LLM_A2
2021575

Inference Time

Zero Shot Prompting

Gemma = 342.8378760814667
LLama = 1677.2967312335968
Phi = 1369.6809136867523

Instruction Prompting

Gemma = 286.11159682273865
LLama = 1395.4024691581726
Phi = 1254.0722184181213

ReAct Prompting

Gemma = 353.0435838699341
LLama = 1678.087652431654
Phi = 1352.674598431756

General Findings:

- **Inference Speed**:
    - **Fastest**: Google GEMMA-2B is the smallest model and performs the fastest in inference time.
    - **Balanced**: Microsoft Phi-3.5B provides a middle ground in terms of speed.
    - **Slowest**: Meta-LLaMA-8B takes the longest time for inference but offers the most sophisticated responses.
- **Inference Accuracy**:
    - **Most Accurate**: Meta-LLaMA-8B stands out for instruction-following and knowledge-intensive tasks due to its larger size.
    - **Moderately Accurate**: Microsoft Phi-3.5B offers good instruction-following accuracy but may struggle with more complex prompts.

    - **Lower Accuracy**: Google GEMMA-2B may not perform as well in English tasks as the other two but is optimized for multilingual contexts.

Reasons for Performance Differences:

- **Model Size**: Larger models like Meta-LLaMA-8B naturally perform better in terms of accuracy due to the number of parameters, but they come at the cost of slower inference times.
- **Training Objective**: Models fine-tuned for instruction-following (Phi-3.5B) will be more accurate in general-purpose tasks but fall behind more specialized or larger models in complex tasks.