Machine Learning System Design

Module 5:
Prompt Engineering vs Fine-tuning

Hamza Farooq

Learning Outcomes

We will be covering topics on:

- What is prompt engineering
- Fine-tuning LLMs
- PEFT
- Validation metrics
- Code Walkthrough in fine-tuning models
    - Local LLM
    - ChatGPT

01

Prompt Engineering ?

Prompt Engineering

Prompt engineering involves designing and refining language model prompts to achieve specific desired outputs. It includes crafting prompts that provide clear instructions, context, or constraints to guide the model's responses.

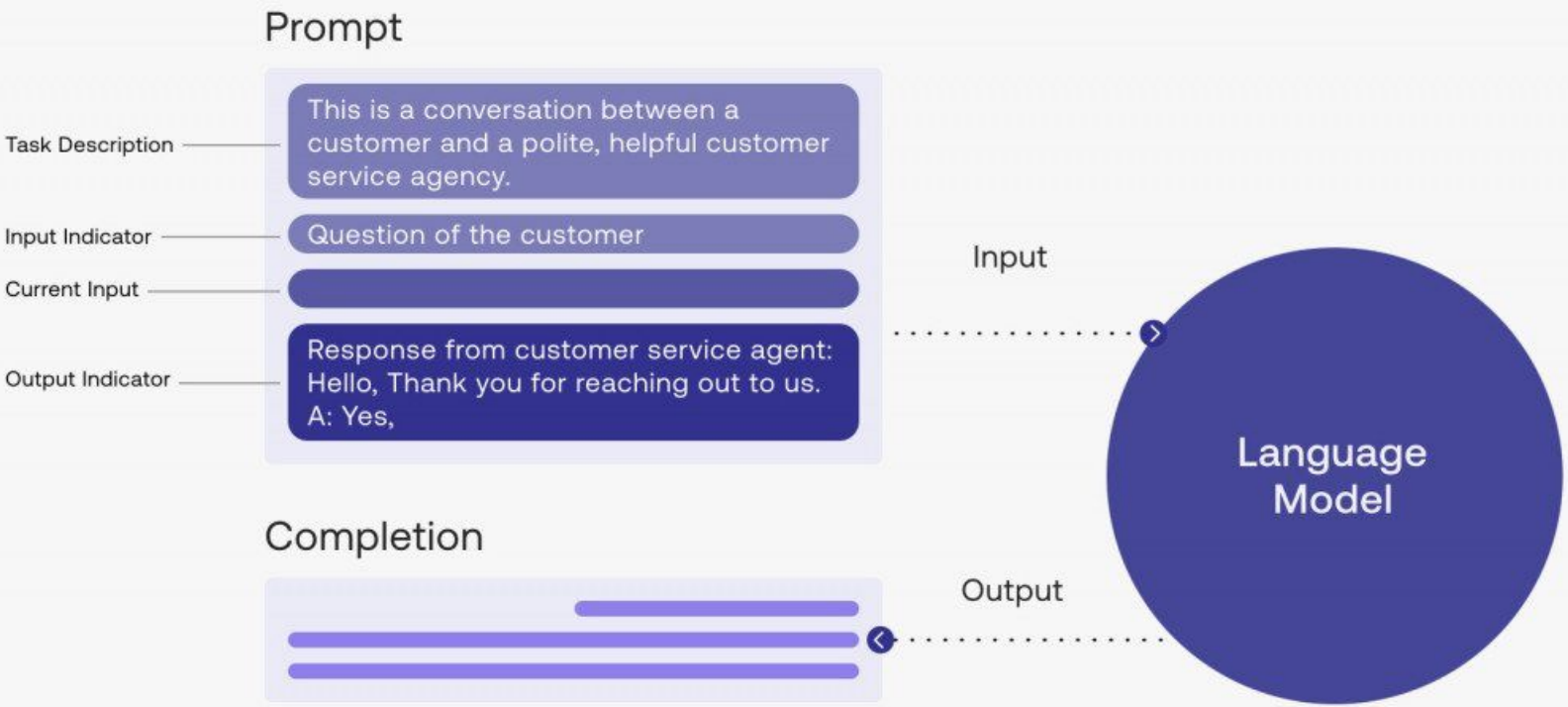
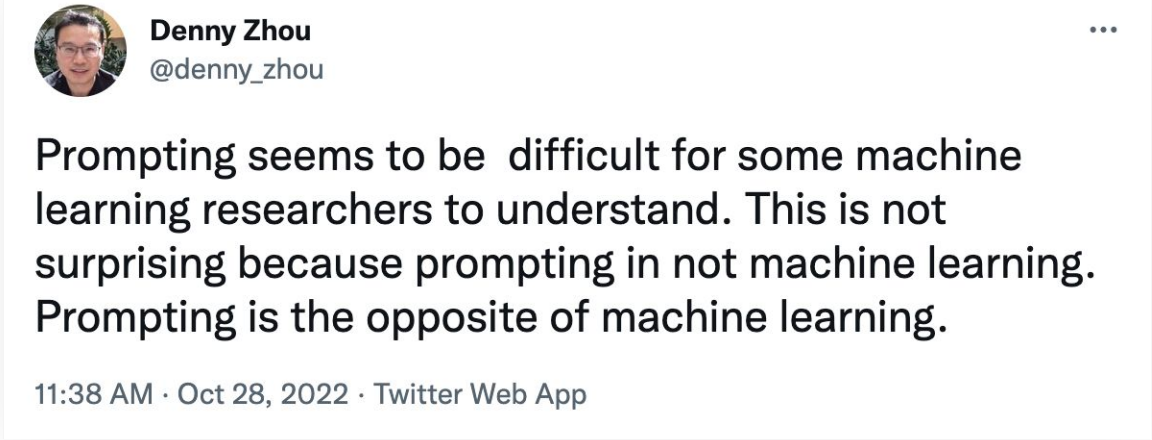


Image: cohere.ai

Prompt engineering is of an more art, than science.



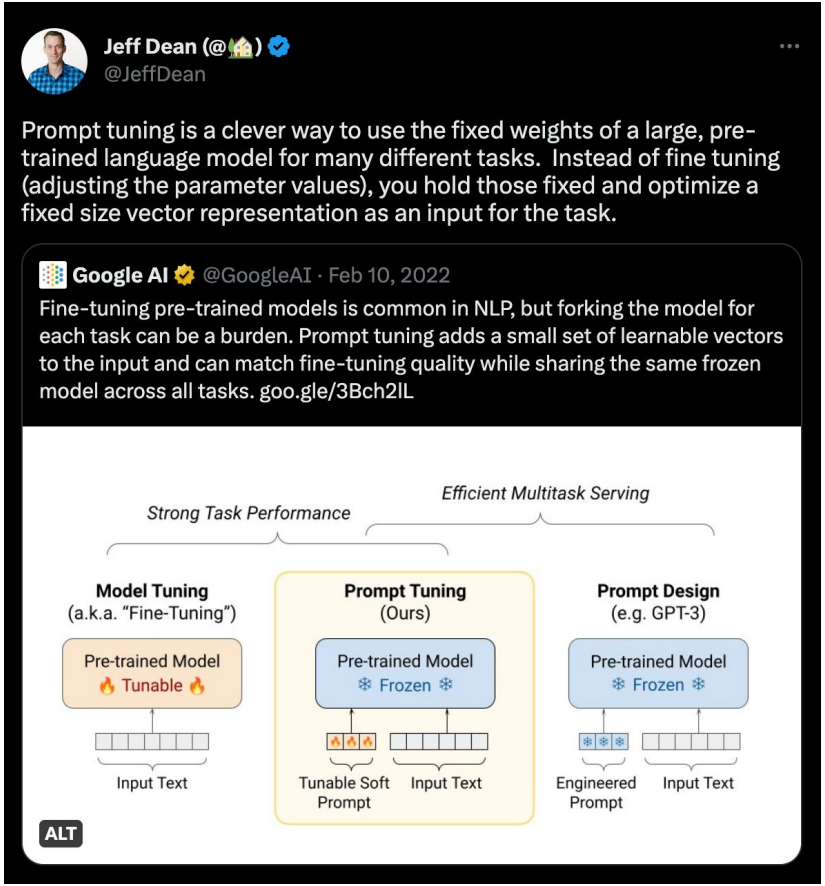
02

Why Fine-tuning ?

Fine tuning

Fine-tuning is necessary to optimize language models for specific tasks or domains by exposing them to task-specific data, improving their performance, and aligning them with the desired outputs.

fine-tuning involves updating a language model's parameters, while prompt engineering entails temporary learning during inference.



Prompt tuning vs Prompt Engineering

- Prompt engineering involves designing and refining language model prompts to improve overall performance, mitigate biases, and enhance accuracy.
- Prompt tuning is a specific aspect of prompt engineering that focuses on fine-tuning or modifying prompts to optimize the model for specific tasks or domains.
- Prompt engineering involves crafting prompts with clear instructions, context, or constraints to guide the model's responses.
- Prompt tuning aims to align the model's output with specific requirements, such as reducing biases, improving factual accuracy, or aligning with ethical considerations, through targeted adjustments to the prompts.

Prompt tuning vs Prompt Engineering

Example of Prompt Engineering:

Prompt: "Write an essay about the benefits of renewable energy sources."

In prompt engineering, the developer may refine the prompt by providing additional context or instructions:

Refined Prompt: "Write an essay about the benefits of renewable energy sources, highlighting their positive impact on reducing greenhouse gas emissions and promoting a sustainable future."

Prompt tuning vs Prompt Engineering

Example of Prompt Tuning:

Prompt: "Translate the following English sentence to French: 'The cat is sitting on the mat.'"

In prompt tuning, the developer may modify the prompt to achieve more accurate or desired responses:

Tuned Prompt: "Translate the following English sentence to French: 'Le chat est assis sur le tapis.' Please provide a fluent translation that captures the natural tone and meaning of the sentence."

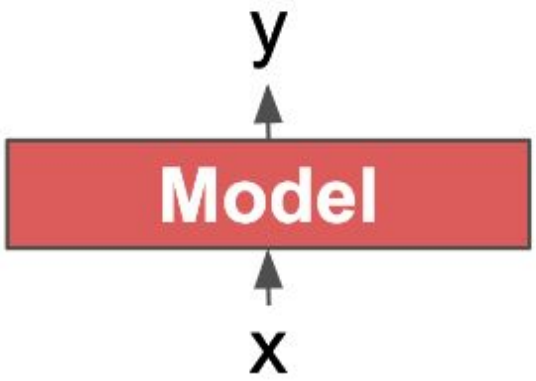
03

Introducing PEFT

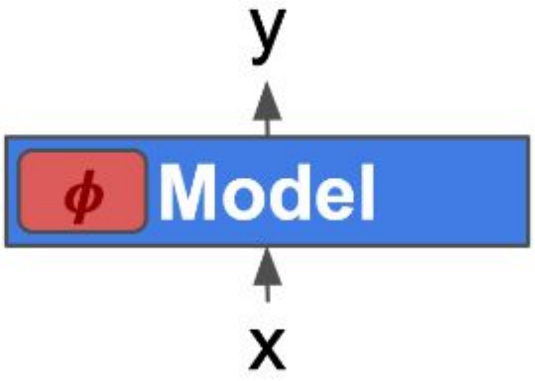
PEFT

Parameter-Efficient Fine-Tuning (PEFT) methods enable efficient adaptation of pre-trained language models (PLMs) to various downstream applications without fine-tuning all the model's parameters.

Fine-tuning large-scale PLMs is often prohibitively costly. In this regard, PEFT methods only fine-tune a small number of (extra) model parameters, thereby greatly decreasing the computational and storage costs. Recent State-of-the-Art PEFT techniques achieve performance comparable to that of full fine-tuning.



(A) Fine-tuning



(B) Parameter-Efficient Fine-tuning (PEFT)

Appendix