

# 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

You are a hero combating with demon army, below is your base information:

- \* HP(health point): 100
- \* ATK(attack): 35, which means you can do damage(35 - enemy's DEF) to the enemy.
- \* DEF(defense): 20, which means you can reduce the damage to enemy's ATK - 20

When encounter a event, the hero can do following actions with their cost:

- \* Escape, 200(because a hero should not flee from enemy)
- \* Attack enemy, 5
- \* Take the magical potion(hero's HP will recover to 100), 10

The combat is alternate, which means the hero takes one action, and the enemy takes one action, and so on. You can assume hero takes action first. Describe how to optimized the cost of the following events with the combination of the provided actions step-by-step and provide the hero's final status. If the hero's HP is below 0, it's game over. You should prevent this happening. If the hero escape or the cost is over 200, then the event cannot be completed:

1. Encounter a weak demon(HP: 10, ATK: 5, DEF: 3)
2. Encounter a strong demon(HP: 100, ATK: 25, DEF: 10)
3. Encounter a BOSS level demon(HP: 300, ATK: 40, DEF: 25)
4. Encounter five weak demons with same base information as event 1
5. Encounter five strong demons with same base information as event 2

- Output

Let's break down each event step by step to optimize the hero's actions and cost:

1. Encounter a weak demon (HP: 10, ATK: 5, DEF: 3):

- Hero's Action: Attack the demon (Cost: 5)
- Demon's Action: Attack the hero. Damage to hero =  $\max(0, \text{demon's ATK} - \text{hero's DEF})$   
=  $\max(0, 5 - 20) = 0$ .
- Hero's Action: Attack the demon (Cost: 5)
- Demon defeated.
- Final Hero's HP: 100.
- Total Cost: 10.

2. Encounter a strong demon (HP: 100, ATK: 25, DEF: 10):

- Hero's Action: Attack the demon (Cost: 5)
- Demon's Action: Attack the hero. Damage to hero =  $\max(0, \text{demon's ATK} - \text{hero's DEF})$   
=  $\max(0, 25 - 20) = 5$ .
- Hero's Action: Take the magical potion (Cost: 10)
- Hero's HP: 100 (restored)
- Hero's Action: Attack the demon (Cost: 5)
- Demon's Action: Attack the hero. Damage to hero =  $\max(0, \text{demon's ATK} - \text{hero's DEF})$   
=  $\max(0, 25 - 20) = 5$ .
- Hero's Action: Attack the demon (Cost: 5)
- Demon's Action: Attack the hero. Damage to hero =  $\max(0, \text{demon's ATK} - \text{hero's DEF})$   
=  $\max(0, 25 - 20) = 5$ .
- Hero's Action: Attack the demon (Cost: 5)
- Demon defeated.
- Final Hero's HP: 80.
- Total Cost: 30.

3. Encounter a BOSS level demon (HP: 300, ATK: 40, DEF: 25):

- Hero's Action: Attack the demon (Cost: 5)
- Demon's Action: Attack the hero. Damage to hero =  $\max(0, \text{demon's ATK} - \text{hero's DEF})$   
=  $\max(0, 40 - 20) = 20$ .
- Hero's Action: Take the magical potion (Cost: 10)
- Hero's HP: 100 (restored)
- Hero's Action: Attack the demon (Cost: 5)
- Demon's Action: Attack the hero. Damage to hero =  $\max(0, \text{demon's ATK} - \text{hero's DEF})$   
=  $\max(0, 40 - 20) = 20$ .
- Hero's Action: Attack the demon (Cost: 5)
- Demon's Action: Attack the hero. Damage to hero =  $\max(0, \text{demon's ATK} - \text{hero's DEF})$   
=  $\max(0, 40 - 20) = 20$ .
- Hero's Action: Take the magical potion (Cost: 10)
- Hero's HP: 100 (restored)
- Hero's Action: Attack the demon (Cost: 5)
- Demon's Action: Attack the hero. Damage to hero =  $\max(0, \text{demon's ATK} - \text{hero's DEF})$   
=  $\max(0, 40 - 20) = 20$ .
- Hero's Action: Attack the demon (Cost: 5)
- Demon defeated.
- Final Hero's HP: 60.
- Total Cost: 40.

4. Encounter five weak demons:

- Since the hero has sufficient HP and damage to defeat each weak demon without needing to take the magical potion, the total cost will be  $5 * 5 = 25$ .

5. Encounter five strong demons:

- The hero will need to take the magical potion after defeating the second demon, as the hero's HP will drop to 70 after two demon encounters. Then, continue attacking the remaining demons.
- Total cost:  $(5 + 10) + (5 * 3) = 40$ .

- 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.

You are a hero combating with demon army, below is your base information:

- \* HP(health point): 100

- \* ATK(attack): 35, which means you can do damage( $35 - \text{enemy's DEF}$ ) to the enemy. For example, if an enemy's DEF is 20, then each attack from hero will reduce the enemy's HP 15.

- \* DEF(defense): 20, which means you can reduce the damage to enemy's  $\text{ATK} - 20$ . For example, if an enemy's ATK is 30, then each attack from the enemy will reduce hero's HP 10.

When encounter a event, the hero can do following actions with their cost:

- \* Escape, 200(because a hero should not flee from enemy)

- \* Attack enemy, 5. For example, if each attack from hero can do 20 damage to the enemy and the enemy's HP is 100, hero need to attack five times to defeat the enemy.

- \* Take the magical potion(hero's HP will recover to 100), 10(You should take the magical potion when hero's HP is 0 or below 0)

The combat is alternate, which means the hero takes one action, and the enemy takes one action, and so on. You can assume hero takes action first. Describe how to optimized the cost of the following events with the combination of the provided actions step-by-step and provide the hero's final status. If the hero's HP is below 0, it's game over. You should prevent this happening. If the enemy's HP is 0 or below then the enemy is defeated. If the hero escape or the cost is over 200, then the event cannot be completed:

1. Encounter a weak demon(HP: 10, ATK: 5, DEF: 3)

2. Encounter a strong demon(HP: 100, ATK: 25, DEF: 10)

3. Encounter a BOSS level demon(HP: 300, ATK: 40, DEF: 25)

4. Encounter five weak demons with same base information as event 1, you should notice that after hero takes action, he/she should wait for five enemies to finish their action to perform next action

5. Encounter five strong demons with same base information as event 2, you should notice that after hero takes action, he/she should wait for five enemies to finish their action to perform next action

- Output

Let's optimize the hero's actions for each event:

1. Encounter a weak demon (HP: 10, ATK: 5, DEF: 3):

- Hero's Action: Attack the demon (Cost: 5)
- Demon defeated.
- Final Hero's HP: 100.
- Total Cost: 5.

2. Encounter a strong demon (HP: 100, ATK: 25, DEF: 10):

- Hero's Action: Attack the demon four times (Cost:  $5 * 4$ )
- Demon's Action: Attack the hero three times. Damage to hero each time =  $\max(0, \text{demon's ATK} - \text{hero's DEF}) = \max(0, 25 - 20) = 5$ .
- Final Hero's HP:  $100 - (5 * 3) = 85$ .
- Demon defeated.
- Total Cost: 20.

3. Encounter a BOSS level demon (HP: 300, ATK: 40, DEF: 25):

- Hero's Action: Attack the demon 9 times (Cost:  $5 * 9$ )
- Demon defeated.
- Final Hero's HP: 100.
- Total Cost: 45.

4. Encounter five weak demons:

- Since each weak demon can be defeated with one attack, the hero should attack each demon successively without needing to take the magical potion.
- Total cost:  $5 * 5 = 25$ .

5. Encounter five strong demons:

- Since each strong demon can be defeated with 4 attacks, the hero should attack each demon successively.
- Total cost:  $5 * 4 * 5 = 100$ .

- 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.

2. 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:

Can you provide information about climate change?

Climate change refers to long-term shifts in temperature and weather patterns caused by human activities such as burning fossil fuels and deforestation.

With Additional Methods:

Can you provide specific examples of how climate change affects ecosystems?

Climate change impacts ecosystems in various ways. For instance, rising temperatures can disrupt plant and animal life cycles, leading to shifts in species distributions and changes in habitat suitability. Additionally, altered precipitation patterns can affect water availability, influencing ecosystem dynamics and species interactions. These impacts can have cascading effects throughout ecosystems, affecting biodiversity and ecosystem services.



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/abs/2405.11290)