Assignment 2

Gemma Inference Times:

```
Average Inference Times:

Zero-Shot Inference Times:
Gemma 2B: 2.05s

Chain of Thought Inference Times:
Gemma 2B: 3.54s

ReAct Prompting Inference Times:
Gemma 2B: 2.93s
```

Phi Inference Times:

```
Average Inference Times:

Zero-Shot Inference Times:
Phi 3.5 Mini: 8.65s

Chain of Thought Inference Times:
Phi 3.5 Mini: 8.64s

ReAct Prompting Inference Times:
Phi 3.5 Mini: 8.61s
```

Gemma 2B consistently outperforms Phi 3.5 Mini in terms of inference speed across all prompting techniques. Phi 3.5 Mini may provide more reliable results in situations requiring

deeper understanding or reasoning, but at the cost of slower speed. Gemma 2B offers faster responses but may require more prompt engineering for complex tasks to ensure accuracy.

Model Size:
Phi 3.5 Mini: The larger model size enables it to generate more accurate and nuanced responses, especially for tasks involving deeper reasoning.

Inference Speed:
Gemma 2B is the clear winner when it comes to speed, making it ideal for real-time applications where rapid response is critical.

Output Quality:
Gemma 2B: Faster with decent accuracy for simpler tasks. In zero-shot prompting, it provides quick and acceptable results for straightforward problems. The smaller size means it may struggle with nuanced or multi-step reasoning.
Phi 3.5: Generates high-quality outputs with more detailed reasoning and context understanding.

The Gemma 2 model focuses on improving small model performance through knowledge distillation, where smaller models (like 2B and 9B) learn from larger models by receiving richer gradients during training. This approach allows the smaller models to perform on par with larger ones across tasks like question answering, commonsense reasoning, mathematics, and coding.
Reference: Gemma 2: Improving Open Language Models at a Practical Size

The phi models leverage a combination of high-quality, filtered web data and synthetic data generated by LLMs. This unique approach allows them to enhance performance through focused training rather than simply scaling up model size, which is a typical limitation of models like Gemma. The introduction of LongRope architecture allows phi-3-mini to extend context lengths significantly (up to 128K tokens), enhancing its ability to process larger contexts effectively. This capability surpasses many competitors, including Gemma, which supports 8192 tokens.
Reference: Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone