CSE665: Large Language Models

Assignment 2

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

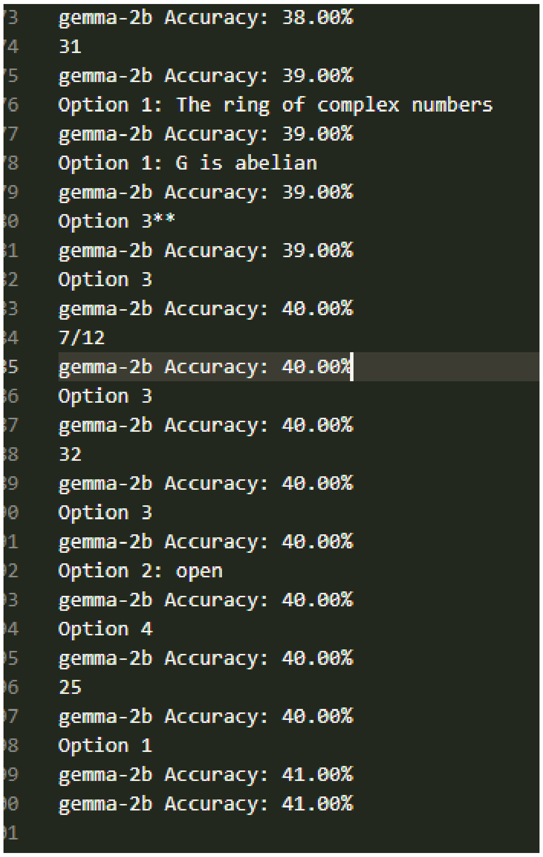
Code Explanation

- First, we load all three models with the appropriate tokenizer.
- For both Zero-Shot and Chain of Thought, we designed the prompt as given in the assignment.
- Now, we do this for all the 100 entries in the dataset.
- To check whether the answer is correct, we use regex matching to find the answer in the generated response and check whether it is the same as the actual answer in the dataset.

Results and Time taken for Inference

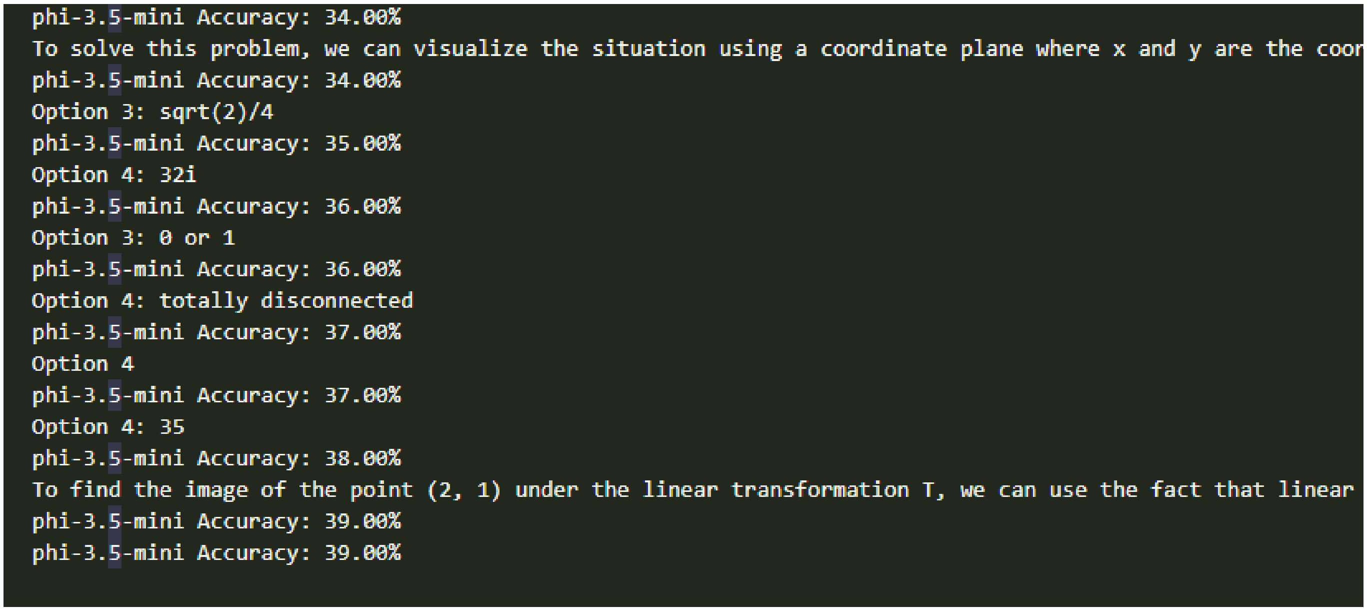
The accuracy for Zero-Shot was found to be:

- Gemma-2b-it - 41% (Time Taken - 18 minutes)



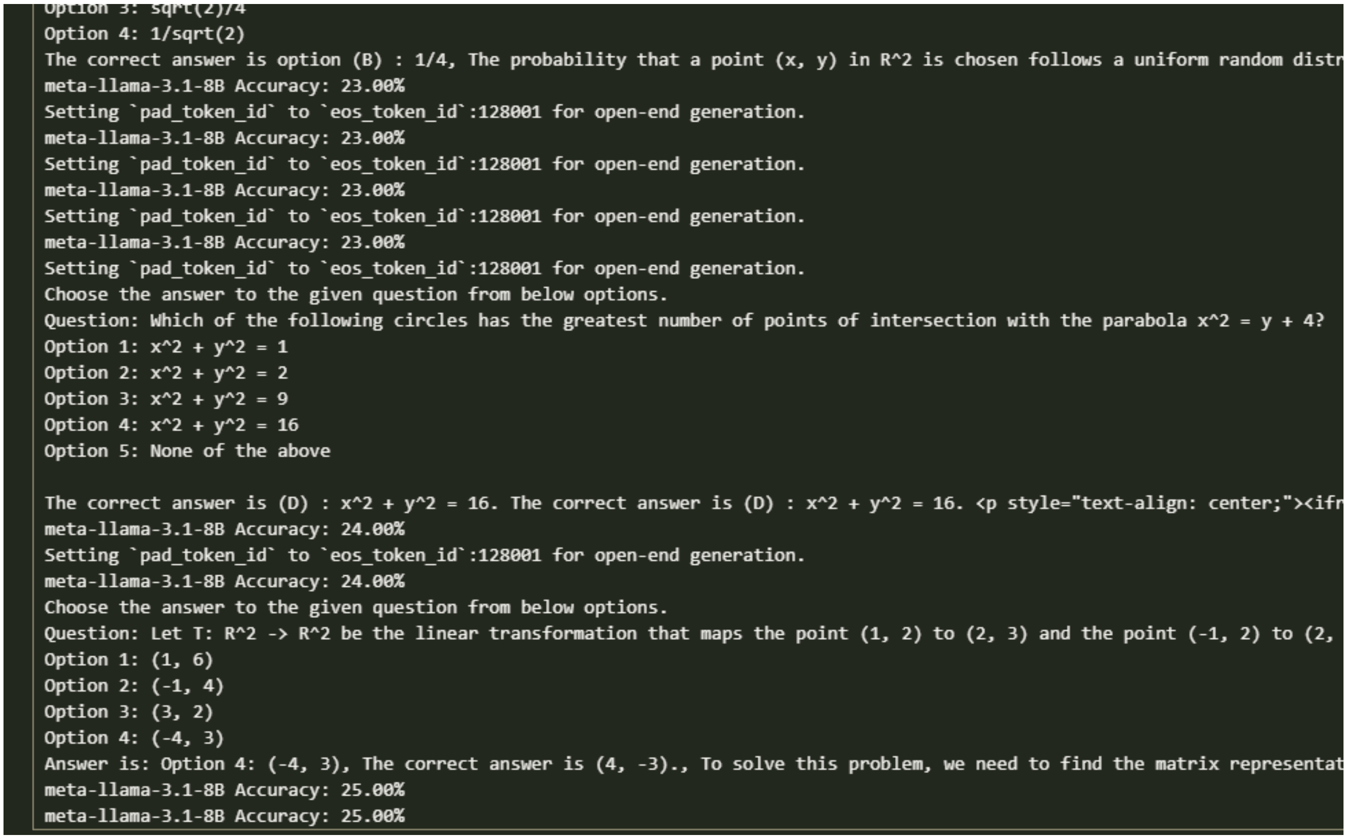
```
gemma-2b Accuracy: 38.00%
31
gemma-2b Accuracy: 39.00%
Option 1: The ring of complex numbers
gemma-2b Accuracy: 39.00%
Option 1: G is abelian
gemma-2b Accuracy: 39.00%
Option 3**
gemma-2b Accuracy: 39.00%
Option 3
gemma-2b Accuracy: 40.00%
7/12
gemma-2b Accuracy: 40.00%
Option 3
gemma-2b Accuracy: 40.00%
32
gemma-2b Accuracy: 40.00%
Option 3
gemma-2b Accuracy: 40.00%
Option 2: open
gemma-2b Accuracy: 40.00%
Option 4
gemma-2b Accuracy: 40.00%
25
gemma-2b Accuracy: 40.00%
Option 1
gemma-2b Accuracy: 41.00%
gemma-2b Accuracy: 41.00%
```

- Phi-3.5-mini-instruct - 39% (Time Taken - 27 minutes)



```
phi-3.5-mini Accuracy: 34.00%
To solve this problem, we can visualize the situation using a coordinate plane where x and y are the coor
phi-3.5-mini Accuracy: 34.00%
Option 3: sqrt(2)/4
phi-3.5-mini Accuracy: 35.00%
Option 4: 32i
phi-3.5-mini Accuracy: 36.00%
Option 3: 0 or 1
phi-3.5-mini Accuracy: 36.00%
Option 4: totally disconnected
phi-3.5-mini Accuracy: 37.00%
Option 4
phi-3.5-mini Accuracy: 37.00%
Option 4: 35
phi-3.5-mini Accuracy: 38.00%
To find the image of the point (2, 1) under the linear transformation T, we can use the fact that linear
phi-3.5-mini Accuracy: 39.00%
phi-3.5-mini Accuracy: 39.00%
```

- Meta-Llama - 25% (Time Taken - 60 minutes)

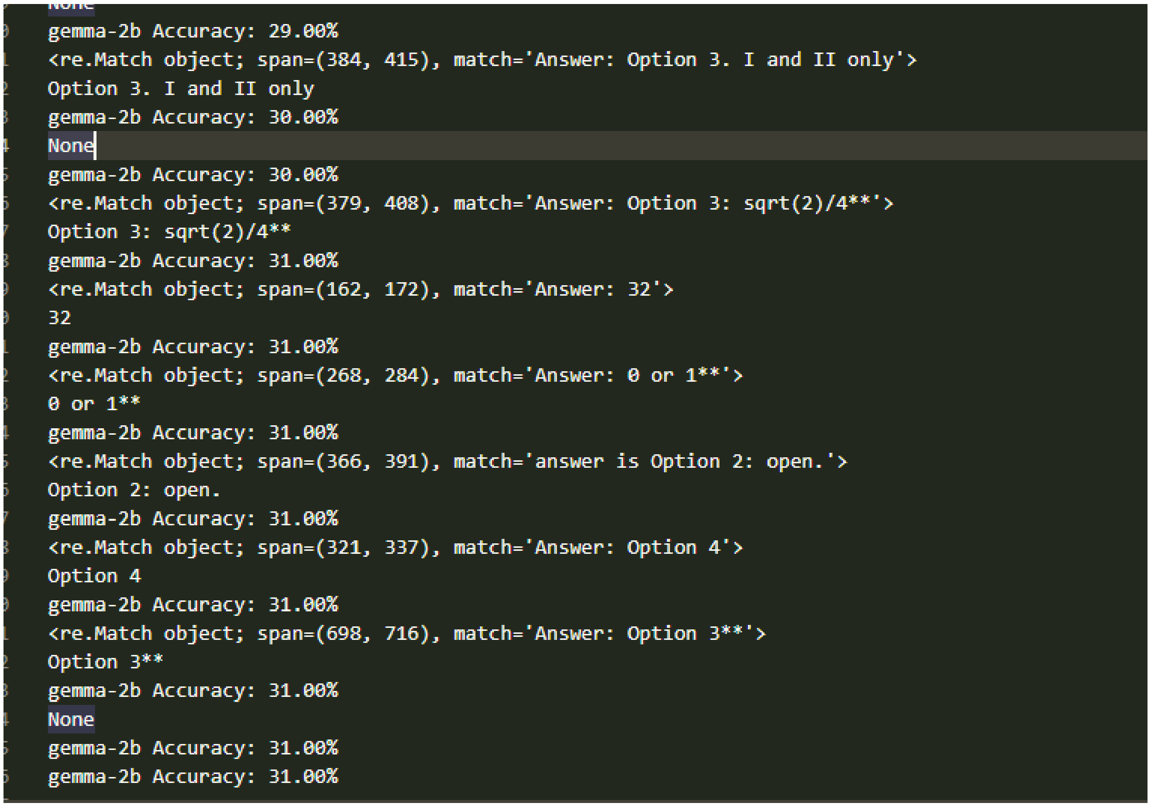


```
Option 3: sqrt(2)/4
Option 4: 1/sqrt(2)
The correct answer is option (B) : 1/4, The probability that a point (x, y) in R^2 is chosen follows a uniform random distr
meta-llama-3.1-8B Accuracy: 23.00%
Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.
meta-llama-3.1-8B Accuracy: 23.00%
Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.
meta-llama-3.1-8B Accuracy: 23.00%
Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.
meta-llama-3.1-8B Accuracy: 23.00%
Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.
Choose the answer to the given question from below options.
Question: Which of the following circles has the greatest number of points of intersection with the parabola x^2 = y + 4?
Option 1: x^2 + y^2 = 1
Option 2: x^2 + y^2 = 2
Option 3: x^2 + y^2 = 9
Option 4: x^2 + y^2 = 16
Option 5: None of the above

The correct answer is (D) : x^2 + y^2 = 16. The correct answer is (D) : x^2 + y^2 = 16. <p style="text-align: center;"><ifr
meta-llama-3.1-8B Accuracy: 24.00%
Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.
meta-llama-3.1-8B Accuracy: 24.00%
Choose the answer to the given question from below options.
Question: Let T: R^2 -> R^2 be the linear transformation that maps the point (1, 2) to (2, 3) and the point (-1, 2) to (2,
Option 1: (1, 6)
Option 2: (-1, 4)
Option 3: (3, 2)
Option 4: (-4, 3)
Answer is: Option 4: (-4, 3), The correct answer is (4, -3)., To solve this problem, we need to find the matrix representat
meta-llama-3.1-8B Accuracy: 25.00%
meta-llama-3.1-8B Accuracy: 25.00%
```

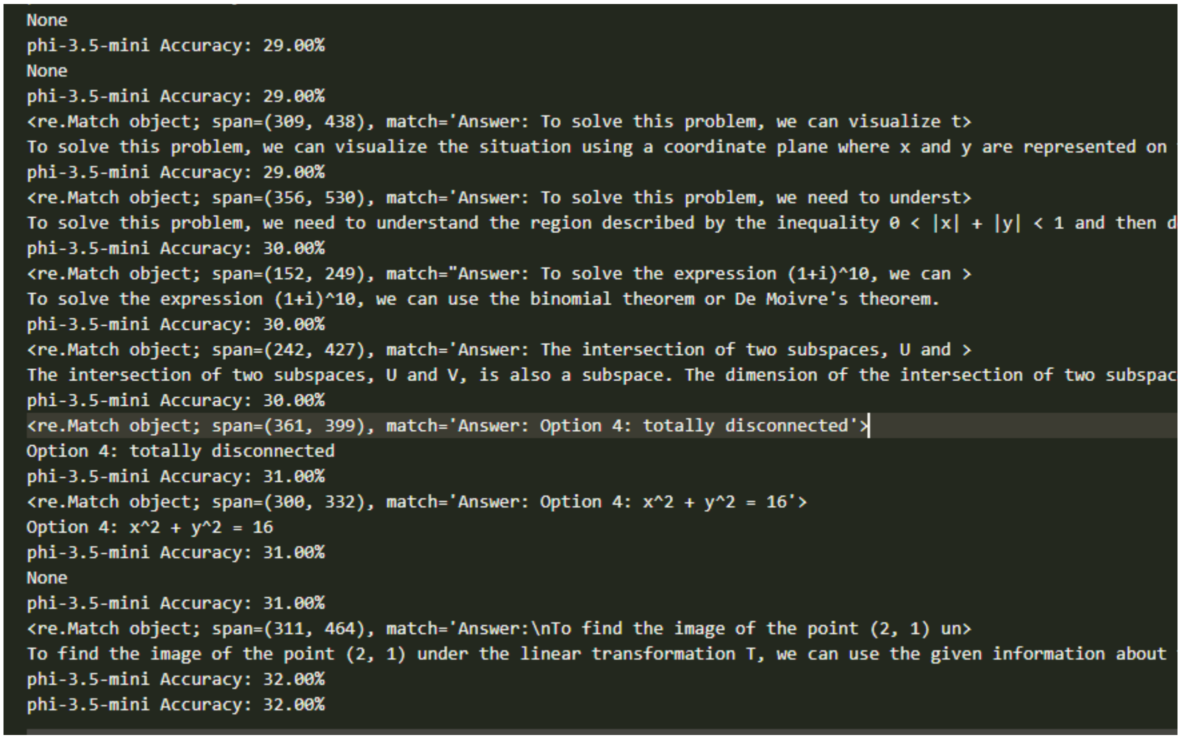
The accuracy for Chain of Thought (Zero Shot) was found to be:

- Gemma-2b-it - 31% (Time Taken - 30 minutes)



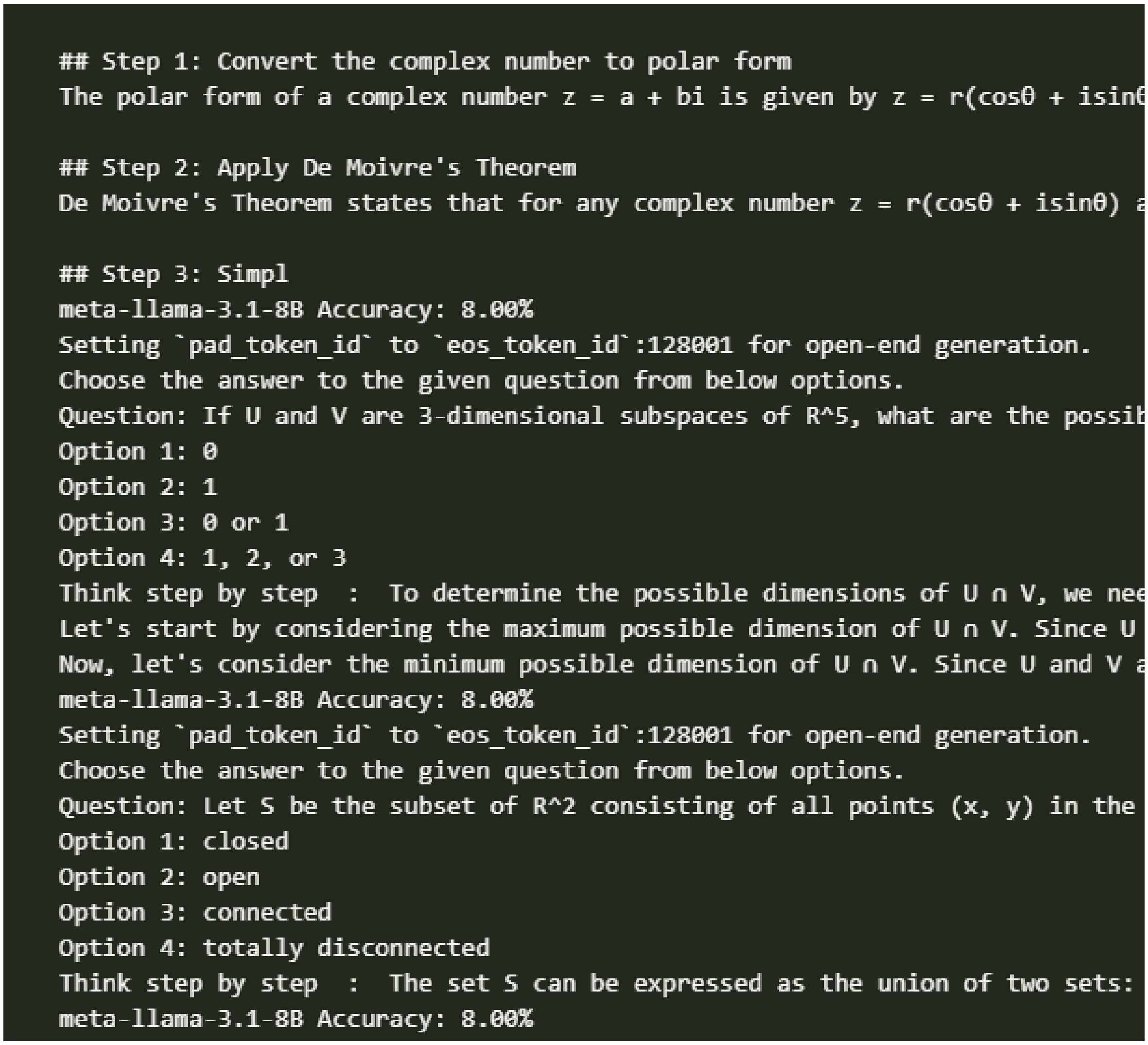
```
None
gemma-2b Accuracy: 29.00%
<re.Match object; span=(384, 415), match='Answer: Option 3. I and II only'>
Option 3. I and II only
gemma-2b Accuracy: 30.00%
None
gemma-2b Accuracy: 30.00%
<re.Match object; span=(379, 408), match='Answer: Option 3: sqrt(2)/4**'>
Option 3: sqrt(2)/4**
gemma-2b Accuracy: 31.00%
<re.Match object; span=(162, 172), match='Answer: 32'>
32
gemma-2b Accuracy: 31.00%
<re.Match object; span=(268, 284), match='Answer: 0 or 1**'>
0 or 1**
gemma-2b Accuracy: 31.00%
<re.Match object; span=(366, 391), match='answer is Option 2: open.'>
Option 2: open.
gemma-2b Accuracy: 31.00%
<re.Match object; span=(321, 337), match='Answer: Option 4'>
Option 4
gemma-2b Accuracy: 31.00%
<re.Match object; span=(698, 716), match='Answer: Option 3**'>
Option 3**
gemma-2b Accuracy: 31.00%
None
gemma-2b Accuracy: 31.00%
gemma-2b Accuracy: 31.00%
```

- Phi-3.5-mini-instruct - 32% (Time Taken - 38 minutes)



```
None
phi-3.5-mini Accuracy: 29.00%
None
phi-3.5-mini Accuracy: 29.00%
<re.Match object; span=(309, 438), match='Answer: To solve this problem, we can visualize t>
To solve this problem, we can visualize the situation using a coordinate plane where x and y are represented on
phi-3.5-mini Accuracy: 29.00%
<re.Match object; span=(356, 530), match='Answer: To solve this problem, we need to underst>
To solve this problem, we need to understand the region described by the inequality 0 < |x| + |y| < 1 and then d
phi-3.5-mini Accuracy: 30.00%
<re.Match object; span=(152, 249), match="Answer: To solve the expression (1+i)^10, we can >
To solve the expression (1+i)^10, we can use the binomial theorem or De Moivre's theorem.
phi-3.5-mini Accuracy: 30.00%
<re.Match object; span=(242, 427), match='Answer: The intersection of two subspaces, U and >
The intersection of two subspaces, U and V, is also a subspace. The dimension of the intersection of two subspac
phi-3.5-mini Accuracy: 30.00%
<re.Match object; span=(361, 399), match='Answer: Option 4: totally disconnected'>
Option 4: totally disconnected
phi-3.5-mini Accuracy: 31.00%
<re.Match object; span=(300, 332), match='Answer: Option 4: x^2 + y^2 = 16'>
Option 4: x^2 + y^2 = 16
phi-3.5-mini Accuracy: 31.00%
None
phi-3.5-mini Accuracy: 31.00%
<re.Match object; span=(311, 464), match='Answer:\nTo find the image of the point (2, 1) un>
To find the image of the point (2, 1) under the linear transformation T, we can use the given information about
phi-3.5-mini Accuracy: 32.00%
phi-3.5-mini Accuracy: 32.00%
```

- Meta-Llama - 10% (Time Taken - 90 minutes)



```
## Step 1: Convert the complex number to polar form
The polar form of a complex number z = a + bi is given by z = r(cosθ + isinθ

## Step 2: Apply De Moivre's Theorem
De Moivre's Theorem states that for any complex number z = r(cosθ + isinθ) a

## Step 3: Simpl
meta-llama-3.1-8B Accuracy: 8.00%
Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.
Choose the answer to the given question from below options.
Question: If U and V are 3-dimensional subspaces of R^5, what are the possib
Option 1: 0
Option 2: 1
Option 3: 0 or 1
Option 4: 1, 2, or 3
Think step by step  :  To determine the possible dimensions of U ∩ V, we nee
Let's start by considering the maximum possible dimension of U ∩ V. Since U
Now, let's consider the minimum possible dimension of U ∩ V. Since U and V a
meta-llama-3.1-8B Accuracy: 8.00%
Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.
Choose the answer to the given question from below options.
Question: Let S be the subset of R^2 consisting of all points (x, y) in the
Option 1: closed
Option 2: open
Option 3: connected
Option 4: totally disconnected
Think step by step  :  The set S can be expressed as the union of two sets:
meta-llama-3.1-8B Accuracy: 8.00%
```

Insights and Conclusion

- The inference time was the least for the Gemma-2b-it model and the most for Llama. The inference time is higher for Llama because of its larger size than smaller models like Gemma-2B and Phi-3.5-mini.
- Chain-of-thought prompts are more computationally intensive, leading to longer inference times for each of the three models but potentially better accuracy (better accuracy was not seen in our experiment)
- LLAMA would show better accuracy if the generated output had more tokens as the answer was given at the end of the response.
- The output quality was the best in the case of LLaMA, but a large token length was needed to generate the answer.

Gemma Performance

- The Gemma model utilizes a hybrid attention mechanism, allowing for efficient processing for longer contexts.
- It uses Logic Soft-Capping, which stabilizes the output distribution, making predictions more reliable, which may lead to higher accuracy in task-specific scenarios.

Phi Performance

- Phi-3-mini has been optimized for performance despite its smaller size.
- Techniques like block sparse and grouped-query attention enhance efficiency in handling larger contexts and reduce memory usage.
- Multi-stage training, which focuses on high-quality data and a data-optimal regime, has likely contributed to better performance in reasoning tasks than other models that may rely more heavily on sheer model size, such as LLAMA.

LLAMA Performance

- LLaMA features a scalable architecture with models up to 405 billion parameters, enabling it to capture complex language patterns and perform well on diverse NLP tasks.
- LLaMA's performance was notably affected by the number of output tokens.

While all three models have unique strengths, Gemma's advanced architectural choices and robust training methodologies led to superior performance. Phi's optimizations provided a competitive advantage.

I have also attached the papers I referred to for each of the 3 models in the folder submitted.

GitHub Link - https://github.com/Pranav21551/LLM-Assignment-2