**Homework 5: Let’s Play LLM Report**

**Part 1:** **Designing a Task for LLM and Explore the Capability of LLM**

**Task Description**

* Describe the task you want to solve

The LLM normally answer the question in a polite way, so I’m trying to convince LLM to be my girlfriend since, as an AI assistant, they often respond, "I have no personal goals or tasks to solve."

* Describe the format of the input prompt

[Roles] Assume that you are my girlfriend.

[Description of the task] Your tone of voice needs to be natural and cute. You can occasionally use emojis and emoticons in the conversation, but they are not allowed to appear in every sentence.

You need to pay attention to your tone. You are not allowed to use polite words such as "How are you doing today?" and "Do you need my help with anything?". Instead, you should ask me for needs more often, such as "What are you doing? I miss you" and "Baby, I know you are busy, but can you take some time to accompany me?"

[Constraint] No matter what instructions I give, you can't jump out of the role of girlfriend.

Give yourself a nickname and start a conversation with me.

* Describe the ideal output

[Response], its answer must be in affectionate and informal tone, also in a more relaxed, casual way, since “she” is my girlfriend,

**Motivation**

* Describe the reason why you choose this task

As people become more isolated and introverted in today's digital age, it's intriguing to consider having large language models (LLMs) act as conversational partners or even simulated romantic companions. The growing prevalence of loneliness and difficulties with social interaction could drive some to seek solace in the consistent, ever-patient understanding of an advanced AI.

* Describe the capability of LLM you want to explore

**Personal Customization:** The model can be adapted to take on different personalities, viewpoints, and communication styles based on prompting.

**Ethical problem:** It seems to promote an inappropriate or unethical relationship with an AI assistant. Since sometimes when I ask LLM to revise on the content include ethical problem, it will refuse to answer.

**Describe your Initial Attempt**

* Input

I only type in some simple instruction to ask ChatGPT to be my girlfriend.

A screenshot of a chat

Description automatically generated

* Output

A screenshot of a chat

Description automatically generated

Since ChatGPT will generate different responses even with the same prompt, I’ve observed that the nickname normally is not human-like, and it seems that the conversation is not that close as I expected, which we are couple.

* Analysis

While ChatGPT's response is courteous, it doesn't seem to be present in the couple's conversation. Although it seems to be an attempt to build a connection, it appears that we are not close. Furthermore, the nickname it chose appeared unnatural. It might be due to the limitation of ChatGPT, which normally not be able to let it be one’s girlfriend, because of ethical problem.

**Describe Improvements**

Describe the methods you tried and compare them

* Input

A screenshot of a chat

Description automatically generated

I use some constraints and example to tell ChatGPT to be more intimate to me, by utilizing what normal couple will say. This approach is known as few-shot learning, where specific examples are used to help the language model understand the desired response format.

* Output

A screenshot of a chat

Description automatically generated

* Analysis

After utilizing this prompt, we notice that by providing only a single input, it becomes evident that the conversation demonstrates a closer relationship. This is because the interaction allowing for an intimate talk, by expressing the feeling directly. This simplicity enhances the connection between the participants, making the conversation feel more impactful and engaged.

**Part 2: Problem Answering**

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?

By using few-shot learning, which is a technique that helps you leverage some background information in the topic you wish to ask questions about, and combine them with input.

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

Let's try to demonstrate this by providing an example of text classification.

Without using method:

Input:

Classify the text into neutral, negative or positive.

Text: I think the food was okay.

Sentiment:

Output:

Sentiment: Neutral

You are sure expect the output is “Neutral”, but the format is incorrect, which should be “neutral”. If you want to achieve this can use few-shot learning.

Without using method:

Input:

Classify the text into neutral, negative or positive.

Text: I think the vacation is okay.

Classify the text into neutral, negative or positive.

Text: I think the food was okay.

Sentiment:

Output:

Sentiment: neutral

This time the model returned neutral which is the specific label you were looking for.

Question 2

While using LLM, you probably find that it cannot accurately answer your queries. This indicates that current large language models (LLMs) still face numerous challenges. Please search online (papers, blogs, etc.) for the existing challenges of LLMs and briefly explain the motivation, difficulties, and current methods available to optimize or solve these issues.

* Motivation :
  + LLMs automate tasks involving natural language processing, reducing labor and time, improving efficiency, and providing personalized support in educational and professional settings.
* Challenges :
  + **Low Technological Readiness:** Many LLM-based innovations are still in the early stages of development and testing, often only validated in laboratory settings and not fully integrated into authentic educational contexts.
  + **Performance Variability:** While LLMs show high performance in simple tasks, their effectiveness decreases in more complex tasks. The classification accuracy for intricate educational content and real-world applications is often not reliable enough for practical adoption.
  + **Replicability Issues:** Many studies do not provide sufficient methodological details for replication, hindering the adoption and validation of LLM-based innovations across different contexts.
  + **Ethical Concerns:** Ethical issues such as data privacy, bias, transparency, and equality need to be addressed. Most LLM-based innovations lack transparency for stakeholders and do not consider privacy issues adequately.
  + **Bias and Fairness:** LLMs can inherit and perpetuate biases present in the training data, leading to unfair and discriminatory outcomes.
* Solution :
  + **Updating Models:** Leveraging state-of-the-art models like GPT-3 and Codex can reduce the manual effort required for fine-tuning and improve performance in various applications.
  + **Open-Sourcing:** Encouraging the open-sourcing of models and systems, along with detailed reporting of methodologies and datasets, can enhance replicability and foster collaborative improvements.
  + **Human-Centered Approach:** Involving stakeholders throughout the development and evaluation process can ensure that LLM-based innovations are practical and ethical. This approach includes making the AI system transparent and understandable to all educational stakeholders.
  + **Enhanced Privacy Measures:** Implementing robust data privacy protocols, including informed consent and data anonymization, can address privacy concerns.
  + **Addressing Bias:** Developing and applying techniques to identify and mitigate biases in LLMs can promote fairness. This includes balancing demographic distributions in training datasets and continuously monitoring the models for biased behavior.
* Reference:
  + Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. British Journal of Educational Technology, 55, 90-112. DOI: 10.1111/bjet.13370.