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

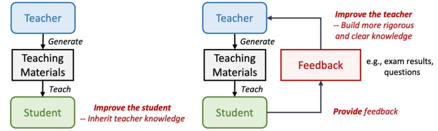
Xuefei Ning*[1], Zifu Wang*[2], Shiyao Li*[1,3], Zinan Lin*[4], Peiran Yao*[3,5], Tianyu Fu[1,3], Matthew B. Blaschko[2], Guohao Dai[6,3], Huazhong Yang[1], Yu Wang[1]

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

Two learning paradigms:

Learn from Teachers (LfT): Use the teacher to improve the student – *Widely explored, e.g., learn from manual labeling, learn from teacher model (knowledge distillation).*

Learn by Teaching (LbT): Use the student feedback to improve the teacher – *This work.*



Learning from Teacher (LfT) Learning by Teaching (LbT)

Why does LbT help?

(a) Increased self-accountability: Introduces social pressure and incentives.

(b) Explicit articulation of implicit and vague thoughts: When preparing teaching materials, the teacher needs to use clear language to convey inner thoughts. (M1 & M2)

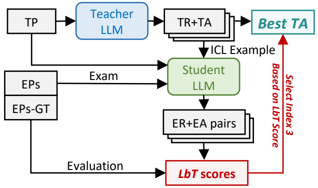
(c) Iterative feedback from diverse students: Interaction with students of varying ability levels and knowledge backgrounds offers valuable feedback. (M3)

TL;DR

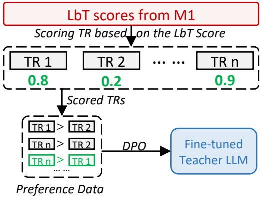
To improve the reasoning abilities of LLMs, we conduct a preliminary exploration of whether LLMs can "learn by teaching" (LbT). If so, we can:

- **Promote knowledge building and reasoning abilities of LLMs (LbT's benefits on human learning).**
- **Evolve stronger LLMs by having them teach weaker ones (weak-to-strong generalization).**

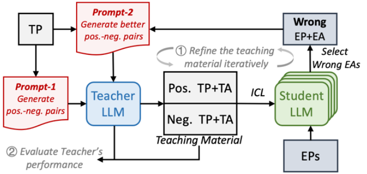
Method Level-1 (M1): Observing students' feedback



Method Level-2 (M2): Learning from the feedback



Method Level-3 (M3): Learning from the feedback iteratively



	M1	M2	M3
Method Details	**Based on the LbT-TMQ assumption:** Good teaching materials is easier for students to learn. => Use the students' exam performance on similar exam problems (EPs) to score the teacher's rationale (TR) & answer (TA) for the teaching problem (TP). Implement a ***search-based output generation pipeline*** with LbT-based scoring mechanism.	Implement a ***generating-scoring-finetuning pipeline*** with LbT-based scoring mechanism.	Implement an ***iterative prompt tuning process*** where the teacher LLM refines ICL exemplars by analyzing the students' failure cases.
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. **Insight: Using TR and ensuring similarity in TP and EP are crucial for successful ICL following.**	• **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.