to modern LLM based system. *CoRR*, abs/2409.09989, 2024. 1
- 658 Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song,
 659 and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In
 660 *NeurIPS Datasets and Benchmarks*, 2021. 3.1, A.5
- 661 Yichen Huang and Lin F. Yang. Gemini 2.5 pro capable of winning gold at IMO 2025. *CoRR*,
 662 abs/2507.15855, 2025. 1, 4
- 663 Zhen Huang, Haoyang Zou, Xuefeng Li, Yixiu Liu, Yuxiang Zheng, Ethan Chern, Shijie Xia, Yiwei
 664 Qin, Weizhe Yuan, and Pengfei Liu. O1 replication journey - part 2: Surpassing o1-preview
 665 through simple distillation, big progress or bitter lesson? *CoRR*, abs/2411.16489, 2024. 2.2
- 666 Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando
 667 Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free
 668 evaluation of large language models for code. In *ICLR*. OpenReview.net, 2025. 3.1, A.5
- 669 Bowen Jin, Hansi Zeng, Zhenrui Yue, Dong Wang, Hamed Zamani, and Jiawei Han. Search-
 670 r1: Training llms to reason and leverage search engines with reinforcement learning. *CoRR*,
 671 abs/2503.09516, 2025. 1, 4
- 672 Kimi-Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun
 673 Xiao, Chenzhuang Du, Chonghua Liao, Chuning Tang, Congcong Wang, Dehao Zhang, Enming
 674 Yuan, Enzhe Lu, Fengxiang Tang, Flood Sung, Guangda Wei, Guokun Lai, Haiqing Guo, Han
 675 Zhu, Hao Ding, Hao Hu, Hao Yang, Hao Zhang, Haotian Yao, Haotian Zhao, Haoyu Lu, Haoze Li,
 676 Haozhen Yu, Hongcheng Gao, Huabin Zheng, Huan Yuan, Jia Chen, Jianhang Guo, Jianlin Su,
 677 Jianzhou Wang, Jie Zhao, Jin Zhang, Jingyuan Liu, Junjie Yan, Junyan Wu, Lidong Shi, Ling Ye,
 678 Longhui Yu, Mengnan Dong, Neo Zhang, Ningchen Ma, Qiwei Pan, Qucheng Gong, Shaowei Liu,
 679 Shengling Ma, Shupeng Wei, Sihan Cao, Siying Huang, Tao Jiang, Weihao Gao, Weimin Xiong,
 680 Weiran He, Weixiao Huang, Wenhai Wu, Wenyang He, Xianghui Wei, Xianqing Jia, Xingzhe Wu,
 681 Xinran Xu, Xinxing Zu, Xinyu Zhou, Xuehai Pan, Y. Charles, Yang Li, Yangyang Hu, Yangyang
 682 Liu, Yanru Chen, Yejie Wang, Yibo Liu, Yidao Qin, Yifeng Liu, Ying Yang, Yiping Bao, Yulun Du,
 683 Yuxin Wu, Yuzhi Wang, Zaida Zhou, Zhaoji Wang, Zhaowei Li, Zhen Zhu, Zheng Zhang, Zhexu
 684 Wang, Zhilin Yang, Zhiqi Huang, Zihao Huang, Ziyao Xu, and Zonghan Yang. Kimi k1.5: Scaling
 685 reinforcement learning with llms. *CoRR*, abs/2501.12599, 2025. 1, 4
- 686 Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward
 687 Grefenstette, and Roberta Raileanu. Understanding the effects of RLHF on LLM generalisation
 688 and diversity. In *ICLR*. OpenReview.net, 2024. 1
- 689 Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph
 690 Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model
 691 serving with pagedattention. In *SOSP*, pp. 611–626. ACM, 2023. 3.1, A.1
- 692 Xuefeng Li, Haoyang Zou, and Pengfei Liu. Torl: Scaling tool-integrated RL. *CoRR*, abs/2503.23383,
 693 2025. 4
- 694 Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan
 695 Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *ICLR*.
 696 OpenReview.net, 2024. A.5

- 702 Jiaheng Liu, Chenchen Zhang, Jinyang Guo, Yuanxing Zhang, Haoran Que, Ken Deng, Zhiqi Bai, Jie
 703 Liu, Ge Zhang, Jiakai Wang, Yanan Wu, Congnan Liu, Jiamang Wang, Lin Qu, Wenbo Su, and
 704 Bo Zheng. DDK: distilling domain knowledge for efficient large language models. In *NeurIPS*,
 705 2024a. 1
- 706 Junnan Liu, Hongwei Liu, Linchen Xiao, Ziyi Wang, Kuikun Liu, Songyang Gao, Wenwei Zhang,
 707 Songyang Zhang, and Kai Chen. Are your llms capable of stable reasoning? *CoRR*, abs/2412.13147,
 708 2024b. 3.1, A.5
- 709 Junnan Liu, Linhao Luo, Thuy-Trang Vu, and Gholamreza Haffari. Situatedthinker: Grounding LLM
 710 reasoning with real-world through situated thinking. *CoRR*, abs/2505.19300, 2025. 1, 2.3, 4
- 711 Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. GAIA: a
 712 benchmark for general AI assistants. In *ICLR*. OpenReview.net, 2024. 4
- 713 Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke
 714 Zettlemoyer, Percy Liang, Emmanuel J. Candès, and Tatsunori Hashimoto. s1: Simple test-time
 715 scaling. *CoRR*, abs/2501.19393, 2025. 1, 2.2
- 716 NovaSky-Team. Think less, achieve more: Cut reasoning costs by 50 <https://novasky-ai.github.io/posts/reduce-overthinking>, 2025. Accessed: 2025-01. 4
- 717 OpenAI. Gpt-4 is openai's most advanced system, producing safer and more useful responses.
<https://openai.com/index/gpt-4/>, 2023. 3.4
- 718 OpenAI. Learning to reason with llms. <https://openai.com/index/learning-to-reason-with-llms/>, 2024a. Accessed: 2024-09. 1, 4
- 719 OpenAI. Introducing openai o3 and o4-mini. <https://openai.com/index/introducing-o3-and-o4-mini/>, 2024b. Accessed: 2024-12. 1, 4
- 720 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
- 721 Xiao Pu, Mingqi Gao, and Xiaojun Wan. Summarization is (almost) dead. *CoRR*, abs/2309.09558,
 722 2023. 1
- 723 Yiwei Qin, Xuefeng Li, Haoyang Zou, Yixiu Liu, Shijie Xia, Zhen Huang, Yixin Ye, Weizhe Yuan,
 724 Hector Liu, Yuanzhi Li, and Pengfei Liu. O1 replication journey: A strategic progress report - part
 725 1. *CoRR*, abs/2410.18982, 2024. 1, 2.1, 2.2, 4
- 726 David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani,
 727 Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark.
 728 *CoRR*, abs/2311.12022, 2023. 3.1, A.5
- 729 John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy
 730 optimization algorithms. *CoRR*, abs/1707.06347, 2017. 2.3, 2.3
- 731 Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li,
 732 Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open
 733 language models. *CoRR*, abs/2402.03300, 2024. 1, 2.1, 3.1
- 734 Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng,
 735 Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient RLHF framework. In *EuroSys*,
 736 pp. 1279–1297. ACM, 2025. 3.1, A.1
- 737 Joykirat Singh, Raghav Magazine, Yash Pandya, and Akshay Nambi. Agentic reasoning and tool
 738 integration for llms via reinforcement learning. *CoRR*, abs/2505.01441, 2025. 1, 4
- 739 Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and
 740 Ji-Rong Wen. R1-searcher: Incentivizing the search capability in llms via reinforcement learning.
 741 *CoRR*, abs/2503.05592, 2025. 2.3, 4

- 756 Zafir Stojanovski, Oliver Stanley, Joe Sharratt, Richard Jones, Abdulhakeem Adefioye, Jean Kaddour,
 757 and Andreas Köpf. REASONING GYM: reasoning environments for reinforcement learning with
 758 verifiable rewards. *CoRR*, abs/2505.24760, 2025. 3.1, A.4
- 759 Jun Wang, Yang Chen, Shuicheng Yan, Heng Ji, Zihu Wang, Xiaohang Yu, Yifan Zhou, Philip Torr,
 760 Hongru Wang, Zhenfei Yin, Guibin Zhang, Lei Bai, Mengyue Yang, Chen Zhang, Zaibin Zhang,
 761 Yue Liao, Xiangyuan Xue, Songtao Huang, Yijiang Li, Michael Littman, Heng Zhou, Zhongzhi Li,
 762 Yutao Fan, Hejia Geng, and Zelin Tan. The landscape of agentic reinforcement learning for llms:
 763 A survey, 2025. 1, 4
- 764 Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang
 765 Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *ACL*
 766 (1), pp. 9426–9439. Association for Computational Linguistics, 2024. 1
- 767 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi,
 768 Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language
 769 models. In *NeurIPS*, 2022. 2.2
- 770 Liang Wen, Yunke Cai, Fenrui Xiao, Xin He, Qi An, Zhenyu Duan, Yimin Du, Junchen Liu, Lifu
 771 Tang, Xiaowei Lv, Haosheng Zou, Yongchao Deng, Shousheng Jia, and Xiangzheng Zhang. Light-
 772 r1: Curriculum sft, DPO and RL for long COT from scratch and beyond. *CoRR*, abs/2503.10460,
 773 2025. 1, 2.1, 4
- 774 Yongliang Wu, Yizhou Zhou, Zhou Ziheng, Yingzhe Peng, Xinyu Ye, Xinting Hu, Wenbo Zhu,
 775 Lu Qi, Ming-Hsuan Yang, and Xu Yang. On the generalization of SFT: A reinforcement learning
 776 perspective with reward rectification. *CoRR*, abs/2508.05629, 2025. 1
- 777 xAI. Grok 4. <https://x.ai/news/grok-4/>, 2025. Accessed: 2025-07. 1, 4
- 778 Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton
 779 Murray, and Young Jin Kim. Contrastive preference optimization: Pushing the boundaries of LLM
 780 performance in machine translation. In *ICML*. OpenReview.net, 2024. 1
- 781 An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li,
 782 Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin
 783 Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang,
 784 Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia,
 785 Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu
 786 Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *CoRR*, abs/2412.15115, 2024a. 3.1,
 787 B.2
- 788 An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu,
 789 Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu,
 790 Xingzhang Ren, and Zhenru Zhang. Qwen2.5-math technical report: Toward mathematical expert
 791 model via self-improvement. *CoRR*, abs/2409.12122, 2024b. 1, 2.1, 4
- 792 An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang
 793 Gao, Chengan Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu,
 794 Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jian Yang,
 795 Jiaxi Yang, Jingren Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu,
 796 Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men,
 797 Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren,
 798 Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yingger Zhang,
 799 Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu.
 800 Qwen3 technical report. *CoRR*, abs/2505.09388, 2025. 1, 3.1, 4
- 801 Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. LIMO: less is more for
 802 reasoning. *CoRR*, abs/2502.03387, 2025. 1, 4
- 803 Huaiyuan Ying, Shuo Zhang, Linyang Li, Zhejian Zhou, Yunfan Shao, Zhaoye Fei, Yichuan Ma,
 804 Jiawei Hong, Kuikun Liu, Ziyi Wang, Yudong Wang, Zijian Wu, Shuaibin Li, Fengzhe Zhou,
 805 Hongwei Liu, Songyang Zhang, Wenwei Zhang, Hang Yan, Xipeng Qiu, Jiayu Wang, Kai Chen,
 806 and Dahua Lin. Internlm-math: Open math large language models toward verifiable reasoning.
 807 *CoRR*, abs/2402.06332, 2024. 1

- 810 Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong
 811 Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi
 812 Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi
 813 Wang, Hongli Yu, Weinan Dai, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying
 814 Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. DAPO: an open-source
 815 LLM reinforcement learning system at scale. *CoRR*, abs/2503.14476, 2025. [3.1](#), [A.4](#)
- 816 Bowen Zhang and Harold Soh. Extract, define, canonicalize: An llm-based framework for knowledge
 817 graph construction. In *EMNLP*, pp. 9820–9836. Association for Computational Linguistics, 2024.
 818 [1](#)
- 819 Yue Zhang, Yafu Li, Ganqu Cui, Yu Cheng, Zhi Wang, Xiaoye Qu, Jianhao Yan, and Zican Hu.
 820 Learning to reason under off-policy guidance. *CoRR*, abs/2504.14945, 2025a. [1](#), [2.3](#)
- 821 Zhong Zhang, Yaxi Lu, Yikun Fu, Yupeng Huo, Shenzhi Yang, Yesai Wu, Han Si, Xin Cong, Haotian
 822 Chen, Yankai Lin, Jie Xie, Wei Zhou, Wang Xu, Yuanheng Zhang, Zhou Su, Zhongwu Zhai,
 823 Xiaoming Liu, Yudong Mei, Jianming Xu, Hongyan Tian, Chongyi Wang, Chi Chen, Yuan Yao,
 824 Zhiyuan Liu, and Maosong Sun. Agentcpm-gui: Building mobile-use agents with reinforcement
 825 fine-tuning. *CoRR*, abs/2506.01391, 2025b. [4](#)
- 826
- 827
- 828
- 829
- 830
- 831
- 832
- 833
- 834
- 835
- 836
- 837
- 838
- 839
- 840
- 841
- 842
- 843
- 844
- 845
- 846
- 847
- 848
- 849
- 850
- 851
- 852
- 853
- 854
- 855
- 856
- 857
- 858
- 859
- 860
- 861
- 862
- 863

864

865

866

867

Appendix

868

869

870

871

Table of Contents

872

A More Implementation Details	18
A.1 Training Details	18
A.2 Full Training Prompt	18
A.3 Inference Prompt	19
A.4 Training Data	19
A.5 Evaluation Benchmarks	22
B Additional Experimental Results and Analysis	22
B.1 Performance of Teacher LLMs	22
B.2 AGENTIC DISTILLATION on Larger Student LLMs	22
B.3 AGENTIC DISTILLATION on Long-CoT Student LLMs	23
B.4 Case Study	23
C Limitations	25
D LLM usage	26

885

886

887

888

889

890

891

892

893

894

895

896

897

898

899

900

901

902

903

904

905

906

907

908

909

910

911

912

913

914

915

916

917

918 A MORE IMPLEMENTATION DETAILS
919920 A.1 TRAINING DETAILS
921922 Training utilized the veRL (Sheng et al., 2025) and vLLM (Kwon et al., 2023) frameworks on the
923 clusters equipped with NVIDIA H100 GPUs. Table 3 present the detailed training parameters for
924 AGENTIC DISTILLATION.
925926 Table 3: Training Parameters.
927

928 Parameters	929 Values
930 Batch Size	256
931 Number of Rollout Per Question	8
932 Rollout Temperature	1.0
933 Rollout Top- p	1.0
934 Rollout Top- k	-1
935 Maximum Number of Generation Tokens	16384
936 Learning Rate	1e-6
937 KL Loss Coefficient	0.001
938 ϵ_{\min}	0.2
939 ϵ_{\max}	0.28
940 Gradient Clipping	1.0
941 Number of Training Steps	300

942 A.2 FULL TRAINING PROMPT
943944 Prompt A.1 illustrates the full training prompt.
945

946 Prompt A.1: Full Training Prompt

947 OBJECTIVE:

948 To answer a User’s question by providing a clear, verifiable reasoning process, potentially interacting with an
949 external environment.
950

951 INTERACTION PROTOCOL:

952 For each question you receive, you MUST follow this two-step process:
953

954 Step 1: Reasoning Process

- 955 •
- Decomposition:**
- Break down the user’s question into a logical, step-by-step sequence of reasoning. Start
-
- 956 from the most basic facts and build upon them.
-
- 957

958 • **External Inquiry (Optional but Encouraged):**

- 959
- 960 – You may issue up to max_turns queries to an External Environment to validate hypotheses, clarify
961 information, or advance your reasoning.
 - 962 – Each query must be a self-contained question enclosed in `<query>...</query>` tags.
 - 963 – **Wait for the `<result>...</result>` block** from the environment before continuing your reasoning.
 - 964 – **Critically analyze and integrate** the content from the `<result>...</result>` block into your reasoning chain.
 - 965 – Do not invent, assume, or hallucinate any `<result>` content. Your reasoning must be grounded in the provided results.

966 Step 2: Final Answer

- 967
- 968 • After your reasoning is complete, state your final answer clearly.
 - 969 • The final answer, and only the final answer, MUST be enclosed in “\boxed{...}”.

972
973

A.3 INFERENCE PROMPT

974
975
976
977

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.

978
979

Prompt A.2: Prompt for Mathematical Reasoning Benchmarks

980
981
982

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

983

A.4 TRAINING DATA

985
986

The training data of AGENTIC DISTILLATION is composed of three parts:

987
988
989
990
991
992

- **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.

993
994
995
996
997

```

1004     tasks = [
1005         ("ab", 1.0, {
1006             "seed": 42,
1007             "length": 10,
1008             "size": size
1009         }),
1010         ("ab", 1.0, {
1011             "seed": 42,
1012             "length": 15,
1013             "size": size
1014         }),
1015         ("acre", 1.0, {
1016             "seed": 42,
1017             "size": size
1018         }),
1019         ("advanced_geometry", 1.0, {
1020             "seed": 42,
1021             "min_coord": -100,
1022             "max_coord": 100,
1023             "size": size
1024         }),
1025         ("aiw", 1.0, {
1026             "seed": 42,
1027             "max_entities": 10,
1028             "size": size
1029         }),
1030         ("cryptarithm", 1.0, {
1031             "seed": 42,
1032             "size": size
1033         })
1034     ]

```

```

1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
    "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,
    "size": size
  })

```


1134 A.5 EVALUATION BENCHMARKS
11351136 The following details describe our evaluation benchmarks:
1137

- **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.

1159 B ADDITIONAL EXPERIMENTAL RESULTS AND ANALYSIS
1160

1162 B.1 PERFORMANCE OF TEACHER LLMs

1163 Table 4 shows the performance of Qwen3-30B-A3B-Instruct-2507 and GPT-OSS-20B, which are
1164 utilized as teacher LLMs in this work. The performance of teacher LLMs can be seen as the upper
1165 bound of the distillation.
11661167 Table 4: Performance of Qwen3-30B-A3B-Instruct-2507 and GPT-OSS-20B.
1168

1169 Methods	1170 Math				1171 Science		1172 Code		1173 Puzzle
	AIME24	AIME25	MATH500	LMB	GPQA-D	MBPP	LCB	RG	
	Avg@16	Avg@16	Avg@4	Avg@4	Avg@4	Avg@4	Avg@4	Avg@4	Avg@4
<i>Qwen3-30B-A3B-Instruct-2507</i>									
-	76.88	63.96	96.75	44.50	55.18	84.05	44.74	19.54	
<i>GPT-OSS-20B</i>									
-	78.62	73.75	96.45	50.50	59.22	93.68	60.53	13.98	

1178 B.2 AGENTIC DISTILLATION ON LARGER STUDENT LLMs

1180 In this section, we evaluate the effectiveness of AGENTIC DISTILLATION on student LLMs with
1181 larger parameter sizes, specifically training Qwen2.5-32B-Instruct (Yang et al., 2024a) with AGENTIC
1182 DISTILLATION. As shown in Table 5, the evaluation results demonstrate that AGENTIC DISTIL-
1183 LATION remains effective for larger-scale models, with AGENTIC DISTILLATION-trained models
1184 outperforming baseline models across all benchmarks. Notably, the performance improvements
1185 for Qwen2.5-32B-Instruct are more pronounced compared to those for Qwen2.5-7B-Instruct. This
1186 enhanced improvement may stem from the 32B model’s stronger baseline capabilities, enabling it
1187 to formulate higher-quality questions and acquire knowledge more efficiently during training with
1188 AGENTIC DISTILLATION.

1188 Table 5: Experimental results of AGENTIC DISTILLATION and baselines with Qwen2.5-32B-Instruct
 1189 as the student LLM. We report the average performance for 16 runs on AIME24 and AIME25, and 4
 1190 runs on others. We abbreviate LMB as LiveMathBench v202505, LCB as LiveCodeBench v6, RG as
 1191 Reasoning Gym, MI as Masked Interaction, and AD as AGENTIC DISTILLATION.

Methods	Math				Science		Code		Puzzle
	AIME24	AIME25	MATH500	LMB	GPQA-D	MBPP	LCB	RG	
	Avg@16	Avg@16	Avg@4	Avg@4	Avg@4	Avg@4	Avg@4	Avg@4	Avg@4
Student LLM: <i>Qwen2.5-32B-Instruct</i> , Teacher LLM: <i>Qwen3-30B-A3B-Instruct-2507</i>									
Original	14.38	13.12	80.85	12.25	47.10	85.12	24.71	13.06	
+SFT	<u>16.83</u>	<u>15.79</u>	<u>83.47</u>	<u>15.33</u>	46.13	84.74	23.15	15.61	
+RL	15.73	14.11	82.51	13.33	<u>50.89</u>	<u>86.71</u>	<u>26.43</u>	<u>17.62</u>	
+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	AIME25	MATH500	LMB	GPQA-D	MBPP	LCB	RG	
	Avg@16	Avg@16	Avg@4	Avg@4	Avg@4	Avg@4	Avg@4	Avg@4	Avg@4
Student LLM: <i>DeepSeek-R1-Distill-Qwen-1.5B</i> , Teacher LLM: <i>Qwen3-30B-A3B-Instruct-2507</i>									
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	<u>28.43</u>	<u>25.70</u>	<u>86.82</u>	<u>17.39</u>	<u>34.68</u>	<u>65.05</u>	<u>14.72</u>	<u>13.53</u>	
+AD	30.56	29.21	88.47	18.90	36.53	67.19	17.06	16.44	

B.4 CASE STUDY

We present examples of AGENTIC DISTILLATION’s training trajectories in Case B.1 and Case B.2. These instances demonstrate the student model’s ability to formulate reasonable and meaningful queries, leverage answers from the teacher model, and successfully complete the reasoning process. Through this interaction, the student effectively acquires new information and enhances its inherent capabilities.

1242

1243

1244

QUESTION

1245

1246

1247

1248

1249

1250

1251

1252

1253

1254

1255

Jen enters a lottery by picking 4 distinct numbers from $S = \{1, 2, 3, \dots, 9, 10\}$. 4 numbers are randomly chosen from S . She wins a prize if at least two of her numbers were 2 of the randomly chosen numbers, and wins the grand prize if all four of her numbers were the randomly chosen numbers. The probability of her winning the grand prize given that she won a prize is $\frac{m}{n}$ where m and n are relatively prime positive integers. Find $m + n$.

ANSWER

116

TRAJECTORY**Step 1: Reasoning Process****• Decomposition:**

1. **Identify the Objective:** Find the sum $m + n$, where $\frac{m}{n}$ is the conditional probability that Jen wins the grand prize given that she wins *any* prize.

2. Define the Events:

- Total pool of numbers: $S = \{1, 2, \dots, 10\}$ (Size 10).
- Jen's pick (J): 4 specific numbers.
- Winning numbers (W): 4 randomly chosen numbers.
- Event E (Prize): The intersection $|J \cap W|$ is at least 2 (i.e., 2, 3, or 4 matches).
- Event G (Grand Prize): The intersection $|J \cap W|$ is exactly 4.

3. **Set up the Formula:** The conditional probability is $P(G|E) = \frac{N(G \cap E)}{N(E)}$. Since G is a subset of E (4 matches implies at least 2), this simplifies to $\frac{N(G)}{N(E)}$.

4. Combinatorial Model: To find the number of ways to match exactly k numbers:

- We need to choose k numbers from the 4 winning numbers: $\binom{4}{k}$.
- We need to choose the remaining $4 - k$ numbers from the 6 non-winning numbers: $\binom{6}{4-k}$.

5. **External Verification:** I need to calculate the specific combinatorial values for $k = 2, 3, 4$ and their sum to determine $N(E)$ and $N(G)$. I will ask an external environment to perform these calculations.

• External Inquiry:

<query>

Calculate the number of ways to match exactly k numbers in a lottery where 4 numbers are drawn from 10. Specifically, compute the values for $C(4,k) * C(6, 4-k)$ for $k=2, k=3$, and $k=4$, and find the sum of these three results.

</query>

<result>

115

</result>

• Analysis and Integration:

- The external result confirms the counts for each case.

- **$N(E)$ (Total ways to win a prize):** This is the sum of outcomes for $k = 2, 3, 4$. According to the result, $90 + 24 + 1 = 115$.

- **$N(G)$ (Ways to win grand prize):** This corresponds to the $k = 4$ case, which is 1.

- **Calculate Probability:** The conditional probability is $\frac{1}{115}$.

- **Identify m and n :** $m = 1, n = 115$.

- **Check Condition:** 1 and 115 are relatively prime positive integers.

- **Final Step:** Calculate $m + n = 1 + 115 = 116$.

Step 2: Final Answer

The probability that Jen wins the grand prize given she wins a prize is $\frac{1}{115}$. Therefore, $m = 1$ and $n = 115$, and the sum $m + n$ is 116.

1296
1297

Case B.2: Case on Bamoogle

1298

QUESTION

1299

What is the capital of the second largest state in the US by area?

1300

1301

ANSWER

1302

austin

1303

1304

TRAJECTORY

1305

Step 1: Reasoning Process

1306

1307

- **Decomposition:** To determine the capital of the second largest state in the US by area, I need to follow a logical sequence.
 - First, I must definitively identify which US state ranks second in terms of total area.
 - Second, once that state is identified, I need to determine its capital city.
- **External Inquiry:** I will start by asking the external environment to identify the second largest state in the US.

1315

1316

1317

1318

1319

1320

1321

1322

1323

```
<query>
  What is the second largest state of US?
</query>
<result>
  Texas
</result>
```

The external result confirms that **Texas** is the second largest state in the US by area. Now I need to find the capital of Texas.

1324

1325

1326

1327

1328

1329

```
<query>
  What is the capital of Texas?
</query>
<result>
  Austin
</result>
```

From the retrieved information, I have verified that the capital of Texas is Austin.

1330

1331

1332

Step 2: Final Answer

1333

1334

1335

1336

1337

1338

1339

C LIMITATIONS

1340

1341

1342

1343

1344

1345

1346

1347

1348

1349

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.

1350 **D LLM USAGE**
13511352 In this paper, the use of LLMs is intentionally restricted to the final stages of the research process,
1353 specifically for refining and proofreading the written content. The LLMs are employed solely
1354 to enhance the clarity, coherence, and grammatical accuracy of the text, ensuring effective and
1355 professional communication of the presented ideas. Importantly, LLMs played no role in the core
1356 components of this work, including the development of the research methodology, the design of
1357 experiments, or the analysis of results. We are aware that we will be responsible for all content in the
1358 paper.
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403