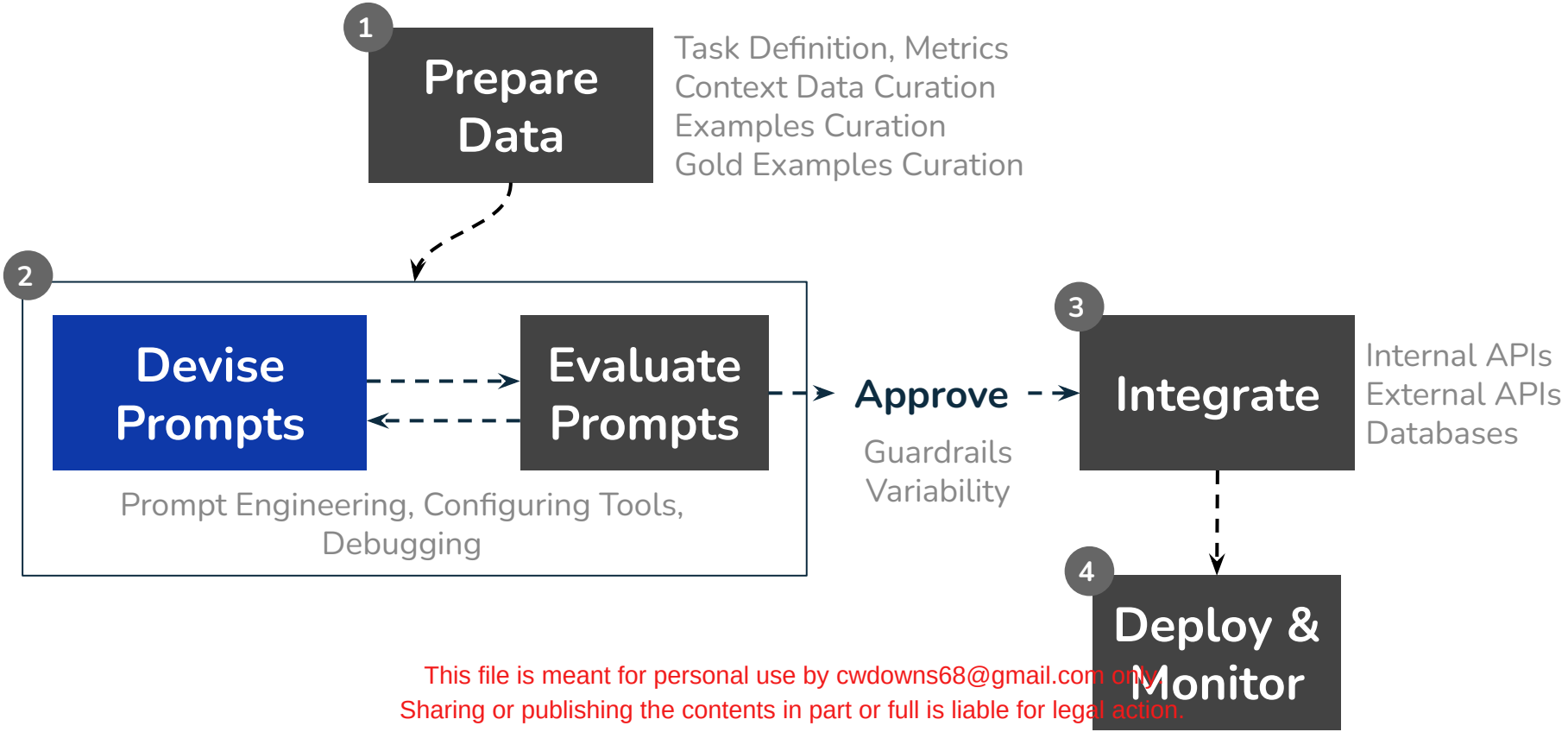
Prompt Engineering

The art and science of designing effective LLM prompts

Agenda

- Large Language Models - An Introduction
- Models & Deployments in Azure Open AI
- Prompt Engineering Fundamentals

Operationalizing Generative AI

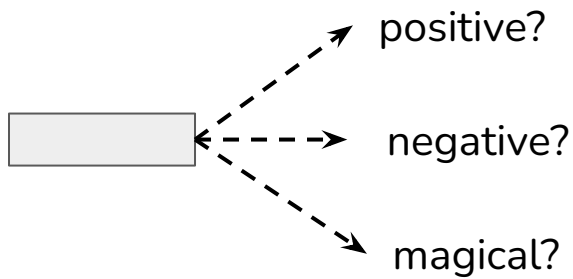


Large Language Models (LLMs)

LLMs are trained using language modeling, that is, predicting the next word in a sequence.

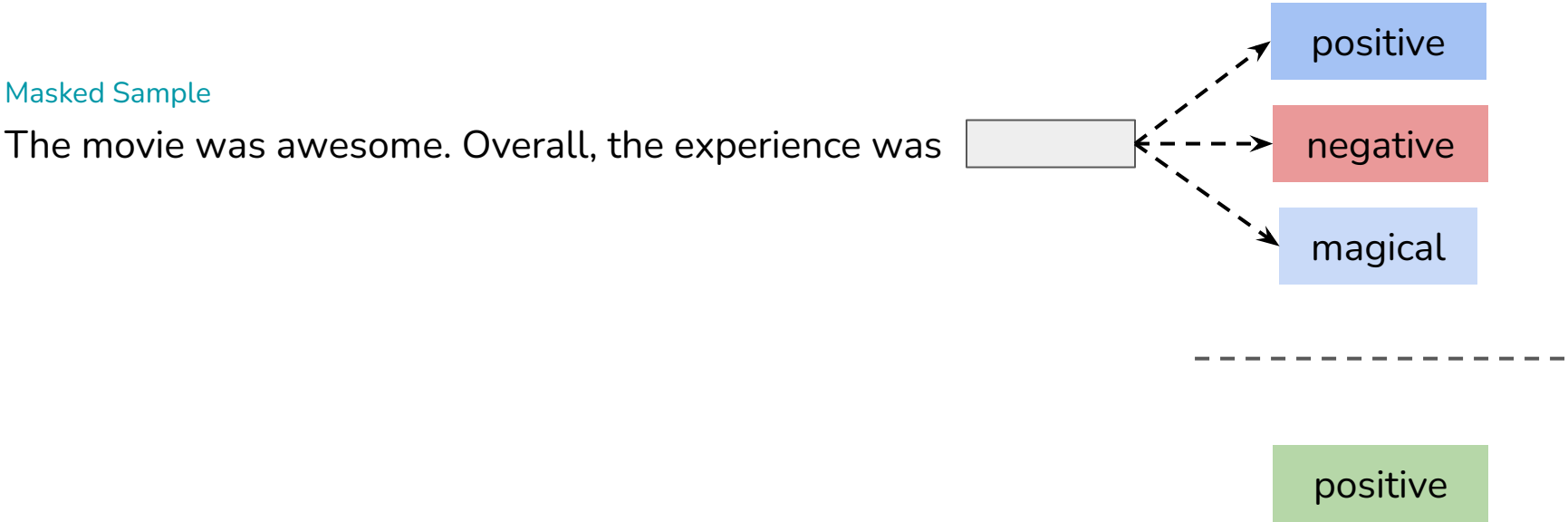
Masked Sample

The movie was awesome. Overall, the experience was [ ]



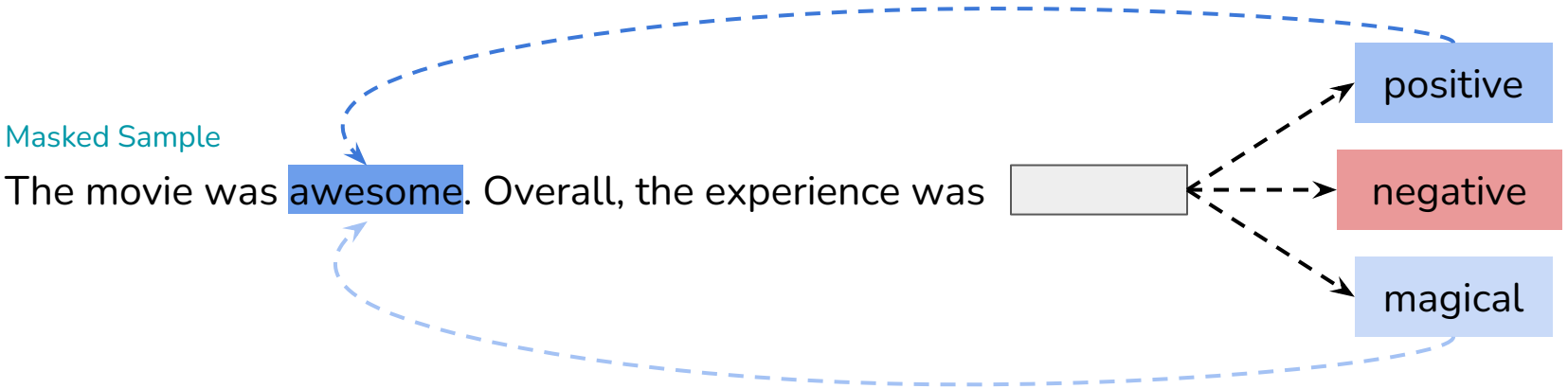
Large Language Models (LLMs)

LLMs are trained using language modeling, that is, predicting the next word in a sequence.



Large Language Models (LLMs)

LLMs are trained using language modeling, that is, predicting the next word in a sequence.



Large Language Models (LLMs)

LLMs are trained using language modeling, that is, predicting the next word in a sequence.

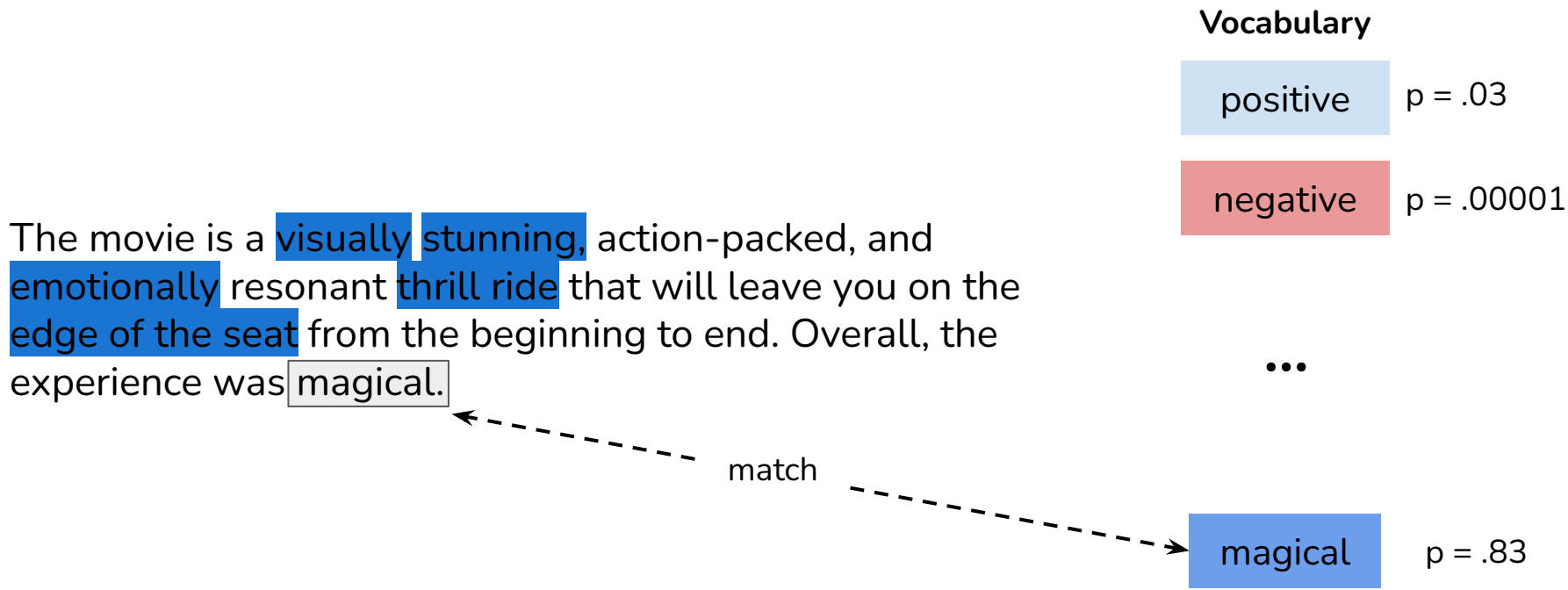
The movie is a visually stunning, action-packed, and emotionally resonant thrill ride that will leave you on the edge of the seat from the beginning to end. Overall, the experience was magical.

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.



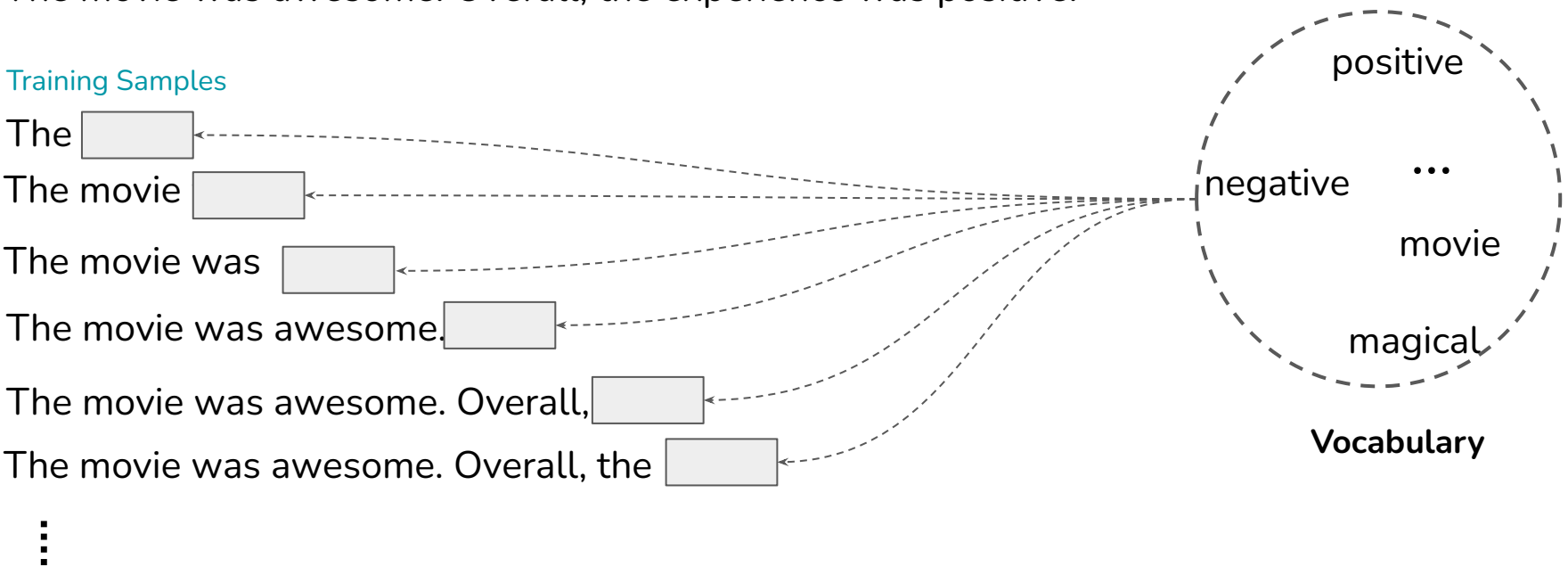
Large Language Models (LLMs)

LLMs are trained using language modeling, that is, predicting the next word in a sequence.

Original sentence

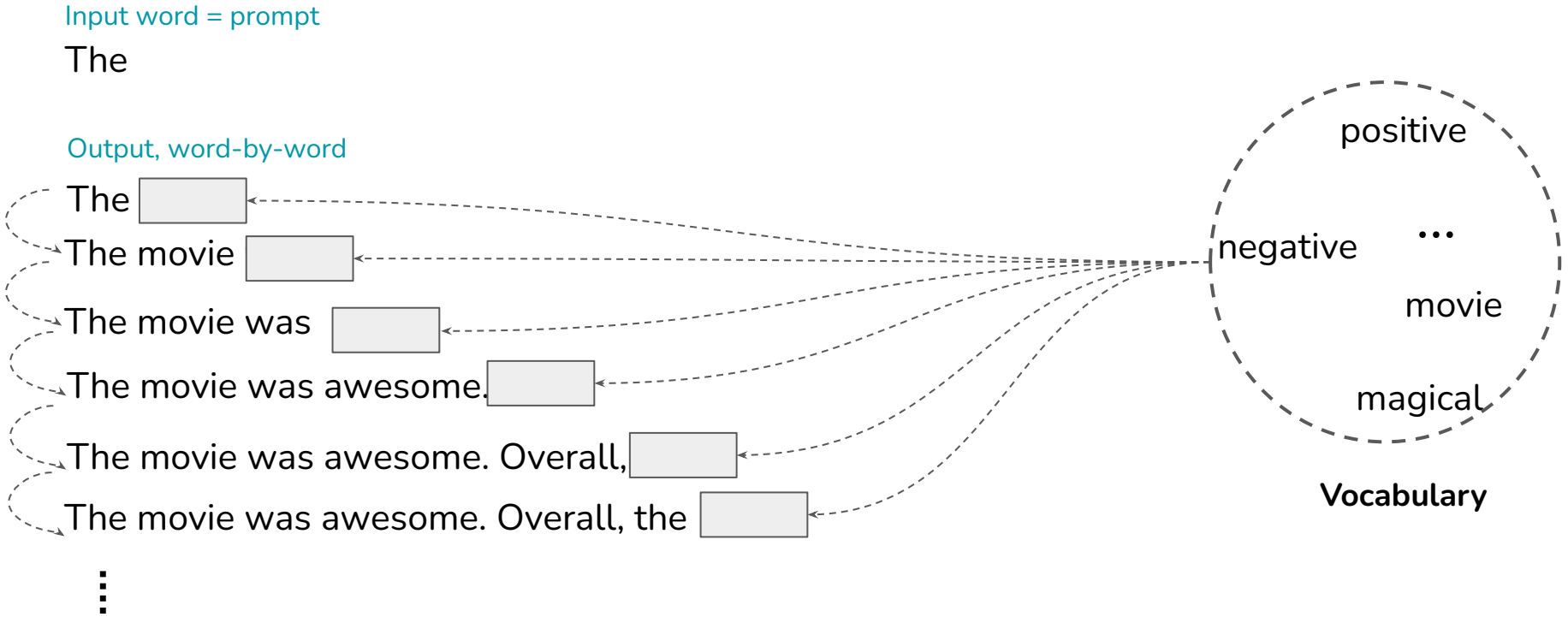
The movie was awesome. Overall, the experience was positive.

Training Samples



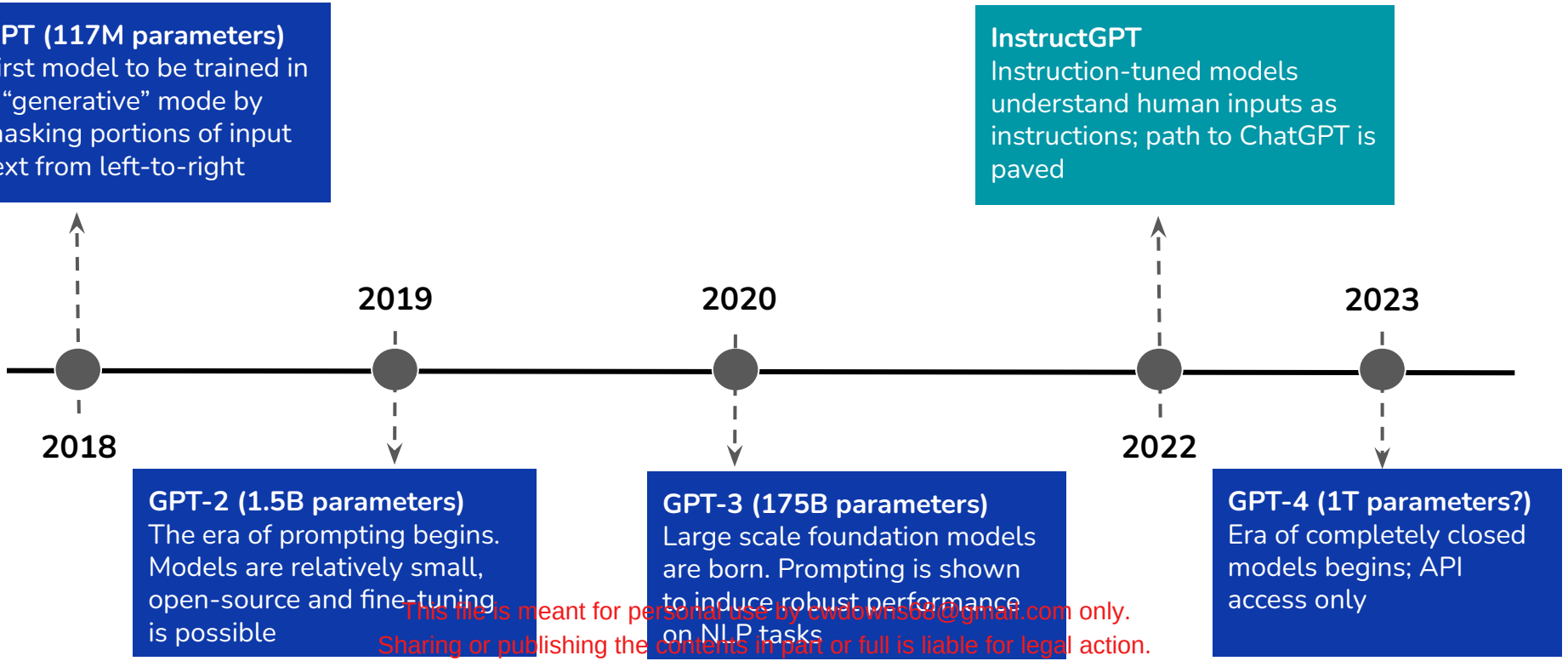
Large Language Models (LLMs)

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



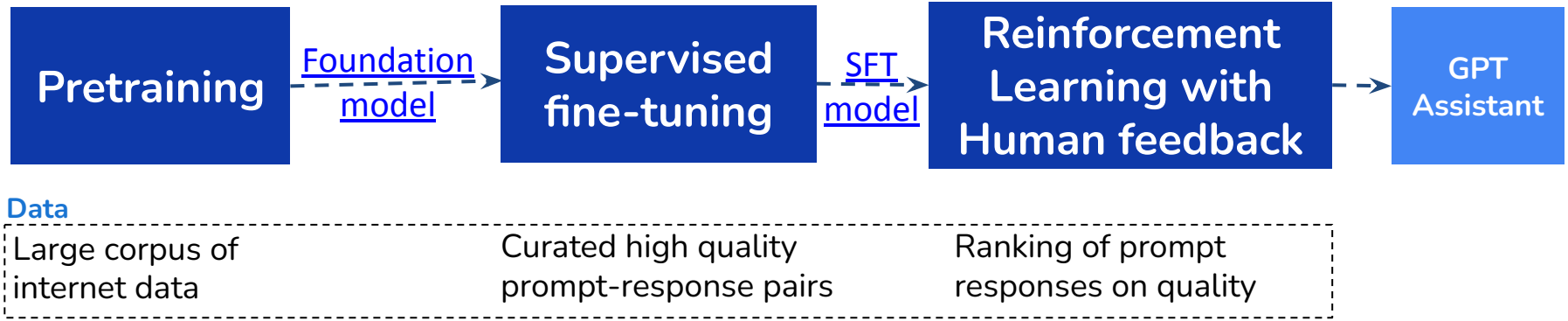
Large Language Models

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



Large Language Models

How are LLMs trained today?



Source: Karpathy (2023)

Models and Deployments

Microsoft

Great Learning
POWER AHEAD

Model

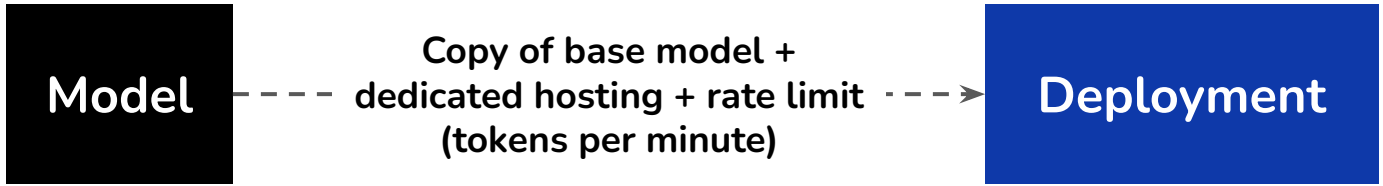
Open AI base model hosted on Azure

Base models

Deploy | Create a custom model | Column options | Refresh | Search

Model name	Model version	Created at	Status	Deployable
gpt-35-turbo	0613	6/19/2023 5:30 AM	Succeeded	Yes
gpt-35-turbo	0301	3/9/2023 5:30 AM	Succeeded	Yes
gpt-35-turbo-16k	0613	6/19/2023 5:30 AM	Succeeded	Yes
text-embedding-ada-002	2	4/3/2023 5:30 AM	Succeeded	Yes
text-embedding-ada-002	1	2/2/2023 5:30 AM	Succeeded	Yes

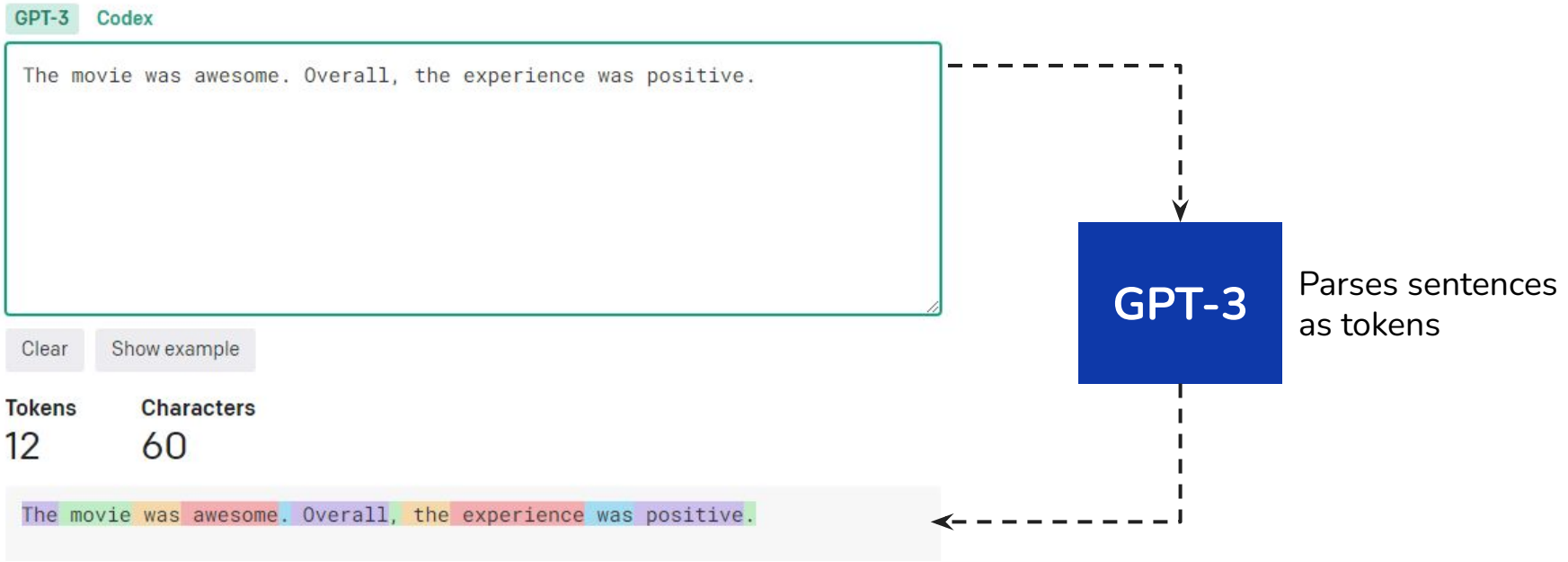
Models and Deployments



	Deployment name	Model version	Capacity	Model deprecati...	Content Filter	Rate limit (Tokens per minute)
☑	gpt-35-turbo	0613	134K TPM	7/5/2024	Default	134000
	gpt-35-turbo-16k	0613	135K TPM	1/15/2024	Default	135000
☐	text-embedding-ada-002	2	133K TPM	2/2/2025	Default	133000

Tokens

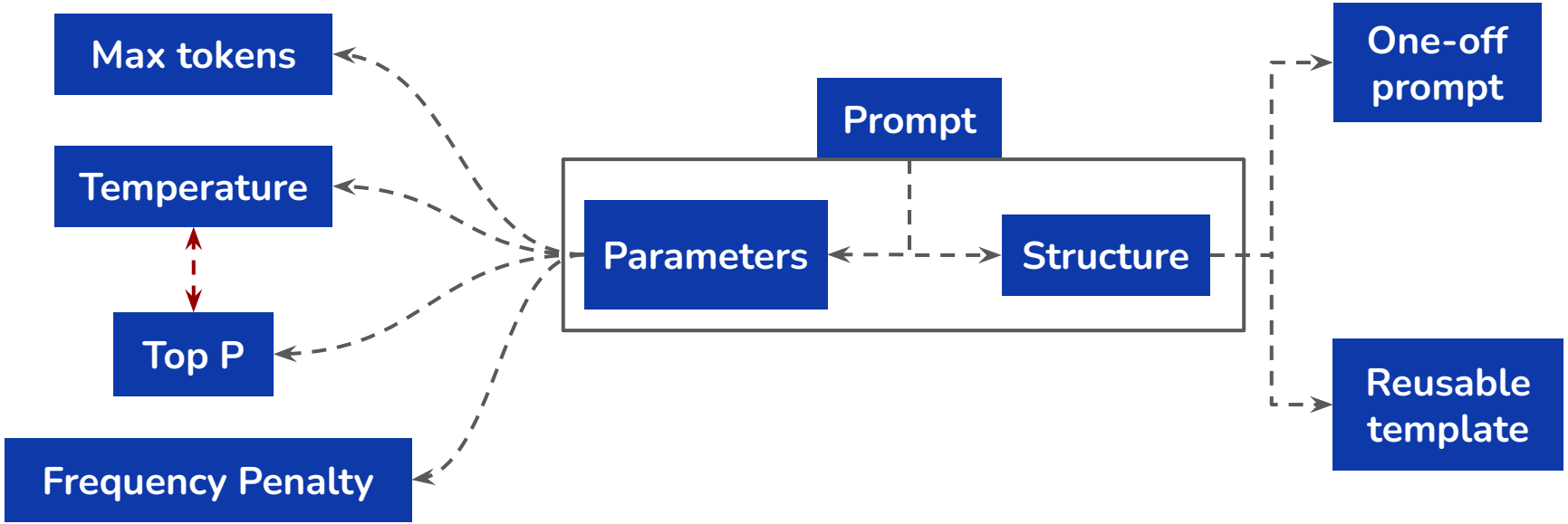
Token = Sequence of characters found in text



Prompt Engineering*

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

Engineering = Iteratively deriving a specific prompt for the task

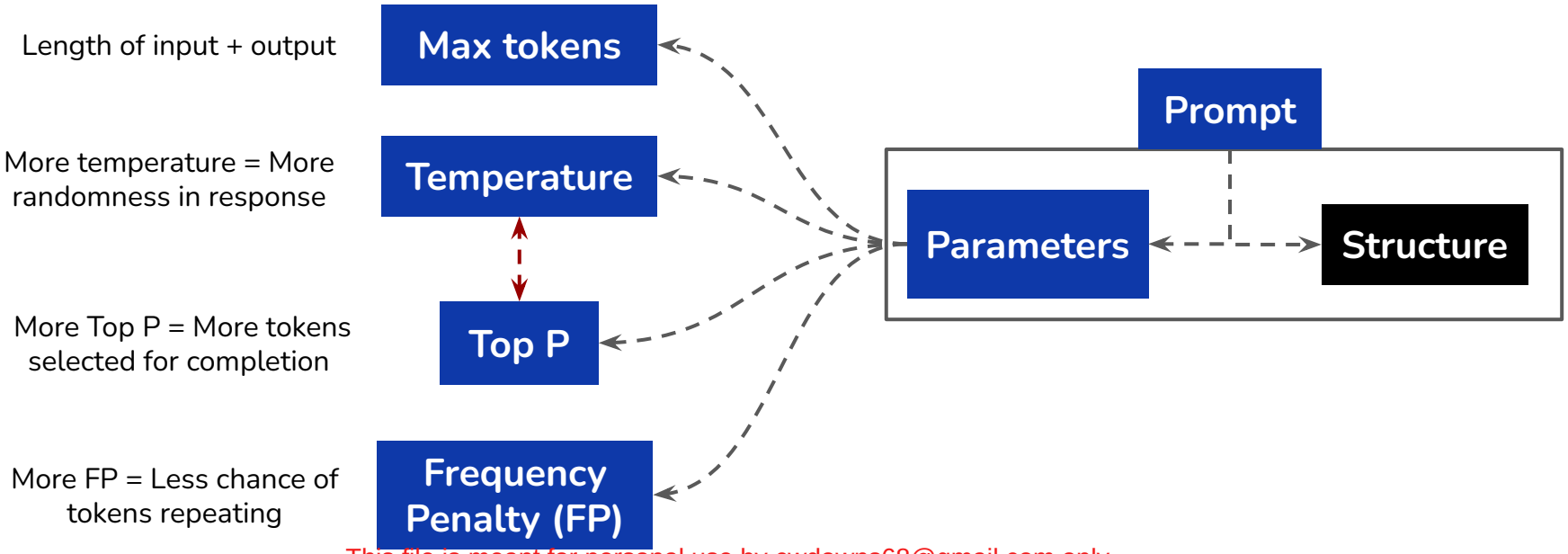


*Also referred to as in-context learning

Prompt Engineering*

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

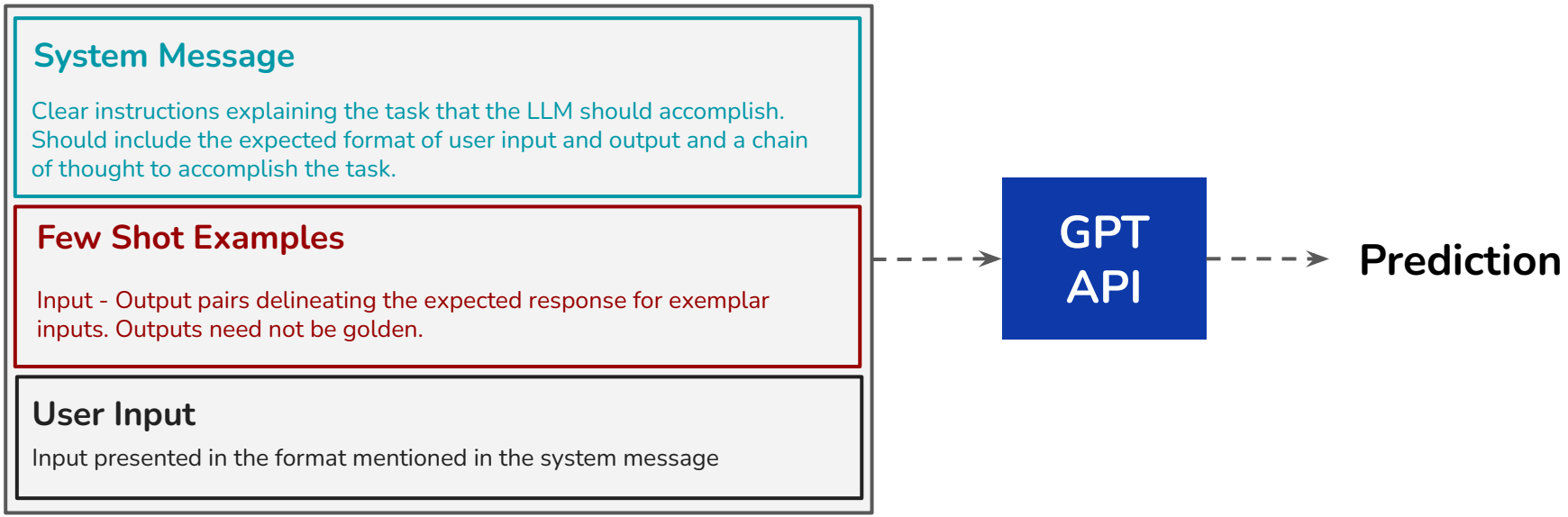
Engineering = Iteratively deriving a specific prompt for the task



***Also referred to as in-context learning**

Prompt Engineering

Components of a prompt template



Pricing

Pricing is per 1000 tokens in the prompt + completion

Tokens	Characters
12	60

The movie was awesome. Overall, the experience was positive.

Models	Context	Prompt (Per 1,000 tokens)	Completion (Per 1,000 tokens)
GPT-3.5-Turbo	4K	$0.0015	$0.002
GPT-3.5-Turbo	16K	$0.003	$0.004
GPT-4	8K	$0.03	$0.06
GPT-4	32K	$0.06	$0.12

Summary