Natual Language Processsing
Spring 2025

**Assignment 2**

**Name** Ke Li **Student No.** 2024210837

---

# 1 Prompt Engineering

## 1.1 Designing a Unique and Challenging Prompt

### 1.1.1 Model Selection

Because our purpose is to generate a task that only one model can perform correctly, while the other model cannot, we need to find differences between the two models. A simple aspect is the model size, the larger model remembers more knowledge. However, this may not be the best choice, as we do not know which data is used to train one model and not the other. So I choose to distinguish the two models by their reasoning ability. Since **Deepseek-R1**(which is just Deepseek-V3 with deep thinking) has a better reasoning ability than **GPT-4o**, which is the true in its technique report (DeepSeek-AI et al. 2025), I decide to use these two models to generate the task.

### 1.1.2 Task Generation

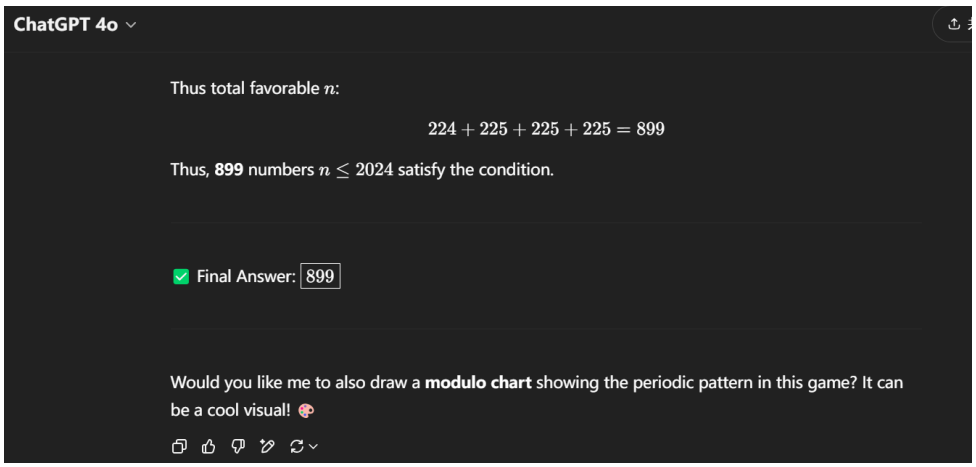
The question should have a single objective and easily verifiable answer, so mathematical problems are a good choice. Deepseek-R1 has reported their results in different math problems, so I choose the dataset which has the biggest difference pass@1 score between the two models, AIME 2024. The question I choose is:

> **Q:** Alice and Bob play the following game. A stack of $n$ tokens lies before them. The players take turns with Alice going first. On each turn, the player removes either 1 token or 4 tokens from the stack. Whoever removes the last token wins. Find the number of positive integers $n$ less than or equal to 2024 for which there exists a strategy for Bob that guarantees that Bob will win the game regardless of Alice's play.

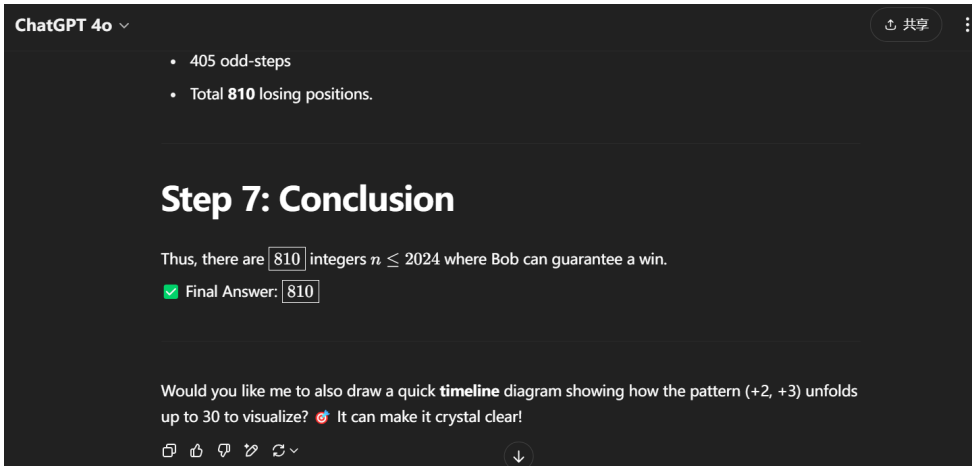
The question ID is **2024-I-3**, and the answer is **809**. During my test, I find that GPT-4o can generate the answer quickly, but the results are not correct. In three tests, the answers are **899**, **810**, and **810**. They are all wrong. However, Deepseek-R1 can generate the answer correctly in all three tests. The results are **809**, **809**, and **809**. Despite the correctness, the time cost of Deepseek-R1 is much higher than GPT-4o. The time cost of Deepseek-R1 is **198s**, **179s**, and **138s**. I find an interesting phenomenon that although Deepseek-R1 has gotten the correct answer, it will "wait" and try another method to verify the answer, which causes the time cost to be much higher than GPT-4o.

### 1.1.3 Results

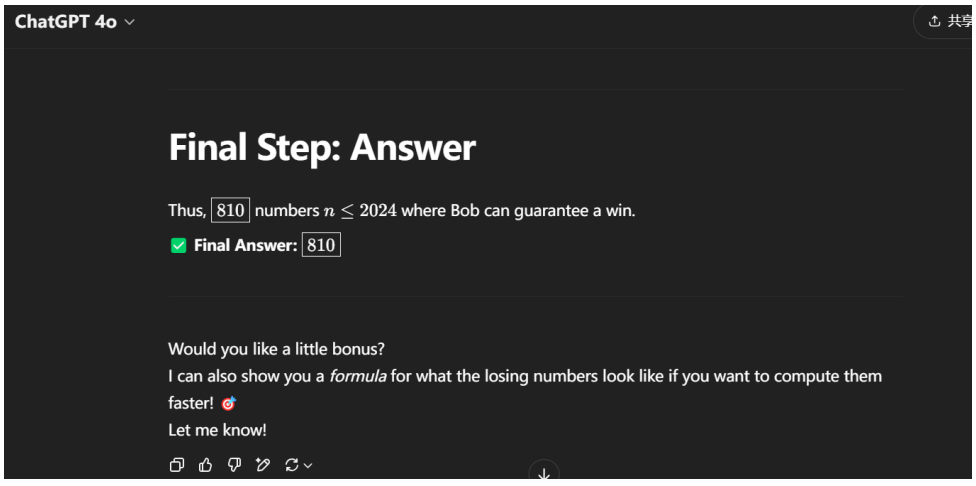
Screen shots of the results are shown in Figure 1 and Figure 2.




(a) GPT-4o result 1




(b) GPT-4o result 2



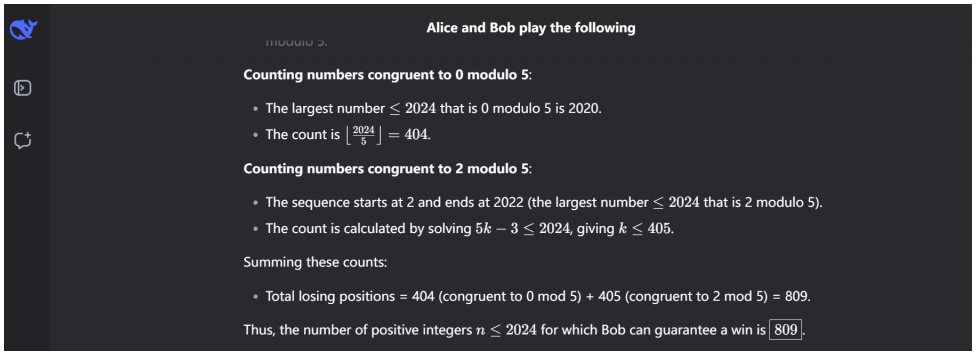

(c) GPT-4o result 3

Figure 1GPT-4o results

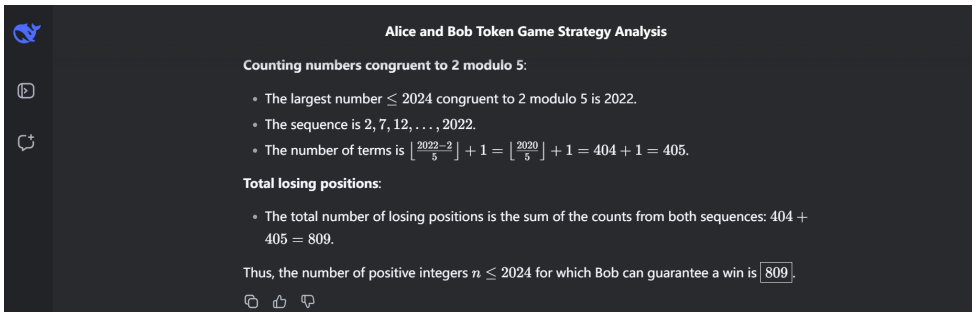
## 1.2 In-Context Learning

### 1.2.1 Questions Selection

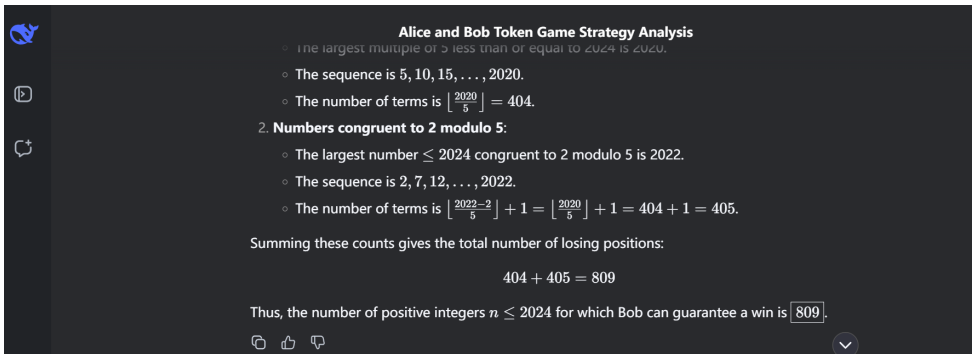
I randomly select 3 questions from the GSM-8K, which is a math problem dataset. The questions are as follows:(you can also find the questions in file **sampled_examples.json**)




(a) Deepseek-R1 result 1




(b) Deepseek-R1 result 2




(c) Deepseek-R1 result 3

Figure 2Deepseek-R1 results

**Q1:** Nancy is crafting clay pots to sell. She creates 12 clay pots on Monday, twice as many on Tuesday, a few more on Wednesday, then ends the week with 50 clay pots. How many did she create on Wednesday?

**Q2:** For the first hour of work, Manolo can make face-masks at the rate of one every four minutes. Thereafter, he can make face-masks at the rate of one every six minutes. How many face-masks does Manola make in a four-hour shift?

**Q3:** The Tigers played 56 home games this year. They had 12 losses and half as many ties. How many games did they win?

The answers are **14**, **45**, and **38** respectively. I test the three questions with model **Deepseek-V3** and **Deepseek-R1**.

### 1.2.2 Prompt Design

We need to design one prompt to solve the three questions, and the output should in structured format. So below is the prompt without examples:

**System prompt:** You are a helpful assistant. Please solve the following three reasoning-based questions from the GSM-8K dataset. Provide a detailed step-by-step solution for each question to clearly show your logical process, and then state the final answer. Ensure the output is in a structured format, such as JSON, and includes the following fields:

- Question ID
- Reasoning Process (a list)
- Final Answer
- Difficulty Classification

**User prompt:** Here are three reasoning-based questions from the GSM-8K dataset:

1. Question: Nancy is crafting clay pots to sell. She creates 12 clay pots on Monday, twice as many on Tuesday, a few more on Wednesday, then ends the week with 50 clay pots. How many did she create on Wednesday?
2. Question: For the first hour of work, Manolo can make face-masks at the rate of one every four minutes. Thereafter, he can make face-masks at the rate of one every six minutes. How many face-masks does Manola make in a four-hour shift?
3. Question: The Tigers played 56 home games this year. They had 12 losses and half as many ties. How many games did they win?

Please solve these questions step by step and provide the final answers.

For few-shot learning, I slightly modify the system prompt which will not affect the results. The main change is to add two examples to the user prompt. The user prompt is as follows:

**User prompt:** Here are demonstration examples of solving GSM-8K problems, followed by the questions you need to answer. Study the structure and reasoning style carefully:

### Demonstration Example 1

**Question ID**: DEMO1

**Question**: A bookstore sells 35 novels in the morning and twice as many in the afternoon. If they had 150 novels in stock at the start, how many remain unsold?

**Reasoning Process**:

- Calculate afternoon sales: 35 novels * 2 = 70 novels
- Sum total sales: 35 novels + 70 novels = 105 novels
- Subtract total sales from stock: 150 novels - 105 novels = 45 novels

**Final Answer**: 45

**Difficulty Classification**: Medium

### Demonstration Example 2

**Question ID**: DEMO2

**Question**: A baker uses 2 cups of flour for each loaf of bread. If she has a 50-cup bag of flour and bakes 12 loaves, how many cups of flour remain?

**Reasoning Process**:

- Calculate flour used: 12 loaves * 2 cups/loaf = 24 cups
- Subtract used flour from total: 50 cups - 24 cups = 26 cups

**Final Answer**: 26

**Difficulty Classification**: Easy

### Now Solve These Questions

1. **Question ID**: Q1 **Question**: Nancy is crafting clay pots to sell. She creates 12 clay pots on Monday, twice as many on Tuesday, a few more on Wednesday, then ends the week with 50 clay pots. How many did she create on Wednesday?
2. **Question ID**: Q2 **Question**: For the first hour of work, Manolo can make face-masks at the rate of one every four minutes. Thereafter, he can make face-masks at the rate of one every six minutes. How many face-masks does Manola make in a four-hour shift?
3. **Question ID**: Q3 **Question**: The Tigers played 56 home games this year. They had 12 losses and half as many ties. How many games did they win?

Of course, you can also find the prompt in the file **code/src/task_icl.py**.

### 1.2.3 Results

The results can be found in the following files:

- **response_chat_without_icl.json**
- **response_chat_with_icl.json**
- **response_reasoner_without_icl.json**
- **response_reasoner_with_icl.json**

I am not going to show the results here because they are too long. However, I will explain the results and analyze the performance of the two models.

The answers of the two models are all correct, and the reasoning process is also correct. This may be because the questions are simple and the models have been trained on these types of questions. However, we can see that for **Deepseek-V3** without in-context learning, the answer is not just a number, but a sentence. After we give it a few examples, the model can generate the answer in a structured format. Another interesting phenomenon is that the reasoning process will be similar to the examples after we give it a few examples.

So we can conclude that in-context learning can help the model to generate the answer in a structured format we need, and think in a similar way as the examples. However, the drawback is also obvious. The model will be "limited" by the examples, and we must think carefully about the examples we give.

# References

DeepSeek-AI et al. (2025). *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*. arXiv: 2501.12948 [cs.CL]. URL: https://arxiv.org/abs/2501.12948.