CSE665: Large Language Models

Assignment 2

Trade off between Model size, Prompt type, Time Taken and Quality

Aryan Dhull | 2021520

Objective:

The primary goal of this task is to evaluate the performance of various publicly available large language models (LLMs) on a multiple-choice mathematics dataset using three different prompting techniques:

- **Zero-Shot** prompting
- **Chain of Thought (CoT)** prompting

The models being evaluated include:

- **Gemma-2B**
- **Phi-3.5-mini**
- **Meta-Llama-3.1-8B**

The evaluation is performed using the **MMLU College Mathematics** dataset from Hugging Face, and the metrics used for comparison include:

- **Accuracy:** How often the model provides the correct answer.
- **Inference Time:** The average time taken by the model to generate a response.

1. Zero-Shot Prompting

Approach:

- **Zero-Shot Prompting** is the simplest form of prompting. In this case, the model is directly given a question and a list of answer options without any guidance or intermediate reasoning steps. The model has to predict the answer without additional context or thought process.

- **Prompt Template**:

  > "Choose the answer to the given question from below options. [Question][Option 1][Option 2][Option 3][Option 4]"

- **Intuition:** Zero-Shot prompting relies solely on the model's pre-trained knowledge and generalization ability. The model must predict the correct answer without any explicit reasoning or decomposition of the problem, making it a quick but often less accurate approach.

2. Chain of Thought (CoT) Prompting

Approach:

- **Chain of Thought (CoT) Prompting** involves asking the model to break down the question and provide a step-by-step reasoning process before selecting an answer. The model is guided to "think step by step" to arrive at the correct conclusion.

- **Prompt Template**:

  > "Choose the answer to the given question from below options. [Question][Option 1][Option 2][Option 3][Option 4] Think step by step"

- **Intuition:** CoT prompting improves a model's performance on complex reasoning tasks by guiding it to think through the problem methodically. This type of prompting can lead to more accurate answers, especially in tasks like mathematics that benefit from logical breakdowns.

Final Comparative Analysis:

MODEL	PROMPT	INFERENCE TIME	ACCURACY
google/gemma-2b-it	Zero Shot	**8.16s per prompt**	**0.3**
	Chain of Thought	**12.02s per prompt**	**0.19**
microsoft/Phi-3.5-mini-instruct	Zero Shot	**15.63s per prompt**	**0.3**
	Chain of Thought	**16.13s per prompt**	**0.19**
meta-llama/Meta-Llama-3.1-8B-Instruct	Zero Shot	**29.00s per prompt**	**0.1**
	Chain of Thought	**28.65s per prompt**	**0.06**

Analysis Of Time:

- **Gemma-2B** had the fastest inference times in both **Zero-Shot and Chain of Thought** prompting, with Zero-Shot being quicker.
- **Phi-3.5-mini-instruct** took approximately twice the time of **Gemma-2B,** but was still faster than **Meta-Llama-3.1-8B.**
- **Meta-Llama-3.1-8B** had the slowest inference time, which reflects its larger model size (8B parameters), leading to increased computational time.

Analysis of Accuracy:

- **Gemma-2B** and **Phi-3.5-mini-instruct** achieved the highest accuracy with **Zero-Shot** prompting, both reaching 0.30.
- **Chain of Thought** prompts led to a drop in accuracy for both **Gemma-2B** and **Phi-3.5-mini-instruct** (0.19), possibly due to the complexity introduced by the reasoning steps.
- **Meta-Llama-3.1-8B** performed the worst across the board, with a significant drop in accuracy. It achieved 0.10 in **Zero-Shot** and 0.06 in **Chain of Thought** prompting.

  One reason for Meta-Llama-3.1-8 B's lower accuracy is that it tends to provide explanations of the solution rather than the solution itself. This behavior leads to inefficiencies in tasks that require concise, direct answers, which contributed to its

reduced performance, especially in tasks expecting clear and correct outputs. The model's design prioritizes detailed reasoning, which, while useful in some contexts, was a disadvantage in this evaluation.

Trade-offs Between Model Size, Inference Speed, and Output Quality:

- **Model Size vs Inference Speed**:
    - **Gemma-2B,** being the smallest model, was the fastest in terms of inference time. It achieved the best balance of speed and accuracy, making it suitable for tasks where time efficiency is critical.
    - **Phi-3.5-mini-instruct** performed slightly slower, likely due to its intermediate size and architectural differences.
    - **Meta-Llama-3.1-8B** had the longest inference time due to its larger parameter size, resulting in slower response times.

- **Prompt Complexity**:
    - **Zero-Shot** prompting was consistently faster across all models, as it relies on direct answers with minimal reasoning steps.
    - **Chain of Thought** prompting, which encourages multi-step reasoning, took longer to process but surprisingly did not enhance accuracy. Instead, it led to a decline in performance for all models. This indicates that CoT may only be beneficial in tasks where the model is already optimized for reasoning.

- **Output Quality**:
    - The best-performing models in terms of accuracy, **Gemma-2B** and **Phi-3.5-mini-instruct,** showed a trade-off between inference speed and accuracy.
    - **Meta-Llama-3.1-8 B's** lower accuracy, combined with a slower response, demonstrates that larger models do not always guarantee better performance. In this case, its focus on explanation over the correct answer led to inefficient outputs, making it less ideal for tasks requiring high accuracy and speed.

When comparing the performance of the three LLMs—**Gemma-2B**, **Phi-3.5-mini-instruct**, and **Meta-Llama-3.1-8B-Instruct**—several factors come into play based on their architecture, fine-tuning methods, and the tasks they're evaluated on. Here's a breakdown of why one model may outperform another:

1. Gemma-2B

Gemma-2B had the fastest inference times and achieved high accuracy in Zero-Shot prompting, often outperforming larger models like Phi-3.5-mini-instruct and Meta-Llama-3.1-8B. With only 2 billion parameters, its smaller size allows it to be highly efficient, especially when fine-tuned, making it effective for tasks like instruction following and multi-step reasoning. Despite its smaller size, it narrows the performance gap with larger models, especially in reasoning tasks.

Paper Analysis:

- Paper 1 explains how Gemma-2B's fine-tuning improves both accuracy and speed, allowing it to remain competitive with larger models despite having fewer parameters.
- Paper 2 details how Gemma models, by balancing size and performance, perform efficiently on complex reasoning tasks such as BBH benchmarks.

Conclusion:

Gemma-2B excels in maintaining both high speed and accuracy by leveraging its smaller model size and fine-tuning capabilities, allowing it to compete with and often outperform larger models like Phi-3.5-mini-instruct and Meta-Llama-3.1-8B in tasks requiring reasoning and instruction following.

2. Phi-3.5-mini-instruct

Phi-3.5-mini-instruct exhibited slower inference times compared to Gemma-2B but showed comparable accuracy, particularly in Zero-Shot tasks. Its larger size (3.5 billion parameters) and use of Reinforcement Learning from Human Feedback (RLHF) improve its ability to follow instructions and perform structured reasoning, especially in Chain of Thought tasks. However, these enhancements lead to longer processing times.

Paper Analysis:

- Paper 1 highlights that instruction tuning helps Phi excel in multi-step reasoning tasks, though it increases complexity and inference time.
- Paper 2 notes that instruction-tuned models like Phi-3.5 perform well in diverse tasks due to sample efficiency, but the trade-off is longer inference times.

Conclusion:

Phi-3.5-mini-instruct excels in instruction-based tasks with higher accuracy but suffers from slower inference due to its larger size and complex tuning techniques.

3. Meta-Llama-3.1-8B

Meta-Llama-3.1-8B demonstrated the slowest inference times and lowest accuracy among the evaluated models, particularly in Chain of Thought tasks. With 8 billion parameters, this model focuses on generating detailed explanations rather than concise answers, which increases its computational burden and leads to reduced performance in tasks requiring quick, direct responses.

Paper Analysis:

- Paper 1 explains that while the LLaMA models are optimized for efficient pre-training and scaling, they prioritize reasoning and detailed explanations, which can hinder performance in tasks that demand brevity and speed.
- Paper 2 discusses how Meta-Llama-3.1-8B excels in generating coherent and reasoned outputs, but the model's focus on detailed explanations results in slower inference times and less accuracy for tasks that require direct answers.
- Paper 3 shows that scaling beyond a certain number of parameters yields diminishing returns. For Meta-Llama-3.1-8B, its size leads to diminishing gains in accuracy while increasing inference time, highlighting the trade-off between scale and efficiency.

Conclusion:

Meta-Llama-3.1-8B is focused on generating detailed explanations, which hampers its speed and accuracy in direct-answer tasks. Its large size (8B parameters) leads to diminishing returns, especially in comparison to smaller models like Gemma-2B, which are faster and more efficient for Zero-Shot and prompt-based tasks.

References:

Zero Shot Prompting

Chain of Thought Prompting

Github:

https://github.com/Aryan-Dhull/LLM-Assignment-2