CSE665: Large Language Models
Assignment 2
Trade off between Model size, Prompt type, Time Taken and Quality

Report

Prompt type:

Zero Shot : The prompting technique to directly instruct the model to perform the task in one prompt without any additional examples and demonstrations.

```
Choose the answer to the given question from below options.\nQuestion:{question}\nOption1:{options[0]}\nOption2:{options[1]}\nOption 3: {options[2]}\nOption 4: {options[3]}.
```

Chain of Thought(Zero Shot): This prompting technique enables the complex reasoning capabilities of the model by directly prompting the model to think step by step.

```
"Choose the answer to the given question from below options.\nQuestion:{question}\nOption1:{options[0]}\nOption2:{options[1]}\nOption 3: {options[2]}\nOption 4: {options[3]}\nThink step by step."
```

Re_Act Prompting: Querying the model to generate both intermediate reasoning steps as well as task specific action steps. This allows the model to interface with and gather information from external sources.

```
"Choose the answer to the given question from below options.\nQuestion:{question}\nOption1:{options[0]}\nOption2:{options[1]}\nOption 3: {options[2]}\nOption 4: {options[3]}\n\nLet's think step by step:\nSteRESULTSp 1 Reasoning: "
```

Further Additions to prompts:

```
action_prompt = reasoning_step_1 + "\nStep 1 Action: "
action_step_1 = generate_response(action_prompt,model,tokenizer,100)

reasoning_step_2 = f"{action_step_1}\n\nStep 2 Reasoning: "
reasoning_step_2_output = generate_response(reasoning_step_2,model,tokenizer,100)

action_step_2 = reasoning_step_2_output + "\nStep 2 Action: "
action_step_2_output = generate_response(action_step_2,model,tokenizer,100)

final_output = f"{reasoning_step_1}\n{action_step_1}\n{reasoning_step_2_output}\n{action_step_2_output}"
```

Source : https://www.promptingguide.ai/

RESULTS

Model : google/gemma-2b-it

Zero Shot:

```
Inference time for zero shot for Gemma: 2.127796 seconds
Accuracy of Gemma for zero shot: 0.39
```

Chain of Thought(Zero Shot):

```
Inference time for chain of thought for Gemma: 5.222095 seconds
Accuracy of Gemma  for cot: 0.21
```

ReAct:

```
Inference time for Re_Act for Gemma: 12.574238 seconds
Accuracy of Gemma for Re_Act: 0.21
```

Model : microsoft/Phi-3.5-mini-instruct

Zero Shot:

```
Inference time for zero shot for Phi-3.5: 4.787268 seconds
Accuracy of Phi-3.5 for zero shot: 0.27
```

Chain of Thought(Zero Shot):

```
Inference time for chain of thought for Phi-3.5: 13.699061 seconds
Accuracy of Phi-3.5  for cot: 0.11
```

Model : meta-llama/Meta-Llama-3.1-8B-Instruct

Zero Shot:

```
 Inference time for zero shot for Llama-3.1: 29.821782 seconds
 Accuracy of Llama-3.1 for zero shot: 0.14
```

Chain of Thought(Zero Shot):

```
Inference time for chain of thought for Llama-3.1: 38.319127 seconds
Accuracy of Llama-3.1  for cot: 0.04
```

Analysis and Trade off :

Meta-Llama-3.1-8B (the largest) had the highest inference times but the lowest accuracy in both zero-shot and chain-of-thought tasks**.** This shows that large model size alone does not guarantee best responses without specific fine-tuning and instructions.
Gemma-2B was the fastest model of them all , this represents the fact that smaller model sizes have faster computations and inference times. But this faster computations results in lower accuracies in complex tasks such as Chain-of-Thought tasks.

Gemma-2B had the best accuracy in Zero Shot prompting which shows that it is better tuned for tasks when it comes to complex college level mathematics.

Finally , the experiment suggests that Gemma-2B might be the best for less complex tasks while for long reasoning required computations larger models like Llama-3.1 might be the one with additional fine-tuning.

Output Quality:
Gemma-2B : The model first gives the answer straightforwardly and then tries to explain the solution. Sometimes the answer is completely different from the options and sometimes it is perfectly alright.

Phi-3.5-mini : The model first provides the answer and then tries to explain the question first and then gives the full step by step solution.

Llama-3.1 : The model starts with the explanation often , sometimes adding its own options to the question asked.The output tends to be overly verbose but lacks clear logical flow. It often goes off-track, generating irrelevant or incorrect information.

Papers and Technical reports:

Gemma -2B:

The model showed fastest inference times compared to larger models such as Llama-3.1. One reason for this can be that Gemma 2B was built with optimizations such as Knowledge Distillation. This technique helps the model to perform better even with small parameter count through learning from larger models.

Paper: Gemma 2: Improving Open Language Models at a Practical Size

Phi-3.5 : Even with larger model size , this model showed comparable inference times , the reason being the internal optimizations for instruction based tasks for the model can allow it to perform well in tasks with zero shot and CoT prompts.

Paper: Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone

Llama-3.1 : The model is fine-tuned more towards instruction-following and general-purpose generation, which may not align perfectly with chain-of-thought reasoning.

Paper : The Llama 3 Herd of Models