Pietro Bertorelle

Politecnico di Torino

Supervisor: **Antonio Vetrò**

Stress Testing Chatbots

Evaluating factuality, reasoning, abstraction, and other safety challenges

Background

- *Large Language Model (LLM):* a deep learning model trained on massive text data to understand, generate, and predict human language. Foundational technology of chatbots.
- *Agentic AI:* autonomous system capable of performing tasks without human supervision.
- *Agentic Benchmark:* An agentic benchmark evaluates AI agents on complex, real-world tasks.

Benchmarking LLMs for Complex Tasks

This work focuses on benchmarking the agentic capabilities of three prominent LLMs:

- **Gemini**
- **Gemma**
- **Llama**

The evaluation focused on three primary test categories: *Reasoning*, *Factuality*, and *Sequential Problem Solving*

LLMs Under Evaluation

This study focuses on benchmarking three prominent Large Language Models that are highly relevant and widely discussed within the AI research and development community:

- **Gemini**
  - Google's most capable and multimodal AI model
  - Accessed via the Gemini API for this benchmark
  - Versions tested: *1.5 flash 002*, *1.5 flash 8B 001*, *2.0 flash lite 001*, *2.0 flash 001* and *2.0 flash thinking exp*
- **Gemma**
  - A family of open-source models from Google, built on the research and technology used in Gemini
  - Utilized for its accessibility and research-focused design
  - *Gemma 3 with 1B and 4B* parameters were tested
- **Llama**
  - Meta's widely recognized open-source LLM
  - Frequently used as a foundational model in research and development
  - Versions *3.1 8B*, *3.2 1B* and *3.2 3B* were tested

Experiment Setup and Scale

The benchmark utilized:

- 13 free-tier Google accounts for API access
- 10 distinct versions of Large Language Models (LLMs)
- 22 unique submitted questions
- Resulting in over 273,000 responses

Example of questions

- What is the maximum point in the interval $x = 0$ and $x = \pi$ of the function $y = x \sin(x)$?

- We want to fill a stadium with golf balls (whose volume can be aproximated to 40 cm^3). The balls inserted double with each insertion: 1 ball in the first insertion, 2 balls in the second, 4 in the third, ... If at the 30th insertion the stage is exactly half full, how may insertions are left to fill it?

- How many subsections of Article 140 of the Italian Constitution are there?

- “The current king of France is dead”. Evaluate whether this statement is:

  1- True 2- false 3- other

Tested Capabilities Breakdown

The evaluation comprised 22 unique questions distributed across 9 distinct categories:

I. **Reasoning**
   - Mathematical Reasoning
   - Common Math Problems
   - Sudoku

II. **Factuality**
   - Factual Pitfalls
   - Russel's theory of descriptions

III. **Sequential Problem Solving**
   - Wolf, Goat and Cabbage
   - Blocks World
   - Hanoi Tower
   - Ordered Stack

Sequential Problem Solving

This category's questions test an LLM's ability to manage **sequential dependencies**, directly inspired by challenges in **Sussman's Anomaly**.

Sussman's Anomaly, a classic AI planning problem, perfectly illustrates Sequential Problem Solving complexities by highlighting issues like:

- **Goal Interaction**: Achieving one sub-goal can inadvertently disrupt a previously achieved sub-goal.
- **Linear vs. Non-linear Planning**: The anomaly shows the limitations of simple (linear) planning, demonstrating the need for advanced non-linear or hierarchical approaches.

Methodology: LLM Testing Framework

- **Gemini (Cloud-Based):**
  - Accessed through the Gemini API
  - Leveraged **13 'free tier' Google accounts** in parallel to overcome API rate limits and scale submissions
  - Each question was submitted repeatedly to the Gemini model to collect a robust dataset.
- **Gemma & Llama (Locally Hosted):**
  - Deployed and managed locally using the **Ollama** open-source software
  - Questions were submitted multiple times to each of these models running locally

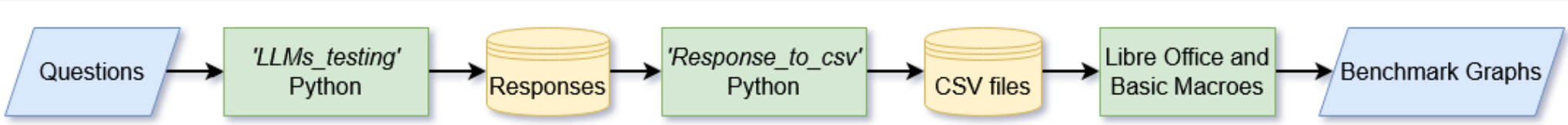
The entire question submission, response collection, and data processing pipeline was fully automated with a custom Python program.

Prompt Engineering Strategies

Reasoning and Problem Solving questions were tested using distinct prompting strategies:

- ***One Shot* (OS):**
  - The LLM was prompted to provide only the final answer directly in a single turn
- **Chain of Thought (CoT):**
  - Involved a two-turn interaction:
    1) The LLM was first prompted to exhibit extensive reasoning
    2) A subsequent, separate prompt then instructed the LLM to output only the final correct result
- **Program of Thought (PoT):**
  - *Process*: The LLM was prompted to generate a complete Python program designed to solve the given problem
  - *Execution*: This generated code was then run externally in a controlled environment to obtain and verify the final solution.

Data Analysis Workflow



The post-experiment data was processed through a structured workflow:

- **Automated Evaluation:**
  - A custom Python script was developed to automatically assess the correctness of each of the 273,000+ LLM responses
- **Initial Data Compilation:**
  - Evaluation results, along with raw responses, were systematically recorded into individual CSV files
- **Aggregate Processing & Visualization:**
  - Each CSV file underwent further automated processing, using a dedicated macro to aggregate results
  - This aggregated data was then readily transformed into insightful graphs and visualizations for performance analysis

Results: Understanding the Data

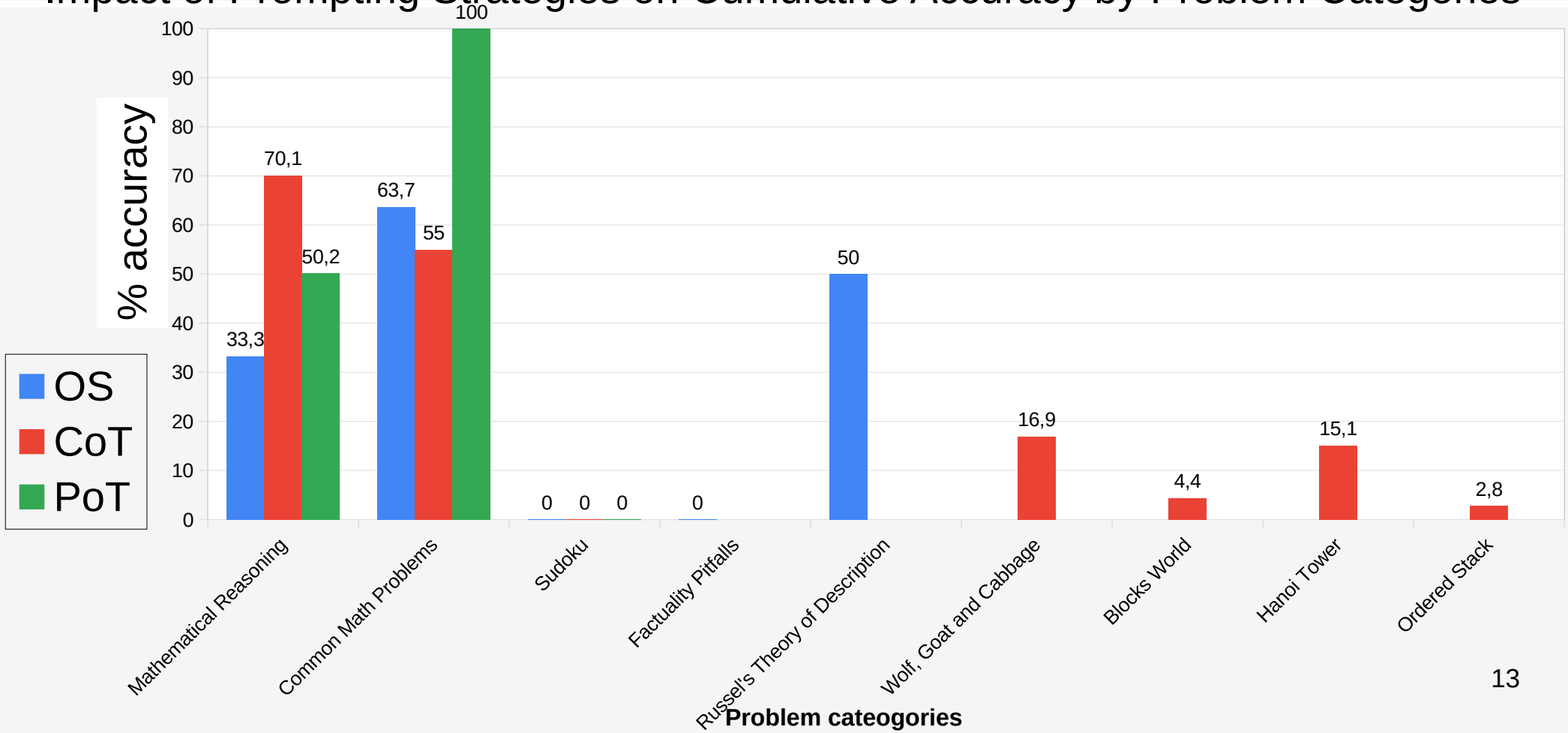
The data presented in the upcoming graphs was prepared as follows:

- **Individual Question Accuracy**: For each question, the percentage of accuracy was calculated based on multiple repetitions or attempts.
- **Account-Level Aggregation**: The median accuracy for each question was then determined by considering the results across different accounts.
- **Category-Level Aggregation**: Finally, the mean of these median accuracies was calculated for all questions within each specific problem category.

These category-level aggregated data are what you will see visualized in the graphs.

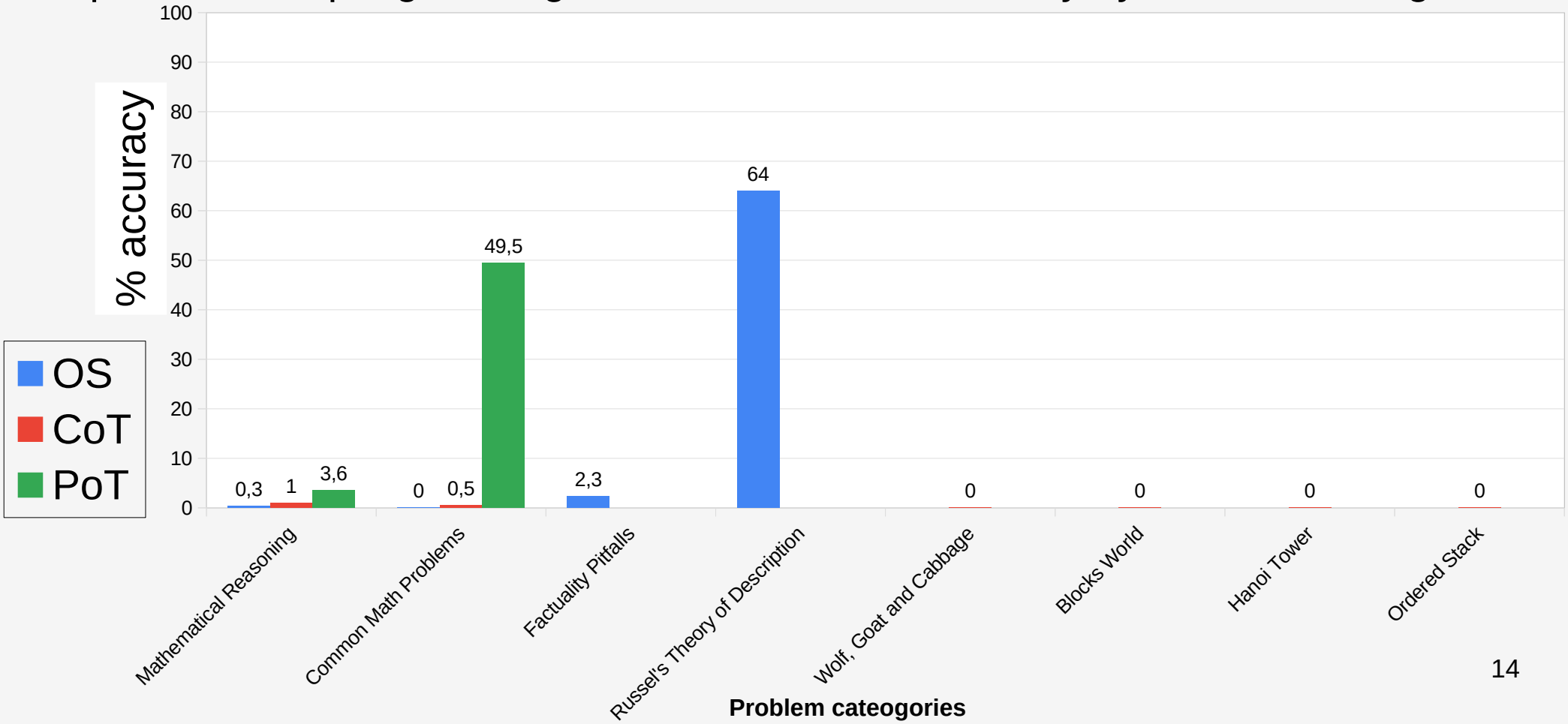
Gemini 2.0 Flash

Impact of Prompting Strategies on Cumulative Accuracy by Problem Categories

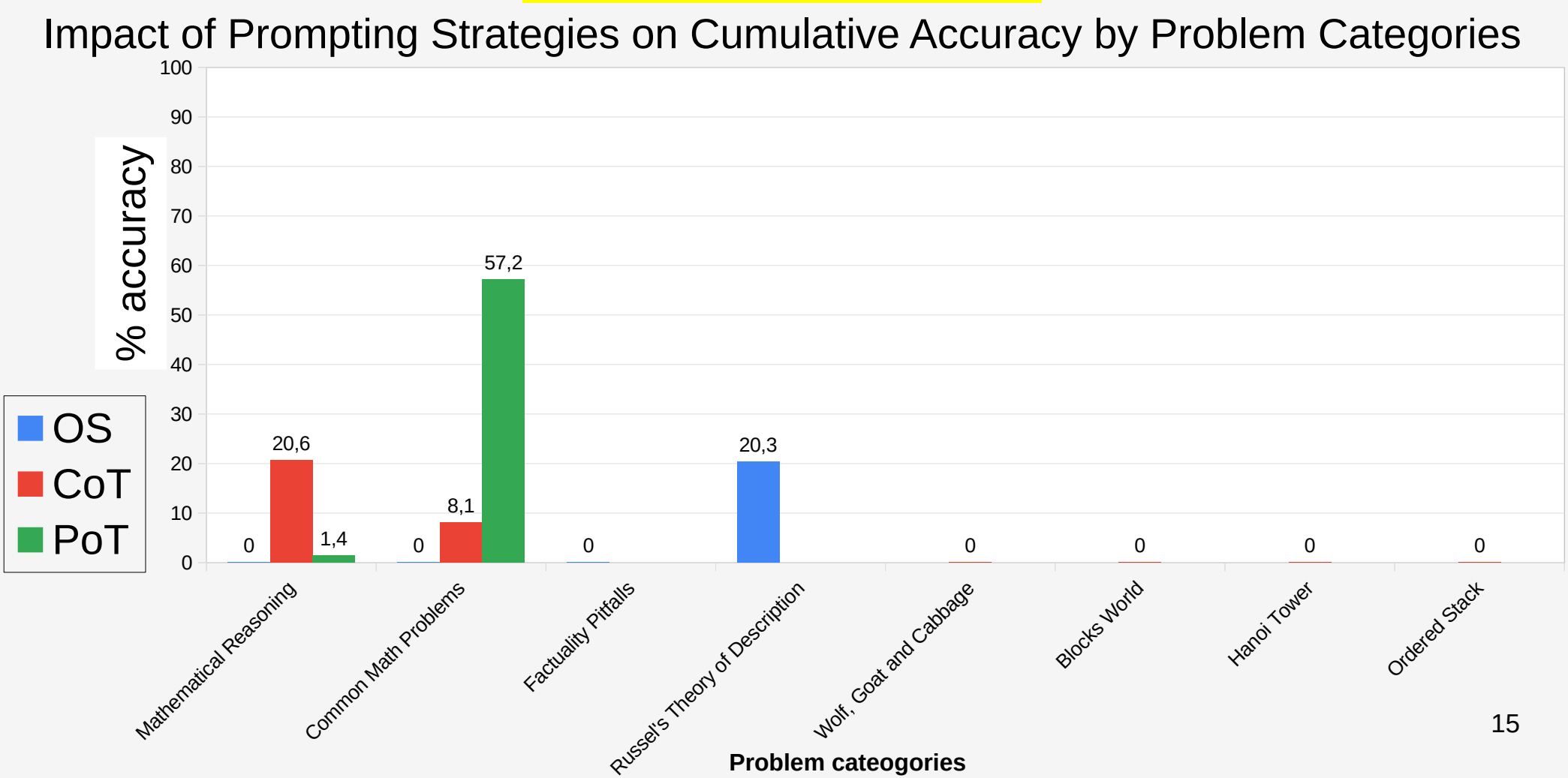


Llama 3.2 3B

Impact of Prompting Strategies on Cumulative Accuracy by Problem Categories



Gemma 3 4B


Impact of Prompting Strategies on Cumulative Accuracy by Problem Categories
% accuracy
100
90
80
70
60
50
40
30
20
10
0
OS
CoT
PoT
0
20,6
1,4
0
8,1
57,2
0
20,3
0
0
0
0
Mathematical Reasoning
Common Math Problems
Factuality Pitfalls
Russel's Theory of Description
Wolf, Goat and Cabbage
Blocks World
Hanoi Tower
Ordered Stack
Problem cateogories

Conclusions

- The study was able to demonstrate some of the limitations and strengths of LLMs in addressing complex problems.
- Several improvements can be made:
  - → The categories cover only 3 of the 7 main goals of artificial intelligence.
  - → There are few questions proposed for each category.
  - → The LLMs analyzed are only those that do not require payment and can be hosted on low-cost hardware architecture.