1. Average Inference Time

- **Zero Shot without Thinking**
  Rank: Llama 3.1 8B > Gemma 2B > Phi 3.5
- **Zero Shot with Thinking**
  Rank : Llama 3.1 8B > Gemma 2B > Phi 3.5
- **React Prompting**
  Rank: Llama 3.1 8B > Gemma 2B > Phi 3.5

The above ranking based on the average Inference time with different prompting methods is due to following reasons:

1. Llama 3.1 8B is taking less time to generate the output just because we are using Together's API to call the Llama model.
2. While Gemma is performing better than Phi3.5 only because Gemma is trained on fewer parameters than Phi.
3. If we were to run the Llama 8B model locally it would rank last just because it is trained on 8B parameters which is greater than any of the two mentioned models.

2. Accuracy

- **Zero Shot without Thinking**
  Rank: Llama 3.1 8B > Gemma 2B = Phi 3.5
- **Zero Shot with Thinking**
  Rank: Llama 3.1 8B > Phi 3.5 > Gemma 2B
- **React Prompting**
  Rank: Llama 3.1 8B > Gemma 2B = Phi 3.5

The above ranking based on the Average accuracy with different prompting methods is due to following reasons:

1. There might be a possibility of Llama being trained on more multiple choice datasets when compared to Gamma and Phi.
2. Another possibility is that Llama may be trained on more Math based questions as compared to Llama and PHi.
3. As the quality of the output generated by Gemma is worse as compared to Llama and Phi, one(or program) may not be able to identify the outcome.

Average Inference Time(Gemma):

- **Zero shot without thinking: 2.02 sec**
- **Zero shot with thinking: 2.13 sec**
- **React Prompting: 105.06 sec**

Average Inference Time(Phi):

- **Zero shot without thinking: 3.94 sec**

- **Zero shot with thinking: 16.95 sec**
- **React Prompting: 395.65 sec**

Average Inference Time(Llama):

- **Zero shot without thinking: 1.4577 sec**
- **Zero shot with thinking: 1.9355 sec**
- **React Prompting: 1.05 seconds**

Accuracy(Gemma):

- **Zero shot without thinking: 16%**
- **Zero shot with thinking: 6%**
- **React Prompting: 10% (Human Evaluation) as answers were very nuanced**

Accuracy(Phi):

- **Zero shot without thinking: 16%**
- **Zero shot with thinking: 12%**
- **React prompting: 60%(Human Evaluation) as answers were very nuanced**

Accuracy(Llama):

- **Zero shot without thinking: 30%**
- **Zero shot with thinking: 36%**
- **React prompting: 32%**

RESEARCH

The Llama 3.1 model stands out against competitors like Gemma and Phi due to several strategic enhancements in its development, particularly in the areas of data diversity, scale of training, and managing complexity. Here’s why Llama 3.1 may be better:

1. **Diverse and Rich Data Sources**: Llama 3.1's training incorporated a wide variety of data sources including web pages, books, articles, and research papers. This diversity ensures comprehensive coverage of different writing styles, terminologies, and subject matters, which is crucial for the model's ability to understand and generate nuanced language. This extensive data set helps the model capture real-world language usage from informal online colloquialisms to the structured and precise language of scientific papers.
2. **Unprecedented Scale**: Llama 3.1 was trained on a significantly larger scale than its predecessors and some contemporaries, boasting about 8 billion trainable parameters. This scale allows Llama 3.1 to adhere closely to the scaling laws that predict improved performance with larger model sizes, which means better understanding and generating capabilities compared to smaller models.
3. **Simplified Complexity Management**: Instead of opting for a more complex mixture-of-experts model, Llama 3.1 utilizes a standard dense Transformer architecture which prioritizes training stability and efficiency. The post-training processes are streamlined to use techniques like supervised fine-tuning, rejection sampling, and direct preference optimization. This approach avoids the pitfalls of more complex and less stable methods like reinforcement learning, leading to a more reliable and stable performance across diverse tasks.
4. **Multilingual Training**: By training on multilingual sources, Llama 3.1 isn't just learning various subjects but also different languages, which enhances its versatility and global applicability. This is especially important in a world where digital content and user interactions span numerous languages.

Reference: Vavekanand, Raja, and Kira Sam. "Llama 3.1: An In-Depth Analysis of the Next-Generation Large Language Model." (2024).

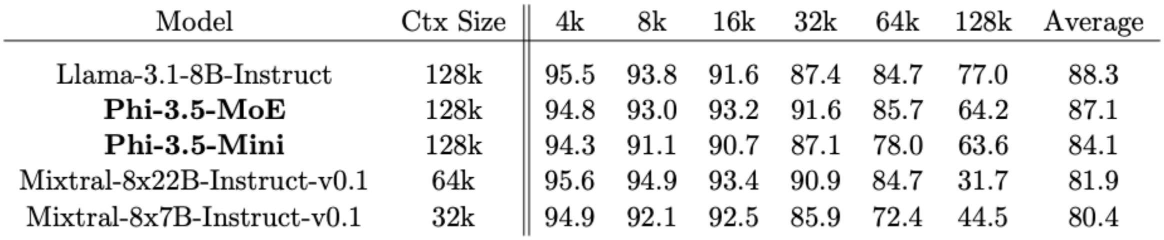
2.

The superior performance of the Phi-3.5 mini model in React Prompting on Math-based questions compared to Llama 3.1-8B and Gemma can be inferred from several enhancements and optimizations mentioned in the text:

1. **Enhanced Multilingual Capabilities**: Phi-3.5 mini includes improvements in handling multilingual data, which could contribute to better understanding and generation of diverse mathematical terminology and formulations that might appear in various languages. This broad linguistic coverage is crucial for math-based questions which can sometimes incorporate terms or concepts from different languages.
2. **Long-Context Data Handling**: The introduction of the long-rope method and mixed context window approach in Phi-3.5 mini allows the model to handle longer text sequences effectively. This ability to manage extended contexts without performance loss is particularly beneficial for complex math-based questions that require maintaining large amounts of contextual information for accurate problem-solving.
3. **Mid-Training Data Expansion**: The Phi-3.5 mini was specifically enhanced with more multilingual and long-text data during mid-training. This approach likely helps the model to better

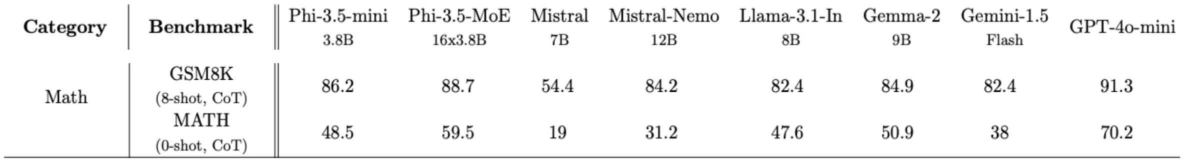
adapt and respond to nuanced and complex mathematical questions, as it would have been exposed to a wider variety of mathematical expressions and problems during its training.

4. **Comparative Model Capacity and Optimization**: Despite being a smaller model, Phi-3.5 mini's performance is enhanced by its architectural optimizations which make it capable of matching or even surpassing the performance of larger models. This indicates efficient use of its model capacity and targeted optimizations that benefit math-based tasks.



Model	Ctx Size	4k	8k	16k	32k	64k	128k	Average
Llama-3.1-8B-Instruct	128k	95.5	93.8	91.6	87.4	84.7	77.0	88.3
Phi-3.5-MoE	128k	94.8	93.0	93.2	91.6	85.7	64.2	87.1
Phi-3.5-Mini	128k	94.3	91.1	90.7	87.1	78.0	63.6	84.1
Mixtral-8x22B-Instruct-v0.1	64k	95.6	94.9	93.4	90.9	84.7	31.7	81.9
Mixtral-8x7B-Instruct-v0.1	32k	94.9	92.1	92.5	85.9	72.4	44.5	80.4

Table 2: Comparison results on RULER benchmark.



Category	**Benchmark**	Phi-3.5-mini 3.8B	Phi-3.5-MoE 16x3.8B	Mistral 7B	Mistral-Nemo 12B	Llama-3.1-In 8B	Gemma-2 9B	Gemini-1.5 Flash	GPT-4o-mini
Math	GSM8K (8-shot, CoT)	86.2	88.7	54.4	84.2	82.4	84.9	82.4	91.3
	MATH (0-shot, CoT)	48.5	59.5	19	31.2	47.6	50.9	38	70.2

Reference:

Abdin, M., Jacobs, S. A., Awan, A. A., Aneja, J., Awadallah, A., Awadalla, H., Bach, N., Bahree, A., Bakhtiari, A., Behl, H., Benhaim, A., Bilenko, M., Bjorck, J., Bubeck, S., Cai, M., Mendes, C. C. T., Chen, W., Chaudhary, V., Chopra, P., . . . Zhou, X. (2024). PHI-3 Technical Report: A highly capable language model locally on your phone. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2404.14219