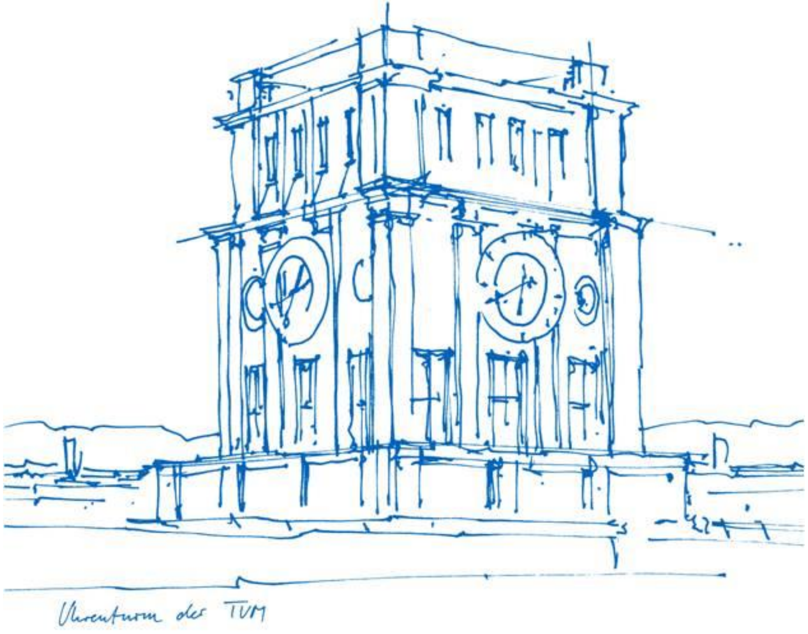
Chair of Computational Imaging and AI in Medicine
TUM School of Computation, Information and Technology
Technical University of Munich

Premise Order Matters in Reasoning with Large Language Models

Presenter: Tuna Karacan

July 02, 2025

Table of Contents



Introduction

- Large language models (LLMs) have demonstrated **impressive performance in reasoning tasks**.
- Even surpassing humans on multiple reasoning tasks, including **STEM problems and code generation**.
- However, LLMs exhibit failure when **distracted or can't form a backwards connection**.
- In this work the **effect of premise ordering on LLM performance** was investigated.

Introduction

Logical Reasoning

- If $\boldsymbol{A}$, then $\boldsymbol{B}$,
- If $\boldsymbol{B}$, then $\boldsymbol{C}$,
- $\boldsymbol{A}$ is true.

- A reasonable human would infer $\boldsymbol{B}$ and $\boldsymbol{C}$ are true as well, regardless of the premise ordering.
- This is called a $\boldsymbol{modus\ ponens}$ which is relatively straight forward for human beings.
- However, LLMs struggle when the **ordering does not match the ground truth ordering**.

Introduction

R-GSM for Mathematical Reasoning

- X has **10** apples,
- They purchase **2** more apples before arriving at school,
- They share **half** their apples after arriving at school,
- How many apples does X have?

- **Grade school math problems** were also tested to confirm and further investigate the premise ordering on LLMs.
- Similar to logical reasoning, the performance of LLMs drop significantly when the **order is different than that of the temporal order**.

Benchmarks

Logical Reasoning

- Each problem is sampled with **SimpleLogic** and consists of 3 parts:
  - Rules in the form of 'If $\boldsymbol{X_1} \dots \boldsymbol{X_n}$ then $\boldsymbol{Y_m}$',
  - Facts $\boldsymbol{A_1} \dots \boldsymbol{A_n}$ that hold **true**,
  - A conclusion '$\boldsymbol{C}$ is **true**' that needs to be proved.

- Each problem has **4-12** rules along with **0**, **5** or **10** distracting rules which are not used in the proof and **5** different ordering of rules.
- **200** different questions are generated and with all of their variants there are **27K** questions in total.

Benchmarks

Logical Reasoning

- Different rule orders are calculated using the **Kendall τ distance** which is then normalized into the **[-1, 1]** range:
  - **1** $\Longrightarrow$ Complete forward order,
  - **-1** $\Longrightarrow$ Complete backward order,
  - **0** $\Longrightarrow$ Complete random order.

- All questions have variants with **τ distances** of **[-1, -0.5, 0, 0.5, 1]**.

Benchmarks

R-GSM for Mathematical Reasoning

- Questions from the **GSM8K** dataset were chosen based on the following rules:
  - Has at least **5** sentences in the problem definition,
  - Different ordering causes an **LLM prediction failure**,
  - Different ordering **does not alter the ground truth**.

- In total **220** question pairs are generated, including the original GSM8K problem description and a rewritten one for grammar and context flow reasons.

Experiments

Experimental Setup

- The problems are evaluated on **GPT-4-turbo**, **GPT-3.5-turbo**, **PaLM 2-L** and **Gemini 1.0 Pro**.
- R-GSM problems have **no additional instruction** other than the problem and logical reasoning problems include an instruction that asks for the **derivation that specifies which premise is used in each step**.
- Decoding is done greedily with **0 temperature** and **zero-shot prompting** is applied in all experiments.

Experiments

Logical Reasoning

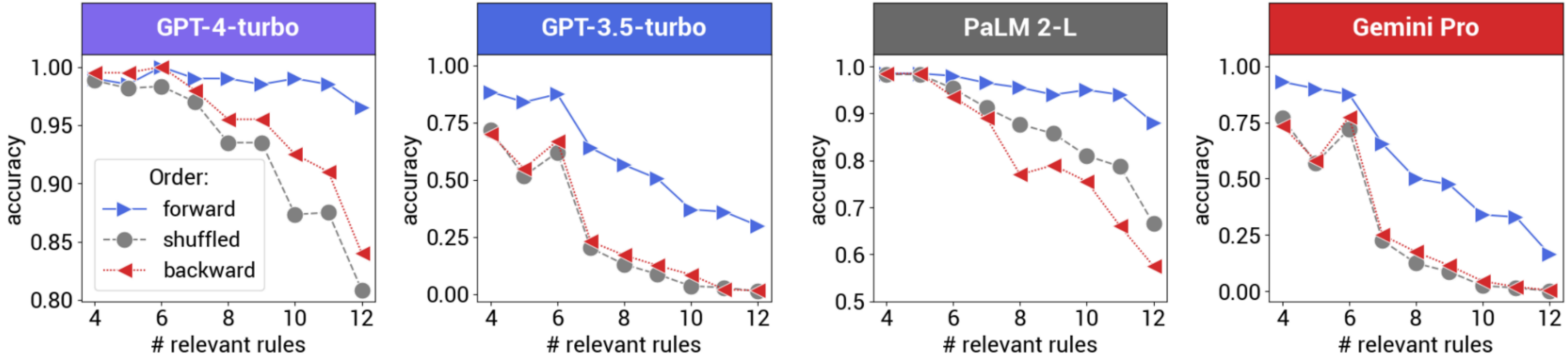


Figure 1: Logical reasoning without distracting rules.

Experiments

Logical Reasoning

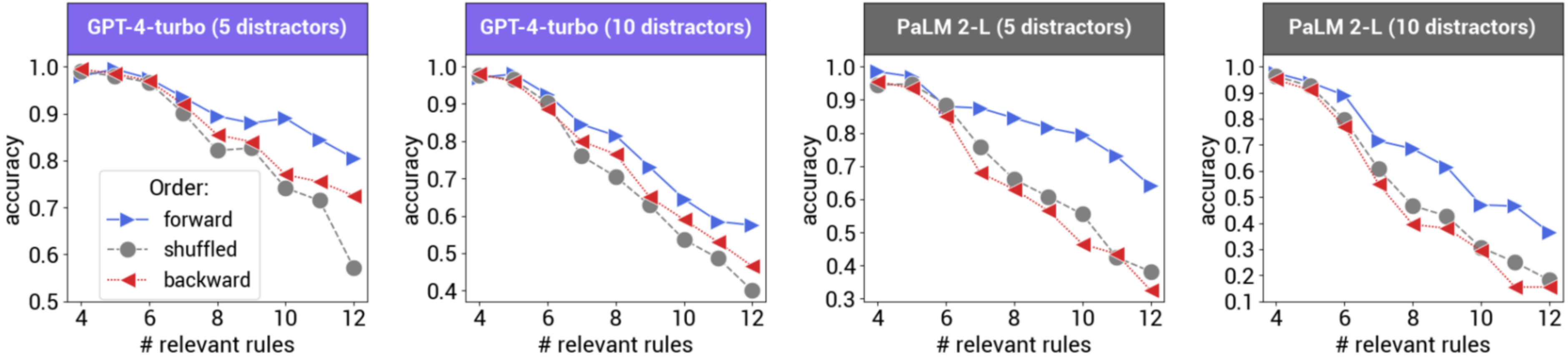


Figure 2: Logical reasoning with distracting rules.

Experiments

Logical Reasoning

	τ	Correct	Wrong Refutation	Hallucination Rule	Hallucination Fact
GPT-4-turbo	1	96.5%	0.5%	1.5%	1.5%
	0.5	76.0%	10.5%	2.0%	11.5%
	0	82.0%	4.5%	3.5%	10.0%
	-0.5	84.5%	1.0%	4.5%	10.0%
	-1	84.0%	0.0%	3.5%	12.5%
GPT-3.5-turbo	1	30.0%	24.5%	9.5%	35.5%
	0.5	1.0%	54.5%	9.5%	33.0%
	0	0.5%	55.0%	7.5%	34.5%
	-0.5	2.0%	50.0%	8.5%	37.5%
	-1	1.5%	34.5%	14.5%	47.0%

	τ	Correct	Wrong Refutation	Hallucination Rule	Hallucination Fact
PaLM 2-L	1	88.0%	0.5%	3.0%	8.5%
	0.5	74.5%	1.5%	9.5%	14.5%
	0	65.5%	2.0%	11.0%	21.5%
	-0.5	59.5%	1.5%	10.0%	29.0%
	-1	57.5%	1.0%	11.5%	30.0%
Gemini 1.0 Pro	1	16.5%	28.0%	5.0%	50.5%
	0.5	0.0%	59.0%	3.5%	37.5%
	0	0.0%	34.0%	9.0%	57.0%
	-0.5	0.5%	24.5%	9.5%	65.5%
	-1	0.5%	27.5%	11.5%	60.5%

Table 1: Error analysis for logical reasoning with 12 relevant rules and no distracting rules.

Experiments

R-GSM for Mathematical Reasoning

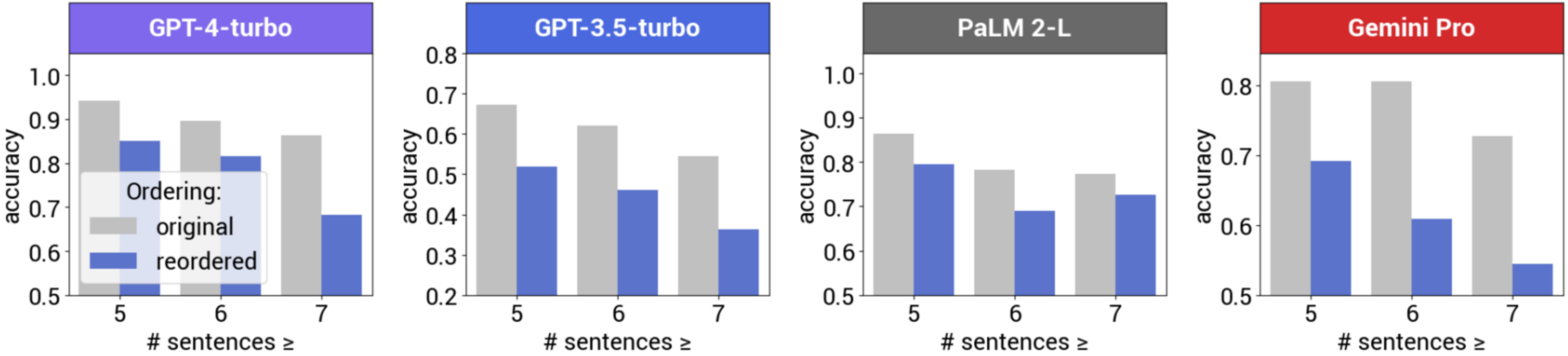


Figure 5: R-GSM results with different problem lengths.

Experiments

R-GSM for Mathematical Reasoning

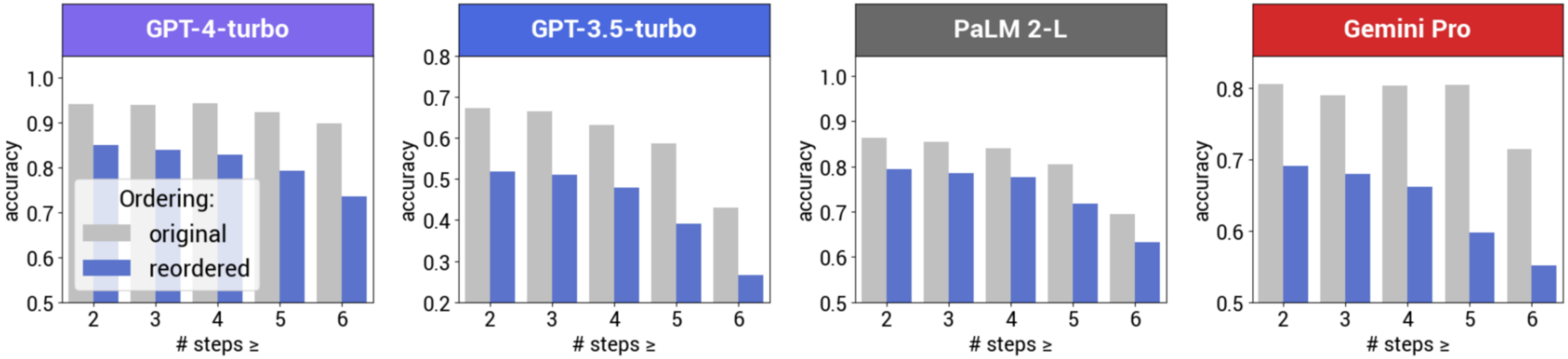


Figure 6: R-GSM results with different numbers of reasoning steps in the ground truth.

Experiments

R-GSM for Mathematical Reasoning

	Init Acc	Reorder Acc
GPT-4-turbo	94.1%	85.0%
PaLM 2-L	86.4%	79.5%
Gemini 1.0 Pro	80.5%	69.1%
GPT-3.5-turbo	67.3%	51.8%

Table 2: Results on the R-GSM dataset accuracies on the full dataset.

	Temporal	Unknown	Others
GPT-4-turbo	45.0%	15.0%	40.0%
GPT-3.5-turbo	21.6%	19.6%	58.8%
PaLM 2-L	34.8%	4.3%	60.9%
Gemini 1.0 Pro	29.5%	18.2%	52.3%

Table 3: Error analysis on R-GSM.

Conclusions

- Premise ordering significantly affects the LLM performance even when the order does not change the underlying task itself.
- LLMs face difficulties when the reasoning problem requires the model to read the problem description back-and-forth, resulting in a performance drop.
- The study was extended to include GSM problems to confirm that the effect is not limited to just logical reasoning.

Thank you for listening!

Do you have any questions?