# Part 1 – Write a detailed explanation of the RLHF process covering the following steps- generating outputs, collecting human feedback, training the reward model and fine-tuning the LLM

Reinforcement Learning with Human Feedback (RLHF) is a form of learning that combines human expertise with reinforcement learning to improve LLMs in any given area. The four main processes involved in RLHF are:

**Generating output** – an output is generated when a user enters a prompt. The output generated can be of various file formats, such as texts, images or a combination, depending on the desired output.

**Collecting human feedback** – the output generated is then assessed based on different criteria such as accuracy, and relevance. Humans can provide feedback on the output through assigning ranks, scores and so on to the generated output.

**Training the reward model** – the feedback scores obtained from humans is used to train the LLM model. The model is used to assign reward tot eh new output.

**Fine-tuning the LLM**- LLM model is fine-tuned using reinforcement learning, which uses the reward model to improve the original model. As a result, the model's ability to generate better results improves over time.

## Part 1a – List three practical applications of RLHF in industries such as healthcare, customer service and creative writing

**Healthcare** – RLHF can be used in healthcare in FAQs (frequently asked questions). The LLM model can use responses generated from healthcare providers to fine-tune medical chatbots to be able to answer patient questions accurately.

**Customer Service** – RLHF can be used as a virtual assistant in retail online platforms to answer questions and provide solution to customer inquiries.

**Creative Writing** – RLHF can assist content creators generate ideas on a desired topic. The content creator can use the output to write stories and/or generate content.

## Reflection – Explain one challenge in scaling RLHF and propose a potential solution.

One of the main challenges is the cost of human evaluators. Since humans can generate different responses to the same input, it is important to have a standardized method each evaluator should maintain. Consequently, the reward model will have the same baseline when it is used to train the original model.

# Part 2: Advanced Prompt Engineering

## Application Task

1. **Chain-of-Thought Prompting**:
    - **Write a CoT prompt for a logic-heavy task (e.g., solving a math problem or analyzing a case study).**
        1. **Prompt** – write a step-by-step sequence on how to solve Fibonacci sequence.
    - **Generate an AI response using the prompt and evaluate whether the step-by-step reasoning improves the clarity of the output**.
        1. Based on the output, four different kinds of solutions were provided from recursion (slowest) to matrix exponentiation(fastest). The output generated states that the number used in generating the sequence will determine which method to use.

2. **Prompt Injection**:
    - **Design a prompt for a customer service chatbot. Include static instructions and inject dynamic inputs based on a user's query (e.g., handling a product refund request).**

      **Example Usage: Status of an Online Order**

      **User Query**:
      "I would like to know the status of my order that was placed yesterday at 1pm."

      **Dynamic Inputs**:
      **Context**: Status request for a recent online purchase.
      **Relevant Policy**: Online orders ship within 48 hours of receipt.

      **Generated Chatbot Response**:
      "Hello! I'm happy to assist you in determining the status of your order. Online purchases ship within 48 hours of receipt. You will receive a second email once your order ships. Orders placed on Friday after 6 pm EST and over the weekend are shipped the next business day!"

3. **Domain-Specific Prompts**:
    - Create three prompts tailored for healthcare, legal, and creative writing domains. For each, specify the tone, structure, and expected output.
        1. **Healthcare** – Assuming the role of a physician, what are the specific symptoms of a heart attack in men and women.
        2. **Legal** – As a defense lawyer, what are the limitation of 'pleading the fifth?"
        3. **Creative writing** – as a poet, construct a poem that revolves around flowers

## Reflection

**In 150–200 words, discuss how advanced prompt engineering can make LLMs more adaptable across different industries**.

Advanced prompt engineering makes LLMs adaptable across industries through the tailoring of question-answer with regards to user input. For example, a user in a doctor's office is going to ask questions with regards to healthcare. In the same manner, the same user can inquire about legal ramifications on injuries sustained at the hands of a medical practitioner. While the context is the same, the output generated is going to be different. The response the model generates is fine-tuned to reflect the context in which the question was posed. Therefore, LLM models can generate output based on feedback responses through user interaction. An example of advanced prompt engineering technique is chain-of-thought prompting. Using chain-of-thought (CoT) prompting, a user can guide a model to generate an output that is 'unique' to their circumstance. For instance, a user can enter a prompt that ask what recipes can be created from a list of ingredients. The user can also control the output, by stating what recipes have already been tried, in this manner the LLM model does not repeat those recipes in its output.

# Part 3: Ethical Considerations in LLMs

## Application Task

1. **Identifying and Mitigating Bias**:
    - **Provide an example of a biased prompt and its output**.
        1. Why are electric cars better for the environment?
    - **Write a revised version of the prompt to remove the bias**.
        1. What are the advantages and disadvantages of electric vehicles when compared to their diesel counterparts?

2. **Fine-Tuned Models in Sensitive Applications**:
    - **Choose a sensitive domain (e.g., finance or healthcare).**
        1. Sensitive domain - **finance**
    - **List three potential risks of deploying a fine-tuned LLM in this field and propose mitigation strategies**.
        1. **Personal identifiable information (PII)-** remove all PIIs before feeding the data into an LLM model. This ensures that the output generated is generic.
        2. **Regulations** – some countries limit how information is shared. Therefore, it is important to adhere to regulation requirement(s) of a country or region when fine-tuning an LLM model
        3. **False information** – LLM can generate false information. Therefore, it is important to verify that the output generated by the model is accurate.
3. **Crafting Responsible Prompts**:
    - **Write a prompt for a potentially controversial topic (e.g., climate change, global conflicts) that ensures neutrality, inclusivity, and ethical considerations.**
        1. How can I explain the effects of climate change to someone who remains skeptical while addressing their concerns on the topic?

# Reflection

## In 150–200 words, explain why ethical considerations are critical for building trust in AI systems.

Ethical considerations are crucial for building trust in AI systems because they ensure that these technologies are used reliably, transparently, and fairly. With the increased influence of AI systems in decision making, its influence across various sectors can impact people and society in a positive or negative manner. Without a strong ethical framework, AI can perpetuate biases, make unfair decisions, or violate privacy rights, thereby eroding public trust. Furthermore, transparency is important, as it helps users understand how AI models work. Ethical AI promotes accountability and transparency and ensures that developers and organizations are responsible for the outcomes of their systems. It also ensures that AI systems are designed to limit bias, so the result produced can be trusted. Lastly, respecting user privacy and ensuring data security are key ethical principles in AI development. By aligning AI with ethical standards, developers can foster positive relationships with users and mitigate the risks of harm, leading to greater adoption and trust.