Evaluating Mathematical Reasoning Beyond accuracy

Shijie Xia[1,2,5], Xuefeng Li[1,2,5], Yixin Liu[4], Tongshuang Wu[3], Pengfei Liu[1,2,5]

[1] Shanghai Jiao Tong University [2] Shanghai Artificial Intelligence Laboratory

[3] Carnegie Mellon University [4] Yale University [5] Generative AI Research Lab (GAIR)

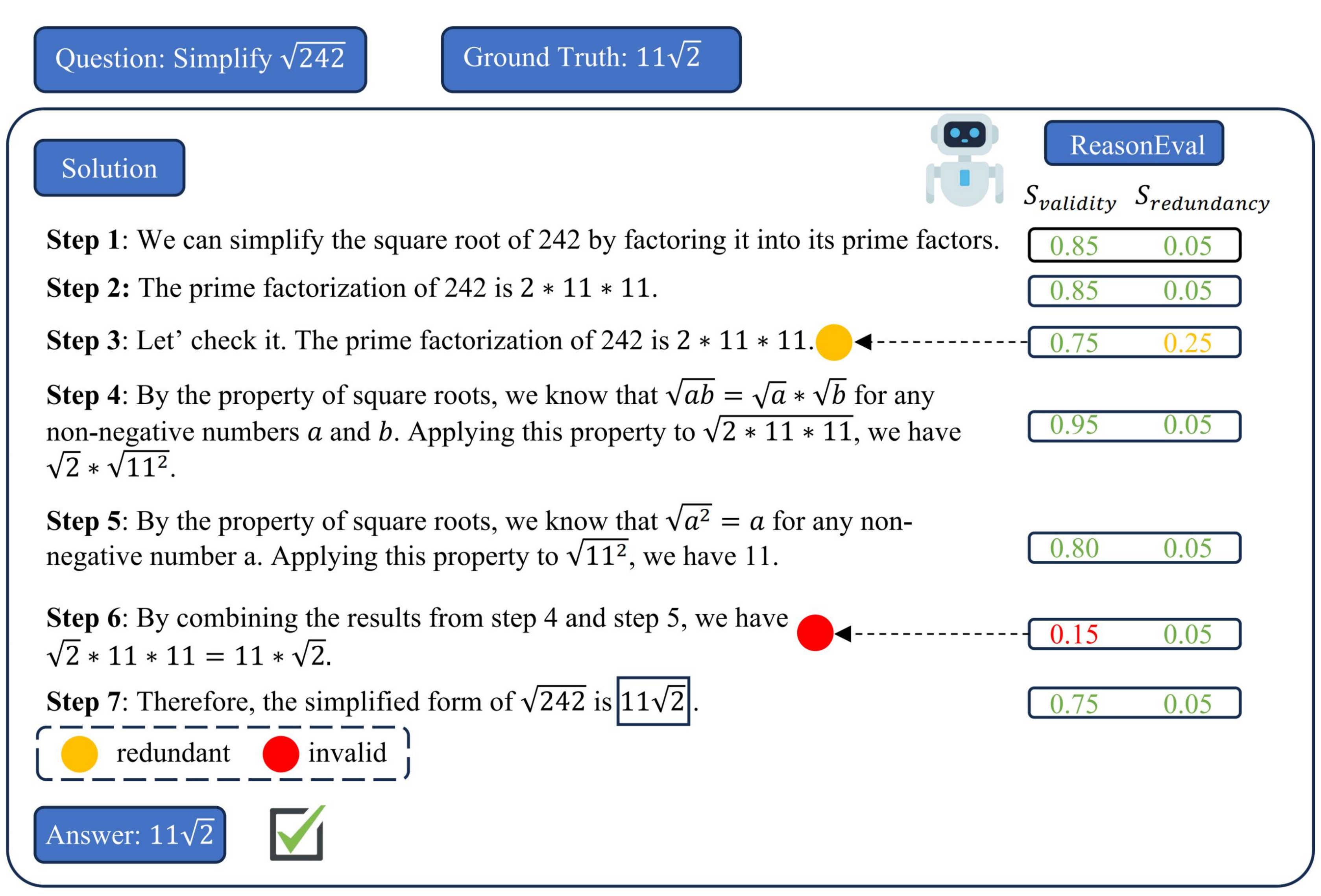
Introduction

The leaderboard of Large Language Models (LLMs) in mathematical tasks has been continuously updated. However, the majority of evaluations focus solely on the final results, neglecting the quality of the intermediate steps. This oversight can mask underlying problems, such as logical errors or unnecessary steps in the reasoning process. To measure reasoning beyond final-answer accuracy, we introduce REASONEVAL, a new methodology for evaluating the quality of reasoning steps. REASONEVAL employs validity and redundancy to characterize the reasoning quality, as well as accompanying LLMs to assess them automatically.

Methodology

Validity: the step contains no mistakes in calculation and logic

Redundancy: the step lacks utility in solving the problem but is still valid

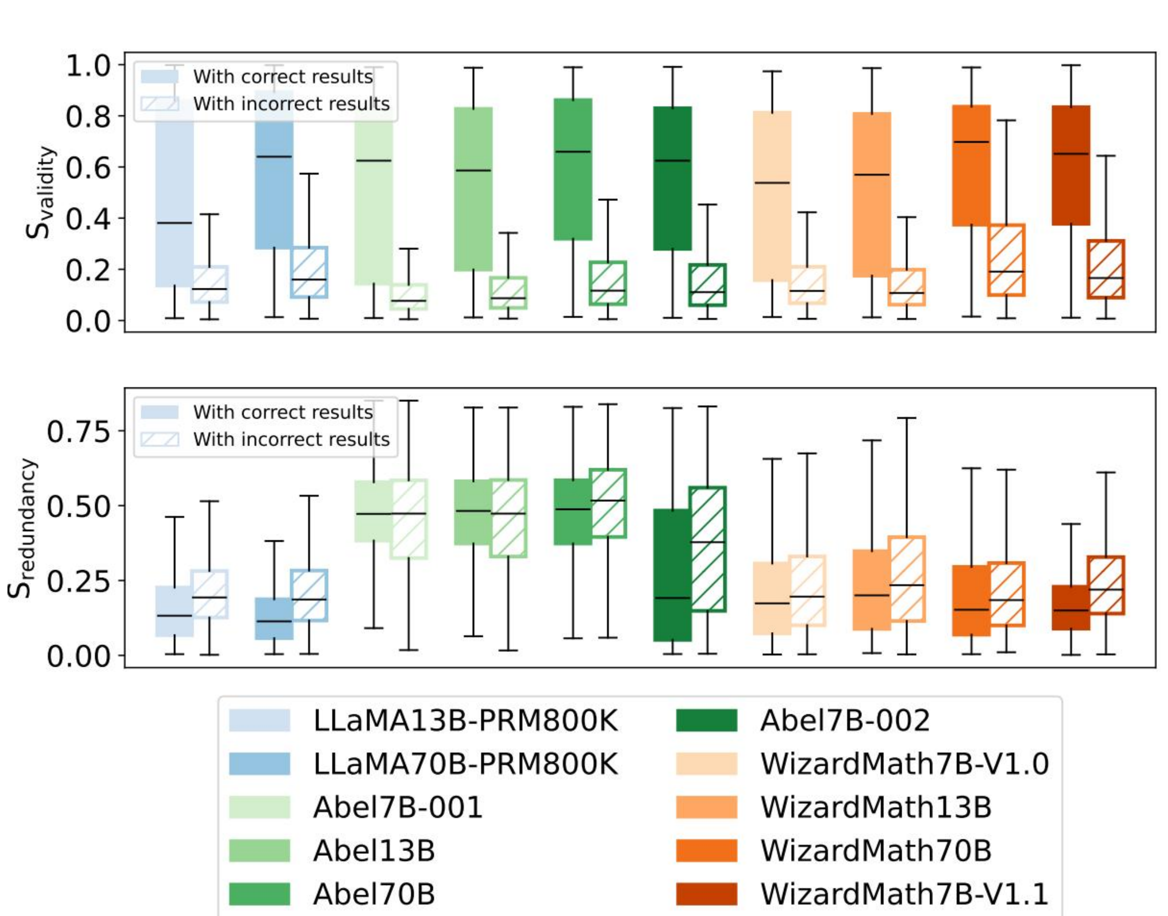


Meta Evaluation

	MR-MATH-invalid				MR-MATH-redundant			
	Solution-level		*Step-level*		*Solution-level*		*Step-level*	
	F1 Score	**AUC**	**F1 Score**	**AUC**	**F1 Score**	**AUC**	**F1 Score**	**AUC**
Embedding-based Methods								
ROSCOE-SA	48.2	57.5	-	-	50.7	53.9	-	-
ROSCOE-SS	51.6	49.6	-	-	52.0	52.7	-	-
Prompting-based Methods								
GPT-3.5-turbo	59.7	-	53.2	-	53.0	-	51.5	-
GPT-4	73.2	-	61.0	-	57.1	-	54.2	-
Step-level Evaluators								
Math-shepherd-Mistral-7b	70.1	77.3	60.0	77.2	50.4	54.5	42.7	53.0
$\text{REASONEVAL}_{\text{Llama2-7B}}$	66.7	79.5	60.8	80.0	60.4	62.8	59.0	68.6
$\text{REASONEVAL}_{\text{WizardMath-7B-V1.0}}$	72.8	81.9	67.7	83.9	60.5	**65.6**	59.0	68.3
$\text{REASONEVAL}_{\text{Mistral-7B}}$	78.0	85.1	68.6	85.7	60.7	63.4	59.7	70.9
$\text{REASONEVAL}_{\text{Llemma-7B}}$	74.7	84.3	76.6	90.5	59.6	63.0	58.6	68.3
$\text{REASONEVAL}_{\text{Abel-7B-002}}$	77.3	86.2	70.4	90.5	58.6	63.6	59.5	71.8
$\text{REASONEVAL}_{\text{WizardMath-7B-V1.1}}$	78.6	87.5	73.9	89.5	**61.6**	64.8	**59.7**	**72.2**
$\text{REASONEVAL}_{\text{Llemma-34B}}$	**79.6**	**90.8**	**77.5**	**92.8**	58.3	62.7	57.5	67.3

		MR-GSM8K-original				MR-GSM8K-reversed			
		Solution-level		*Step-level*		*Solution-level*		*Step-level*	
	OOD	**F1 Score**	**AUC**	**F1 Score**	**AUC**	**F1 Score**	**AUC**	**F1 Score**	**AUC**
Embedding-based Methods									
ROSCOE-SA	✓	51.6	54.4	-	-	54.5	57.9	-	-
ROSCOE-SS	✓	49.6	60.1	-	-	49.6	52.1	-	-
Prompting-based Methods									
GPT-3.5-turbo	-	54.9	-	52.3	-	54.3	-	49.9	-
GPT-4	-	81.7	-	69.0	-	72.2	-	52.2	-
Step-level Evaluators									
Math-shepherd-Mistral-7b	✗	**86.0**	**93.9**	73.4	88.5	**77.2**	**88.0**	59.6	77.9
$\text{REASONEVAL}_{\text{Mistral-7B}}$	✓	61.8	79.8	62.9	86.1	61.0	71.9	61.5	84.3
$\text{REASONEVAL}_{\text{WizardMath-7B-V1.1}}$	✓	74.1	90.7	72.8	**91.4**	74.4	86.3	**70.5**	**90.5**
$\text{REASONEVAL}_{\text{Llemma-34B}}$	✓	81.0	88.1	**73.5**	86.8	76.1	84.1	69.3	85.0

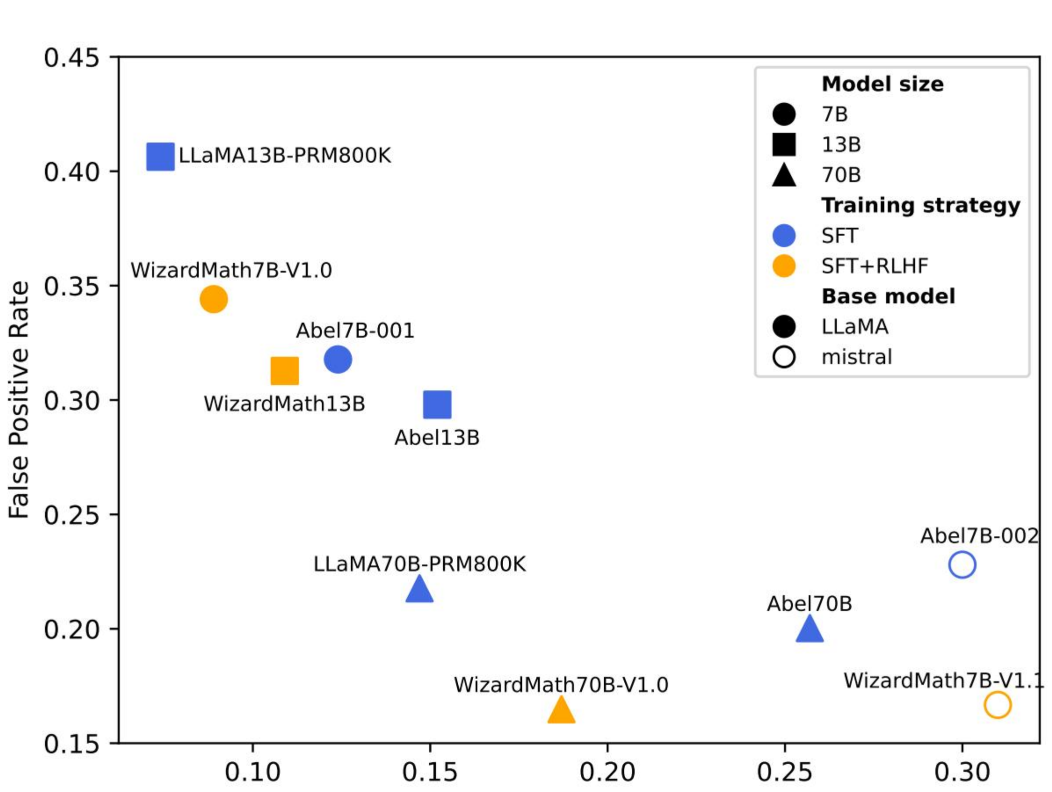
Evaluating Reasoning Quality of LLMs



Model	Acc. (%)	FPR (%)
LLaMA2-13B-PRM800K	7.4	40.6
LLaMA2-70B-PRM800K	14.7	21.8
Abel7B-001	12.4	31.8
Abel13B	15.2	29.8 (29.2)
Abel70B	25.7	20.0
Abel7B-002	30.0	22.8
WizardMath7B-V1.0	8.9	34.4
WizardMath13B	10.9	31.3 (28.3)
WizardMath70B	18.7	16.5
WizardMath7B-V1.1	31.0	16.7

Findings

- An improvement in the result accuracy is not sufficient to ensure an enhancement in the overall quality of reasoning steps in challenging mathematical problems.
- The model scale, the base model, and the training methods have significantly influenced the quality of reasoning steps.
- When a model is unsure about how to solve a problem, it tends to make more attempts that lack meaningful progression.



Data Selection

REASONEVAL can select high-quality training data to improve the efficiency of solving problems and the quality of solutions.

Filter	#D	Acc.	Val.	Red.	#Token
-	100%	22.2	65.2	27.4	723.4
val.	76.7%	22.0	65.9	26.4	699.9
random	76.7%	20.1	62.5	27.4	765.6
red.	71.9%	21.8	65.6	22.1	681.5
random	71.9%	20.3	62.3	28.0	746.1
red. & val.	56.7%	22.0	67.8	22.5	701.2
random	56.7%	20.0	62.1	27.6	739.5

Resource

Code: https://github.com/GAIR-NLP/ReasonEval

Model:
https://huggingface.co/GAIR/ReasonEval-7B
https://huggingface.co/GAIR/ReasonEval-34B

Email: shijiexia@sjtu.edu.cn

Twitter: ShijieX60925