

CSE665: Large Language Models

Assignment 2

Trade off between Model size, Prompt type, Time Taken and Quality

Github:

https://github.com/skyscrappers/LLM/tree/main/LLM_A2

Results:

NOTE: All Models are loaded in 4-bit Quantization.

Average Inference time comparison:

	Llama 3.1	Phi-3.5	gemma-2b
Zero-Shot	1.5209	0.8302	1.1779
Chain of Thought (Zero-Shot)	50.6914	39.8789	5.3387
Reactive Prompting	31.4786	29.4721	3.8054

Model accuracy comparison:

	Llama 3.1	Phi-3.5	gemma-2b
Zero Shot	35%	33%	31%
Chain of Thought (Zero Shot)	37%	39%	33%
Reactive Prompting	32%	34%	32%

Prompts:

Zero-Shot:

Choose the answer to the given question from below options.

Question: {ds['test'][i]['question']}

option 1: {ds['test'][i]['choices'][0]}

option 2: {ds['test'][i]['choices'][1]}

option 3: {ds['test'][i]['choices'][2]}

option 4: {ds['test'][i]['choices'][3]}

Just output the correct option number like: option 1, option 2, option 3, option 4 without any explanation.

Chain of Thought:

Choose the answer to the given question from below options.

Question: {ds['test'][i]['question']}

option 1: {ds['test'][i]['choices'][0]}

option 2: {ds['test'][i]['choices'][1]}

option 3: {ds['test'][i]['choices'][2]}

option 4: {ds['test'][i]['choices'][3]}

Explain step by step and then Output the correct option number like: option 1, option 2, option 3, option 4.

ReAct:

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 can repeat N times)

Thought: I now know the final answer

Final Answer: the final answer to the original input question

Choose the answer to the given question from below options;

Question: {ds['test'][i]['question']}

option 1: {ds['test'][i]['choices'][0]}

option 2: {ds['test'][i]['choices'][1]}

option 3: {ds['test'][i]['choices'][2]}

option 4: {ds['test'][i]['choices'][3]}

Make sure to give output in format Option number. Example if Option 1 is

correct give output "Final Answer: Option 1"

1. Model size vs Inference time:

As the Model size (number of parameters) decreases i.e; **Llama 3.1 > Phi-3.5 > gemma-2b**, the average inference time decreases and Inference speed increases for each prompting technique, which should actually be the case as number of computation increases with model size.

Only gemma-2b takes a bit more inference time than Phi-3.5 in case of zero shot.

Higher the Inference time implies lower the inference speed hence,

Order of inference speed : **Llama 3.1 < Phi-3.5 < gemma-2b**

2. Prompt technique vs Inference time:

Also, the order of inference time based on prompting technique for each model is:

Chain of Thought > Reactive prompting > Zero-Shot.

CoT requires the model to explain its reasoning step-by-step before providing the final answer. This forces the model to generate multiple intermediate outputs as it processes the question logically, which greatly increases the computational load and, consequently, the inference time.

Reactive prompting requires the model to dynamically interact with the prompt context, potentially adapting its reasoning based on new information or constraints. While this involves some level of reasoning and contextual adaptation, it is less

resource-intensive than CoT, as it doesn't require the same depth of explanation for each step.

Zero-shot prompting involves simply selecting the correct answer without any explanation or reasoning. Since the model doesn't need to engage in complex reasoning or justify its answer, it only needs to process the question and output the final result, leading to the shortest inference time.

3. Model size vs accuracy:

A general trend is that the accuracy increases as the model size increase which is also the case for provided results except for CoT technique.

LLAMA-3.1-8B-Instruct:

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
Reasoning						
ARC Challenge (0-shot)	83.4	87.6	74.2	94.8	88.7	83.7
GPQA (0-shot, CoT)	32.8	-	28.8	46.7	33.3	30.8
Tool use						
BFCL	76.1	-	60.4	84.8	-	85.9
Nexus	38.5	30.0	24.7	56.7	48.5	37.2
Long context						
ZeroSCROLLS/QuALITY	81.0	-	-	90.5	-	-
InfiniteBench/En.MC	65.1	-	-	78.2	-	-
NIH/Multi-needle	98.8	-	-	97.5	-	-
Multilingual						
Multilingual MGSM (0-shot)	68.9	53.2	29.9	86.9	71.1	51.4

Gemma-2b:

Benchmark	metric	Gemma-1 2B	Gemma-2 2B	Mistral 7B	LLaMA-3 8B	Gemma-1 7B	Gemma-2 9B	Gemma-2 27B
MMLU	5-shot	42.3	52.2	62.5	66.6	64.4	71.3	75.2
ARC-C	25-shot	48.5	55.7	60.5	59.2	61.1	68.4	71.4
GSM8K	5-shot	15.1	24.3	39.6	45.7	51.8	68.6	74.0
AGIEval	3-5-shot	24.2	31.5	44.0 [†]	45.9 [†]	44.9 [†]	52.8	55.1
DROP	3-shot, F1	48.5	51.2	63.8*	58.4	56.3	69.4	74.2
BBH	3-shot, CoT	35.2	41.9	56.0 [°]	61.1 [°]	59.0 [°]	68.2	74.9
Winogrande	5-shot	66.8	71.3	78.5	76.1	79.0	80.6	83.7
HellaSwag	10-shot	71.7	72.9	83.0	82.0	82.3	81.9	86.4
MATH	4-shot	11.8	16.0	12.7	-	24.3	36.6	42.3
ARC-e	0-shot	73.2	80.6	80.5	-	81.5	88.0	88.6
PIQA	0-shot	77.3	78.4	82.2	-	81.2	81.7	83.2
SIQA	0-shot	49.7	51.9	47.0*	-	51.8	53.4	53.7
Boolq	0-shot	69.4	72.7	83.2*	-	83.2	84.2	84.8
TriviaQA	5-shot	53.2	60.4	62.5	-	63.4	76.6	83.7
NQ	5-shot	12.5	17.1	23.2	-	23.0	29.2	34.5
HumanEval	pass@1	22.0	20.1	26.2	-	32.3	40.2	51.8
MBPP	3-shot	29.2	30.2	40.2*	-	44.4	52.4	62.6
Average (8)		44.0	50.0	61.0	61.9	62.4	70.2	74.4
Average (all)		44.2	48.7	55.6	-	57.9	64.9	69.4

PHI Technical Report:

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
Popular	Arena Hard	37	37.9	18.1	39.4	25.7	42	55.2	75
	BigBench Hard CoT (0-shot)	69	79.1	33.4	60.2	63.4	63.5	66.7	80.4
MMLU	MMLU (5-shot)	69	78.9	60.3	67.2	68.1	71.3	78.7	77.2
	MMLU-Pro (0-shot, CoT)	47.5	54.3	18	40.7	44	50.1	57.2	62.8
Reasoning	ARC Challenge (10-shot)	84.6	91.0	77.9	84.8	83.1	89.8	92.8	93.5
	BoolQ (2-shot)	78	84.6	80.5	82.5	82.8	85.7	85.8	88.7
	GPQA (0-shot, CoT)	27.2	36.8	15.6	28.6	26.3	29.2	37.5	41.1
	HellaSwag (5-shot)	69.4	83.8	71.6	76.7	73.5	80.9	67.5	87.1
	OpenBookQA (10-shot)	79.2	89.6	78	84.4	84.8	89.6	89	90
	PIQA (5-shot)	81	88.6	73.4	83.5	81.2	83.7	87.5	88.7
	Social IQA (5-shot)	74.7	78.0	73	75.3	71.8	74.7	77.8	82.9
	TruthfulQA (10-shot, MC2)	64	77.5	64.7	68.1	69.2	76.6	76.6	78.2
	WinoGrande (5-shot)	68.5	81.3	58.1	70.4	64.7	74	74.7	76.9
Multilingual	Mi MMLU (5-shot)	55.4	69.9	47.4	58.9	56.2	63.8	77.2	72.9
	MGSM (0-shot CoT)	47.9	58.7	31.8	63.3	56.7	76.4	75.8	81.7
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
Long context	Qasper	41.9	40.0	31.4	30.7	37.2	13.9	43.5	39.8
	SQuALITY	24.3	24.1	25.9	25.8	26.2	0	23.5	23.8
Code	HumanEval (0-shot)	61.5	70.7	35.4	63.4	66.5	61	74.4	86.6
	MBPP (3-shot)	68.6	80.8	50.4	68.1	69.4	69.3	77.5	84.1
Average		61.1	69.2	48.5	61.3	61.0	63.3	68.5	74.9

In the <https://arxiv.org/pdf/2404.14219> Table 3 shows that **Phi-3.5** (Image of PHI Technical Report) performs better than **Llama-3.1** and **gemma-2b** in the Category - Math which can also be seen in LLAMA-3.1-8B-Instruct Image same is also the case in table of results for Chain of Thought prompting technique, but the **gemma-2b** we use contains 2-billion parameters which is the reason of it's lowest accuracy.

However, we can observe that **Llama 3.1** performs better than **Phi-3.5** and **gemma-2b** in case of zeroshot and reactive prompting which should be the general case based on Model size and Prompting technique.

4. Prompting technique vs accuracy:

The general trend observed for accuracy based on different prompting techniques is:

Chain of Thought > Reactive Prompting > Zero-Shot except for the Llama 3.1 which 32% in reactive prompting and 35% in Zeroshot.

However these results also depends on the model parameters and tokens used by model while generating the responses, for **Llama 3.1** with **reactive prompting** I observed that Response is incomplete in many prompt and hence leading to the deterioration in the accuracy.

5. Model size and prompting technique vs Output quality:

LLAMA 3.1:

Output for LLama 3.1 is consistent for each prompting technique:

ZeroShot: Option number (Like option 1, option 2)

Chain of Thought:

To solve this problem, let's analyze the given equation: $e^x + x - 2 = 0$.

First, we'll consider the interval $[0, 1]$. To find the number of real solutions in this interval, we can use the concept of the Lambert W function or analyze the graph of the function $f(x) = e^x + x - 2$.

The Lambert W function is defined as the inverse function of $f(x) = xe^x$. It's a transcendental function that can be used to solve equations involving exponentials.

However, we can also use graphical analysis or numerical methods to understand the behavior of the function.

For the interval $[0, 1]$, let's analyze the function:

$$f(0) = e^0 + 0 - 2 = -2$$

$$f(1) = e^1 + 1 - 2 \approx 0.71828$$

Since the function is continuous and differentiable, we can use the intermediate value theorem, which states that if a function is continuous on a closed interval $[a, b]$, then it takes on every value between $f(a)$ and $f(b)$.

In this case, since $f(0) < 0$ and $f(1) > 0$, we can conclude that there exists at least one real root in the interval $(0, 1)$.

To find the exact number of real roots in the interval $[0, 1]$, we can use numerical methods or graphical analysis. However, since we don't have the exact number of real roots, let's analyze the options:

Option 1: $k = 0$ and $n = 1$

Option 2: $k = 1$ and $n = 0$

Option 3: $k = n = 1$

Option 4: $k > 1$

Based on our analysis, we can see that option 2 is the most plausible answer. The function $f(x) = e^x + x - 2$ has at least one real root in the interval $[0, 1]$, and there are no real roots outside this interval.

Therefore, the correct option is:

option 2.

Reactive Prompt:

Parsing LLM output produced both a final answer and a parse-able action:: Thought: To solve this problem, we need to find the number of real solutions of the equation $e^x + x - 2 = 0$ in the interval $[0, 1]$ and outside this interval.

Action: Use the Calculator to find the number of real solutions.

Action Input: The equation $e^x + x - 2 = 0$.

Observation: The equation $e^x + x - 2 = 0$ can be rewritten as $e^x + (x - 1) = 1$. We know that e^x is always greater than or equal to 0, and the minimum value of e^x is 1 when $x = 0$.

Thought: Since e^x is always greater than or equal to 0, we can conclude that $e^x + (x - 1)$ will be greater than 1 when $x > 1$, and less than 1 when $x < 1$.

Action: Use the Calculator to find the number of real solutions in the interval $[0, 1]$.

Action Input: The equation $e^x + x - 2 = 0$ in the interval $[0, 1]$.

Observation: By analyzing the graph of the function $e^x + x - 2$, we can see that there is only one real solution in the interval $[0, 1]$.

Thought: Since there is only one real solution in the interval $[0, 1]$, we can conclude that there are no real solutions outside this interval.

Action: Use the Calculator to verify the number of real solutions.

Action Input: The equation $e^x + x - 2 = 0$.

Observation: The equation $e^x + x - 2 = 0$ has only one real solution in the interval $[0, 1]$, and there are no real solutions outside this interval.

Thought: I now know the final answer.

Final Answer: option 2

PHI-3.5:

ZeroShot: Option number (Like option 1, option 2)

Chain of Thought:

To solve this problem, let's analyze the given equation:

$$e^x + x - 2 = 0$$

We need to find the number of real solutions (k) in the interval [0, 1], and the number of real solutions (n) that are not in [0, 1].

Step 1: Check for solutions in the interval [0, 1]

We can use the Intermediate Value Theorem, which states that if a continuous function has values of opposite signs at the endpoints of an interval, then there exists at least one root within that interval.

Let's evaluate the function at the endpoints of the interval [0, 1]:

$$f(0) = e^0 + 0 - 2 = 1 - 2 = -1$$

$$f(1) = e^1 + 1 - 2 = e - 1$$

Since $e > 2$, $f(1) > 0$. We have opposite signs at the endpoints ($f(0) = -1$ and $f(1) > 0$), which means there is at least one root in the interval [0, 1].

Step 2: Check for solutions outside the interval [0, 1]

To check for solutions outside the interval, we can analyze the behavior of the function at the endpoints of the interval [0, 1] and its derivative.

The derivative of the function is:

$$f'(x) = e^x + 1$$

At $x = 0$, $f'(0) = e^0 + 1 = 2 > 0$, which means the function is increasing at $x = 0$.

...

option 2: $k = 1$ and $n = 0$

Output: option 2.

Reactive Prompt:

Parsing LLM output produced both a final answer and a parse-able action:: Thought: To solve this problem, I need to analyze the behavior of the function $f(x) = e^x + x - 2$

within the interval $[0, 1]$. I should check for sign changes and critical points to determine the number of real solutions.

Action: I will use the Calculator tool to evaluate the function at critical points and endpoints of the interval.

Action Input: I will evaluate $f(x) = e^x + x - 2$ at $x = 0$, $x = 1$, and approximate critical points within the interval $[0, 1]$.

Observation:

- $f(0) = e^0 + 0 - 2 = 1 - 2 = -1$
- $f(1) = e^1 + 1 - 2 \approx 2.71828 + 1 - 2 = 0.71828$

Now, I will check for critical points by taking the derivative of $f(x)$ and finding its roots.

$$f'(x) = e^x + 1$$

To find critical points, I need to solve $e^x + 1 = 0$.

Using the Calculator tool:

Action Input: Solve $e^x + 1 = 0$

Observation: There is no real solution for $e^x + 1 = 0$, as the exponential function is always positive.

Thought: Since there are no critical points within the interval $[0, 1]$, the function must either be always increasing or always decreasing. I will now check the values of $f(x)$ at the endpoints to determine the behavior.

$$f(0) = -1$$

...

Thought: I now know the final answer.

Final Answer: Option 2: k

GEMMA-2B:

ZeroShot: option 3: $k = n = 1$ The equation has one real solution in the interval

Chain of Thought:

*Option 3: $k = n = 1$ *

Explanation:

* The equation $e^x + x - 2 = 0$ has one real solution in the interval $[0, 1]$ because e^0

$= 1$ and 1 is in the interval.

* The equation has no real solutions outside the interval $[0, 1]$ because e^x is always positive for all real values of x .

* The number of real solutions that are not in $[0, 1]$ is equal to the number of real solutions in the interval $[0, 1]$, which is 1 .

* Therefore, $k = n = 1$.

Reactive Prompt:

Parsing LLM output produced both a final answer and a parse-able action:: Question: Up to isomorphism, how many additive abelian groups G of order 16 have the property that $x + x + x + x = 0$ for each x in G ?

Thought: I now know that the answer is 3 .

Action: Calculate the number of elements in G that satisfy the property.

Action Input: The number of elements in G that satisfy the property.

Observation: The result is 3 .

Thought: I now know that the final answer is 3 .

Final Answer: 3

Conclusion: Output format is more consistent for LLAMA-3.1 and PHI-3.5 than GEMMA-2B, again LLAMA-3.1 and PHI-3.5 are more robust models and have larger number of parameters than GEMMA-2B and hence they follow the instruction provided in the prompt more accurately than GEMMA-2B and hence leading to generate more consistent output.

However the output quality is more dependent on the 'prompting technique', LLAMA-3.1 and PHI-3.5 generate direct correct option in **zeroshot prompting**, on other hand GEMMA-2B is still generating the some line of explanation along with correct answers.

For, **Chain of Thought**, Output is consistent in all models, however in some cases LLAMA-3.1 tends to generate larger number of explanation steps and exhausting all tokens and leading to incomplete responses.

For **Reactive prompting** also the output is consistent for LLAMA-3.1 in majority cases but PHI-3.5 and GEMMA-2B generate some incomplete (Action and Thought process doesn't reach the final answer) and inconsistent (some responses contain the mathematical expression for final answer and some contain the correct option number) which leads to manual verification of correctness for calculation of accuracy.

FINAL CONCLUSION:

- **Gemma 2B:** Best suited for speed-critical applications but at the cost of reduced accuracy and useful for edge devices with limited computational resources because of less number of parameters.
- **Llama 3.1:** A better choice for tasks where accuracy is prioritized over speed, but it is computationally expensive in terms of resources and time.
- **Phi 3.5:** Offers a middle ground, excelling slightly in CoT reasoning but with a similar speed-accuracy trade-off as **Llama 3.1**, It is better than **Llama 3.1** in Math category as per PHI Technical Report.

References:

1. [Phi-3 Technical Report](#)
2. <https://ai.meta.com/blog/meta-llama-3-1/>
3. [Gemma 2: Improving Open Language Models at a Practical Size](#)
4. [React Prompting Guide](#)