Natual Language Processsing
Spring 2025

Assignment 2

Name Ke Li **Student No.** 2024210837

---

1 Prompt Engirneering

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

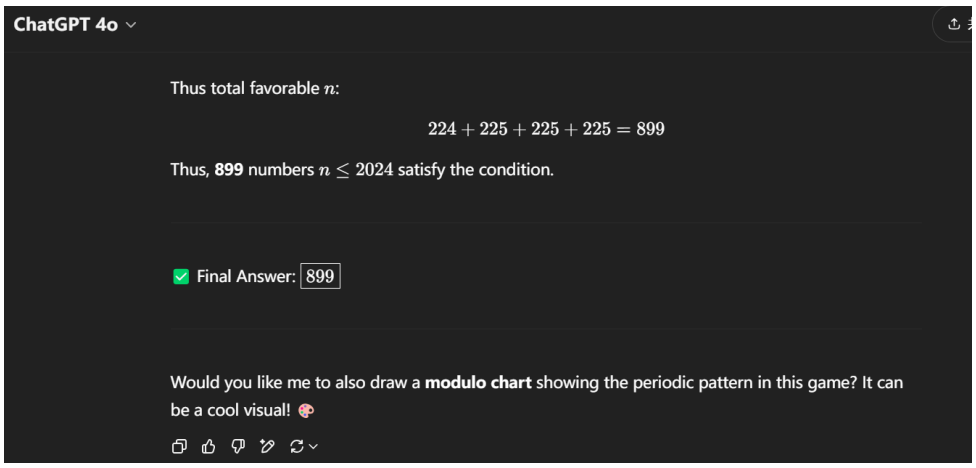
The question should have a single objective and easily verifiable answer, so mathematical problems are a good choice. Deepseek-R1 has reported their results in different math problems, so I choose the dataset which has the biggest difference pass@1 score between the two models, AIME 2024. The question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

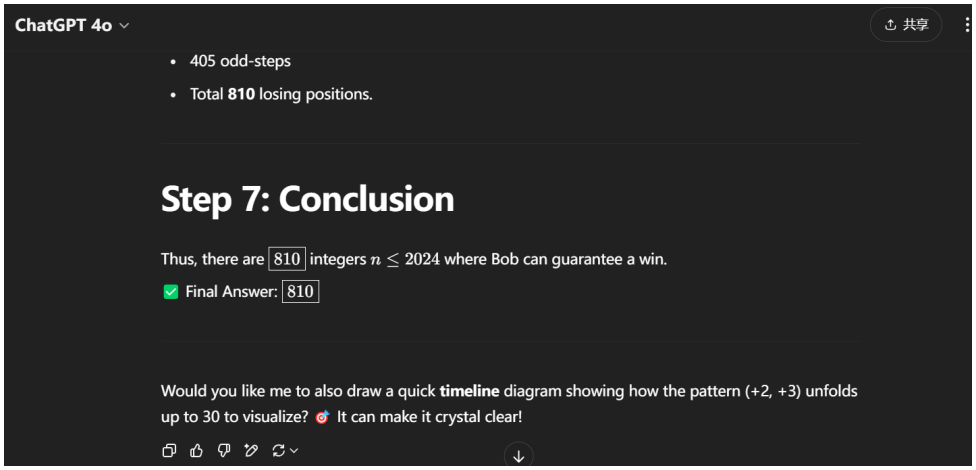
The question ID is **2024-I-3**, and the answer is **809**. During my test, I find that GPT-4o can generate the answer quickly, but the results are not correct. In three tests, the answers are **899**, **810**, and **810**. They are all wrong. However, Deepseek-R1 can generate the answer correctly in 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

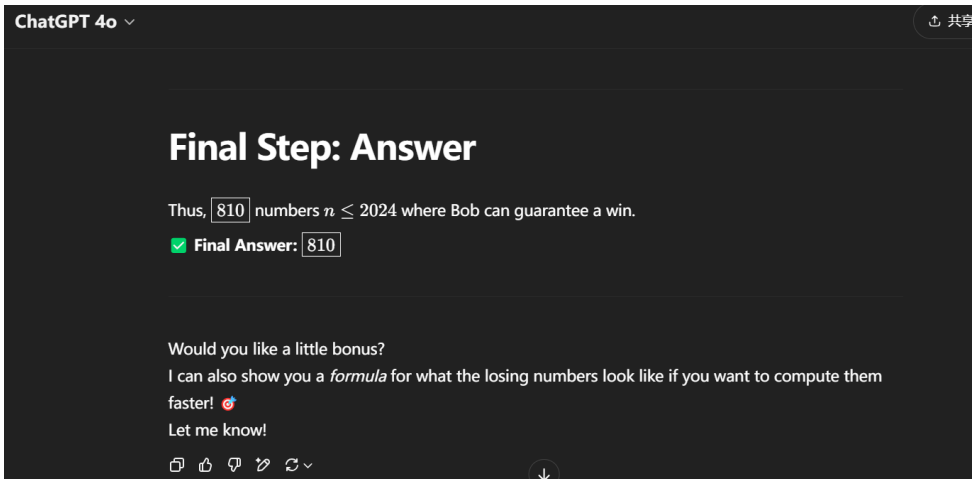
Screen shots of the results are shown in Figure 1 and Figure 2.



(a) GPT-4o result 1



(b) GPT-4o result 2

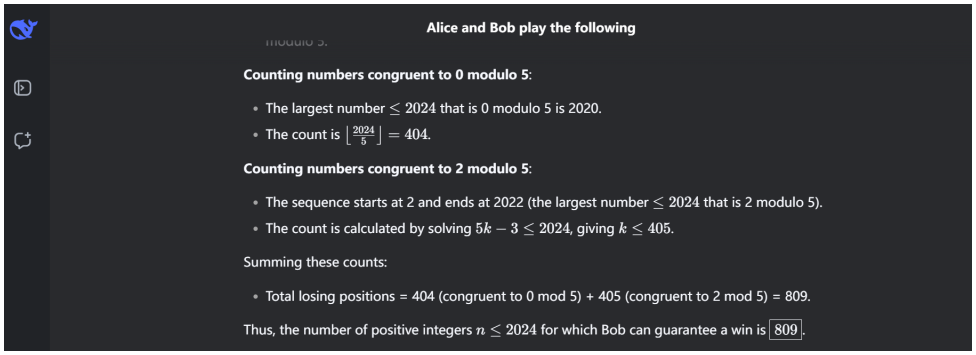


(c) GPT-4o result 3

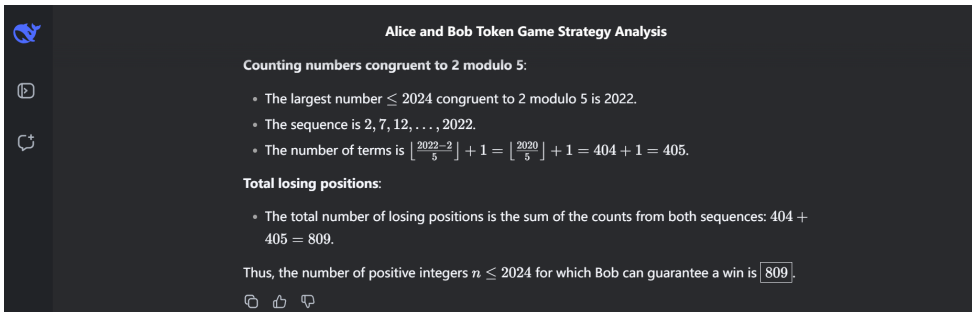
Figure 1GPT-4o results

References

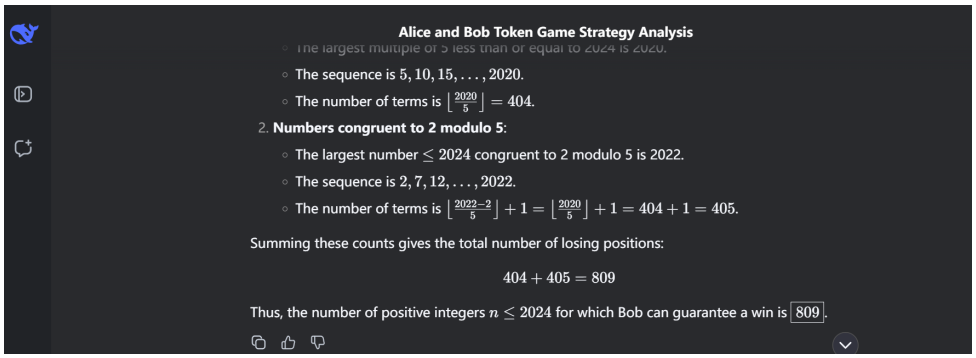
DeepSeek-AI et al. (2025). *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*. arXiv: 2501.12948 [cs.CL]. URL: https://arxiv.org/abs/2501.12948.



(a) Deepseek-R1 result 1



(b) Deepseek-R1 result 2



(c) Deepseek-R1 result 3

Figure 2Deepseek-R1 results