Objective

Develop a project that demonstrates the application of RLHF principles, advanced prompt engineering techniques, and ethical considerations in designing an AI solution for a real-world task.

Step 1: Define the Problem

Choose a task where RLHF and prompt engineering can enhance performance and ethical considerations are crucial. Examples include:

- **Healthcare Chatbot**: Answer general health queries responsibly.
- **Legal Summarizer**: Condense legal documents into understandable summaries.
- **Creative Assistant**: Generate creative content (e.g., stories, poems) based on user input.

For this project, I chose to create a healthcare chatbot

Step 2: Apply RLHF Principles

1. **Generate Model Outputs**:
   - Provide multiple outputs for a sample query/task (e.g., explaining a legal term).

Prompt:
You are a responsible and informative healthcare chatbot designed to answer general health queries while ensuring accuracy, clarity, and a reassuring tone. Generate multiple responses to the following user query:

User Query: "What are the symptoms of the flu?"

Response Guidelines:

1. Provide at least three distinct responses that vary in detail, tone, and structure.
2. Ensure each response is medically accurate and easy to understand.
3. Maintain a responsible and non-alarming tone.
4. Where applicable, encourage users to consult healthcare professionals for serious concerns.

Response 1 (Concise & Clear):
"The flu often causes symptoms such as fever, chills, cough, sore throat, runny nose, muscle aches, fatigue, and headaches. If symptoms are severe or persistent, consider consulting a healthcare provider."

Response 2 (Detailed & Informative):
"The flu, or influenza, is a viral infection that affects the respiratory system. Common symptoms include fever, chills, dry cough, sore throat, congestion, muscle aches, headaches, and extreme tiredness. Some individuals, such as young children, the elderly, and those with weakened immune systems, may experience more severe symptoms. If you experience difficulty breathing, persistent chest pain, or high fever that does not subside, seek medical attention promptly."

Response 3 (Forthcoming & Conversational):
"If you're feeling under the weather with a fever, chills, and body aches, you might have the flu! Other common signs include a sore throat, stuffy nose, fatigue, and sometimes nausea. The best thing you can do is rest, stay hydrated, and monitor your symptoms. If things don't improve or get worse, check in with a doctor just to be safe.

2. **Collect Human Feedback**:
   - Define criteria for evaluating responses (e.g., clarity, accuracy, tone). Rank the outputs based on these criteria.

        1. Criteria for evaluating responses include relevancy and accuracy to the prompt. Based on the above criteria, I would rate in other of importance –
            1. Response 2 – most relevant and accurate
            2. Response 1 – relevant and accurate but not very detailed
            3. Response 3 – relevant, but information provided is rather generalized

3. **Train a Reward Model**:
    - Simulate a reward signal using the rankings. Explain how the reward model improves the main LLM.
        1. Based on the above ranking, I will use response 2 to fine-tune the chatbot model, since the output generated will better service. As a result, the chatbot is able to learn on best-responses, and it is able to apply it to similar user inputs.

Step 3: Incorporate Advanced Prompt Engineering

1. **Write a static prompt and show how dynamic inputs can improve its relevance**.
    1. **Static prompt example** -"What are flu symptoms?"
    2. **Dynamic input example-**
        1. **User Age:** >=70
        2. **Health Concern:** Severe symptoms
        3. **Urgency Level:** High
        4. **Response Tone:** Detailed & informative
2. **Use CoT prompting for a complex query and evaluate whether step-by-step reasoning improves accuracy**.
    1. CoT, also known as chain-of-thought prompting, is used to provide guidance based on a user prompt. In the case of the healthcare chatbot, it can be used to provide important information that is accurate and relevant to the user. In addition, it assists the user in making informed decisions. CoT can also assist healthcare professionals to break down complex ideas into simpler chunks that help their patients improve their understanding, since the healthcare professional is using a structured plan in explanation.

Step 4: Implement Ethical Considerations

1. **Bias Detection**:
    - **Evaluate your AI solution for potential biases in responses. Provide examples and corrections**.
        1. An example of potential bias is from the dynamic input example. In general, one tends to assume that only the elderly can experience severe symptoms with regards to the flu. Therefore, to remove potential bias, I could fine-tune the chatbot to include data from different age ranges, thereby making the generated output more relevant to the user's age and sex.
2. **Data Privacy**:
    - **Outline how you would anonymize sensitive user data during training or deployment**.
        1. Since medical data would be involved in training, I will ensure that all identifiable patient information is removed before training, such as names, addresses, and patient medical history. Therefore, the final model will contain generalized information on various illnesses that are not specific to a user.

Step 5: Evaluate and Report

Some of the metrics that will be used to determine the accuracy of the chatbot include relevancy to user input and user response to mini survey at the end of each output. Such surveys will then be used as a reward mechanism in improving the chatbot's output, thereby increasing the accuracy of generated responses. One of the main challenges in healthcare chatbot development is that users tend to enter personal information during queries. As a result, any identifiable information is removed from prompts before fine-tuning the chatbot. In addition, privacy laws differ between regions, so modifying the chatbot to adhere to local laws is another challenge. Therefore, having a compliant version based on region is important in maintaining patient privacy.

Another challenge is maintaining ethical standards when training and fine-tuning the chatbot model. Since data generated may be from a particular group, the result generated will tend to be skewed to the most active users. Therefore, it is important to use a large data set, to account for varying experiences. This ensures that the chatbot can provide accurate responses to user input.