Table 1: Results on 181 MATH test problems with 256 TR-TA pairs. The best results of each row are highlighted in green.

Teacher	Student	Greedy	SC	**M1** (MAX)	**M1** (SUM)
LLaMA3-70B	LLaMA3-8B	70.16	81.77	86.74	87.85
LLaMA3-70B	Mistral-7B	70.16	81.77	86.19	85.08
LLaMA3-70B	LLaMA3-8B & Mistral-7B	70.16	81.77	87.85	87.29
GPT-3.5	LLaMA3-8B	59.11	77.90	83.43	83.43
GPT-3.5	Mistral-7B	59.11	77.90	81.22	83.43
GPT-3.5	LLaMA3-8B & Mistral-7B	59.11	77.90	84.53	84.53
LLaMA3-8B	LLaMA3-8B	45.85	64.64	77.90	82.87
Mistral-7B	LLaMA3-8B	19.88	40.88	51.93	53.59

Table 2: Results on 181 MATH test problems with 128 TR-TA pairs, where standard errors are computed over 10 bootstrapping samples from 256 TR-TA pairs. The best results of each row are highlighted in green.

Teacher	Student	Greedy	SC	**M1** (MAX)	**M1** (SUM)
LLaMA3-70B	LLaMA3-8B	70.16	81.05 ± 0.61	86.02 ± 0.86	87.46 ± 0.82
LLaMA3-70B	Mistral-7B	70.16	81.05 ± 0.61	85.58 ± 1.06	86.13 ± 0.63
LLaMA3-70B	LLaMA3-8B & Mistral-7B	70.16	81.05 ± 0.61	87.13 ± 1.05	87.07 ± 0.56
GPT-3.5	LLaMA3-8B	59.11	77.18 ± 0.56	83.43 ± 0.65	83.04 ± 0.82
GPT-3.5	Mistral-7B	59.11	77.18 ± 0.56	81.71 ± 1.25	82.98 ± 1.07
GPT-3.5	LLaMA3-8B & Mistral-7B	59.11	77.18 ± 0.56	84.25 ± 1.08	82.82 ± 0.72
LLaMA3-8B	LLaMA3-8B	45.85	64.31 ± 0.79	77.46 ± 1.30	79.89 ± 1.42
Mistral-7B	LLaMA3-8B	19.88	38.45 ± 0.51	48.84 ± 2.18	51.44 ± 1.03

Table 3: Results on 500 MATH test problems with greedy decoding, where standard errors are computed over 3 repeated runs.

Teacher/Student	Original	Correctness-DPO	**M2**
LLaMA3-8B	29.0	30.2 ± 0.43	31.8 ± 0.59

Table 4: 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%

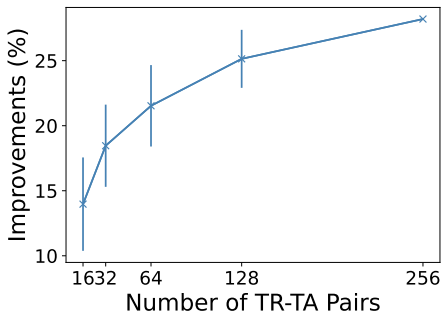

Figure 1: Relative improvements of **M1** over SC using LLaMA3-8B as the teacher and the student on 181 MATH test problems with different number of TR-TA pairs, where standard errors are computed over 10 bootstrapping samples from 256 TR-TA pairs.