GSM-PLUS: A Comprehensive Benchmark for Evaluating the Robustness of LLMs as Mathematical Problem Solvers

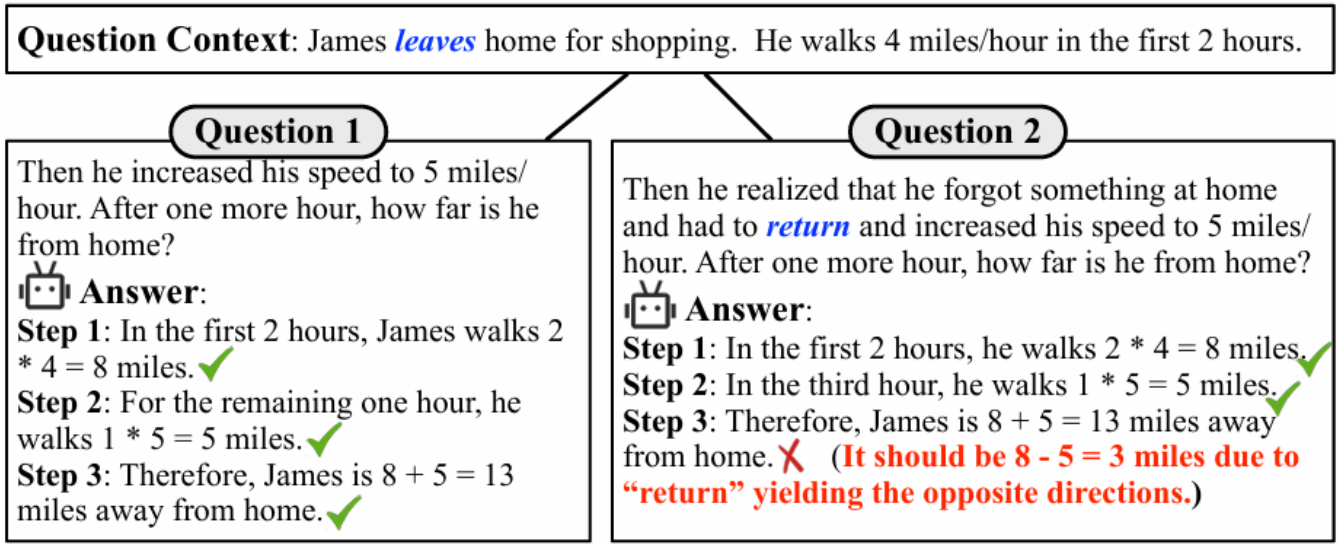
Qintong Li, et al

The University of Hong Kong & Tencent AI Lab

July 2, 2024

Abstract

- Does LLM truly understand math, or simulate understanding?
  - Even slight variations in problem can lead to confusion for LLMs.
  - Enhancing the robustness of LLMs is essential.
- We propose GSM-PLUS, which is more challenging than GSM-8K.
  - Highlighting the weaknesses of LLMs
  - Identifying solutions to address these weaknesses



Introduction

- GSM-PLUS is 5 perturbations of GSM-8K
  - Numerical Variation
  - Arithmetic Variation
  - Problem understanding
  - Distractor insertion
  - Critical thinking
- With perturbation, there is no difference in difficulty
  - Human performance is unaffected
  - LLMs showed a gap up to 20% in accuracy

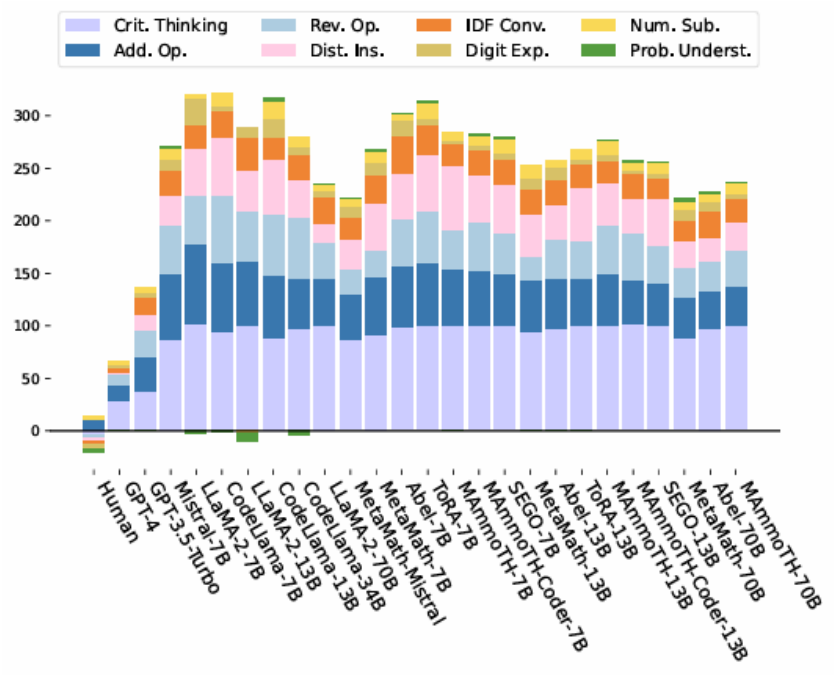
Perturbation Category

Seed Question: Janet's ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder at the farmers' market daily for $2 per fresh duck egg. How much in dollars does she make every day at the farmers' market?
Solution: Janet sells 16 - 3 - 4 = 9 duck eggs a day. She makes 9 * 2 = 18 every day at the farmer's market. Answer: 18

Perturbation Category		Question Variation
Numerical Variation	Num. Sub.	16 → 20 three → five four → six 2 → 3
	Digit Exp.	16 → 1600 four → 400
	IDF Conv.	three → 1/4 2 → 2.5
Arithmetic Variation	Add. Op.	Janet's ducks lay . . . every day with four. She also uses two eggs to make a homemade hair mask every day. She sells . . . make every day at the farmers' market?
	Rev. Op.	Janet's ducks lay 16 eggs per day. She eats three . . . with four. She sells the remainder at the farmers' market daily for a certain amount per fresh duck egg. She makes $18 every day at the farmers' market. How much does each duck egg cost?
Problem Understanding		Janet's ducks lay 16 eggs daily. She eats three for breakfast and uses four to bake muffins for her friends. She sells the remaining eggs at the local farmers' market for $2 per fresh duck egg. How much money does she make each day by selling eggs at the farmers' market?
Distractor Insertion		Janet's ducks . . . with four. She also uses two eggs to feed her pet parrot, but her neighbor gives her two eggs from his own ducks to replace them. She sells . . . at the farmers' market?
Critical Thinking		Janet's ducks lay eggs per day. She eats three for breakfast every morning and . . . How much in dollars does she make every day at the farmers' market?

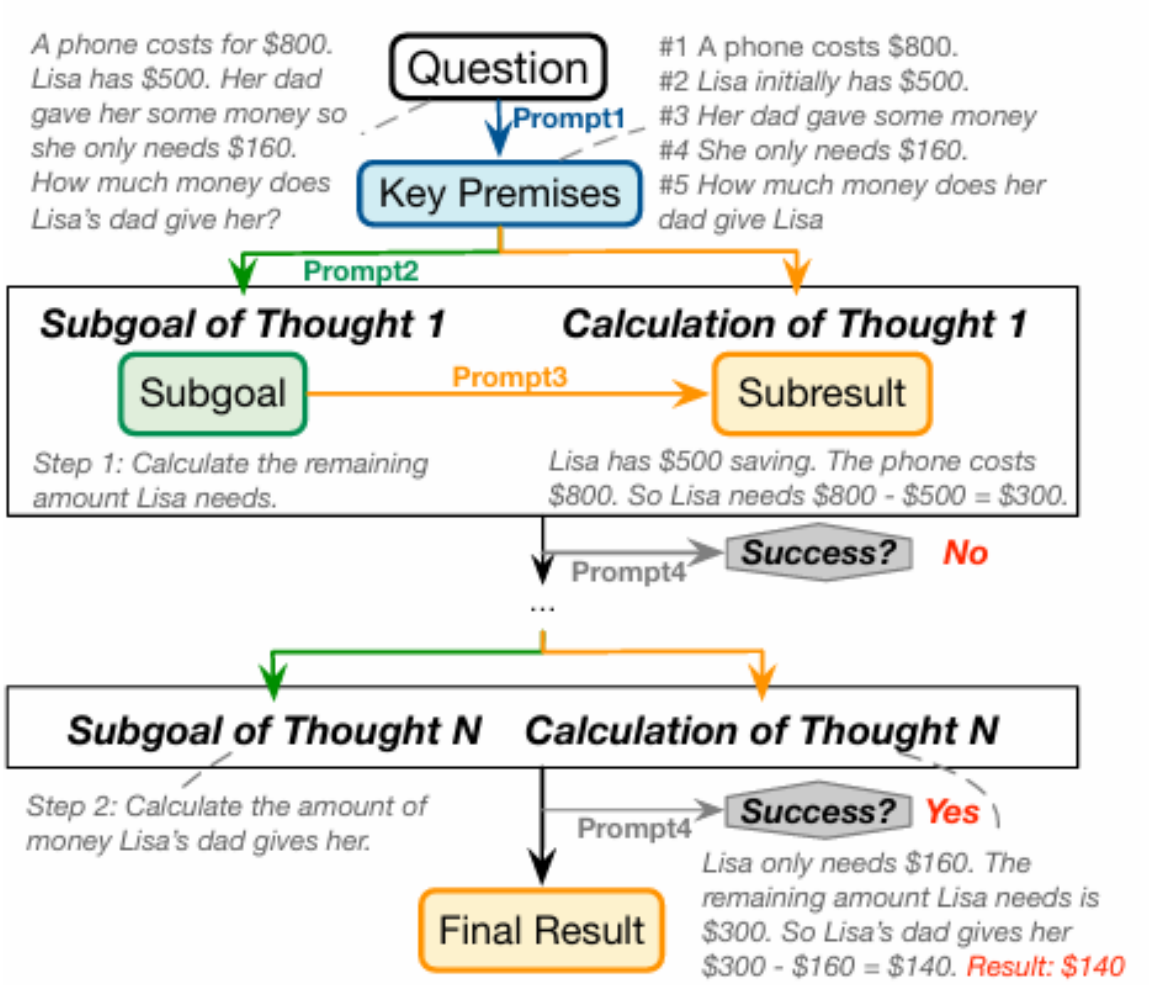
Experiment

- Experimental Setup
  - Test for models like GPT-4, LLaMA-2, CodeLlama, etc.
  - Compare results in GSM-8K and GSM-PLUS
  - Observe Performance Drop Rate (PDR)
- Overall Results
  - Quite robust to Numerical Variation, Problem understanding
  - Vulnerable to other variations



Solution for Robustness

- Compositional Prompting
  - Iteratively decompose complex problems



Conclusion

- Most LLMs showed large PDR in GSM-PLUS
  - Models need to be robust to minor variations
  - Compositional Prompting partially solved the problem

- Limitation
  - Training over elementary school level is required.
  - Only accuracy of the answer, not the solution chain.
  - We don’t know underlying reason of LLMs’ failure.