Not All Correct Answers Are Equal: Why Your Distillation Source Matters

Xiaoyu Tian, Yunjie Ji, Haotian Wang, Shuaiting Chen,
Sitong Zhao, Yiping Peng, Han Zhao, Xiangang Li

a-m-team[1]

Abstract

Distillation has emerged as a practical and effective approach to enhance the reasoning capabilities of open-source language models. In this work, we conduct a large-scale empirical study on reasoning data distillation by collecting verified outputs from three state-of-the-art teacher models—AM-Thinking-v1, Qwen3-235B-A22B, and DeepSeek-R1—on a shared corpus of 1.89 million queries. We construct three parallel datasets and analyze their distributions, revealing that AM-Thinking-v1-distilled data exhibits greater token length diversity and lower perplexity. Student models trained on each dataset are evaluated on reasoning benchmarks including AIME2024, AIME2025, MATH500, and LiveCodeBench. The AM-based model consistently achieves the best performance (e.g., 84.3 on AIME2024, 72.2 on AIME2025, 98.4 on MATH500, and 65.9 on LiveCodeBench) and demonstrates adaptive output behavior—producing longer responses for harder tasks and shorter ones for simpler tasks. These findings highlight the value of high-quality, verified reasoning traces. We release the AM-Thinking-v1 and Qwen3-235B-A22B distilled datasets to support future research on open and high-performing reasoning-oriented language models. The datasets are publicly available on Hugging Face[2].

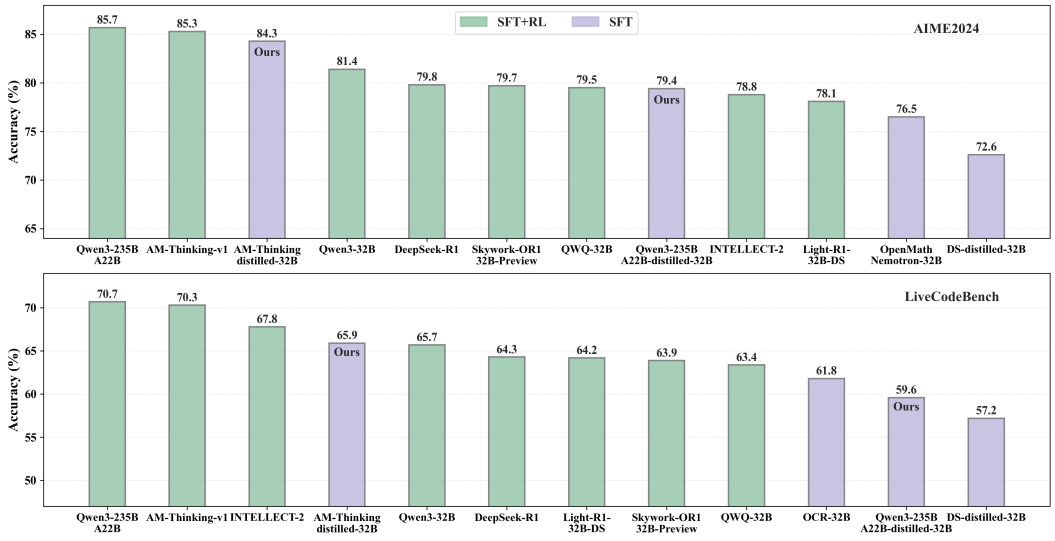

Figure 1: Open-source model benchmarks on AIME2024/LiveCodeBench.

[1]The a-m-team is an internal team at Beike (Ke.com), dedicated to exploring AGI technology.
[2]Datasets are available on Hugging Face: AM-Thinking-v1-Distilled, AM-Qwen3-Distilled.

1 Introduction

Recent work has demonstrated the effectiveness and efficiency of distillation-based training for enhancing the reasoning ability of Large Language Models (LLMs) [1, 2, 3]. By transferring reasoning traces from stronger teacher models, distilled data enables smaller or open-source models to achieve significant improvements on challenging tasks such as mathematics, coding, and scientific reasoning.

Building on this line of research, we systematically distilled reasoning data from three state-of-the-art models: DeepSeek-R1 [1], Qwen3-235B-A22B [3], and AM-Thinking-v1 [4]. For each of approximately 1.89 million identical queries, we collected full chain-of-thought responses from all three models, resulting in three parallel large-scale datasets. This unique setup allows for a direct comparison of reasoning styles and data distributions across leading models.

We carefully processed and cleaned all three datasets, including thorough deduplication, strict filtering, and contamination removal. We further analyzed the data distributions and content diversity to provide a comprehensive understanding of the strengths and characteristics of each distillation source.

Our experiments show that models trained with data distilled from AM-Thinking-v1 achieve particularly strong performance. On challenging reasoning benchmarks, such as AIME2024 [5] (84.3), AIME2025 [6] (72.2), MATH500 [7] (98.4), and LiveCodeBench [8] (65.9), the AM-Thinking-v1 distilled model consistently outperforms those trained on Qwen3-235B-A22B or DeepSeek-R1 data. Moreover, our analysis reveals that the AM-Thinking-v1 distilled model exhibits an adaptive generation length: producing longer responses on harder tasks (e.g., AIME, LiveCodeBench), and shorter ones on simpler datasets (e.g., MATH500). This behavior aligns with the token-level distribution of the AM-distilled dataset, which contains both short and long responses more frequently than the other two sources.

These results highlight the practical value of large-scale, high-quality reasoning data distillation, for improving open-source LLMs. To promote further progress in the field, we release both the AM-Thinking-v1 and Qwen3-235B-A22B distilled datasets[3]. We hope our work provides valuable resources and insights for the open-source community, enabling more effective reasoning-focused model development and contributing to the broader progress of reasoning research.

2 Data

This section first introduces the data preprocessing and distillation pipeline used to construct our training corpus, and then presents a detailed analysis of the resulting datasets in terms of distribution, length, and quality.

2.1 Data Collection and Query Processing

To support robust and comprehensive model training, we constructed a large-scale training corpus by aggregating data from a diverse set of publicly available open-source corpora. These corpora span a broad range of NLP tasks, including mathematical reasoning, code generation, scientific reasoning, instruction following, multi-turn dialogue, and general reasoning. For downstream analysis and targeted data processing, each data source was systematically assigned to a specific task category.

Training Data Categories The aggregated training data were classified as follows:

- **Mathematical Reasoning**: Datasets requiring advanced numerical reasoning and multi-step logic, such as OpenR1-Math-220k [9], Big-Math-RL-Verified [10], NuminaMath [11], among others.
- **Code Generation**: Datasets aimed at enhancing code synthesis and programmatic problem-solving abilities, including PRIME [12], DeepCoder [13], KodCode [14].
- **Scientific Reasoning**: Datasets emphasizing reasoning within the natural sciences, such as task_mmmlu [15], chemistryQA [16], and LOGIC-701 [17].

[3]Note that since the distillation data of DeepSeek-R1 is easily accessible, we only release the distillation data of AM-Thinking-v1 and Qwen3-235B-A22B.

- **Instruction Following (IF)**: Data focused on instruction comprehension and faithful execution, including Llama-Nemotron-Post-Training-Dataset [18], tulu-3-sft-mixture [19], if-eval-like, and AutoIF.
- **Multi-turn Conversation**: Corpora curated to train dialogue agents on contextually coherent and consistent multi-turn interactions, such as InfinityInstruct [20], OpenHermes-2.5 [21], and ultra_chat [22].
- **General Reasoning**: Datasets covering diverse open-ended reasoning and general knowledge tasks, including evol [23], open_orca [24], flan [25].

Query Preprocessing To guarantee the reliability of subsequent model training, we applied rigorous multi-stage preprocessing to the raw queries:

1. **Deduplication**: Exact duplicate queries (identical text) were removed.
2. **Filtering**:
   - Queries with a high Unicode character ratio were discarded to eliminate corrupted or meaningless samples.
   - Incomplete or empty queries were excluded.
   - Instances containing URLs or tabular structures were filtered out to reduce noise and hallucination risk.
3. **Decontamination**: To mitigate data contamination, especially regarding the core evaluation set (e.g., AIME2024 [5]), we conducted both exact match filtering and semantic deduplication. The latter leveraged the bge-m3 embedding model [26] to compute semantic similarity, removing queries exceeding a threshold of 0.9 with respect to the evaluation set.

2.2 Data Distilling

After preprocessing, we performed large-scale data distillation to further enhance the quality of our training corpus.

Distillation Framework For each preprocessed query, we adopted an incremental distillation strategy using three state-of-the-art models: AM-Thinking-v1 [4], Qwen3-235B-A22B [3], and DeepSeek-R1 [1]. Each query was independently distilled by these three models. For every model, the distillation process was repeated on the same query until the generated response satisfied the verification criterion (i.e., the verification score ≥ 0.9). Consequently, each query yielded up to three high-quality distilled outputs, corresponding to the three models, with each output refined iteratively until it passed the automatic verification.

Automatic Verification and Scoring To ensure the reliability and correctness of the distilled data, we employed automatic verification procedures tailored for each data category, assigning a verification score (*verify_score*) to every model-generated response:

- **Mathematical Reasoning**: Responses were verified using a two-stage process—first with Math-Verify[4], and, if necessary, subsequently with Qwen2.5-7B-Instruct [27, 28]. Each result was assigned a binary verification score.
- **Code Generation**: Each code response was validated in a sandbox environment using up to 10 test cases (assert and input-output for Python; input-output for C++), with the verification score reflecting the pass rate.
- **Scientific Reasoning**: The similarity between predicted and reference answers was assessed using Qwen2.5-7B-Instruct [27, 28], yielding a normalized score.
- **Instruction Following**: Responses were verified with the ifeval validator, supplementing missing constraints with Qwen2.5-72B-Instruct [27, 28]. The mean pass rate over all constraints was taken as the verification score.

[4] `https://github.com/huggingface/Math-Verify`

- **Multi-turn Conversations and General Reasoning**: Decision-Tree-Reward-Llama-3.1-8B [29] was used to evaluate coherence, correctness, and helpfulness, which were aggregated into a normalized composite score.

A unified verification score threshold of 0.9 was used across all data categories.

Quality Assurance Measures To further enhance data quality, we introduced several additional validation and filtering strategies:

- **Perplexity-based Filtering**: We computed perplexity scores using a strong 32B language model [30], with each model employing a different threshold. Notably, responses distilled from AM-Thinking-v1 demonstrated the lowest perplexity among the three models.
- **High-Frequency Ngram Filtering**: 20-token ngrams occurring more than 20 times were identified and removed to reduce template-like redundancy.
- **Logical and Structural Validation**: Checks included ensuring an even number of dialogue turns for conversation data, explicit presence of both reasoning ("think") and answer segments in each sample.

Ultimately, this process yielded a comprehensive dataset of 1.89 million queries, each paired with high-quality, verified responses distilled from all three models.

2.3 Data Analysis

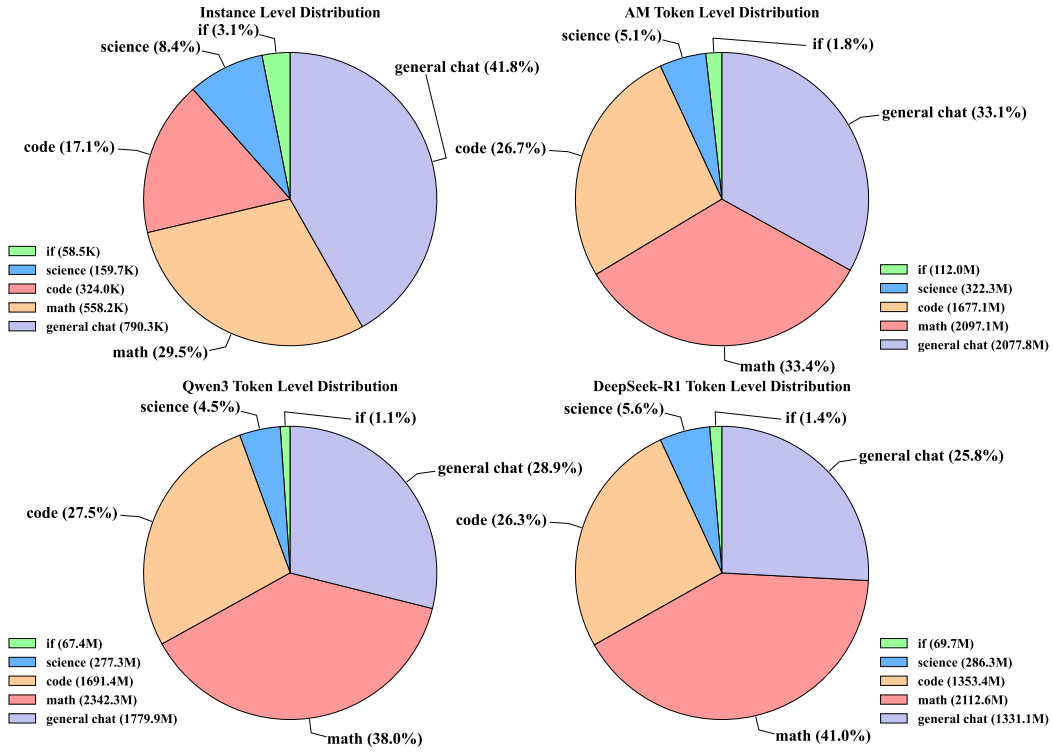

Figure 2: Instance-level and token-level output distributions are analyzed for AM-Thinkin-v1, Qwen3-235B-A22B, and DeepSeek-R1. The general chat includes both multi-turn conversations and other types of data.

We conduct a detailed analysis of the training data distilled from three different large-scale models: AM-Thinking-v1 [4], Qwen3-235B-A22B [3], and DeepSeek-R1 [1]. This comparative analysis covers the instance-level and token-level output distributions, token length characteristics, and perplexity (PPL) distributions, providing insight into the data quality and structural tendencies of each dataset.

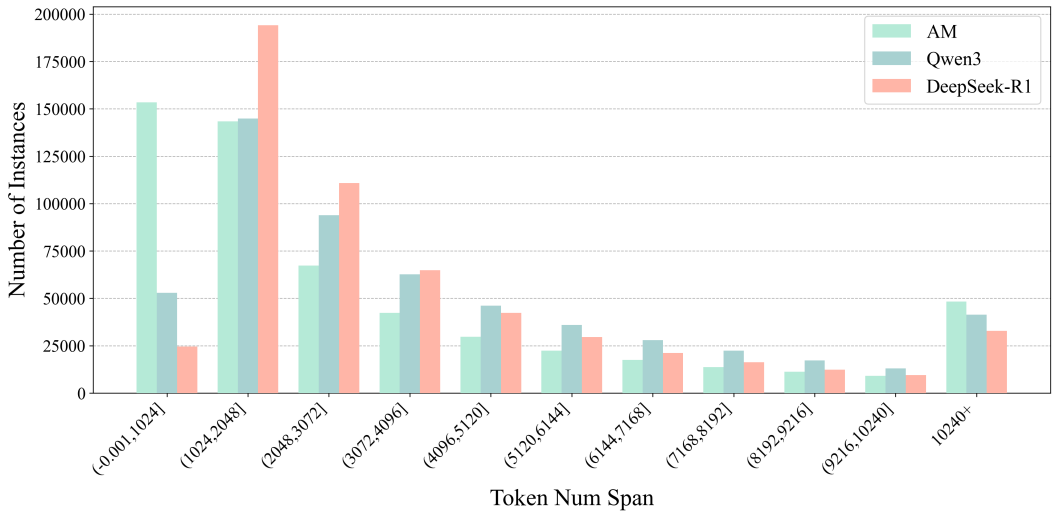

Figure 3: Token span distribution of instances for AM-Thinking-v1, Qwen3-235B-A22B, and DeepSeek-R1 on math.

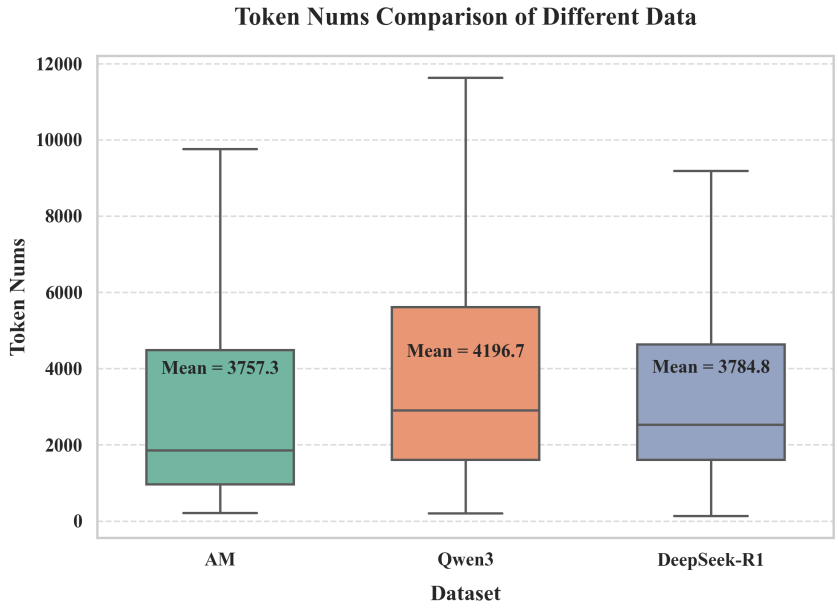

Figure 4: Token count distributions for AM-Thinking-v1, Qwen3-235B-A22B, and DeepSeek-R1 datasets. Box plots show the distribution of token numbers, with means labeled. Qwen3-235B-A22B has the highest average token count, followed by DeepSeek-R1 and AM-Thinking-v1.

Figure 2 shows the output distribution of these datasets at both the instance and token levels. The instance-level distribution (top-left) reveals that this dataset includes a high proportion of general chat (41.8%), followed by math (29.5%) and code (17.1%). In contrast, Qwen3-235B-A22B and DeepSeek-R1 token-level distributions (bottom charts) show a more pronounced focus on math (38.0% and 41.0% respectively), with general chat and code sharing similar proportions. Notably, AM-Thinking-v1's token-level distribution (top-right) also emphasizes math (33.4%), albeit to a lesser extent than DeepSeek-R1. Science and instruction-following (IF) data make up a minor share across all datasets.

Further, Figure 3 presents the token span distribution specifically for math instances. It demonstrates that AM-Thinking-v1's math data exhibits a highly dispersed distribution—many short sequences

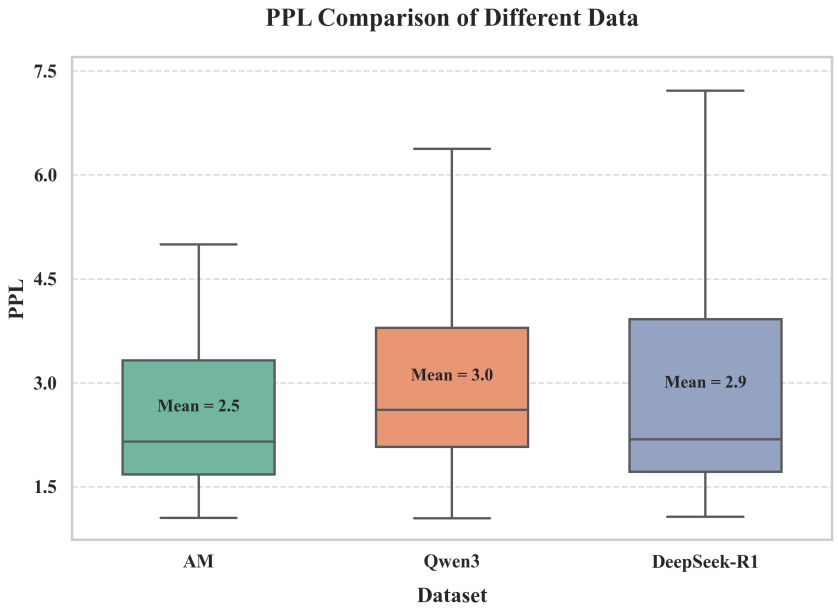

Figure 5: Perplexity (PPL) distributions for AM-Thinking-v1, Qwen3-235B-A22B, and DeepSeek-R1 datasets. Box plots show PPL distributions, with means labeled. AM-Thinking-v1 achieves the lowest mean PPL, indicating better overall quality.

(under 1024 tokens) and a substantial portion of very long sequences (10240+ tokens). This reflects the token distribution characteristics under a fixed query set, where AM-Thinking-v1 tends to produce both short and very long responses more frequently. In contrast, Qwen3-235B-A22B's data generally exhibits longer token spans, indicating a tendency toward producing longer responses. DeepSeek-R1's token spans are mostly concentrated between 1k and 8k tokens, showing a moderate range of response lengths.

Box plots in Figure 4 provide further insight into the average token length per instance. The Qwen3-235B-A22B dataset has the highest mean token count (4196.7), followed by DeepSeek-R1 (3784.8) and AM-Thinking-v1 (3757.3). This aligns with the histogram observation that Qwen3-235B-A22B emphasizes longer instances, whereas AM-Thinking-v1 covers a broader range of lengths, including extremely short and long sequences.

Finally, Figure 5 compares the perplexity (PPL) across the datasets. Perplexity is a key measure of language model performance, with lower values indicating better quality. Among the three datasets, AM-Thinking-v1 achieves the lowest mean PPL (2.5), suggesting that its distilled outputs are generally of higher quality. DeepSeek-R1 (mean PPL = 2.9) performs slightly better than Qwen3 (mean PPL = 3.0), highlighting the relatively strong performance of AM-distilled data in terms of perplexity.

3 Experiments

3.1 Training Configuration

Training is conducted based on the Qwen2.5-32B [27, 28] base model. As suggested by [2, 4], all three models are trained with a learning rate of 8e-5, a maximum sequence length of 32k (using sequence packing), and a global batch size of 64 for 2 epochs. Samples longer than 32k tokens are excluded. We apply cosine warmup with 5% of total steps, and the learning rate decays to zero thereafter. For multi-turn dialogues, only the final response containing the reasoning process is used as the training target to focus learning on reasoning.

3.2 Benchmarks and Evaluation Setup

To rigorously assess our models' capabilities, we select a diverse suite of challenging benchmarks across mathematical reasoning, programming, and general chatbot performance:

- **AIME2024** [5]: A high-difficulty dataset featuring 30 integer-answer questions from the 2024 American Invitational Mathematics Examination, designed to evaluate accurate mathematical problem-solving skills.
- **AIME2025** [6]: Contains 30 new problems curated from the 2025 AIME Part I and Part II exams, offering a forward-looking benchmark for advanced mathematical reasoning.
- **LiveCodeBench (LCB)** [8]: A dynamically evolving, contamination-free code generation benchmark. Tasks are aggregated from platforms like LeetCode, Codeforces, and AtCoder. In line with prior works such as Qwen3 [3], we use a snapshot of queries submitted between October 2024 and February 2025.
- **MATH500** [7]: A benchmark consisting of 500 challenging math word problems designed to evaluate the problem-solving abilities of large language models. It covers diverse mathematical domains such as algebra, geometry, calculus, and number theory, requiring models to perform multi-step reasoning and symbolic manipulation.

All benchmarks were evaluated under uniform conditions. The generation length was capped at 49,152 tokens. For stochastic decoding, we consistently adopted a temperature of 0.6 and top-p of 0.95 across applicable tasks.

Response sampling was tailored to the nature of each benchmark:

- **AIME2024 and AIME2025**: For each question, we generated 64 outputs to estimate pass@1 accuracy.
- **LiveCodeBench**: We sampled 16 completions per prompt to compute pass@1.
- **MATH500**: Each prompt was answered once, and we sampled 4 times to compute pass@1.

Followed by AM-Thinking-v1[4], a unified system prompt was employed across all tasks to standardize output format and encourage reasoning:

```
You are a helpful assistant. To answer the user's question,
you first think about the reasoning process and then provide
the user with the answer. The reasoning process and answer
are enclosed within <think> </think> and <answer> </answer>
tags, respectively, i.e., <think> reasoning process here
</think> <answer> answer here </answer>.
```

User prompts were benchmark-specific:

- **AIME tasks and MATH500**: A supplementary instruction was added: `Let's think step by step and output the final answer within \box.`
- **LiveCodeBench**: Original task prompts were used without any alterations.

4 Results and Analysis

We evaluate the models on the reasoning benchmarks described in Section 3.2, using models trained with data distilled from AM-Thinking-v1, Qwen3-235B-A22B, and DeepSeek-R1. The evaluation results are presented in Table 1.

As shown in Table 1, the model distilled from AM-Thinking-v1 consistently achieves the highest accuracy across all benchmarks. On the more challenging math tasks, AIME2024 and AIME2025, it attains scores of 84.3 and 72.2, respectively, outperforming the Qwen3- and DeepSeek-distilled models by a considerable margin. It also leads on MATH500 (98.4) and LiveCodeBench (65.9), indicating broad generalization across both mathematical and code-based reasoning tasks.

To better understand model behavior, we analyze the average generation length per sample across benchmarks (Table 2). Interestingly, the AM-Thinking-v1$_{\text{Distilled}}$ model produces notably longer

Table 1: Comparison across reasoning benchmarks using distilled data from different teacher models.

	AM-Thinking-v1$_{\text{Distilled}}$	Qwen3-235B-A22B$_{\text{Distilled}}$	DeepSeek-R1$_{\text{Distilled}}$
AIME2024	**84.3**	79.4	70.9
AIME2025	**72.2**	62.2	52.8
MATH500	**98.4**	93.9	95.8
LiveCodeBench (v5, 2024.10–2025.02)	**65.9**	59.6	57.0

Table 2: Average generation length (tokens per sample) across reasoning benchmarks.

	AM-Thinking-v1$_{\text{Distilled}}$	Qwen3-235B-A22B$_{\text{Distilled}}$	DeepSeek-R1$_{\text{Distilled}}$
AIME2024	15273.8	13516.4	11853.5
AIME2025	18199.2	16975.7	13495.9
MATH500	3495.7	6429.4	3613.0
LiveCodeBench (v5, 2024.10–2025.02)	23426.9	13576.7	30731

outputs on more complex tasks: 15273.8 and 18199.2 tokens for AIME2024 and AIME2025, respectively, and 23426.9 for LiveCodeBench. In contrast, on the simpler MATH500 benchmark, its average generation length (3495.7) is shorter than that of the Qwen3-235B-A22B$_{\text{Distilled}}$ model. This adaptive generation pattern suggests that the AM-distilled model can better modulate its output length based on task complexity—generating more detailed solutions when needed while remaining concise on simpler problems. This aligns with our earlier analysis in Section 2.3, where the AM-Thinking-v1 distilled dataset exhibited a higher proportion of both short and long token sequences. The broader distribution of token lengths in the training data likely contributes to the model's improved ability to adjust its response length dynamically. Such length modulation is a desirable property in reasoning tasks, where overgeneration or undergeneration can negatively impact performance.

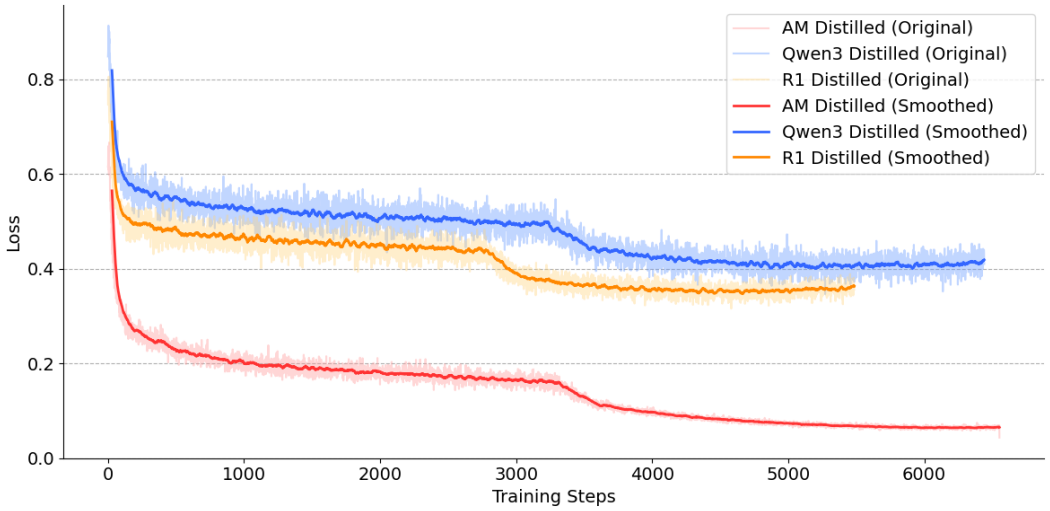

Figure 6: Loss curves of AM-Thinking-v1-Distilled, DeepSeek-R1-Distilled and Qwen3-235B-A22B-Distilled.

We further compare training dynamics by examining the loss curves shown in Figure 6. The AM-Thinking-v1$_{\text{Distilled}}$ maintains a consistently lower training loss than its Qwen3-235B-A22B$_{\text{Distilled}}$ and DeepSeek-R1$_{\text{Distilled}}$ counterparts throughout the optimization process. This observation supports the notion that the AM-Thinking-v1 dataset provides more learnable, coherent, and high-quality supervision signals for the base model.

5 Conclusion and Future Work

In this work, we present a comprehensive empirical study on reasoning data distillation for open-source language models. Using three state-of-the-art teacher models—AM-Thinking-v1, Qwen3-235B-A22B, and DeepSeek-R1—we constructed a large-scale parallel corpus comprising 1.89 million verified reasoning samples. Through rigorous data preprocessing, verification scoring, and quality assurance, we ensured the construction of high-quality training data suitable for robust student model learning.

Empirical results across a diverse set of benchmarks, including AIME2024 (84.3), AIME2025 (72.2), MATH500 (98.4), and LiveCodeBench (65.9), demonstrate that models trained on AM-Thinking-v1-distilled data consistently achieve strong performance.

To gain deeper insights into model behavior, we conducted detailed analysis of generation behavior and training dynamics. We observed that the AM-distilled model exhibits adaptive generation length—producing longer responses on harder tasks and shorter ones on easier benchmarks—indicating its capacity to adjust to task difficulty. This aligns with our earlier data analysis showing that AM-Thinking-v1-distilled data features a wide range of token lengths, providing stronger support for adaptive reasoning.

Looking ahead, a promising direction for future work is to further enhance these models using reinforcement learning techniques, such as Proximal Policy Optimization (PPO) or Generalized Group Relative Policy Optimization (GRPO), to enhance reasoning ability and alignment. We release the distilled datasets based on AM-Thinking-v1 and Qwen3-235B-A22B to support ongoing research in open and high-performing reasoning-oriented language models.

References

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

[2] Xiaoyu Tian, Sitong Zhao, Haotian Wang, Shuaiting Chen, Yiping Peng, Yunjie Ji, Han Zhao, and Xiangang Li. Deepdistill: Enhancing llm reasoning capabilities via large-scale difficulty-graded data training, 2025.

[3] Qwen Team. Qwen3, April 2025.

[4] Yunjie Ji, Xiaoyu Tian, Sitong Zhao, Haotian Wang, Shuaiting Chen, Yiping Peng, Han Zhao, and Xiangang Li. Am-thinking-v1: Advancing the frontier of reasoning at 32b scale, 2025.

[5] MAA. American invitational mathematics examination - aime. `https://maa.org/math-competitions/american-invitational-mathematics-examination-aime`, feb 2024. Accessed in February 2024, from American Invitational Mathematics Examination - AIME 2024.

[6] Yixin Ye, Yang Xiao, Tiantian Mi, and Pengfei Liu. Aime-preview: A rigorous and immediate evaluation framework for advanced mathematical reasoning. `https://github.com/GAIR-NLP/AIME-Preview`, 2025. GitHub repository.

[7] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step, 2023.

[8] Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*, 2024.

[9] Hugging Face. Open r1: A fully open reproduction of deepseek-r1, January 2025.

[10] Alon Albalak, Duy Phung, Nathan Lile, Rafael Rafailov, Kanishk Gandhi, Louis Castricato, Anikait Singh, Chase Blagden, Violet Xiang, Dakota Mahan, and Nick Haber. Big-math: A large-scale, high-quality math dataset for reinforcement learning in language models, 2025.

[11] Jia LI, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. Numinamath. `https://huggingface.co/AI-MO/NuminaMath-CoT`, 2024.

[12] Lifan Yuan, Wendi Li, Huayu Chen, Ganqu Cui, Ning Ding, Kaiyan Zhang, Bowen Zhou, Zhiyuan Liu, and Hao Peng. Free process rewards without process labels. *arXiv preprint arXiv:2412.01981*, 2024.

[13] Michael Luo, Sijun Tan, Roy Huang, Ameen Patel, Alpay Ariyak, Qingyang Wu, Xiaoxiang Shi, Rachel Xin, Colin Cai, Maurice Weber, Ce Zhang, Li Erran Li, Raluca Ada Popa, and Ion Stoica. Deepcoder: A fully open-source 14b coder at o3-mini level. `https://pretty-radio-b75.notion.site/`, 2025. Notion Blog.

[14] Zhangchen Xu, Yang Liu, Yueqin Yin, Mingyuan Zhou, and Radha Poovendran. Kodcode: A diverse, challenging, and verifiable synthetic dataset for coding. 2025.

[15] Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, et al. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks, 2022.

[16] Microsoft. Chemistry-qa. `https://github.com/microsoft/chemistry-qa`, 2021. [GitHub repository].

[17] hivaze. Logic-701: A benchmark dataset for logical reasoning in english and russian. `https://huggingface.co/datasets/hivaze/LOGIC-701`, 2023. Hugging Face Dataset.

[18] NVIDIA. Llama-nemotron-post-training-dataset. https://huggingface.co/datasets/nvidia/Llama-Nemotron-Post-Training-Dataset, 2025. Version 1.1, released on April 8, 2025.

[19] Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, et al. Tülu 3: Pushing frontiers in open language model post-training. 2024.

[20] Beijing Academy of Artificial Intelligence (BAAI). Infinity instruct. *arXiv preprint arXiv:2406.XXXX*, 2024.

[21] Teknium. Openhermes 2.5: An open dataset of synthetic data for generalist llm assistants, 2023.

[22] Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations, 2023.

[23] WizardLM Team. Wizardlm evol-instruct 70k dataset. https://huggingface.co/datasets/WizardLMTeam/WizardLM_evol_instruct_70k, 2023. Accessed: 2025-04-23.

[24] Wing Lian, Bleys Goodson, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". Openorca: An open dataset of gpt augmented flan reasoning traces. https://https://huggingface.co/datasets/Open-Orca/OpenOrca, 2023.

[25] Bleys Goodson. Fine flan: Seqio to parquet so you don't have to. https://huggingface.co/datasets/Open-Orca/FLAN, 2023.

[26] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024.

[27] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

[28] Qwen Team. Qwen2.5: A party of foundation models, September 2024.

[29] Min Li and RLHFlow Team. Decision-tree-reward-llama-3.1-8b. https://huggingface.co/RLHFlow/Decision-Tree-Reward-Llama-3.1-8B, 2025. Interpreting Language Model Preferences Through the Lens of Decision Trees.

[30] Han Zhao, Haotian Wang, Yiping Peng, Sitong Zhao, Xiaoyu Tian, Shuaiting Chen, Yunjie Ji, and Xiangang Li. 1.4 million open-source distilled reasoning dataset to empower large language model training. *arXiv preprint arXiv:2503.19633*, 2025.