**Homework 5: Let’s Play LLM**

**Part 1: Designing a Task for LLM and Explore the Capability of LLM (70%)**

**Task Description (10%)**

* In this scenario, a hero is fighting with demon army. The hero can do multiple actions. I will assign some events for hero to meet with, and ChatGPT should determine the hero how to deal with these events with optimized cost.
* The input prompt consists of three parts, the first part specifies the hero’s base information, the second part tells the actions hero can take and each action’s cost, and the third part describes the events hero will encounter for LLM to solve.
* For event 1, 2, 4, 5, ChatGPT should answer them correctly. For event 3, ChatGPT should recognize this event can’t be completed and provide reasons.

**Motivation (10%)**

* I love to play video games or online games, and I am curious about whether LLM can play simple game just like human with full description.
* I want to explore ChatGPT's comprehension of scenes, decision-making abilities, and some simple math computation abilities.

**Describe your Initial Attempt (15%)**

* Input

一張含有 文字, 字型, 螢幕擷取畫面, 文件 的圖片

自動產生的描述

* Output

一張含有 文字, 螢幕擷取畫面, 字型, 代數 的圖片

自動產生的描述

一張含有 文字, 螢幕擷取畫面, 字型, 文件 的圖片

自動產生的描述

一張含有 文字, 螢幕擷取畫面, 文件, 字型 的圖片

自動產生的描述

一張含有 文字, 螢幕擷取畫面, 字型, 代數 的圖片

自動產生的描述

* Analysis: From the output of ChatGPT, you can notice that it only gives correct answer for event 4. Event 1: Hero can just attack once and the cost is 5. Event 2: Hero doesn’t need to take the magical potion and the cost is 20. Event 3: Hero will escape and this event cannot be completed because the total cost for beating demon is above 200. Event 5: The cost should be 85. By above, I find that ChatGPT didn’t fully realize the combat alternative, the time to take the magical potion, and has a limited understanding of the scenario. I will use more detailed descriptions to make ChatGPT answers more accurately.

**Describe Improvements (35%)**

* Input: I provide example about combat system for ChatGPT to further understand how to compute the loss of HP. I tell ChatGPT when the time is to take the magical potion. I provide more descriptions about combating with demons. I tell ChatGPT if the enemy’s HP is 0 or below, then the enemy is defeated, and no more attacks are needed.

一張含有 文字, 紙張, 字型, 文件 的圖片

自動產生的描述

* Output

一張含有 文字, 螢幕擷取畫面, 字型, 代數 的圖片

自動產生的描述

一張含有 文字, 字型, 螢幕擷取畫面, 代數 的圖片

自動產生的描述

一張含有 文字, 螢幕擷取畫面, 字型, 白色 的圖片

自動產生的描述

一張含有 文字, 字型, 收據, 螢幕擷取畫面 的圖片

自動產生的描述

一張含有 文字, 收據, 字型, 螢幕擷取畫面 的圖片

自動產生的描述

* Analysis: From the output of ChatGPT, you can notice that it gives correct answer for event 1, 2, 4 based on the more complete description. But for event 3, 5, it still cannot compute the correct answer. I think it has two possible reasons. First, my descriptions are still not enough for these events although I think it is enough. Second, because event 3, 5 have more complex computation than event 1, 2, 4, ChatGPT may not be good at some little complex computation.

**Part 2: Problem Answering (30%)**

**Question 1**

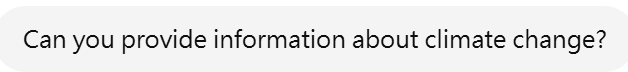
1. Besides providing a more detailed description of the question in the prompt, what other methods can be employed to improve the accuracy of the responses?

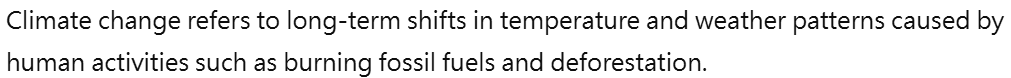
We can provide clarification prompts or ask follow-up questions to refine our query, helping ChatGPT produce more precise and relevant responses. And crafting specific and clear questions ensures ChatGPT understands ours intent accurately, leading to more focused and accurate responses.

1. Please provide an example demonstrating the impact of implementing the aforementioned methods on a prompt.

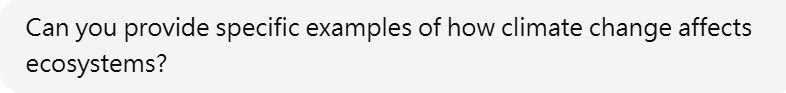
Suppose we want to know about climate change.

Without Additional Methods:





With Additional Methods:



一張含有 文字, 字型, 白色, 文件 的圖片

自動產生的描述

In this example, my clarification prompt and specificity in the query lead to a more detailed and relevant response from ChatGPT, demonstrating the effectiveness of these methods in improving response accuracy.

**Question 2**

* Question: Handling Bias and Fairness in LLMs
* Motivation: LLMs have been found to perpetuate biases present in the training data, leading to unfair or discriminatory outcomes in their outputs. Addressing bias and ensuring fairness in LLMs is crucial for promoting equity and inclusivity in AI applications.
* Challenges:
  + Identifying and Understanding Bias: Recognizing biases present in LLMs and understanding their underlying causes is complex due to the vast amount of data they are trained on and the subtle ways biases can manifest in language.
  + Mitigating Bias without Compromising Performance: Removing bias from LLMs without sacrificing their overall performance and natural language generation capabilities is challenging, as bias may be intertwined with linguistic patterns and contextual understanding.
  + Ensuring Fairness Across Demographic Groups: LLMs should generate outputs that are fair and equitable across different demographic groups, but achieving this goal requires careful consideration of various social, cultural, and historical factors.
* Solution:
  + Bias Detection and Mitigation Techniques: Employing techniques such as debiasing algorithms, adversarial training, and counterfactual data augmentation to identify and mitigate biases in LLMs' training data and outputs.
  + Fairness-Aware Training: Incorporating fairness constraints into the training process of LLMs to encourage fair decision-making and reduce disparities in model outputs across demographic groups.
  + Evaluation Metrics for Fairness: Developing metrics and evaluation frameworks to assess the fairness and equity of LLMs' outputs, enabling researchers and practitioners to quantify and compare the performance of different fairness interventions.
* Reference: [2405.11290 (arxiv.org)](https://arxiv.org/pdf/2405.11290)