LLM - Assignment 2 Report

Manav Mittal - 2021538

Prompts Used

Zero Shot

```
Choose the answer to the given question from below options.
No explaination required.

    {question}

    Option 1: {options[0]}
    Option 2: {options[1]}
    Option 3: {options[2]}
    Option 4: {options[3]}

    Answer:
```

Chain of Thought

```
Choose the answer to the given question from below options.
Explain the solution step by step.

    {question}

    Option 1: {options[0]}
    Option 2: {options[1]}
    Option 3: {options[2]}
    Option 4: {options[3]}

    Answer:
```

ReAct

```
Answer the following questions as best you can.

        Use the following format:

        Question: the input question you must answer
        Thought: you should always think about what to do
        Action: the action to take
        Action Input: the input to the action
        Observation: the result of the action
        ... (this Thought/Action/Action Input/Observation c
an repeat N times)
        Thought: I now know the final answer
        Final Answer: the final answer to the original inpu
t question



        Choose the answer to the given question from below
options. I dont need explanation.
        Keep the Answer in format "Option [x]", where x is
1,2,3,4.

        {question}

        Option 1: {options[0]}
        Option 2: {options[1]}
        Option 3: {options[2]}
        Option 4: {options[3]}

        Answer:
```

Inference Time

I have calculated average inference time over 10 samples for each case.

Model	Zero Shot	Chain of Thought	ReAct
gemma-2b	1.3 s	7.8 s	5.5 s
phi-3.5	1.4 s	54.8 s	31 s

Model	Zero Shot	Chain of Thought	ReAct
llama-3.1	1.9 s	44.7 s	31.2 s

1. Trade-off between Inference time and Model size

   In case of zero shot prompting we can clearly see that larger the model size more is the inference time. Similar trend can be seen other type of prompting techniques Phi being an exception in case of CoT prompting technique. The general trend is **Llama ≥ Phi > Gemma**.

2. Trade-off between Inference time prompt type

   We can clearly see that inference time is highly dependent upon the prompt type. The trend in this case is **CoT > ReAct > Zero Shot**. The reason behind this behaviour is Chain of Thought is more explanatory and solution driven while ReAct is not explanation driven which leads to less number of token generation and in case of Zero shot we prompted the model with “No explanation required” thus very few tokens were generated leading to least inference time.

Accuracy

Model	Zero Shot	Chain of Thought	ReAct
gemma-2b	25%	29%	26%
phi-3.5	28%	41%	32%
llama-3.1	35%	37%	31%

1. Trade-off between Accuracy (Output quality) and Model Size

   The usual trend should be the larger the model the better is the accuracy and output quality. Similar trend can seen here also larger the model size better the accuracy. Gemma has least number of parameters i.e 2 billion have least accuracy among the three. Phi being having 3.5 billion parameter have accuracy better than gemma and worse than llama (in zero shot). Llama being largest model having 8 billion parameters is the better performing than gemma in every case and Phi in case of zero shot. The reason behind Phi being better than Llama in CoT and ReAct is discussed later in the report. So the trends can be seen here are **Llama > Gemma Phi > Gemma**

   .

2. Trade-off between Accuracy (Output quality) and Prompt Type

The clear trend we can see in this case is **CoT > ReAct > Zero Shot.** The exception being Llama in case ReAct. It performed poorer in comparison to its zero-shot. The reason behind that can be in case of zero shot no explanations were given by the model while in case of ReAct model will generate lengthy outputs and in several cases due to this max-token-limit was reached and Llama was terminated before generating the final answer.

3. Trade-off between Accuracy (Output quality) and inference time

   Comparing both the tables we can clearly see that the more is the inference time more is the accuracy and output quality. As we can see CoT has the highest inference time but it also has the highest accuracy. The reason behind behind this is more tokens are being generated in CoT consisting the reasoning leading to better next word generation but more inference time.

Output Quality

Commenting on output quality (in terms of quality of the response not accuracy) I would like to sate the trend as follows **Llama ≥ Phi > Gemma**. In almost every case the quality of Gemma was worse than Llama and Phi. Though the quality of the output of Llama and Phi were comparable. Llamas responses were more detailed if have better compute resources it may have performed better. Phi's output were appropriately explained though if demanded more explanation accuracy can improve.

Analysis and Reasoning

Benchmark	metric	Gemma-1 2B	Gemma-2 2B	Mistral 7B	LLaMA-3 8B	Gemma-1 7B	Gemma-2 9B
MMLU	5-shot	42.3	**52.2**	62.5	66.6	64.4	**71.3**
ARC-C	25-shot	48.5	**55.7**	60.5	59.2	61.1	**68.4**
GSM8K	5-shot	15.1	**24.3**	39.6	45.7	51.8	**68.6**
AGIEval	3-5-shot	24.2	**31.5**	44.0†	45.9†	44.9†	**52.8**
DROP	3-shot, F1	48.5	**51.2**	63.8*	58.4	56.3	**69.4**
BBH	3-shot, CoT	35.2	**41.9**	56.0*	61.1*	59.0*	**68.2**
Winogrande	5-shot	66.8	**71.3**	78.5	76.1	79.0	**80.6**
HellaSwag	10-shot	71.7	**72.9**	**83.0**	82.0	82.3	81.9
MATH	4-shot	11.8	**16.0**	12.7	-	24.3	**36.6**
ARC-e	0-shot	73.2	**80.6**	80.5	-	81.5	**88.0**
PIQA	0-shot	77.3	**78.4**	**82.2**	-	81.2	81.7
SIQA	0-shot	49.7	**51.9**	47.0*	-	51.8	**53.4**
Boolq	0-shot	69.4	**72.7**	83.2*	-	83.2	**84.2**
TriviaQA	5-shot	53.2	**60.4**	62.5	-	63.4	**76.6**
NQ	5-shot	12.5	**17.1**	23.2	-	23.0	**29.2**
HumanEval	pass@1	**22.0**	20.1	26.2	-	32.3	**40.2**
MBPP	3-shot	29.2	**30.2**	40.2*	-	44.4	**52.4**
Average (8)		44.0	**50.0**	61.0	61.9	62.4	**70.2**
Average (all)		44.2	**48.7**	55.6	-	57.9	**64.9**

Fig 1 - Gemma 2b Evaluation [2]

Prompt Method[a]	HotpotQA (EM)	Fever (Acc)
Standard	28.7	57.1
CoT (Wei et al., 2022)	29.4	56.3
CoT-SC (Wang et al., 2022a)	33.4	60.4
Act	25.7	58.9
ReAct	27.4	60.9
CoT-SC → ReAct	34.2	**64.6**
ReAct→ CoT-SC	**35.1**	62.0
Supervised SoTA[b]	67.5	89.5

Fig 3 - ReAct Prompting Comparison [4]

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
Popular	Arena Hard	37	37.9	18.1	39.4	25.7	42
	BigBench Hard CoT (0-shot)	69	79.1	33.4	60.2	63.4	63.5
MMLU	MMLU (5-shot)	69	78.9	60.3	67.2	68.1	71.3
	MMLU-Pro (0-shot, CoT)	47.5	54.3	18	40.7	44	50.1
Reasoning	ARC Challenge (10-shot)	84.6	91.0	77.9	84.8	83.1	89.8
	BoolQ (2-shot)	78	84.6	80.5	82.5	82.8	85.7
	GPQA (0-shot, CoT)	27.2	36.8	15.6	28.6	26.3	29.2
	HellaSwag (5-shot)	69.4	83.8	71.6	76.7	73.5	80.9
	OpenBookQA (10-shot)	79.2	89.6	78	84.4	84.8	89.6
	PIQA (5-shot)	81	88.6	73.4	83.5	81.2	83.7
	Social IQA (5-shot)	74.7	78.0	73	75.3	71.8	74.7
	TruthfulQA (10-shot,MC2)	64	77.5	64.7	68.1	69.2	76.6
	WinoGrande (5-shot)	68.5	81.3	58.1	70.4	64.7	74
Multilingual	MI MMLU (5-shot)	55.4	69.9	47.4	58.9	56.2	63.8
	MGSM (0-shot CoT)	47.9	58.7	31.8	63.3	56.7	76.4
Math	GSM8K (8-shot, CoT)	86.2	88.7	54.4	84.2	82.4	84.9
	MATH (0-shot, CoT)	48.5	59.5	19	31.2	47.6	50.9

Fig 2 - Phi 3.5 Evaluation [3]

Category / Benchmark	Llama 3.1 8B	Gemma 2 9B IT	Mistral 7B Instruct	Llama 3.1 70B	Mixtral 8x22B Instruct	GPT 3.5 Turbo
General MMLU (0-shot, CoT)	73.0	72.3 (5-shot, non-CoT)	60.5	86.0	79.9	69.8
MMLU PRO (5-shot, CoT)	48.3	-	36.9	66.4	56.3	49.2
IFEval	80.4	73.6	57.6	87.5	72.7	69.9
Code HumanEval (0-shot)	72.6	54.3	40.2	80.5	75.6	68.0
MBPP EvalPlus (base) (0-shot)	72.8	71.7	49.5	86.0	78.6	82.0
Math GSM8K (8-shot, CoT)	84.5	76.7	53.2	95.1	88.2	81.6
MATH (0-shot, CoT)	51.9	44.3	13.0	68.0	54.1	43.1

Fig 4 - Llama 3.1 8B Evaluation [1]

1. Comparing ReAct and CoT

   In Fig. 3 we can clearly see that CoT has better performance than ReAct on QA task. CoT has got a EM of 29.4 while on the hand ReAct got EM of 27.4 on HotpotQA. For these evaluations PaLM-540B was used. This similar trend can also be seen in our case. Since our task is QA we are getting better accuracies in case of CoT than ReAct.

2. Phi 3.5 v/s Llama 3.1

   In Fig. 2 we can clearly see that Phi-3.5 mini (CoT) has better performance than Llama-3.1 on Maths dataset. This behaviour can also be seen in out case since we are getting significantly better with Phi than Llama in case of CoT. While in case of ReAct we are getting close performance for both Llama and Phi still Phi performed better since ReAct also includes

explanation and reasoning. On the other hand in case of Zero-shot Llama performed better than Phi.

3. Llama 3.1 v/s Gemma

   In Fig. 4 we can clearly see that Llama 3.1 8B performs significantly better than Gemma 2 9B IT. And in Fig. 1 we can clearly see that Gemma 2 2B performs significantly poor in comparison to Gemma 2 9B on the same datasets. Thus we can conclude that Llama 3.1 8B has better performance than Gemma 2 2B for mathematical data. This observation we can also see in our analysis too Gemma 2 has performed significantly poorer in our case in comparison to Llama 3.1.

4. Phi 3.5 v/s Gemma 2

   From the above points 2 and 3 we can conclude that Phi 3.5 has better performance than Gemma 2 for mathematical data.

Conclusion

1. If inference time is the priority then Gemma is the model to choose since it has a very significantly lower inference time in comparison to the other 2. It also has less number of parameters making it more suitable for less performance devices (e.g IoT, Mobile phones).
2. If performance is the priority and inference time does not matter much Phi 3.5 is the best model to choose. Since it performs very good on CoT and ReAct techniques.
3. If one wants something in the middle they can use Llama 3.1 with Zero-shot prompting. Since it has approximately same inference time as other two and performs better than Gemma in all categories and Phi in Zero-shot.

References

[1] - https://ai.meta.com/blog/meta-llama-3-1/

[2] - Gemma 2: Improving Open Language Models at a Practical Size

[3] - Phi-3 Technical Report

[4] - ReAct Prompting - Prompt Engineering Guide