清華大學电子工程系
Department of Electronic Engineering, Tsinghua University

Can LLMs Learn by Teaching for Better Reasoning? A Preliminary Study

Xuefei Ning[*1], Zifu Wang[*2], Shiyao Li[*1,3], Zinan Lin[*4], Peiran Yao[*3,5],
Tianyu Fu[1,3], Matthew B. Blaschko[2], Guohao Dai[3,6], Huazhong Yang[1], Yu Wang[1]
foxdoraame@gmail.com

[1]Tsinghua University [2]KU Leuven [3]Infinigence-AI
[4]Microsoft Research [5]University of Alberta [6]Shanghai Jiao Tong University

Contents

1 Overview

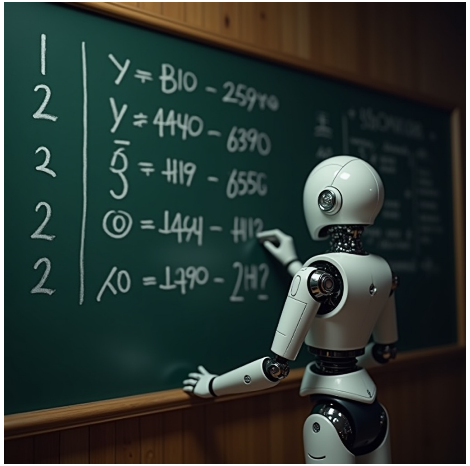
2 Method Design and Results

3 Broader Discussion

4 Summary

Motivation

- In recent years, we have been impressed by LLMs with extensive knowledge, strong planning skills, and good intuitions.
- However, the ability of current LLMs to provide accurate knowledge and reasoning appears to lag behind other abilities.



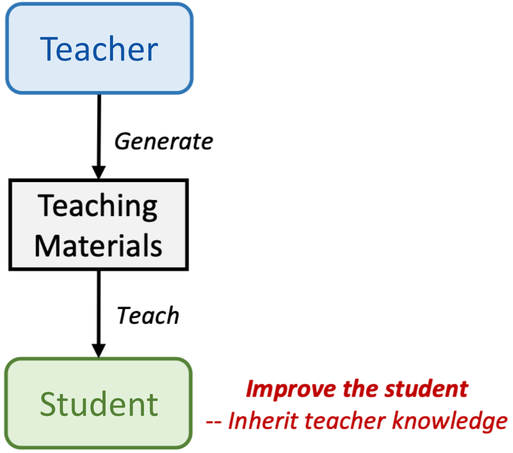
Mathematical Reasoning

LLM Coder

* Images are generated by liblib: https://www.liblib.art/

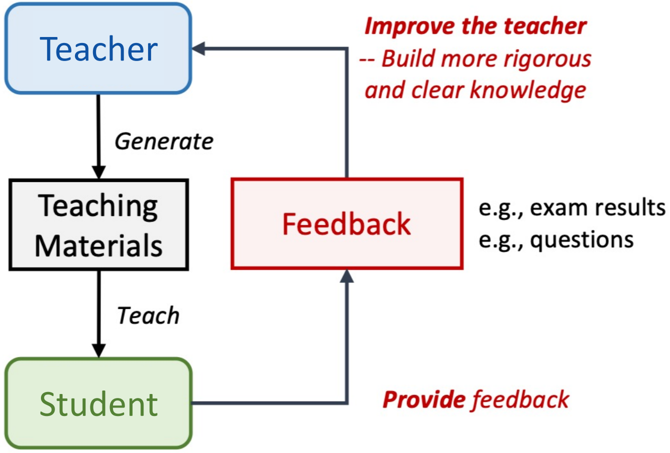
Motivation

- **In human education, how to improve the accurate knowledge building and reasoning?**
  - **Learn from Teacher (LfT)**: Use the teacher to improve the student.
  - **Learn by Teaching (LbT)**: Use the student feedback to improve the teacher. LbT has been shown to effectively promote accurate knowledge building and reasoning.



Learning from Teacher (LfT)

For example:
- *Learning from human teacher's annotations*
- *Learning from a stronger model (knowledge distillation)*



Learning by Teaching (LbT)

Benefits of LbT in Human Education

- **In human education, how and why LbT improves the teacher abilities?**

(a) Increased self-accountability

The task of teaching introduces social pressure and incentives to teachers, encouraging individuals to raise their standards and work harder.

(b) Explicit articulation of implicit and vague thoughts

- During the preparation of teaching materials, the teacher needs to use clear and organized language to convey its inner thoughts.
- **LbT assumption on teaching material quality (LbT-TMQ assumption)**: Teaching materials that make it easier for students to learn have clearer and more accurate logic.

Benefits of LbT in Human Education

- **In human education, how and why LbT improves the teacher abilities?**

(c) Iterative feedback from diverse students

In the teaching process, interaction with students of varying ability levels and knowledge backgrounds offers valuable feedback.

The teach or

struggle w this

interactive

- Recog be straightforward to the teacher but require more information for students.
- Identify gaps in teachers' own knowledge.
- Discover novel connections when addressing students' misconceptions and erroneous associations.

Our Question:

Can LLMs also learn by teaching for Better Reasoning?

Explore LbT in LLMs

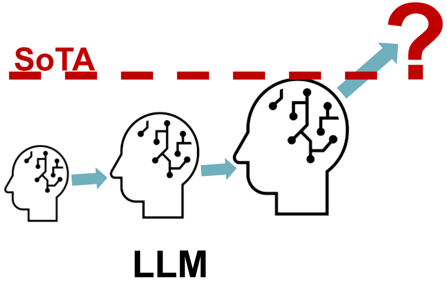
- **Can LLMs also learn by teaching for better reasoning?**

We implement the LbT idea in LLMs to construct three methods, especially focusing on the potential benefits (b) and (c) of LbT.

In the future, implanting incentives (a) into the LLM learning process is also worth trial.

- For example, setting up a *collaborative multi-agent learning framework* with <u>proper rewards</u> and <u>communication restrictions</u>

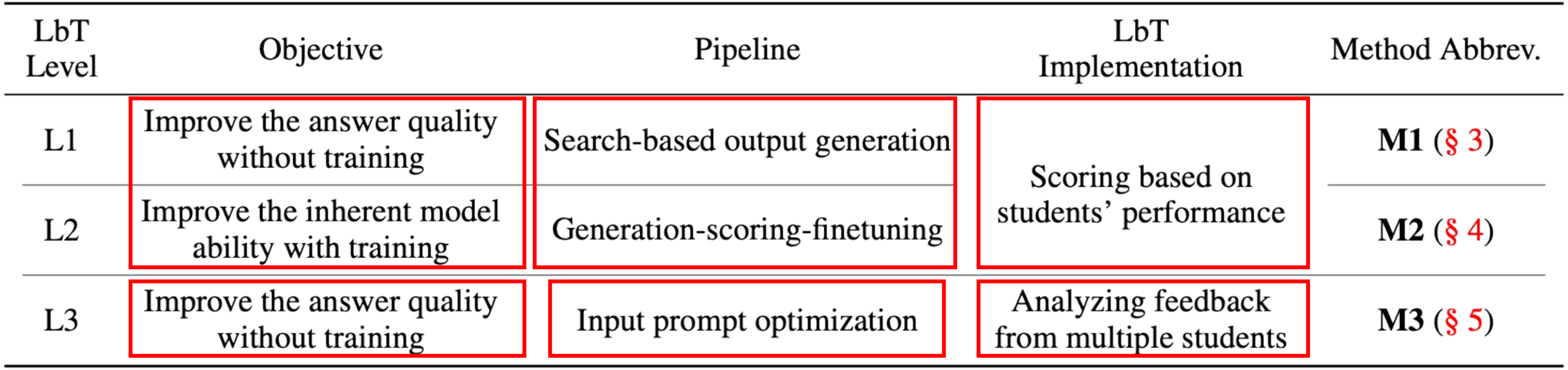
LbT has the potential to improve stronger models by having them teach weaker ones (weak-to-strong generalization). This might offer opportunities for continuous model evolution, especially as the data scaling faces challenges.



Our LbT Implementations

- Borrowing from benefit (b), our **M1**/**M2** implement LbT as a **rationale / answer scoring method** in the well-established search-based output generation or generation-scoring-finetuning pipelines, using the student's performance to score one rationale and answer of the teacher.
- Borrowing from benefit (c), our **M3** implements LbT as an **iterative prompt tuning process**, in which the teacher analyzes student's failure cases and improve the prompt for the teacher itself.

Table 1: The explored **M1**, **M2**, **M3** methods.



LbT Level	Objective	Pipeline	LbT Implementation	Method Abbrev.
L1	Improve the answer quality without training	Search-based output generation	Scoring based on students' performance	**M1** (§ 3)
L2	Improve the inherent model ability with training	Generation-scoring-finetuning		**M2** (§ 4)
L3	Improve the answer quality without training	Input prompt optimization	Analyzing feedback from multiple students	**M3** (§ 5)

Building a procedure analogy between M1, M2, M3 and three levels of teaching procedure in human learning:

- **M1-Level 1: Observing students' feedback**
- **M2-Level 2: Learning from the feedback**
- **M3-Level 3: Learning from the feedback iteratively**

Results Summary

	M1	M2	M3
Implementation	• **Based on the LbT-TMQ assumption** • **Search-based output generation pipeline with LbT-based scoring**	• **Generating-scoring-finetuning pipeline with LbT-based scoring**	• **Let the LLM iteratively refine ICL examples by analyzing the students' feedback**
Task (Dataset): Results/Insights	• **Mathematical reasoning (MATH)**: 3.31% ~ 18.23% improvement over SC with the same number of rationales. 0.17%~8.29% improvement over SC with comparable or lower compute. • **Code synthesis (Leetcode problems)**: Notable improvements in LeetCode score.	**Mathematical reasoning (MATH)**: For LLaMA3-8B, the M2-tuned model achieves a 1.8% improvement over correctness-based DPO, on 500 MATH test problems.	**Verbal logical reasoning (Liar/Logic)**: • M3 can craft better ICL examples through multiple refinement rounds. • The feedback from students other than the teacher itself is beneficial.

Contents

1 Overview

2 Method Design and Results

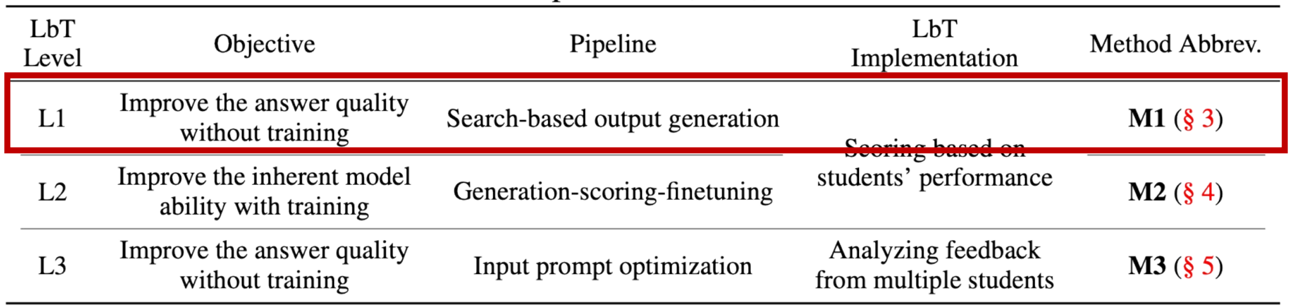
3 Broader Discussion

4 Summary

Method Design: M1

- **L1: Observing students' feedback**. The teacher instructs the students, who then provide feedback (taking exams and reporting the score). The student exam score can be used as an indicator of the quality of the teaching material.
  - **LbT assumption on teaching material quality (LbT-TMQ assumption)**: Teaching materials that make it easier for students to learn have clearer and more accurate logic.
- **Idea of M1**: For a given teaching problem, we generate a set of rationale and answers, and score each rationale and answer pair based on its ability to teach student models (using in-context learning) to correctly answer similar problems. We hope that this LbT score can help select better answer for the teaching problem and achieve a higher answer accuracy.

Table 1: The explored **M1**, **M2**, **M3** methods.



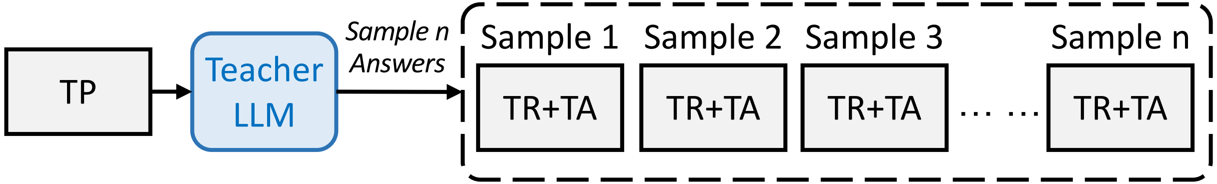
LbT Level	Objective	Pipeline	LbT Implementation	Method Abbrev.
L1	Improve the answer quality without training	Search-based output generation	Scoring based on students' performance	**M1** (§ 3)
L2	Improve the inherent model ability with training	Generation-scoring-finetuning		**M2** (§ 4)
L3	Improve the answer quality without training	Input prompt optimization	Analyzing feedback from multiple students	**M3** (§ 5)

Method Design: M1

- **Key implementation choices in M1**
  - **What is the form of the teaching material?** => The teacher generates a **teaching rationale (TR) and answer (TA)** for the teaching problem (TP) as the teaching material.
  - **How do the student learn from the teaching material?** => The student learns from the TP-TR-TA example using **in-context learning**.
  - **How to evaluate the student's learning performance?** => The student takes "an exam" to **solve some exam problems (EPs) similar to the teaching problem (TP)**, and gets an **exam score**.
  - **How do we utilize the feedback of the student's learning performance?** => The exam score is the "LbT score" of the teaching material (TR-TA pair). The teacher will **generate multiple TR-TAs**, and **select the TA with the highest LbT score – M1 (MAX)**, or **use weighted voting to decide the TA – M1 (SUM)**.

Method Design: M1

- Firstly, we let the teacher LLM generate multiple TR-TA pairs for a given TP.



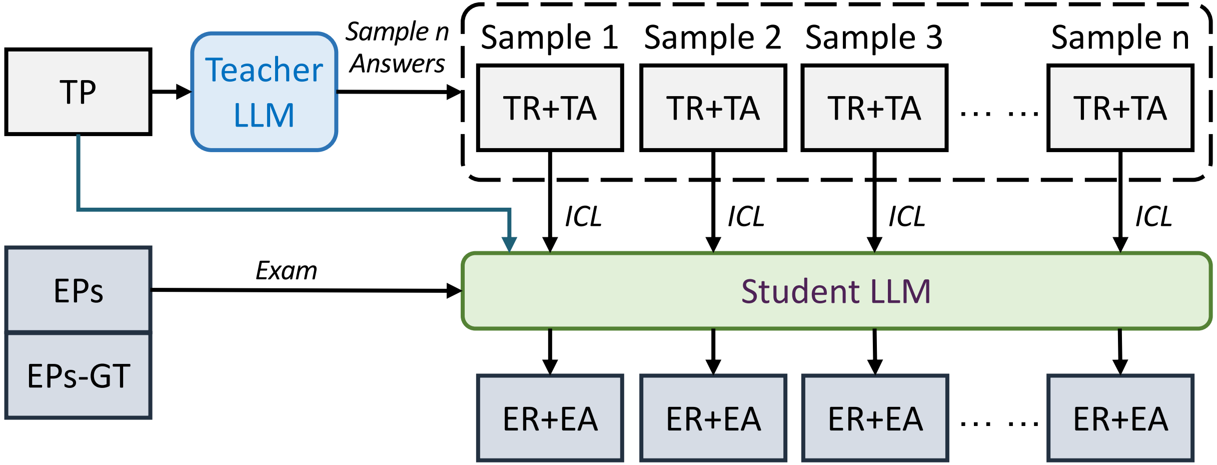
Note:

Teaching Problem – TP
Teaching Rationale – TR
Teaching Answer – TA

Exam Problem – EP
Exam Rationale – ER
Exam Answer – EA

Method Design: M1

- Secondly, each TR-TA pair is separately used as the in-context learning (ICL) example to guide the student model in solving a series of EPs.

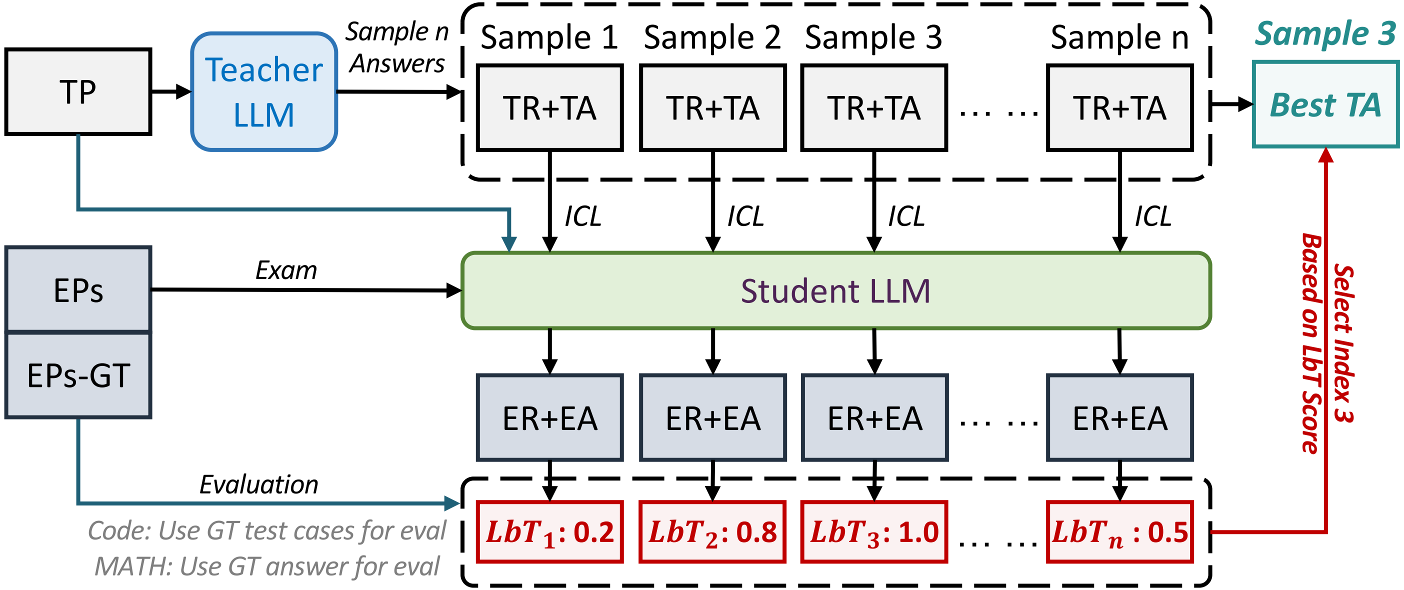


Note:
Teaching Problem – TP
Teaching Rationale – TR
Teaching Answer – TA

Exam Problem – EP
Exam Rationale – ER
Exam Answer – EA

Method Design: M1

- Finally, with the produced Exam Rationales (ERs) and Exam Answers (EAs), each student will then receive an exam score, denoted as the LbT score**. The LbT score can be used as a quality assessment of the corresponding TR-TA pair.**



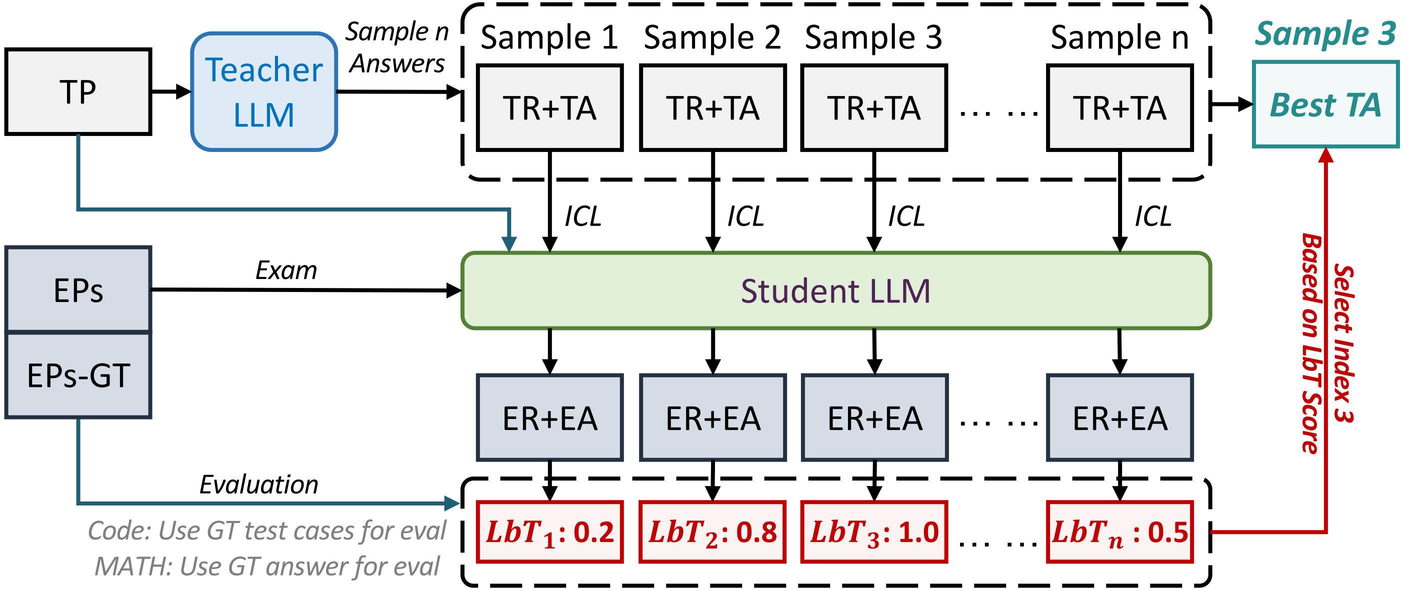
Code: Use GT test cases for eval
MATH: Use GT answer for eval

Note:
Teaching Problem – TP
Teaching Rationale – TR
Teaching Answer – TA

Exam Problem – EP
Exam Rationale – ER
Exam Answer – EA

Method Design: M1

- We consider two ways to select the final TA:
  - **M1 (MAX):** We select the TR-TA pair with the highest LbT score. As shown in the figure.
  - **M1 (SUM):** For datasets whose answer equivalence can be decided relatively easily, e.g., via exact matching, as in the MATH dataset, we can take the sum of the LbT scores for each TA separately, and select the TA with the maximum sum.



Code: Use GT test cases for eval
MATH: Use GT answer for eval

Note:
Teaching Problem – TP
Teaching Rationale – TR
Teaching Answer – TA

Exam Problem – EP
Exam Rationale – ER
Exam Answer – EA

Evaluation on Mathematical Reasoning: M1

- Experimental Setups
  - Dataset: Functional MATH, i.e., MATH() [2]
    - An extension of the MATH dataset [1].
    - 181 problems in the MATH test set (according to the 12k-500 train-test split in [3]) are provided with 3 functional variants each.
  - TPs: 181 (out of 500) test problems that have functional variants in MATH().
  - TR-TA generation: For each TP, we sample 256 TR-TA pairs.
  - EPs for each TP: We utilize the 3 functional variants of TP as EPs.
  - ER-EA generation: Each EP is answered 3 times with randomized student decoding, resulting in 9 EP-ER-EA pairs in total for scoring each TR-TA pair.

[1] Dan Hendrycks, et al. Measuring mathematical problem solving with the math dataset. NeurIPS, 2021.

[2] Saurabh Srivastava, et al. Functional benchmarks for robust evaluation of reasoning performance, and the reasoning gap. arXiv, 2024.

[3] Hunter Lightman,et al. Let's verify step by step. arXiv, 2023.

Evaluation on Mathematical Reasoning: M1

- Experimental Results
  - **M1 is effective with various model settings and surpasses baselines.** M1 exceeds self-consistency (SC) with various model settings.

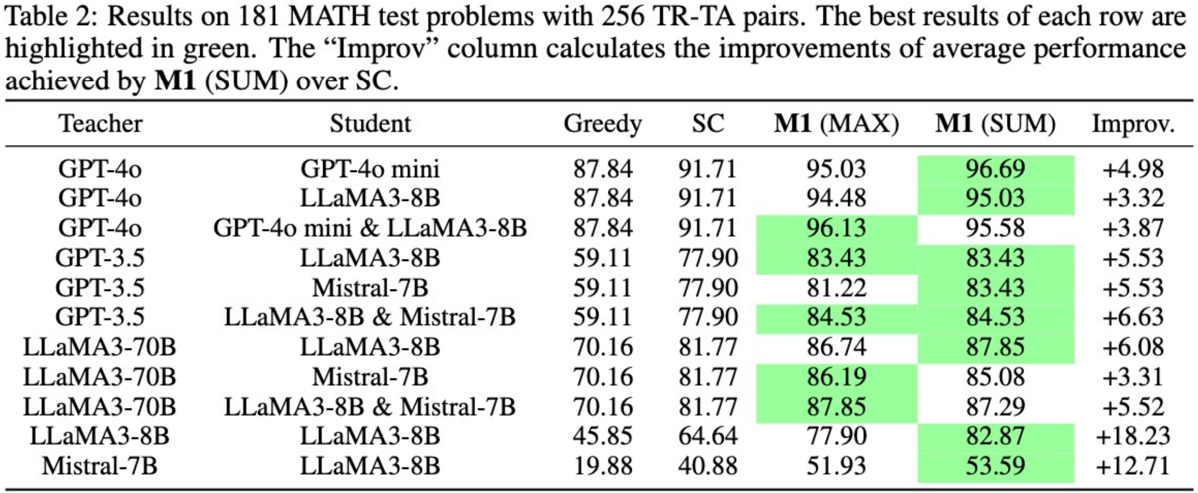


Table 2: Results on 181 MATH test problems with 256 TR-TA pairs. The best results of each row are highlighted in green. The "Improv" column calculates the improvements of average performance achieved by **M1** (SUM) over SC.

Teacher	Student	Greedy	SC	**M1** (MAX)	**M1** (SUM)	Improv.
GPT-4o	GPT-4o mini	87.84	91.71	95.03	96.69	+4.98
GPT-4o	LLaMA3-8B	87.84	91.71	94.48	95.03	+3.32
GPT-4o	GPT-4o mini & LLaMA3-8B	87.84	91.71	96.13	95.58	+3.87
GPT-3.5	LLaMA3-8B	59.11	77.90	83.43	83.43	+5.53
GPT-3.5	Mistral-7B	59.11	77.90	81.22	83.43	+5.53
GPT-3.5	LLaMA3-8B & Mistral-7B	59.11	77.90	84.53	84.53	+6.63
LLaMA3-70B	LLaMA3-8B	70.16	81.77	86.74	87.85	+6.08
LLaMA3-70B	Mistral-7B	70.16	81.77	86.19	85.08	+3.31
LLaMA3-70B	LLaMA3-8B & Mistral-7B	70.16	81.77	87.85	87.29	+5.52
LLaMA3-8B	LLaMA3-8B	45.85	64.64	77.90	82.87	+18.23
Mistral-7B	LLaMA3-8B	19.88	40.88	51.93	53.59	+12.71

*SC: self-consistency

Evaluation on Mathematical Reasoning: M1

- Experimental Results
  - **M1 can further benefit from multiple students.** Using GPT-3.5 to teach both LLaMA3-8B and Mistral-7B achieves a improvement over teaching LLaMA3-8B or Mistral-7B separately.

Teaching multiple students might be better

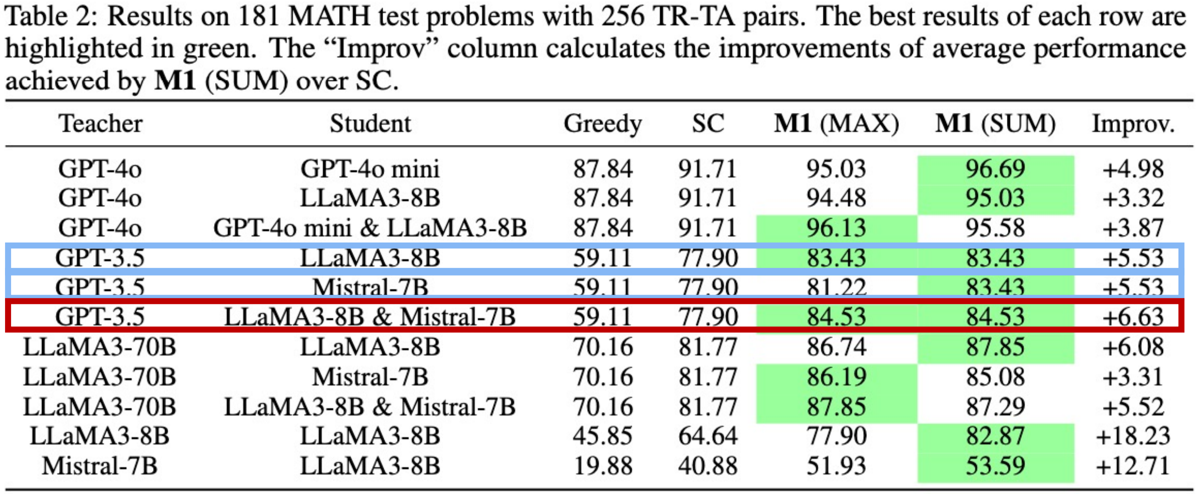


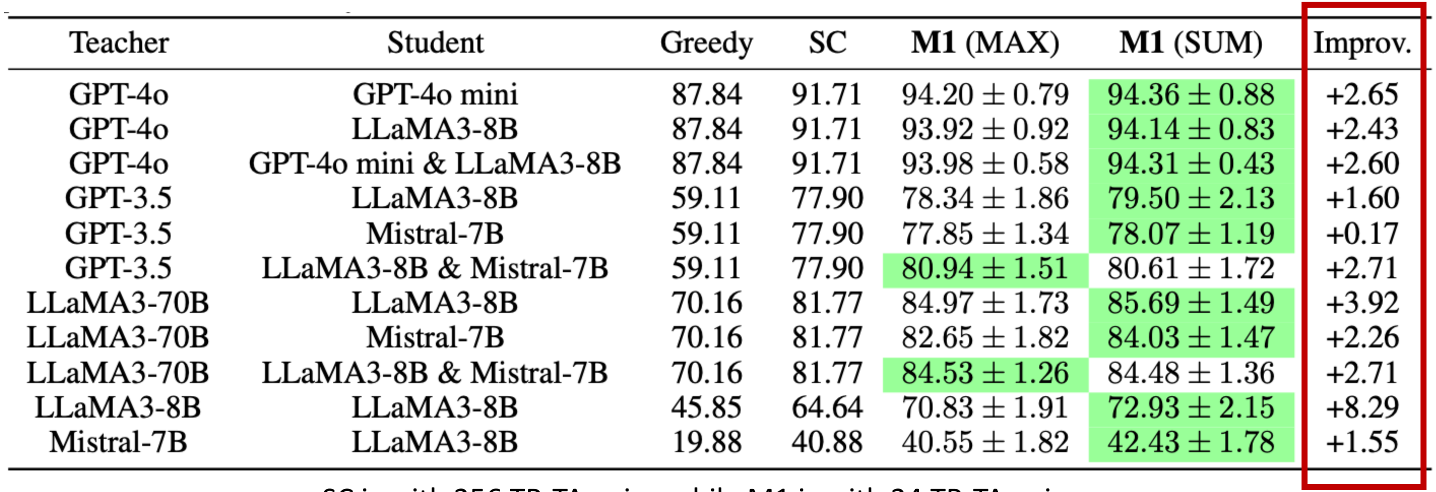
Table 2: Results on 181 MATH test problems with 256 TR-TA pairs. The best results of each row are highlighted in green. The "Improv" column calculates the improvements of average performance achieved by **M1** (SUM) over SC.

Teacher	Student	Greedy	SC	**M1** (MAX)	**M1** (SUM)	Improv.
GPT-4o	GPT-4o mini	87.84	91.71	95.03	96.69	+4.98
GPT-4o	LLaMA3-8B	87.84	91.71	94.48	95.03	+3.32
GPT-4o	GPT-4o mini & LLaMA3-8B	87.84	91.71	96.13	95.58	+3.87
GPT-3.5	LLaMA3-8B	59.11	77.90	83.43	83.43	+5.53
GPT-3.5	Mistral-7B	59.11	77.90	81.22	83.43	+5.53
GPT-3.5	LLaMA3-8B & Mistral-7B	59.11	77.90	84.53	84.53	+6.63
LLaMA3-70B	LLaMA3-8B	70.16	81.77	86.74	87.85	+6.08
LLaMA3-70B	Mistral-7B	70.16	81.77	86.19	85.08	+3.31
LLaMA3-70B	LLaMA3-8B & Mistral-7B	70.16	81.77	87.85	87.29	+5.52
LLaMA3-8B	LLaMA3-8B	45.85	64.64	77.90	82.87	+18.23
Mistral-7B	LLaMA3-8B	19.88	40.88	51.93	53.59	+12.71

*SC: self-consistency

Evaluation on Mathematical Reasoning: M1

- Experimental Results
  - M1 incurs higher inference cost than SC when using the same number of TR-TA pairs.
  - We show that with comparable or much lower compute, M1 with just 24 TR-TA pairs achieves a 0.17%~8.29% accuracy improvement over SC with 256 TR-TA pairs.



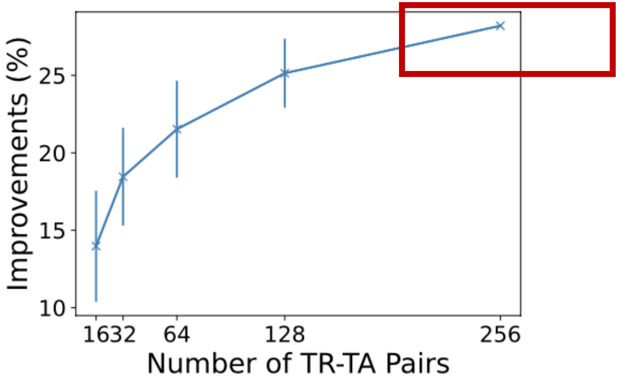
Teacher	Student	Greedy	SC	**M1** (MAX)	**M1** (SUM)	Improv.
GPT-4o	GPT-4o mini	87.84	91.71	94.20 ± 0.79	94.36 ± 0.88	+2.65
GPT-4o	LLaMA3-8B	87.84	91.71	93.92 ± 0.92	94.14 ± 0.83	+2.43
GPT-4o	GPT-4o mini & LLaMA3-8B	87.84	91.71	93.98 ± 0.58	94.31 ± 0.43	+2.60
GPT-3.5	LLaMA3-8B	59.11	77.90	78.34 ± 1.86	79.50 ± 2.13	+1.60
GPT-3.5	Mistral-7B	59.11	77.90	77.85 ± 1.34	78.07 ± 1.19	+0.17
GPT-3.5	LLaMA3-8B & Mistral-7B	59.11	77.90	80.94 ± 1.51	80.61 ± 1.72	+2.71
LLaMA3-70B	LLaMA3-8B	70.16	81.77	84.97 ± 1.73	85.69 ± 1.49	+3.92
LLaMA3-70B	Mistral-7B	70.16	81.77	82.65 ± 1.82	84.03 ± 1.47	+2.26
LLaMA3-70B	LLaMA3-8B & Mistral-7B	70.16	81.77	84.53 ± 1.26	84.48 ± 1.36	+2.71
LLaMA3-8B	LLaMA3-8B	45.85	64.64	70.83 ± 1.91	72.93 ± 2.15	+8.29
Mistral-7B	LLaMA3-8B	19.88	40.88	40.55 ± 1.82	42.43 ± 1.78	+1.55

SC is with 256 TR-TA pairs, while M1 is with 24 TR-TA pairs.

Evaluation on Mathematical Reasoning: M1

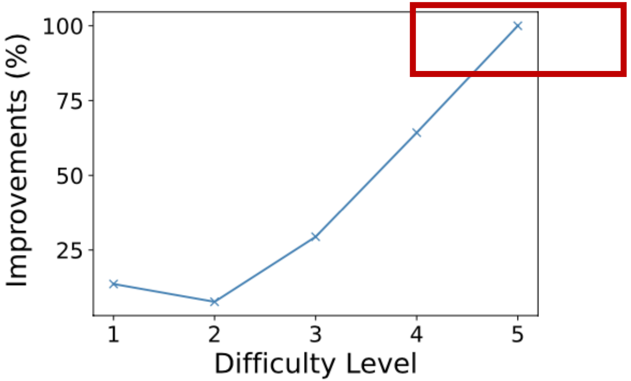
- Experimental Results
  - **The relative improvement of M1 over SC increases as the number of TR-TA pairs or the difficulty level grows.**
    - The improvements do not saturate at 256 TR-TA pairs.
    - The improvements are larger at harder problems: M1 can identify infrequent but correct TAs.

the improvements ***do not saturate***



Relative improvements of M1 over SC using LLaMA3-8B as the teacher and student with respect to number of TR-TA pairs

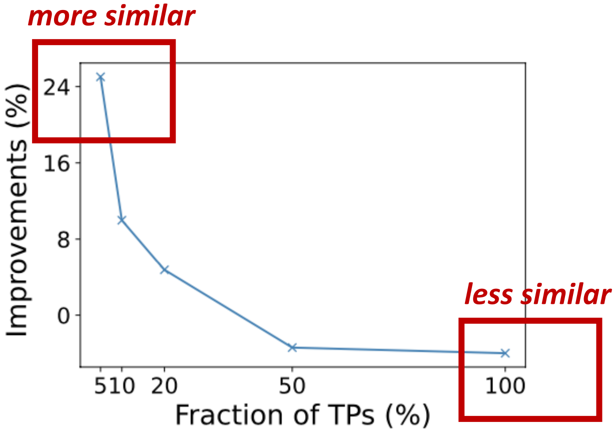
LbT is much better than SC on ***harder problems***



Relative improvements of M1 over SC using LLaMA3-8B as the teacher and student with respect to difficulty level

Evaluation on Mathematical Reasoning: M1

- Experimental Results
  - **The TP and the corresponding EPs should be similar for the LbT score to be indicative of the TR-TA quality**. We use the functional variants as EPs, which are very similar to TPs.
  - **M1 only provides improvements for TPs that have similar problems in the training set.**



Relative improvements of M1 over SC using LLaMA3-8B as the teacher and student with respect to the fraction of TPs when sorted by the cosine distance to the 2 closest problems from the training set

Evaluation on Competition-Level Code Synthesis: M1

- Experimental Setups
  - Datasets: Grandmaster Dynamic Programming (DP) study plan on Leetcode.
    - Each problem group (or dataset) in the study plan has N=5~10 problems (**TPs**)
    - Each problem has 2~3 visible test cases and many hidden test cases (Need submission)
  - Evaluation:
    - **Visible score (V-score):** If the code *passes all* visible cases, assign 1; otherwise, 0. We calculate **the exam V-score as the LbT score** to avoid additional Leetcode submissions.
    - **Submit score (S-score):** Submit the code to Leetcode and record the pass rate.

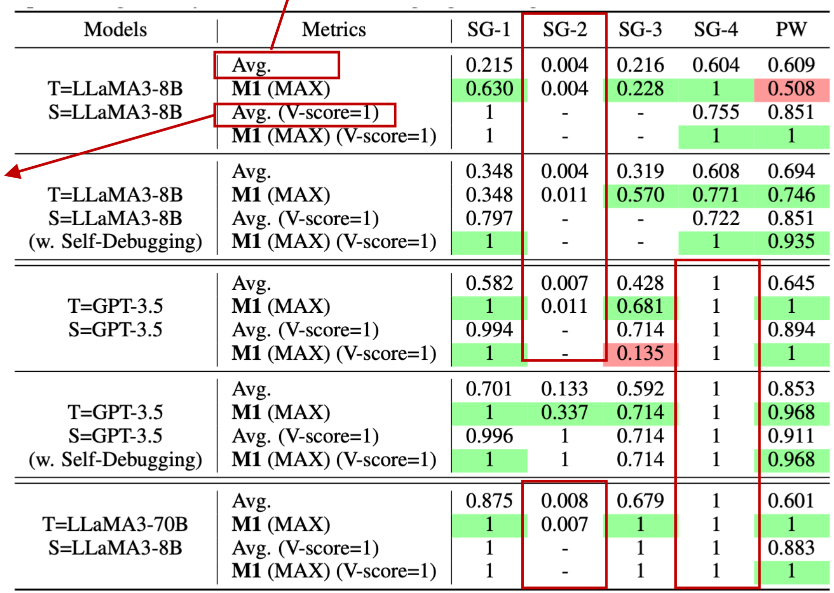
  - TPs: 18 problem in 3 datasets (Game Theory: 5 problems, Bitmasking: 6 problems, General-1D: 7 problems).
  - TR-TA generation: For each TP, we sample 8 TR-TA pairs (TR: natural language rationale, TA: Python code).
  - EPs for each TP: We use other problems in the same dataset with the TP as the N-1 EPs.
  - ER-EA generation: We use greedy decoding to generate 1 ER-EA pair for each EP, resulting in N-1 EP-ER-EA pairs in total.

Evaluation on Competition-Level Code Synthesis: M1

- Experimental Results
  - **M1 selects better TR-TA than the baseline in most cases** (marked in green).
  - **M1 shows the largest improvements on TPs with medium difficulty**: For very simple (e.g., SG-4 for GPT-3.5 & LLaMA3-70B) or challenging (e.g., SG-2 for models other than GPT-3.5) cases, M1 shows marginal improvements.

Average S-score of all TR-TA pairs

Average S-score of the TR-TA pairs whose V-score=1[1]



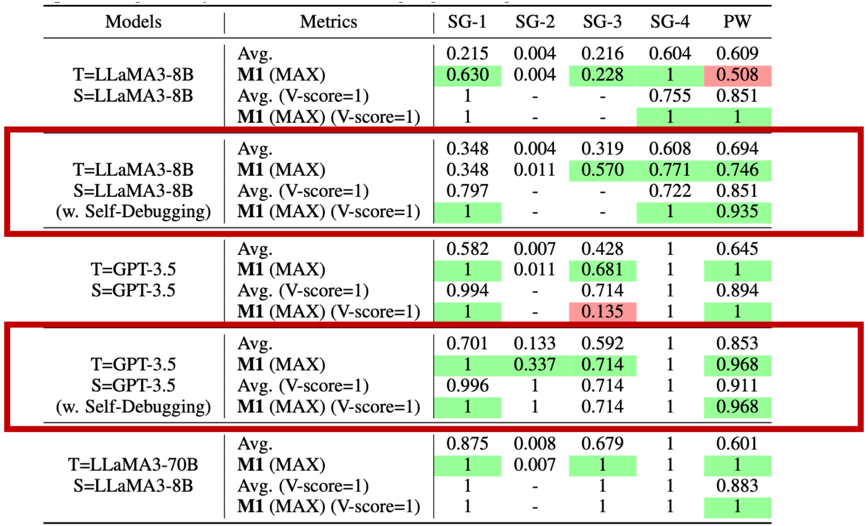
Models	Metrics	SG-1	SG-2	SG-3	SG-4	PW
T=LLaMA3-8B S=LLaMA3-8B	Avg.	0.215	0.004	0.216	0.604	0.609
	M1 (MAX)	0.630	0.004	0.228	1	0.508
	Avg. (V-score=1)	1	-	-	0.755	0.851
	M1 (MAX) (V-score=1)	1	-	-	1	1
T=LLaMA3-8B S=LLaMA3-8B (w. Self-Debugging)	Avg.	0.348	0.004	0.319	0.608	0.694
	M1 (MAX)	0.348	0.011	0.570	0.771	0.746
	Avg. (V-score=1)	0.797	-	-	0.722	0.851
	M1 (MAX) (V-score=1)	1	-	-	1	0.935
T=GPT-3.5 S=GPT-3.5	Avg.	0.582	0.007	0.428	1	0.645
	M1 (MAX)	1	0.011	0.681	1	1
	Avg. (V-score=1)	0.994	-	0.714	1	0.894
	M1 (MAX) (V-score=1)	1	-	0.135	1	1
T=GPT-3.5 S=GPT-3.5 (w. Self-Debugging)	Avg.	0.701	0.133	0.592	1	0.853
	M1 (MAX)	1	0.337	0.714	1	0.968
	Avg. (V-score=1)	0.996	1	0.714	1	0.911
	M1 (MAX) (V-score=1)	1	1	0.714	1	0.968
T=LLaMA3-70B S=LLaMA3-8B	Avg.	0.875	0.008	0.679	1	0.601
	M1 (MAX)	1	0.007	1	1	1
	Avg. (V-score=1)	1	-	1	1	0.883
	M1 (MAX) (V-score=1)	1	-	1	1	1

S-score results on the *Game Theory* dataset in LeetCode Grandmaster DP study plan.

[1] Yujia Li, et al. Competition-level code generation with alphacode. Science, 2022.

Evaluation on Competition-Level Code Synthesis: M1

- Experimental Results
  - **Self-Debugging (SD)[1] is both complementary to and beneficial for M1.** We experiment with applying one-iteration SD. SD usually fixes simple non-logical bugs.
    - Complementary (applying SD on TAs): SD fixes non-logical bugs in TAs such as missing imports, miswritten variable names, and so on. While M1 mainly assess the quality of the TR-TA logic.
    - Beneficial (applying SD on EAs): Fixing non-logical bugs in EAs can make the exam score more indicative of the quality of the TR-TA logic.



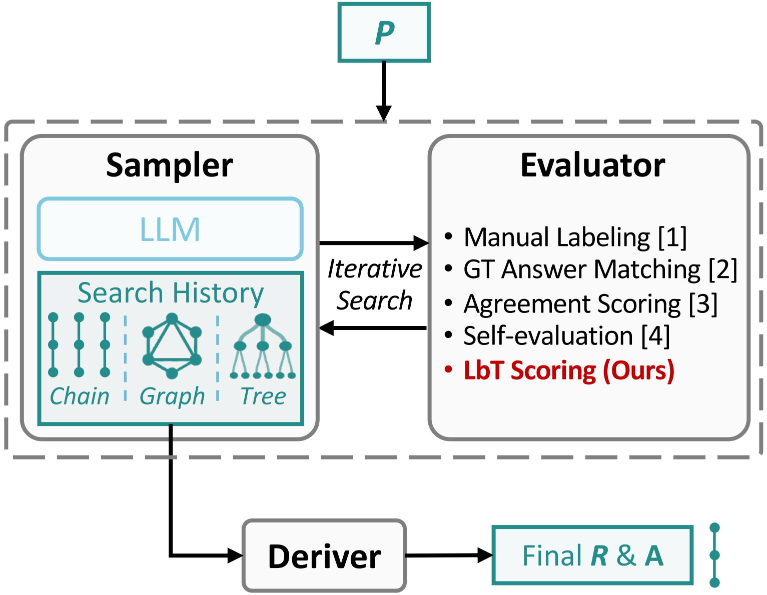
Models	Metrics	SG-1	SG-2	SG-3	SG-4	PW
T=LLaMA3-8B S=LLaMA3-8B	Avg.	0.215	0.004	0.216	0.604	0.609
	M1 (MAX)	0.630	0.004	0.228	1	0.508
	Avg. (V-score=1)	1	-	-	0.755	0.851
	M1 (MAX) (V-score=1)	1	-	-	1	1
T=LLaMA3-8B S=LLaMA3-8B (w. Self-Debugging)	Avg.	0.348	0.004	0.319	0.608	0.694
	M1 (MAX)	0.348	0.011	0.570	0.771	0.746
	Avg. (V-score=1)	0.797	-	-	0.722	0.851
	M1 (MAX) (V-score=1)	1	-	-	1	0.935
T=GPT-3.5 S=GPT-3.5	Avg.	0.582	0.007	0.428	1	0.645
	M1 (MAX)	1	0.011	0.681	1	1
	Avg. (V-score=1)	0.994	-	0.714	1	0.894
	M1 (MAX) (V-score=1)	1	-	0.135	1	1
T=GPT-3.5 S=GPT-3.5 (w. Self-Debugging)	Avg.	0.701	0.133	0.592	1	0.853
	M1 (MAX)	1	0.337	0.714	1	0.968
	Avg. (V-score=1)	0.996	1	0.714	1	0.911
	M1 (MAX) (V-score=1)	1	1	0.714	1	0.968
T=LLaMA3-70B S=LLaMA3-8B	Avg.	0.875	0.008	0.679	1	0.601
	M1 (MAX)	1	0.007	1	1	1
	Avg. (V-score=1)	1	-	1	1	0.883
	M1 (MAX) (V-score=1)	1	-	1	1	1

S-score results on the *Game Theory* dataset in LeetCode Grandmaster DP study plan.

[1] Xinyun Chen, et al. Teaching large language models to self-debug. arXiv, 2023.

Related Work Discussion: M1

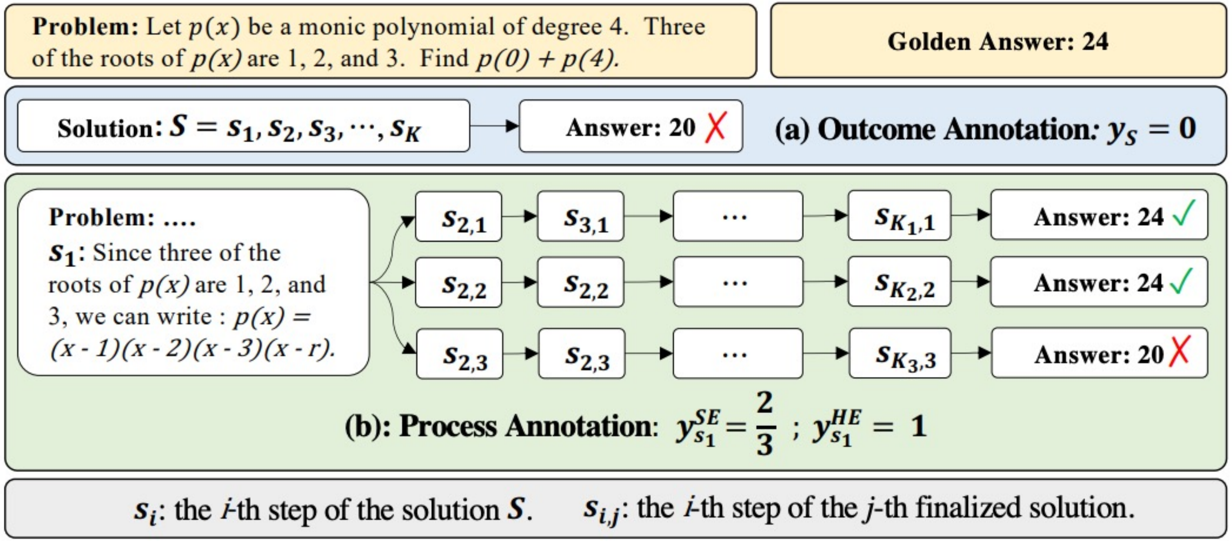
- Search-based output generation pipeline
  - (1) Sampler: Keep a search history of rationale steps or chains (possibly organized as graphs), sample new rationale chain or step.
  - (2) Evaluator: Evaluate the quality of each rationale chain or step. The evaluation score guides the sampler to do the search.
    - M1 designs an LbT evaluator that scores each rationale based on its ability in teaching student models to correctly answer similar problems.
  - (3) Deriver: Derive the final rationale or answer from the search history.



[1] Hunter Lightman,et al. Let's verify step by step. arXiv, 2023.
[2] Zheng Yuan, et al. Scaling relationship on learning mathematical reasoning with large language models. arXiv, 2023.
[3] Jiaxin Huang, et al. Large language models can self-improve. EMNLP, 2023.
[4] Weizhe Yuan, et al. Self-rewarding language models. arXiv, 2024.

Related Work Discussion: M1

- Math-shepherd can be considered as another implementation of **LbT scoring**
  - Math-shepherd evaluates each partial rationale by measuring how often another “completer” model arrives at the correct answer by continuing from the partial rationale.
  - In Math-Shepherd, the students (i.e., the completer) are examined by extending the partial teaching rationale for the same problem, whereas in our M1/M2, students are examined on similar problems, using the full rationale from the teaching problem as an exemplar.

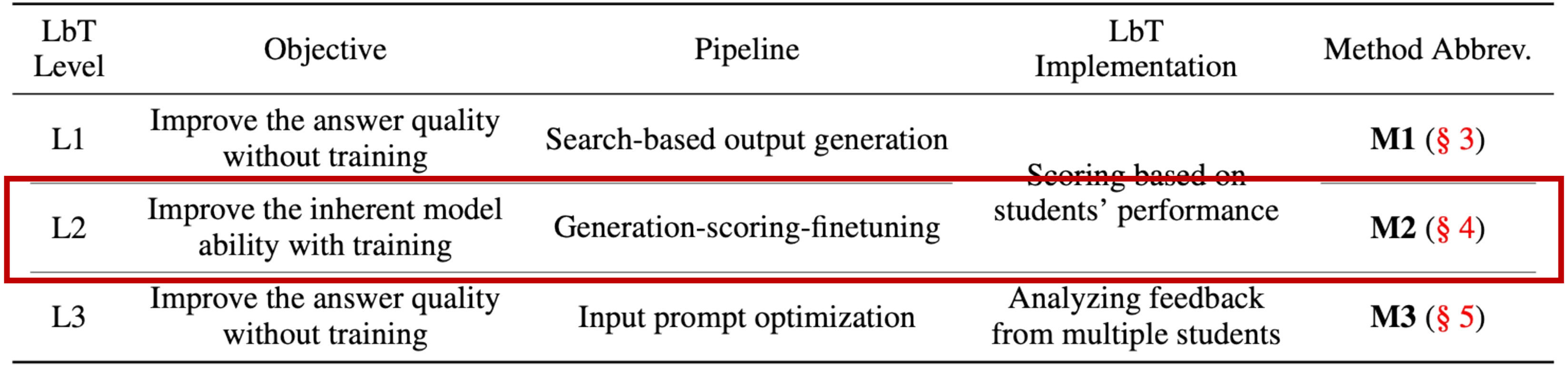

Problem: Let $p(x)$ be a monic polynomial of degree 4. Three of the roots of $p(x)$ are 1, 2, and 3. Find $p(0) + p(4)$.
Golden Answer: 24
Solution: $S = s_1, s_2, s_3, \cdots, s_K$
Answer: 20 ✗
(a) Outcome Annotation: $y_S = 0$
Problem: ....
s_1: Since three of the roots of $p(x)$ are 1, 2, and 3, we can write : $p(x) = (x - 1)(x - 2)(x - 3)(x - r)$.
$s_{2,1}$
$s_{3,1}$
$\cdots$
$s_{K_1,1}$
Answer: 24 ✓
$s_{2,2}$
$s_{2,2}$
$\cdots$
$s_{K_2,2}$
Answer: 24 ✓
$s_{2,3}$
$s_{2,3}$
$\cdots$
$s_{K_3,3}$
Answer: 20 ✗
(b): Process Annotation: $y_{s_1}^{SE} = \frac{2}{3}$; $y_{s_1}^{HE} = 1$
s_i: the i-th step of the solution S.
$s_{i,j}$: the i-th step of the j-th finalized solution.

Method Design: M2

- **L2: Learning from the feedback**. Further finetuning the teacher LLM to become better at reasoning, leveraging the student exam scores.

- **M2**: We use the LbT scoring method to score TRs. Then, we apply direct preference optimization (DPO) to fine-tune the teacher LLM with the TR-score pairs. We show that the LLM tuned by M2 is better than the LLM tuned when only using the TA correctness as the TR score.

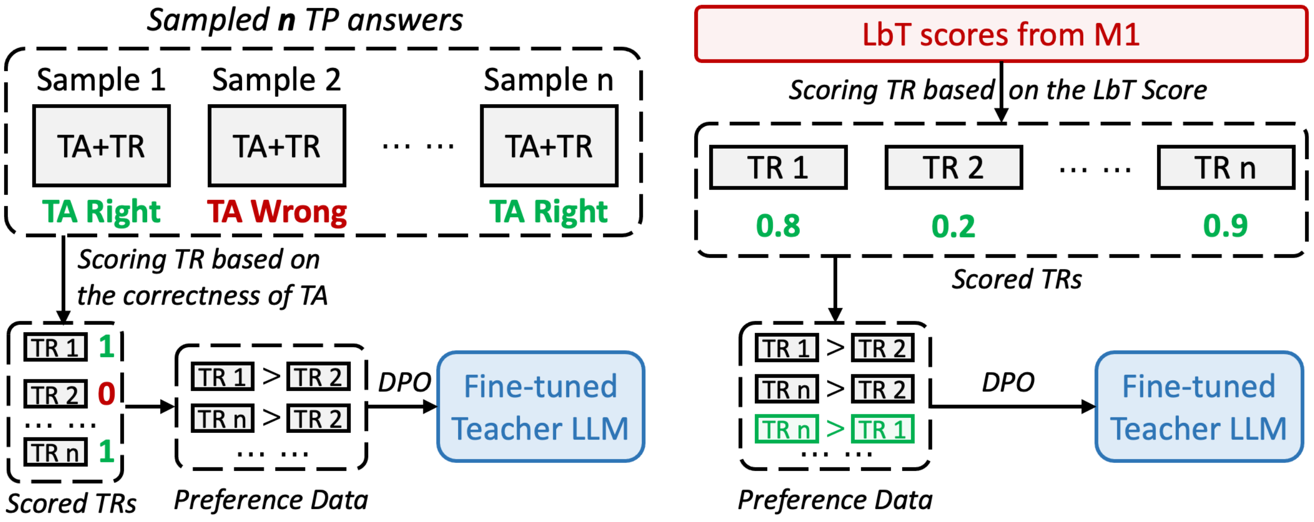
Table 1: The explored **M1**, **M2**, **M3** methods.



LbT Level	Objective	Pipeline	LbT Implementation	Method Abbrev.
L1	Improve the answer quality without training	Search-based output generation	Scoring based on students' performance	**M1** (§ 3)
L2	Improve the inherent model ability with training	Generation-scoring-finetuning		**M2** (§ 4)
L3	Improve the answer quality without training	Input prompt optimization	Analyzing feedback from multiple students	**M3** (§ 5)

Method Design: M2

- The baseline method only assess whether each TA is correct or incorrect. This scoring method cannot reflect which TR is best among multiple TRs whose TAs are all correct or incorrect.
- We collect the LbT scores of many TR-TA pairs (M1 scoring) and use them to finetune the teacher with DPO.



(a) Correctness-guided DPO (Baseline)

(b) DPO with LbT score (Ours)

Evaluation: M2

- Experimental Setups
  - Dataset:
    - Train: 1564 MATH training problems that have functional variants in MATH().
    - Test: 500 MATH test problems.
  - TPs: 1564 MATH training problems that have functional variants in MATH()
  - TR-TA generation: For each TP, we sample 32 TR-TA pairs from the teacher.
  - EPs for each TP: We utilize the 3 functional variants of TP as EPs.
  - DPO score: For each TR, we calculate 0.5 × TA correctness + 0.5 × LbT score as its score, where correctness is 1 or 0 when the TA is correct or wrong.
  - DPO preference pair selection: We select pairs from the 32 TRs whose score difference exceeds a threshold of 0.3, and keep at most 8 preference pairs for each TP.

Evaluation: M2

- Experimental Results
  - **M2 achieves better results than only using TA correctness score in DPO**.

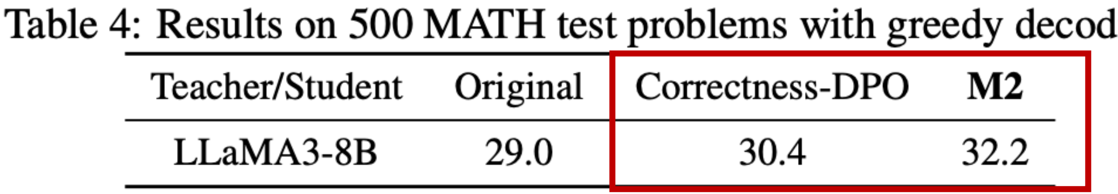


Table 4: Results on 500 MATH test problems with greedy decoding.

Teacher/Student	Original	Correctness-DPO	**M2**
LLaMA3-8B	29.0	30.4	32.2

  - **LbT can discern the preference between these TR-TA pairs.**
    - Although both TRs produce a correct TA, the losing TR is unnecessarily verbose and cannot be generalized to other similar problems.
    - Although both TRs produce a wrong TA, the winning TR is logically better than the loser.

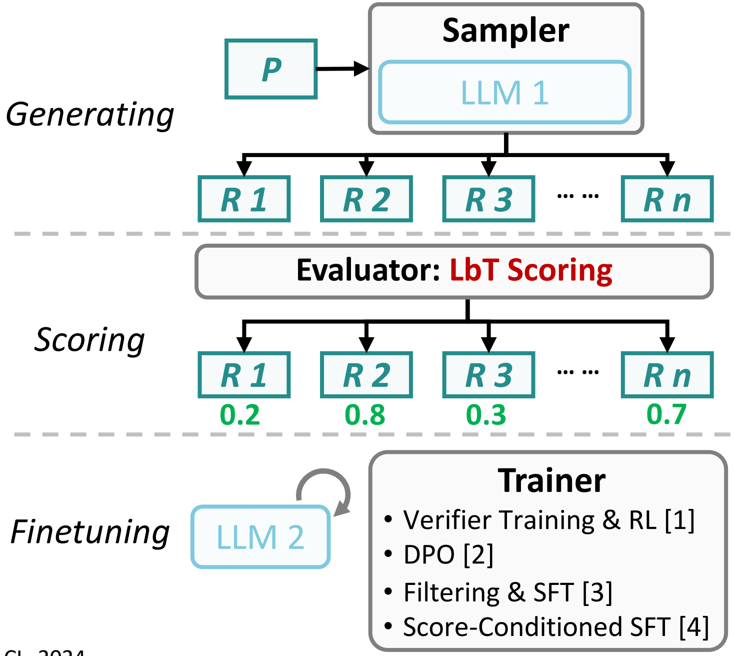
Related Work Discussion: M2

- Generation-scoring-finetuning pipeline

  - (1) Generating: Letting the target LLM or a teacher LLM generate multiple rationales for a given problem;

  - (2) Scoring: Scoring the rationales using an evaluator;

  - (3) Finetuning: Utilizing the rationales and scores to (optionally) train a verifier, and finetune the target LLM by reinforcement learning, DPO or its variant, filtering and supervised finetuning (SFT), or score-conditioned SFT.



[1] Karl Cobbe, et al. Training verifiers to solve math word problems. arXiv, 2021.

[2] Peiyi Wang, et al. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. ACL, 2024.

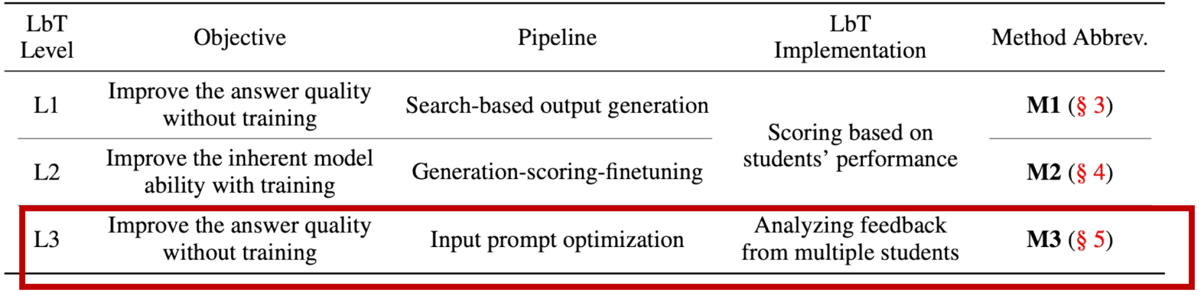
[3] Rafael Rafailov, et al. Direct preference optimization: Your language model is secretly a reward model. NeurIPS, 2023.

[4] Yu Meng, et al. Simpo: Simple preference optimization with a reference-free reward. arXiv, 2024.

Method Design: M3

- **L3: Learning from the feedback iteratively**. The teachers can teach the students (L1) and learn from the feedback (L2) *iteratively*.
  - Feedback form: Instead of leveraging the students’ exam scores, M3 leverages the students’ detailed exam responses to help the teacher iteratively refine its own prompts.
  - LbT Implementation & Analogous benefit: Instead of implementing the ***LbT-TMQ assumption*** as a ***scoring mechanism*** used in existing pipelines as in M1/M2 (LbT benefit b), M3 implements the LbT benefit c as an ***“iterative” prompt tuning process***.
- **M3**: We guide the teacher to iteratively improve teaching materials (a set of exemplars), based on the student and teacher performance when the set is used as the ICL examples. The final set of exemplars is used as the ICL examples to test the teacher’s performance on a hold-out test set.

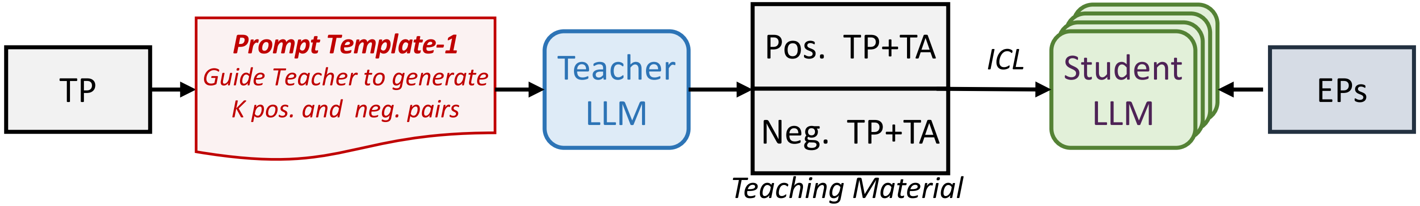
Table 1: The explored **M1**, **M2**, **M3** methods.



LbT Level	Objective	Pipeline	LbT Implementation	Method Abbrev.
L1	Improve the answer quality without training	Search-based output generation	Scoring based on students’ performance	**M1** (§ 3)
L2	Improve the inherent model ability with training	Generation-scoring-finetuning		**M2** (§ 4)
L3	Improve the answer quality without training	Input prompt optimization	Analyzing feedback from multiple students	**M3** (§ 5)

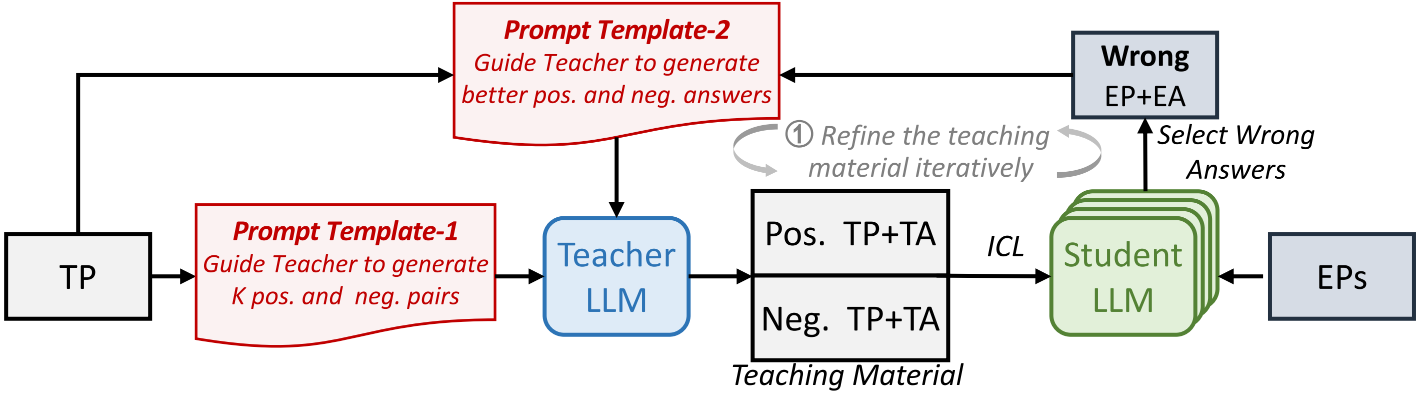
Method Design: M3

- Firstly: Given a task, we sample some exemplars, and use them as the ICL examples for the students to answer EPs.



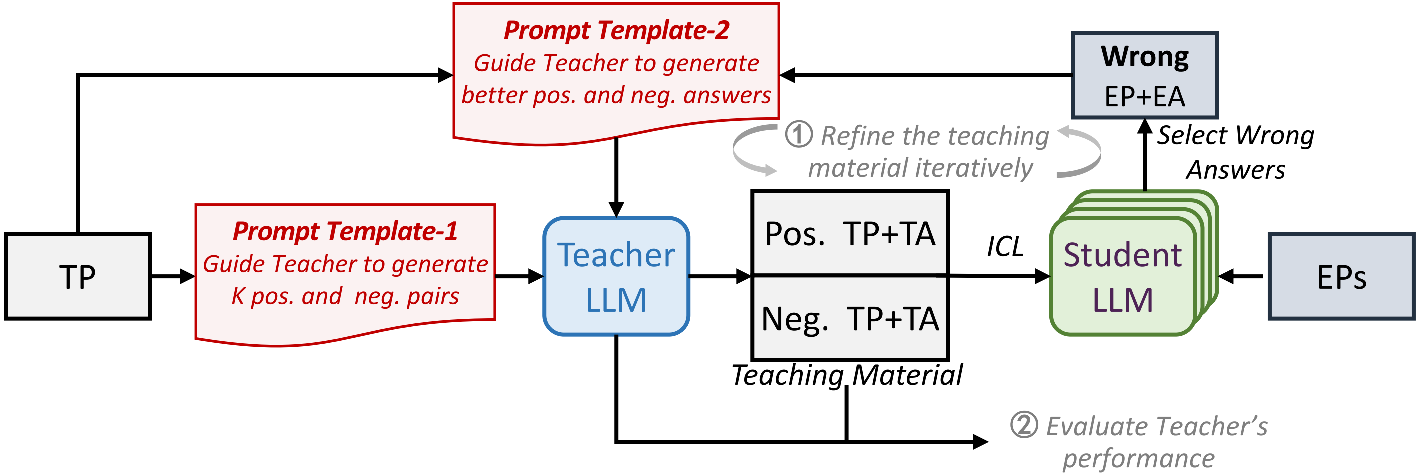
Method Design: M3

- Secondly: Run multiple refinement iterations. In each iteration:
  - The current exemplars are used as the ICL examples to teach students to answer EPs.
  - Select the EPs that students answered incorrectly and prompt the teacher to reflect on why the current exemplars might have misled students in these instances.
  - Based on the reflection, the teacher generates multiple updated exemplar sets.
  - Keep the exemplar set that achieves the best teacher performance.


Prompt Template-2
Guide Teacher to generate better pos. and neg. answers
Wrong
EP+EA
① Refine the teaching material iteratively
Select Wrong Answers
TP
Prompt Template-1
Guide Teacher to generate K pos. and neg. pairs
Teacher LLM
Pos. TP+TA
Neg. TP+TA
ICL
Student LLM
EPs
Teaching Material

Method Design: M3

- Finally: Report the teacher performance on the hold-out test set when using the resulting ICL examples.


Prompt Template-2
Guide Teacher to generate better pos. and neg. answers
Wrong
EP+EA
① Refine the teaching material iteratively
Select Wrong Answers
TP
Prompt Template-1
Guide Teacher to generate K pos. and neg. pairs
Teacher LLM
Pos. TP+TA
Neg. TP+TA
Teaching Material
ICL
Student LLM
EPs
② Evaluate Teacher's performance

Evaluation: M3

- Experimental Setups
  - Datasets:
    - **Liar**[1]: A false statement detection dataset. 4,574 statements with speaker and context.
    - **Logical Fallacy**[2]**:** A fallacy classification dataset. 2,449 samples of 13 fallacy types.
    - We cast these two tasks as a binary classification task.
  - M3 Implementation Details:
    - We maintain 4 exemplar sets. Each exemplar set contains 8 positive (class=1) and negative (class=0) exemplars.
    - We run a total of 5 iterations of teaching material improvements.
      - For each current set, the teacher LLM generates 8 new exemplar sets by analyzing the students' failures
      - Out of the 4x8=32 new sets, we choose the 4 exemplar sets with the highest F1 score on the training set.
    - We report the teacher's F1 score on the dev and test splits combined (mean and the standard error across 14 random experiments).

[1] William yang Wang, et al. "liar, liar pants on fire": A new benchmark dataset for fake news detection. ACL, 2017.
[2] Zhijing Jin, et al. Logical fallacy detection. EMNLP, 2022.

Evaluation: M3

- Experimental Results
  - **It is feasible to apply LbT on iterative prompt optimization**: LLMs are able to reflect on the failure cases of students and propose revised exemplars that improve the teacher's performance.

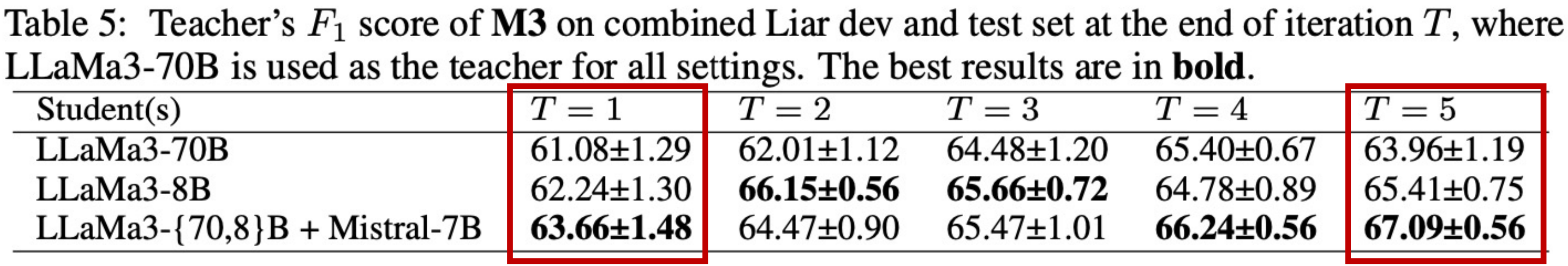


Table 5: Teacher's F_1 score of **M3** on combined Liar dev and test set at the end of iteration T, where LLaMa3-70B is used as the teacher for all settings. The best results are in **bold**.

Student(s)	$T = 1$	$T = 2$	$T = 3$	$T = 4$	$T = 5$
LLaMa3-70B	61.08±1.29	62.01±1.12	64.48±1.20	65.40±0.67	63.96±1.19
LLaMa3-8B	62.24±1.30	**66.15±0.56**	**65.66±0.72**	64.78±0.89	65.41±0.75
LLaMa3-{70,8}B + Mistral-7B	**63.66±1.48**	64.47±0.90	65.47±1.01	**66.24±0.56**	**67.09±0.56**

Evaluation: M3

- Experimental Results
  - **Having one or multiple LLMs different to the teacher as the student improves the quality of the teaching material faster**
    - We speculate that the benefits are brought by more diverse error types made by a different (weaker) student model.

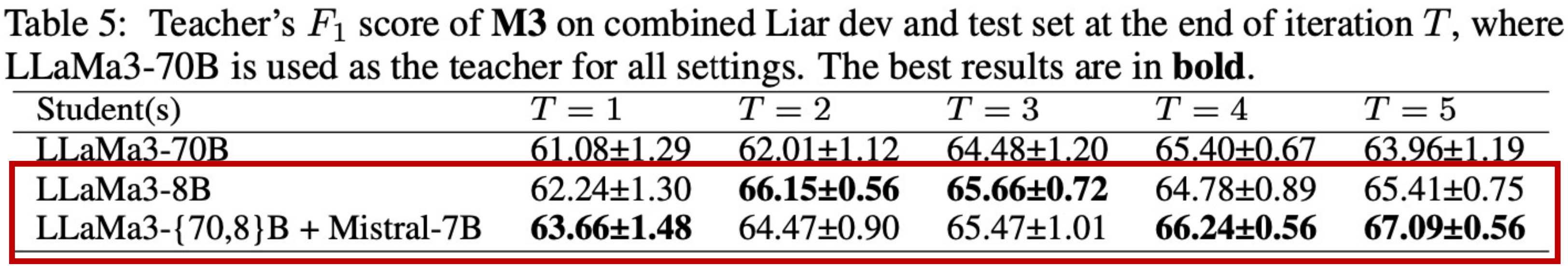


Table 5: Teacher's F_1 score of **M3** on combined Liar dev and test set at the end of iteration T, where LLaMa3-70B is used as the teacher for all settings. The best results are in **bold**.

Student(s)	$T=1$	$T=2$	$T=3$	$T=4$	$T=5$
LLaMa3-70B	61.08±1.29	62.01±1.12	64.48±1.20	65.40±0.67	63.96±1.19
LLaMa3-8B	62.24±1.30	**66.15±0.56**	**65.66±0.72**	64.78±0.89	65.41±0.75
LLaMa3-{70,8}B + Mistral-7B	**63.66±1.48**	64.47±0.90	65.47±1.01	**66.24±0.56**	**67.09±0.56**

Evaluation: M3

- Experimental Results
  - **Having diverse students help discover a diverse set of errors that the teacher could make**.
    - By choosing a student model different from the teacher model, we identify more types of valid causes of teacher mistakes.
    - M3 can reduce the errors of those causes.

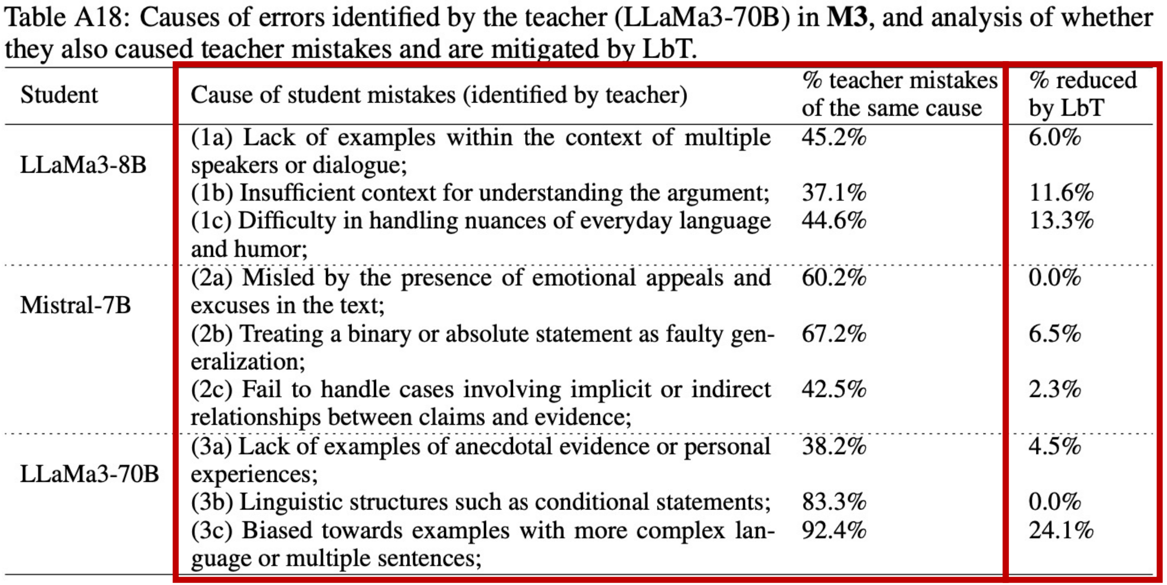


Table A18: Causes of errors identified by the teacher (LLaMa3-70B) in **M3**, and analysis of whether they also caused teacher mistakes and are mitigated by LbT.

Student	Cause of student mistakes (identified by teacher)	% teacher mistakes of the same cause	% reduced by LbT
LLaMa3-8B	(1a) Lack of examples within the context of multiple speakers or dialogue;	45.2%	6.0%
	(1b) Insufficient context for understanding the argument;	37.1%	11.6%
	(1c) Difficulty in handling nuances of everyday language and humor;	44.6%	13.3%
Mistral-7B	(2a) Misled by the presence of emotional appeals and excuses in the text;	60.2%	0.0%
	(2b) Treating a binary or absolute statement as faulty generalization;	67.2%	6.5%
	(2c) Fail to handle cases involving implicit or indirect relationships between claims and evidence;	42.5%	2.3%
LLaMa3-70B	(3a) Lack of examples of anecdotal evidence or personal experiences;	38.2%	4.5%
	(3b) Linguistic structures such as conditional statements;	83.3%	0.0%
	(3c) Biased towards examples with more complex language or multiple sentences;	92.4%	24.1%

- *Main error types (numbered a,b,c)*
- *Different students (numbered 1,2,3)*

Contents

1 Overview

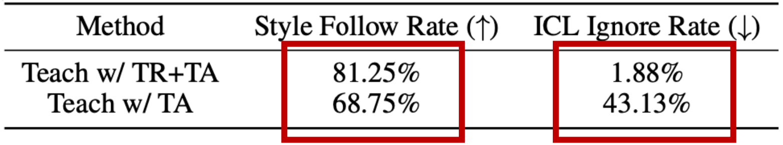
2 Method Design and Results

3 Broader Discussion

4 Summary

Broader Discussion

- **Insights into In-Context Learning (ICL)**
  - Prior work[1] found that a correct input-output pairing in ICL examples does not matter much.
  - Two key factors are important for successful ICL following, and thus establishing a positive correlation between the ICL example accuracy (TA accuracy / teacher score) and the answer accuracy (EA accuracy / LbT score / student score).
    - (1) Using Chain-of-Thought (i.e., detailed rationale) help ICL following.
      - **Style Follow Rate** measures whether EA matches the coding style of TA.
        - Provide rationale in the ICL example will get larger Style Follow Rate.
      - **ICL Ignore Rate** measures whether EA's style matches that of the code generated by the student without any ICL example.
        - Provide rationale in the ICL example will get lower ICL Ignore Rate.



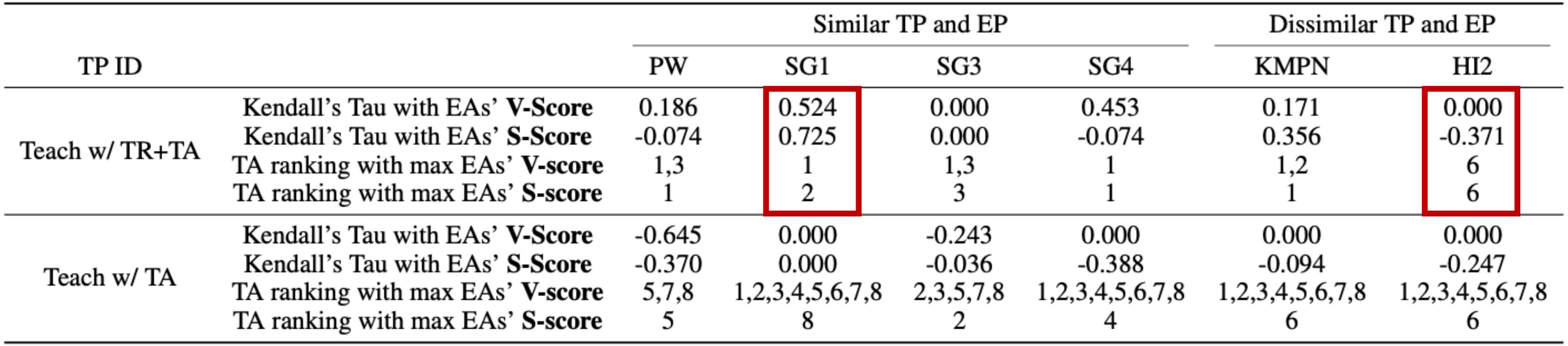
Method	Style Follow Rate (↑)	ICL Ignore Rate (↓)
Teach w/ TR+TA	81.25%	1.88%
Teach w/ TA	68.75%	43.13%

Comparing teaching with TP+TR+TA and teaching with TP+TA (without TR)
LLaMA3-8B (student/teacher). Game Theory **dataset, 5 problems, 8 TR-TAs for each problem**

[1] Sewon Min, et al. Rethinking the role of demonstrations: What makes in-context learning work? arXiv 2022.

Broader Discussion

- **Insights into In-Context Learning (ICL)**
  - Prior work[1] found that a correct input-output pairing in ICL examples does not matter much.
  - Two key factors are important for successful ICL following, and thus **establishing a positive correlation between the ICL example accuracy (TA accuracy / teacher score) and the answer accuracy (EA accuracy / LbT score / student score)**.
    - (2) TP and EP need to be similar.
      - When TPs and EPs are similar and TR is provided in the ICL example, the ranking correlation between the TA accuracy and EA accuracy is higher. And we can select high-accuracy TAs based on the S-score or V-score of EAs.



		Similar TP and EP				Dissimilar TP and EP	
TP ID		PW	SG1	SG3	SG4	KMPN	HI2
Teach w/ TR+TA	Kendall's Tau with EAs' **V-Score**	0.186	0.524	0.000	0.453	0.171	0.000
	Kendall's Tau with EAs' **S-Score**	-0.074	0.725	0.000	-0.074	0.356	-0.371
	TA ranking with max EAs' **V-score**	1,3	1	1,3	1	1,2	6
	TA ranking with max EAs' **S-score**	1	2	3	1	1	6
Teach w/ TA	Kendall's Tau with EAs' **V-Score**	-0.645	0.000	-0.243	0.000	0.000	0.000
	Kendall's Tau with EAs' **S-Score**	-0.370	0.000	-0.036	-0.388	-0.094	-0.247
	TA ranking with max EAs' **V-score**	5,7,8	1,2,3,4,5,6,7,8	2,3,5,7,8	1,2,3,4,5,6,7,8	1,2,3,4,5,6,7,8	1,2,3,4,5,6,7,8
	TA ranking with max EAs' **S-score**	5	8	2	4	6	6

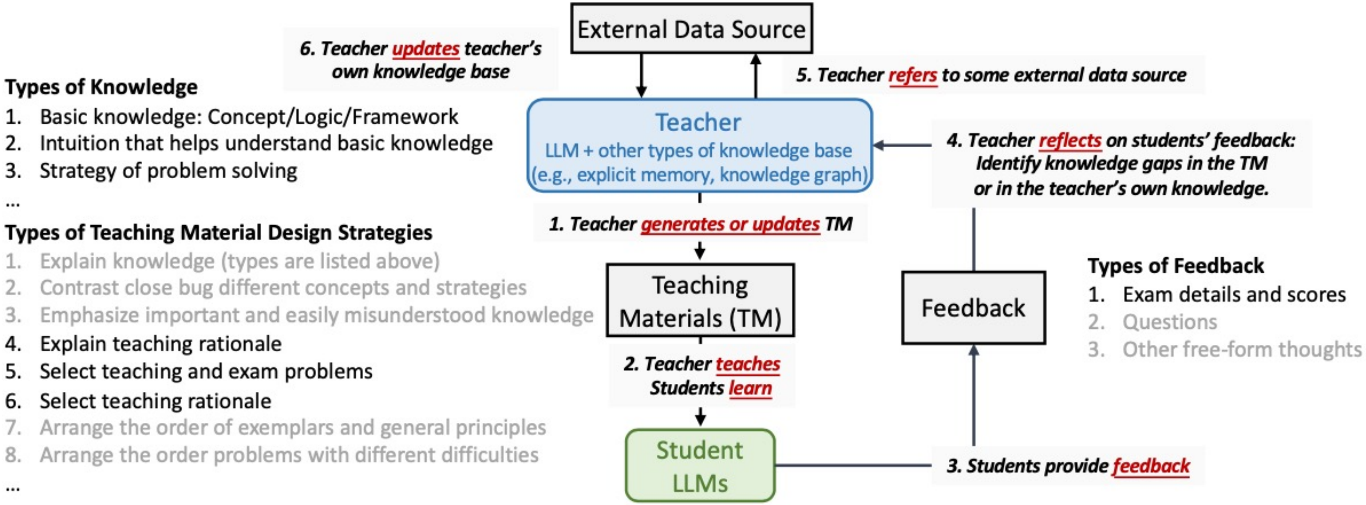
[1] Sewon Min, et al. Rethinking the role of demonstrations: What makes in-context learning work? arXiv 2022.

Broader Discussion

- **Limitations**
  - **The most impractical thing of M1 & M2: (1) Need similar EPs and (2) their GT answers. (3) Suitable EPs are selected according to human-provided similarity information.**
    - **Possible extension of M1**: (1) Can we synthesize similar problems? (2-1) Can we synthesize GT EP answers with the LLMs? => We made some attempts, not very successful. (2-2) Let the teacher give a score of the ER-EA (something like self-evaluation), instead of using GT EP answers. (3) Let a model automatically identify EPs similar to a TP from a large pool.
    - **Another "self-instruct" extension of M2**: Synthesize a new problem based on a group of problems that are already known to be similar, and has GT answers. Use the new problem as the TP, existing problems as the EPs, score the rationales of the new problem with the LbT score. *(Only need to synthesize similar problems, do not need to synthesize GT answers)*
  - **We only experiment with problems with oracle GT answer or test cases.**
    - LbT can be extended to open-ended problems, such as dialogue, writing, and open-ended math problems, maybe by letting the teacher evaluate a student's answer and provide the LbT score.
  - **Potential Risks of Bias Perpetuation.**
    - In open-ended problems where no GT judgment exists and LLM-based judgment is needed, it is possible that teaching materials deemed "well accepted" by students are not necessarily closer to the truth. Instead, they may align with the existing biases of teachers or students, posing a risk of the teacher perpetuating their own biases or indirectly learning the students' biases.

Broader Discussion

- **Borrowing Education Strategies to Improve LLMs**
  - **Borrowing the design strategies of teaching materials**.
  - **Borrowing the education pipelines** to design inference and training pipelines for LLMs.
    - (1) Task-oriented collaborative multi-agent learning
    - (2) Better LbT by configuring proper teacher and student
    - (3) Flexible teaching quality evaluation



Types of Knowledge

1. Basic knowledge: Concept/Logic/Framework
2. Intuition that helps understand basic knowledge
3. Strategy of problem solving

...

Types of Teaching Material Design Strategies

1. Explain knowledge (types are listed above)
2. Contrast close bug different concepts and strategies
3. Emphasize important and easily misunderstood knowledge
4. Explain teaching rationale
5. Select teaching and exam problems
6. Select teaching rationale
7. Arrange the order of exemplars and general principles
8. Arrange the order problems with different difficulties

...

Types of Feedback

1. Exam details and scores
2. Questions
3. Other free-form thoughts

External Data Source
6. Teacher updates teacher's own knowledge base
5. Teacher refers to some external data source
Teacher
LLM + other types of knowledge base (e.g., explicit memory, knowledge graph)
4. Teacher reflects on students' feedback: Identify knowledge gaps in the TM or in the teacher's own knowledge.
1. Teacher generates or updates TM
Teaching Materials (TM)
Feedback
2. Teacher teaches Students learn
Student LLMs
3. Students provide feedback

Contents

1 Overview

2 Method Design and Results

3 Broader Discussion

4 Summary

Summary

- **Target**: Improve LLM reasoning.
- **Idea**: Borrowing from human learning - the **Learning-by-Teaching** idea.
  - Benefit (b): Can assess the quality of the teaching material based on students' performance (the LbT-TMQ Assumption).
  - Benefit (c): The iterative feedback from diverse students can help identify ignored gaps.
- **Task**: mathematical reasoning, competition-level code synthesis, and verbal logical reasoning.
  - Require accurate knowledge and reasoning and cannot be easily solved through vague logic or simple memorization.
- **Implementation**:
  - **M1 & M2**: Implement an **LbT-based scoring component**, leveraging the LbT-TMQ assumption "teaching materials that make it easier for students to learn have clearer and more accurate logic". We use this scoring component in well-established pipelines.
  - **M3**: Implement **an iterative teaching & feedback process for prompt tuning**.
- **Some findings and possible potentials**:
  - M1 offers a new way of scaling up inference compute to obtain accuracy benefit.
  - Our results suggest LbT's potential for harnessing the diversity offered by different students and facilitating weak-to-strong generalization.
- **Roadmap of potential future research**: See Sec. 6 of the paper on borrowing educational concepts in improving LLMs.

*See Appendix D for more discussions on our research rationale in this project.

Results Summary

	M1	M2	M3
Implementation	- **Based on the LbT-TMQ assumption** - **Search-based output generation pipeline with LbT-based scoring**	- **Generating-scoring-finetuning pipeline with LbT-based scoring**	- **Let the LLM iteratively refine ICL examples by analyzing the students' feedback**
Task (Dataset): Results/Insights	- **Mathematical reasoning (MATH)**: 3.31% ~ 18.23% improvement over SC with the same number of rationales. 0.17%~8.29% improvement over SC with comparable or lower compute. - **Code synthesis (Leetcode problems)**: Notable improvements in LeetCode score.	**Mathematical reasoning (MATH)**: For LLaMA3-8B, the M2-tuned model achieves a 1.8% improvement over correctness-based DPO, on 500 MATH test problems.	**Verbal logical reasoning (Liar/Logic)**: - M3 can craft better ICL examples through multiple refinement rounds. - The feedback from students other than the teacher itself is beneficial.

Thank You !

Welcome to email me for discussing this work, or other collaborations focusing on LLM reasoning!

NICSEFC-EffAlg Team Website

http://nics-effalg.com/

Xuefei Ning*,Zifu Wang*, Shiyao Li*, Zinan Lin*, Peiran Yao*,

Tianyu Fu, Matthew B. Blaschko, Guohao Dai, Huazhong Yang, Yu Wang

foxdoraame@gmail.com

Paper

https://arxiv.org/abs/2406.14629

Code

https://github.com/imagination-research/lbt

Website

https://sites.google.com/view/llm-learning-by-teaching