Lab 6: Summarize Dialogue with Generative AI

This lab demonstrates **dialogue summarization** using generative AI models, emphasizing the role of **prompt engineering** in improving model outputs. It compares **zero-shot**, **one-shot**, and **few-shot** inference methods, showing how prompt designs influence the performance of large language models.

Student Name	ID	Section
Lujain Bukassim Almarri	S20106753	1

Objective

The lab aims to:

1. **Dialogue Summarization:**

- **Extractive: Selects and combines sentences directly from the input.**
- **Abstractive: Generates new sentences to summarize the content concisely.**

2. **Prompt Engineering:**

- **Modifies prompts (inputs) to guide the model in performing specific tasks.**

3. **Inference Methods:**

- **Zero-shot: No examples are provided; the model relies entirely on its pretraining.**

- **One-shot: A single example is included in the prompt.**
- **Few-shot: Multiple examples are provided to improve contextual understanding.**

Steps in the Lab

1. Setup

- **Use a pre-trained language model, FLAN-T5, and load a dialogue dataset from Hugging Face.**
- **The dataset contains conversations and corresponding human-written summaries.**

2. Zero-Shot Inference

- **The model summarizes the dialogue without any explicit instructions.**
- **The outputs may lack clarity since the model isn't explicitly directed to summarize the input.**

3. Prompt Engineering

- **Adding clear instructions like "Summarize the following conversation" significantly improves the model's performance.**
- **A properly designed prompt can align the model's output with the desired summary format.**

4. One-Shot Inference

- **Includes one example of a dialogue with its summary as part of the prompt.**
- **This example helps the model understand the structure and nature of the task.**
- **The output becomes more contextually relevant and accurate.**

5. Few-Shot Inference

- **Provides multiple examples of dialogues and summaries in the prompt.**
- **More examples enable the model to better grasp nuances, improving the quality of the generated summary.**
- **It's crucial to balance the number of examples with the model's input limitations (e.g., token limit).**

6. Tuning Generative Configurations

- **Parameters such as:**
    - **Maximum Tokens: Limits the length of the generated summary.**
    - **Temperature: Controls the randomness of predictions.**
    - **Top-p Sampling: Adjusts the probability distribution for selecting tokens.**
- **Adjusting these parameters allows fine-tuning of the model's behavior to balance creativity and accuracy.**

Applications

- **Customer Support**: Summarize long customer interactions into concise reports.
- **Meeting Summaries**: Condense discussions into actionable notes.
- **Virtual Assistants**: Summarize conversations for efficient responses.
- **Healthcare**: Create medical summaries from doctor-patient dialogues.

Key Takeaways

1. **Prompt Engineering**:
    - Clear instructions improve the AI's ability to perform tasks effectively.
    - Examples in prompts help the model contextualize the task better.
2. **Inference Techniques**:
    - Zero-shot is fast but less nuanced.
    - One-shot and few-shot improve understanding and output quality, especially for complex tasks.
3. **Configuration Tuning**:
    - Fine-tuning generative settings helps balance brevity, relevance, and creativity in outputs.

By combining thoughtful prompt engineering, examples, and parameter adjustments, the lab showcases how to leverage generative AI for high-quality dialogue summarization.