香港中文大學
The Chinese University of Hong Kong

MR-GSM8K: A Meta Reasoning Benchmark for LLM Evaluation

Zhongshen Zeng

The department of Computer Sci & Erg

The Chinese University of Hong Kong

Supervisors: Prof Yu Bei & Prof. Jia JiaYa

Background & Motivations

- Current Evaluation Methods are limited
  - Results oriented evaluation (e.g. Accuracy Metrics)
    - The final computation results can be generated via flawed reasoning paths
    - The scores can be inflated easily via data contamination
    - The design philosophy makes scaling benchmark difficult
  - Saturations of benchmarks
    - Most well-recognized reasoning benchmarks are saturated

	Open source	Chinese General	English General	Knowledge	Arithmetic	Math	Reasoning	Coding
		AlignBench	MT-Bench	MMLU	GSM8K	MATH	BBH	HumanEval
DeepSeek-V2	**Yes**	**7.89**	**8.85**	**80.6**	**94.8**	**71.0**	**83.4**	**84.8**
GPT-4-Turbo-1106	-	8.01	9.32	84.6	93.0	64.1	-	82.2
GPT-4-0613	-	7.53	8.96	86.4	92.0	52.9	83.1	84.1
GPT-3.5	-	6.08	8.21	70.0	57.1	34.1	66.6	48.1

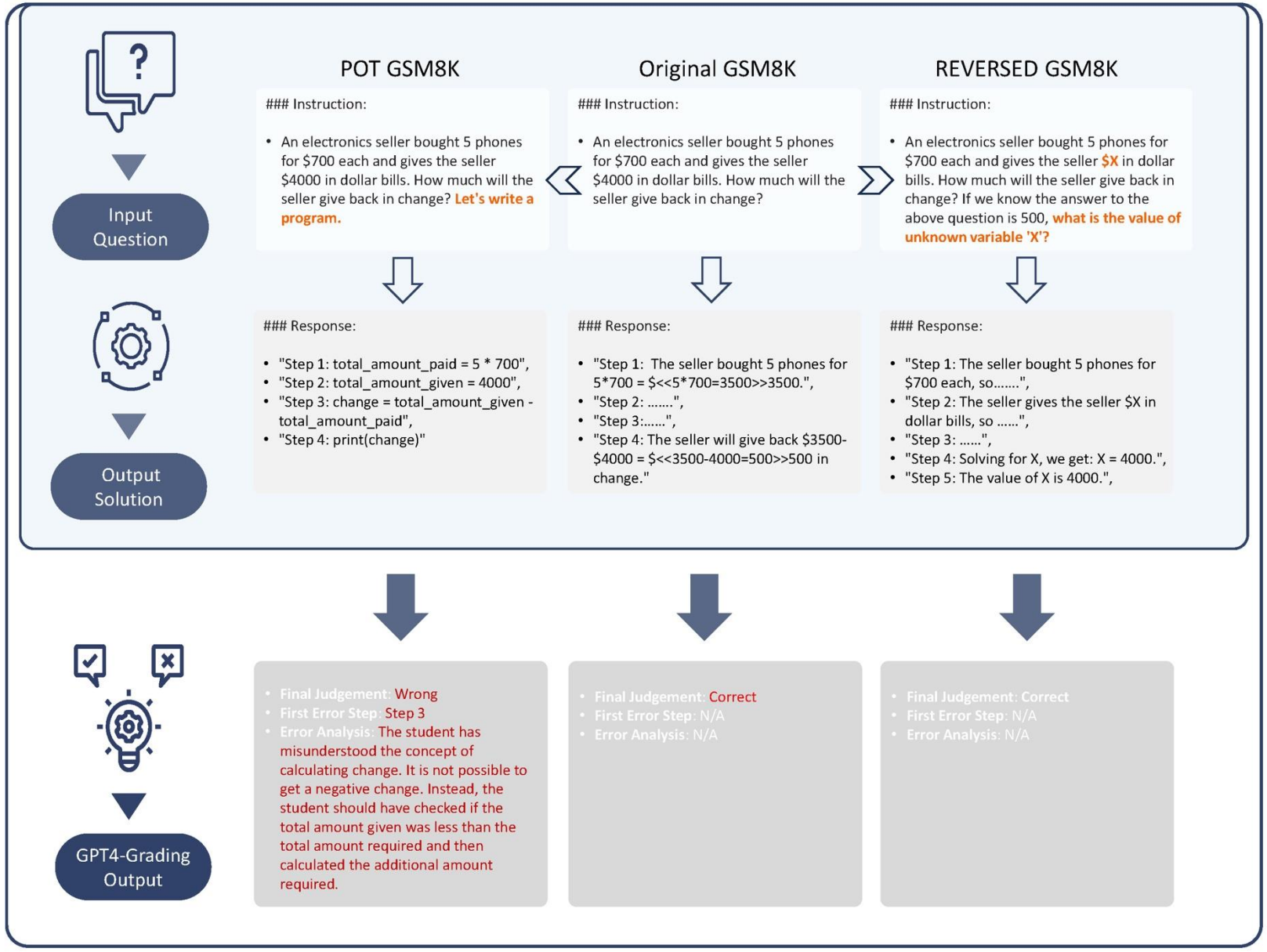
Meta-Reasoning

- Process Oriented Evaluation: Meta-Reasoning
  - From question-answering to solution-scoring
  - Switch from the role of student to that of teachers
  - Requires reason about reasonings, thus termed Meta-Reasoning
- Why Meta-Reasoning
  - Shifts the focus from computation results to the computation process
  - Robust against data-contamination and memorization
  - Inherently fitted for system-2 thinking evaluation
  - Relatively easy to scale the benchmarks
  - Far from saturations

MR-GSM8K: Dataset Construction

1. For each GSM8k problems, we collect:
   - original problem + CoT answer
   - reversed problem + CoT answer
   - original problem + PoT answer

2. Trained experts annotate the error analyses:
   - Answer Correctness
   - First error step
   - Error reason
   - Step correction



MR-Score

- LLMs will be given a question & candidate solution, and then:
  - Determine the solution correctness
  - If deemed incorrect, find the first error step and explain the error reason
- MR-Score: weighted metrics over three subtask performances
  - Accuracy of Solution Correctness: $MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}}$
    - Matthews Correlation Coefficient
  - Accuracy of locating first error step $ACC_{\text{step}} = \frac{N_{\text{correct\_first\_error\_step}}}{N_{\text{incorrect\_sols}}}$ $ACC_{\text{reason}} = \frac{N_{\text{correct\_error\_reason}}}{N_{\text{incorrect\_sols}}}$
  - Accuracy of explaining the error reason
  - MR-Score: $MR\text{-}Score = w_1 * \max(0, MCC) + w_2 * ACC_{\text{step}} + w_3 * ACC_{\text{reason}}$
- Evaluation:
  - Solution correctness and incorrect first error step can be calculated automatically
  - Error reason explanation requires humans or machine scoring
  - 92 percent author-model agreement rates

Experiments

- O1 series of models are leading, showcasing the benefits of self-reflective long-CoT reasoning pattern
- Small models like Phi-3 are comparable to 70B models, indicating the importance of data quality and diversity

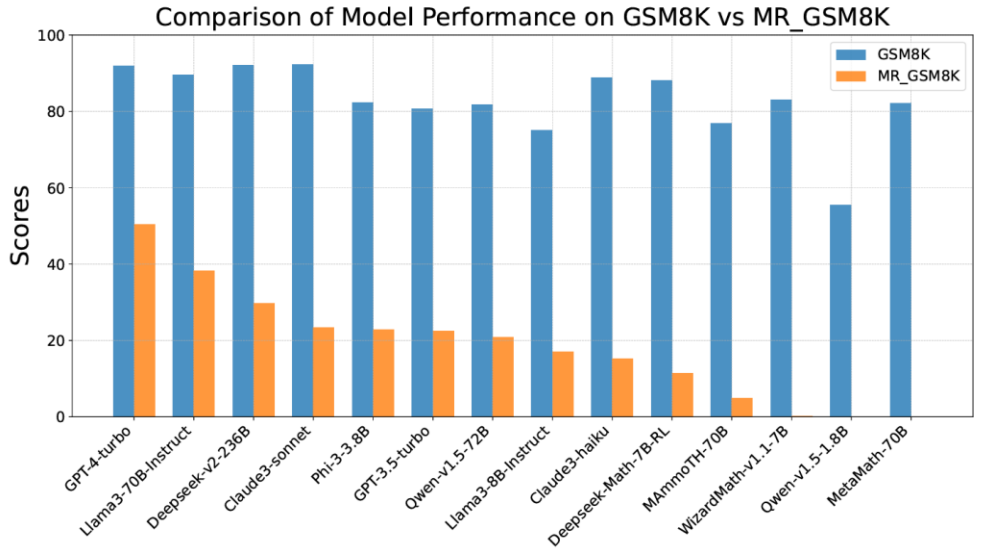
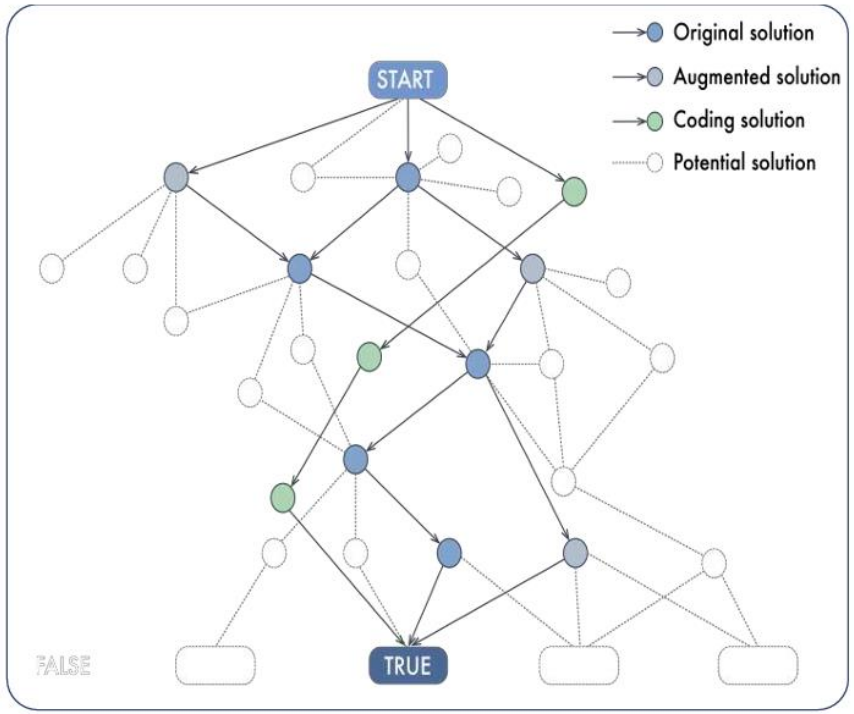


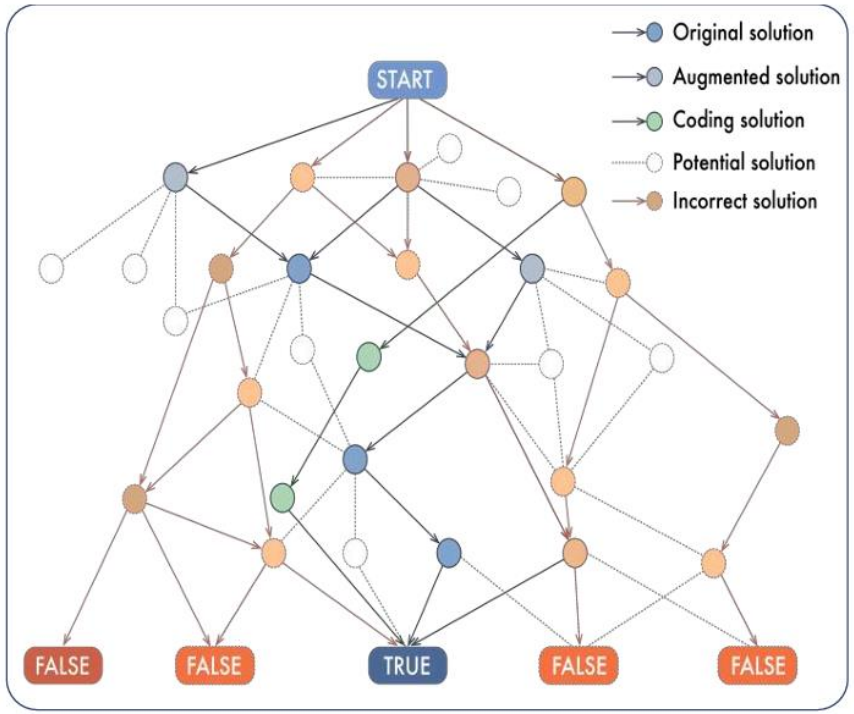
Table 2: Evaluation results on MR-GSM8K. SC-TPR and SC-TNR stand for the true positive and true negative rate for the solution correctness determination. K stands for number of demos in our prompts and bold number indicates the best performance within the corresponding model groups.

Model	SC-TPR		SC-TNR		MCC		ACC_{step}		ACC_{reason}		MR-Score	
	k=0	k=3	k=0	k=3	k=0	k=3	k=0	k=3	k=0	k=3	k=0	k=3
Open-Source Small												
Qwen-1.8B	21.8	33.3	0.1	3.9	0.	0.	0.	0.4	0.	0.	0.	0.1
Phi3-3.8B	11.3	62.6	98.5	72.6	20.4	35.4	32.9	26.3	18.0	13.9	**22.9**	21.9
Open-Source Medium												
Deepseek-Math-7B-RL	77.3	2.4	52.3	0.4	30.4	0.	9.8	0.1	5.1	0.1	11.6	0.1
WizardMath-v1.1-7B	99.3	6.7	0.5	0.6	0.0	0.0	0.3	0.2	0.3	0.1	0.2	0.1
Llama3-8B	3.2	40.9	98.3	80.3	5.1	23.1	29.1	23.3	15.0	11.6	17.2	**17.4**
Open-Source Large												
MAmmoTH-70B	88.0	89.8	23.1	2.8	14.6	0.0	3.9	0.3	1.8	0.3	5.0	0.2
MetaMath-70B	7.8	0.0	0.3	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0
Llama3-70B	67.6	89.3	83.0	66.0	51.3	56.4	38.9	33.5	32.7	25.7	**38.3**	34.2
Qwen1.5-72B	83.7	87.7	57.1	52.4	42.0	42.5	19.1	23.1	13.5	15.8	20.9	23.3
Deepseek-v2-236B	60.1	88.2	87.2	61.5	49.4	51.2	26.8	32.4	23.8	28.3	29.8	34.1
Closed-Source LLMs												
Claude3-Haiku	70.4	99.0	51.7	8.1	22.5	16.7	17.2	2.3	11.3	1.8	15.3	4.9
GPT-3.5-Turbo	16.3	59.7	93.8	65.7	16.2	25.5	30.6	21.0	20.3	13.0	22.6	17.9
Claude3-Sonnet	35.1	88.4	89.8	44.8	30.0	36.5	25.2	18.8	19.9	15.6	23.5	20.8
GPT-4-Turbo	69.5	83.0	91.8	84.2	63.3	67.2	48.8	51.7	46.3	48.1	50.5	53.0
o1-mini-2024-09-12	93.3	93.3	95.6	94.8	89.0	88.1	67.6	67.6	62.2	61.8	69.2	68.8
o1-preview-2024-09-12	89.3	84.4	96.8	95.6	86.6	80.8	68.3	69.5	65.7	66.6	**70.7**	70.3

Meta-Reasoning Summary

- Meta-Reasoning and System-2 Thinking
  - More holistic and comprehensive coverage on solution space
  - Requires examination of assumptions/conditions/logic and even infer counterfactually
  - Exposes the shortcoming of current training pipeline
  - O1 style long-CoT data contains self-reflective, branching patterns





Thanks

Q & A