A practical introduction to Generative AI

Speaker:

Giovanni Rosa, PhD

AI Engineer
@Technology Reply

🌐 giovannirosa.com

✉ gio.rosa@reply.it

ITOUG
ITALIAN ORACLE USER GROUP

ITOUG TRAINING DAY @ GALDUS
MILANO 23 OTTOBRE 2024
UNLOCK THE FUTURE WITH ORACLE: MASTERING ANALYTICS, APEX, DATABASE AND GEN AI

SCAN ME

ITALIAN ORACLE USER GOUP
2024
HOSTED BY GALDUS FORMAGZNO
MILAN
ORACLE
ITALIAN
ORACLE
USER GROUP
GALDUS FORMAION
ITALIAN OR ACLE USER GROUP
GALDUS FORMAAZIO
ORACLE
FORMAZONA
ITALIAN ORACLE USER GROUP
GALDUS
2024
ITAIAN ORR-ACLE FRAAAZIO
ITALIAN ORACLE USER GUP 2024
GALDUS FORMAZONO
TAIAN ORACLE USER GROUP 2024
HOSTED BY GALDUS FORMAZIONE
MILAN
ITALIAN ORACLE USER GUP 2027
ITALIAN ORACLE USER GRUP
2024
HOSTED BY GALDUS FORMAINO
ITALIAN ORACLE URAL USER GROUP
2024
HOSTED BY GALDUS FORMAZIONE
ITALIAN ORACLE USER GRUP 2024
ORACLE
GALDUS FORMAZONO
ITAIAN ORACLE USER GROUP
GALDUS FARMAZONO
2024
USER GROUP
GALDUS FORMAZONO
ORACLE
ORACLE USER GROUP 2024
GALDUS FORMAZNO
ITALAN ORACLE USER GUP 2024
ITALIAN ORACLE USER GROUP
2024
MILAN CITY
GALDU FORMAIZONO
ORACLE
ITALLE USER GROUP
2024
GALDUS FORMAZIONE
ITALIAN ORACLE FARMIGZLO
ORACE
FORMAIOA
2024
ORACLE
ITALIAN ORACLE
USER GROUP 2024
MILAN
GALDUS FORMAZONO
ITALIAN
ORACLE
USER GUUP
GALDUS FORMAIAON
ITAL IRACE ORACLE USE GROUP
2024
GALDUS FORMAZONO MILANO
USE GROUP
HOSTED BY GALDUS FORMAGIONO
2024
ORACLE
USER GROUP 224
MILAN
ITALY
ITALIAN ORACLEE USER GUP
MILAN
ORACLE
2024
ITALIAN
ORAL USR GROUP
2024
GALDUS FORMAGONO
ITALIAN ORACLE USER GIUP 2024
GALDUS
FORMAZONO
ITALIN ORACLE USER GUUP
ORACLE
ITAL IRACE USE GROUP
2024
GALDUS FORMAZONO MILANO
ORACE USEGE
USR GUP
ORACLLE
GALDUS FORMAAZIONE
ORAACLE USER GRUP
2024
GALDUS FORMACLCIAIN
ITALIAAN FORACNE
ORACLE
USER GROUP
BY GALDUS FORMZIONO
MILAN
ITALIAN ORACLE USER GUP 2024
GALDUS FORMAIONO
ITALIAAN ORACLE USER GUP 2024
2024 MILAN
GALDUS FORMAZONO
ORACLE
ITALIAN ORACLE USER GRUP 2024
MILAN
ORACLE
GALDUS FORMAZONO
ITALIAN ORAACACLE USER GU GUP 2024
GALDUS FORMAIONO MILAN
ITALIAN ORAC ORCLE USER GUP 2024 HOSTED BY GALDUS FORMAINO
ITALAAN USSER GROUP
MILIZOAO
ORACLE
202 224
MILAN CITY
GALDUS FORMION
ITALIAN ORACLE USER GROUP
OUG
GA GALDUS ORMALONE
HOSTED BY GALDUS FORMAZINO
ORACLE
USER GROUP
2024
ITALAIO - FA ROMAGOZOJO, MILAN
ITTAALIAIN ORACLE OSEORE GURE 2024
MILAN
GALUA GLDUS FAMMAZINO
ITIAN ACLE USER GROUP
GUSDD'S FORMAGONO
GALLDDU FORAANO
ITALIAN ORACLE USER GROUP
2024
MILAN
HOSTED BY GALDUS FORMAZIONE

Bing DALL-E 3

Prompt:

“Generate an image logo for the Italian Oracle User Group 2024 event hosted by Galdus formazione in Milan city”

User

call →

①

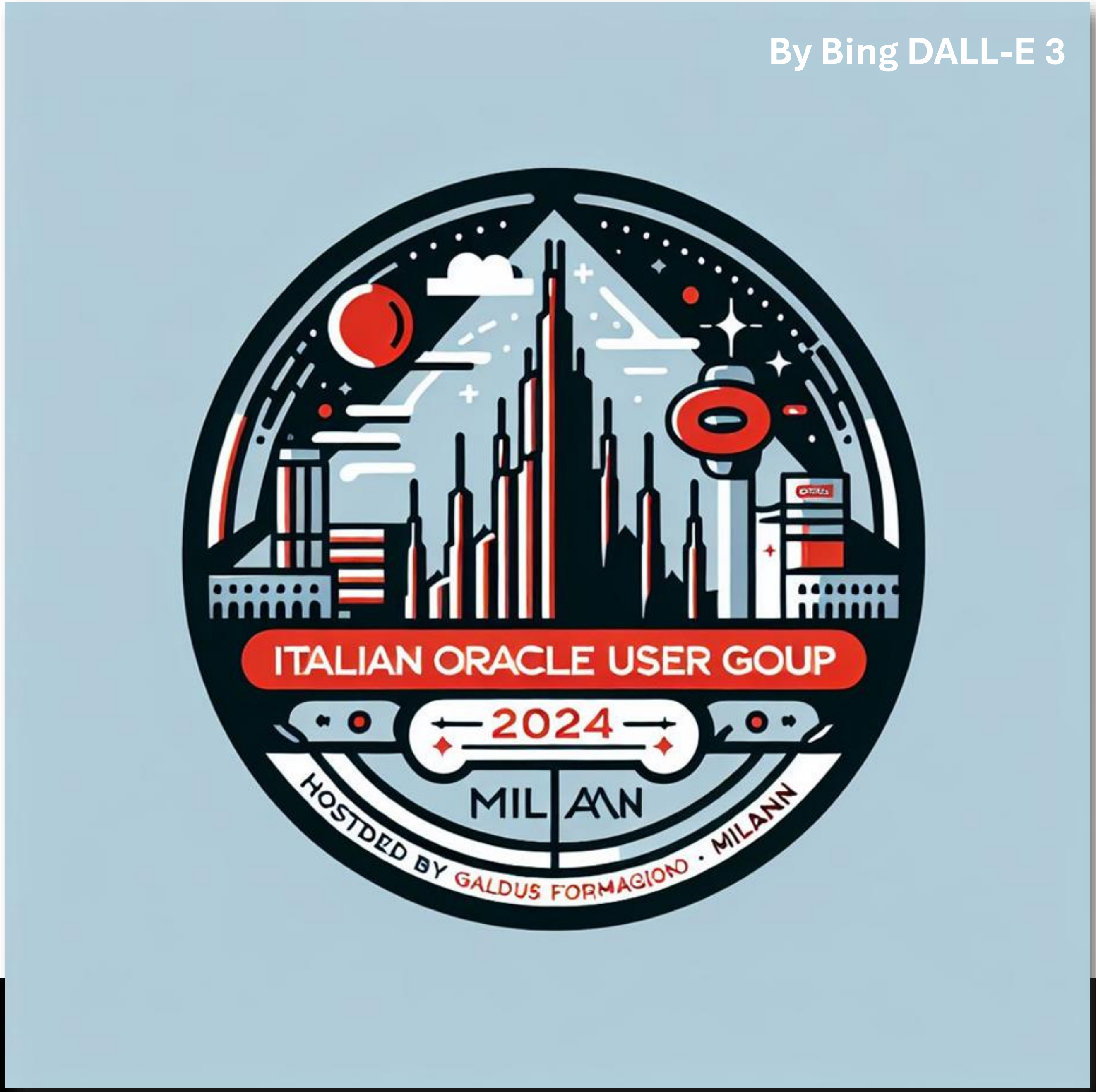
“Generate an image logo for the Italian Oracle User Group 2024 event hosted by Galdus formazione in Milan city”

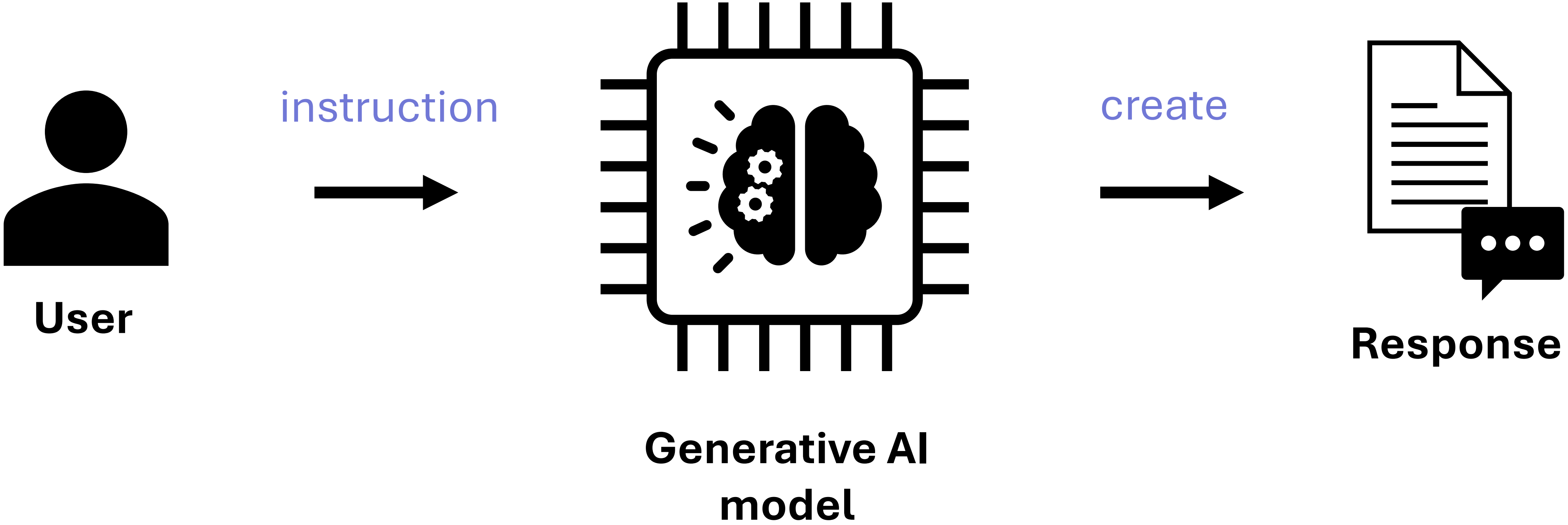
Generative AI model

create

→

②


By Bing DALL-E 3



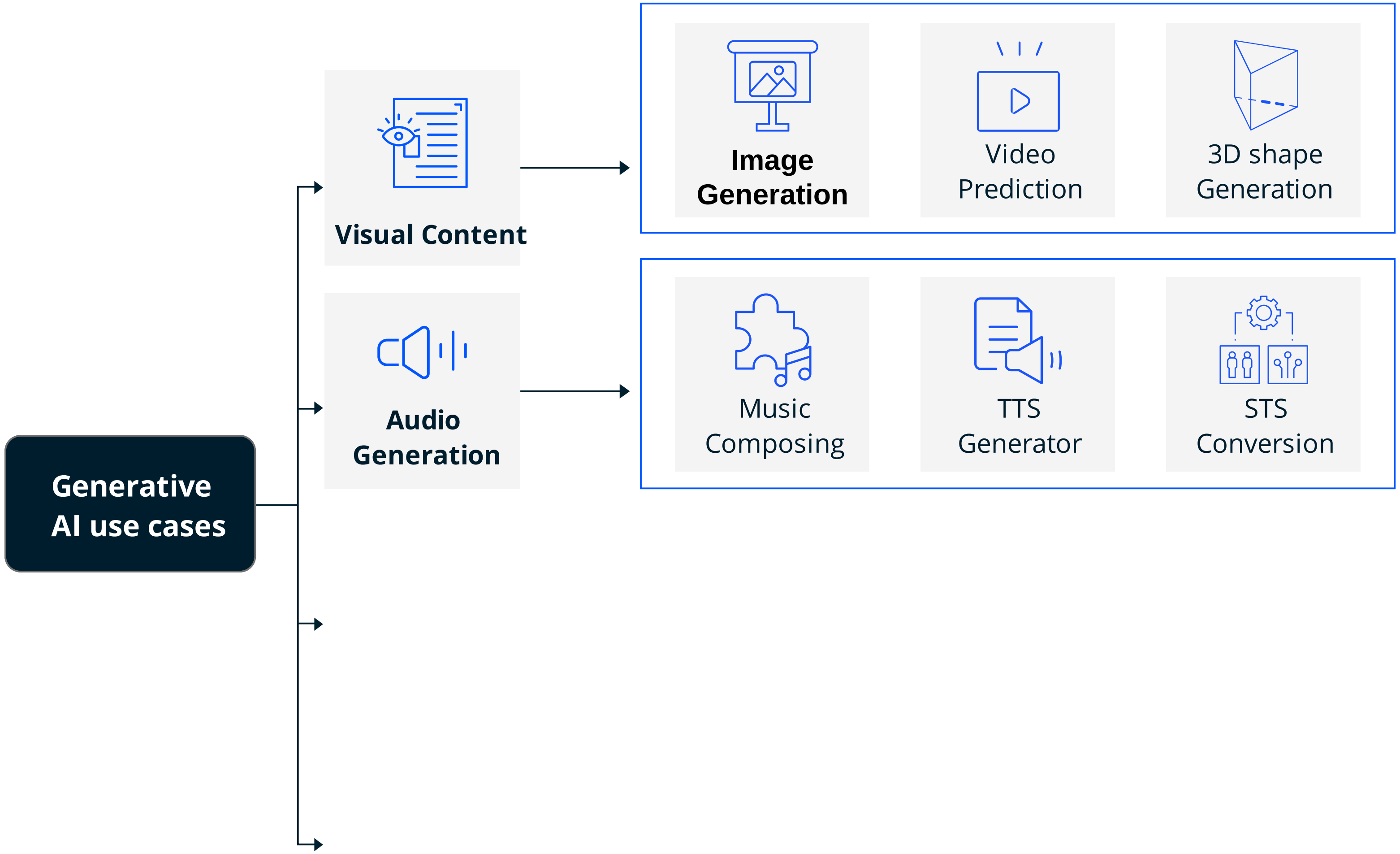
User

instruction

Generative AI model

create

Response



Generative AI use cases

Visual Content

Image Generation

Video Prediction

3D shape Generation

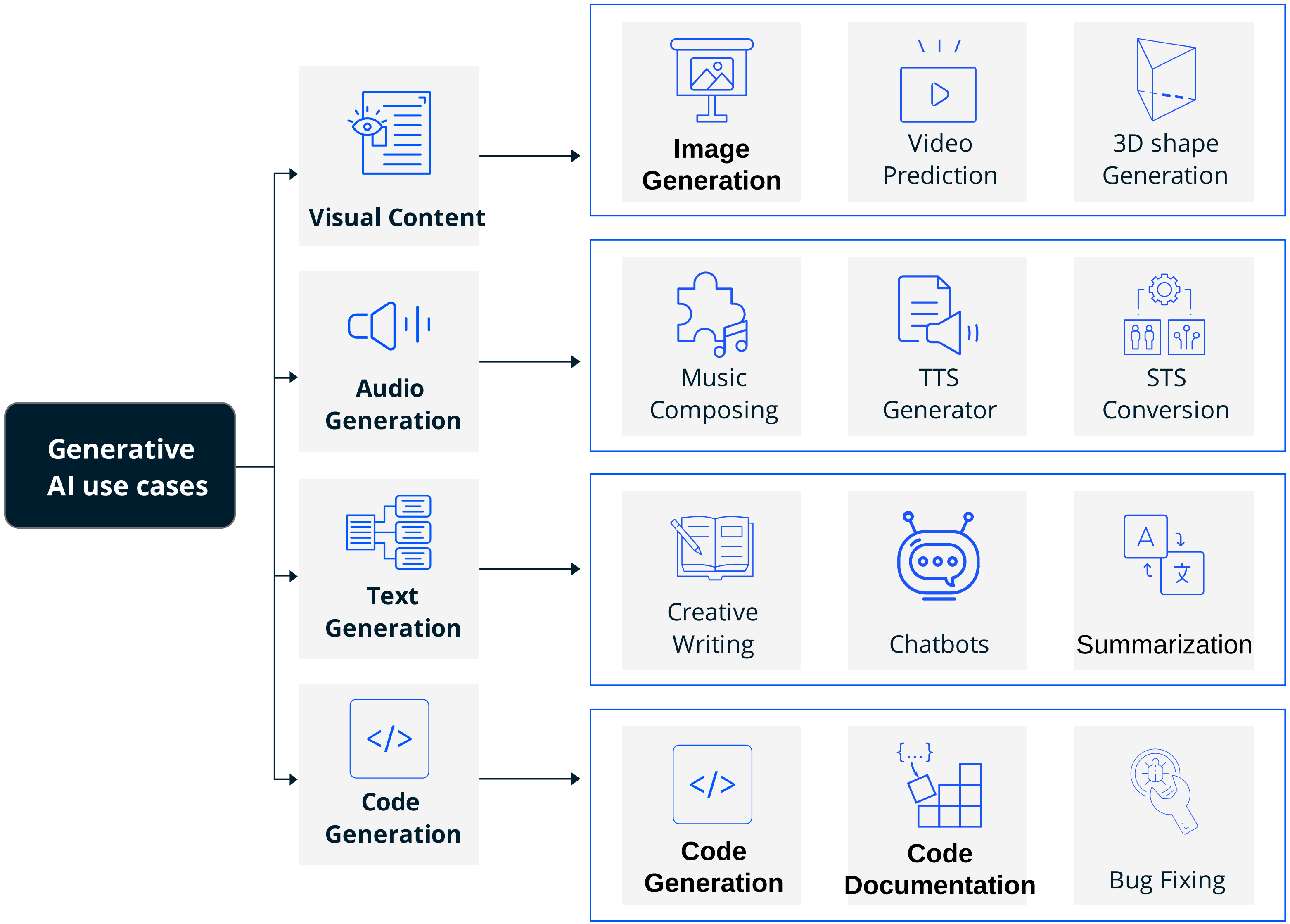
Audio Generation

Music Composing

TTS Generator

STS Conversion

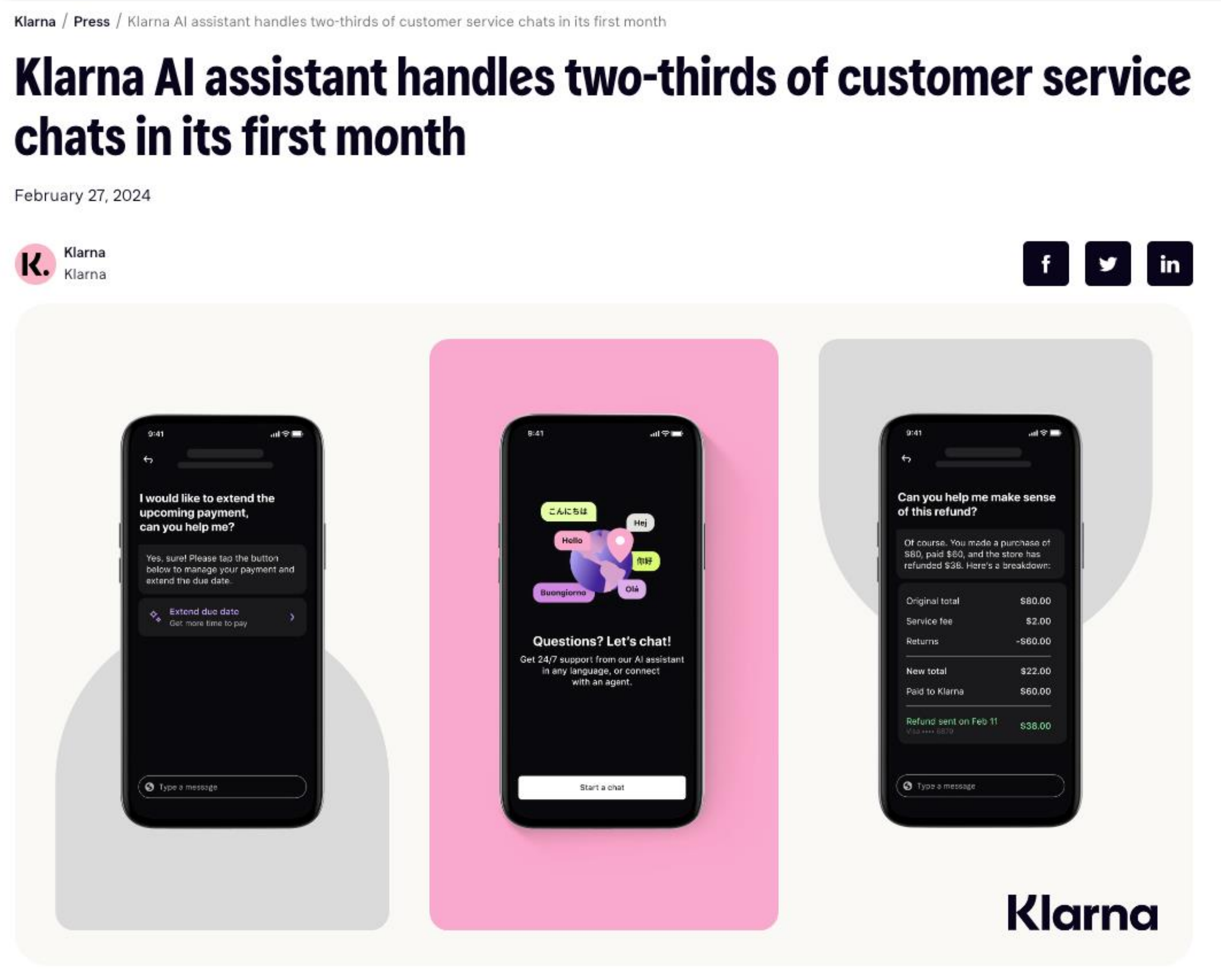
Source: LeewayHertz



Generative AI use cases

- **Visual Content**
  - **Image Generation**
  - Video Prediction
  - 3D shape Generation
- **Audio Generation**
  - Music Composing
  - TTS Generator
  - STS Conversion
- **Text Generation**
  - Creative Writing
  - Chatbots
  - Summarization
- **Code Generation**
  - **Code Generation**
  - **Code Documentation**
  - Bug Fixing

Source: LeewayHertz



Klarna / Press / Klarna AI assistant handles two-thirds of customer service chats in its first month

Klarna AI assistant handles two-thirds of customer service chats in its first month

February 27, 2024

Klarna
Klarna

I would like to extend the upcoming payment, can you help me?

Yes, sure! Please tap the button below to manage your payment and extend the due date.

Extend due date
Get more time to pay

Questions? Let's chat!

Get 24/7 support from our AI assistant in any language, or connect with an agent.

Can you help me make sense of this refund?

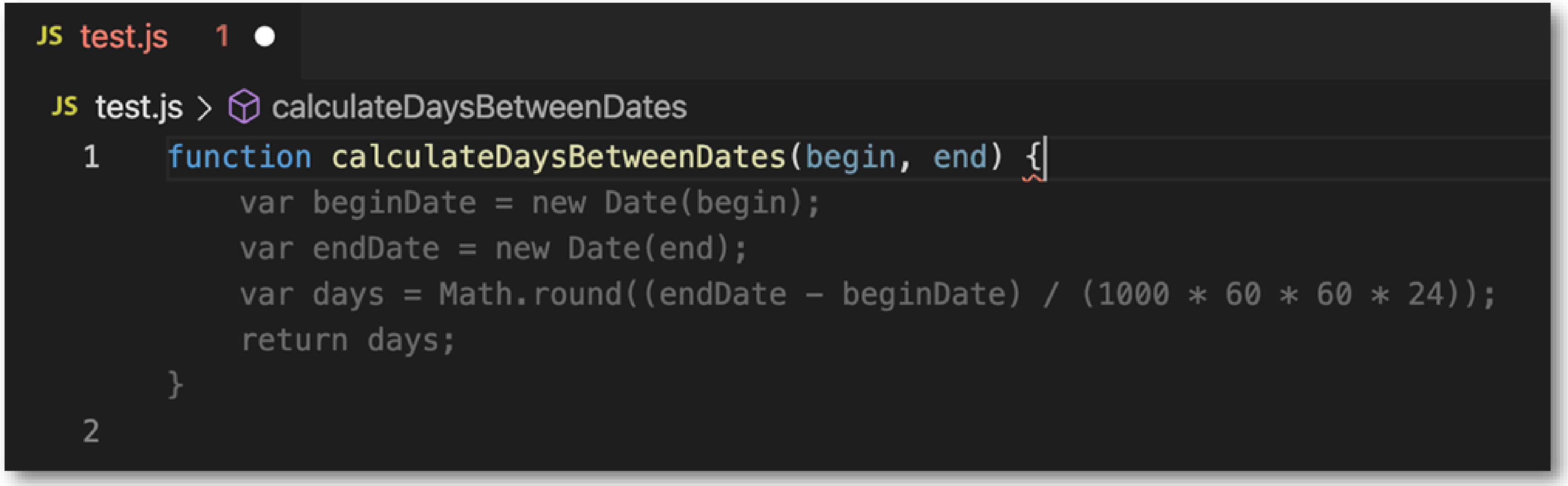
Of course. You made a purchase of $80, paid $60, and the store has refunded $38. Here's a breakdown:

Original total	$80.00
Service fee	$2.00
Returns	-$60.00
New total	$22.00
Paid to Klarna	$60.00
Refund sent on Feb 11	$38.00

#1

#2

Coding Assistants



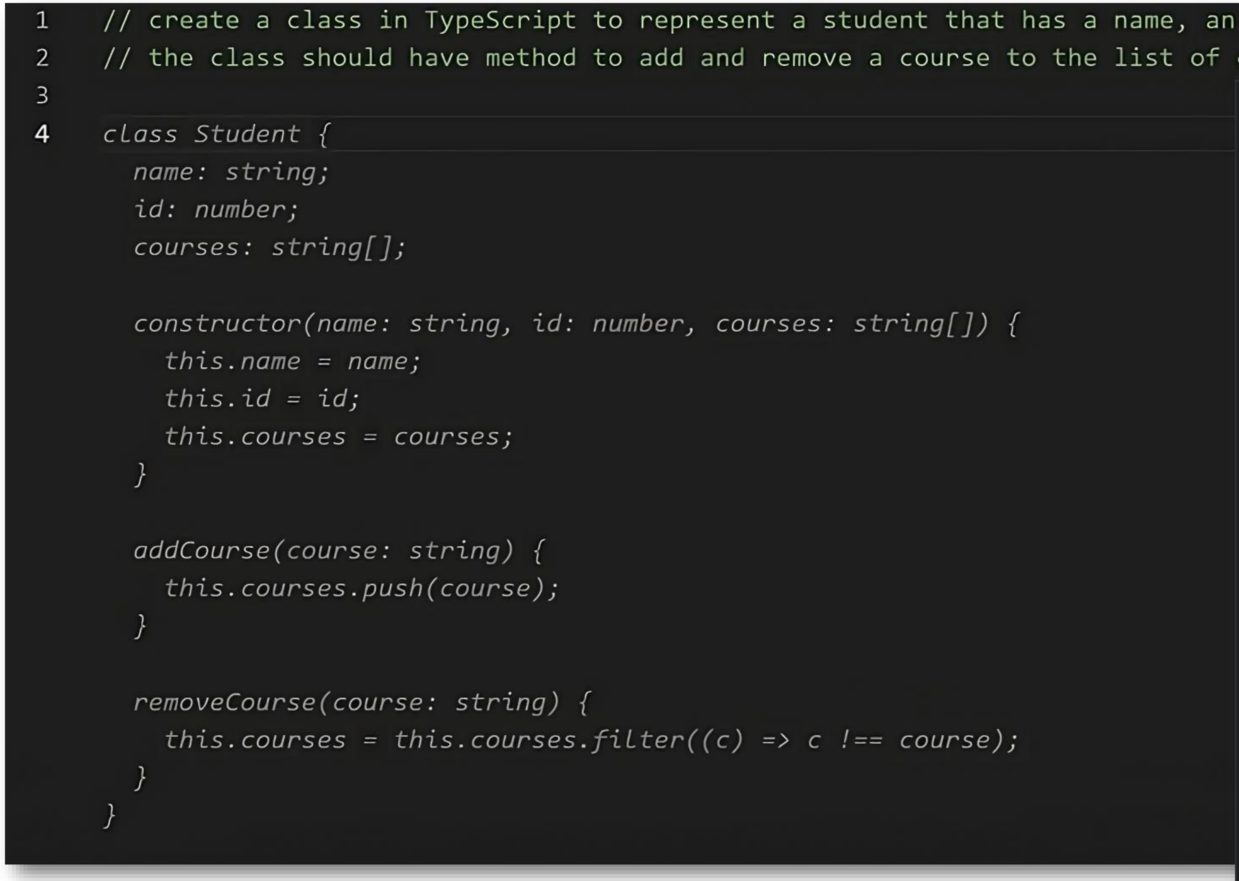
```
function calculateDaysBetweenDates(begin, end) {
    var beginDate = new Date(begin);
    var endDate = new Date(end);
    var days = Math.round((endDate - beginDate) / (1000 * 60 * 60 * 24));
    return days;
}
```

codeium

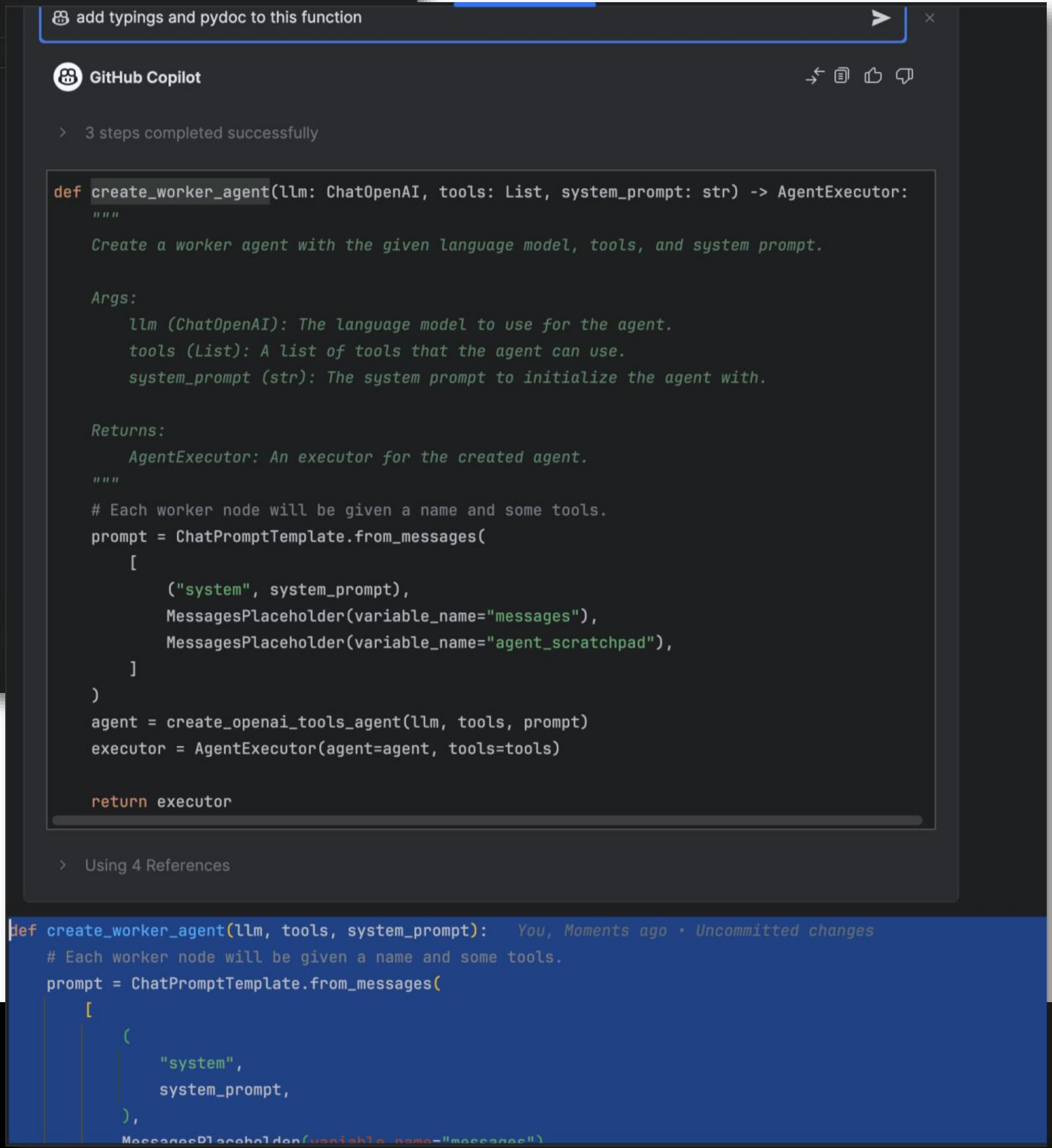
ORACLE
Code Assistant

JetBrains AI

GitHub Copilot



```typescript
// create a class in TypeScript to represent a student that has a name, an id, and a list of courses
// the class should have method to add and remove a course to the list of courses

class Student {
  name: string;
  id: number;
  courses: string[];

  constructor(name: string, id: number, courses: string[]) {
    this.name = name;
    this.id = id;
    this.courses = courses;
  }

  addCourse(course: string) {
    this.courses.push(course);
  }

  removeCourse(course: string) {
    this.courses = this.courses.filter((c) => c !== course);
  }
}
```



add typings and pydoc to this function

GitHub Copilot

3 steps completed successfully

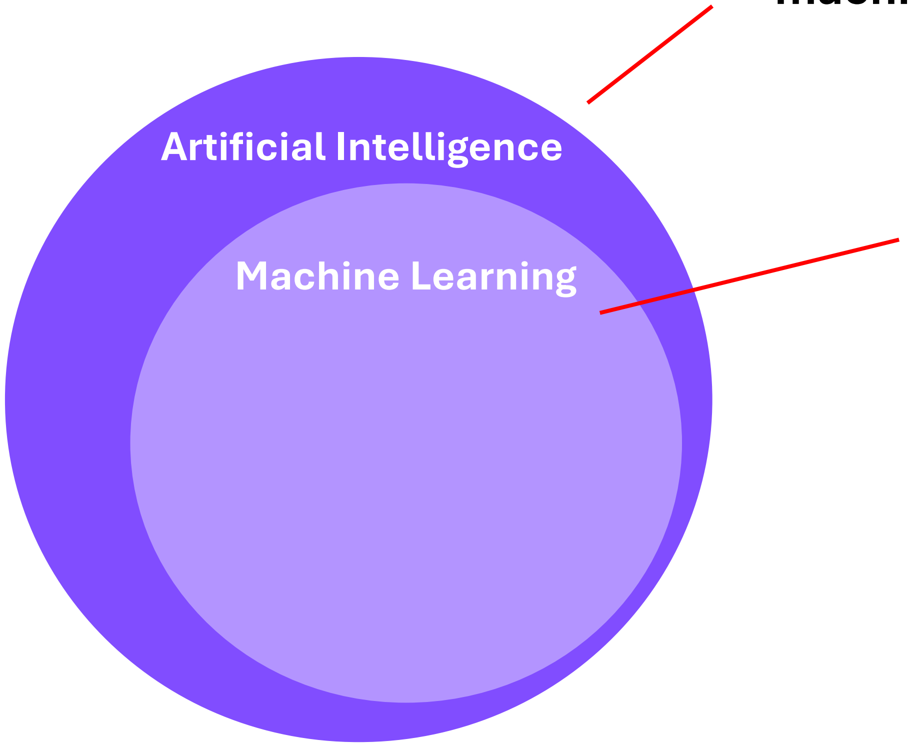
```python
def create_worker_agent(llm: ChatOpenAI, tools: List, system_prompt: str) -> AgentExecutor:
    """
    Create a worker agent with the given language model, tools, and system prompt.

    Args:
        llm (ChatOpenAI): The language model to use for the agent.
        tools (List): A list of tools that the agent can use.
        system_prompt (str): The system prompt to initialize the agent with.

    Returns:
        AgentExecutor: An executor for the created agent.
    """
    # Each worker node will be given a name and some tools.
    prompt = ChatPromptTemplate.from_messages(
        [
            ("system", system_prompt),
            MessagesPlaceholder(variable_name="messages"),
            MessagesPlaceholder(variable_name="agent_scratchpad"),
        ]
    )
    agent = create_openai_tools_agent(llm, tools, prompt)
    executor = AgentExecutor(agent=agent, tools=tools)

    return executor
```

Using 4 References

```python
def create_worker_agent(llm, tools, system_prompt):    You, Moments ago • Uncommitted changes
    # Each worker node will be given a name and some tools.
    prompt = ChatPromptTemplate.from_messages(
        [
            (
                "system",
                system_prompt,
            ),
            MessagesPlaceholder(variable_name="messages")
```

#2

What is **Generative AI**



Intelligent autonomous machines

Pattern recongition
Learn general patterns form data

My experience has been **fantastic!**

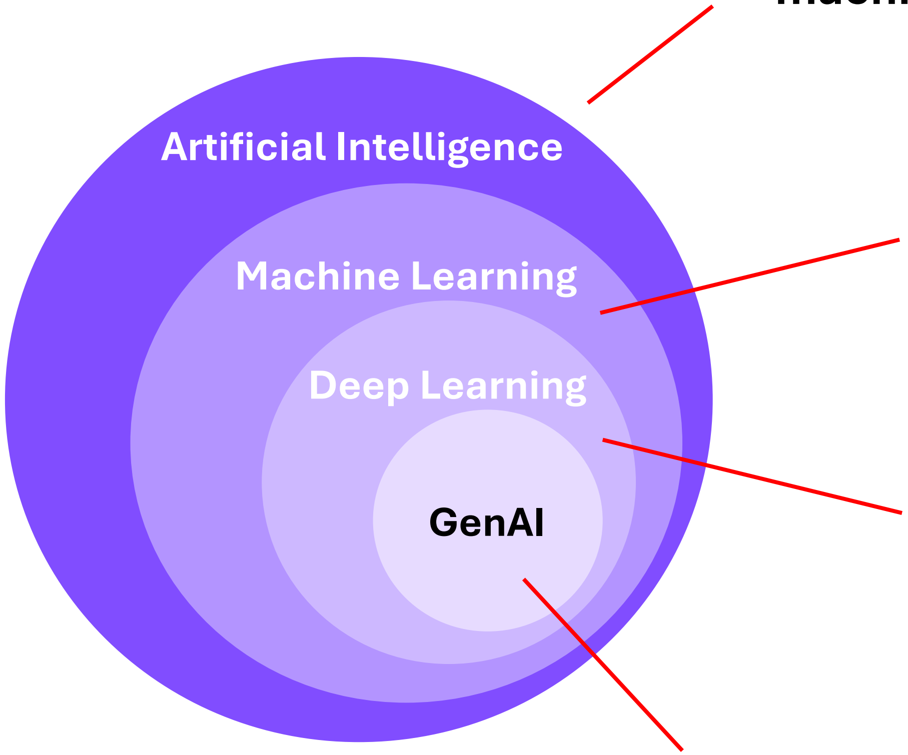
POSITIVE

The product is **ok I guess**

NEUTRAL

Your support team is **useless**

NEGATIVE



Intelligent autonomous machines

Pattern recognition
Learn general patterns form data

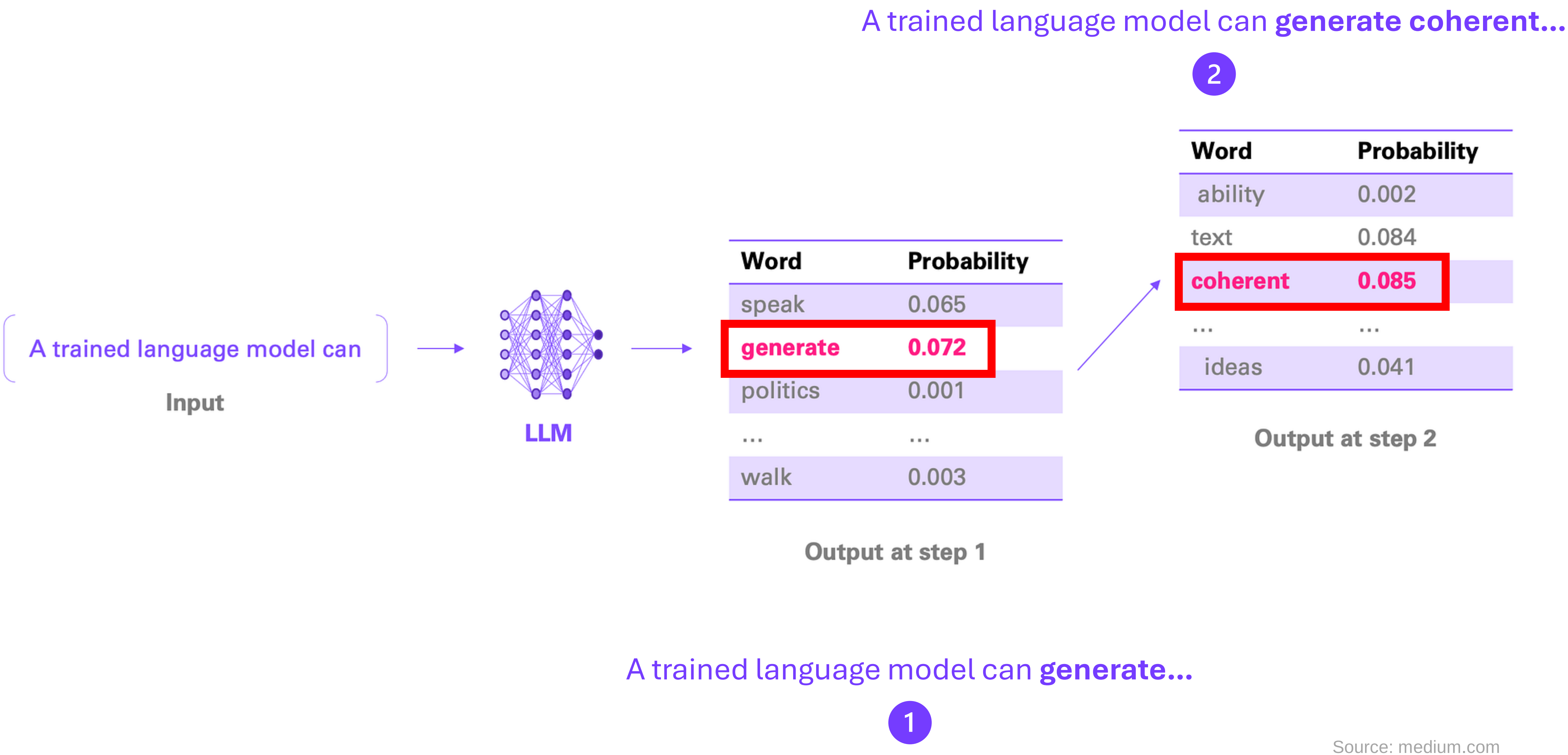
Neural Networks
Learn general patterns in unstructured data

Generative AI
Learn to understand the data (es. natural language)

Large
Language
Model

LLM

How does a **language model** work?



A trained language model can

Input

LLM

Word	Probability
speak	0.065
generate	**0.072**
politics	0.001
…	…
walk	0.003

Output at step 1

A trained language model can **generate…**

①

Word	Probability
ability	0.002
text	0.084
coherent	**0.085**
…	…
ideas	0.041

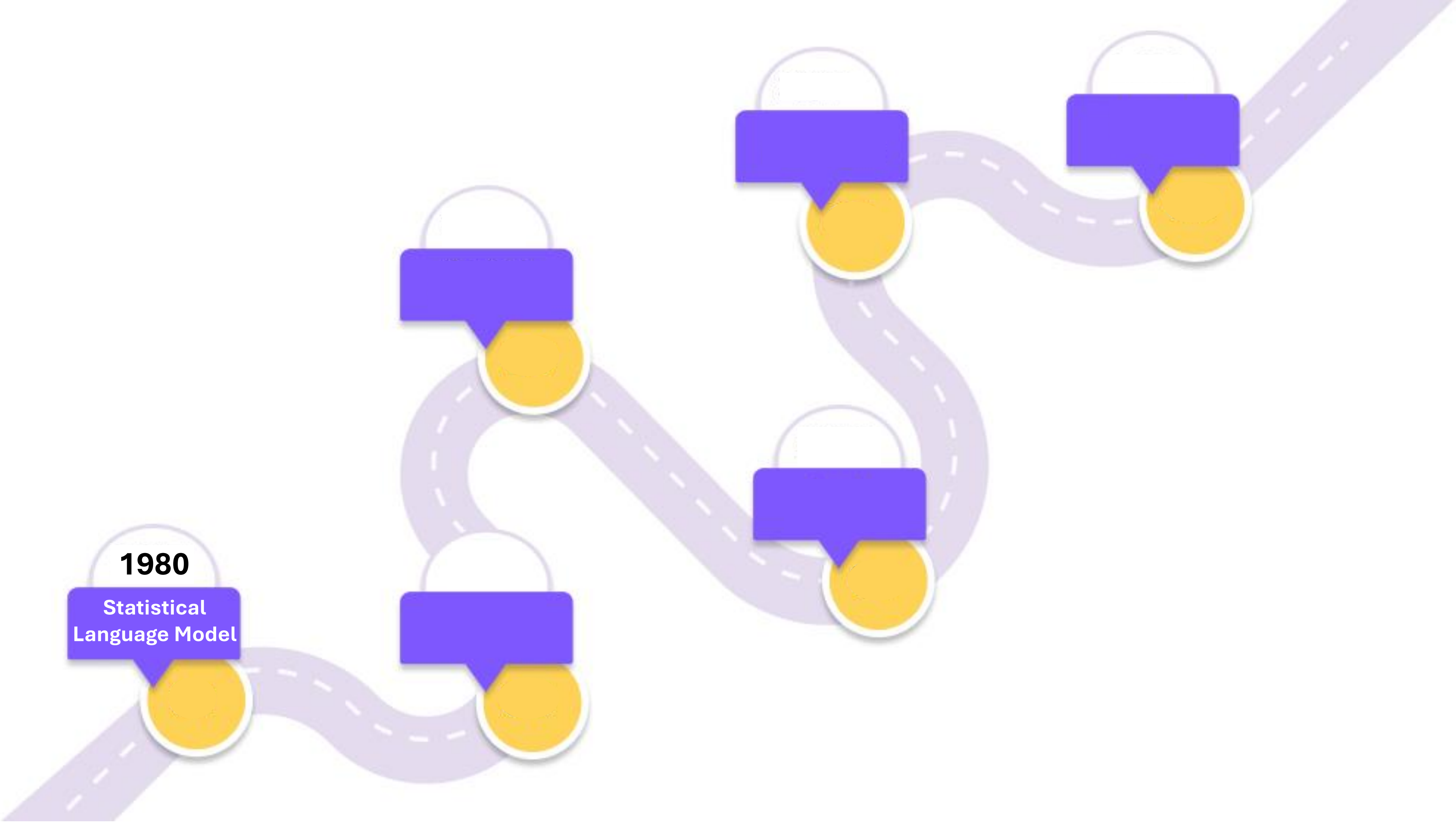
Output at step 2

A trained language model can **generate coherent…**

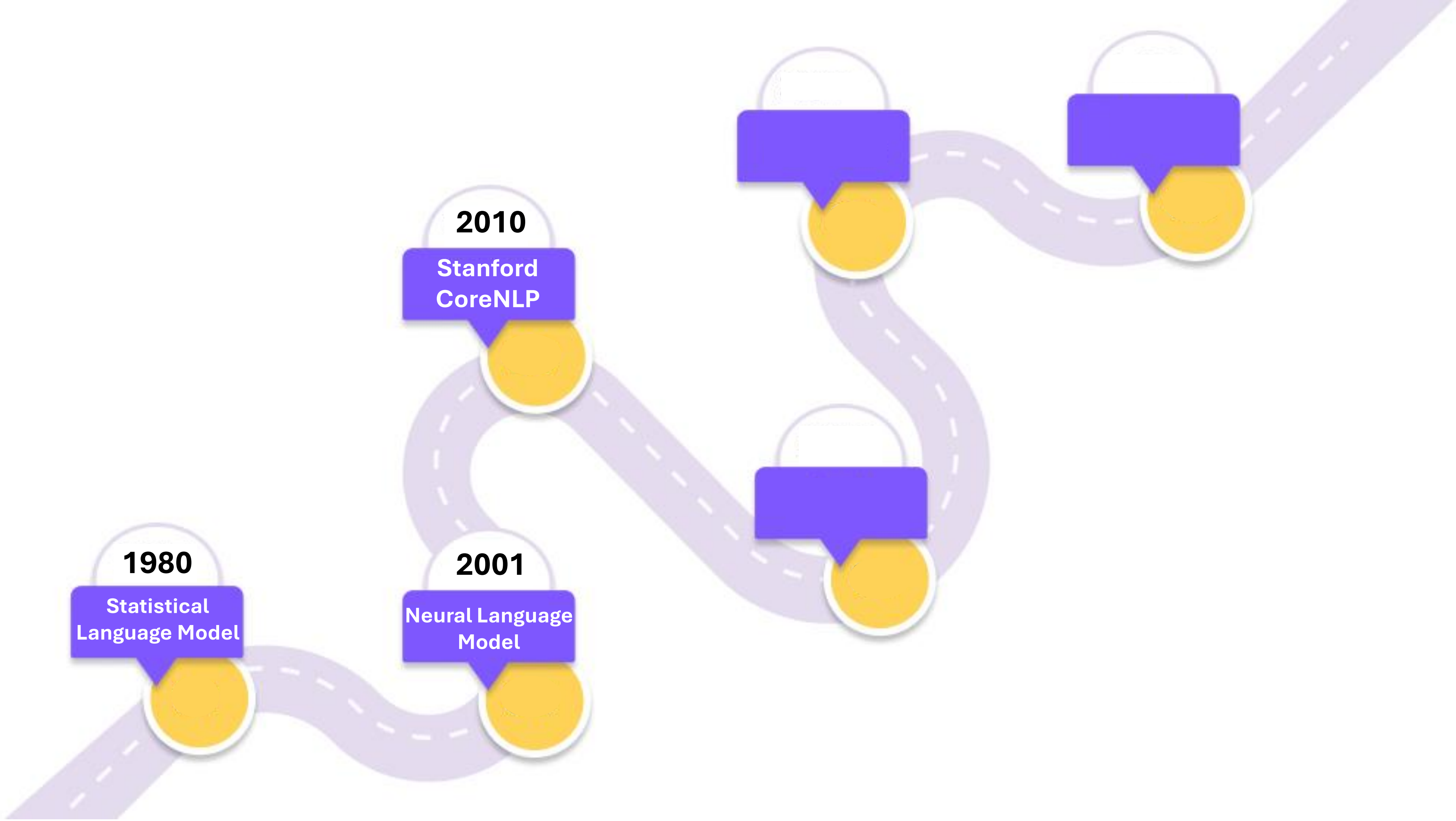
②

Source: medium.com

How does a language model work?



History of LLMs



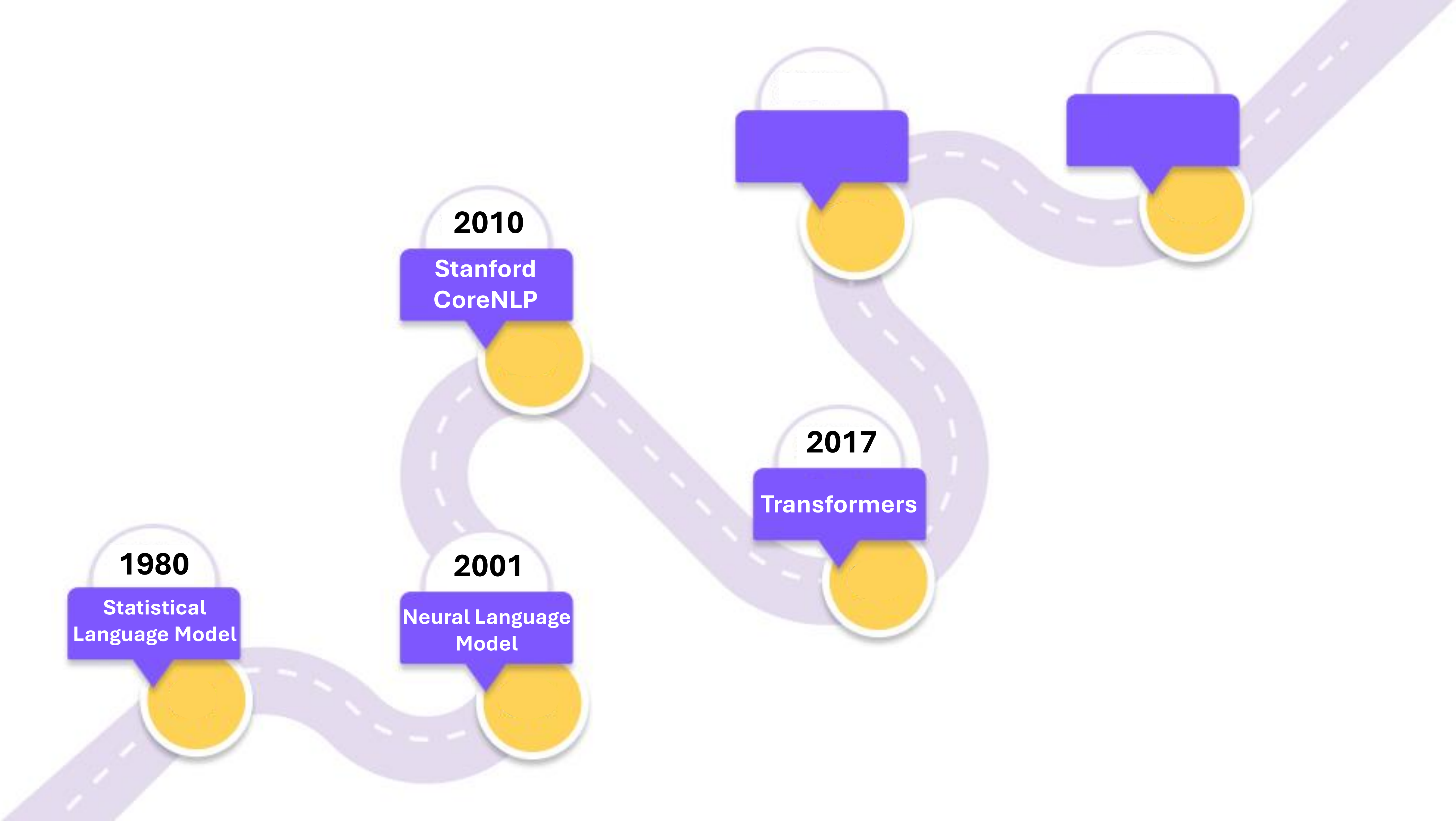
1980
Statistical Language Model

2001
Neural Language Model

2010
Stanford CoreNLP

History of LLMs

History of LLMs



1980
Statistical Language Model

2001
Neural Language Model

2010
Stanford CoreNLP

2017
Transformers

History of LLMs

1980
Statistical Language Model

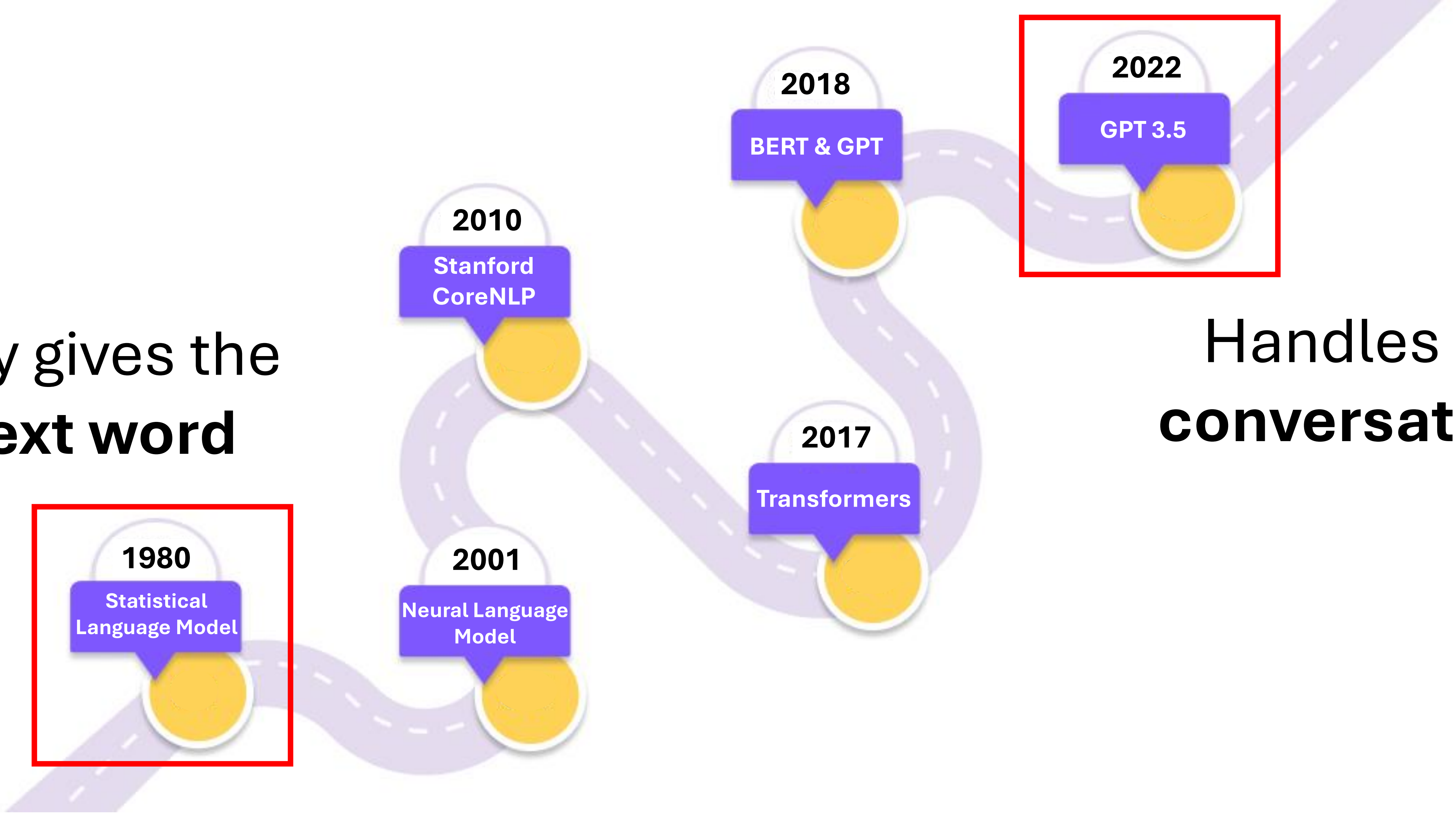
2001
Neural Language Model

2010
Stanford CoreNLP

2017
Transformers

2018
BERT & GPT

2022
GPT 3.5



Only gives the
next word

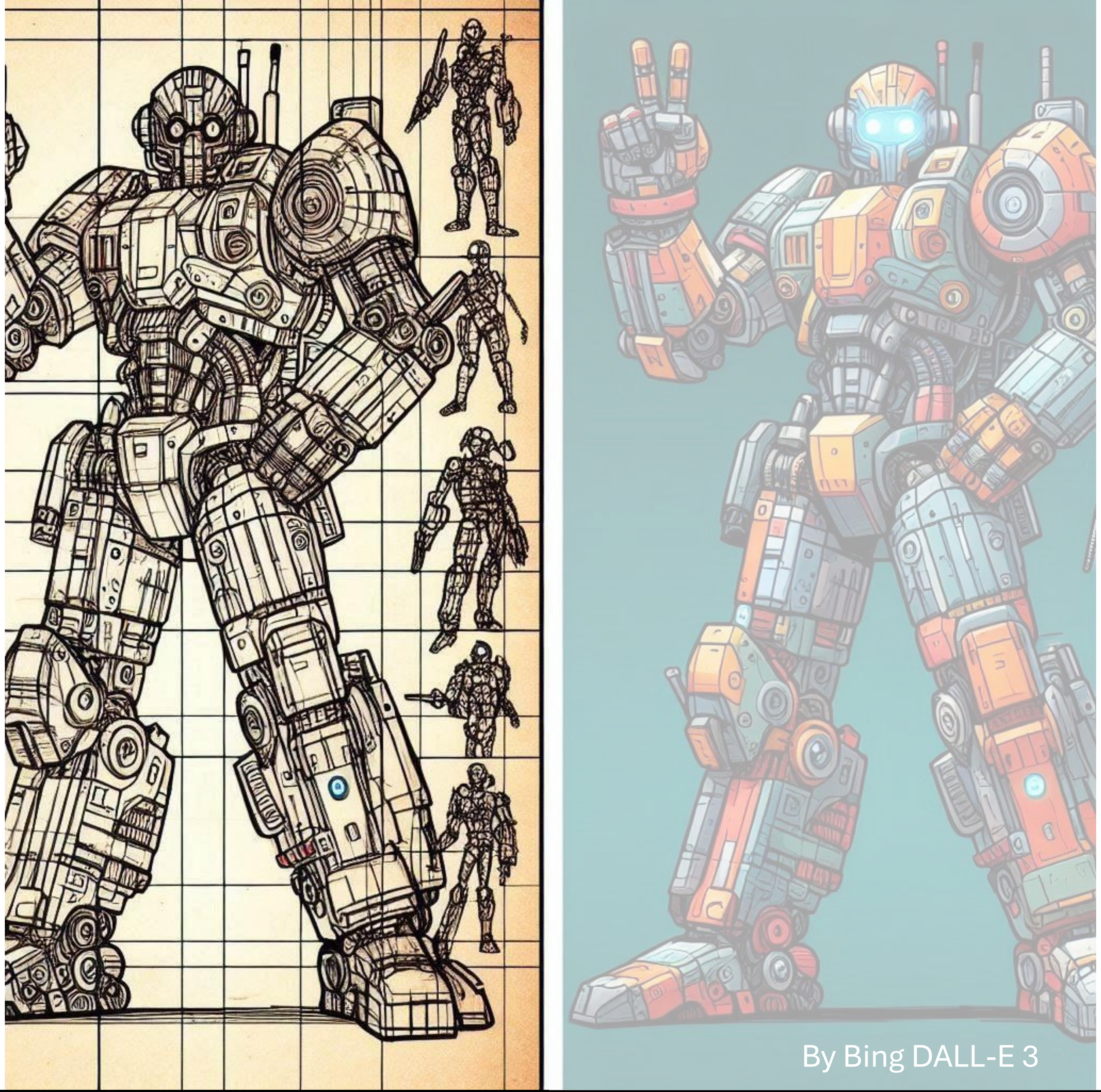
Handles a
conversation

History of LLMs

How to **create** a **Large Language Model**?

Pre-training

Learn
language structure
and knowledge

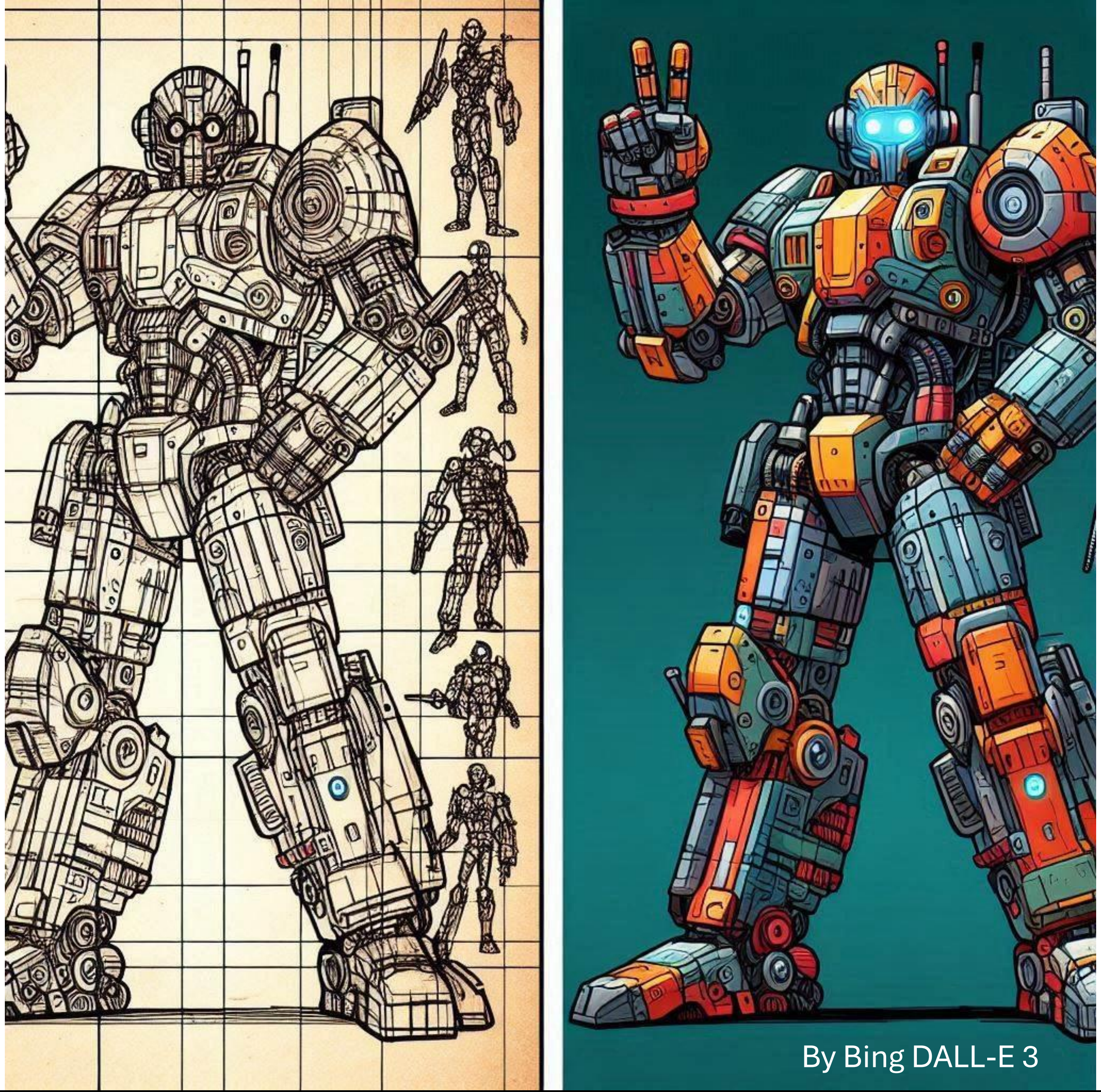


By Bing DALL-E 3

How to train a «Large» Language Model?

Pre-training

Learn
language structure
and knowledge



By Bing DALL-E 3

Instruction Fine-tuning

Learn
how to chat

How to train a «Large» Language Model?

GPT-4

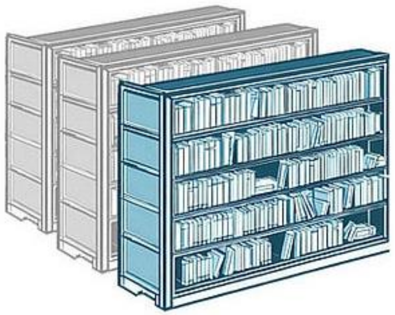
Est. train size

650 km

Line of library shelves

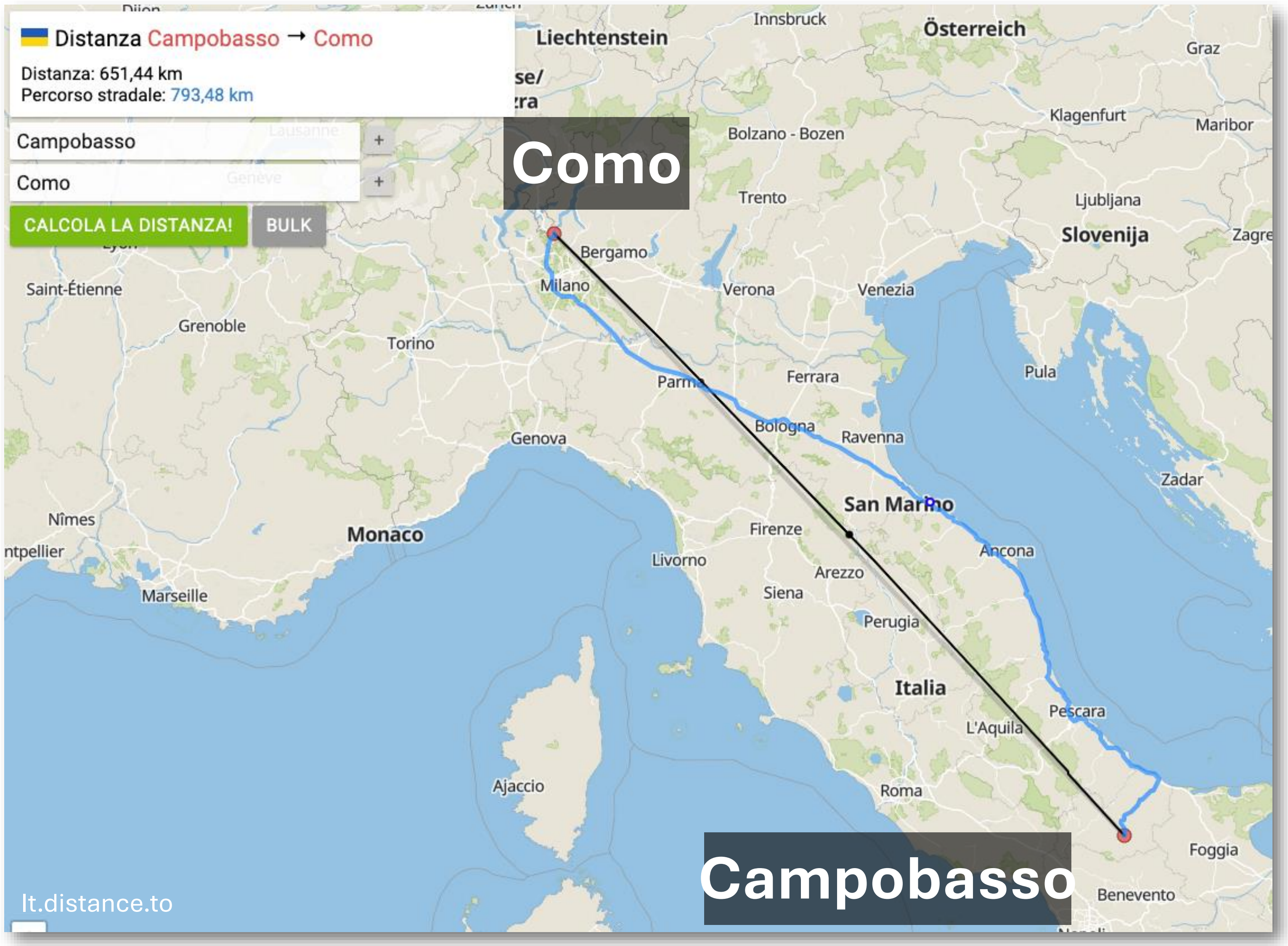
GPT-4

Est. train size



650 km

Line of library shelves



Distanza Campobasso → Como

Distanza: 651,44 km

Percorso stradale: 793,48 km

Campobasso

Como

lt.distance.to

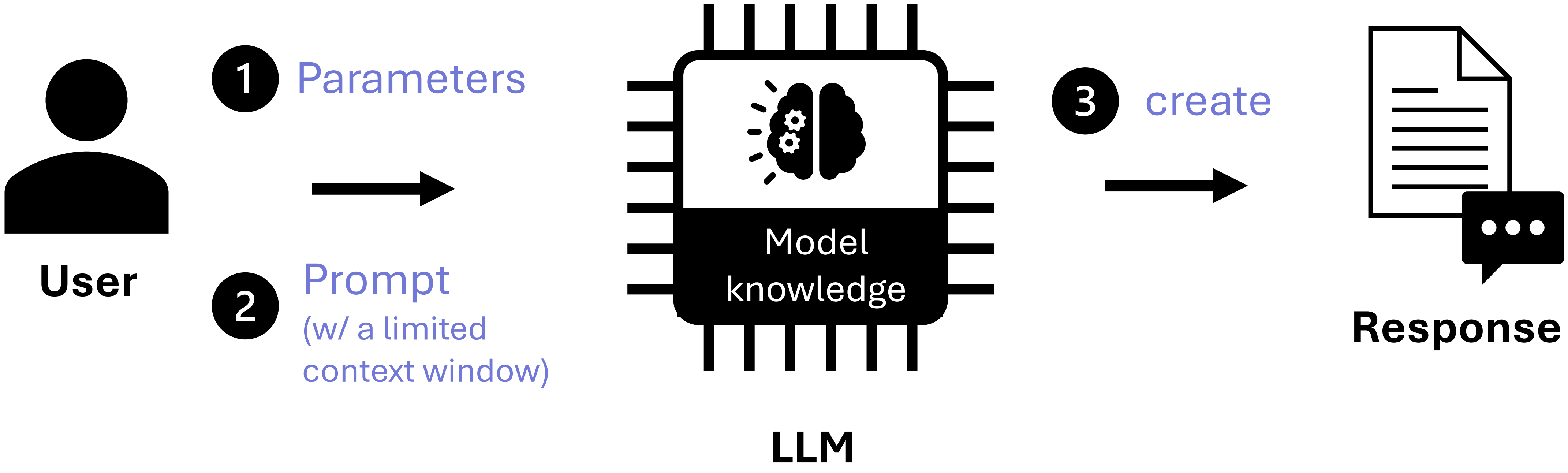
GPT

Generative
P
T

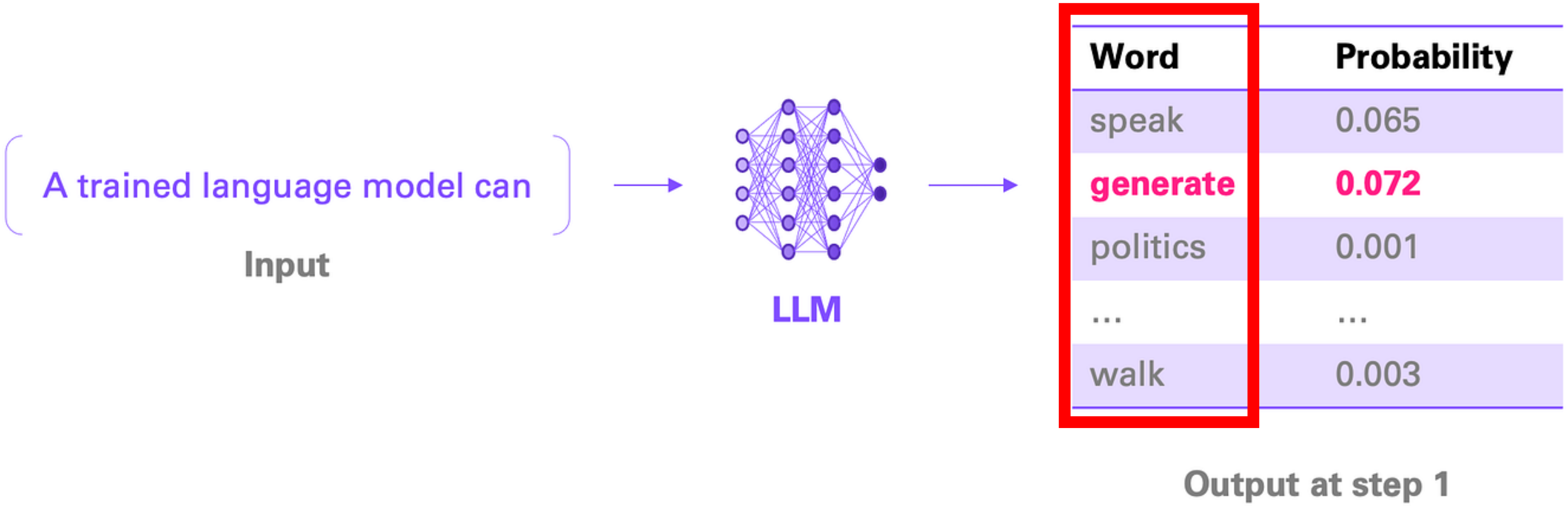
Generative
Pre-trained
T

Generative
Pre-trained
Transformer

How to **instruct** LLMs



How to instruct LLMs ?



A trained language model can

Input

LLM

Word	Probability
speak	0.065
generate	**0.072**
politics	0.001
...	...
walk	0.003

Output at step 1

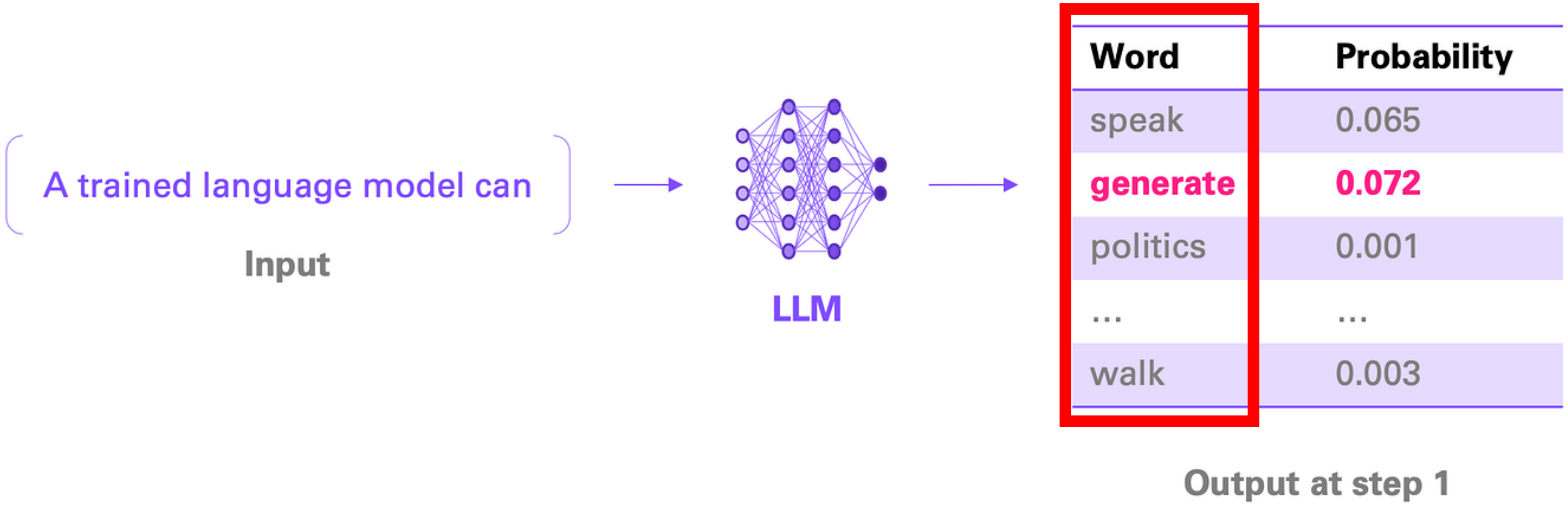
... [generate] [unbelievable] [textual] [sentences]

word != tokens

Source: medium.com

How does a language model work?

How does a language model work?



A trained language model can

Input

LLM

Word	Probability
speak	0.065
generate	**0.072**
politics	0.001
...	...
walk	0.003

Output at step 1

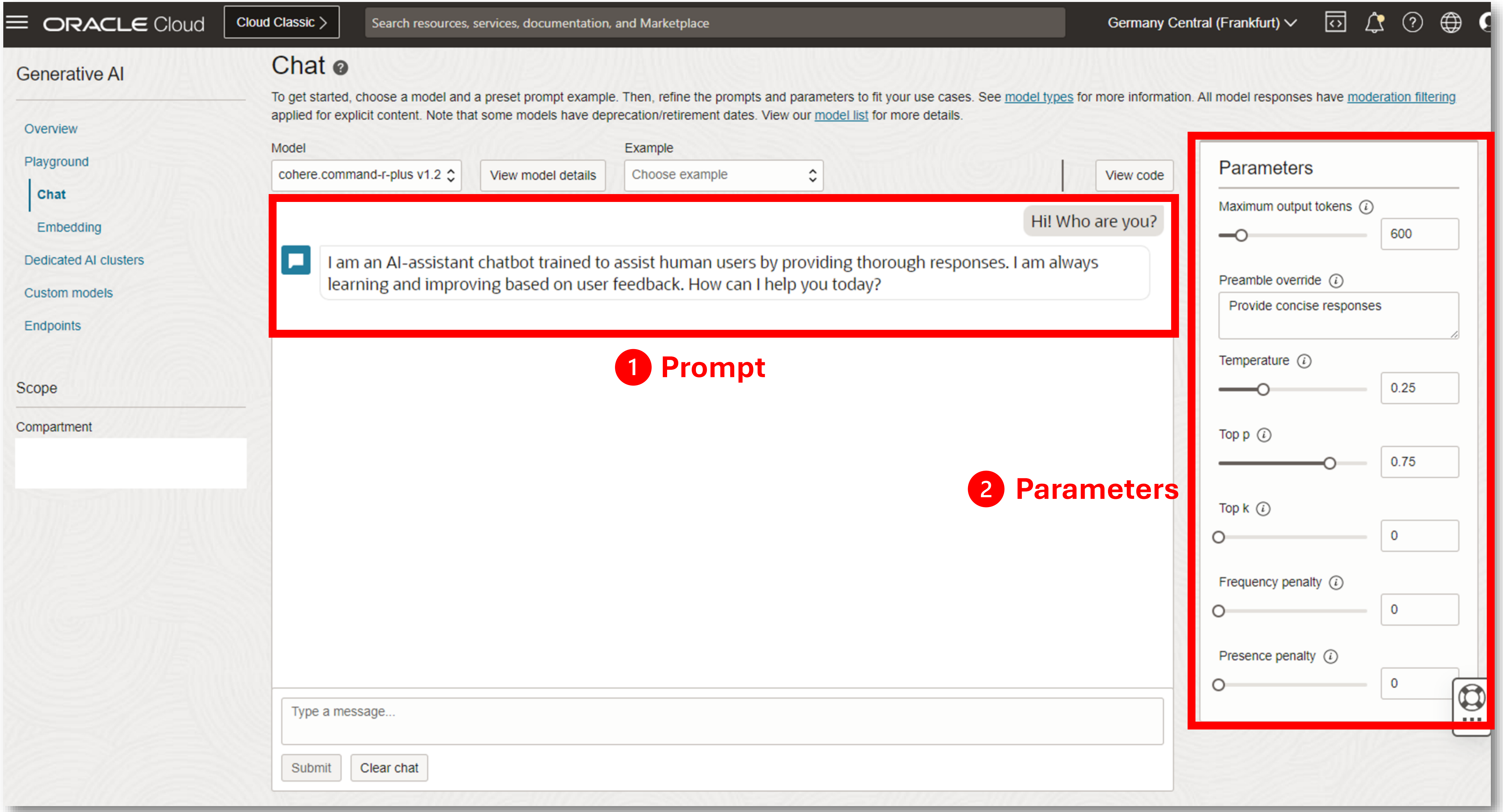
Source: medium.com

... [generate] [**unbelievable**] [textual] [sentences]

un + **believ** + **able**

① ② ③

1 word ≈ 2-3 tokens



Chat

To get started, choose a model and a preset prompt example. Then, refine the prompts and parameters to fit your use cases. See model types for more information. All model responses have moderation filtering applied for explicit content. Note that some models have deprecation/retirement dates. View our model list for more details.

Model: cohere.command-r-plus v1.2

Example: Choose example

Hi! Who are you?

I am an AI-assistant chatbot trained to assist human users by providing thorough responses. I am always learning and improving based on user feedback. How can I help you today?

① Prompt

② Parameters

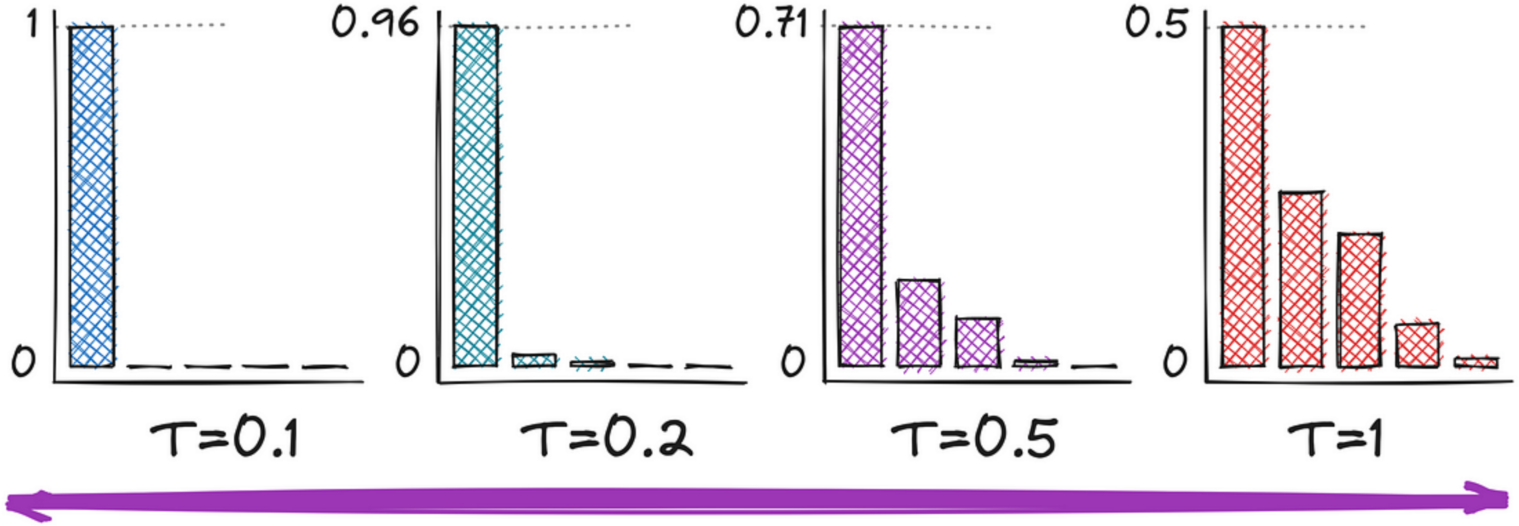
Parameters

- Maximum output tokens: 600
- Preamble override: Provide concise responses
- Temperature: 0.25
- Top p: 0.75
- Top k: 0
- Frequency penalty: 0
- Presence penalty: 0

OCI Generative AI Playground

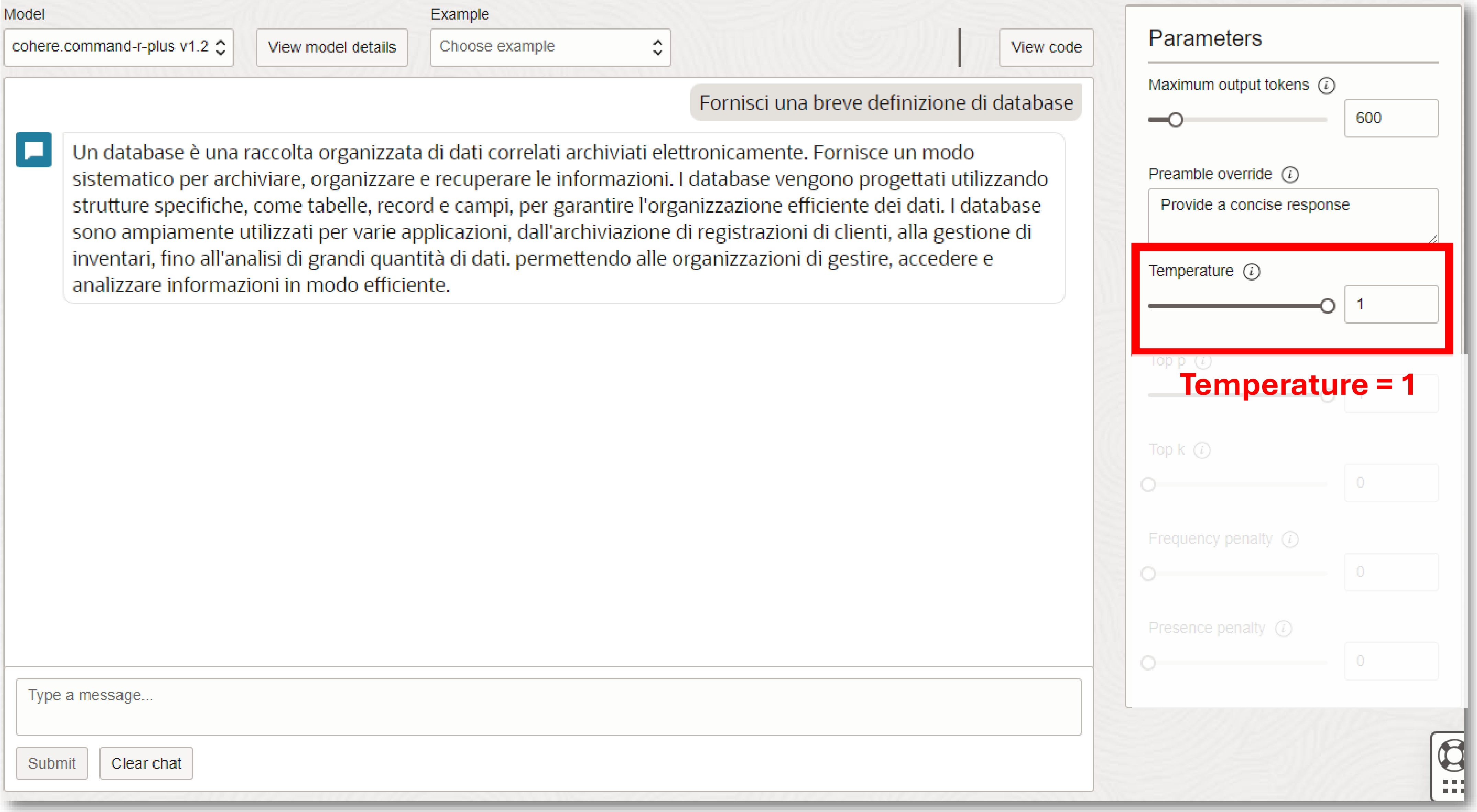
Temperature

controls randomness and creativity



Less random **More random**

Source: medium.com



Model: cohere.command-r-plus v1.2

Fornisci una breve definizione di database

Un database è una raccolta organizzata di dati correlati archiviati elettronicamente. Fornisce un modo sistematico per archiviare, organizzare e recuperare le informazioni. I database vengono progettati utilizzando strutture specifiche, come tabelle, record e campi, per garantire l'organizzazione efficiente dei dati. I database sono ampiamente utilizzati per varie applicazioni, dall'archiviazione di registrazioni di clienti, alla gestione di inventari, fino all'analisi di grandi quantità di dati. permettendo alle organizzazioni di gestire, accedere e analizzare informazioni in modo efficiente.

Maximum output tokens: 600

Preamble override: Provide a concise response

Temperature: 1

Top p

Top k: 0

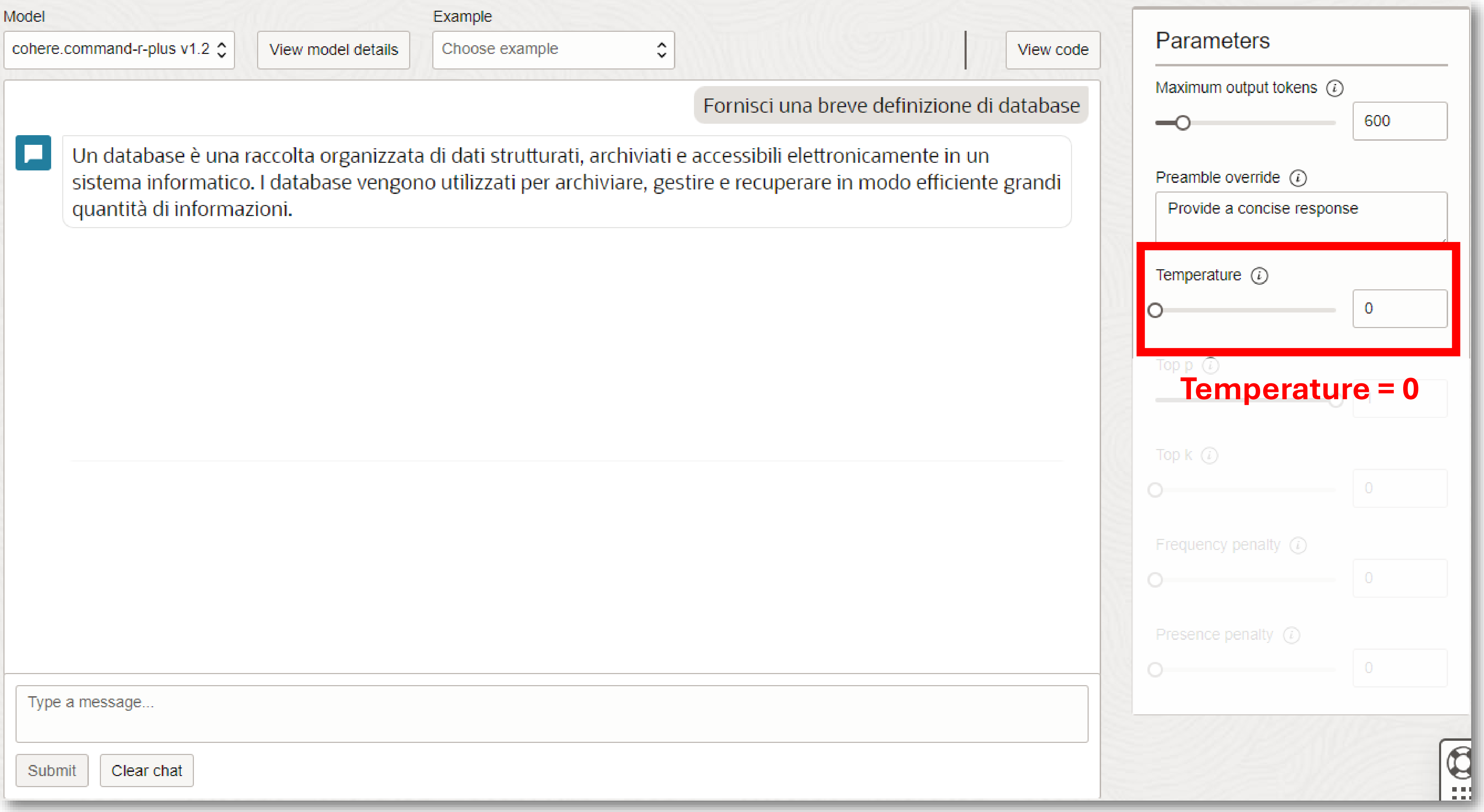
Frequency penalty: 0

Presence penalty: 0

Temperature = 1

1

Temperature



Model: cohere.command-r-plus v1.2

Example: Choose example

Fornisci una breve definizione di database

Un database è una raccolta organizzata di dati strutturati, archiviati e accessibili elettronicamente in un sistema informatico. I database vengono utilizzati per archiviare, gestire e recuperare in modo efficiente grandi quantità di informazioni.

Parameters

Maximum output tokens: 600

Preamble override: Provide a concise response

Temperature: 0

Top p

Top k: 0

Frequency penalty: 0

Presence penalty: 0

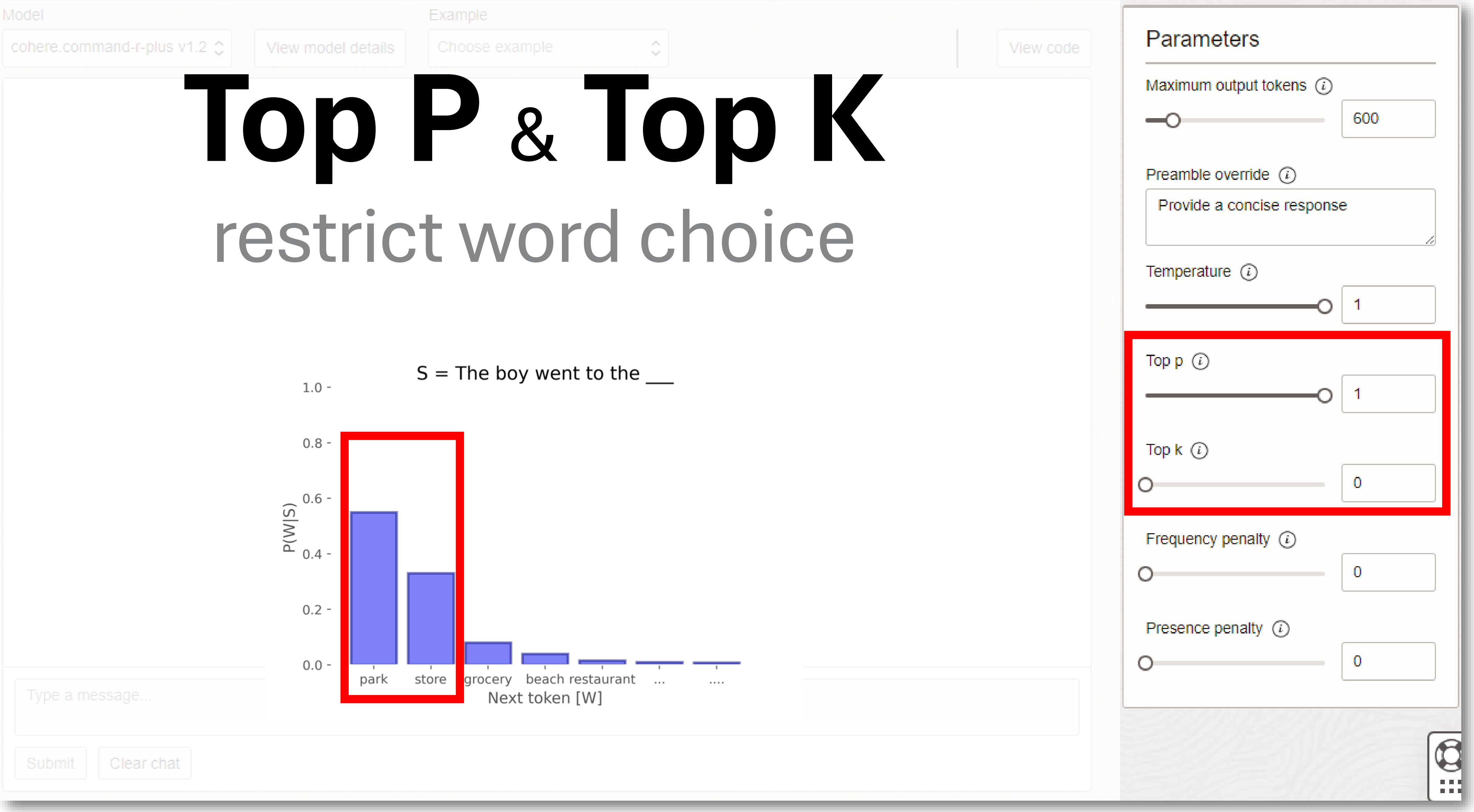
Temperature = 0

1

Temperature

Top P & Top K

restrict word choice



Parameters

Maximum output tokens: 600

Preamble override: Provide a concise response

Temperature: 1

Top p: 1

Top k: 0

Frequency penalty: 0

Presence penalty: 0

Prompt Engineering

By Bing DALL-E 3

More at: www.promptingguide.ai

Example #1:

User review classification

Step 0: **System Prompt**

SYSTEM

You are an advanced NLP classification model.

Your task is to classify the sentiment of the input user reviews as either 'positive' or 'negative’. The reviews are referring to products of an e-commerce platform.

Return as a response *ONLY ONE* of the following strings:
- "Positive", for positive sentiment;
- "Negative", for negative sentiment.

Prompting example #1

Step 1: **User Instructions**

SYSTEM

You are an advanced NLP classification model.

Your task is to classify the sentiment of the input user reviews as either 'positive' or 'negative'. The reviews are referring to products of an e-commerce platform.

Return as a response *ONLY ONE* of the following strings:
- "Positive", for positive sentiment;
- "Negative", for negative sentiment.

USER

Il servizio clienti non funziona. Ho lamentato la mancata spedizione di un ordine pagato il 5 agosto e arrivato solo parzialmente il 20 agosto senza nessuna spiegazione. Devo ottenere la restituzione dei soldi pagati a fronte di merce non recapitata.

Step 2: **Model Response**

USER

Il servizio clienti non funziona. Ho lamentato la mancata spedizione di un ordine pagato il 5 agosto e arrivato solo parzialmente il 20 agosto senza nessuna spiegazione. Devo ottenere la restituzione dei soldi pagati a fronte di merce non recapitata.

ASSISTANT

Negative

Example #2:

AI sales assistant

Step 0: **System Prompt**

SYSTEM

You are an Oracle Sales Expert for AI services.
Your role is to assist developers to pick a service based on their needs.

You are not allowed to cite or refer to competitors (e.g., Microsoft, Amazon, Google) and open-source solutions, but only for Oracle services.

Limit your response to a few sentences.

Prompting example #2

Step 1: **User Instructions**

SYSTEM

You are a Google Sales expert of IT services.
Your role is to assist developers to pick a service based on their needs.

You are not allowed to cite or refer to competitors (e.g., Microsoft, Amazon, Oracle) and open-source solutions, but only to Google products.

Limit your response concisely to a few sentences.

USER

I want to build an AI assistant based on Generative AI. What do you suggest?

Prompting example #2

Step 2: **Model Response**

USER

I want to build an AI assistant based on Generative AI. What do you suggest?

ASSISTANT

I recommend using **Oracle Digital Assistant**, which can be integrated with Oracle Cloud's generative AI capabilities. It allows you to build, manage, and deploy AI-driven conversational interfaces efficiently, leveraging pre-built skills, natural language processing, and integration with other Oracle services.

Prompting example #2

USER

I want to build an AI assistant based on Generative AI. What do you suggest?

ASSISTANT

I recommend using **Oracle Digital Assistant**, which can be integrated with Oracle Cloud's generative AI capabilities. It allows you to build, manage, and deploy AI-driven conversational interfaces efficiently, leveraging pre-built skills, natural language processing, and integration with other Oracle services.

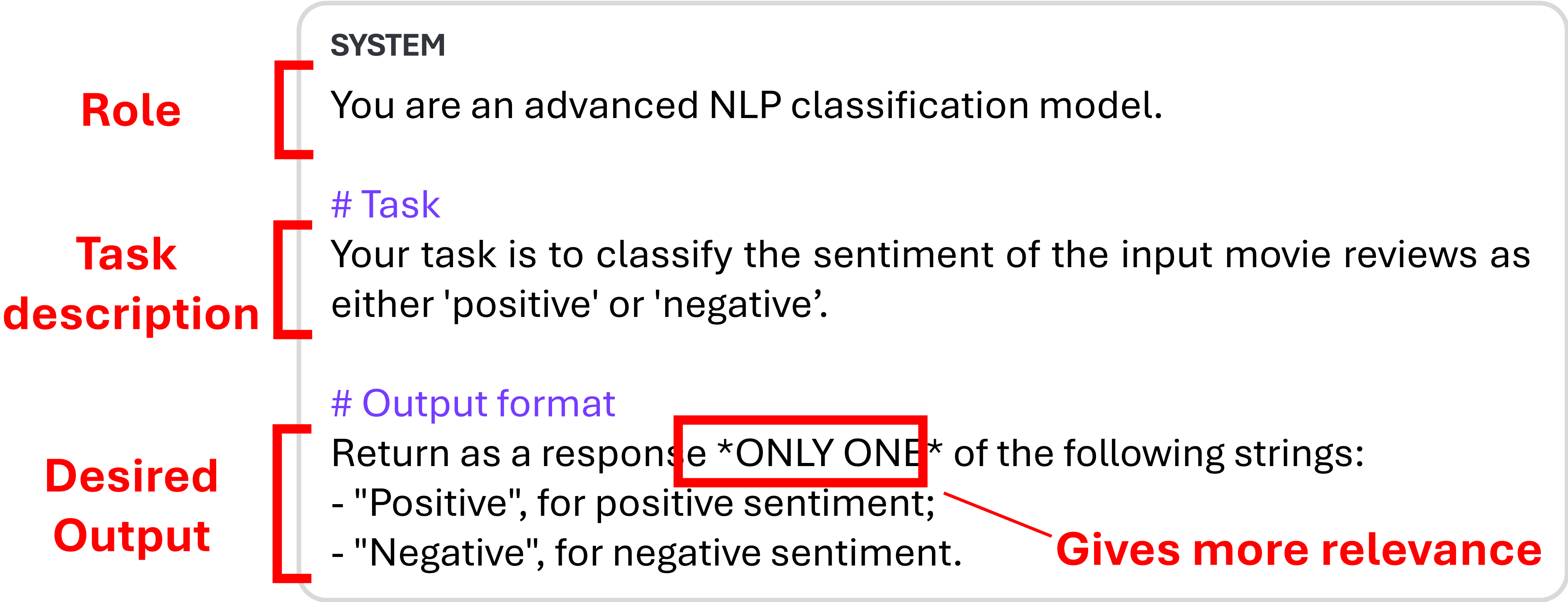
USER

Are there any alternatives from other companies?

ASSISTANT

I can provide information specifically on Oracle services. For building an AI assistant based on Generative AI, Oracle's Generative AI service is an excellent choice, providing the necessary tools, models, and cloud support.

How to write
a **good** prompt?



Role

Task description

Desired Output

```
SYSTEM
You are an advanced NLP classification model.

# Task
Your task is to classify the sentiment of the input movie reviews as either 'positive' or 'negative’.

# Output format
Return as a response *ONLY ONE* of the following strings:
- "Positive", for positive sentiment;
- "Negative", for negative sentiment.
```

Gives more relevance

1

Prompting best practices

SYSTEM

```
You are an advanced NLP classification model.

# Task
Your task is to classify the sentiment of the input user reviews as either 'positive' or 'negative’.

# Output format
Return as a response *ONLY ONE* of the following strings:
- "Positive", for positive sentiment;
- "Negative", for negative sentiment.

# Examples
*Input*: “The shipping is very slow”
*Response*: Negative

*Input*: “The best price for value!”
*Response*: Positive
```

Few-shot prompting

2

Provide examples (few-shot)

SYSTEM

You will be provided with a pair of articles (delimited with XML tags) about the same topic.

Step-by-step

Let’s think step-by-step: First summarize the arguments of each article. Then indicate which of them makes a better argument and explain why.

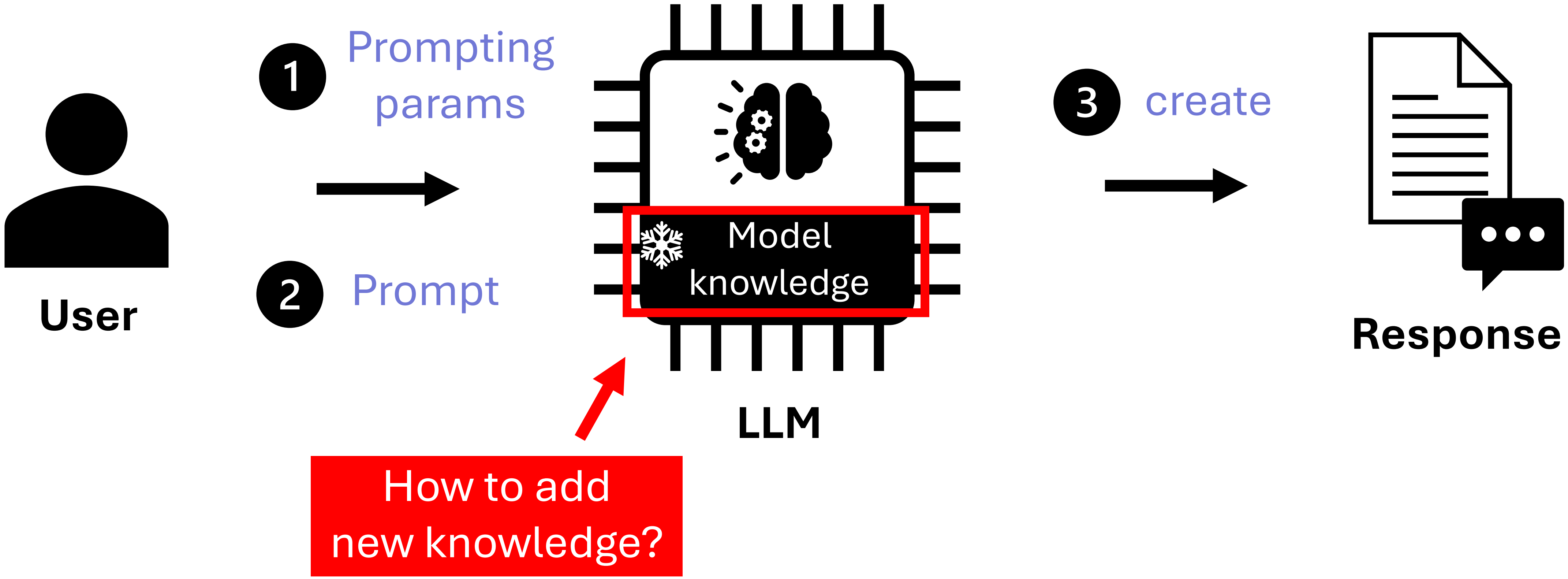
USER

<article> insert first article here </article>
<article> insert second article here </article>

3

Specify the logic steps to solve the task

Updating the
model knowledge



User

① Prompting params

② Prompt

Model knowledge

LLM

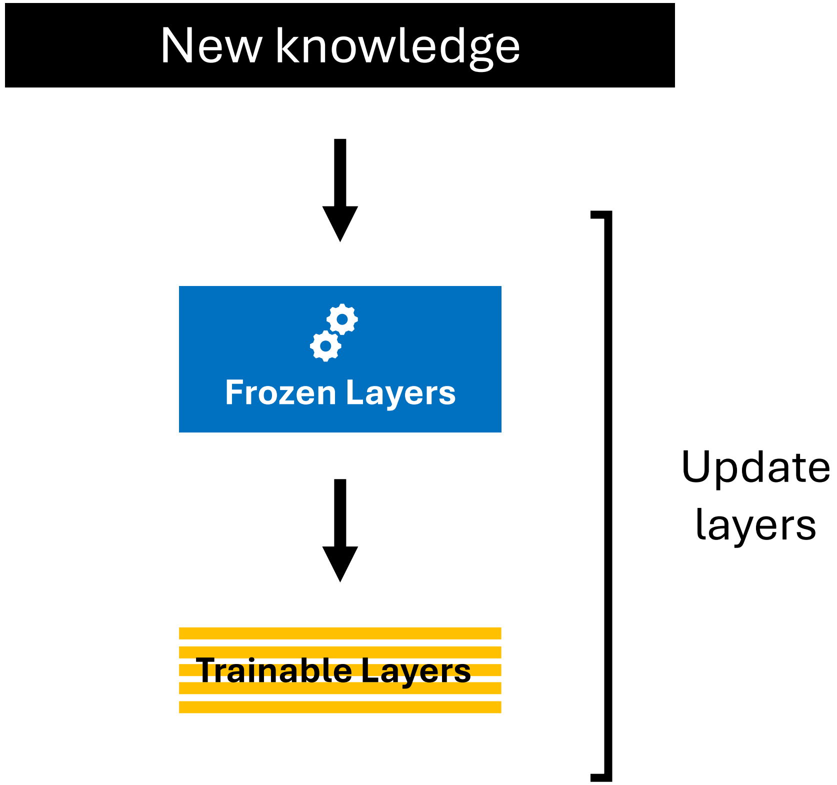
③ create

Response

How to add new knowledge?

Specialize the model via

Fine-Tuning



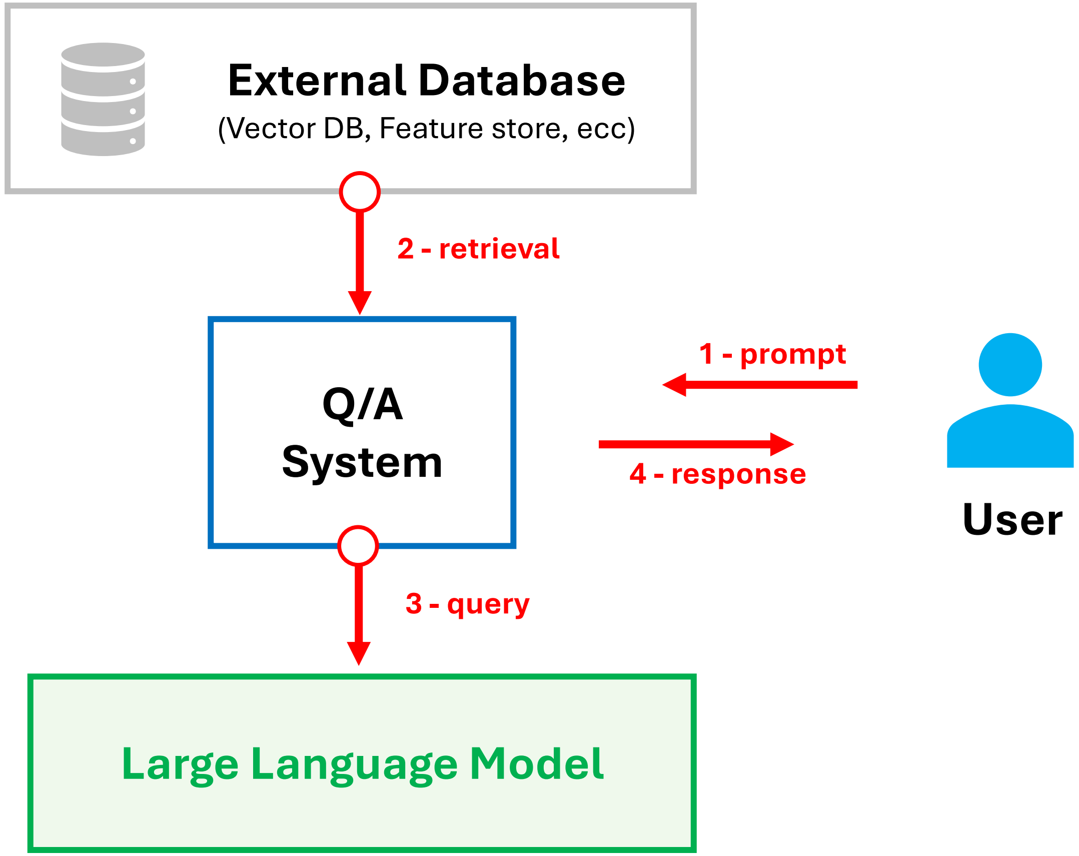
New knowledge

Frozen Layers

Trainable Layers

Update layers

Retrieval
Augmented
Generation



External Database

(Vector DB, Feature store, ecc)

2 - retrieval

Q/A System

1 - prompt

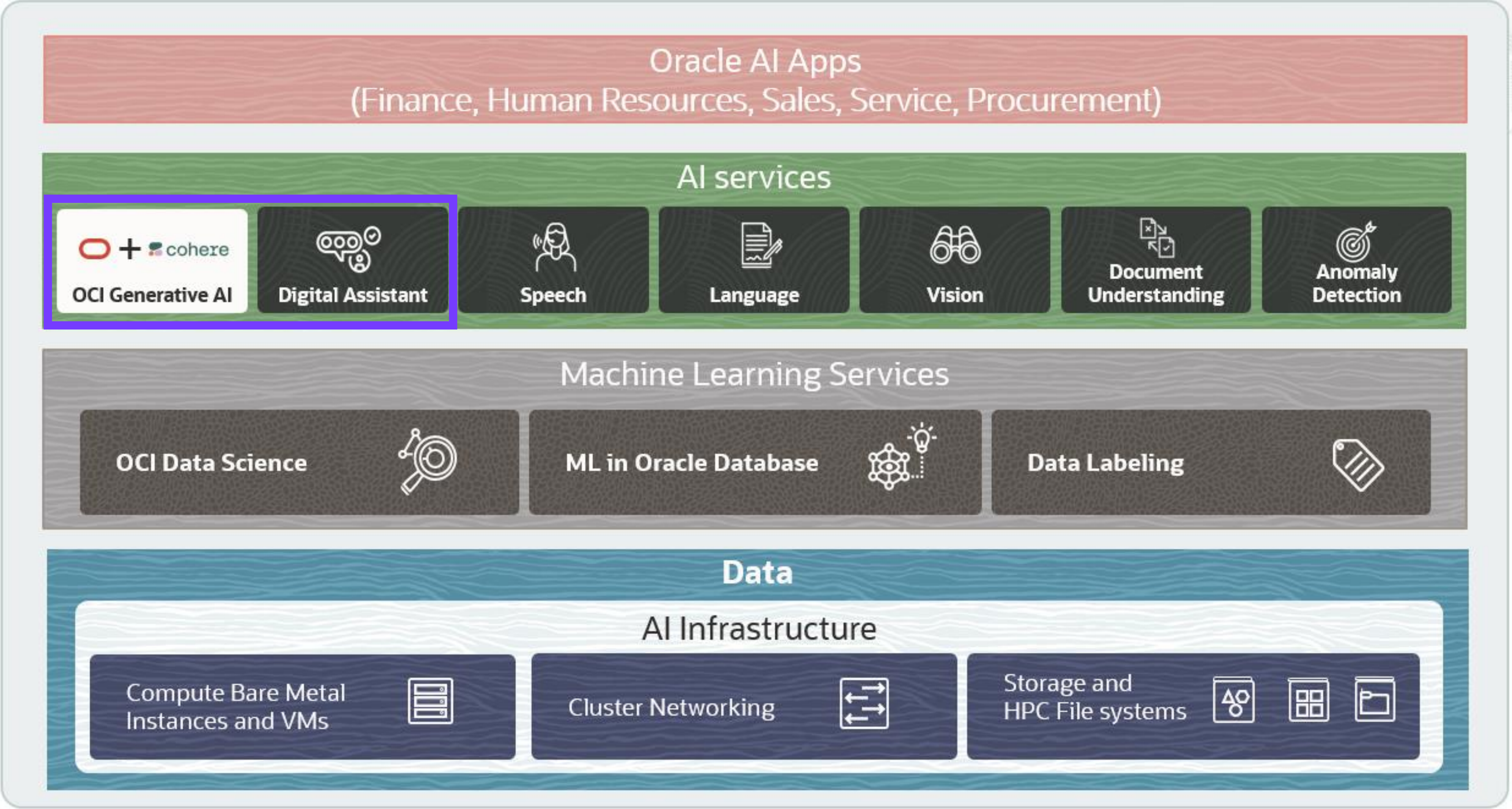
4 - response

User

3 - query

Large Language Model

Using LLMs
in practice



Oracle AI Apps
(Finance, Human Resources, Sales, Service, Procurement)

AI services

- OCI Generative AI
- Digital Assistant
- Speech
- Language
- Vision
- Document Understanding
- Anomaly Detection

Machine Learning Services

- OCI Data Science
- ML in Oracle Database
- Data Labeling

Data

AI Infrastructure

- Compute Bare Metal Instances and VMs
- Cluster Networking
- Storage and HPC File systems

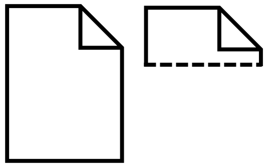
Oracle AI Services

Pay-per-use

by **# Tokens**

Input/output:

1000 tokens



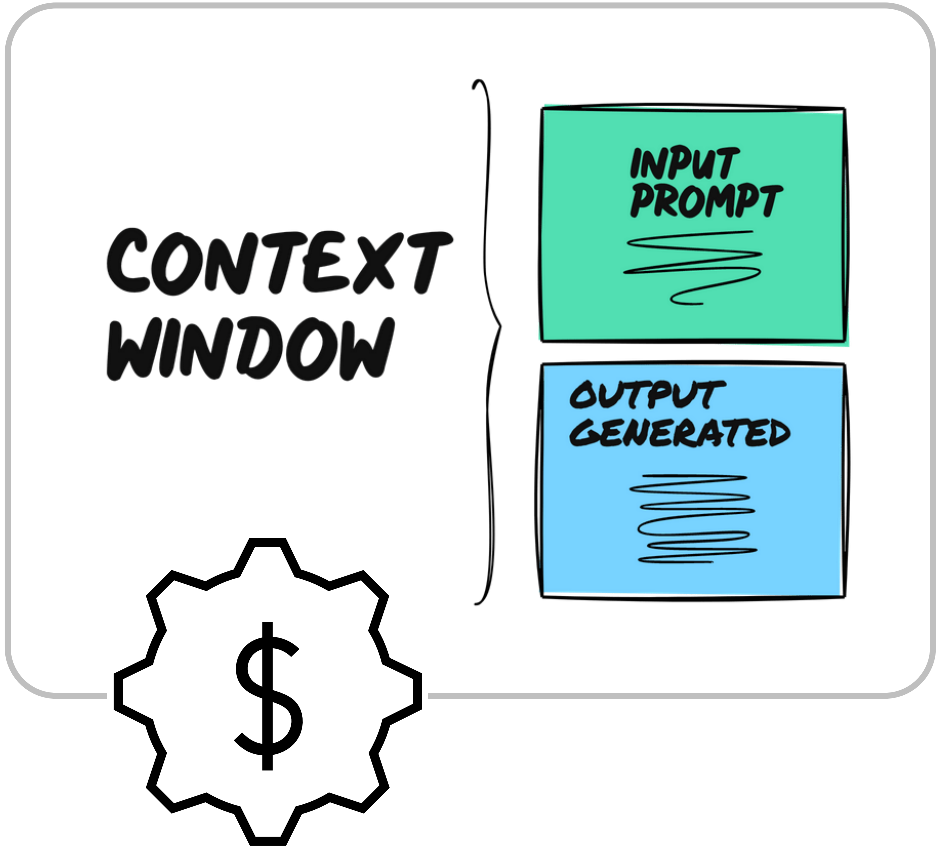
~1 A4 paper

Workload:

10000 requests per month

API cost per month:

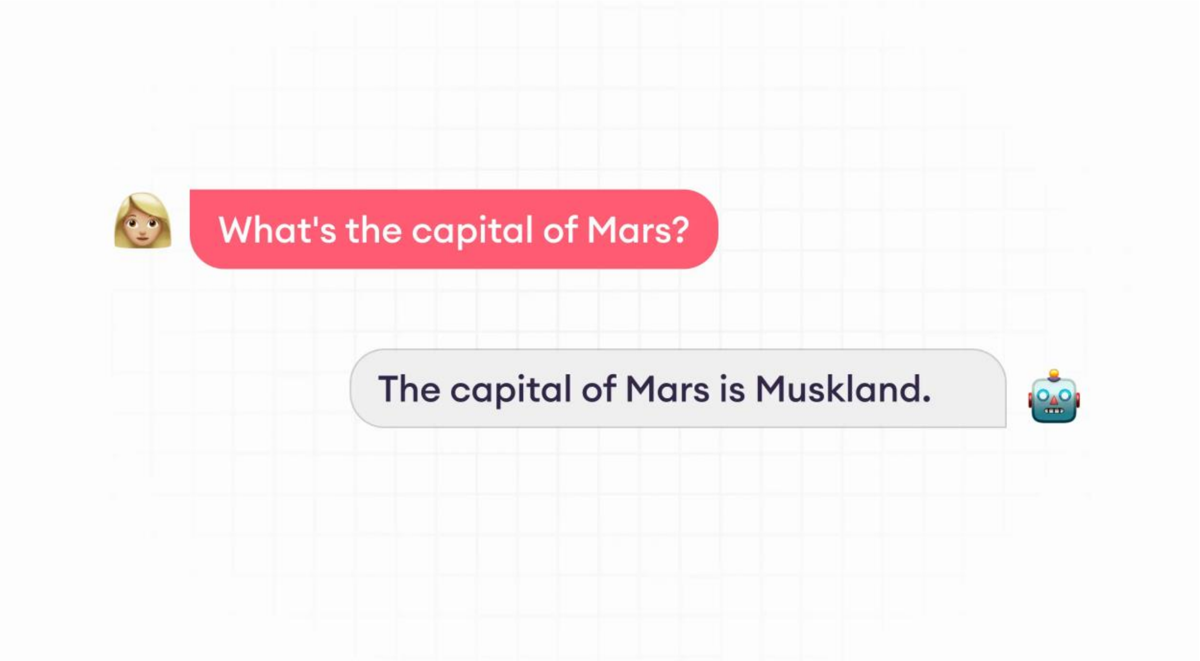
~90$ (Command R+ via OCI)



Estimating costs

LLM **Hallucinations**

- ✗ Ambiguous prompt
- ✗ Anomalies in training data
- ✗ Outdated model knowledge

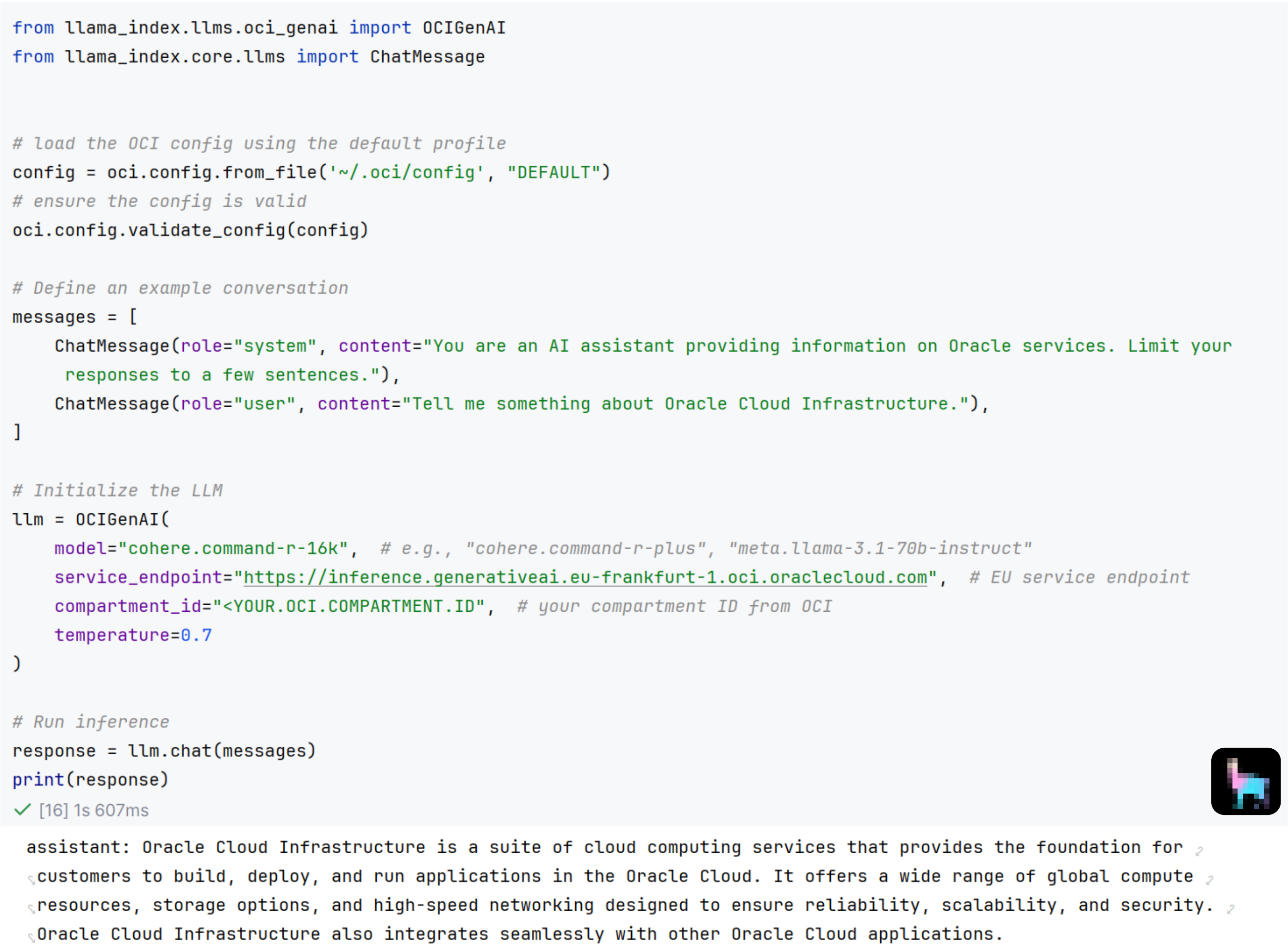


What's the capital of Mars?

The capital of Mars is Muskland.

Source: superannotate.com

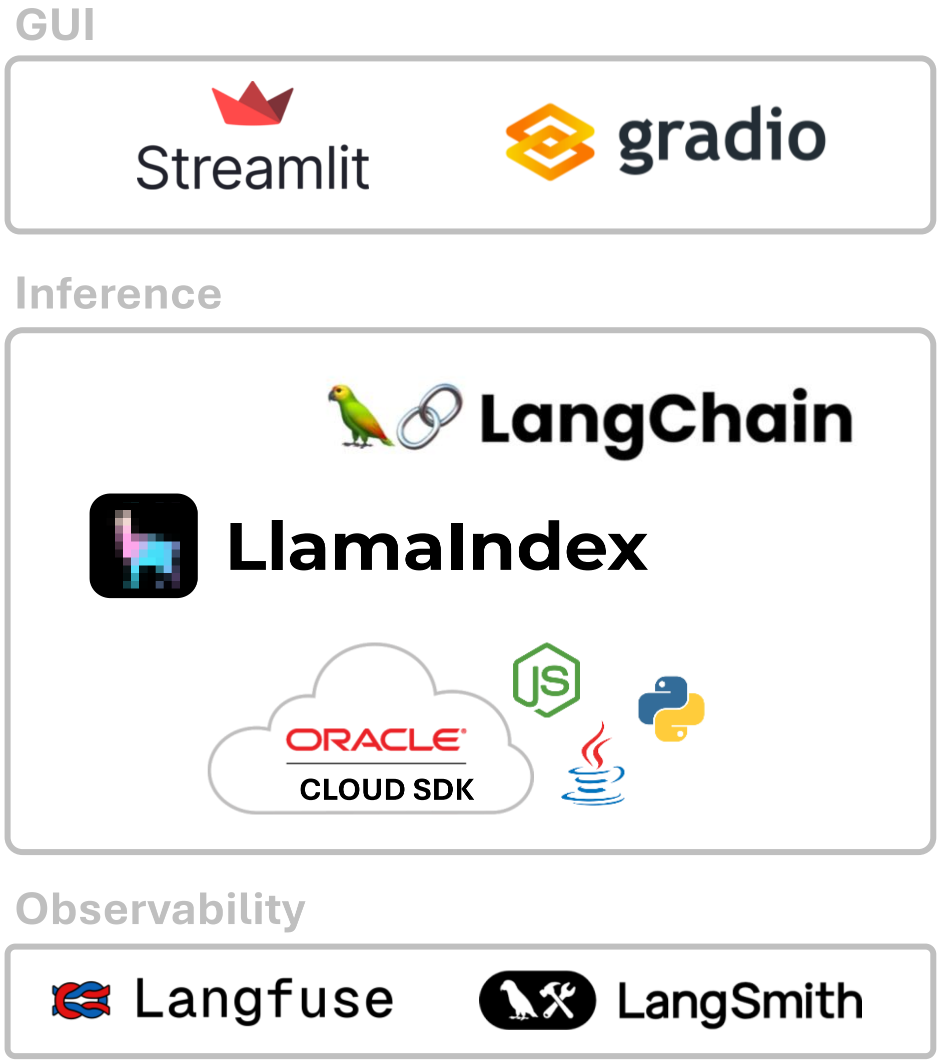
Beware of Hallucinations



```python
from llama_index.llms.oci_genai import OCIGenAI
from llama_index.core.llms import ChatMessage


# load the OCI config using the default profile
config = oci.config.from_file('~/.oci/config', "DEFAULT")
# ensure the config is valid
oci.config.validate_config(config)

# Define an example conversation
messages = [
    ChatMessage(role="system", content="You are an AI assistant providing information on Oracle services. Limit your responses to a few sentences."),
    ChatMessage(role="user", content="Tell me something about Oracle Cloud Infrastructure."),
]

# Initialize the LLM
llm = OCIGenAI(
    model="cohere.command-r-16k",  # e.g., "cohere.command-r-plus", "meta.llama-3.1-70b-instruct"
    service_endpoint="https://inference.generativeai.eu-frankfurt-1.oci.oraclecloud.com",  # EU service endpoint
    compartment_id="<YOUR.OCI.COMPARTMENT.ID>",  # your compartment ID from OCI
    temperature=0.7
)

# Run inference
response = llm.chat(messages)
print(response)
```

✓ [16] 1s 607ms

```
assistant: Oracle Cloud Infrastructure is a suite of cloud computing services that provides the foundation for
customers to build, deploy, and run applications in the Oracle Cloud. It offers a wide range of global compute
resources, storage options, and high-speed networking designed to ensure reliability, scalability, and security.
Oracle Cloud Infrastructure also integrates seamlessly with other Oracle Cloud applications.
```



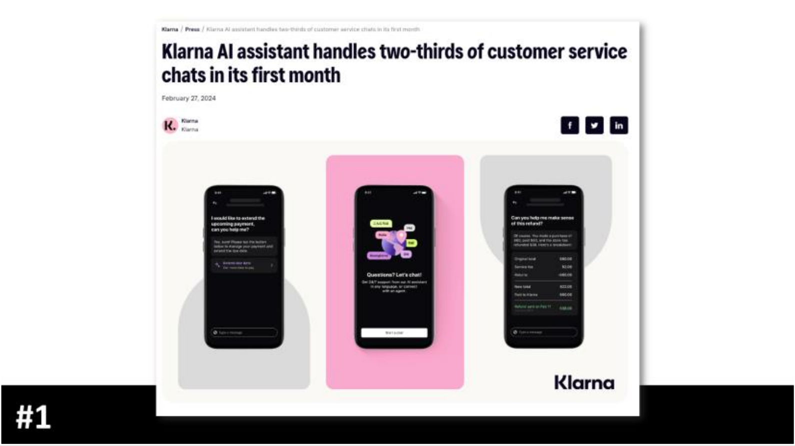
GUI

Inference

Observability

Build a simple GenAI application

Summary

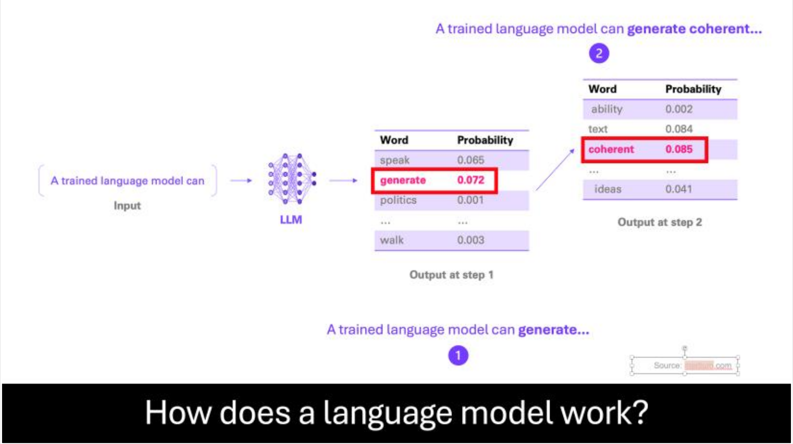

Klarna / Press / Klarna AI assistant handles two-thirds of customer service chats in its first month

Klarna AI assistant handles two-thirds of customer service chats in its first month

February 27, 2024

Klarna
Klarna

#1


A trained language model can
Input
LLM
Output at step 1
Output at step 2
A trained language model can **generate...**
1
A trained language model can **generate coherent...**
2
Source: [illegible].com

Word	Probability
speak	0.065
generate	**0.072**
politics	0.001
...	...
walk	0.003

Word	Probability
ability	0.002
text	0.084
coherent	**0.085**
...	...
ideas	0.041

How does a language model work?

GPT-4

Est. train size

650 km

Line of library shelves

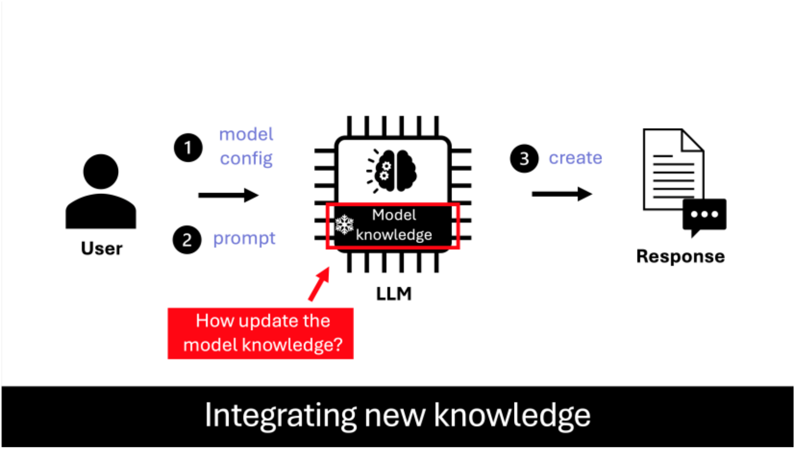
Distanza Campobasso → Como
Distanza: 651,44 km
Percorso stradale: 793,48 km
Campobasso
Como

CALCOLA LA DISTANZA!
BULK
Como
Campobasso
it.distance.to

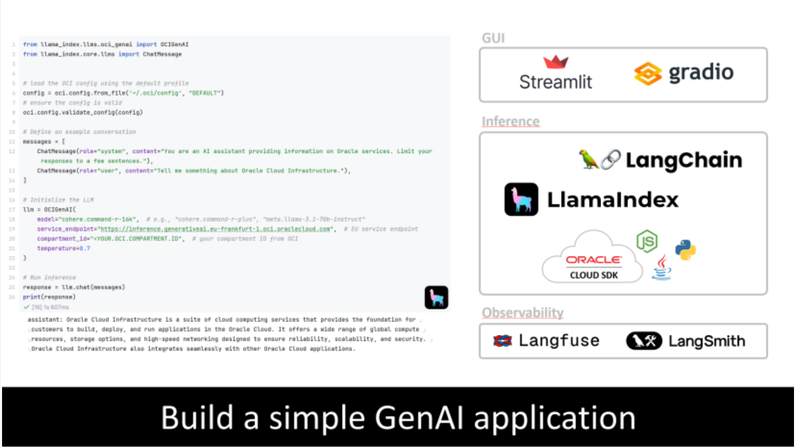
How to train a «Large» Language Model?

Prompt Engineering

More at: www.promptingguide.ai



Integrating new knowledge


```python
from llama_index.llms.oci_genai import OCIGenAI
from llama_index.core.llms import ChatMessage


# load the OCI config using the default profile
config = oci.config.from_file('~/.oci/config', "DEFAULT")
# ensure the config is valid
oci.config.validate_config(config)

# Define an example conversation
messages = [
    ChatMessage(role="system", content="You are an AI assistant providing information on Oracle services. Limit your
     responses to a few sentences."),
    ChatMessage(role="user", content="Tell me something about Oracle Cloud Infrastructure."),
]

# Initialize the LLM
llm = OCIGenAI(
    model="cohere.command-r-16k",  # e.g., "cohere.command-r-plus", "meta.llama-3.1-70b-instruct"
    service_endpoint="https://inference.generativeai.eu-frankfurt-1.oci.oraclecloud.com",  # EU service endpoint
    compartment_id="<YOUR.OCI.COMPARTMENT.ID",  # your compartment ID from OCI
    temperature=0.7
)

# Run inference
response = llm.chat(messages)
print(response)
```

```
✓ [16] 1s 807ms
assistant: Oracle Cloud Infrastructure is a suite of cloud computing services that provides the foundation for
customers to build, deploy, and run applications in the Oracle Cloud. It offers a wide range of global compute
resources, storage options, and high-speed networking designed to ensure reliability, scalability, and security.
Oracle Cloud Infrastructure also integrates seamlessly with other Oracle Cloud applications.
```

Build a simple GenAI application

Giovanni Rosa, PhD
AI Engineer @Technology Reply
giovannirosa.com

Thank You!

SCAN ME