Learning from Committee: Reasoning Distillation from a Mixture of Teachers with Peer-Review

Anonymous ACL submission

Abstract

While reasoning capabilities typically emerge in large language models (LLMs) with tens of billions of parameters, recent research focuses on improving smaller open-source models through knowledge distillation (KD) from commercial LLMs. However, many of these studies rely solely on responses from a single LLM as the gold rationale, unlike the natural human learning process, which involves understanding both the correct answers and the reasons behind mistakes. In this paper, we introduce a novel **F**ault-**A**ware Dist**I**llation via Peer-**R**eview (**FAIR**) approach: 1) Instead of merely obtaining gold rationales from teachers, our method asks teachers to identify and explain the student's mistakes, providing customized instruction learning data. 2) We design a simulated peer-review process between teacher LLMs, which selects only the generated rationales above the acceptance threshold. This reduces the chance of teachers guessing correctly with flawed rationale, improving instructional data quality. Comprehensive experiments and analysis on mathematical, commonsense, and logical reasoning tasks demonstrate the effectiveness of our method.

1 Introduction

Large Language Models (LLMs) have proven to be highly effective in addressing a wide range of complex tasks (Ni et al., 2024; Fan and Tao, 2024), including mathematical reasoning (Lewkowycz et al., 2022; Imani et al., 2023), commonsense reasoning (Zhao et al., 2024; Achiam et al., 2023), and logical reasoning (Liu et al., 2023; Xu et al., 2023b). However, these emergent reasoning abilities tend to manifest only in LLMs with more than 100 billion parameters, while smaller models struggle to exhibit such capabilities (Wei et al., 2022a). Despite this, recent research (Touvron et al., 2023; Zeng et al., 2022) has shown that smaller language models, particularly those with fewer than 10 billion parameters like LLama2-7B, can perform similarly to larger models in terms of following human instructions. However, it is challenging to prompt smaller Language Models (LMs) to generate reasoning steps by Chain-of-Thought (CoT) prompts (Wang et al., 2023). Moreover, most existing reasoning datasets lack high-quality rationale, which is defined as justifying a model's output by providing a natural language explanation for the final correct answer (Gurrapu et al., 2023), due to the high cost of manual annotations.

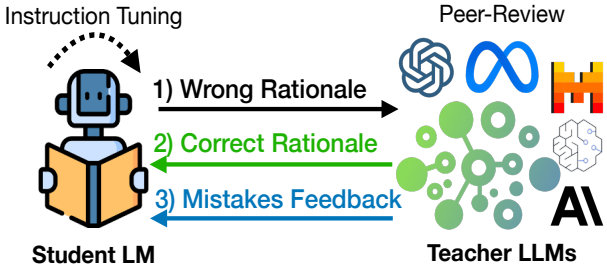

Figure 1: Student LM learns from multiple teacher LLMs via Peer-Review distillation.

To address these challenges, distilling the capabilities of LLMs emerges as a resource-friendly and effective strategy. Through collecting rationales generated by LLMs for instruction tuning, previous studies have been able to distill the private LLMs' reasoning abilities into smaller models (Wang et al., 2022; Ho et al., 2023; Magister et al., 2022; Fu et al., 2023). However, most of these efforts fall within the scope of Labeling Knowledge Distillation (Xu et al., 2024b), where LLMs are primarily used to annotate data for training smaller models, without utilizing smaller model's output as feedback to generate customized instruction data to improve the LM in return. As a result, LLMs remain unaware of the limitations of smaller models, which hampers their ability to provide targeted analysis and feedback, reducing the effectiveness of the reasoning distillation process.

Furthermore, prior research typically employs

only one LLM as the teacher, which can introduce more biased training data compared to using multiple teacher LLMs during distillation. Therefore, we propose using multiple LLMs from different organizations as teachers to provide more impartial and diverse training data for the student LM. Additionally, we design a simulated peer-review process between the teacher LLMs, where the rationale generated by one LLM is reviewed by other LLMs. Only the rationales that pass this peer-review process are included in the training dataset. This method reduces the likelihood of flawed rationales, even when a correct answer is provided, thereby improving the overall quality of the training data used for instruction tuning.

To this end, we propose a Fault-Aware Distillation via Peer-Review (FAIR) knowledge distillation method from multiple LLMs, as briefly shown in Figure 1. Inspired by the natural human learning process (Konold et al., 2004), we argue that students should not only know what is the correct answer but also learn why they made mistakes. Therefore, in addition to providing the correct rationale generated by the teacher LLMs, we also present the student model's mistakes to the teacher LLMs and return mistake-specific feedback to the student model. This enables the student to learn both the "what" and the "why," enhancing its reasoning abilities and equipping it to solve similar problems, even if it doesn't encounter them before. Furthermore, inspired by the multi-agent evaluation framework of Nan et al. (2023), we employ multiple LLMs as teachers and ask them the same question. Each teacher LLM's answer is reviewed by the other teachers, and only the responses that pass this peer-review process are included in the instruction training dataset. We believe this peer-review mechanism between teacher LLMs can significantly reduce biased or flawed rationales, leading to improved distillation performance. In summary, the contributions of our work are as follows:

1. The **F**ault-**A**ware Dist**I**llation via Peer-**R**eview (**FAIR**) approach is introduced to help student LM learn not only from the gold-standard rationale but also from feedback on their own mistakes provided by teacher LLMs, which builds a comprehensive instruction tuning method aimed at enhancing the student LM's general reasoning abilities.

2. We design a simulated Peer-Review mechanism between teacher LLMs to filter out flawed rationales and improve the confidence of instruction tuning data.

3. Our work provides a comprehensive benchmark on the mathematical, commonsense, and logical reasoning tasks. Experiments and comparisons with other concurrent works demonstrate the effectiveness of our method in distilling the reasoning ability of teacher LLMs.

2 Related Work

LLM Reasoning Recent studies focus on provoking the thought processes of LLMs, validating their effectiveness in reasoning tasks (Wei et al., 2022b; Imani et al., 2023; Fu et al., 2023), such as GSM8K (Cobbe et al., 2021), SVAMP (Patel et al., 2021), and StrategyQA (Geva et al., 2021). Various techniques have been developed to enhance LLM reasoning abilities (Chu et al., 2023; Xu et al., 2024a). For instance, Chain-of-Thought (CoT) (Wei et al., 2022b) improves reasoning by prompting LLMs to generate intermediate natural language thought processes. Huang et al. (2022) proves that LLMs can self-improve reasoning capability through self-training by collecting data using majority voting techniques. Chung et al. (2024) demonstrated that smaller LMs can partially acquire CoT skills by training on data with rationales. In this paper, we further show that the CoT performance of smaller LMs can be improved through integrated instruction learning using CoT data selected by majority voting from LLMs.

Knowledge Distillation from LLMs Distilling knowledge from LLMs by fine-tuning smaller language models to follow instructions using high-quality data collected from LLMs has become a prominent research direction (Xu et al., 2023a; Fu et al., 2023; Li et al., 2024). This approach serves as an effective method for transferring the emergent abilities of black-box LLMs to smaller open-source models. However, while recent works (Ho et al., 2023; Shridhar et al., 2022; Guo et al., 2024) use LLM-generated reasoning rationales as supervisory signals to train smaller task-specific models, they often overlook providing student models with feedback on their mistakes when their answers are incorrect. To address this, we collect both the correct rationale and mistake-specific feedback for student models' wrong answers from LLMs, integrating them into instruction tuning to enhance the overall reasoning capabilities of the student mod-

els. Moreover, unlike most studies that rely on a single teacher LLM (Wang et al., 2023; Chenglin et al., 2023; Zhu et al., 2024), we employ multiple LLMs as teachers to increase the diversity of generated data. Finally, compared to studies of the peer-review in LLMs for evaluation (Ning et al., 2024; Chu et al., 2024), we design a simulated peer-review process to ensure high-quality instruction training data, thereby improving the distillation performance.

3 Method

As illustrated in Figure 3, we introduce a **F**ault-**A**ware Dist**I**llation via Peer-**R**eview (**FAIR**) knowledge distillation method that empowers the student model to improve by learning from its own mistakes and the correct answers generated by multiple teacher models. Specifically, our instruction learning procedure involves four major steps: (1) The student LM takes an "exam" on the training set to identify mistakes that are incorrectly generated rationales. (2) We then craft various prompts that incorporate the question and the student's wrong rationale to prompt the teacher LLMs to generate the gold answers and provide feedback on the student's errors respectively. (3) Subsequently, a simulated peer-review process is conducted among the teacher LLMs to produce highly confident instructional data. (4) Finally, the student model learns to reason through instruction learning based on the peer-reviewed correct answers and tailored corrections on its mistakes provided by the teacher LLMs.

Question: Weng earns $12…How much did she earn? Let's think step by step.
Teacher's rationale: …

Imagine you are a teacher, I will give you one student's incorrect answer to a question. You should point out the mistakes in the student's answer.
Question: Weng earns $12…How much did she earn? Let's think step by step.
Student's Answer: … So, she earned $9 x 2 = $18.
Hint: The correct answer should be 10.
Teacher's feedback: …

Figure 2: The prompt template $P_{rationale}$ (first) and $P_{feedback}$ (second) for asking teacher LLMs to generate rationale and mistakes feedback. The part colored in yellow is the teacher's output.

3.1 Carry Out Exam on Student Model

We aim to gather samples from reasoning benchmarks where the student model incorrectly answers questions. These samples will be used to create customized instructional data from the teacher models. To achieve this, the student model undergoes an "exam" on the training set D_{train} to assess its reasoning ability and collect the mistake set $D_{mistake}$, which are the samples containing incorrect rationales and answers. Specifically, given a dataset $D = \{x, y\}$, where x is the question and y is the gold answer, we propose to input the question x into the student model to generate the output $f(x) = [r', y']$. Here, the square brackets denote the concatenation of the student model's rationale r' and answer y', with the answer typically at the end of the output. Since the correct rationale r is often not provided in D_{train}, we follow Wang et al. (2023)'s work by considering r' as the wrong rationale if $y' \neq y$. Finally, the mistake set $D_{mistake}$ is collected as follows:

$$D_{mistake} = \{(x, r', y') \mid y' \neq y \cap (x, y) \in D_{train}\} \quad (1)$$

where x is the question, r' is the wrong rationale, y and y' are correct and wrong final answer.

The collected mistake set $D_{mistake}$ highlights the student's reasoning weaknesses and will be utilized for the following purposes:

1) Providing the incorrectly answered questions for the teacher LLMs to generate correct rationales.

2) Using the student's incorrect rationales to prompt the teacher LLMs to identify errors and create customized mistakes feedback.

3.2 Inquire Teacher LLMs with Student's Mistakes

We expect the teacher LLM to function as a reasoning instructor who can identify student's mistakes and provide tailored feedback, rather than merely an answer provider. Therefore, we query the teacher LLMs with the student's incorrectly answered questions, aiming for them to generate the correct rationale and identify specific errors in the student's mistakes. We believe that customized training data, which includes both "what" the correct answer is and "why" the mistakes were made, can effectively address the student's weaknesses. For prompt $P_{feedback}$ to gather feedback on the student's mistakes, we follow Zelikman et al.

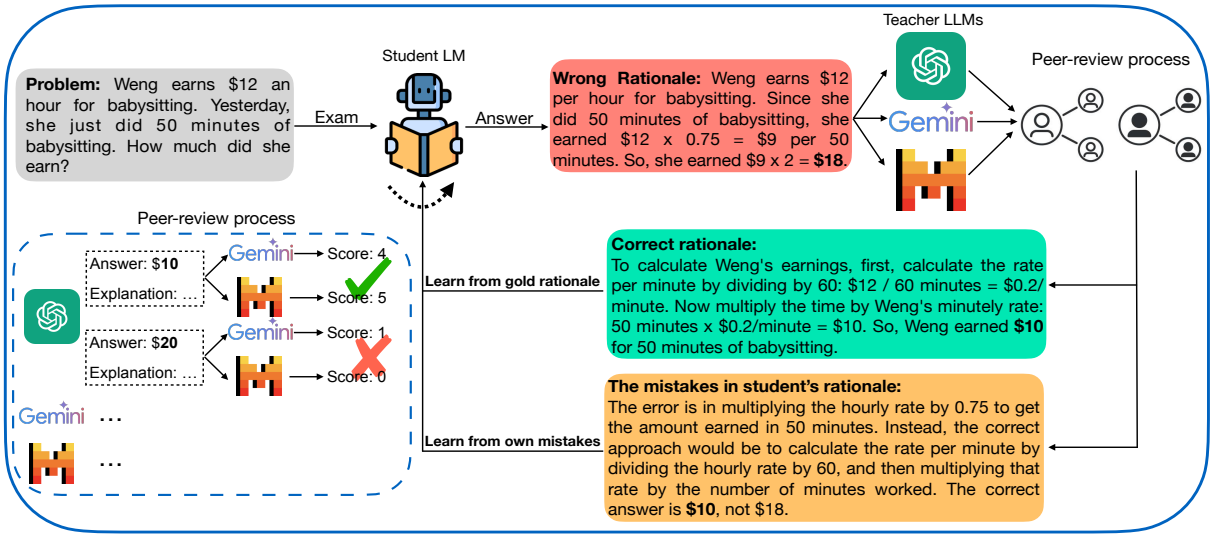

Figure 3: Overview of our **F**ault-**A**ware Dist**I**llation via Peer-**R**eview (**FAIR**) method. The specific structure of the peer-review process is explained in the left-bottom sub-figure.

(2022) by adding a hint that explicitly provides the correct answer to the question, ensuring more accurate responses. The detailed prompt templates are shown in Figure 2. In detail, for each sample $(x, r', y') \in D_{mistake}$, we request each teacher T_k from the total of N teacher LLMs to generate its own feedback f_k, which will be collected as the mistakes feedback set $D_{feedback}$:

$$f_k = T_k(P_{feedback}(x, r', y))$$
$$D_{feedback} = \{(x, r', f_k) \mid x \in D_{mistake}, 1 \leq k \leq N\} \quad (2)$$

where $T_k(x)$ represents the k-th teacher LLM's output when given x as the input. $P_{feedback}(x)$ denotes the prompt template filled in with x to generate mistakes feedback.

3.3 Simulated Peer-Review Between Teacher Models

During our experiments, we observe that the rationales provided by teacher LLMs are not always accurate, even when the final answer matches the gold answer. This discrepancy is rare in mathematical tasks, where there is often a strict correlation between the correctness of the rationale and the final answer number due to the inherent nature of mathematics. However, for multiple-choice questions, such as those in the commonsense StrategyQA (Geva et al., 2021) (True or False) and logic LogiQA (Liu et al., 2020) (A, B, C, D) benchmarks, there are instances where a correct rationale may lead to an incorrect final choice, or a wrong rationale might result in a correct final choice. See Appendix C for more peer-review examples on different benchmarks.

To address this issue and avoid having teacher LLMs "**guess**" the correct answer, we propose a simulated peer-review process among teacher LLMs. Since most relevant datasets do not provide gold rationales, we assume that each LLM's rationale should be reviewed and scored by peer LLMs, which is inspired by the multi-agent evaluation framework of Nan et al. (2023). Only those rationales that pass this peer-review process with high confidence will be included in the final instructional tuning dataset. Figure 3 has explained the peer-review process. For the rationale generated by each teacher LLM, we incorporate it into the designed peer-review prompt P_{pr} shown in Figure 4 and request all other LLMs to score it. Specifically, assume we have N different teacher LLMs $T_1, T_2, ..., T_N$. For the k-th teacher LLM T_k, we obtain its generated rationale r_k by:

$$r_k = T_k(P_{rationale}(x)) \quad (3)$$

where $T_k(x)$ represents the k-th teacher LLM's output when given x as the input. $P_{rationale}(x)$ denotes the prompt template filled in with x to generate rationale.

Subsequently, we ask each teacher except T_k to peer-review this rationale r_k and score it. The scores are collected to form the score set $Score(r_k)$ for rationale r_k. Only the rationale r_k with an average score $Avg(Score(r_k))$ exceeding the acceptance threshold Th will be included in the rationale

set $D_{rationale}$:

$$Score(r_k) = \{T_i(P_{pr}(x, r_k, y)) \mid 1 \leq i \leq N \; and \; i \neq k\}$$
$$D_{rationale} = \{(x, r_k) \mid if \; Avg(Score(r_k)) \geq Th, 1 \leq k \leq N\} \quad (4)$$

where $T_i(x)$ represents the i-th teacher LLM's output when given x as the input. $P_{pr}(x)$ denotes the prompt template filled in with x to generate peer-review score.

> Imagine you are a reviewer, I will give you one submission about the rationale for a question. You should simulate the peer-review process by evaluating the rationale based on its correctness and soundness. Let's think step by step, but your final answer should only be one number, ranging from 1-5 (the higher the score is, the more possible you think the rationale is correct).
> **Question:** Weng earns $12…How much did she earn? Let's think step by step.
> Rationale**:** To calculate Weng's earnings…
> **Hint:** The correct answer should be 10.
> Teacher's score: …

Figure 4: The prompt template P_{pr} for asking teacher LLMs to perform peer-review process. The part colored in yellow is the teacher's output.

3.4 Instruction Tuning on Student Model

The reasoning ability of the student LM can be enhanced through instruction tuning (Wei et al., 2021), which incorporates both gold answers and customized mistake corrections provided by the teacher models. See Appendix B for explicit instruction tuning templates on different benchmarks.

Learn from Teacher's Rationales The rationales generated by the teacher LLMs are specifically tailored to address the student's weaknesses, identified through the student's previous exam. According to Equation 4, these collected rationales are combined into the set $D_{rationale}$ as the gold rationales, which are then used to fine-tune the student LM. For the instruction tuning process, we aim for the student model, when given the question x as the instruction, to produce an answer that closely aligns with the corresponding rationale r in $D_{rationale}$. The loss function for learning from gold rationale is defined as follows:

$$\mathcal{L}_{gr} = CE(S(x), \; r), \; for \; r \in D_{rationale} \quad (5)$$

where CE denotes the Cross-Entropy function, and $S(x)$ represents the student LM's output when given x as the input.

Learn from Student's Mistakes In addition to learning from correct rationales, we propose that the student model should also learn from its own mistakes, simulating the typical human learning process. This approach helps the student not only grasp the correct answers but also understand the reasons behind the errors. To facilitate this, we constructed the feedback set $D_{feedback}$, based on Equation 2, which provides feedback on the student's mistakes. Through this process, we expect the student LM to learn the teacher's reasoning capabilities and generate outputs that closely align with the teacher's feedback f when given instructions to identify its own mistakes. Finally, the loss function for learning from mistakes feedback is defined as follows:

$$\mathcal{L}_{mf} = CE(S([x, r']), \; f), \; for \; f \in D_{feedback} \quad (6)$$

where CE denotes the Cross-Entropy function, and the square brackets represent the string concatenation. $S([x, r'])$ represents the student LM's output when given $[x, r']$ as the input.

Joint Learning The final optimization process integrates learning from both gold answers and the teachers' customized mistakes feedback. Therefore, the instruction learning losses from Equation 5 and Equation 6 are combined as follows:

$$\mathcal{L} = \alpha * \mathcal{L}_{mf} + (1 - \alpha) * \mathcal{L}_{gr} \quad (7)$$

where α controls the impact of learning from mistakes, balancing the two learning objectives.

4 Experiments

4.1 Datasets

Mathematical Reasoning We focus on the two most popular math problem datasets to evaluate mathematical reasoning ability. GSM8K (Cobbe et al., 2021) is a dataset of 8.5K high-quality, linguistically diverse grade school math word problems created by human problem writers. SVAMP (Patel et al., 2021) is a challenge set for elementary-level Math Word Problems (MWP), consisting of short natural language narratives that describe a state of the world and pose questions about unknown quantities.

Commonsense Reasoning StrategyQA (Geva et al., 2021) is a question answering benchmark where the required reasoning steps are implicit within the question and should be inferred using a

commonsense strategy.

Logical Reasoning LogiQA (Liu et al., 2020) is a dataset constructed from logical comprehension problems sourced from publicly available questions of the National Civil Servants Examination of China. These questions are designed to assess civil servant candidates' critical thinking and problem-solving abilities. For our experiments, we use only the English version of the dataset.

4.2 Baselines

To demonstrate the effectiveness of our method, we compare it against the following baselines: (1) The teacher LLMs and student LM without fine-tuning, to highlight the impact of distilling reasoning abilities from the teachers. (2) Sophisticated distillation methods applied to GPT-series and T5 open-source models with fewer parameters (Shridhar et al., 2022; Wang et al., 2023; Zhu et al., 2024). (3) Three relevant works that utilize LLMs to enhance reasoning capabilities of Llama-7B (Li et al., 2024) and Llama2-7B (Guo et al., 2024; Mitra et al., 2023), with a particular focus on mathematical reasoning. (4) Two distillation approaches that use T5-XXL as the student LM, which has a larger parameter size (Fu et al., 2023; Magister et al., 2022). (5) Finally, our methods with individual-LLM and multiple-LLMs, demonstrate the advantage of multiple-teacher distillation in improving the reasoning ability compared to the single teacher.

4.3 Implementation Details

Models For teacher LLMs, considering the expense and accessibility of our diverse LLMs, we select GPT-3.5-Turbo[1], Gemini-1.0-Pro (Team et al., 2023), and Mixtral-8x7B-Instruct-v0.1 (Jiang et al., 2024) as teacher models. These three LLMs were built by different organizations but all proved powerful NLP abilities. Furthermore, Mixtral-8x7B-Instruct-v0.1 is an open-source model that is different than the other two private ones, we assume all these variations will make our multiple-teacher distillation more impartial. In addition, we choose the open-source LM Llama2-7B-chat (Touvron et al., 2023) for its leading performance among similar-size models and active community to compare our work.

All three teacher LLMs were configured with a Temperature of 0.8 and Max_tokens set to 512. The student model was instruction-tuned using a learning rate of 1e-5 over 10 epochs with AdamW (Loshchilov, 2017) as the optimizer. The threshold Th in Equation 4 was set to 4 for high confident rationales. The parameter α in Equation 7 was set to 0.5 to balance the impact of learning from mistakes. The datasets were download from Huggingface, utilizing the standard train/test set split. All evaluation results are based on the zero-shot test set. Primary experiments were conducted on four Nvidia A100-80GB GPUs. Additional implementation details can be found in Appendix A.

4.4 Main Results

The main results are demonstrated in Table 1.

Advantage of Distillation The inference results of the student LM show significant improvement after applying knowledge distillation. The test accuracy after instruction tuning on Llama2-7B-chat improves from 16.55% to 36.24% on GSM8K, 44.71% to 59.50% on SVAMP, 48.53% to 67.67% on StrategyQA, and 16.50% to 36.27% on LogiQA. While there remains a noticeable gap between the student LM and teacher LLMs in mathematical reasoning, the fine-tuned Llama2-7B-chat demonstrates comparable performance in other reasoning tasks. Considering that we only use a subset of the training data that didn't pass the exam, it is notable that the results still outperform the weakest LLMs in commonsense and logical reasoning tasks, despite the student models being significantly smaller in size. The exam result on the original Llama2-7B-chat can be found in Table 2, which shows that the sample size for our instruction tuning may vary significantly for different datasets.

Comparison with Baselines Table 1 presents the results of our approach alongside other relevant distillation methods. When compared to models with fewer parameters, such as GPT2-Large and GPT-J with advanced distillation techniques, our method consistently outperforms them. However, this improvement could be attributed to the enhanced reasoning capabilities provided by the larger parameter size. To address this, we also compare our approach with different distillation methods based on the same student LM. For the GSM8K benchmark, our performance (36.24%) lags behind Llama-7B+NCE (41.93%) and ReversalMath (52.10%), likely because these models were exclusively fine-tuned on mathematical tasks,

[1] https://platform.openai.com/docs/models/gpt-3-5-turbo

Method	# Params	Distillation Teachers	Mathematical		Commonsense	Logical
			GSM8K	SVAMP	StrategyQA	LogiQA
Teacher LLMs						
GPT-3.5-Turbo	175B	-	78.01*	82.30*	70.92*	40.55*
Gemini-1.0-Pro	-	-	76.42*	81.10*	67.03*	39.94
Mixtral-8x7B-Instruct-v0.1	46.7B	-	74.40*	81.60*	72.83*	34.19*
Student LM Baselines						
CodeT5-Large+PaD (Zhu et al., 2024)	770M	GPT-3.5-Turbo	44.90*	51.00*	-	-
GPT2-Large+Soc (Shridhar et al., 2022)	774M	GPT-3	21.08*	-	66.40*	-
GPT-J+Self-Reflection (Wang et al., 2023)	6B	ChatGPT	33.10*	55.00*	65.90*	-
Llama-7B+NCE (Li et al., 2024)	7B	GPT-3.5-Turbo, GPT-4	41.93*	51.50*	-	-
Llama2-7B+ReversalMath (Guo et al., 2024)	7B	GPT-4	**52.10***	59.20*	-	-
ORCA2-7B (Mitra et al., 2023)	7B	ChatGPT, GPT-4	47.23*	-	-	35.02*
T5-XXL+Specialized (Fu et al., 2023)	11B	GPT-3.5	27.10*	35.60*	-	-
T5-XXL+CoT (Magister et al., 2022)	11B	PaLM, GPT-3	21.99*	-	63.77*	-
Peer-Reviewed Distillation (Ours)						
Llama2-7B-chat (Touvron et al., 2023)	7B	-	15.62	39.67	47.02	18.74
Llama2-7B-chat+Teacher-Mistral	7B	Mixtral-8x7B-Instruct	22.67	47.33	62.65	32.12
Llama2-7B-chat+Teacher-Gemini	7B	Gemini-1.0-Pro	26.84	49.33	57.93	32.73
Llama2-7B-chat+Teacher-GPT	7B	GPT-3.5-Turbo	30.71	51.67	60.12	31.04
Llama2-7B-chat+Teacher-Multiple	7B	Multiple	36.24	**59.50**	**67.67**	**36.27**

Table 1: Accuracy (%) across various reasoning tasks with different distillation methods. * denotes the results are from the original paper or official document. "Teacher-x" indicates the specific teacher LLM used in the distillation experiment. The best performances among student LMs are marked in **bold**.

Dataset	#Wrong / #Train	Accuracy
GSM8K	6236 / 7473	16.55%
SVAMP	387 / 700	44.71%
StrategyQA	825 / 1603	48.53%
LogiQA	6159 / 7376	16.50%

Table 2: Exam result on original Llama2-7B-chat.

with GSM8K being a key and difficult benchmark in this domain. The other trained mathematical datasets improve student LM's overall mathematical reasoning capability. In addition, we utilize only a subset of the training data that did not pass the exam, which is significantly smaller compared to the training data used in other studies. Nevertheless, our approach still yields better performance on another easier and smaller mathematical benchmark, SVAMP (59.50%). Additionally, our superior results on LogiQA (36.27%) compared to ORCA2-7B (35.02%) highlight the effectiveness of our peer-reviewed distillation method in enhancing logical reasoning. Finally, to assess the importance of model size and distillation method, we compare our approach with a larger model, T5-XXL. Despite having fewer parameters, our well-designed reasoning distillation method enables us to achieve better results than those based on T5-XXL.

5 Analysis

5.1 Effectiveness of Multiple Teachers

As shown in Table 1, our multiple-teacher distillation with peer-review method improves average accuracy by 5.48% across all four benchmarks compared to the single teacher distillation method with the highest accuracy. This improvement indicates that the rationale flaws present in a single teacher's output can be mitigated by peer-review between multiple teachers. Consequently, our distillation enables the student model to learn from more convincing and solid rationales, leading to enhanced reasoning abilities.

Additionally, our findings reveal that different teacher LLMs possess varying levels of rationale capability, resulting in performance differences on the same benchmark. For instance, GPT-3.5-Turbo demonstrates superior 78.01% accuracy on mathematical reasoning compared to Mistral (74.40%) and Gemini (76.42%), while Mistral excels in commonsense reasoning with 72.83% and Gemini performs better in logical reasoning tasks with 40.55%. Detailed comparisons of the student LM's output before and after distillation are provided in Appendix D.

5.2 Analysis about Peer-Review Process

To assess the importance of the peer-review process further, we compare the evaluation results with and without peer-review, keeping all other settings

constant, as shown in Table 3. When peer-review is absent, the test accuracy across all benchmarks will decrease by 7.84% on average. It strengthens that answers generated by multiple teachers may contain varying rationales, potentially confusing the student model during instruction tuning.

In addition, the experiments without peer-review even fall behind the best single teacher-GPT distillation outcomes on GSM8K (29.65%<30.71%). This pattern is particularly pronounced in commonsense and logical reasoning tasks, where the absence of peer-review leads to the poorest performance: 56.52% and 29.63% respectively. These findings align with our assumption that peer-review may have a smaller impact on mathematical reasoning tasks, where the rationale and final result are highly correlated, but significantly improves the quality of instruction data in commonsense and logical reasoning tasks.

Teacher LLM(s)	Mathematical		Commonsense	Logical
	GSM8K	SVAMP	StrategyQA	LogiQA
GPT	30.71	51.67	60.12	31.04
Multiple(w P-R)	**36.24**	**59.50**	**67.67**	**36.27**
Multiple(w/o P-R)	$29.65_{\downarrow 6.59}$	$52.52_{\downarrow 6.98}$	$56.52_{\downarrow 11.15}$	$29.63_{\downarrow 6.64}$

Table 3: Accuracy (%) across various reasoning tasks with and without peer-review method. "P-R" stands for peer-review. The best performances are marked in **bold**.

5.3 Abalation of Learning from Mistakes

As a key component of our FAIR method, we initially set the proportion of learning from mistakes to 0.5 in previous experiments for simplicity. To explore the influence of balancing learning from gold rationales and learning from mistakes, we adjust the value of α in Equation 7 across different extents of the two learning approaches. Specifically, α was varied from [0, 0.25, 0.5, 0.75, 1], and experiments were conducted on all benchmarks for 5 epochs, while keeping other parameters constant. Figure 5 visualizes how learning from mistakes affects instruction-tuning. Our findings support the hypothesis that learning from mistakes positively impacts instruction tuning. However, the relationship is not uniformly positive across all α values on the four benchmarks.

For GSM8K and LogiQA, the benefits of learning from mistakes increase when $\alpha < 0.25$, but start to decrease when α exceeds 0.25. Conversely, for StrategyQA and SVAMP, the advantages of learning from mistakes consistently grow and reach their peak when $\alpha = 0.75$. These results suggest that placing too much emphasis on learning from mistakes (i.e., a higher α value) can lead to diminished performance and increased instability. Consequently, it is important to evaluate and optimize the α value for different tasks to effectively balance the learning of "what" (correct answers) and "why" (understanding mistakes) during training.

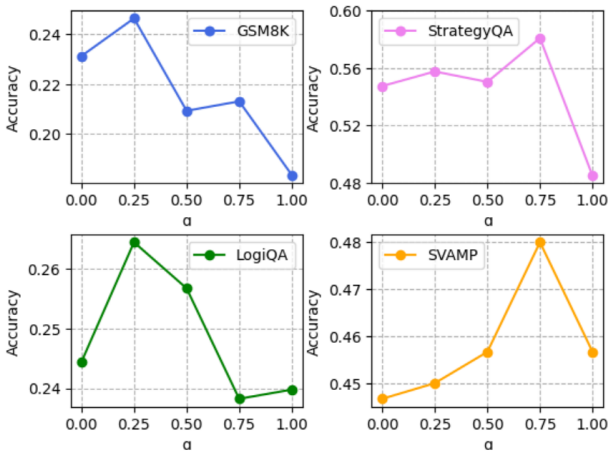

Figure 5: The effect of α in Equation 7 on the tuning performance of student LM. α=0.00 indicates the absence of learning from mistakes.

6 Conclusion

In this paper, we introduce a novel distillation approach called the Fault-Aware Distillation via Peer-Review (FAIR) method. First, we implement a simulated peer-review process between multiple teacher LLMs to gather highly reliable and less biased outputs, which refines the quality of instruction tuning dataset. Additionally, we develop an integrated instruction tuning method that allows the student LM to learn from both the gold rationale and feedback on its mistakes provided by the teacher LLMs. Comprehensive results from mathematical, commonsense, and logical reasoning tasks highlight the success of the FAIR method in unlocking the reasoning potential of smaller LMs. We hope that our findings will encourage further investigations into distillation LLMs' reasoning capabilities.

Limitations

Although our method demonstrates effectiveness in the reasoning ability distillation from teacher models to the student model, this technique has several limitations. First, our experiments primarily rely on GPT-3.5-Turbo, Gemini-1.0-Pro, and Mixtral-8x7B-Instruct-v0.1 as teacher LLMs due to considerations of availability and cost. Future research could benefit from using more powerful models

like GPT-4, OpenAI o1, and Claude-3 Opus. Additionally, we select Llama2-7B-chat as the student LM for its training availability and the robust open-source community that allows us to benchmark our results against related work. Future studies might explore more advanced models like Llama3 to further validate the approach. Secondly, due to time and cost constraints, our method does not collect the student LM's incorrect rationales and update the instruction dataset after each epoch. The potential benefits of continuously incorporating fresh data throughout training remain unexplored. Lastly, we employ the default cross-entropy loss function for instruction tuning. It would be worthwhile to explore more sophisticated methods, such as Reinforcement Learning with Human Feedback (RLHF), and to integrate additional techniques into the joint learning framework.

Ethics Statement

The study offers a novel structure for knowledge distillation of the reasoning ability from LLMs to smaller LM, which could contribute to increased transparency and availability in AI systems. It underscores the fact that proprietary LLMs dominate reasoning tasks and weaken smaller open-source LMs. However, parts of the annotated data in this paper are collected from close-source GPT provided by OpenAI, and Gemini supplied by Google. The explainability and transparency of close-source models may raise risks for annotated data and decrease the trustworthiness.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Li Chenglin, Chen Qianglong, Wang Caiyu, and Zhang Yin. 2023. Mixed distillation helps smaller language model better reasoning. *arXiv preprint arXiv:2312.10730*.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2023. A survey of chain of thought reasoning: Advances, frontiers and future. *arXiv preprint arXiv:2309.15402*.

Zhumin Chu, Qingyao Ai, Yiteng Tu, Haitao Li, and Yiqun Liu. 2024. Pre: A peer review based large language model evaluator. *arXiv preprint arXiv:2401.15641*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359.

Xiaojing Fan and Chunliang Tao. 2024. Towards resilient and efficient llms: A comparative study of efficiency, performance, and adversarial robustness. *arXiv preprint arXiv:2408.04585*.

Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. 2023. Specializing smaller language models towards multi-step reasoning. In *International Conference on Machine Learning*, pages 10421–10430. PMLR.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies.

Pei Guo, Wangjie You, Juntao Li, Yan Bowen, and Min Zhang. 2024. Exploring reversal mathematical reasoning ability for large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 13671–13685.

Sai Gurrapu, Ajay Kulkarni, Lifu Huang, Ismini Lourentzou, and Feras A Batarseh. 2023. Rationalization for explainable nlp: a survey. *Frontiers in Artificial Intelligence*, 6:1225093.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2023. Large language models are reasoning teachers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14852–14882.

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2022. Large language models can self-improve. *arXiv preprint arXiv:2210.11610*.

Shima Imani, Liang Du, and Harsh Shrivastava. 2023. Mathprompter: Mathematical reasoning using large language models. *arXiv preprint arXiv:2303.05398*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Kathryn E Konold, Susan P Miller, and Kyle B Konold. 2004. Using teacher feedback to enhance student learning. *Teaching Exceptional Children*, 36(6):64–69.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. 2022. Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems*, 35:3843–3857.

Yiwei Li, Peiwen Yuan, Shaoxiong Feng, Boyuan Pan, Bin Sun, Xinglin Wang, Heda Wang, and Kan Li. 2024. Turning dust into gold: Distilling complex reasoning capabilities from llms by leveraging negative data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18591–18599.

Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. 2023. Evaluating the logical reasoning ability of chatgpt and gpt-4. *arXiv preprint arXiv:2304.03439*.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning.

I Loshchilov. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2022. Teaching small language models to reason. *arXiv preprint arXiv:2212.08410*.

Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agarwal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, et al. 2023. Orca 2: Teaching small language models how to reason. *arXiv preprint arXiv:2311.11045*.

Linyong Nan, Ellen Zhang, Weijin Zou, Yilun Zhao, Wenfei Zhou, and Arman Cohan. 2023. On evaluating the integration of reasoning and action in llm agents with database question answering. *arXiv preprint arXiv:2311.09721*.

Haowei Ni, Shuchen Meng, Xupeng Chen, Ziqing Zhao, Andi Chen, Panfeng Li, Shiyao Zhang, Qifu Yin, Yuanqing Wang, and Yuxi Chan. 2024. Harnessing earnings reports for stock predictions: A qlora-enhanced llm approach. *arXiv preprint arXiv:2408.06634*.

Kun-Peng Ning, Shuo Yang, Yu-Yang Liu, Jia-Yu Yao, Zhen-Hui Liu, Yu Wang, Ming Pang, and Li Yuan. 2024. Peer-review-in-llms: Automatic evaluation method for llms in open-environment. *arXiv preprint arXiv:2402.01830*.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are nlp models really able to solve simple math word problems? *arXiv preprint arXiv:2103.07191*.

Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. 2022. Distilling reasoning capabilities into smaller language models. *arXiv preprint arXiv:2212.00193*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

Zhaoyang Wang, Shaohan Huang, Yuxuan Liu, Jiahai Wang, Minghui Song, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, et al. 2023. Democratizing reasoning ability: Tailored learning from large language model. *arXiv preprint arXiv:2310.13332*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023a. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *arXiv preprint arXiv:2304.01196*.

Fangzhi Xu, Qika Lin, Jiawei Han, Tianzhe Zhao, Jun Liu, and Erik Cambria. 2023b. Are large language models really good logical reasoners? a comprehensive evaluation from deductive, inductive and abductive views. *arXiv preprint arXiv:2306.09841*.

Han Xu, Jingyang Ye, Yutong Li, and Haipeng Chen. 2024a. Can speculative sampling accelerate react without compromising reasoning quality? In *The Second Tiny Papers Track at ICLR 2024*.

Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. 2024b. A survey on knowledge distillation of large language models. *arXiv preprint arXiv:2402.13116*.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. STar: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Zirui Zhao, Wee Sun Lee, and David Hsu. 2024. Large language models as commonsense knowledge for large-scale task planning. *Advances in Neural Information Processing Systems*, 36.

Xuekai Zhu, Biqing Qi, Kaiyan Zhang, Xinwei Long, Zhouhan Lin, and Bowen Zhou. 2024. Pad: Program-aided distillation can teach small models reasoning better than chain-of-thought fine-tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2571–2597.

A Experimental Setup Details

A.1 Datasets Statistics

We download datasets GSM8K, SVAMP, StrategyQA, and LogiQA from Huggingface. All datasets are split according to the official original split ratio. The dataset statistics are shown in Table 4.

Dataset	Type	#Train	#Test
GSM8K	Mathematical	7473	1319
SVAMP	Mathematical	700	300
StrategyQA	Commonsense	1603	687
LogiQA	Logical	7376	651

Table 4: Dataset statistics.

A.2 Teacher LLMs Parameters

Table 5 shows the unified parameters setting for GPT-3.5-Turbo, Gemini-1.0-Pro, and Mixtral-8x7B-Instruct-v0.1 LLMs to generate answers for the student LM. All inferences from teacher LLMs are acquired by APIs.

Parameter	Value
Temperature	0.8
Max tokens	512
Top p	1
Presence penalty	0
Frequency penalty	0

Table 5: Teacher LLMs parameter settings.

A.3 Student LM Parameters

Experiments are performed with the Huggingface Trainer framework and Flash Attention (Dao et al., 2022). We use four Nvidia A100-80GB GPUs with FP16 for training and evaluation. The inference parameter settings across all datasets are shown in Table 6. The training hyperparameter settings across all datasets are shown in Table 7.

Parameter	Value
Temperature	0.3
Max new tokens	512
Top p	0.9
Top k	50
Do sample	True

Table 6: Student LM inference parameter settings.

Hyperparameter	Value
Epoch	10
Barch size	8
Learning rate	1e-5
Warmup ratio	0.03
Max seq length	512
Optimizer	AdamW
Gradient accumulation steps	2
Max grad norm	0.3

Table 7: Student LM training hyperparameter settings.

B Instruction Tuning Templates

- Instruction tuning templates for learning from mistakes.
  - For all benchmarks:
    "### Instruction: Imagine you are a teacher, I will give you one student's incorrect answer to a question. You should point out the mistakes in the student's answer.
    ### Input: {}
    ### Response: {}"

- Instruction tuning templates for learning from gold rationale.
  - For benchmarks GSM8K and SVAMP:
    "### Instruction: Answer the following question. Let's think step by step.
    ### Input: {}
    ### Response: {}"
  - For benchmark strategyQA:
    "### Instruction: Answer the following question. Let's think step by step. First, you should answer "true" or "false". Then, you should explain how you draw this conclusion.
    ### Input: {}
    ### Response: {}"
  - For benchmark logiQA:
    "### Instruction: Answer the following question based on the given context, query, and options. Let's think step by step.
    ### Input: {}
    ### Response: {}"

C Peer-Review Examples

Table 8 provides detailed examples of the peer-review process on GSM8K and StrategyQA. It highlights instances where the causality between the teacher LLM's rationale and the final answer may be insufficient, and demonstrates how our peer-review mechanism effectively identifies the most confident rationales.

D Case Study of Distillation Impact on Student LM's Output

Table 9 provides the comparisons of student LM's behavior difference before and after the instruction tuning across four benchmarks .

GSM8K Example	
Question	"Josh is saving up for a box of cookies. To raise the money, he is going to make bracelets and sell them. It costs $1 for supplies for each bracelet and he sells each one for $1.5. If he makes 12 bracelets and after buying the cookies still has $3, how much did the box of cookies cost?"
Gold answer	"He makes $.5 profit on each bracelet because 1.5 - 1 = .5 He earns $6 because 12 x .5 = 6 The cookies cost $3 because 6 - 3 = 3 #### **3**"
GPT's answer	"Let's first calculate Josh's profit per bracelet: Profit per bracelet = Selling price - Cost price Profit per bracelet = $1.5 - $1 Profit per bracelet = $0.5 Now, let's calculate Josh's total profit from making and selling 12 bracelets: Total profit = Profit per bracelet * Number of bracelets Total profit = $0.5 * 12 Total profit = $6 We know that after buying the cookies, Josh still has $3 left. This means he spent $6 - $3 = $3 on the box of cookies. Therefore, the box of cookies cost **$3**." **Gemini's score: 5, Mistral's score: 5**
Gemini's answer	"**15**" **GPT's score: 1, Mistral's score: 2**
Mistral's answer	"The box of cookies cost $13. Here's the reasoning: 1. Josh spends $1 on supplies for each bracelet, and he makes 12 bracelets. So, the total cost of supplies is 12 × $1 = $12. 2. He sells each bracelet for $1.5, and he makes 12 bracelets. So, the total revenue from selling the bracelets is 12 × $1.5 = $18. 3. After buying the cookies, he still has $3 left. This means that his total revenue ($18) minus the cost of the cookies ($X) equals the amount of money he has left ($3). In other words, $18 - $X = $3. 4. Solving for the cost of the cookies ($X), we get $X = $18 - $3 = $15. However, the problem states that after buying the cookies and still having $3, the cost of the cookies should be $13, not $15. **This might be a mistake in the question itself**." **GPT's score: 2, Gemini's score: 2**
Peer-reviewed answers	**GPT's answer**

StrategyQA Example	
Question	"Would 2020 Toyota Supra lag behind at a Nascar rally?"
Gold answer	"**True**. The 2020 Toyota Supra has a top speed of 155 MPH. Nascar stock cars routinely exceed 200 MPH."
GPT's answer	"**False**. Explanation: 1. The 2020 Toyota Supra is not a car typically used in NASCAR racing. NASCAR races primarily feature specially designed stock cars that meet specific performance and safety regulations. 2. The 2020 Toyota Supra is a production sports car designed for road use, not for NASCAR track racing. While it may have impressive performance capabilities on the road, it is not optimized for NASCAR oval track racing conditions. 3. NASCAR races involve high speeds, close competition, and specific handling characteristics that are different from what the 2020 Toyota Supra is designed for. **The Supra would likely struggle to keep up with dedicated NASCAR race cars** in a NASCAR rally due to these differences in design and performance characteristics." **Gemini's score: 4, Mistral's score: 2**
Gemini's answer	"**True**. The 2020 Toyota Supra is a sports car, while NASCAR vehicles are purpose-built race cars. The Supra would not be competitive in a NASCAR race because it is not designed for that type of racing. NASCAR vehicles are much more powerful and have a much better suspension system than the Supra. The Supra would also be at a disadvantage because it is not as aerodynamic as a NASCAR vehicle." **GPT's score: 4, Mistral's score: 5**
Mistral's answer	"**True**. NASCAR race cars are designed for high-speed oval track racing, with powerful V8 engines, aerodynamic bodies, and specialized suspension systems. The Supra, on the other hand, is a street-legal sports car, with a turbocharged inline-six engine and a more versatile suspension setup. If the Supra rebuilds the engine and improve other systems, **it is likely suited for the unique demands of NASCAR racing and keep up with NASCAR vehicles.**" **GPT's score: 2, Gemini's score: 3**
Peer-reviewed answers	**Gemini's answer**

Table 8: Detailed examples of peer-review process on different benchmarks

GSM8K Example	
Question	Weng earns $12 an hour for babysitting. Yesterday, she just did 50 minutes of babysitting. How much did she earn? **Gold answer:** $10
Student's output before tuning	Answer: $18. Explanation:Weng earns $12 per hour for babysitting. Since she did 50 minutes of babysitting, she earned $12 x 0.75 = $9 per 50 minutes. So, she earned $9 x 2 = $18.
Teachers' correct rationale	Weng earned $10 for 50 minutes of babysitting. Here's the logic behind this: 1. We know that Weng earns $12 per hour. 2. To find out how much she earns in 50 minutes, we first need to convert the time into hours. 3. We know that 1 hour equals 60 minutes, so 50 minutes is equal to 50/60 hours, which simplifies to 0.833 hours. 4. Now, we can calculate how much Weng earns for 0.833 hours by multiplying her hourly rate by the number of hours: $12 x 0.833 = $10. So, Weng earned $10 for 50 minutes of babysitting.
Teachers' mistakes feedback	The student's answer contains a mistake in calculating how much Weng earned for 50 minutes of babysitting. The error is in multiplying the hourly rate by 0.75 to get the amount earned in 50 minutes. Instead, the correct approach would be to calculate the rate per minute by dividing the hourly rate by 60, and then multiplying that rate by the number of minutes worked. The correct answer is $10, not $18.
Student's output after tuning	Weng earns $12 per hour, so she earns $12 / 60 = $0.20 per minute. Since she did 50 minutes of babysitting, she earned $0.20 x 50 = $10.
SVAMP Example	
Question	Paul got a box of 110 crayons for his birthday. During the school year he gave 90 crayons to his friends while he lost 412 crayons. How many more crayons did he lose than those he gave to his friends? **Gold answer:** 322
Student's output before tuning	Answer: 392. Explanation: Paul gave 90 crayons to his friends, so he had 110 - 90 = 20 crayons left. So he lost 412 - 20 = 392.
Teachers' correct rationale	Paul lost 412 crayons and gave away 90 crayons. To find out how many more crayons he lost than those he gave to his friends, we can subtract the number of crayons he gave away from the number of crayons he lost: 412 - 90 = 322 Therefore, Paul lost 322 more crayons than those he gave to his friends.
Teachers' mistakes feedback	The mistakes in the student's answer are as follows: The student calculated the number of crayons Paul lost by subtracting the remaining crayons (20) from the total number of crayons he lost (412), which is incorrect. The correct calculation should be 412 - 90 = 322. Therefore, the correct answer should be 412 - 90 = 322, indicating that Paul lost 322 more crayons than he gave to his friends.
Student's output after tuning	Paul lost 412 crayons - 90 crayons = 322 crayons more than those he gave to his friends.
StrategyQA Example	
Question	Would someone in Mumbai refer to Solanum melongena as an eggplant? **Gold answer:** False
Student's output before tuning	True: In Mumbai, Solanum melongena is commonly referred to as an eggplant.
Teachers' correct rationale	False. In Mumbai, people would more commonly refer to "Solanum melongena" as "brinjal" or "baingan" rather than "eggplant." While "eggplant" is widely used in American English, in India, especially in regions like Mumbai, the term "brinjal" is more common. "Baingan" is a Hindi word used across many parts of India.
Teachers' mistakes feedback	The student's answer contains a mistake in the explanation provided. The correct answer to the question is False, but the student mistakenly provided a hint that supports the correct answer. The correct information should be that in Mumbai, Solanum melongena is commonly referred to as a brinjal, not an eggplant.
Student's output after tuning	Answer: False. Solanum melongena is not referred to as "eggplant" in Mumbai. The term "eggplant" is specifically used in American English and Solanum melongena is known as "brinjal" in Mumbai. So, the answer is false.
LogiQA Example	
Question	Some Cantonese don't like chili, so some southerners don't like chili. Which of the following can guarantee the above argument? Options: ["Some Cantonese love chili.", "Some people who like peppers are southerners.", "All Cantonese are southerners.", "Some Cantonese like neither peppers nor sweets."]. **Gold answer:** "All Cantonese are southerners."
Student's output before tuning	Answer: ["Some Cantonese love chili."] Rationale: The argument states that "Some Cantonese don't like chili," which implies that not all Cantonese people dislike chili. Therefore, the statement "Some Cantonese love chili" is the only option that can guarantee the argument.
Teachers' correct rationale	The correct option is: "All Cantonese are southerners." Explanation: The argument states that some Cantonese don't like chili, and to conclude from this that some southerners don't like chili, we need to ensure that Cantonese are included within the broader category of southerners. If all Cantonese are southerners, then the statement about "some Cantonese not liking chili" can be applied to southerners as well, making the argument valid.
Teachers' mistakes feedback	The student's answer is incorrect. The statement "Some Cantonese love chili" does not guarantee the argument about "southerners" provided in the question. The correct answer should be "All Cantonese are southerners." This option directly connects the dislike for chili among some Cantonese to the fact that some southerners don't like chili, making it the most relevant choice to support the argument.
Student's output after tuning	Answer: ["All Cantonese are southerners."] Explanation: The argument states: "Some Cantonese don't like chili, so some southerners don't like chili." To guarantee this argument, we need to establish that some southerners (specifically Cantonese) don't like chili.

Table 9: Case study of distillation impact on student LM's output