LLM Assignment - 2
Efforts By - Ieshaan Awasthy (2021054)

Q1. Evaluate and compare the inference time for each LLM using above prompts. Additionally, assess the accuracy of the generated outputs and discuss the trade-offs between model size, inference speed, prompt used and output quality.

Ans 1.

Inference Time on 100 problems (seconds)

Models	Model Size	Zero Shot	CoT	ReAct
Gemma	2 Billion	1.9	3.6	3.7
Phi 3.5 mini instruct	3.5 Billion	13.1	15.2	18.5
Llama 3.1 instruct	8 Billion	2.1	4.9	7.2

Accuracy of the Models in percentage (%)

Models	Model Size	Zero Shot	CoT	ReAct
Gemma	2 Billion	40%	42%	36%
Phi 3.5 mini instruct	3.5 Billion	34%	44%	40%
Llama 3.1 instruct	8 Billion	37%	49%	43%

Clear trade-off between speed and quality: smaller models are faster but produce slightly less accurate results, while larger models can produce more accurate outputs at the cost of increased inference time.

Q2. Read technical reports and papers related to given 3 LLM's and give reasons why Model X performed better than Model Y or comparable in your case by citing relevant reasons from verified resources.

Ans 2. Detailed analysis of inferences:

Inference Time:

- **Gemma (2B)** is the fastest across all prompting methods, with an average inference time of around 1.9 to 3.7 seconds.
- **Llama 3.1 (8B)** has slower inference times compared to Gemma, but still performs significantly faster than **Phi 3.5 mini instruct (3.5B)**.
- **Phi 3.5 mini instruct (3.5B)** is the slowest, taking between 13.1 and 18.5 seconds depending on the prompting method. This shows a substantial increase in inference time as the model size grows.

Accuracy:

- For all models, **Chain of Thought (CoT)** prompting consistently yields the highest accuracy.
- **Zero Shot** prompting generally performs the weakest, except in the case of Gemma, where it is comparable to ReAct.
- **ReAct** prompting improves accuracy but tends to be slightly weaker than CoT for most models, though it's better than Zero Shot.

Trade-offs:

- **Model Size**: Larger models like Llama 3.1 (8B) and Phi 3.5 mini (3.5B) generally perform better in terms of accuracy compared to smaller models like Gemma (2B). However, the accuracy improvement comes at the cost of increased inference time.

- **Inference Speed**: The smaller Gemma (2B) model is much faster, but its accuracy is slightly lower compared to larger models when advanced prompts (CoT, ReAct) are used.
- **Prompt Used**:
    - CoT enhances reasoning and leads to better accuracy for most models but increases inference time.
    - ReAct improves accuracy as well, but it is generally slower than Zero Shot while offering a trade-off between speed and better output quality.

To support the claim regarding the trade-offs and performance of Gemma, Phi, and LlaMa models, I will be providing articles and research papers to justify it.

Analyzes the performance of the models:

Llama 3.1 excels in complex reasoning tasks due to its larger size and architecture, offering superior accuracy with techniques like Chain of Thought (CoT), though it has slower inference speeds. In contrast, Gemma, with its 2B and 7B versions, focuses on speed and efficiency, achieving competitive results through architectural improvements and instruction tuning. Phi 3.5 Mini Instruct, while smaller, struggles with speed and complexity compared to both models. Despite employing advanced prompting strategies, it lags behind in efficiency. Ultimately, the trade-offs between model size, architecture, and optimization dictate their performance across tasks. (https://ar5iv.labs.arxiv.org/html/2403.08295, https://www.ibm.com/blog/meta-releases-llama-3-1-models-405b-parameter-variant/)

Chain-of-Thought (CoT) Prompting and Reasoning

It was found that CoT has greatly enhanced the performance of language models on tasks requiring multi-step reasoning. In my experience, CoT prompts deliver the highest accuracy for LLaMa, which supports the findings of this research. It demonstrates that while CoT significantly boosts

reasoning capabilities in larger models, it also leads to longer inference times.

[2201.11903] Chain-of-Thought Prompting Elicits Reasoning in Large Language Models (arxiv.org)