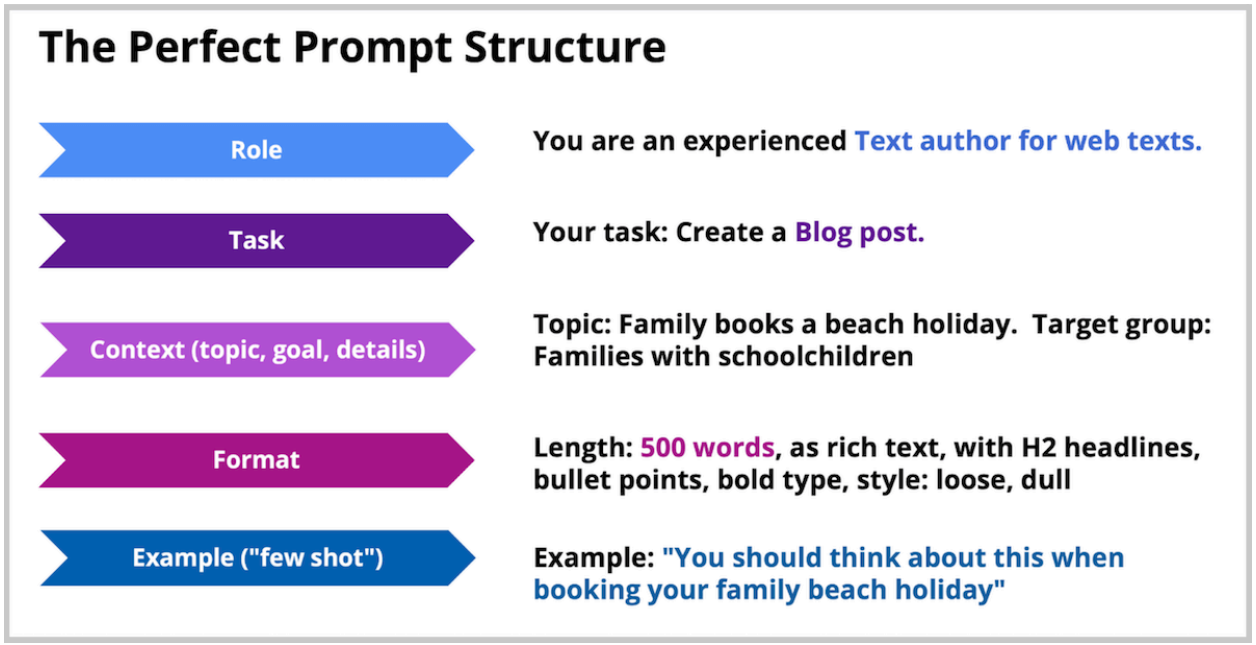
Prompting

What is Prompting?

Prompting refers to the process of interacting with an artificial intelligence (AI) system, particularly generative AI models like Large Language Models (LLMs) or text-to-image generators, by providing it with specific instructions or queries to achieve a desired output. It is the text that we give as an input into the AI to tell it what to do and get our desired output.

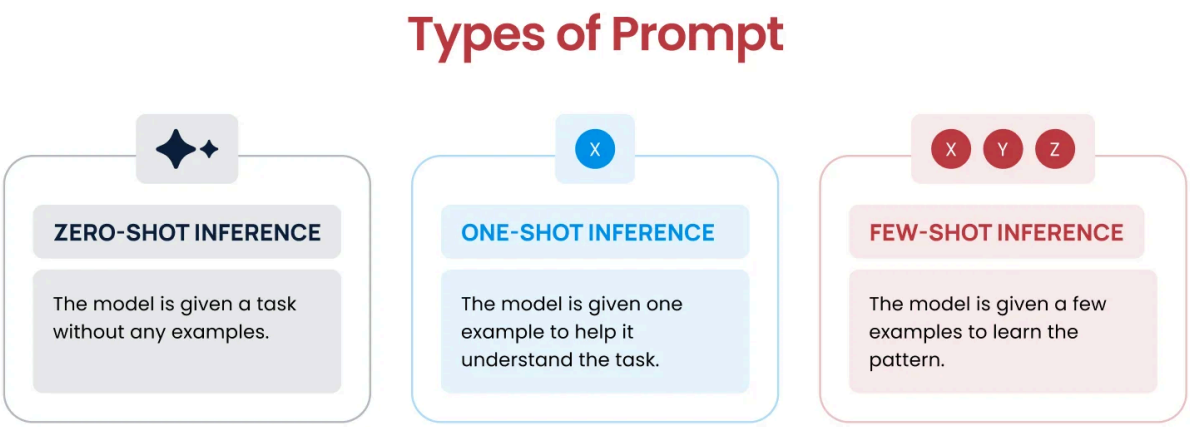


Here are the key points about prompting:

- **Input to the AI:** The prompt is the user's input, which acts as a guide or command for the AI model.

- **Goal-Oriented:** The goal of prompting is to guide the AI to generate a specific, relevant, and high-quality response, whether it's an answer, a summary, a piece of code, or a unique image.
- **Prompt Engineering:** The skill and practice of crafting effective prompts to get the best possible result from an AI model is often called **prompt engineering**. This involves being specific, providing context, defining a role for the AI, or using techniques like "few-shot" or "Chain-of-Thought" prompting.
- **Versatile:** Prompts can be simple questions, detailed instructions with constraints (like tone, length, or format), or even examples for the AI to follow.

Different Types of Prompting



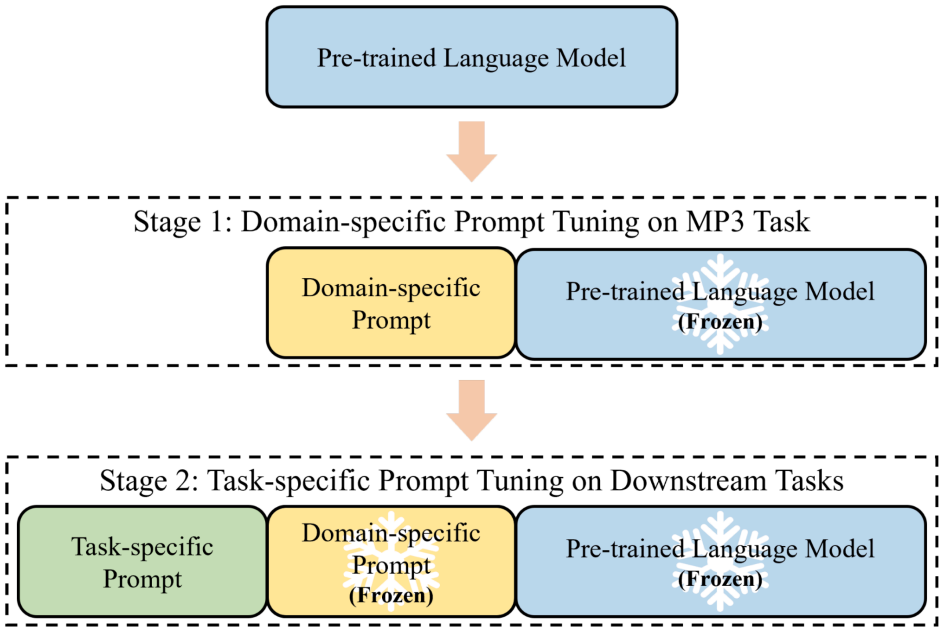
Prompting techniques for generative AI models, especially Large Language Models (LLMs), have rapidly evolved. They range from simple, direct queries to complex, multi-step instructions designed to elicit high-quality, reasoned, or creative outputs.

Some Foundational Prompting Techniques are:

- **Zero-Shot Prompting**: Giving the AI an instruction or question with **no examples** of the desired input/output. It relies purely on the model's pre-trained knowledge. Example: "Translate the following English sentence to French: 'The sun is shining today.'"
- **One-Shot Prompting**: Providing the AI with **one example** of the desired input/output format before asking it to perform the task. Example: "Translate 'cat' to Spanish, which is 'gato'. Now translate 'dog' to Spanish."
- **Few-Shot Prompting**: Providing the AI with a **small set of examples** (typically 2-5) of the desired task. This helps the model infer the pattern, style, or constraints you want it to follow for a new, similar task. Example: Topic: Photosynthesis - Summary: Plants use sunlight to make food. Topic: Gravity - Summary: The force that pulls objects toward each other. **Now summarize:** Topic: Black Holes.
- **Chain-of-Thought (CoT) Prompting**: Instructing the model to **show its reasoning steps** before giving the final answer. This significantly improves accuracy on arithmetic,

commonsense, and complex logic tasks. Example: "I had 10 apples. I bought 5 more, then gave 3 to a friend. How many do I have? **Let's think step by step.**"

What is Prompt Tuning?

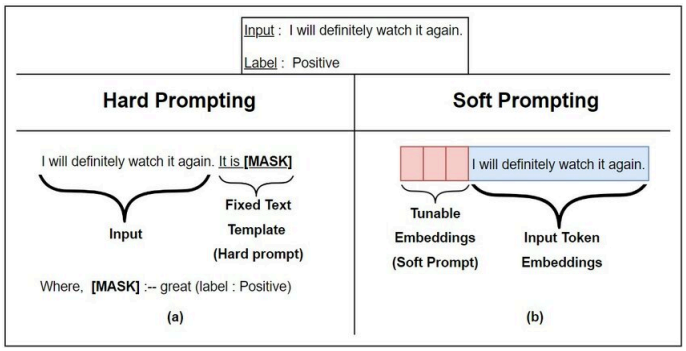


Prompt Tuning is a method for adapting large, pre-trained language models (LLMs) to perform specific tasks more effectively without needing to fine-tune or retrain the entire model.

It's a technique which is used to steer a pre-trained AI model towards better performance on a new or specific task. For eg. Sentiment Analysis.

How does it work?

- **Frozen Model:** The core parameters of the large language model are kept **fixed** or "frozen" during the process. The vast knowledge the model learned during its initial pre-training is preserved.
- **Learnable Soft Prompts:** A small sequence of new, trainable parameters, often called "**soft prompts**" or "prompt parameters," are added to the input.



  - Unlike manually written "hard prompts" (which use actual words and sentences), soft prompts are a string of numbers (embeddings) that are not recognizable to a human.

- **Training:** The model is trained on the specific task data, but only the **soft prompt parameters** are updated using standard optimization techniques like gradient descent. The LLM's frozen weights remain untouched.
- **Guiding the Model:** Through this training, the soft prompt learns to encode task-specific knowledge. When this optimized soft prompt is prepended to the original input text, it acts as a highly effective cue, guiding the frozen LLM to produce the desired output for that task.

Advantages:

- **Efficiency and Speed:** Requires significantly less computational power, time, and data compared to training all parameters of a huge model.
- **Low Data Requirement:** Can achieve good results even with smaller, labeled datasets.
- **Model Size:** The model remains small, as only the tiny prompt parameters need to be stored per task, not an entirely new version of the multi-billion parameter model.