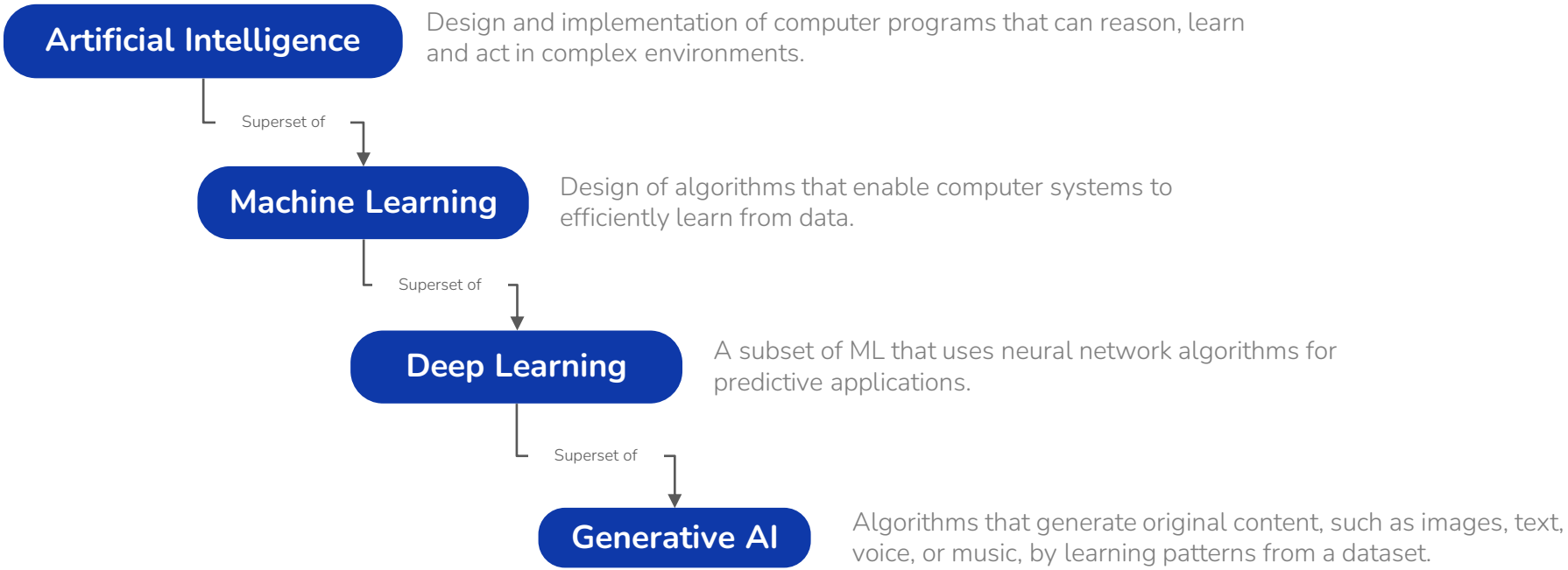
Week 13: Python for Prompt Engineering

Agenda

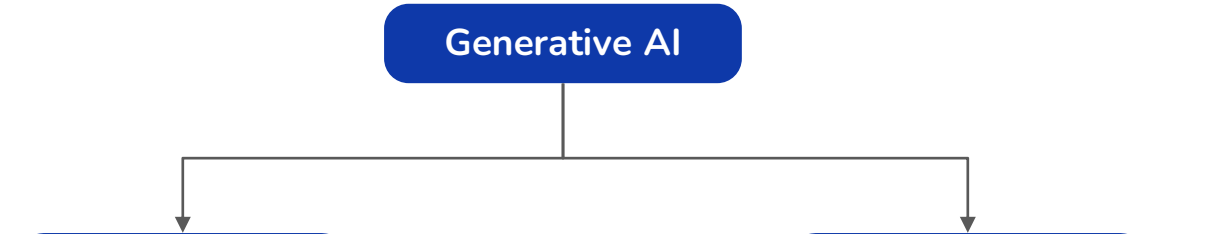
In this session, we will discuss:

- Introduction to Generative AI
- LLMs: A Deep Dive
- Using LLMs with Anyscale APIs
- Prompt Engineering Fundamentals
- Using LLMs for classification and summarization

Generative AI - An Introduction



Generative AI - An Introduction



Large Language Models (LLMs)

LLMs are trained using language modeling, that is, predicting the next word in a sequence. They do so by assigning probabilities to a fixed vocabulary.

The movie is a visually stunning, action-packed, and emotionally resonant thrill ride that will leave you on the edge of the seat from the beginning to end. Overall, the experience was magical.

Vocabulary

positive p = .03

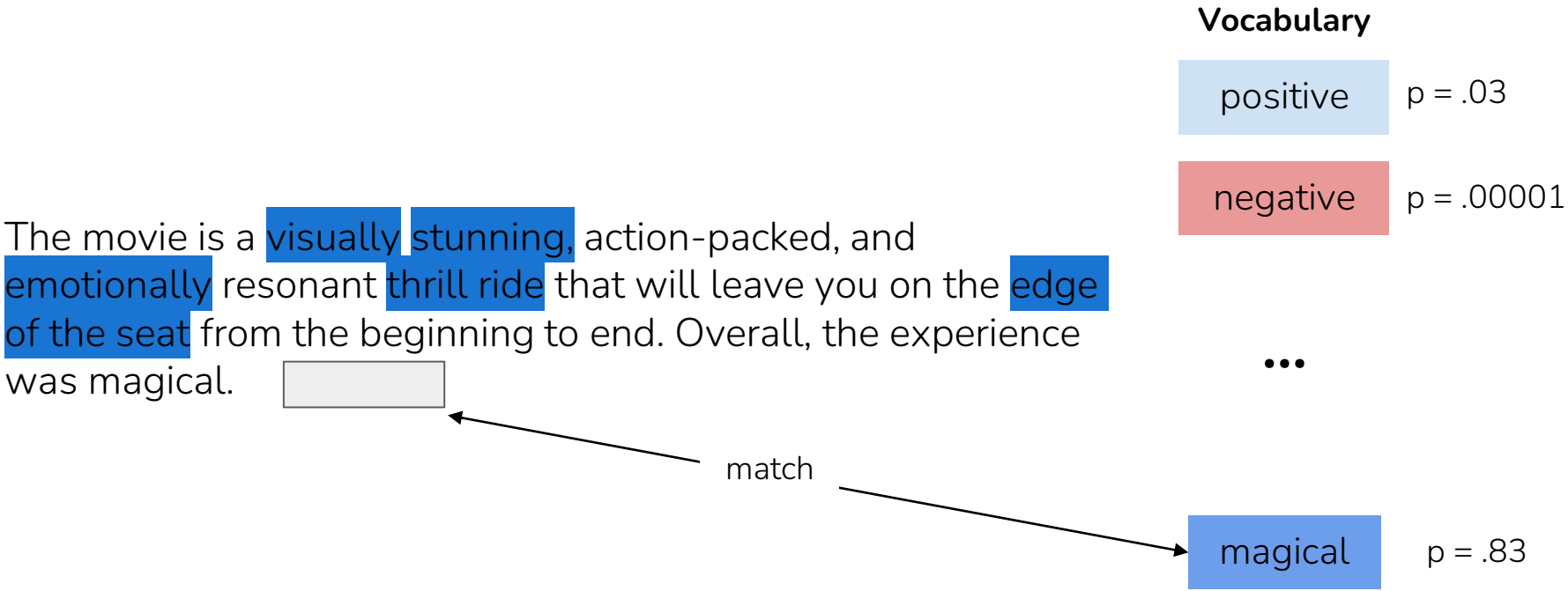
negative p = .00001

...

magical p = .83

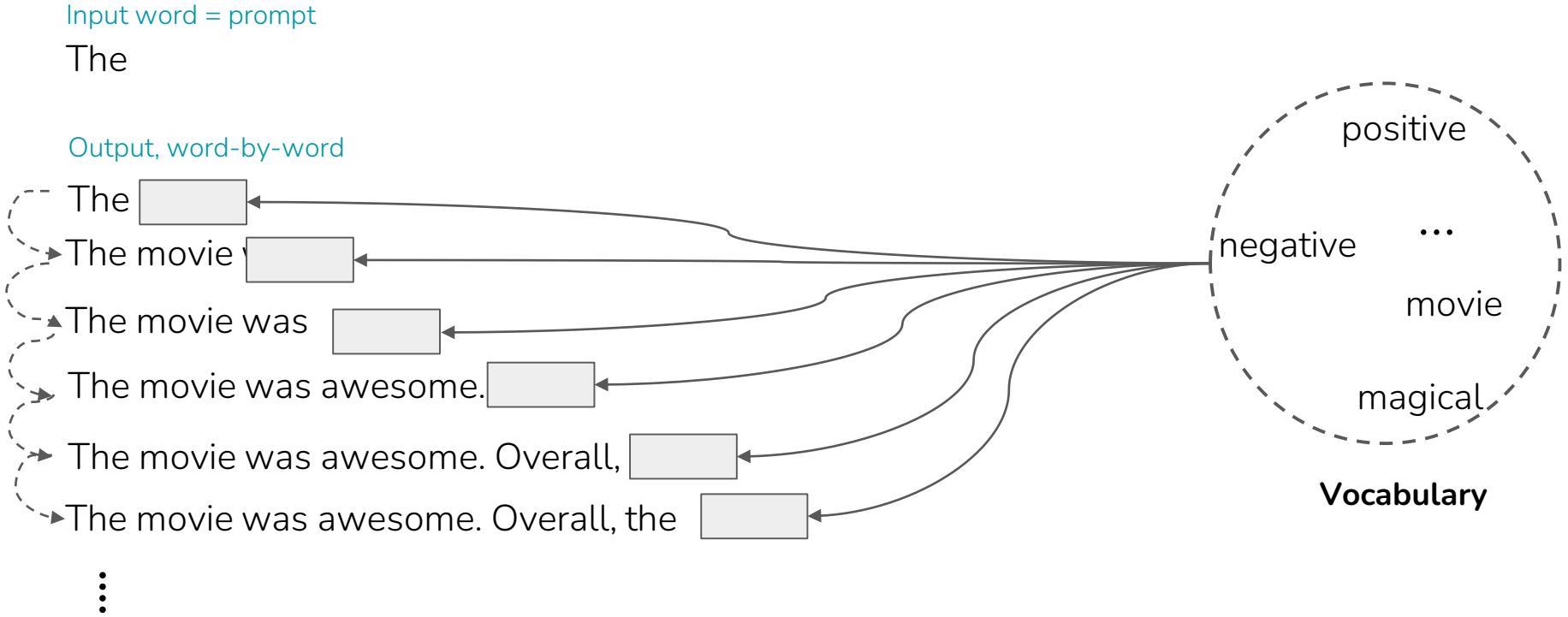
Large Language Models (LLMs)

LLMs are trained using language modeling, that is, predicting the next word in a sequence. They do so by assigning probabilities to a fixed vocabulary.



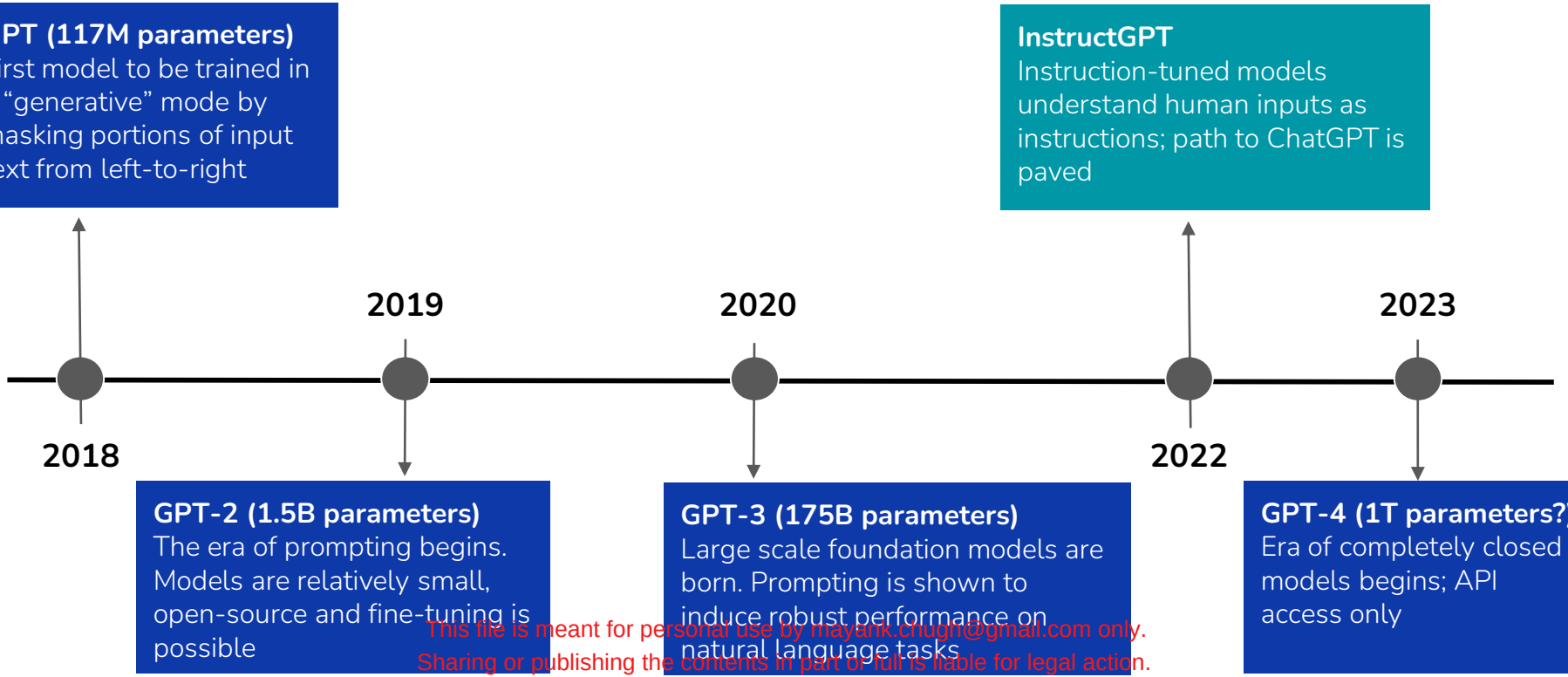
Large Language Models (LLMs)

During inference, the LLM predicts the next word in the input sequence.



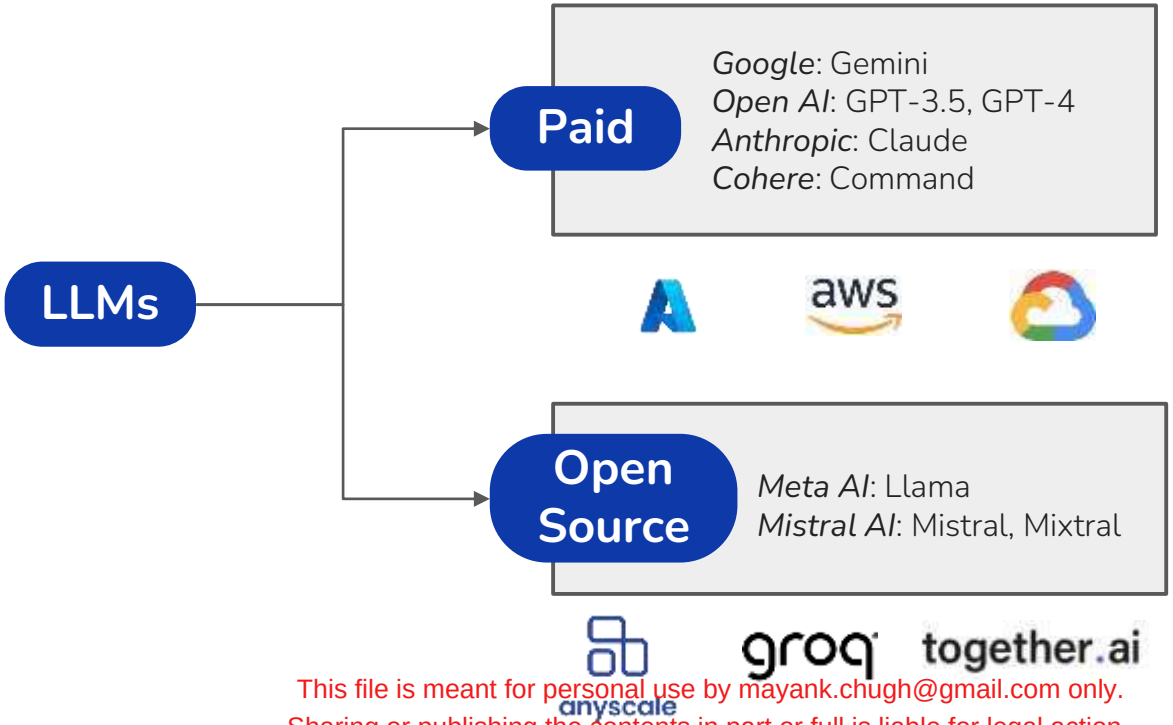
Large Language Models (LLMs)

Over the last 2 years, LLMs (e.g., Open AI GPT) have evolved to be AI assistants



Accessing Large Language Models (LLMs)

LLMs (both paid and open source) can be accessed either through public cloud providers, LLM vendors or using self-hosted company servers.



Accessing LLMs using Anyscale APIs

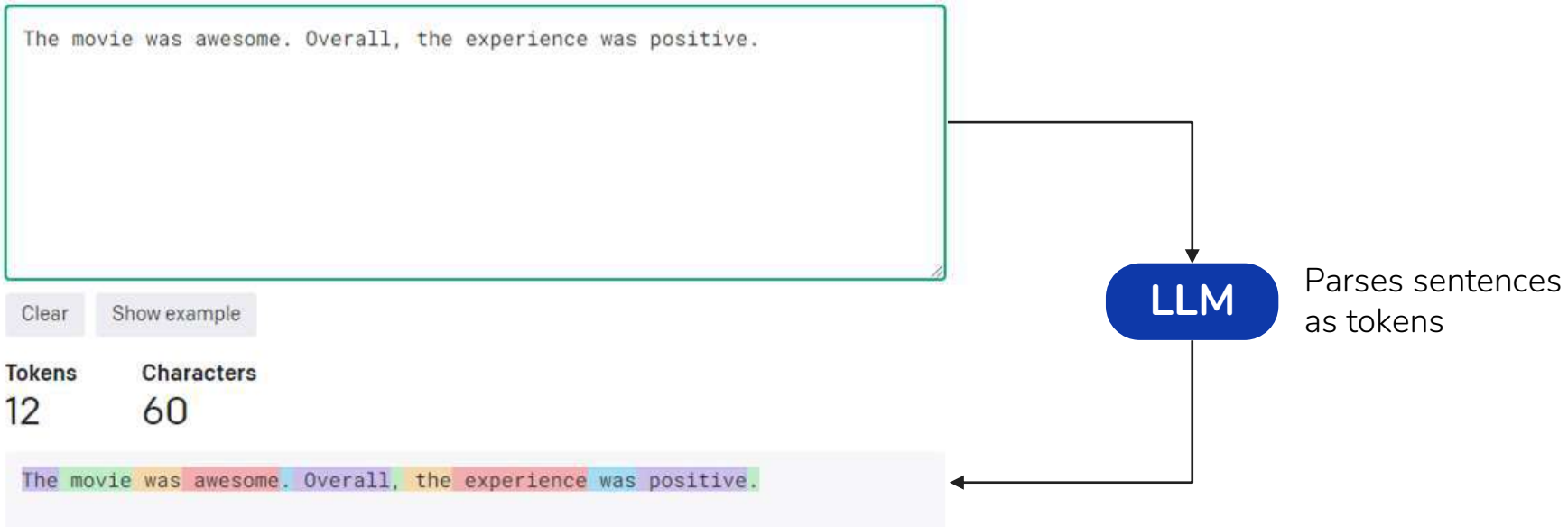
Anyscale provides fast access to a host of open-source large language models.

Model	Price ($/M tokens)
Mistral-7B-Instruct-v0.1	0.15
Llama-2-7b-chat-hf	0.15
Llama-3-8b-chat-hf	0.15
gemma-7b-it	0.15
NeuralHermes-2.5-Mistral-7B	0.15
Llama-2-13b-chat-hf	0.25
Mixtral-8x7B-Instruct-v0.1	0.50
Mixtral-8x22B-Instruct-v0.1	0.90
Llama-2-70b-chat-hf	1.0
Llama-3-70b-chat-hf	1.0
CodeLlama-70b-Instruct-hf	1.0

We use Mistral 7B Instruct and NeuralHermes 2.5 in this course. These models are priced at $ 0.15 per 1 million tokens. Both models have excellent reasoning capabilities and are popular enterprise-grade LLMs.

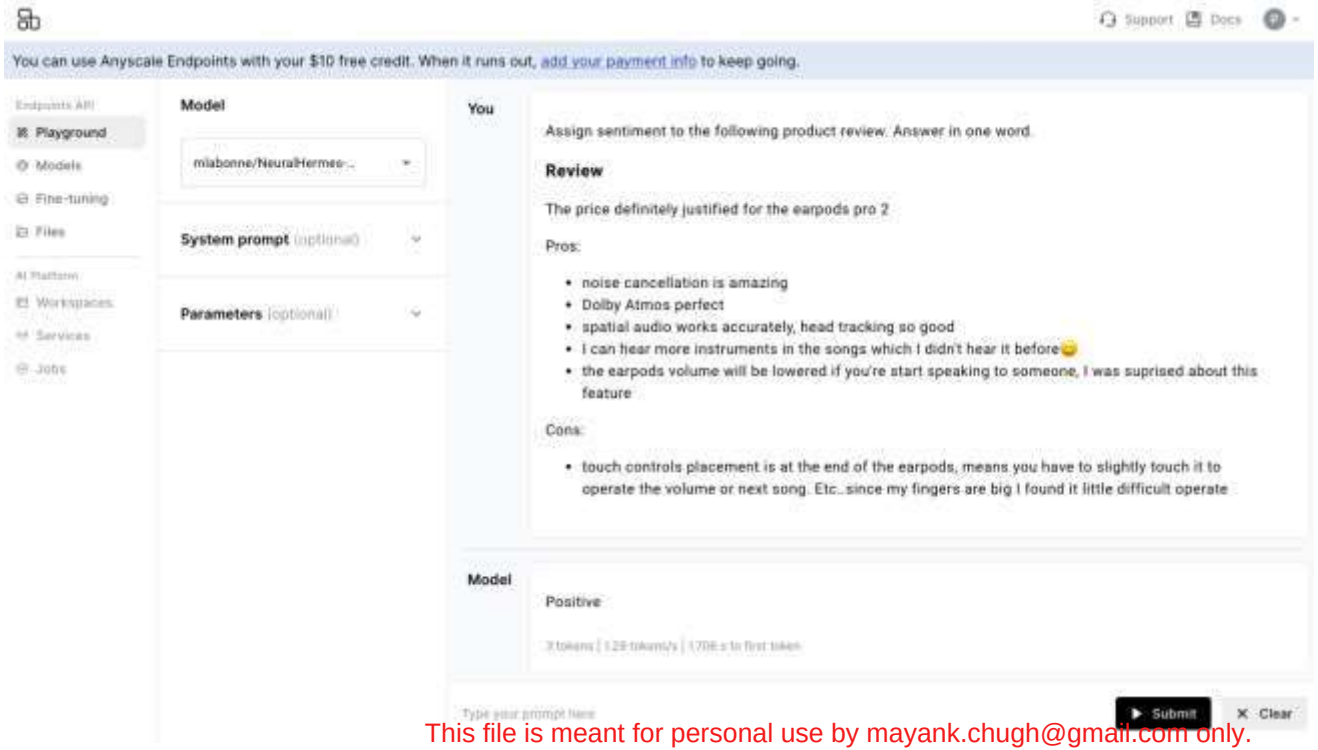
Accessing LLMs using Anyscale APIs

A token refers to a segment or piece of data, such as a word, punctuation, or other meaningful element, into which input text is divided for processing by the model.



Accessing LLMs using Anyscale APIs

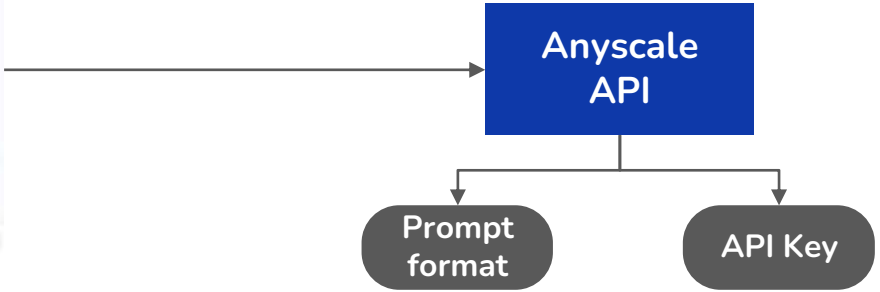
The Anyscale Playground enables iterative development for prompt engineering, that is, designing specific instructions for LLMs to accomplish a task.



Accessing LLMs using Anyscale APIs

The Anyscale Playground enables iterative development for prompt engineering, that is, designing specific instructions for LLMs to accomplish a task.

Playground enables quick iterations on prompts. Once an effective prompt is discovered, it is translated to code for efficient API access

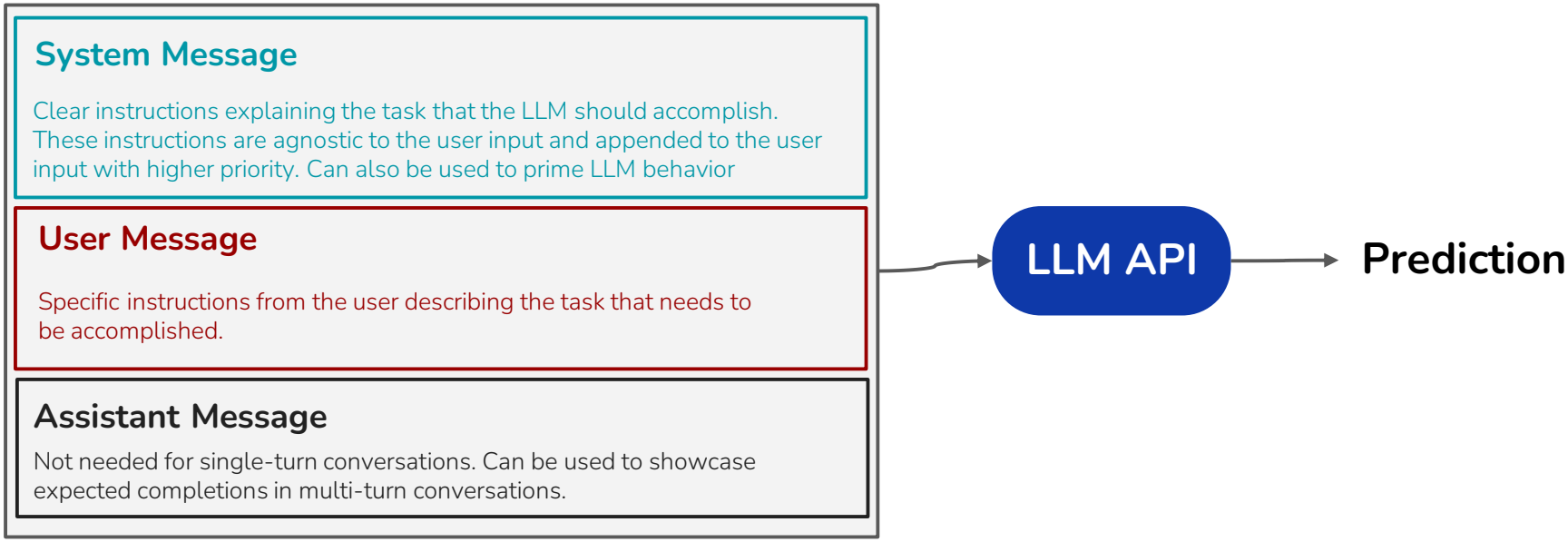


Anyscale API's are compatible with Open AI. This allows seamless switch over between Anyscale and Open AI if needed.

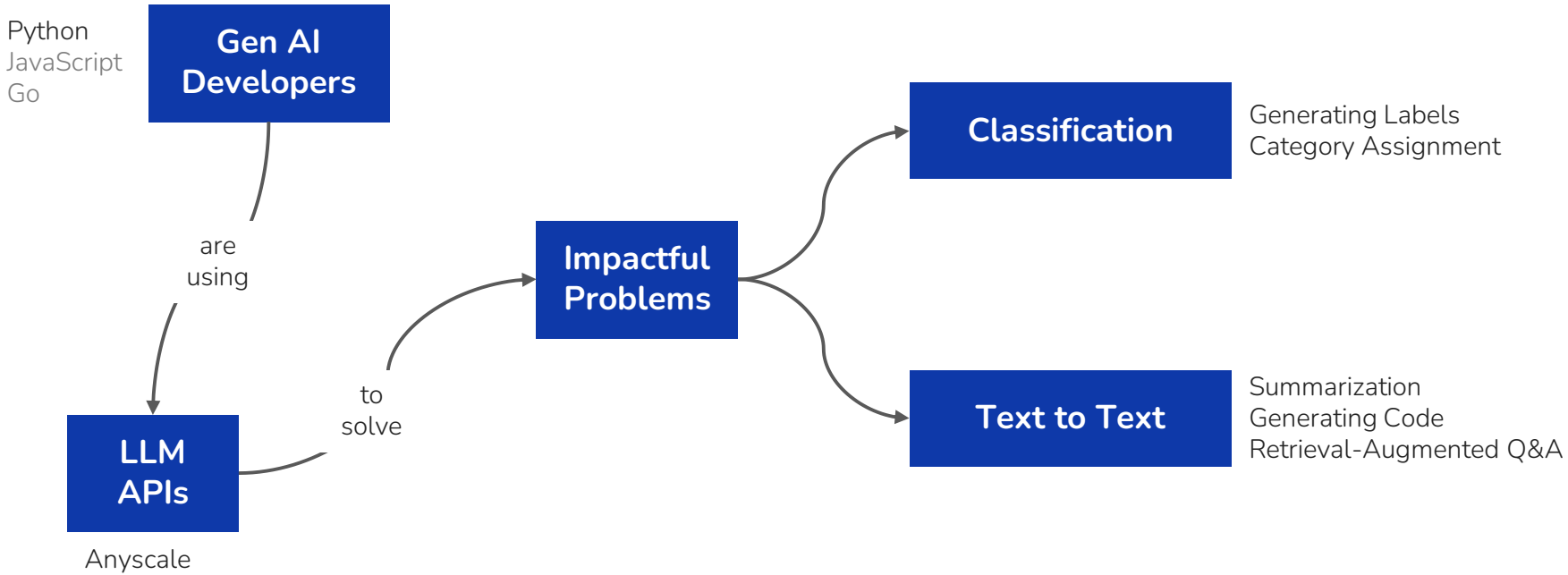
[Notebook]

Anyscale API Prompt Format

Anyscale APIs are compatible with the Open AI APIs and have the following three components.



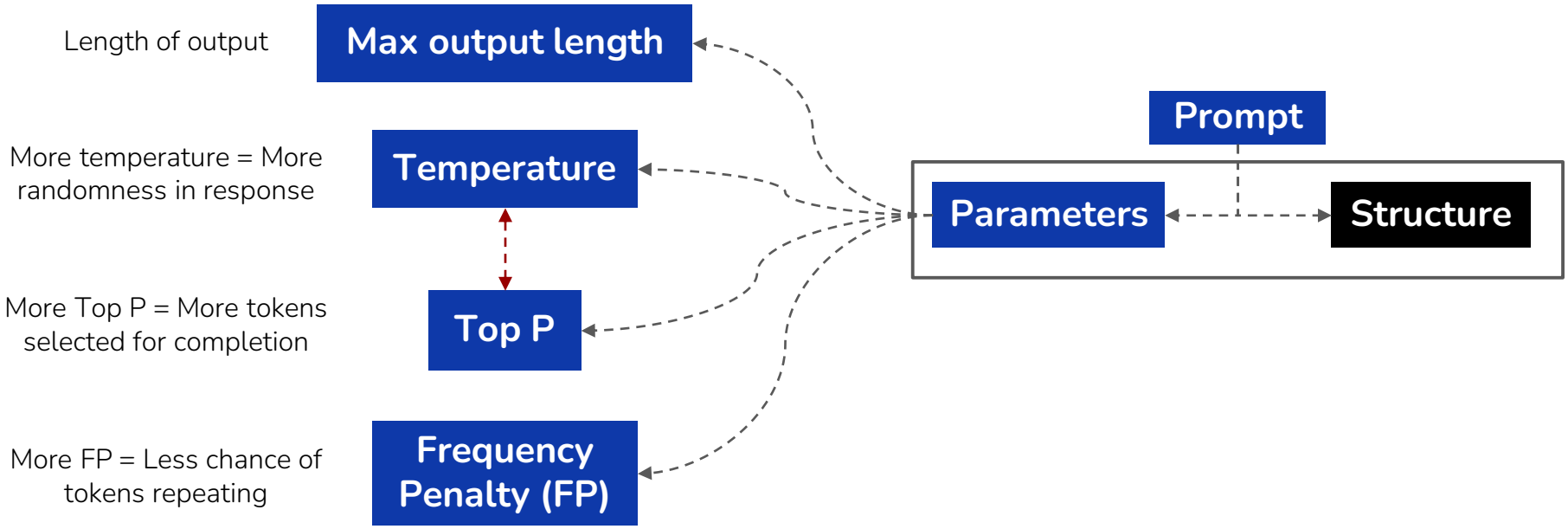
Applications of Generative AI



Prompt Engineering Fundamentals

Prompt = Specific set of instructions sent to a LLM to accomplish a task

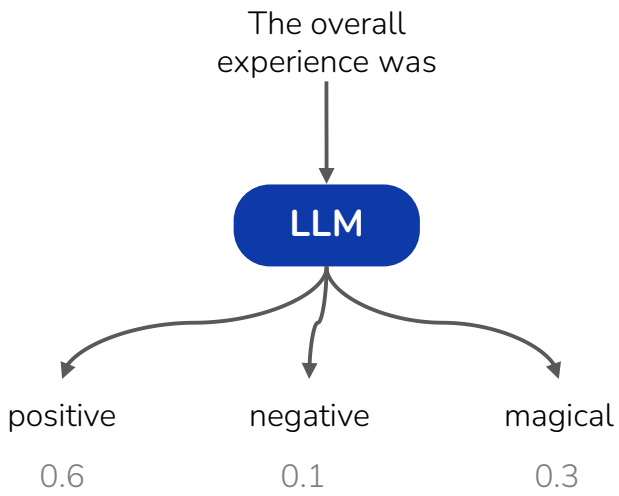
Engineering = Iteratively deriving a specific prompt for the task



[Playground Demo]

Prompt Engineering Fundamentals

Understanding temperature



Temperature = 0

The overall experience was positive
The overall experience was positive
The overall experience was positive

Repeated execution produces the same results

Temperature = 1

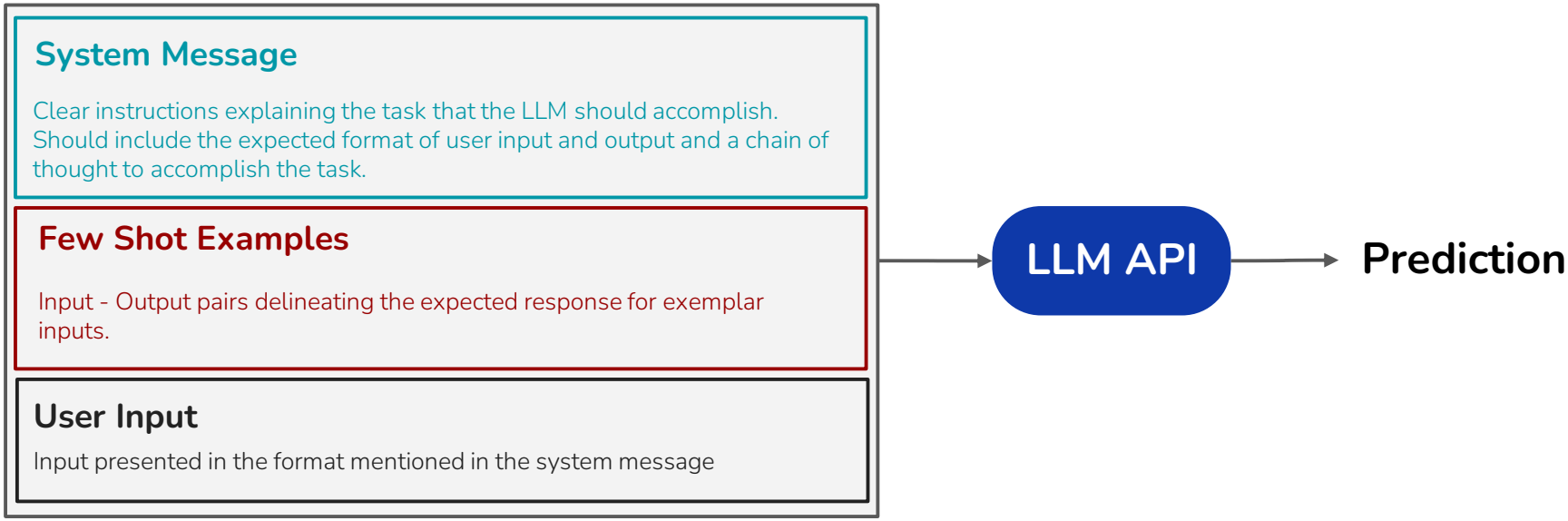
The overall experience was positive
The overall experience was magical
The overall experience was negative

Repeated execution can produce different results

[Playground Demo]

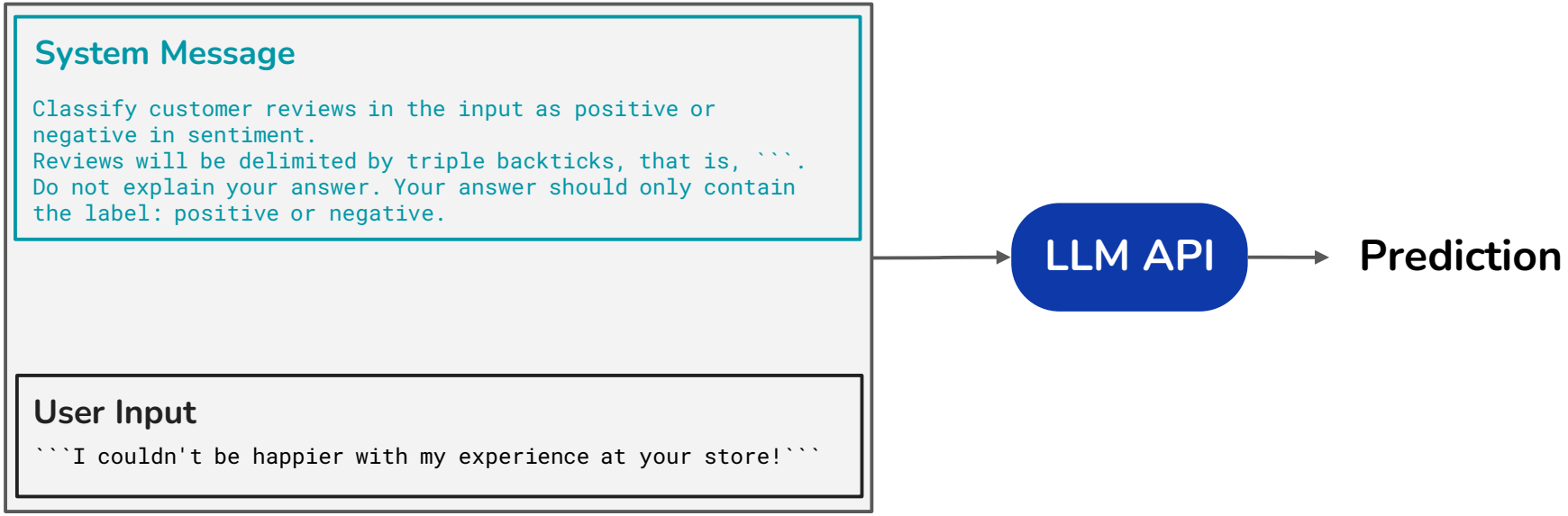
Prompt Engineering Fundamentals

Components of a prompt template



Prompt Engineering Fundamentals

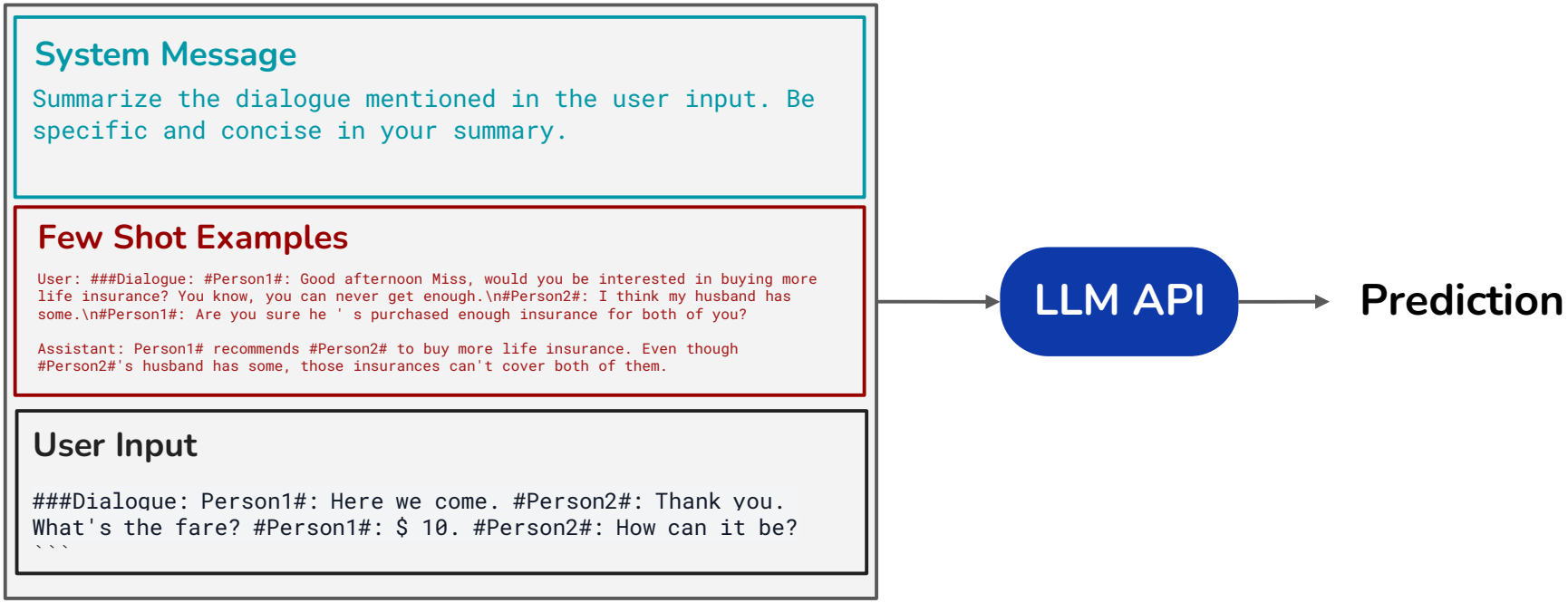
Zero-shot prompt template example - sentiment analysis



[Notebook]

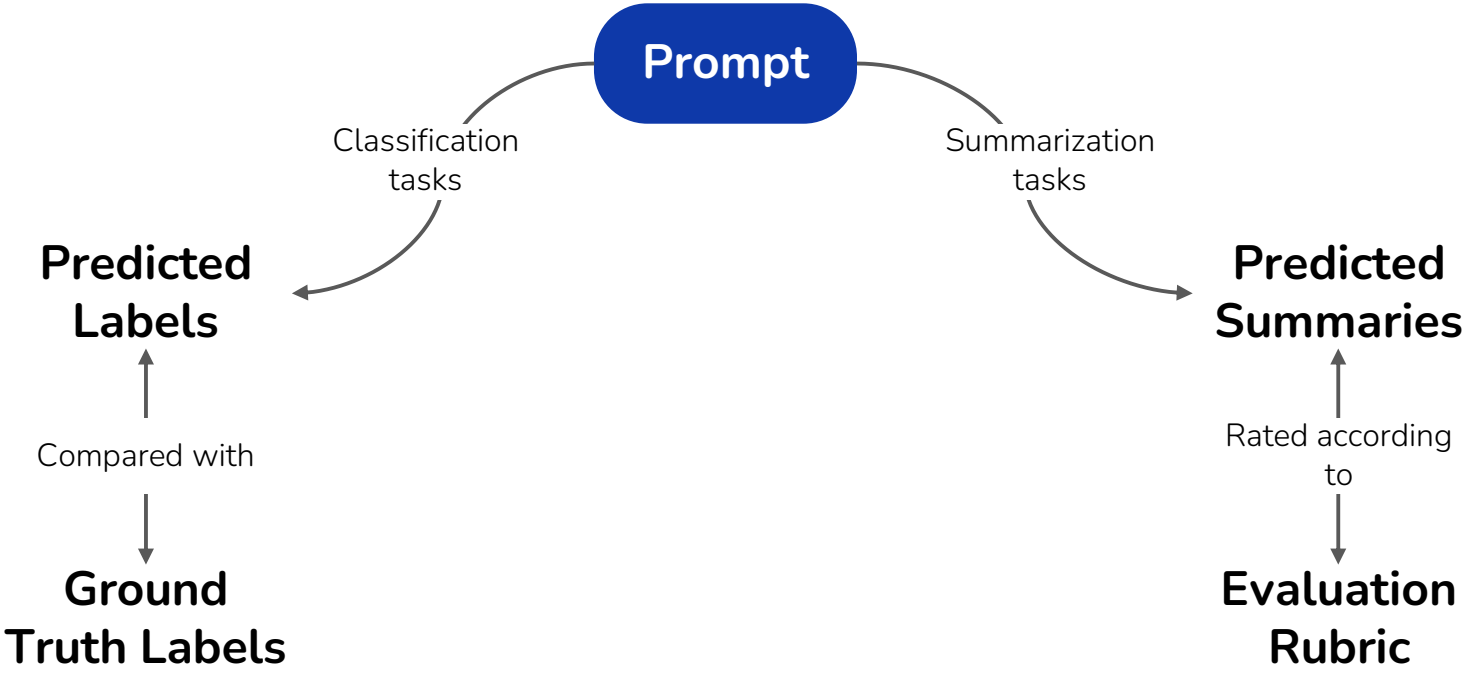
Prompt Engineering

Few-shot prompt template example - Summarization



[Notebook]

Prompt Evaluation



[Notebook]

Summary