Enhancing Multi-Step Reasoning in Language Models with Synthetic Math Data Augmentation

Team: HP_Fighters

Presenter: Jieqing Mei

Candidate solutions for enhancing math reasoning

- Construct reasoning dataset
- Knowledge distillation[1]
- Construct Knowledge Base for math problem solving[2][3]
- Leverage tool to enhance inference performance

[1] Li, Junlong, et al. "CodeI/O: Condensing Reasoning Patterns via Code Input-Output Prediction." arXiv preprint arXiv:2502.07316 (2025).
[2] Wang, Zeqing, et al. "Towards top-down reasoning: An explainable multi-agent approach for visual question answering." arXiv preprint arXiv:2311.17331 (2023).
[3] Li, Hang, et al. "Knowledge tagging system on math questions via llms with flexible demonstration retriever." arXiv preprint arXiv:2406.13885 (2024).

Problem of existing math datasets

- Existing Math Datasets (contain CoT / PoT)
  BelleGroup/school_math_0.25M
  TIGER-Lab/MathInstruct

Q

```
Sides of a rectangular park are in the ratio 3: 2 and its area is 3750
sq m, the cost of fencing it at 50 ps per meter is?
Answer Choices: (A) Rs.122 (B) Rs.129 (C) Rs.125 (D) Rs.120 (E) Rs.121
```

A

```
Let's solve the multi-choice question step by step. 3x * 2x = 3750 => x
= 25 2(75 + 50) = 250 m 250 * 1/2 = Rs.125
The answer is C
```

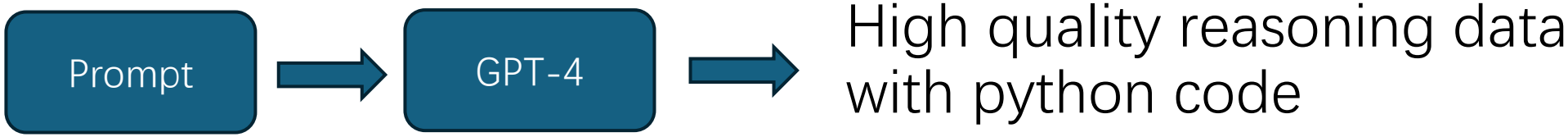
Problem 1 : lack of clear multi-step reasoning process

Problem 2 : calculate only with model's own generalization ability

Solution

- Multi-step reasoning: simulate the human thinking process
- Utilize Python code as computational tool

An example of Data Synthesis



Three implemented approaches for data synthesis

- **Using Prompt Engineering to Generate Math Questions**

  Pros: Generate math problems, reasoning steps efficiently from scratch.
  Cons: Less diverse data with suboptimal quality, requiring manual review which remains time-consuming.

- **Few-Shot Learning & Manual Verification**

  Pros: Ensures diverse, high-quality math problems.
  Cons: Labor-intensive and not scalable for large datasets.

- **Automatic Data Synthesis**

  Pros: Enables automated generation while ensuring answer accuracy through predefined datasets.
  Cons: Limited by scarcity of high-quality math datasets and depended on existing datasets.

Results

model	math_100	gsm8k_1319	gsm8k_ja_1069
base_model	3(3%)	324(24.6%)	278(26%)
our model (submitted, with vllm inference)	5(5%)	230(17.4%)	171(16%)
our model (local, with hg inference + python execution)	-	383(29%)	-

Approaches in progress

- Distill knowledge from SOTA open source LLM

  Soft label generation is required, resulting in significant time costs

- Translate high quality math reasoning dataset

  Demand amount of time to perform the translation task

- Build a RAG component aimed at solving similar math problem

Thanks for your attention