Large Language Models

Assignment-2

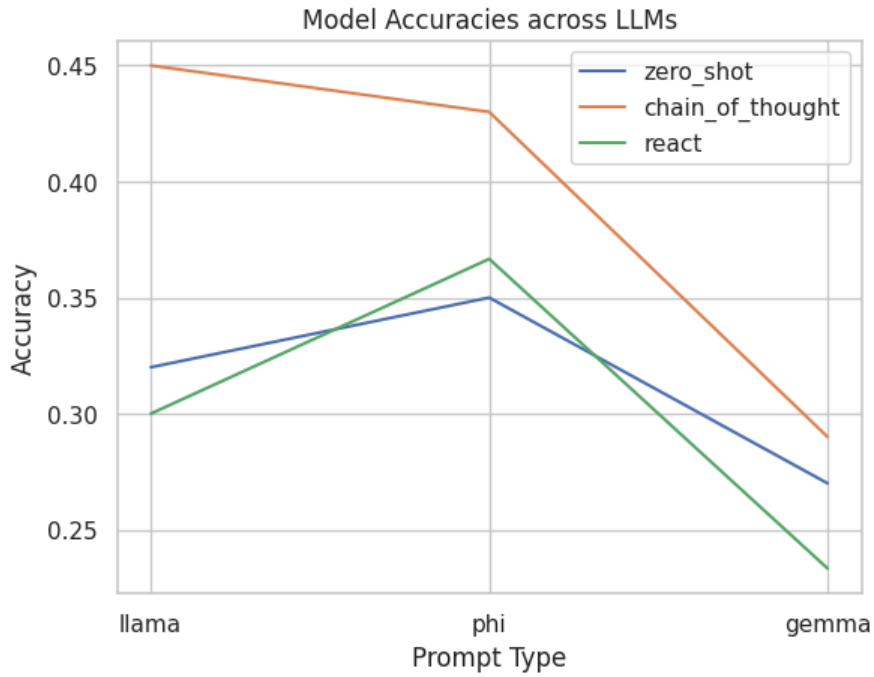
Sachin Sharma, 2021559

LLMs

1. gemma-2b-it
2. Phi-3.5-mini-instruct
3. Meta-Llama-3.1-8B-Instruct

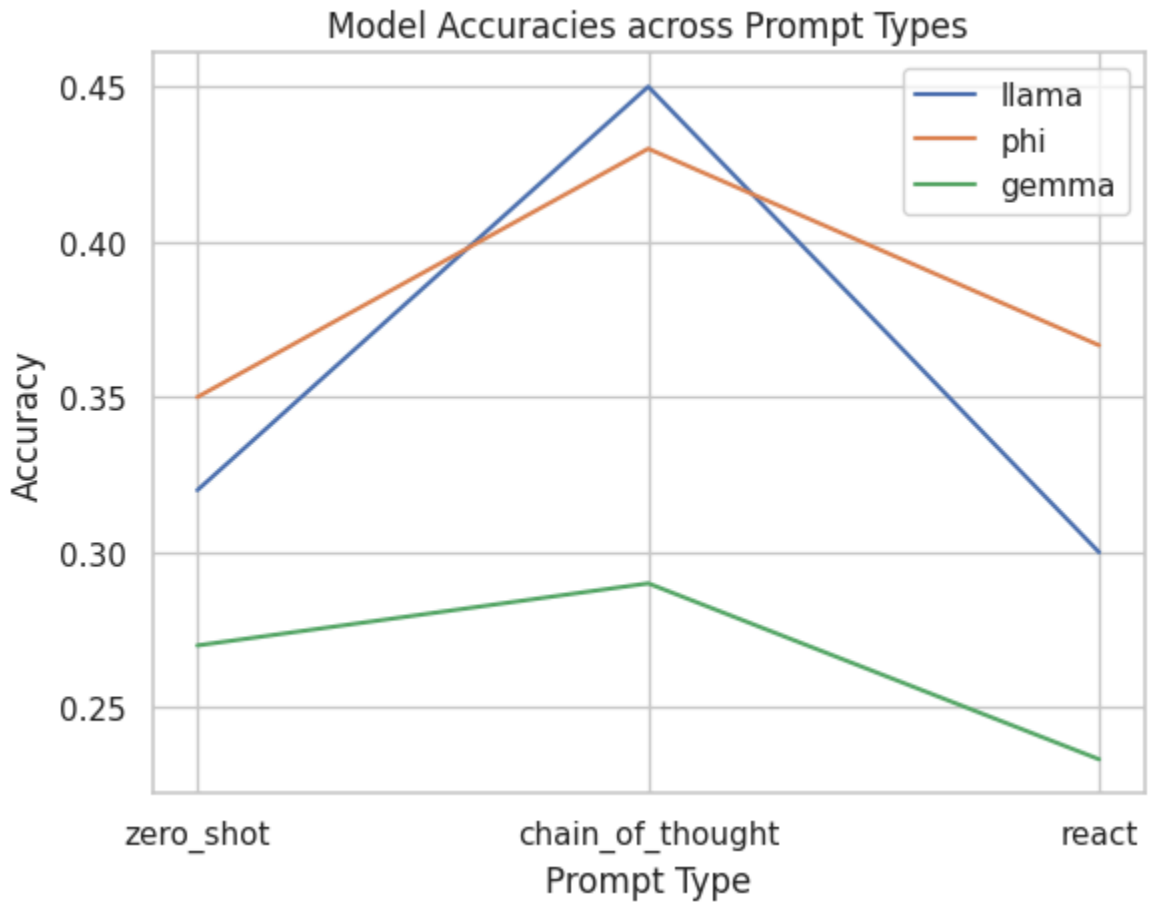
Evaluation of LLMs

Accuracy Plot



The performance of all the LLMs varied significantly. Accuracies of Llama-3.1 and Phi-3.5 were comparable, with Phi having slightly better Zero-Shot accuracy, but Llama has better

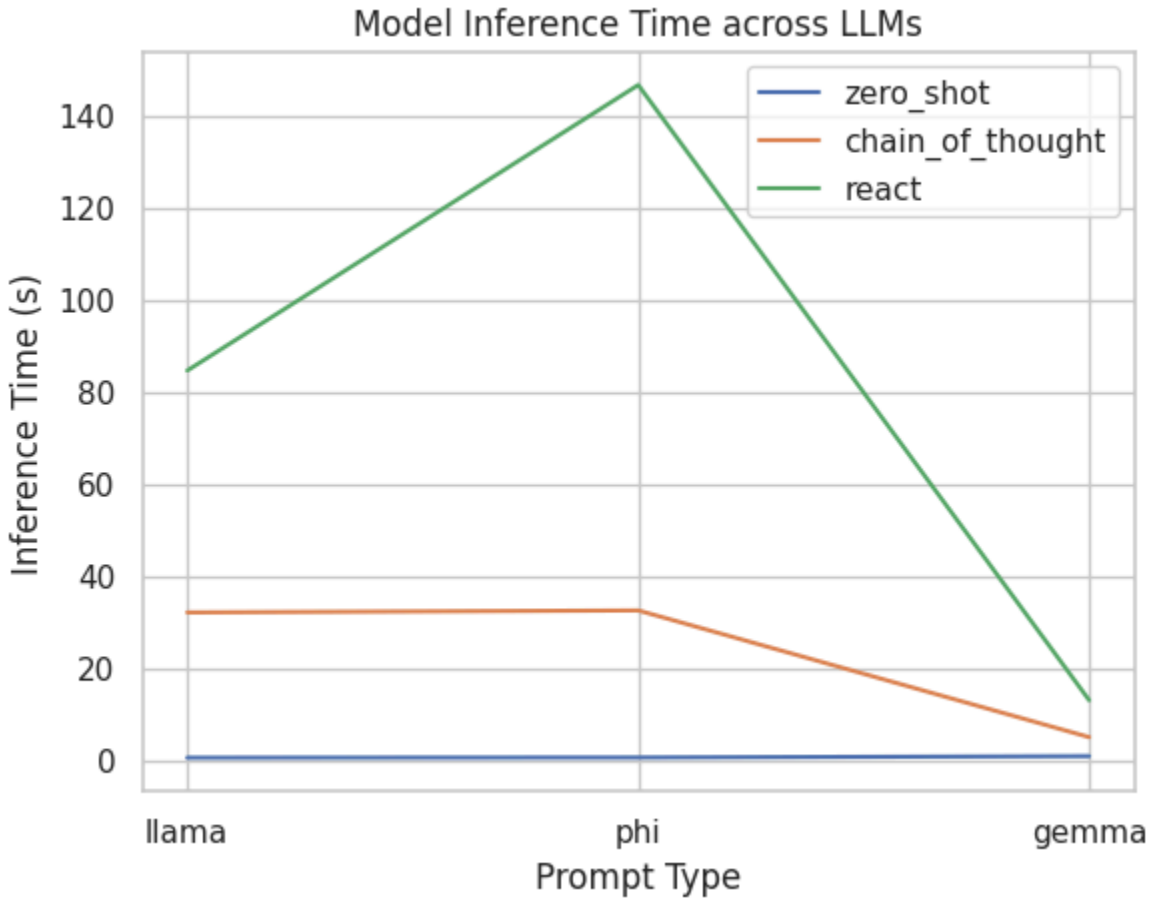
chain-of-thought performance. This suggests that both the models are at par with each other, but Llama improves drastically with chain-of-thought prompting. Gemma, being the smallest model, has the least accuracy due to a smaller number of parameters and, thus, lower learning capacity.



Llama performed best with chain-of-thought prompting with 45% accuracy, however, Phi is almost comparable with its performance. This suggests that even though Phi has a smaller number of parameters, its architecture enables good reasoning ability, potentially because of `LongRoPEScaledRotaryEmbedding` and Residual Attention Dropout layers, which are not present in Llama.

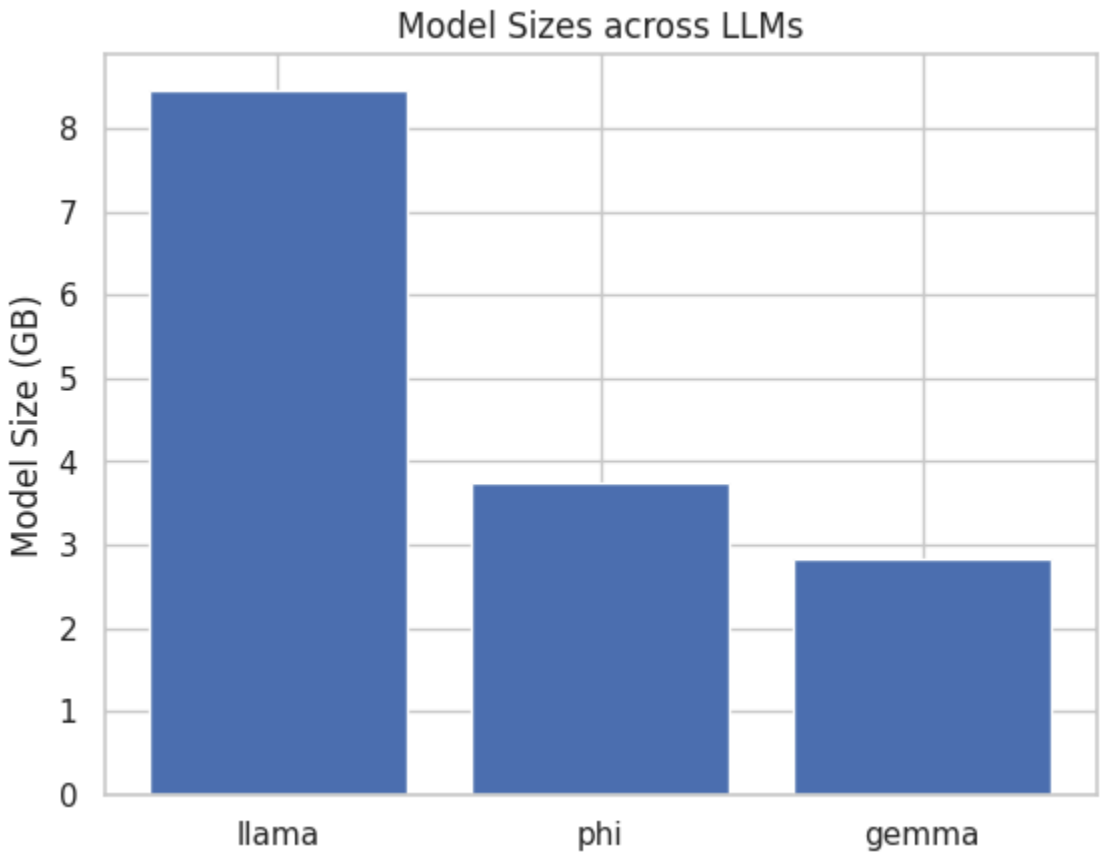
The performance of all 3 models improved from zero-shot to chain-of-thought, but reduced from chain-of-thought to ReAct prompting. As given in the https://www.promptingguide.ai/techniques/react, ReAct works better than COT for knowledge-intensive tasks but lacks slightly in Arithmetic and Mathematics problems. A possible reason for the drop in accuracy could be that COT involves a much continuous stream of reasoning and steps. At the same time, ReAct breaks down the reasoning into interactive steps, which causes many tokens and ideas to overlap and some tokens to miss out in subsequent steps. Thus, complex reasoning tasks suffer from discontinuity in reasoning.

Inference Time



Llama, being the biggest model, generally takes more time to generate inference. However, Phi takes an abnormally much longer time in the case of ReAct prompting, maybe due to specific implementation of the ReAct technique. Apart from that, inference times for Llama and Phi are similar, and Gemma is much faster due to its smaller size.

Model Size



Llama is the biggest model with 8.5 GB of memory footprint, followed by Phi, and then Gemma, which is the decreasing order of the number of parameters.

Comparison Review of LLM Performance

Meta-Llama-3.1-8B-Instruct:

- It's the largest model amongst the 3, with 8 billion parameters (Meta Llama 3 Herd of Models).
- Llama-3.1 uses Grouped-Query Attention (GQA), enabling large context window size and instruction-tuning for improved inference scalability. (Hugging Face).
- It is optimized for broader multilingual instruction tasks.
- Due to a much larger number of parameters than its competitors in the assignment, its reasoning ability increases with an increase in the problem context. That's why its performance increases dramatically from zero-shot to chain-of-thought.

Phi-3.5-mini-Instruct:

- This lightweight model has 3.5 billion parameters and is designed for resource-constrained environments like mobile or embedded systems (ar5iv).
- Despite its smaller size, its architecture enables it to accomplish complex reasoning tasks, potentially due to `LongRoPEScaledRotaryEmbedding` and Residual Attention Dropout layers, which enable it to attend to more specific parts of the prompt (Phi-3 Technical Report).

Gemma-2B-it:

- The smallest model amongst given ones with only 2 billion parameters.
- Gemma is an instruction-tuned LLM that is lightweight and performs slightly better in regional languages than its counterparts due to a more significant proportion of non-English text in the training data. Despite the small size, it is competitive with much larger models (Gemma Paper).
- Due to its significantly small size and small context length, its learning capacity is fairly limited. Thus, it performance doesn't improve when applied chain-of-thought prompting.