LLM Assignment 2 Report

1. Introduction

- **Task Overview**: We evaluate the performance of three LLMs on a mathematics question-answering task using Zero-Shot and Chain of Thought (CoT) prompts. The aim was to assess their accuracy and inference speed across different prompt types, and draw conclusions based on their model sizes, capabilities, and trade-offs.
- **Dataset**: The dataset used for evaluation is the **MMLU College Mathematics** dataset from Hugging Face.

2. Experiment Setup

- **Models**:
    - Google Gemma-2B: A 2-billion parameter LLM developed by Google.
    - Microsoft Phi-3.5-mini: A smaller instruct-tuned LLM from Microsoft.
    - Meta LLaMA-3.1-8B: A larger 8-billion parameter LLM from Meta.

3. Results Summary

Model	Prompt Type	Total Correct (Out of 100)	Avg. Inference Time (seconds)
Google Gemma-2B	Zero-Shot	28/100	3.5551
Google Gemma-2B	CoT	30/100	2.8052
Meta LLaMA-3.1-8B	Zero-Shot	26/100	7.6411
Meta LLaMA-3.1-8B	CoT	23/100	7.6279
Microsoft Phi-3.5-mini	Zero-Shot	42/100	7.4295
Microsoft Phi-3.5-mini	CoT	42/100	7.4782

4. Analysis

a. Inference Time Comparison

- **Gemma-2B**: Showed the fastest average inference time, especially with CoT prompting (2.8052s). This might be attributed to its relatively smaller size (2B parameters), making it more computationally efficient for inference. CoT prompting reduced inference time compared to Zero-Shot, suggesting that the model processes structured thought faster than a direct answer.
- **Meta LLaMA-3.1-8B**: The largest model had the slowest inference times, around 7.6 seconds per example, regardless of the prompt used. This is consistent with the expected overhead from its 8B parameter size.
- **Microsoft Phi-3.5-mini**: While slightly larger than Gemma-2B, its inference times were comparable to Meta LLaMA-3.1-8B, likely due to its tuning and architecture optimizations.

b. Accuracy Comparison

- **Phi-3.5-mini** consistently performed better than the other models, scoring 42/100 correct in both Zero-Shot and CoT prompts. Its instruct tuning and optimization for general-purpose tasks might explain this superior performance.
- **Gemma-2B** showed a modest increase in performance with CoT prompting, moving from 28/100 to 30/100 correct. This suggests that CoT helped it reason slightly better.
- **Meta LLaMA-3.1-8B** surprisingly underperformed relative to its size. In fact, its CoT performance dropped to 23/100 from 26/100 in Zero-Shot, suggesting that it might have struggled with the step-by-step reasoning process on this specific dataset.

5. Trade-offs and Model Comparison

a. Model Size vs. Inference Time

- **Gemma-2B**: Demonstrated the best balance between inference speed and moderate accuracy. Its smaller size allows for faster processing, making it suitable for tasks where time is critical.
- **Meta LLaMA-3.1-8B**: Despite being the largest model, it did not show a significant improvement in accuracy over smaller models. Its large size led to slower inference, which might not justify its use if time and computational resources are limited.
- **Phi-3.5-mini**: While not the fastest, it provided the best accuracy, suggesting that its instruct-tuning and optimizations make it a strong candidate for tasks requiring reliable reasoning with some tolerance for higher inference time.

b. Prompt Type Comparison

- **Zero-Shot vs. Chain of Thought**:
    - For **Gemma-2B**, CoT improved both accuracy and inference time, possibly because this model benefits from the structured nature of CoT.
    - **Phi-3.5-mini** showed no difference in performance between Zero-Shot and CoT, which might indicate that it processes both types of prompts similarly.

- **Meta LLaMA-3.1-8B** experienced a slight performance degradation with CoT, which could be related to overfitting or inefficiency in handling CoT on this dataset.

6. Technical Insights and Relevant Research

- **Google Gemma-2B**:
  - Source: https://arxiv.org/abs/2408.00118
  - According to technical reports on the model, its relatively small size and efficient transformer architecture contribute to its quick inference time. It is optimized for efficiency rather than high complexity, which might explain its moderate accuracy but strong performance in terms of speed.
- **Microsoft Phi-3.5-mini**:
  - Source: https://arxiv.org/abs/2404.14219
  - Research on instruct-tuned models suggests that they are highly effective in generalizing across tasks. Phi-3.5-mini's tuning likely plays a significant role in its ability to consistently produce better results, even in a Zero-Shot setting. This explains its higher accuracy despite not being the largest model.
- **Meta LLaMA-3.1-8B**:
  - Source: https://ai.meta.com/blog/meta-llama-3/
  - As per LLaMA reports, the model architecture excels at handling a wide range of language tasks, but its performance might be hindered on more structured tasks like mathematics, where fine-tuning plays a crucial role. The slight performance drop in CoT suggests that its large size might introduce some inefficiency in reasoning-heavy tasks.

7. Conclusion

- **Best Accuracy**: Microsoft Phi-3.5-mini, due to its instruct tuning, offers the best balance between prompt understanding and performance.
- **Best Speed**: Google Gemma-2B, owing to its smaller size, is the fastest in terms of inference time, making it a good choice where speed is crucial.
- **Trade-off Discussion**: Larger models like Meta LLaMA-3.1-8B do not always outperform smaller models, especially when the task does not heavily benefit from their larger capacity. The choice of model should depend on whether inference speed or accuracy is the priority, with Phi-3.5-mini striking the best balance.

Github repo: AbhayChowdhry/LLM_projects (github.com)