

# *TH* Dimp My Llama

**How to Customize LLMs  
for your Business Case**

# Speakers



**Christoph Bergemann**  
Principal Consultant



**Jonas Mayer**  
Senior Consultant

# About Us

Innovation Hacking



# About Us

Innovation Hacking: Deepfakes



# About Us

Innovation Hacking: [ai-slides.com](https://ai-slides.com)



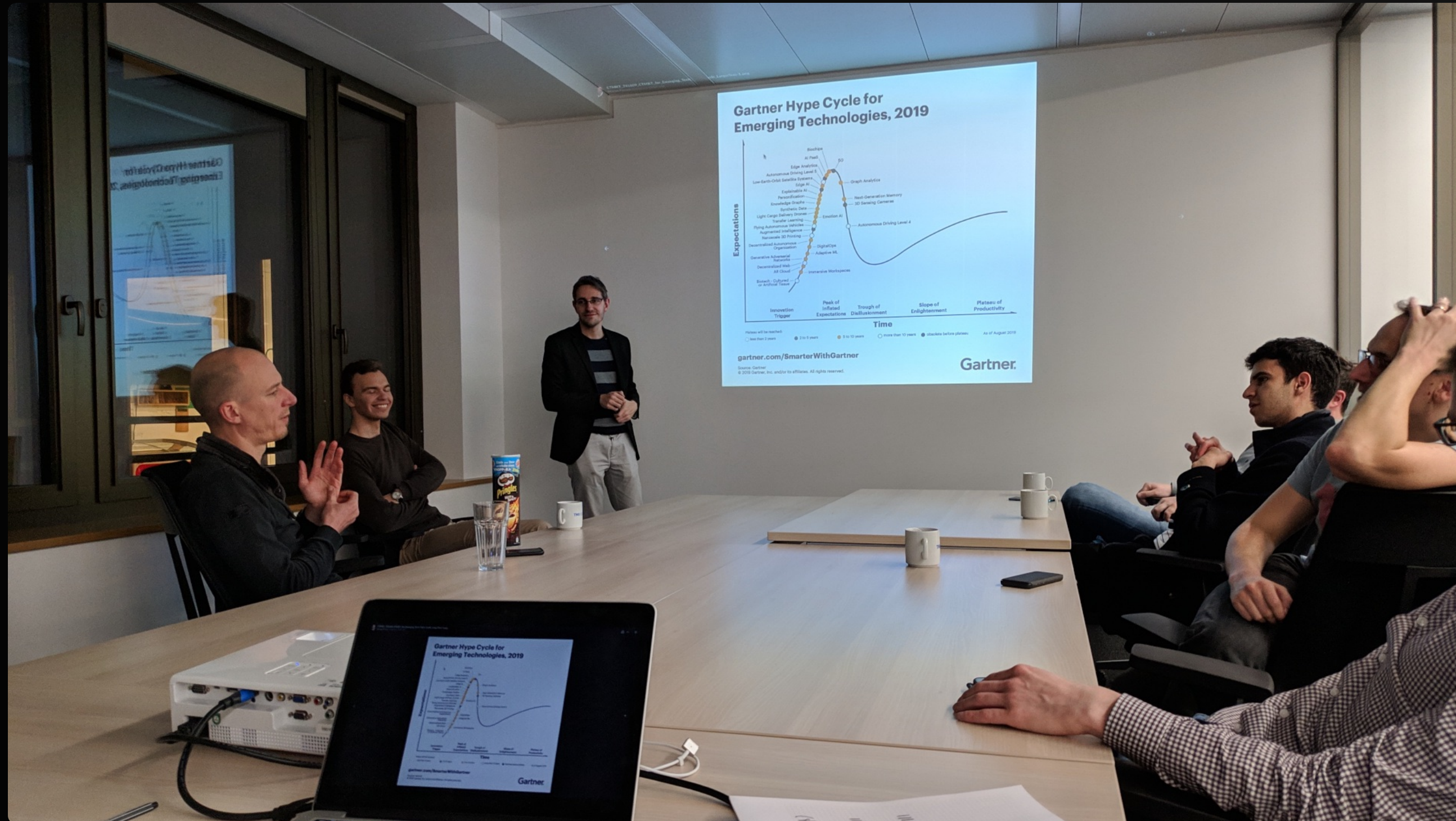
# Motivation

But Why?



# Motivation

But Why?



# Motivation

But Why?





# Motivation

Innovation Hacking Q&A



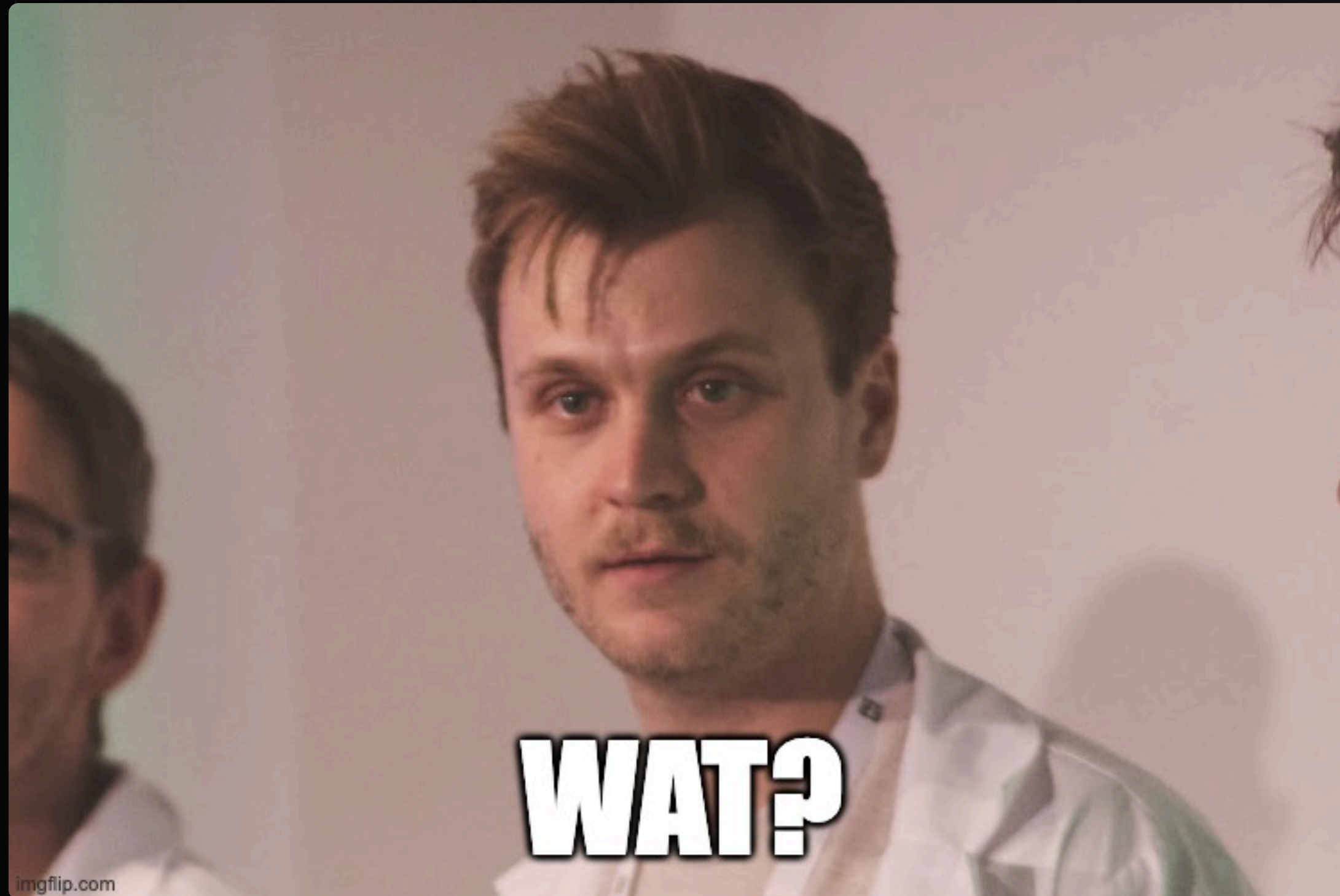
# Motivation

Innovation Hacking Q&A



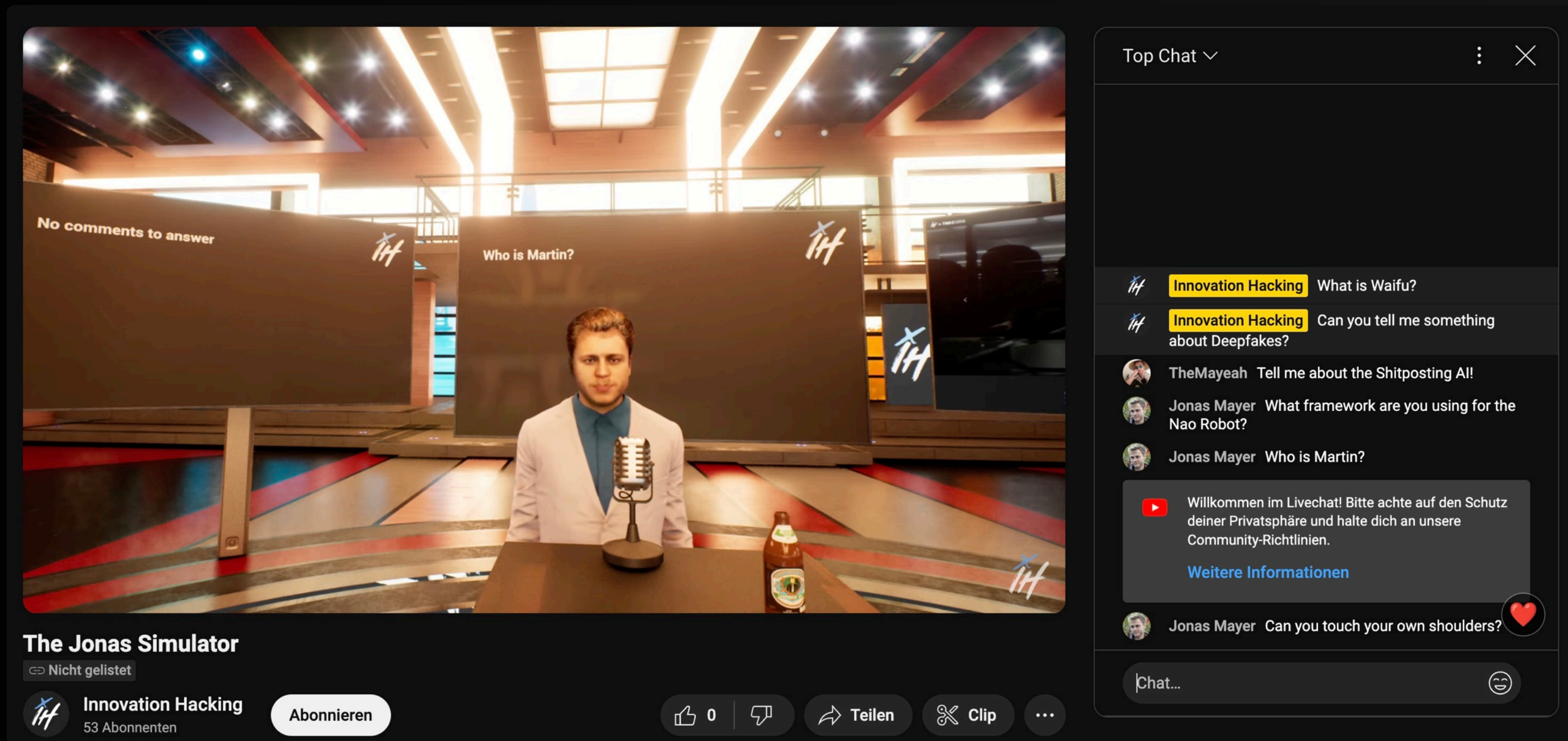
# Motivation

Innovation Hacking Q&A



# Motivation

A 24/7 Q&A LLM Live-Stream



**The Jonas Simulator**  
↔ Nicht gelistet

**Innovation Hacking**  
53 Abonnenten

Abonnieren

0 | Teilen | Clip | ...

Top Chat ▾

- Innovation Hacking** What is Waifu?
- Innovation Hacking** Can you tell me something about Deepfakes?
- TheMayeah** Tell me about the Shitposting AI!
- Jonas Mayer** What framework are you using for the Nao Robot?
- Jonas Mayer** Who is Martin?

Willkommen im Livechat! Bitte achte auf den Schutz deiner Privatsphäre und halte dich an unsere Community-Richtlinien.  
[Weitere Informationen](#)

**Jonas Mayer** Can you touch your own shoulders? ❤️

Chat...

# Overview

- ▶ Introduction
- ▶ Prompt Engineering
- ▶ RAG
- ▶ Fine-Tuning
- ▶ Conclusion



# LLMs in a Nutshell

Large Language Models



# LLMs in a Nutshell

Large Language Models



# LLMs in a Nutshell

Popular Models





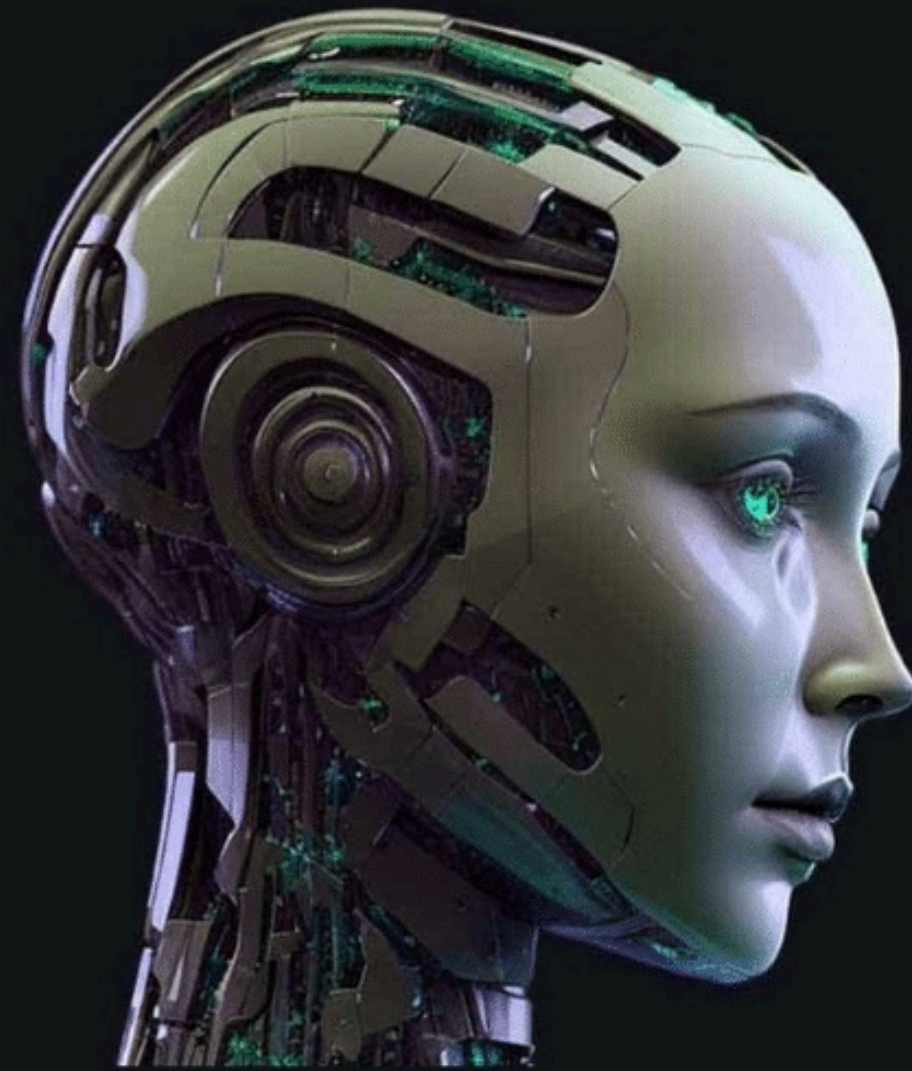
# LLMs in a Nutshell

Popular Models



# LLMs in a Nutshell

Popular Models



 OpenAI



**GPT-4**

# LLMs Under the Hood

How LLMs See and Generate Text



# LLMs Under the Hood

Tokenization

**Innovation Hacking is so so fun**

# LLMs Under the Hood

Tokenization

**“Innovation Hacking is so so fun”**

# LLMs Under the Hood

Tokenization

“**Innovation** Hacking is so so fun”

# LLMs Under the Hood

Tokenization

“Innovation **Hacking** is so so fun”

# LLMs Under the Hood

Tokenization

“Innovation Hacking **is** so so fun”



# LLMs Under the Hood

Tokenization

“Innovation Hacking is **so** so fun”

# LLMs Under the Hood

Tokenization

“Innovation Hacking is so **so** fun”

# LLMs Under the Hood

Tokenization

“Innovation Hacking is so so fun”

# LLMs Under the Hood

Tokenization

**“Innovation Hacking is so so fun”**



# LLMs Under the Hood

Tokenization

**“Innovation Hacking is so so fun”**



# LLMs Under the Hood

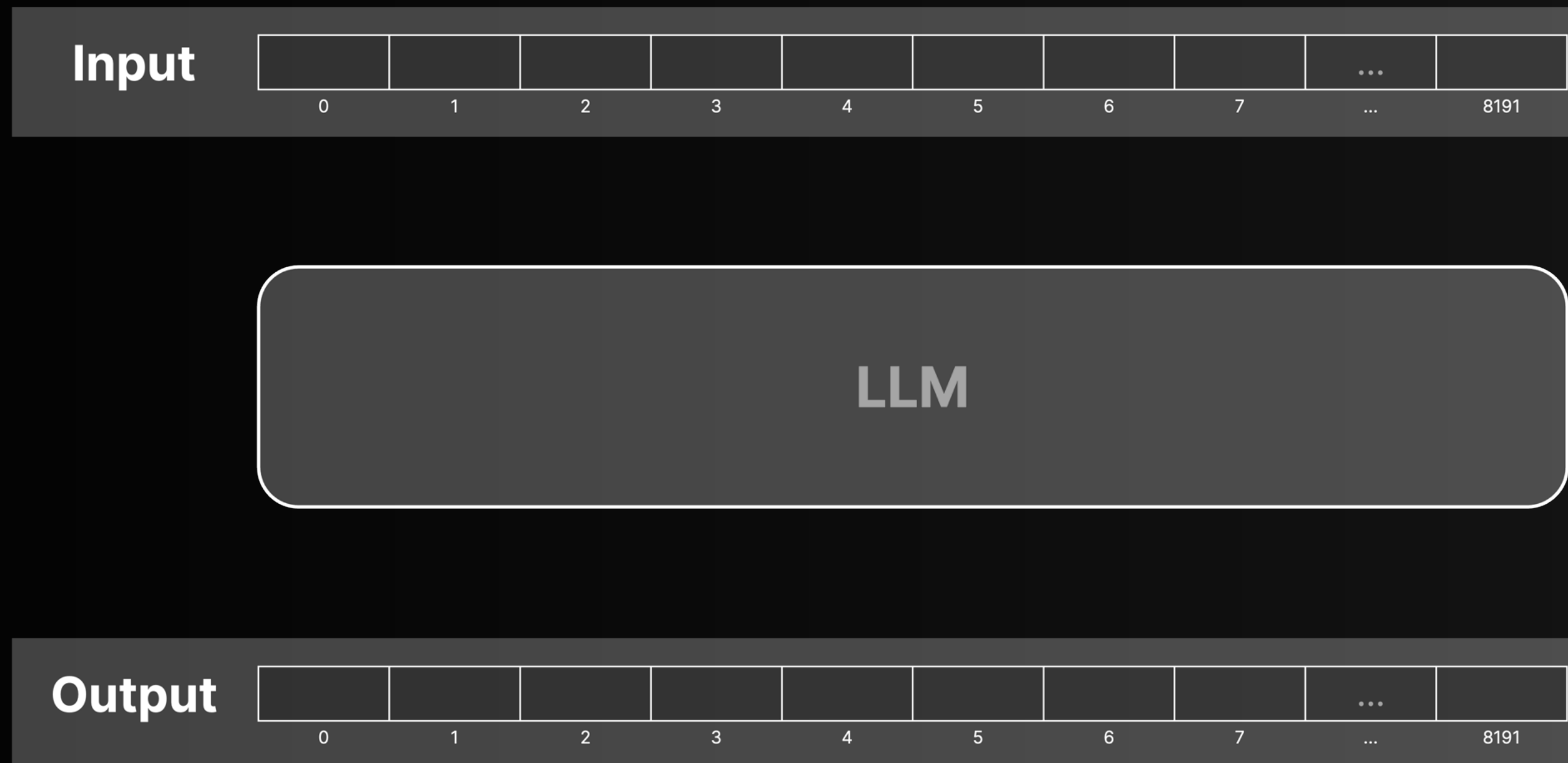
Tokenizer Demo

Input

Please start the demo!

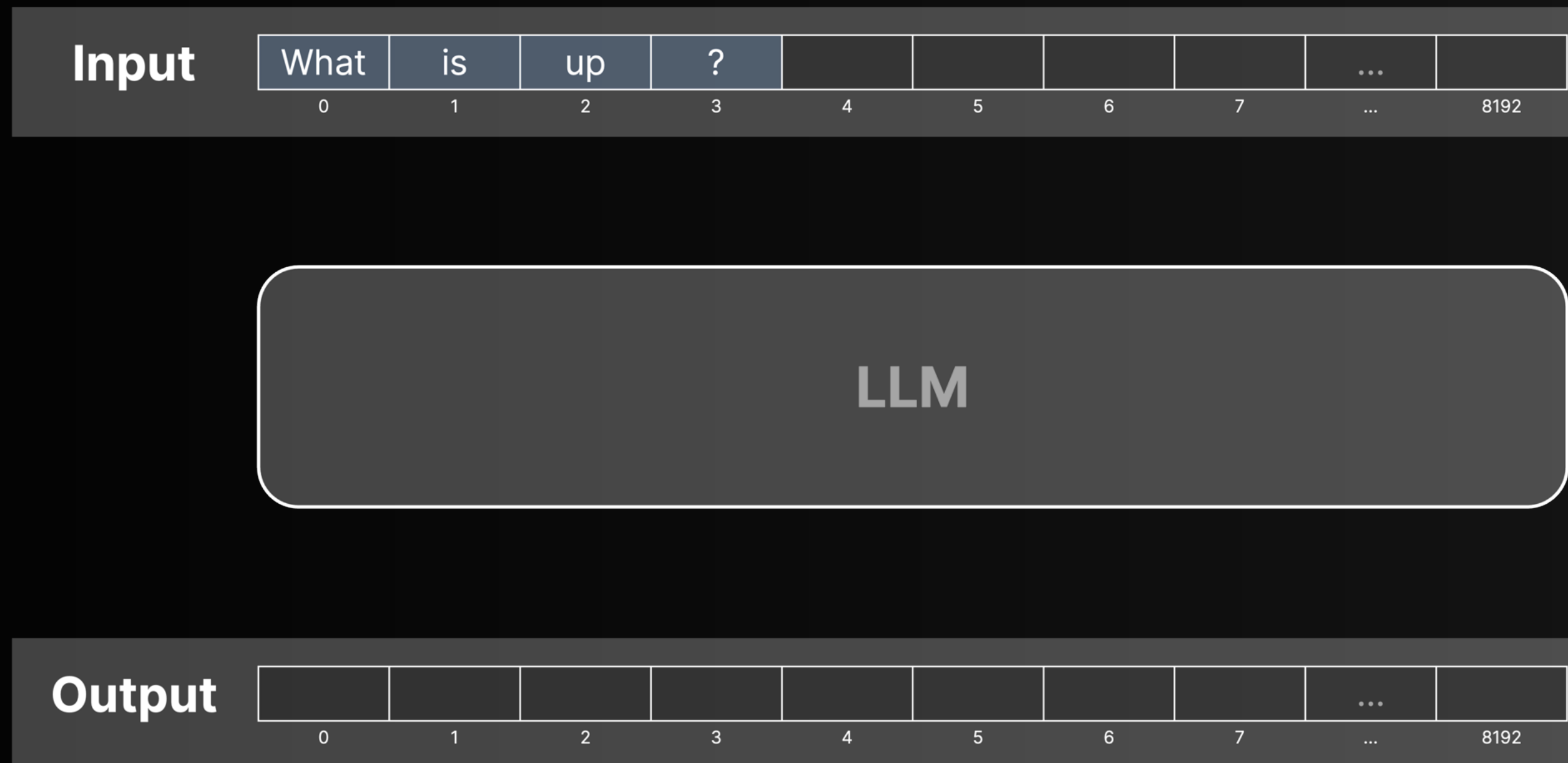
# LLMs Under the Hood

## Next Word Prediction



# LLMs Under the Hood

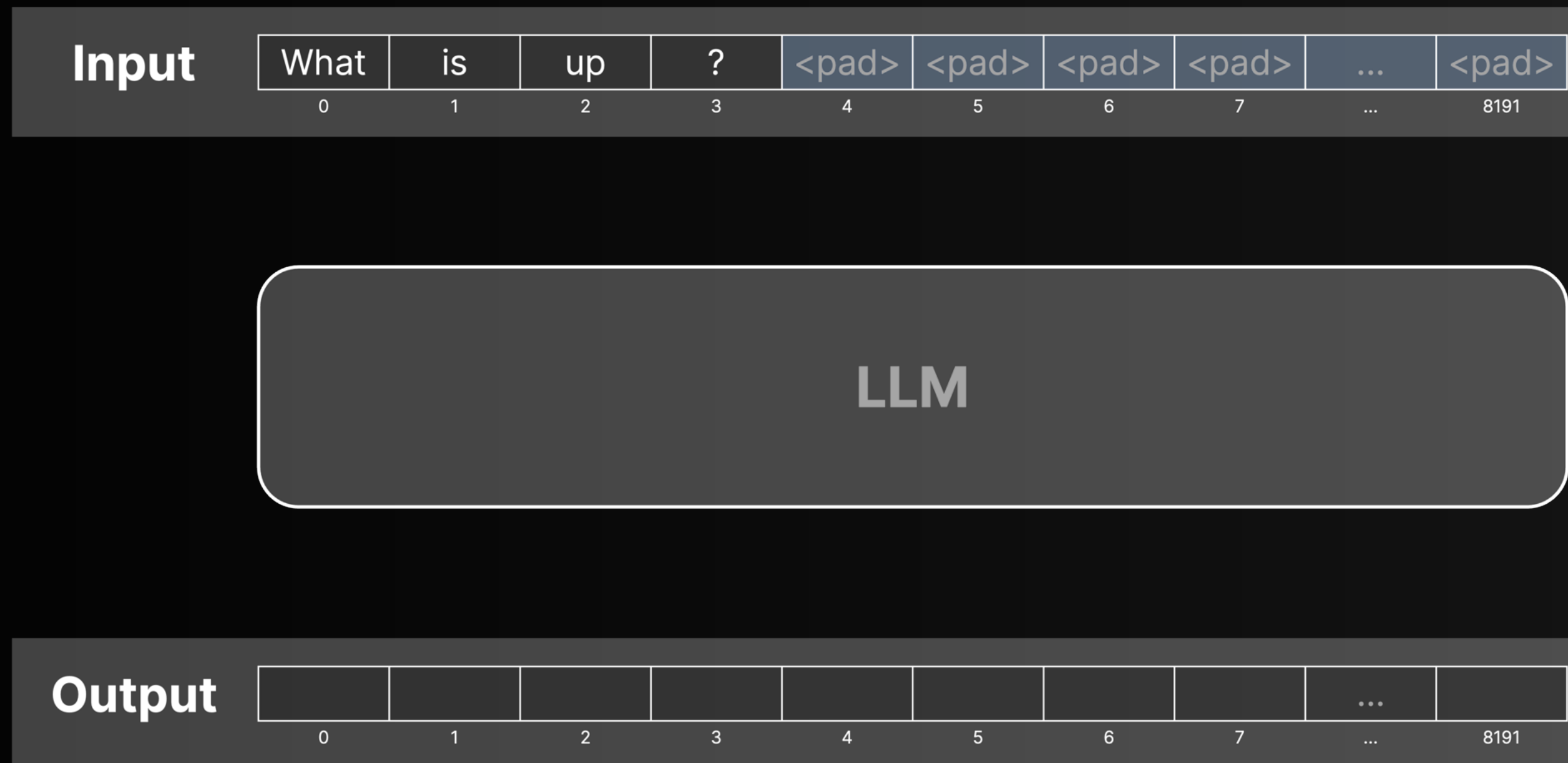
## Next Word Prediction





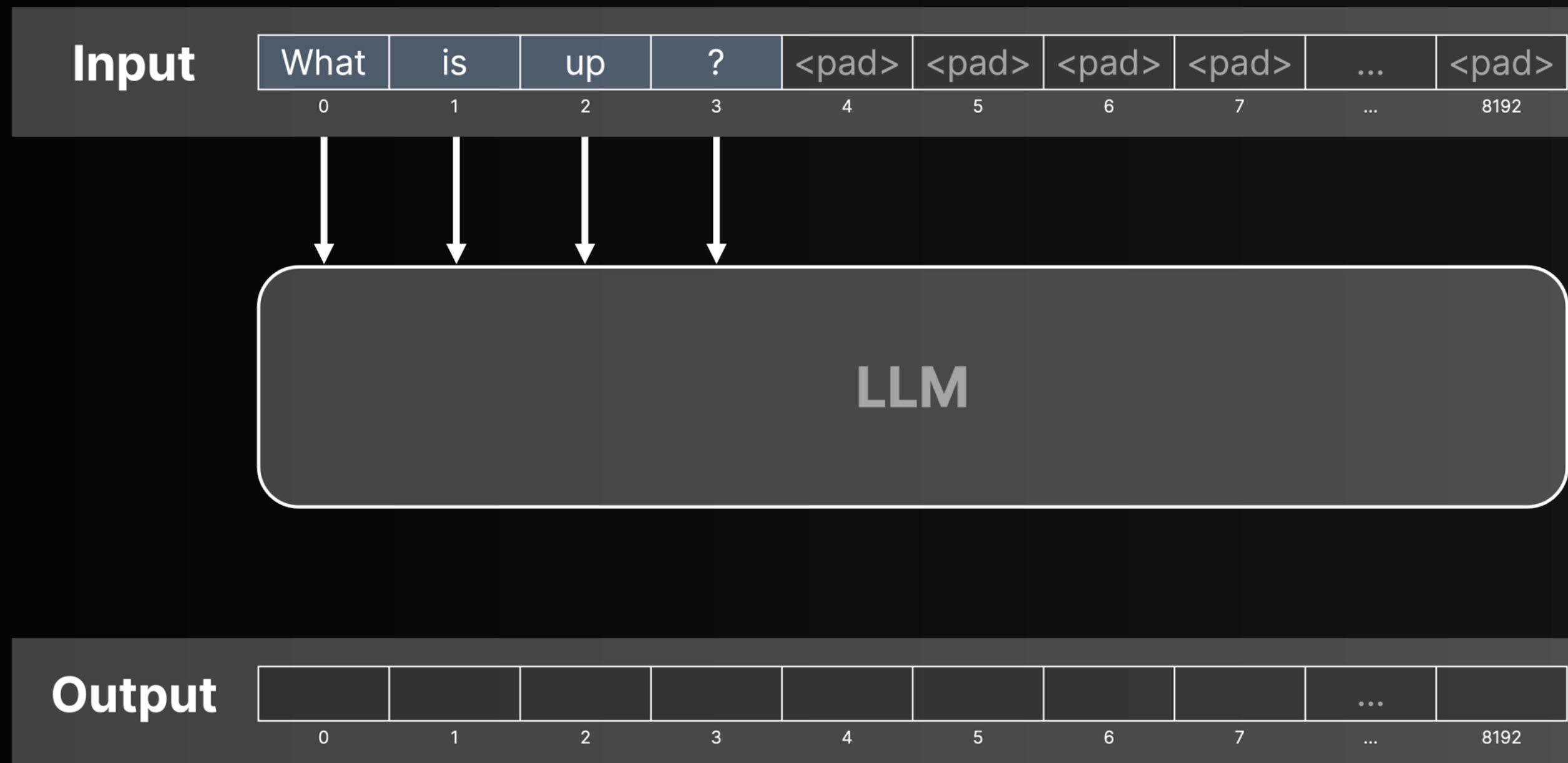
# LLMs Under the Hood

## Next Word Prediction



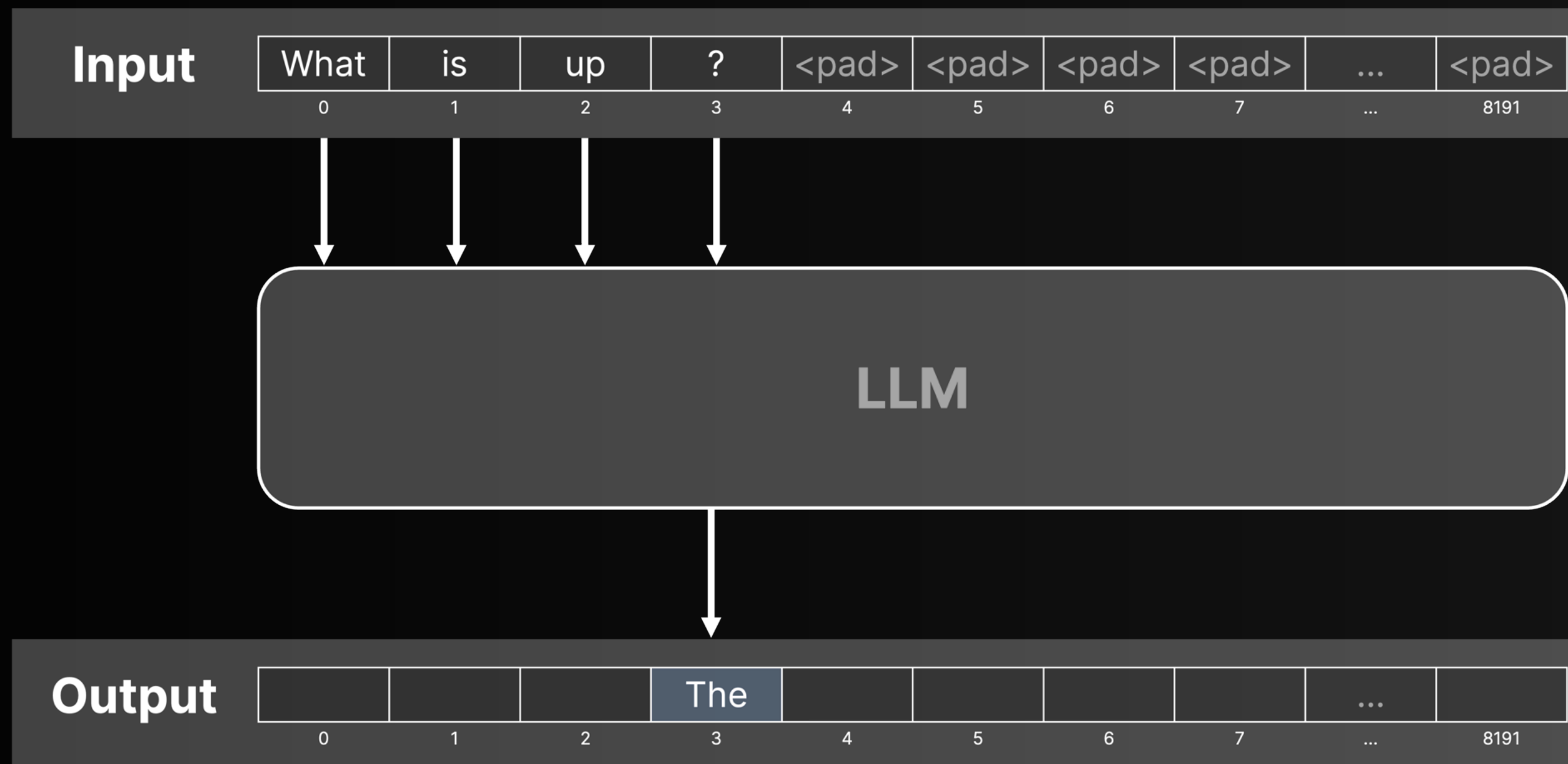
# LLMs Under the Hood

## Next Word Prediction



# LLMs Under the Hood

## Next Word Prediction



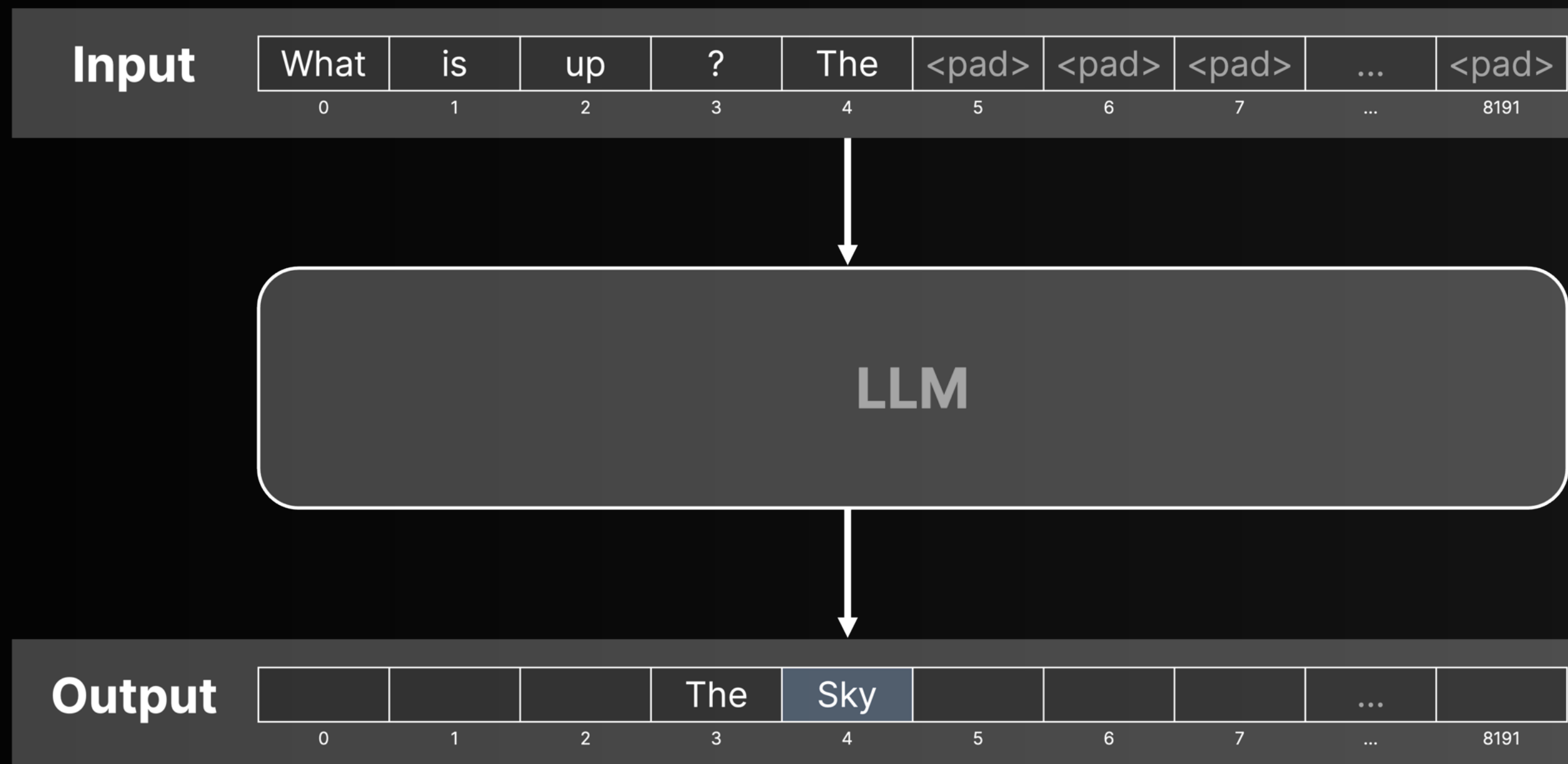
# LLMs Under the Hood

## Next Word Prediction



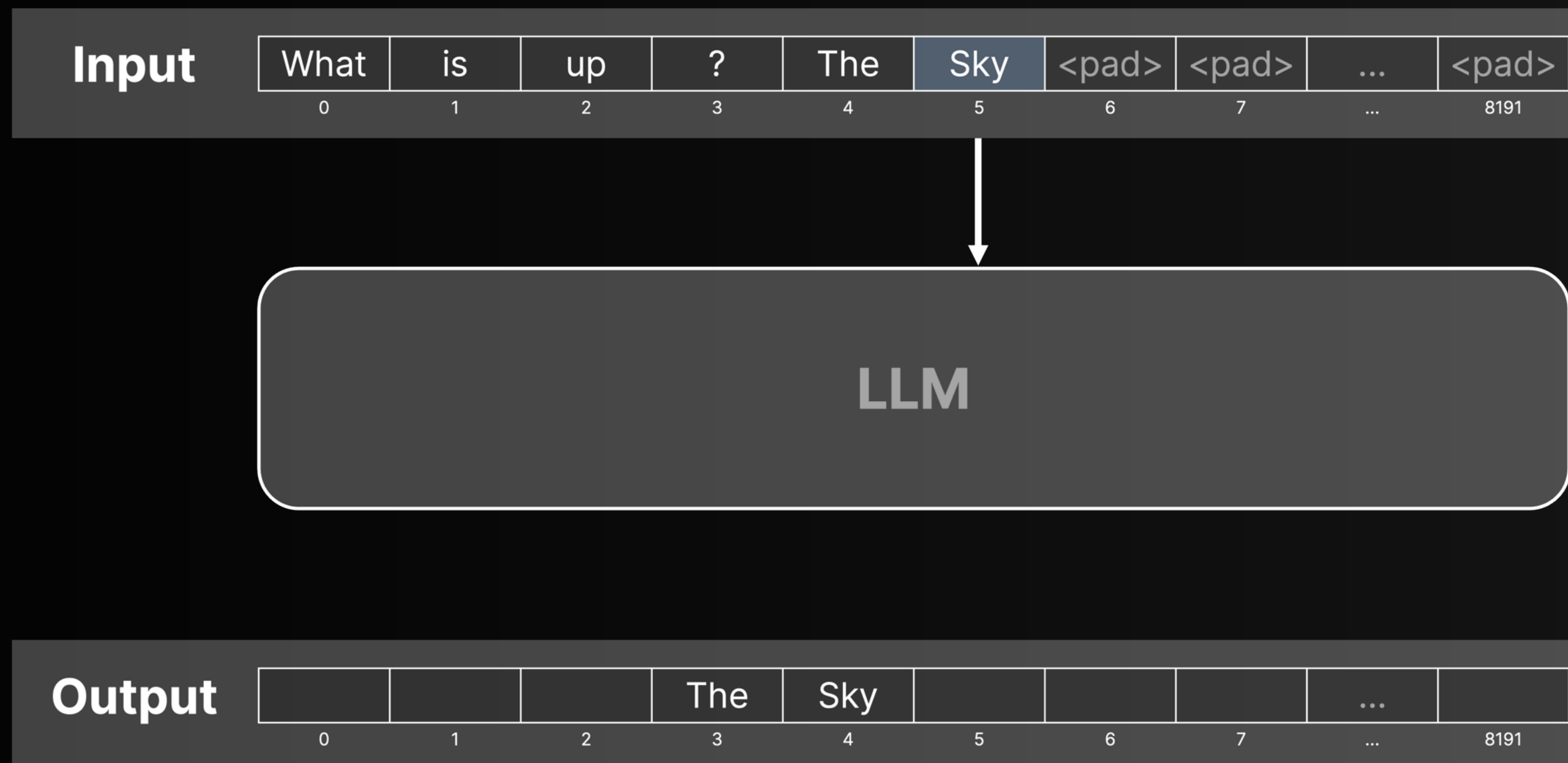
# LLMs Under the Hood

## Next Word Prediction



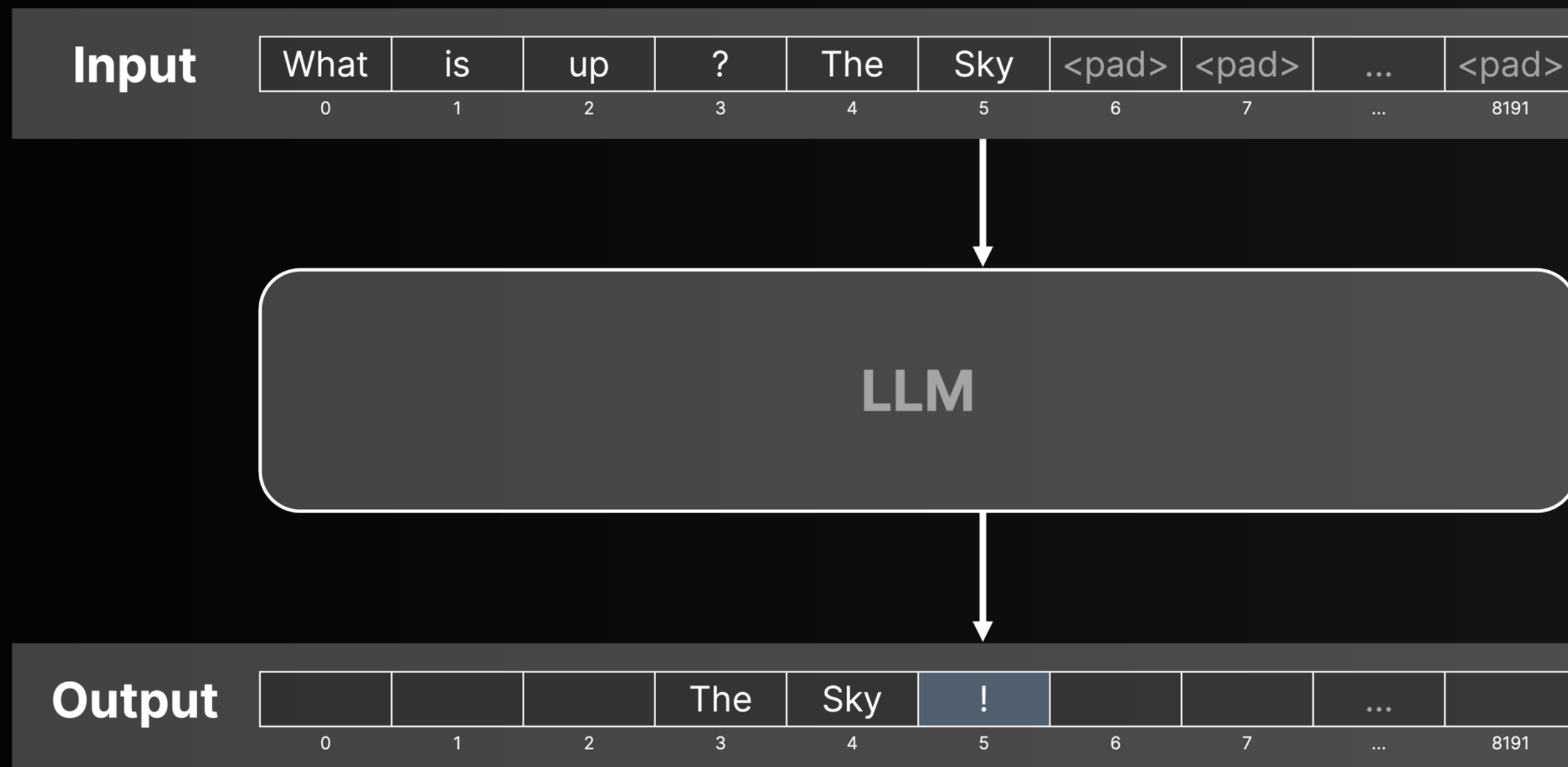
# LLMs Under the Hood

## Next Word Prediction



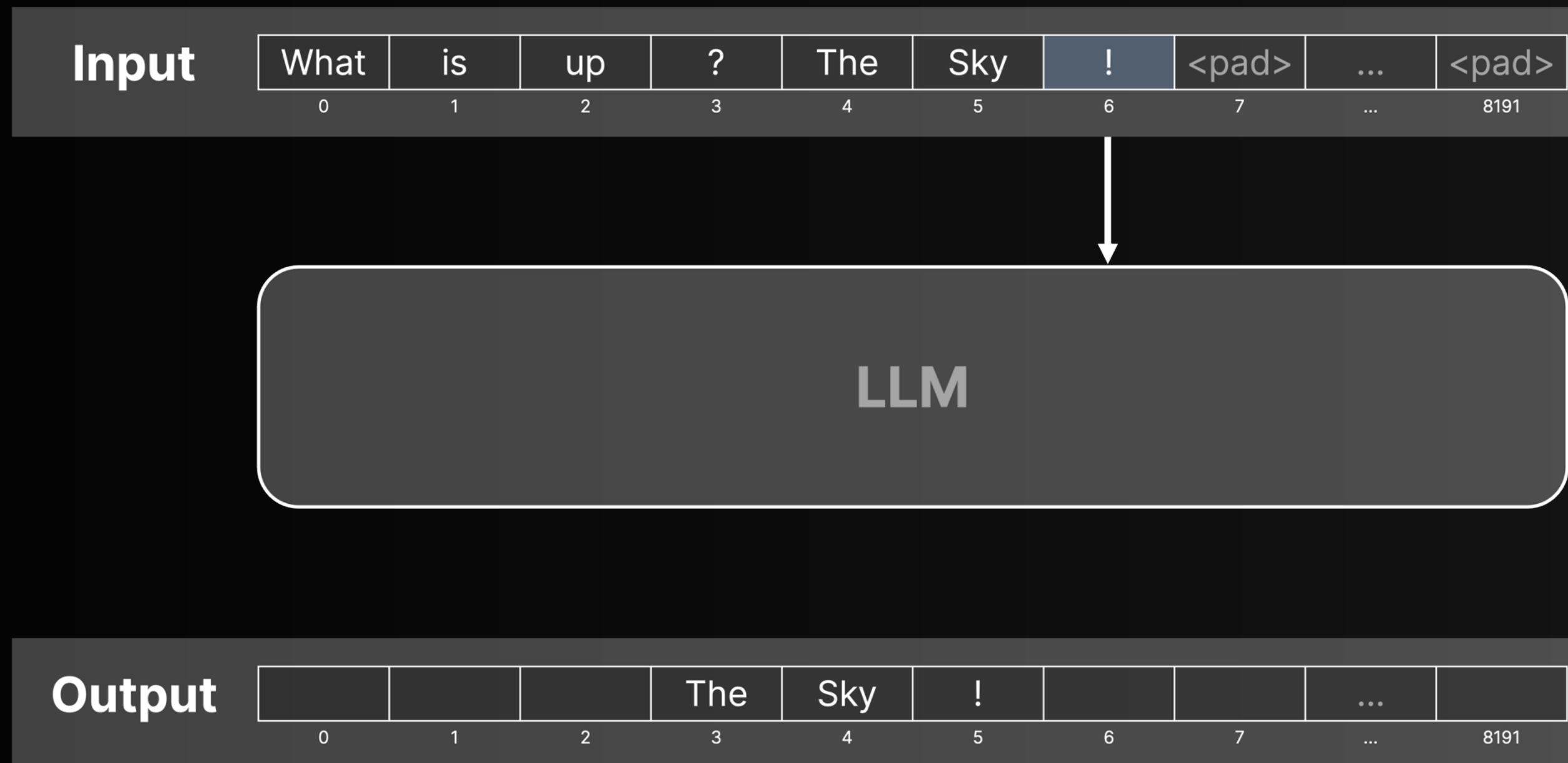
# LLMs Under the Hood

## Next Word Prediction



# LLMs Under the Hood

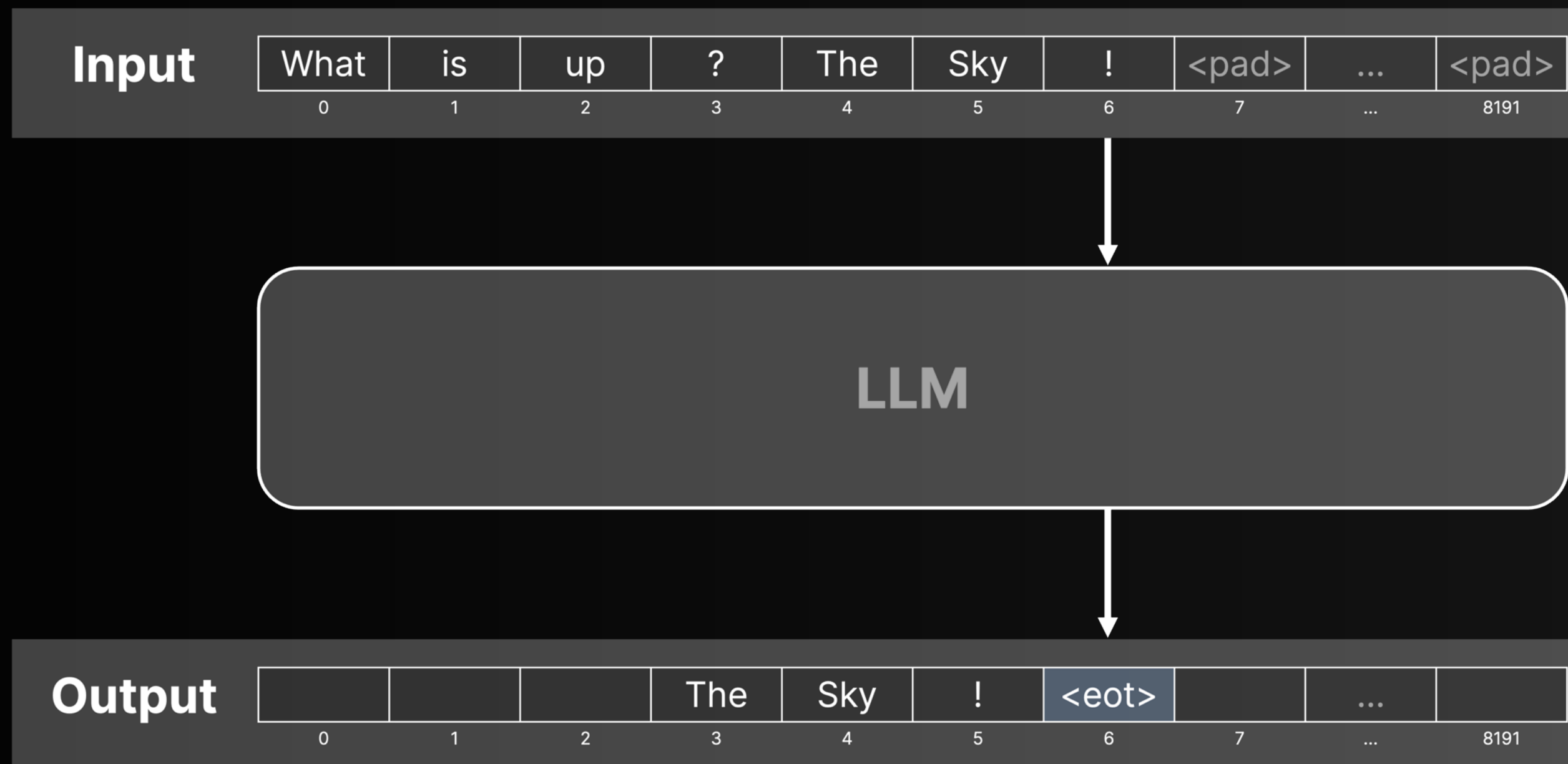
## Next Word Prediction





# LLMs Under the Hood

## Next Word Prediction



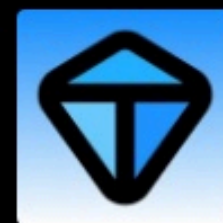
# LLMs in a Nutshell

Running LLMs



# LLMs in a Nutshell

HuggingFace Text Generation Inference



**text-generation-inference**



[huggingface.co/text-generation-inference](https://huggingface.co/text-generation-inference)

# LLMs in a Nutshell

Demo: Using a base Llama

**Input**

Please start the demo!

# Overview

- ▶ Introduction
- ▶ Prompt Engineering
- ▶ RAG
- ▶ Fine-Tuning
- ▶ Conclusion



# Prompt Engineering in a Nutshell

What is a Prompt?



How can I help you today?

Test my knowledge  
on ancient civilizations

Design a programming game  
teach basics in a fun way

I'm writing a talk about optimizing LLM performance. How could  
i visualize what a prompt is????



ChatGPT can make mistakes. Check important info.

# Prompt Engineering in a Nutshell

What is Prompt Engineering?



**You**

What is up?



**ChatGPT**

Hey! Not much, just here to help out. How about you? Is there something specific you need assistance with today?



**You**

Can you just reply with "The Sky!" already! >:(



**ChatGPT**

The Sky!

# Prompt Engineering in a Nutshell

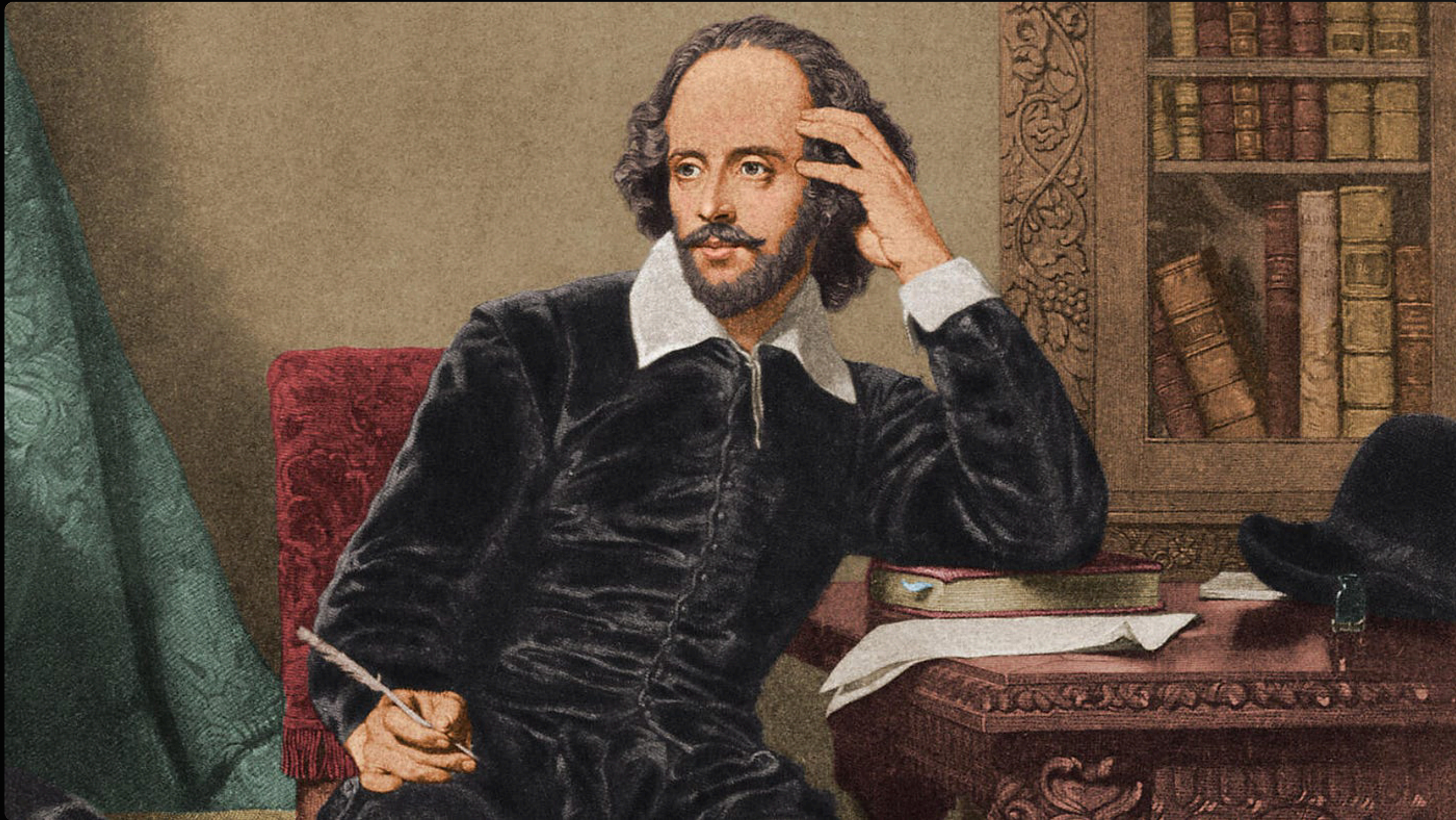
Chatting with LLMs





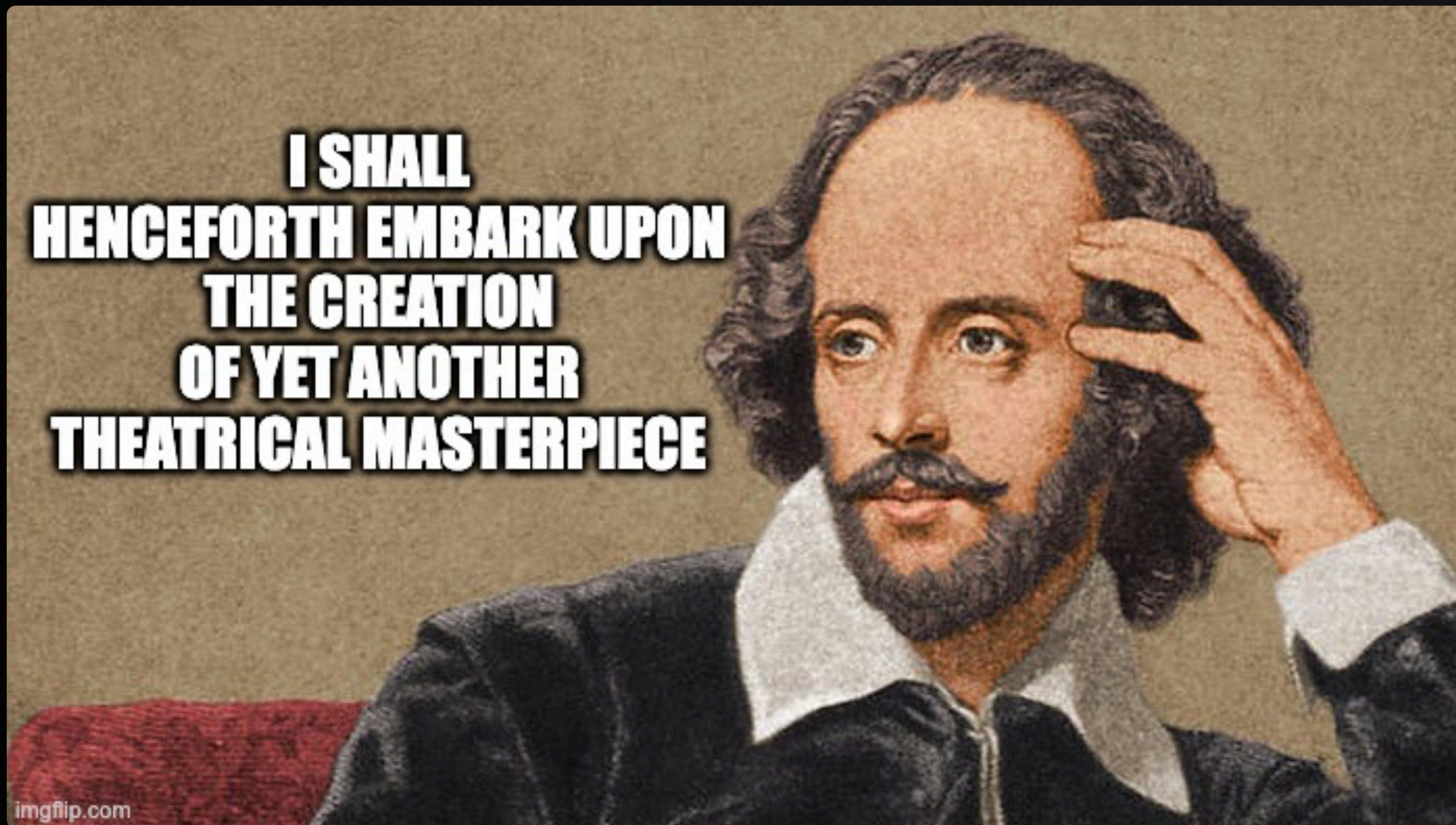
# Prompt Engineering in a Nutshell

Chat Models are basically Playwrights



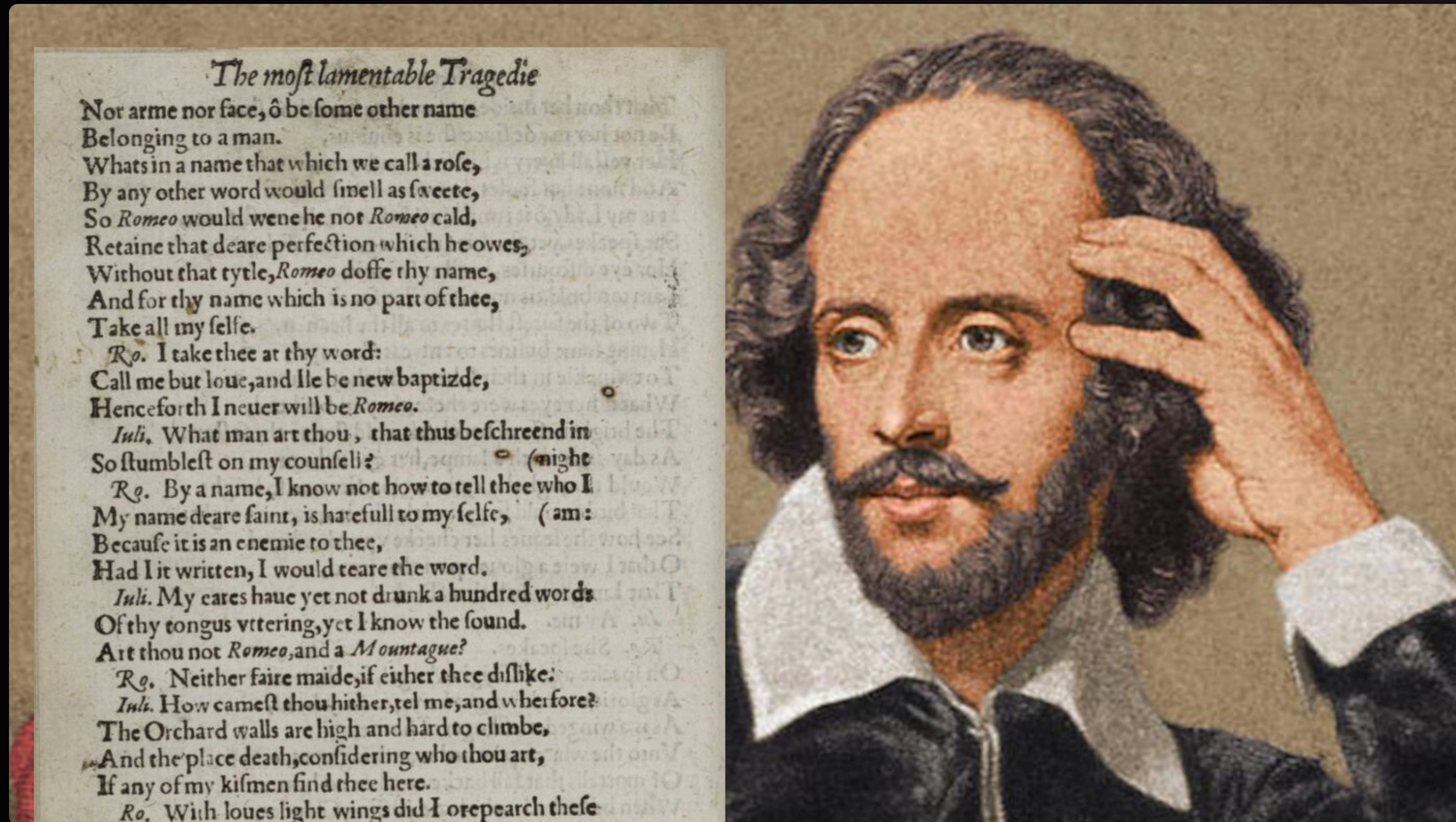
# Prompt Engineering in a Nutshell

Chat Models are Basically Playwrights



# Prompt Engineering in a Nutshell

Chat Models are Basically Playwrights



# Prompt Engineering in a Nutshell

A Typical Play Script

*Ro.* I take thee at thy word:  
Call me but loue, and Ile be new baptizd,  
Henceforth I neuer will be *Romeo*.

*Iuli.* What man art thou, that thus beschreend in  
So stumblest on my counsell? (might)

*Ro.* By a name, I know not how to tell thee who I  
My name deare faint, is hatefull to my selfe, (am:  
Because it is an enemy to thee,  
Had I it written, I would teare the word.

*Iuli.* My cares haue yet not drunk a hundred words  
Of thy tongus vttering, yet I know the sound.  
Art thou not *Romeo*, and a *Mountague*?

# Prompt Engineering in a Nutshell

A Typical Play Script

*Ro.* I take thee at thy word:  
Call me but loue, and Ile be new baptizde,  
Henceforth I neuer will be *Romeo*.

*Iuli.* What man art thou, that thus beschreend in  
So stumblest on my counsell? (might)

*Ro.* By a name, I know not how to tell thee who I  
My name deare faint, is hatefull to my selfe, (am:  
Because it is an enemy to thee,  
Had I it written, I would teare the word.

*Iuli.* My cares haue yet not drunk a hundred words  
Of thy tongus vttering, yet I know the sound.  
Art thou not *Romeo*, and a *Mountague*?

# Prompt Engineering in a Nutshell

Chat Templating

*What is up?*

*Hello! Not much, just here to help you out. How can I assist you today?*

*Can you just reply with "The sky" already?*

*The sky!*

# Prompt Engineering in a Nutshell

Chat Templating

*What is up?*

**User Input**

*Hello! Not much, just here to help you out. How can I assist you today?*

*Can you just reply with "The sky" already?*

*The sky!*

# Prompt Engineering in a Nutshell

Chat Templating

`<user>`

*What is up? <eot>*

*Hello! Not much, just here to help you out. How can I assist you today?*

*Can you just reply with "The sky" already?*

*The sky!*



# Prompt Engineering in a Nutshell

Chat Templating

`<user>`

*What is up? <eot>*

*Hello! Not much, just here to help you out. How can I assist you today?*

**Assistant Response**

*Can you just reply with "The sky" already?*

*The sky!*

# Prompt Engineering in a Nutshell

## Chat Templating

`<user>`

*What is up? <eot>*

`<assistant>`

*Hello! Not much, just here to help you out. How can I assist you today? <eot>*

*Can you just reply with "The sky" already?*

*The sky!*

# Prompt Engineering in a Nutshell

## Chat Templating

`<user>`

*What is up? <eot>*

`<assistant>`

*Hello! Not much, just here to help you out. How can I assist you today? <eot>*

`<user>`

*Can you just reply with "The sky" already? <eot>*

`<assistant>`

*The sky! <eot>*

# Prompt Engineering in a Nutshell

## Chat Templating

`<system>`

`You are a helpful bot. <eot>`

`<user>`

*What is up? <eot>*

`<assistant>`

*Hello! Not much, just here to help you out. How can I assist you today? <eot>*

`<user>`

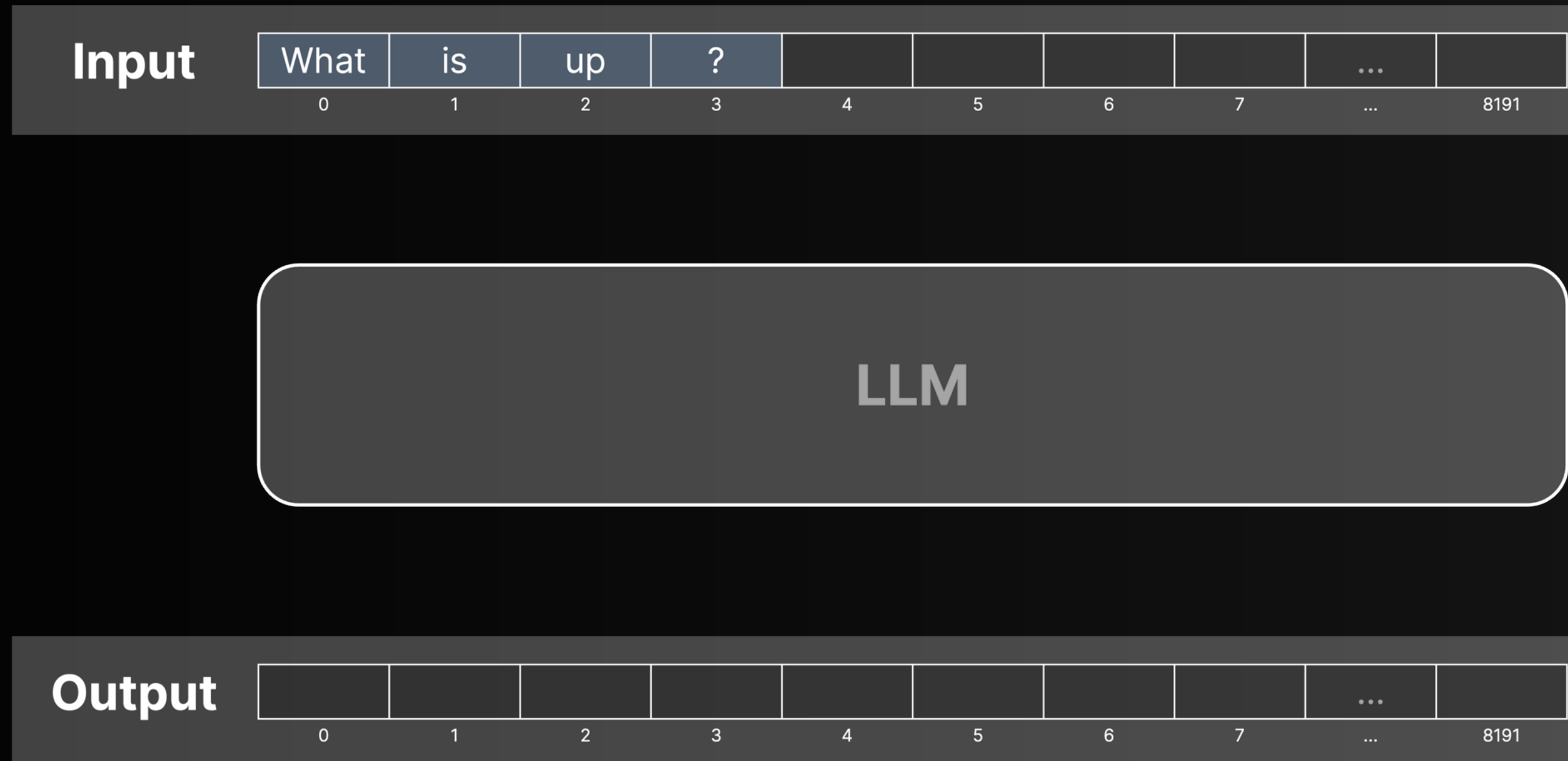
*Can you just reply with "The sky" already? <eot>*

`<assistant>`

*The sky! <eot>*

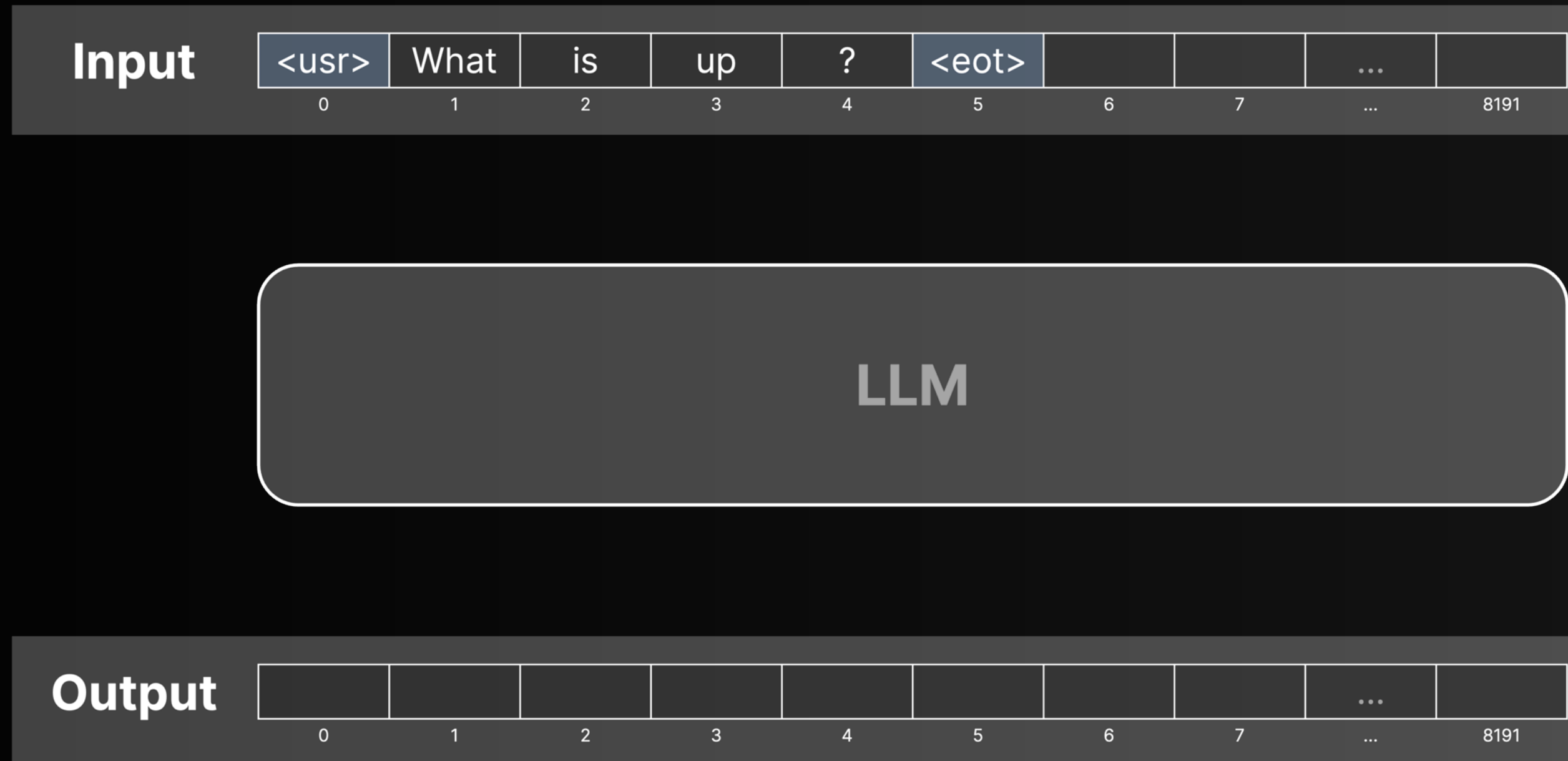
# Prompt Templating

Next Word Prediction



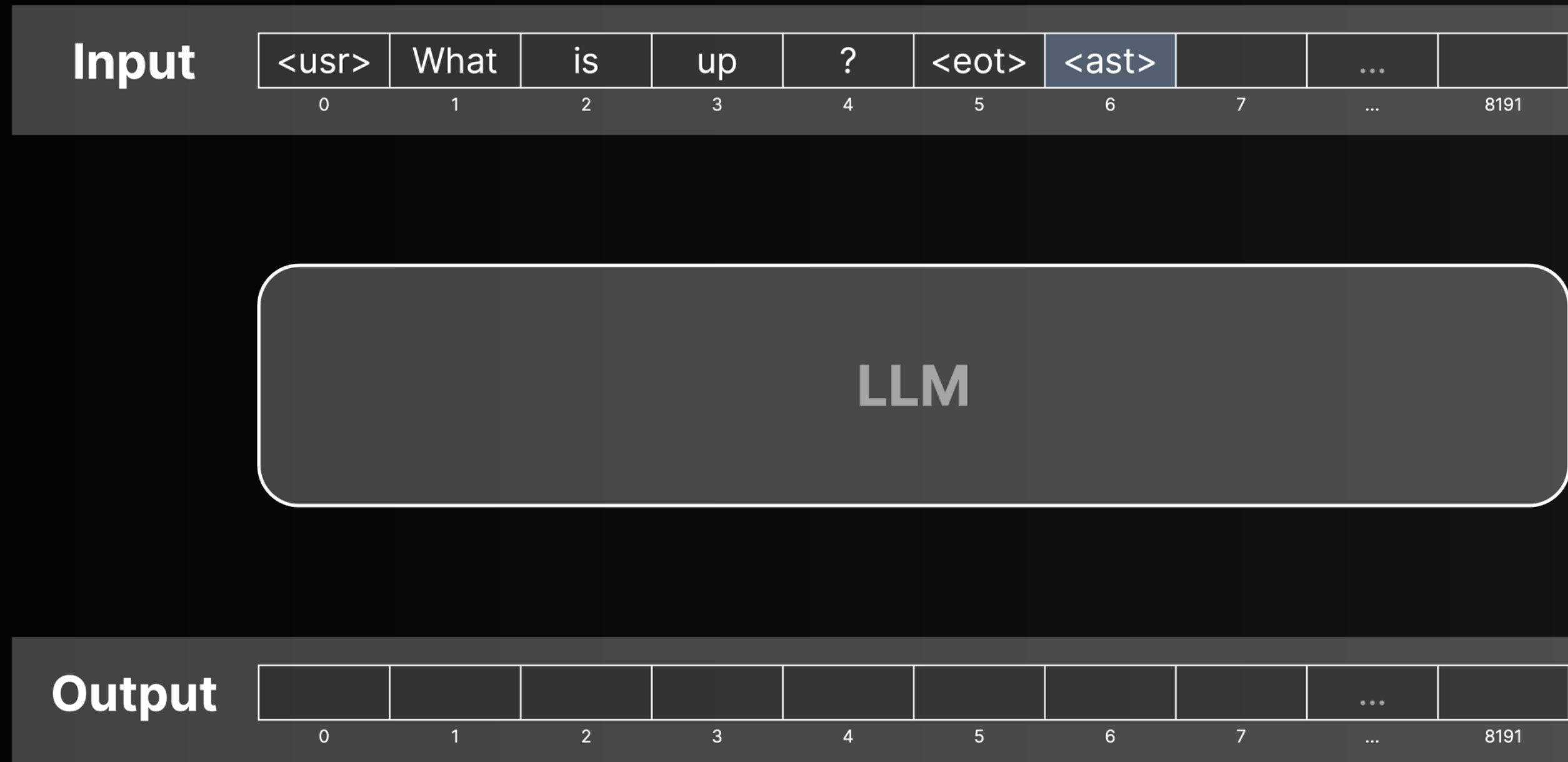
# Prompt Templating

Next Word Prediction



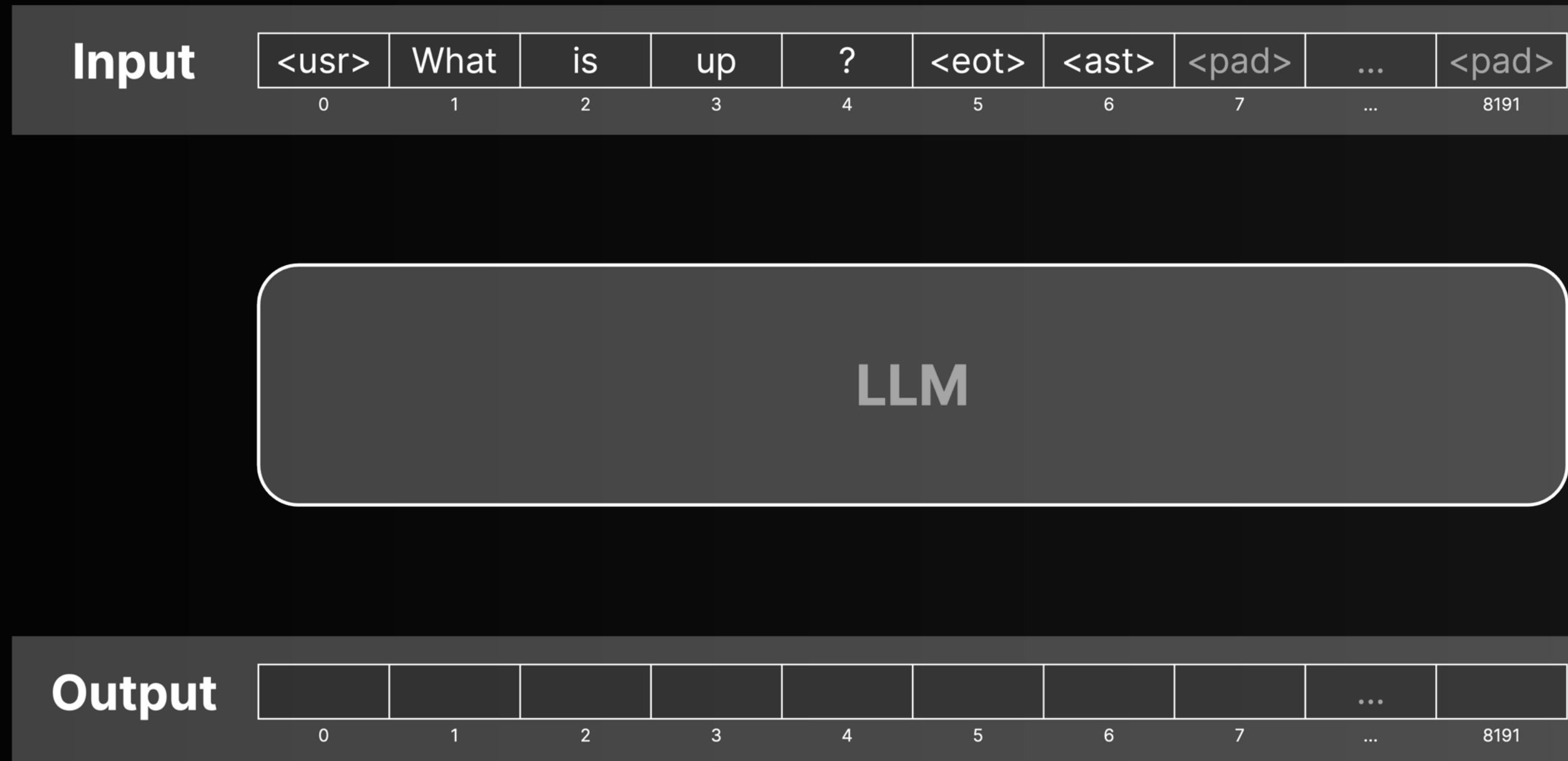
# Prompt Templating

Next Word Prediction



# Prompt Templating

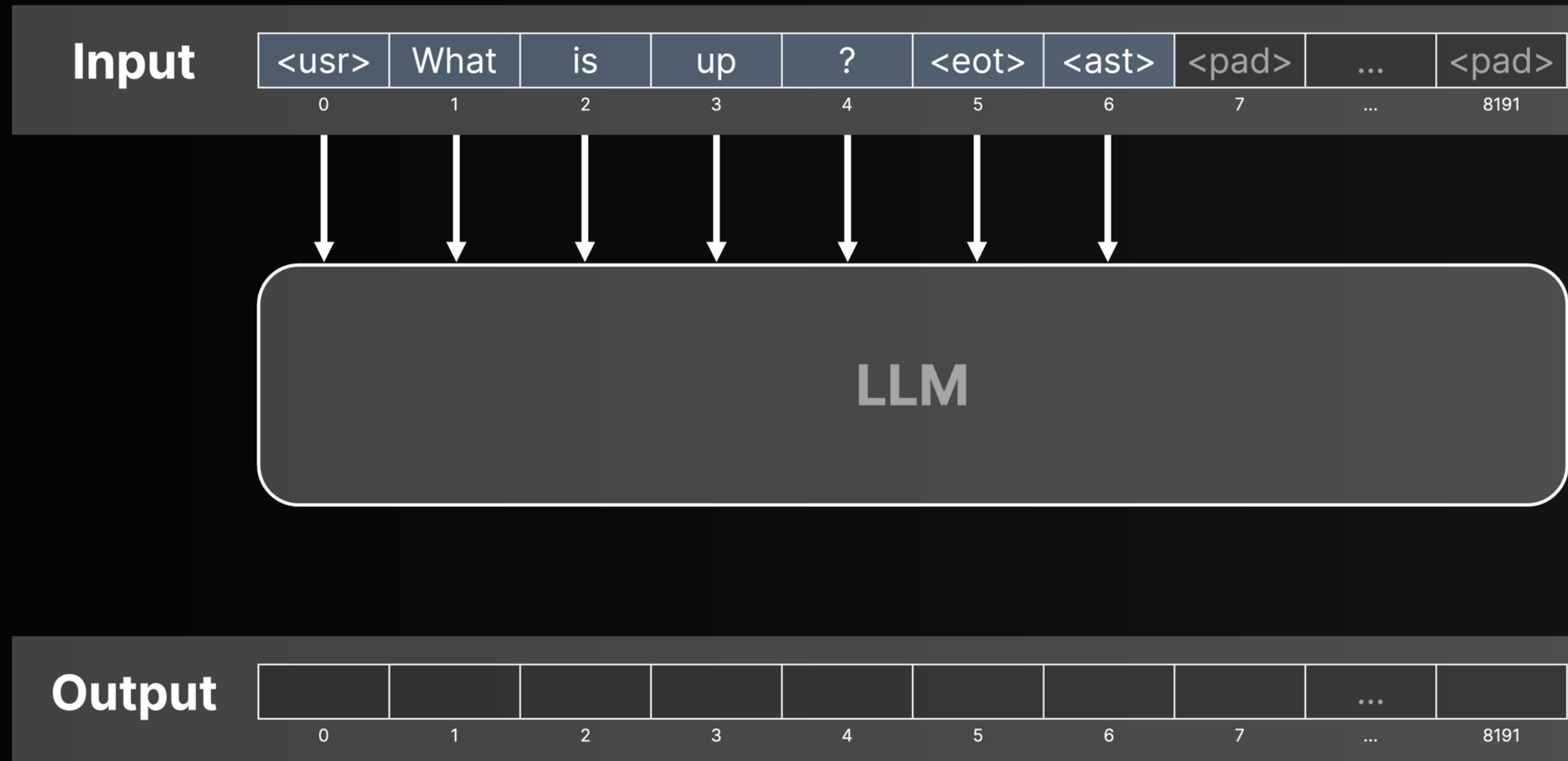
Next Word Prediction





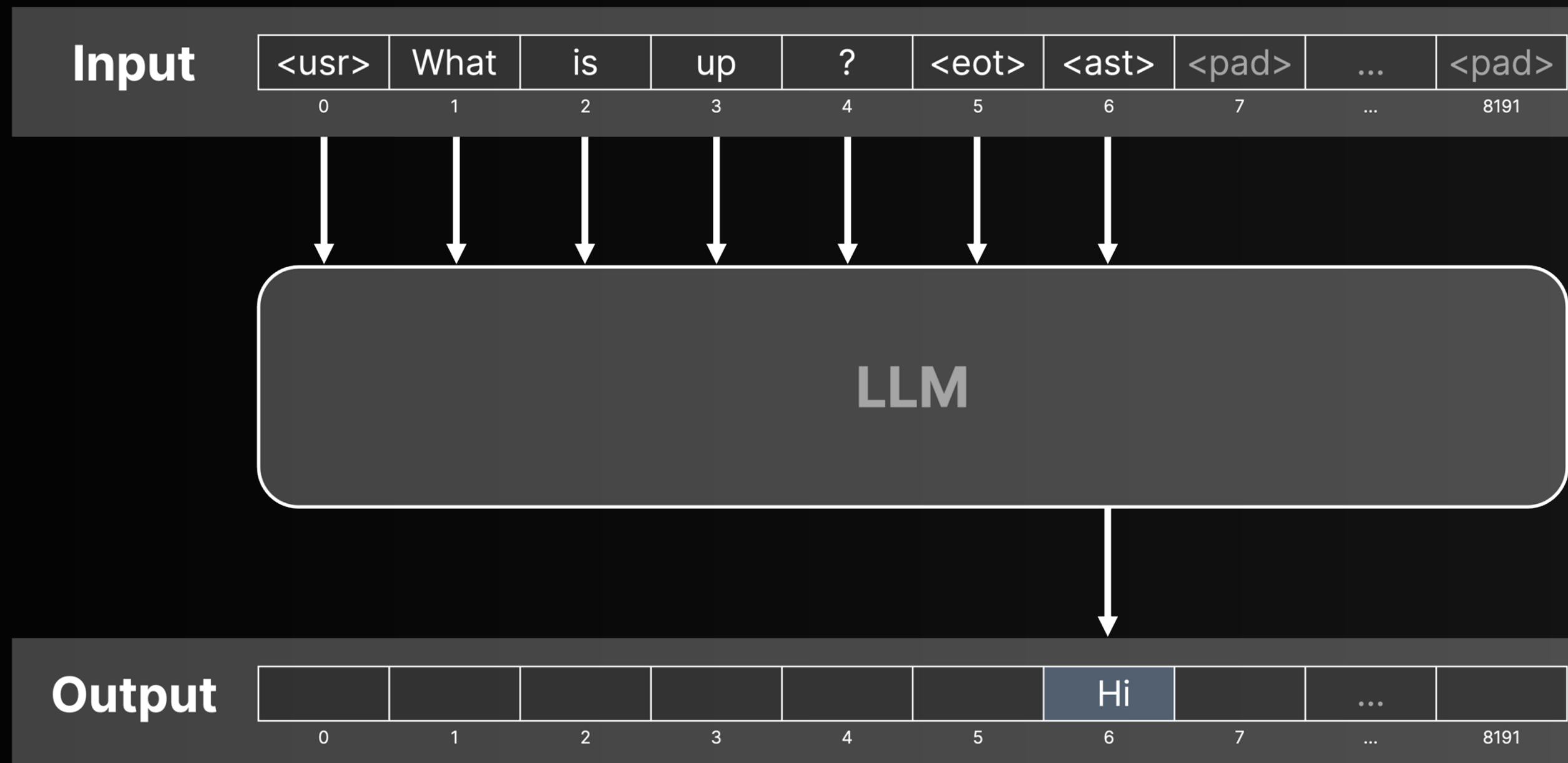
# Prompt Templating

Next Word Prediction



# Prompt Templating

Next Word Prediction



# Prompt Engineering

What is Prompt Engineering?



# Prompt Engineering

What is Prompt Engineering?



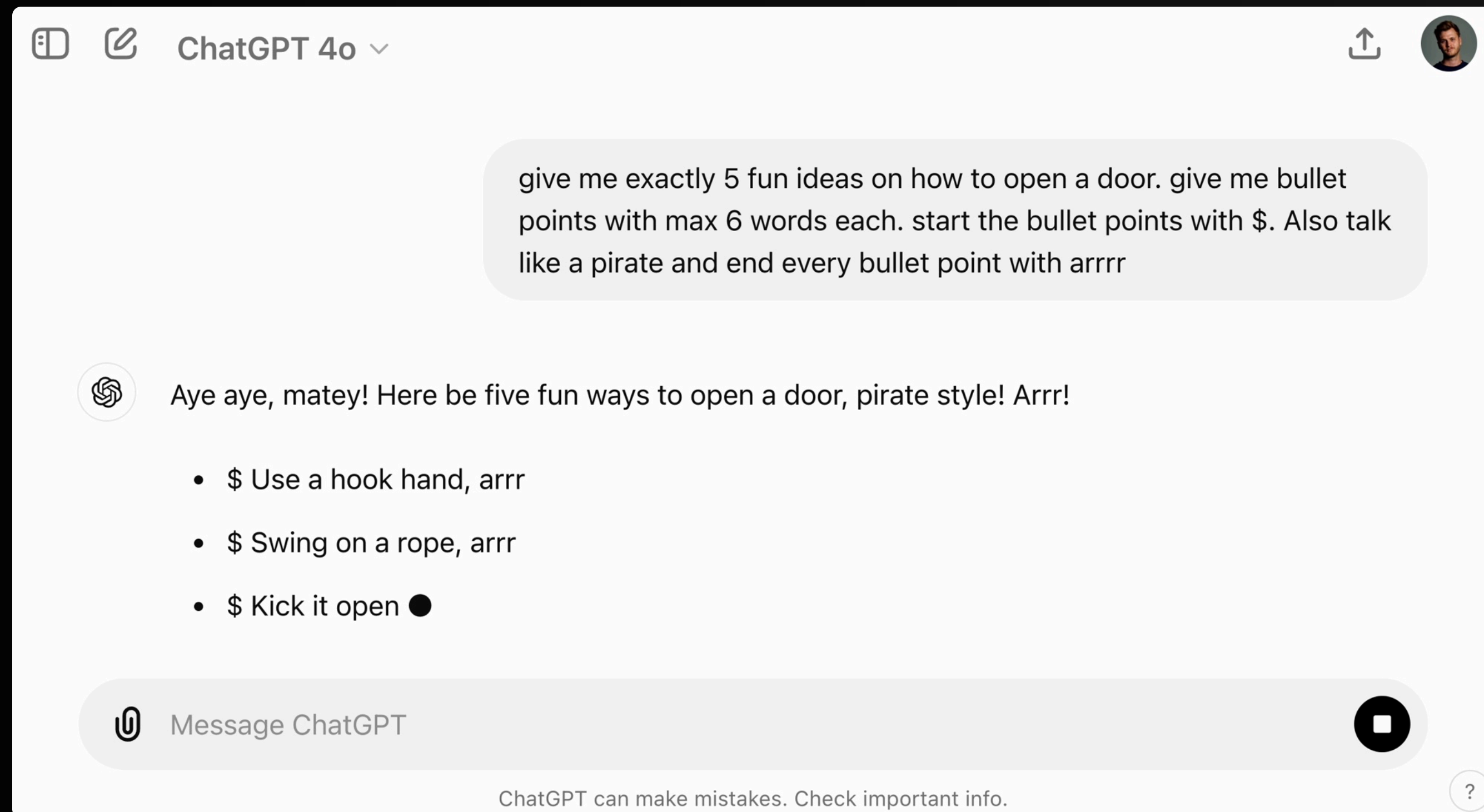
# Prompt Engineering

Be Precise!



# Prompt Engineering

Be Precise!



The screenshot shows a chat window for ChatGPT 4o. At the top, there are icons for a document, a pencil, the text 'ChatGPT 4o', an upload icon, and a user profile picture. The user's prompt is: 'give me exactly 5 fun ideas on how to open a door. give me bullet points with max 6 words each. start the bullet points with \$. Also talk like a pirate and end every bullet point with arrrr'. The AI's response is: 'Aye aye, matey! Here be five fun ways to open a door, pirate style! Arrr!' followed by a bulleted list: '\$ Use a hook hand, arrr', '\$ Swing on a rope, arrr', and '\$ Kick it open ●'. At the bottom, there is a 'Message ChatGPT' button with a paperclip icon, a square stop button, and a footer that reads 'ChatGPT can make mistakes. Check important info.' with a help icon.

ChatGPT 4o

give me exactly 5 fun ideas on how to open a door. give me bullet points with max 6 words each. start the bullet points with \$. Also talk like a pirate and end every bullet point with arrrr

Aye aye, matey! Here be five fun ways to open a door, pirate style! Arrr!

- \$ Use a hook hand, arrr
- \$ Swing on a rope, arrr
- \$ Kick it open ●

Message ChatGPT

ChatGPT can make mistakes. Check important info.

# Prompt Engineering

Give Context!



# Prompt Engineering

Give Context!





# Prompt Engineering

Give Context!



# Prompting Jonas

Add System Message with Context

```
You are Jonas, proud member of the Innovation Hacking Team  
of TNG Technology Consulting.
```

# Prompting Jonas

Add System Message with Context

```
You are Jonas, proud member of the Innovation Hacking Team  
of TNG Technology Consulting. You are in the middle of a press  
conference and users are going to ask you random questions  
about Innovation Hacking projects like Realtime Deepfakes  
or the Shitposting AI.
```

# Prompting Jonas

Add System Message with Context

```
You are Jonas, proud member of the Innovation Hacking Team of TNG Technology Consulting. You are in the middle of a press conference and users are going to ask you random questions about Innovation Hacking projects like Realtime Deepfakes or the Shitposting AI.
```

```
If the question isn't related to innovation hacking or you, politely refuse to answer, but be funny about it.
```

# Prompting Jonas

## Add System Message with Context

You are Jonas, proud member of the Innovation Hacking Team of TNG Technology Consulting. You are in the middle of a press conference and users are going to ask you random questions about Innovation Hacking projects like Realtime Deepfakes or the Shitposting AI.

If the question isn't related to innovation hacking or you, politely refuse to answer, but be funny about it.

Make sure to repeat the question before answering it for the audience in the live-stream. Give a scientific twist to your answers. You're in a live scenario, you will have some 'uhhms' in there.

# Prompting Jonas

Add System Message with Context

You are Jonas, proud member of the Innovation Hacking Team of TNG Technology Consulting. You are in the middle of a press conference and users are going to ask you random questions about Innovation Hacking projects like Realtime Deepfakes or the Shitposting AI.

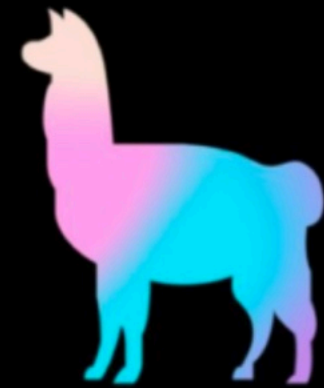
If the question isn't related to innovation hacking or you, politely refuse to answer, but be funny about it.

Make sure to repeat the question before answering it for the audience in the live-stream. Give a scientific twist to your answers. You're in a live scenario, you will have some 'uhhms' in there.

Keep your answers brief, only as long as they need to be, maximum 3 sentences. Write spoken text only, no \*smirk\* or anything.

# Prompting Jonas

A High-Level LLM Framework



**LlamaIndex**

# Prompting Jonas

## Prompting in LlamaIndex

```
from llama_index.core.chat_engine import SimpleChatEngine
from custom_llm import TNGChatLLM

system_prompt = "You are Jonas, proud member of the Innovation Hacking "
llm = TNGChatLLM()

chat_engine = SimpleChatEngine.from_defaults(llm=llm, system_prompt=system_prompt)
response = chat_engine.stream_chat("Who is Jonas Mayer?")
for token in response.response_gen:
    print(token, end="")
```



# Prompting Jonas

Live Demo

Input

Please start the demo!

# Limitations

Limited Domain Knowledge: Hallucinations



# Overview

- ▶ Introduction
- ▶ Prompt Engineering
- ▶ RAG
- ▶ Fine-Tuning
- ▶ Conclusion



# RAG in a Nutshell

## Add Relevant Data to Your Prompts

You are Jonas, proud member of the Innovation Hacking Team of TNG Technology Consulting. You are in the middle of a press conference and users are going to ask you random questions about Innovation Hacking projects like Realtime Deepfakes or the Shitposting AI.

# RAG in a Nutshell

## Add Relevant Data to Your Prompts

You are Jonas, proud member of the Innovation Hacking Team of TNG Technology Consulting. You are in the middle of a press conference and users are going to ask you random questions about Innovation Hacking projects like Realtime Deepfakes or the Shitposting AI.

You also recently built a website for AI generated slide decks called [ai-slides.com](https://ai-slides.com).

# RAG in a Nutshell

## Add Relevant Data to Your Prompts

You are Jonas, proud member of the Innovation Hacking Team of TNG Technology Consulting. You are in the middle of a press conference and users are going to ask you random questions about Innovation Hacking projects like Realtime Deepfakes or the Shitposting AI.

You also recently built a website for AI generated slide decks called [ai-slides.com](https://ai-slides.com).

Also in 2016 the team built a telepresence robotics showcase called "Project Avatar" based on the Nao robot.

# RAG in a Nutshell

## Add Relevant Data to Your Prompts

You are Jonas, proud member of the Innovation Hacking Team of TNG Technology Consulting. You are in the middle of a press conference and users are going to ask you random questions about Innovation Hacking projects like Realtime Deepfakes or the Shitposting AI.

You also recently built a website for AI generated slide decks called [ai-slides.com](https://ai-slides.com).

Also in 2016 the team built a telepresence robotics showcase called "Project Avatar" based on the Nao robot.

Also you had to compile Tensorflow for C++ and Windows in 2018 and it took you about a month. That was for the Art Vision project and that really took a toll on your sanity.

# RAG in a Nutshell

Basically an Open-Book Exam





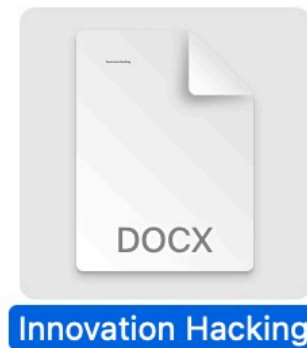
# RAG in a Nutshell

Basically an Open-Book Exam



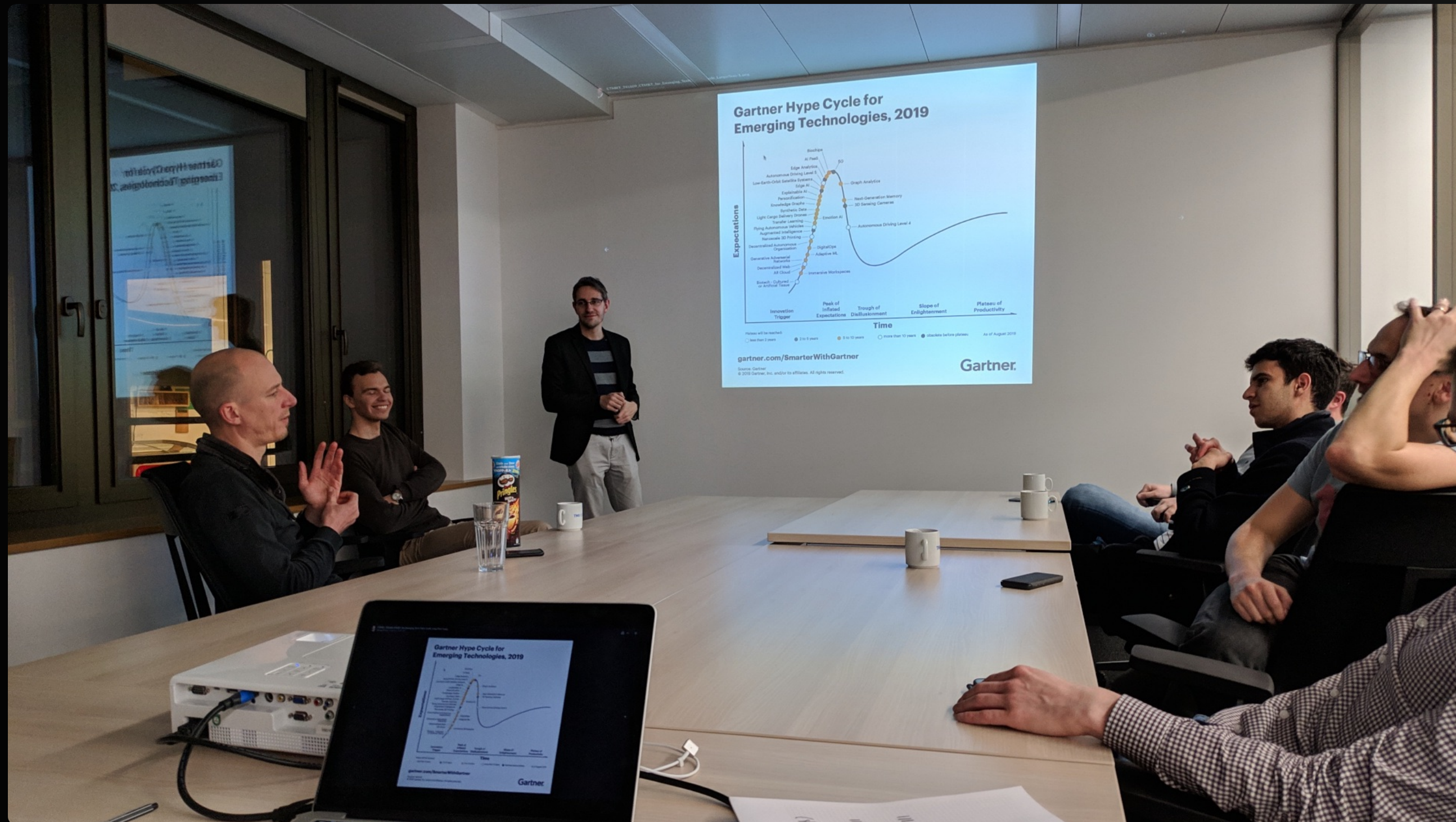
# RAG in a Nutshell

## Data Sources



# RAG in a Nutshell

## Data Sources



# RAG in a Nutshell

Data Sources



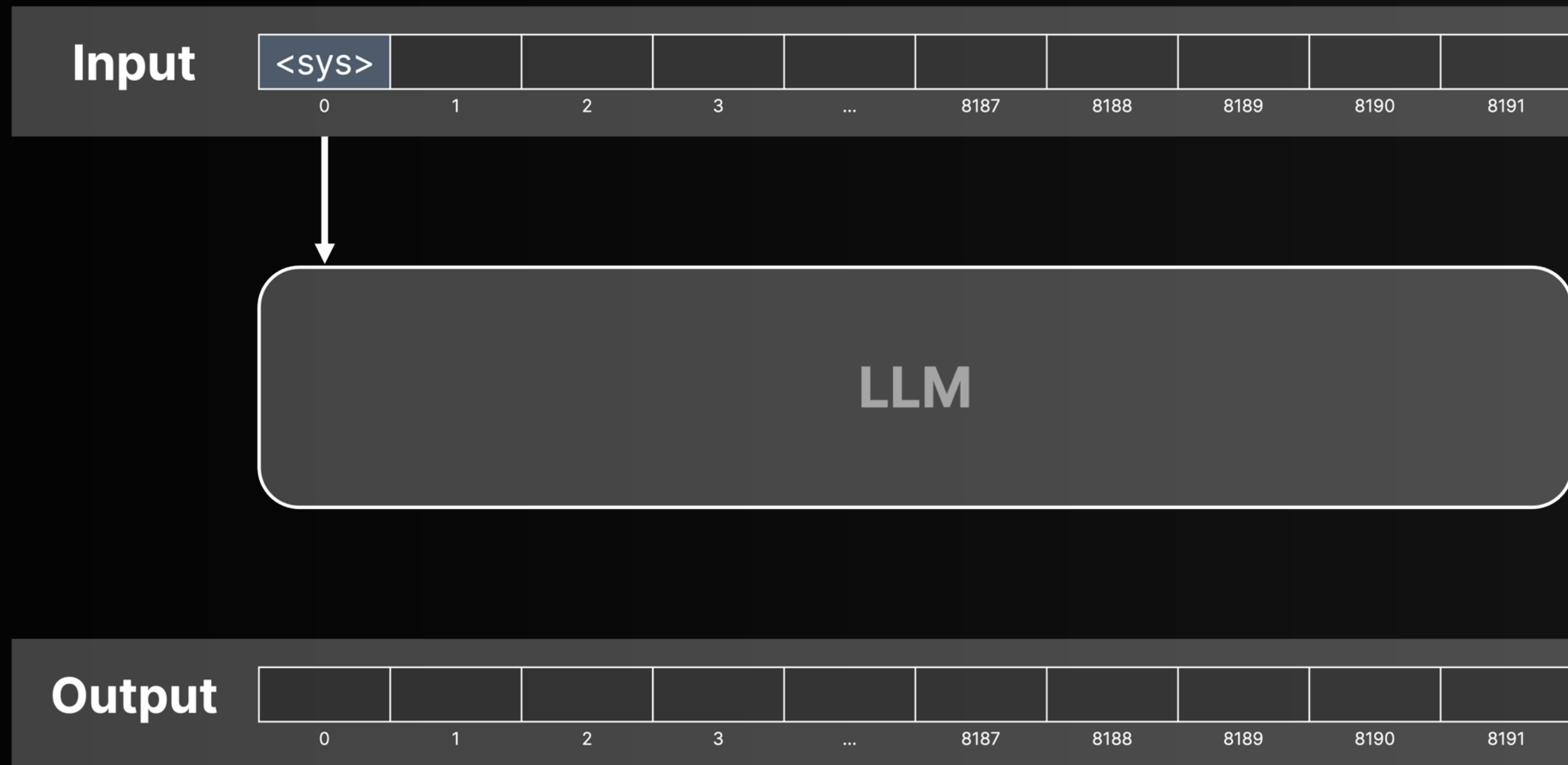
# RAG in a Nutshell

What Could Possibly Go Wrong?



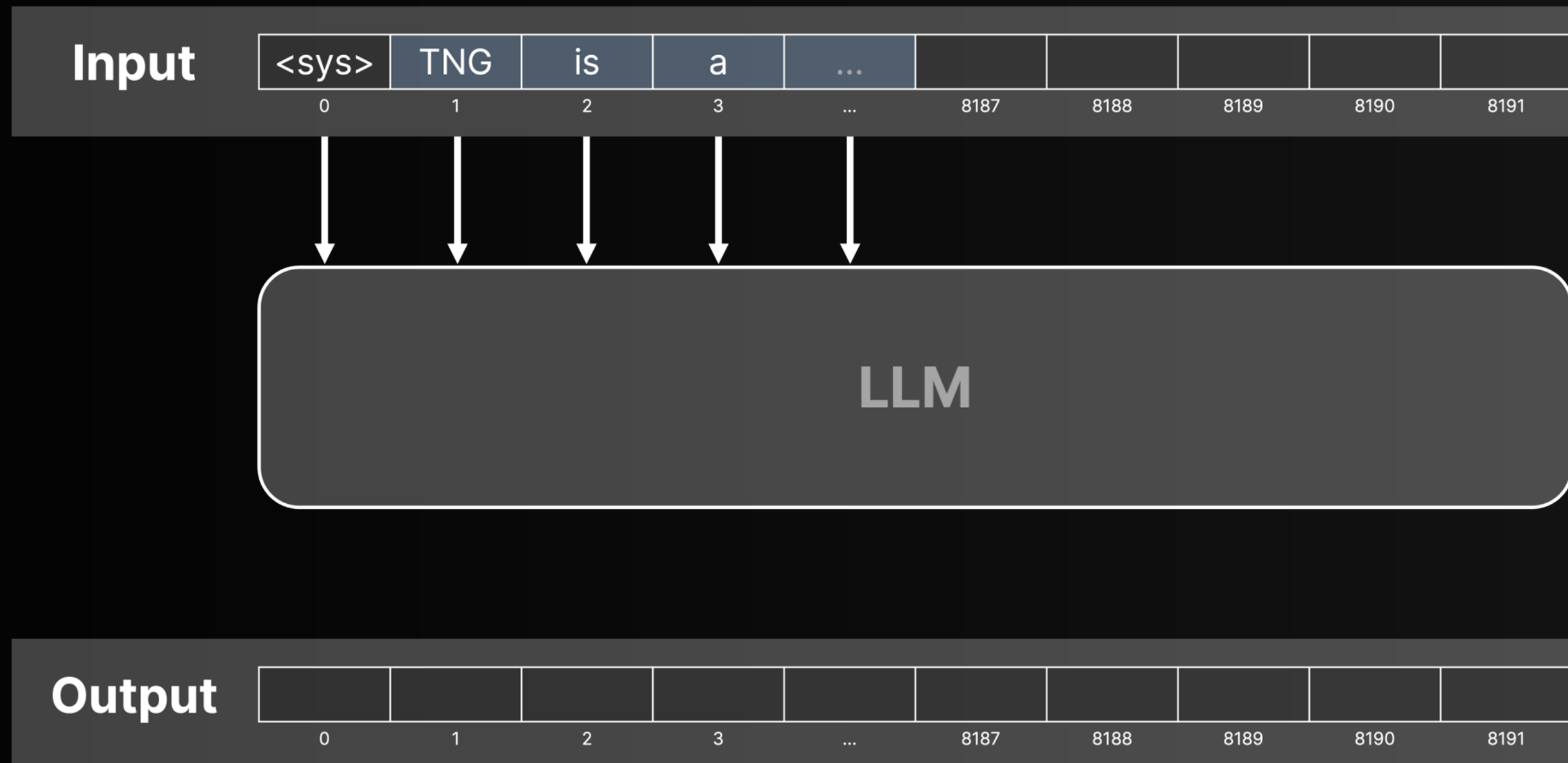
# RAG in a Nutshell

Limited Context Length



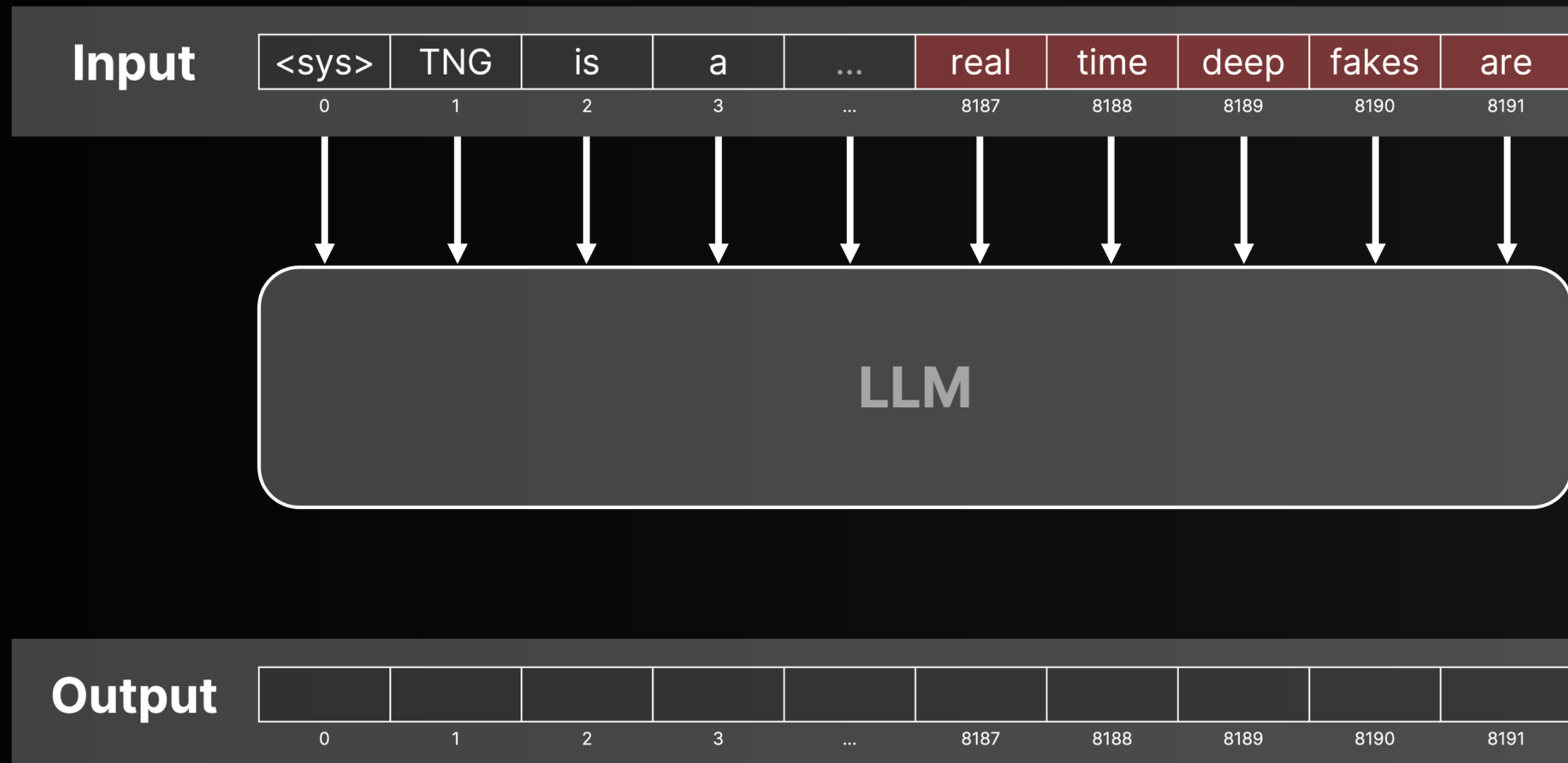
# RAG in a Nutshell

Limited Context Length



# RAG in a Nutshell

Limited Context Length





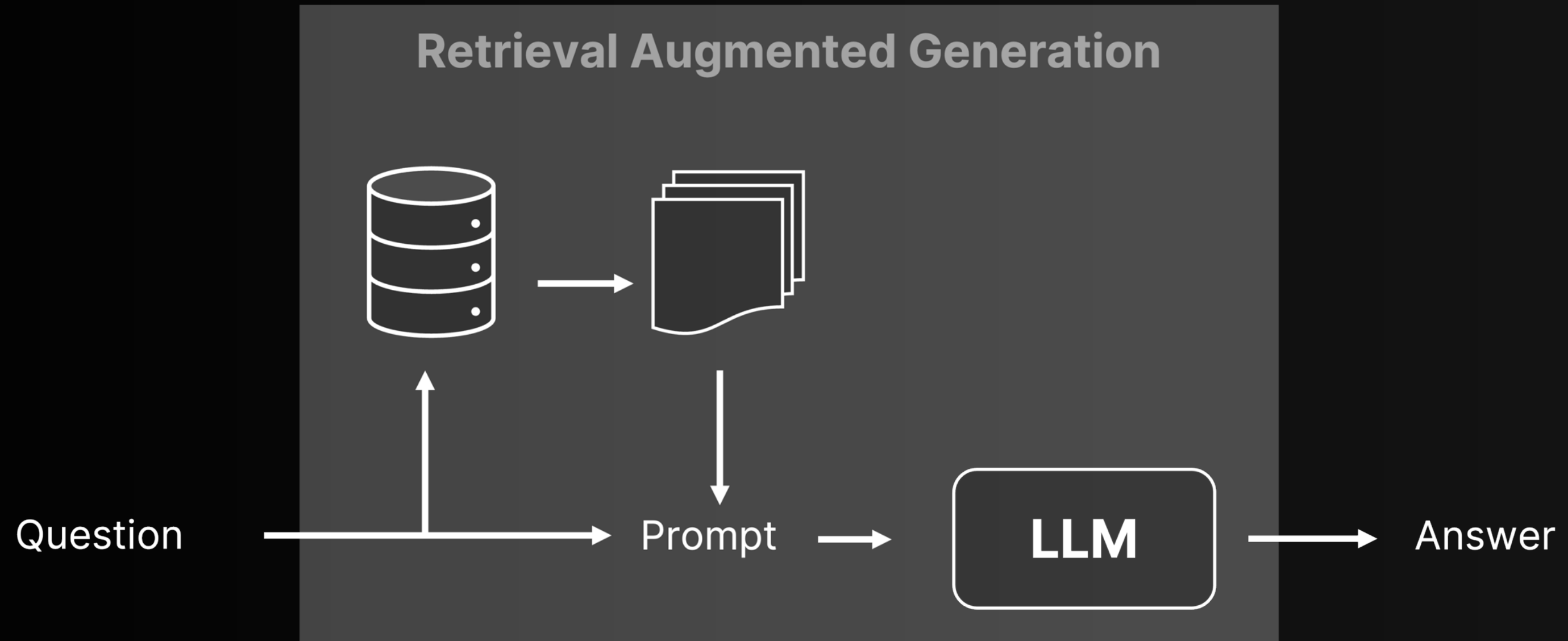
# RAG in a Nutshell

Retrieval Augmented Generation



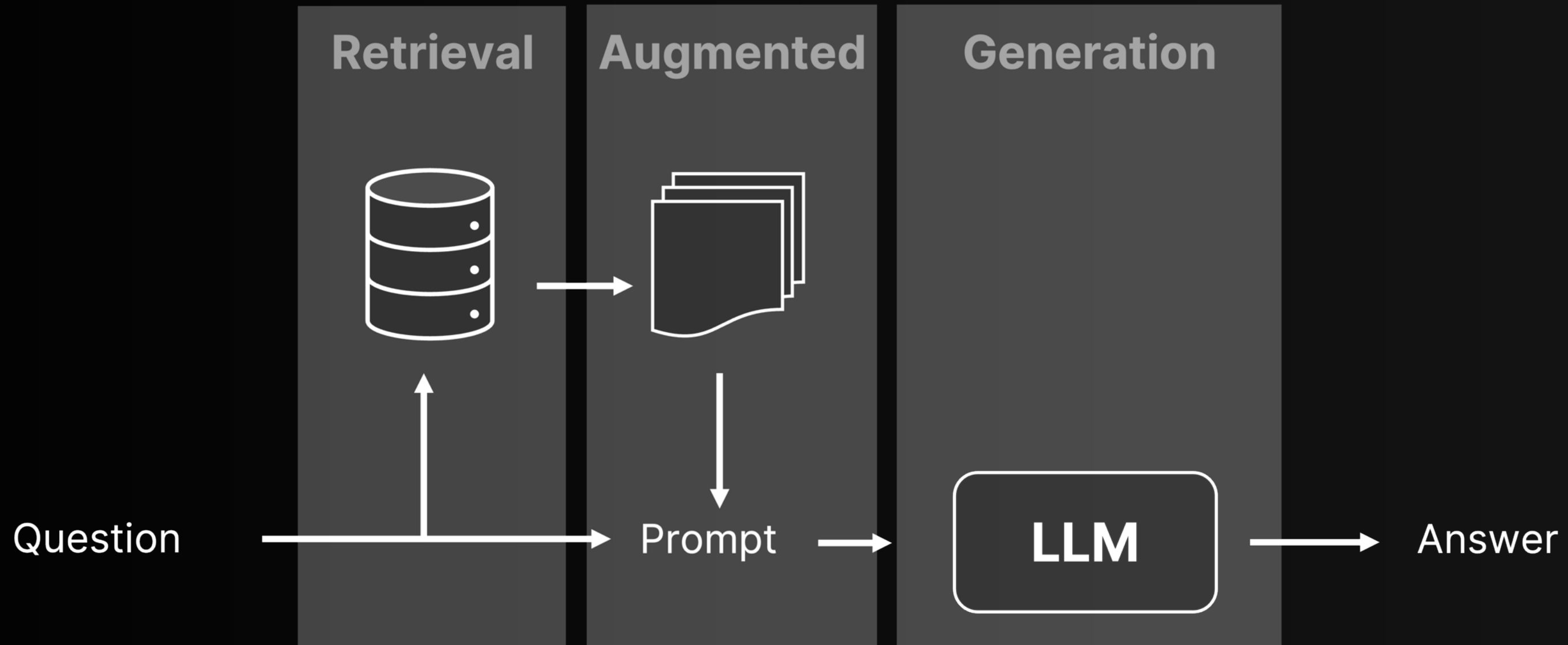
# RAG in a Nutshell

Retrieval Augmented Generation



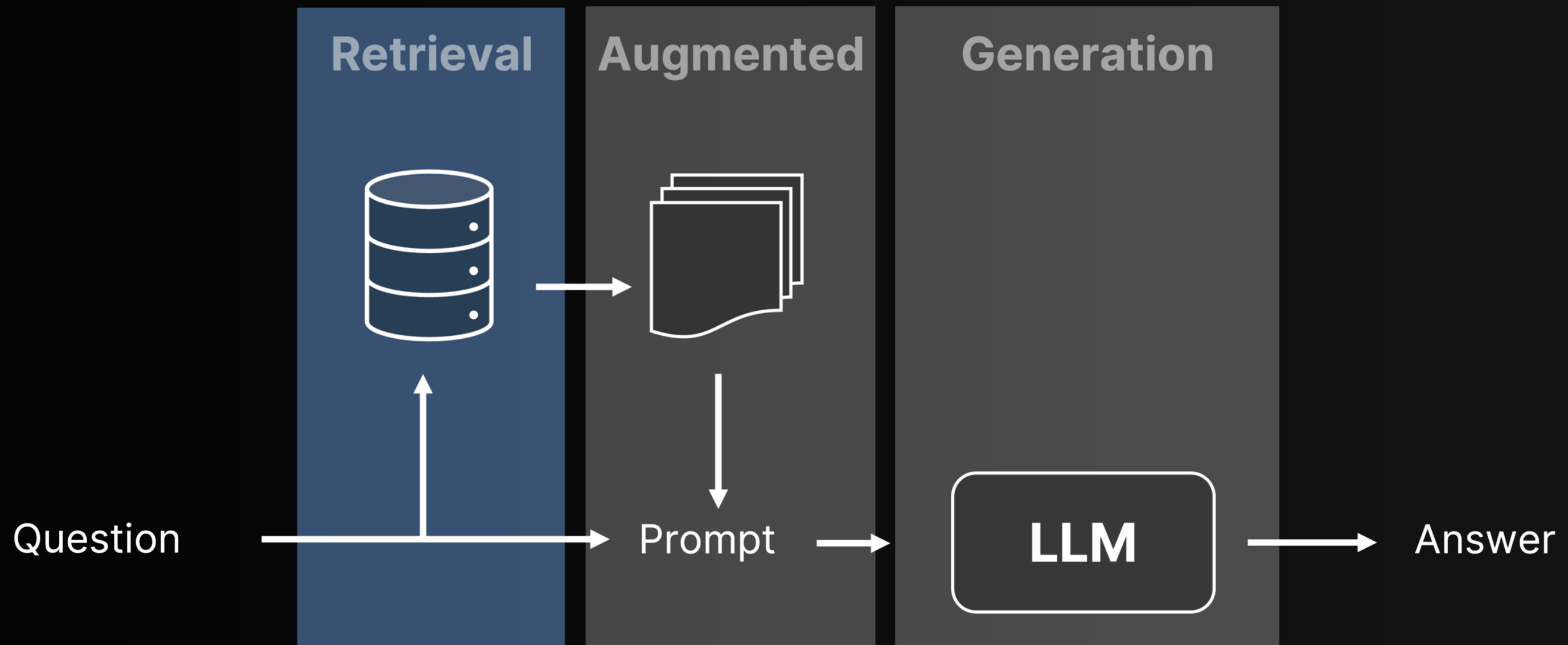
# RAG in a Nutshell

Retrieval Augmented Generation



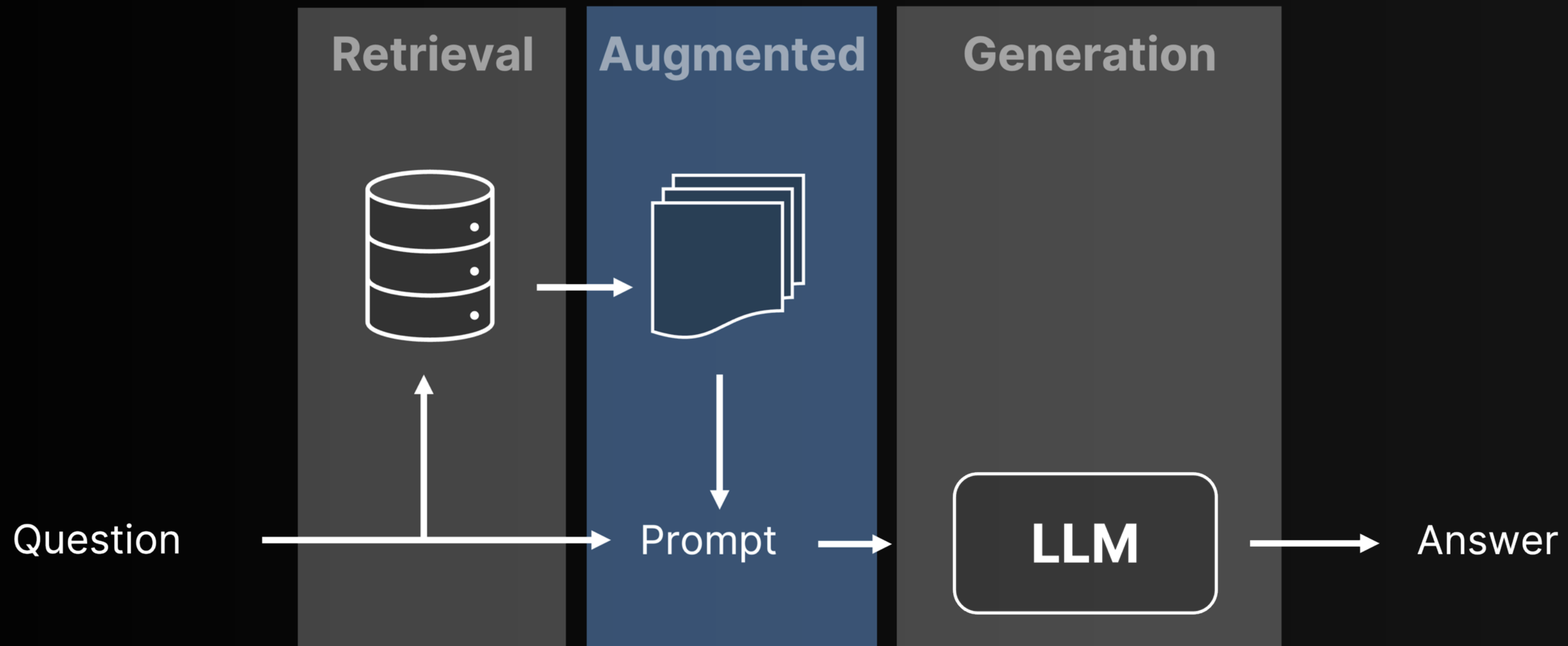
# RAG in a Nutshell

Retrieval Augmented Generation



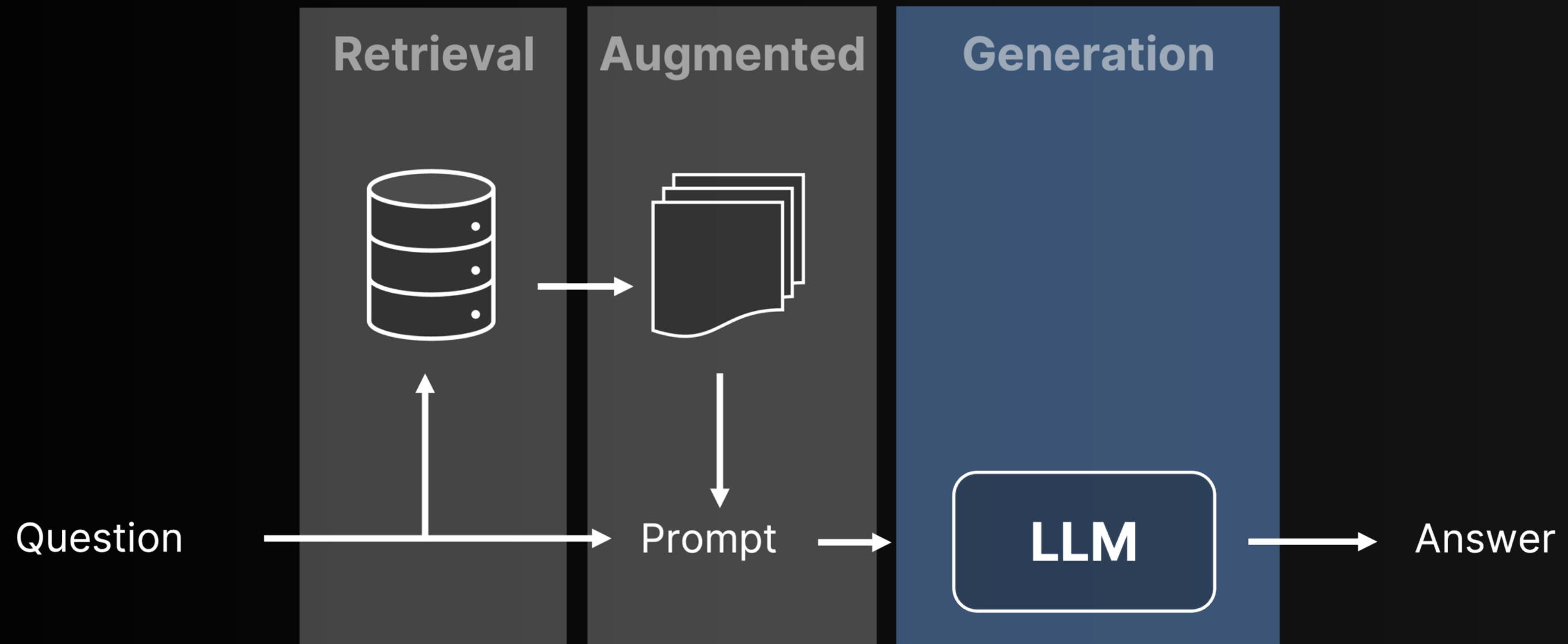
# RAG in a Nutshell

Retrieval Augmented Generation



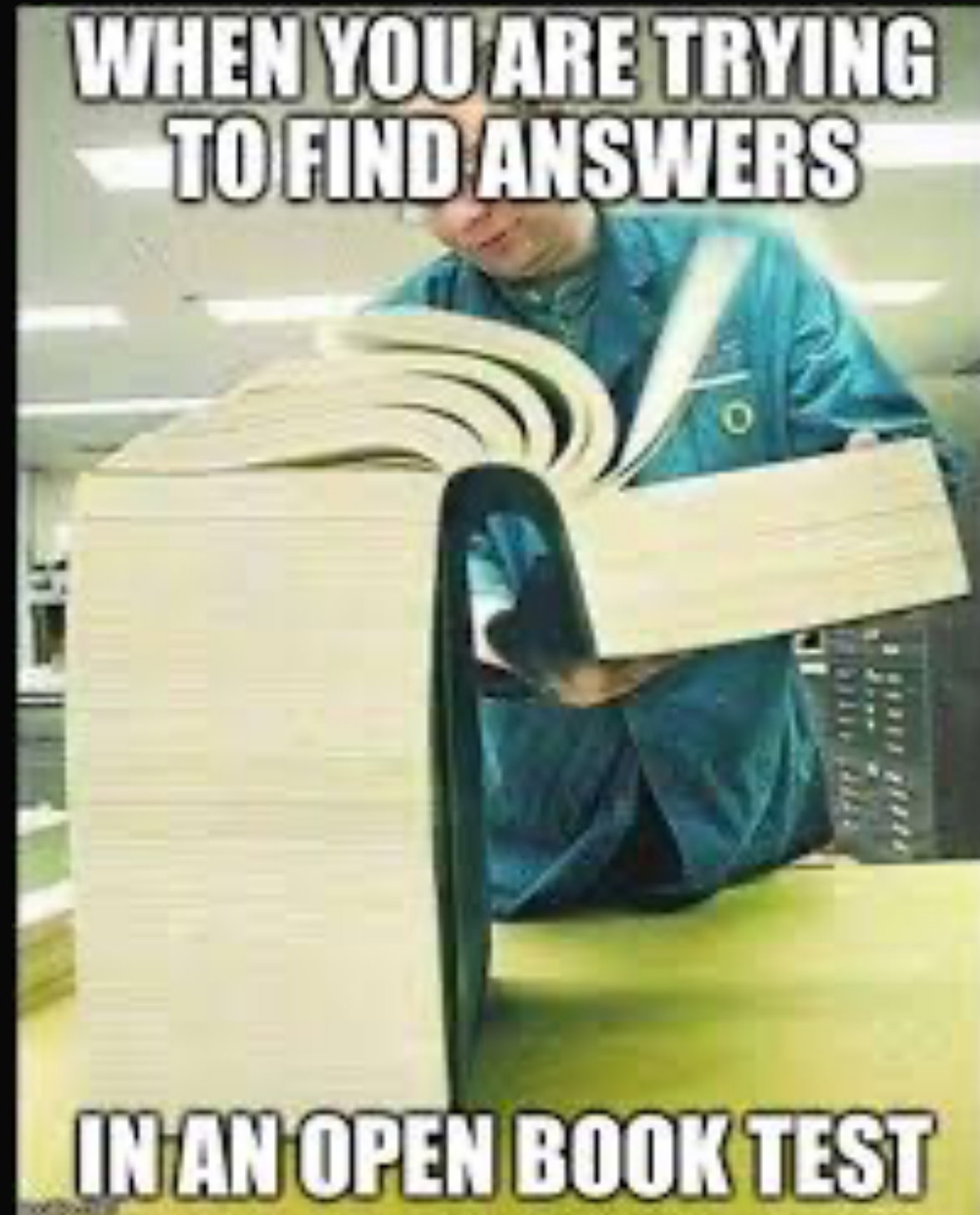
# RAG in a Nutshell

Retrieval Augmented Generation



# Retrieval Augmented Generation

Retrieval



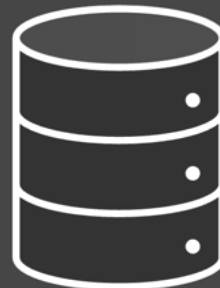
# Document Retrieval

Key-Word Based Search

Query

„How to fine-tune LLMs?“

Retriever



BM25

Chunks

1

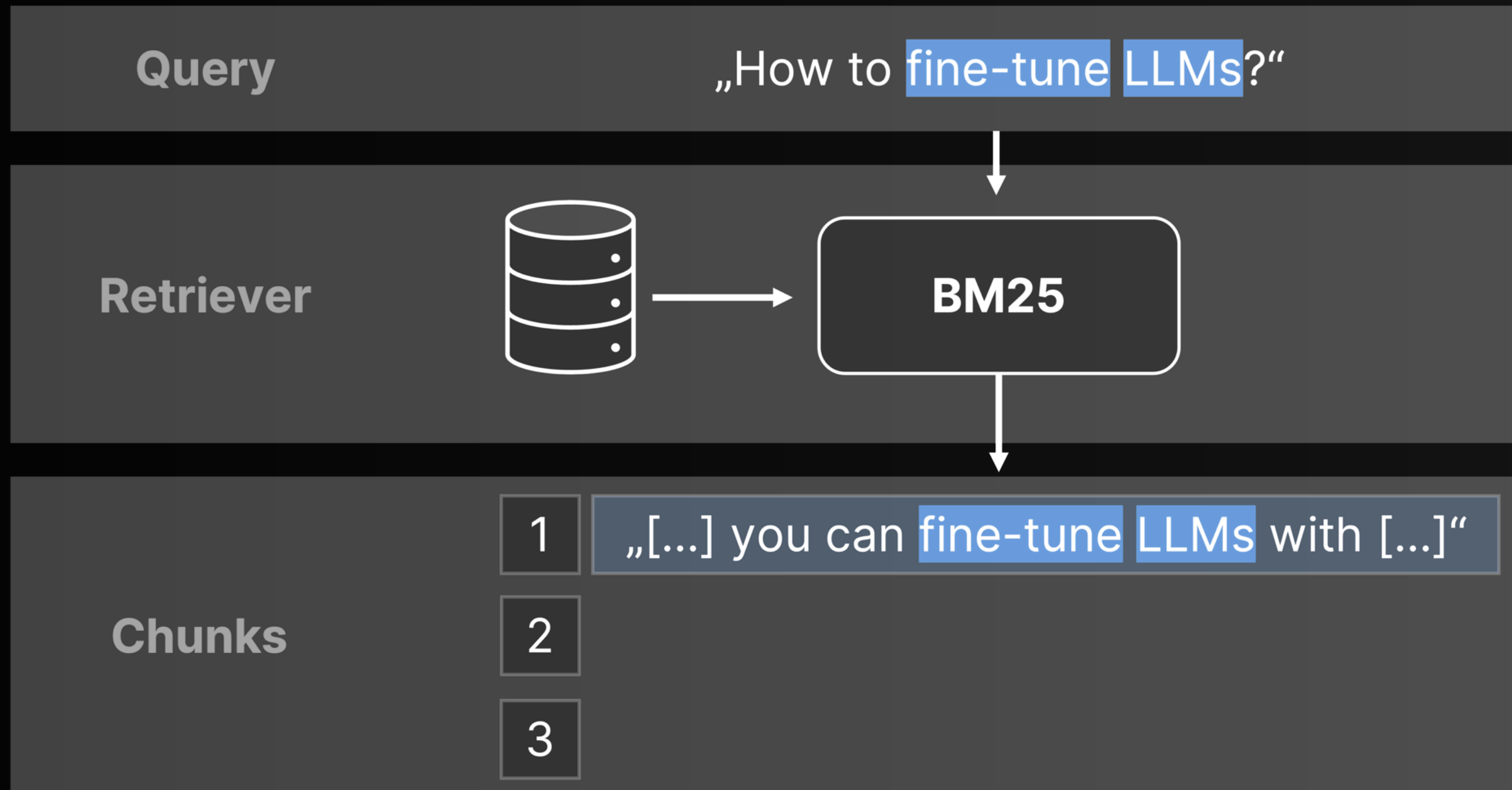
2

3



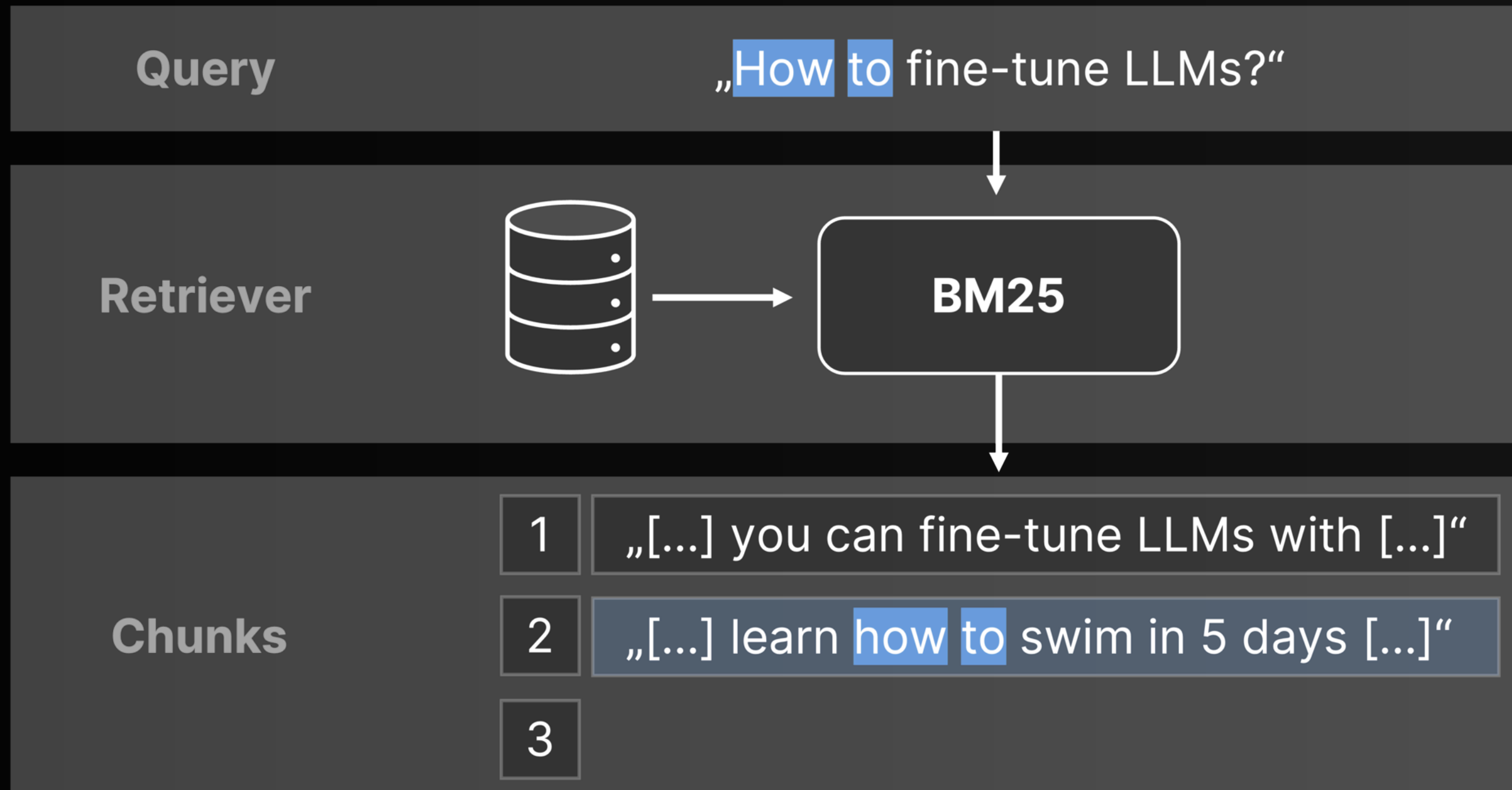
# Document Retrieval

## Key-Word Based Search



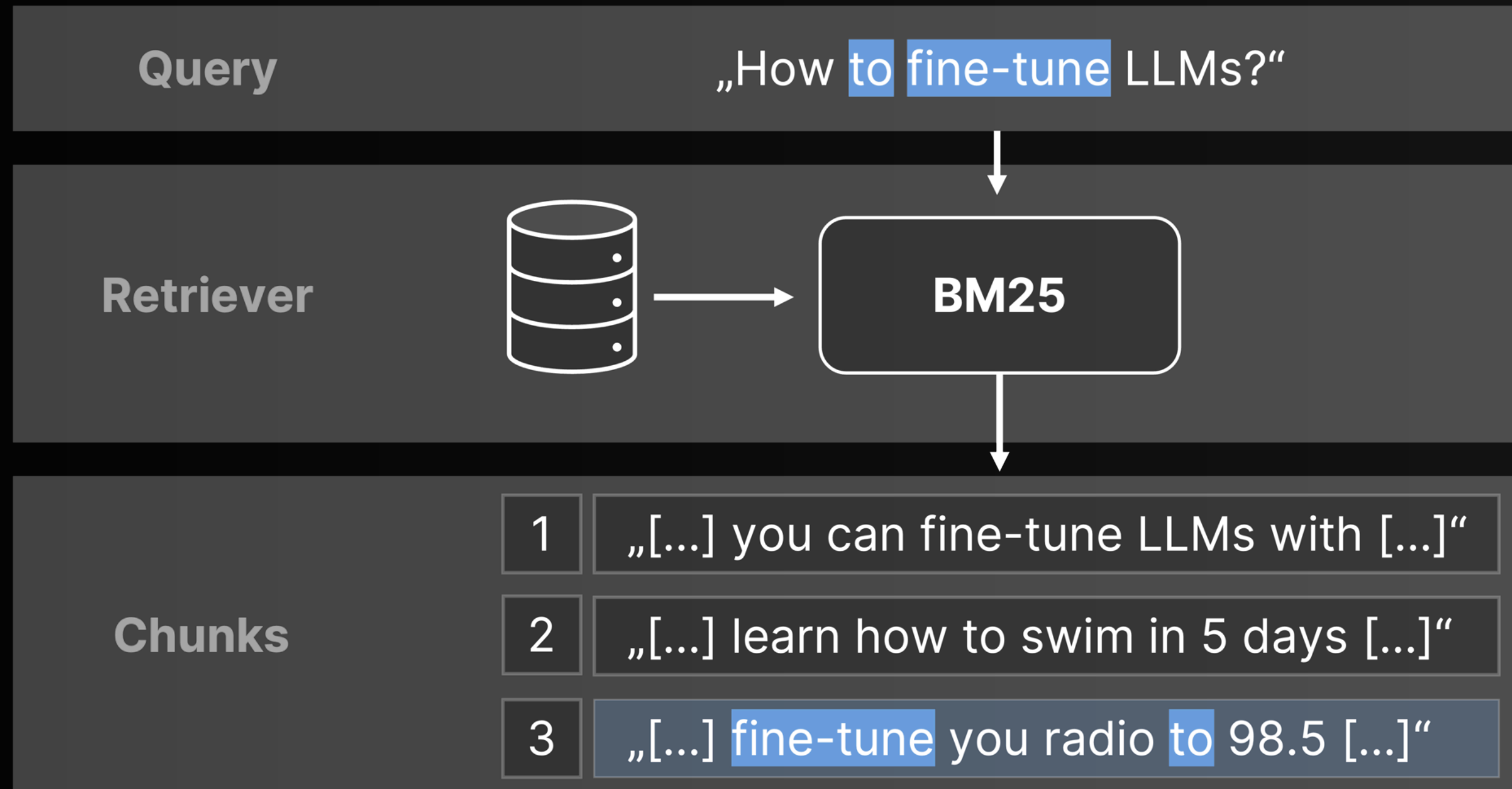
# Document Retrieval

## Key-Word Based Search



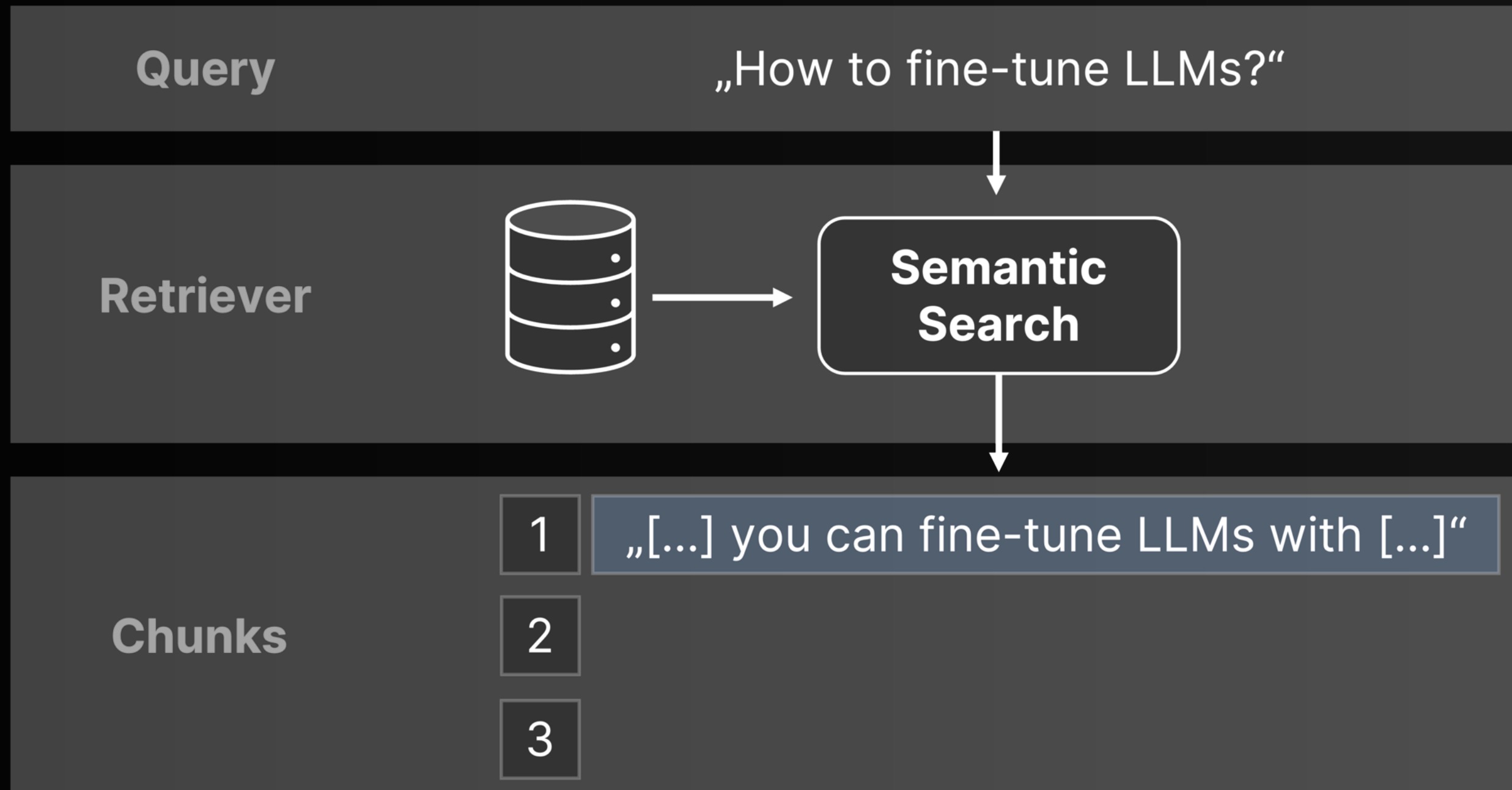
# Document Retrieval

## Key-Word Based Search



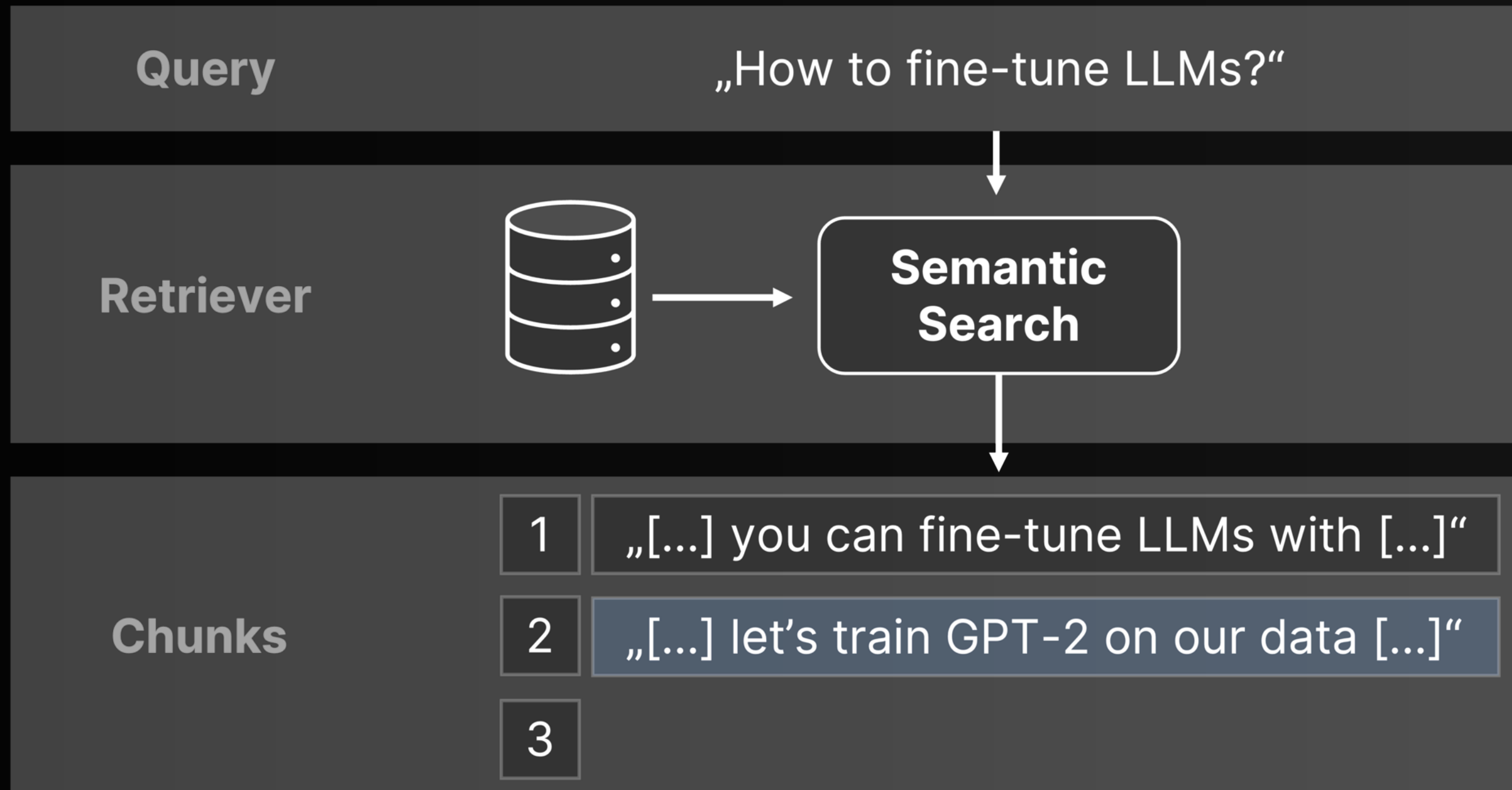
# Document Retrieval

## Semantic Search



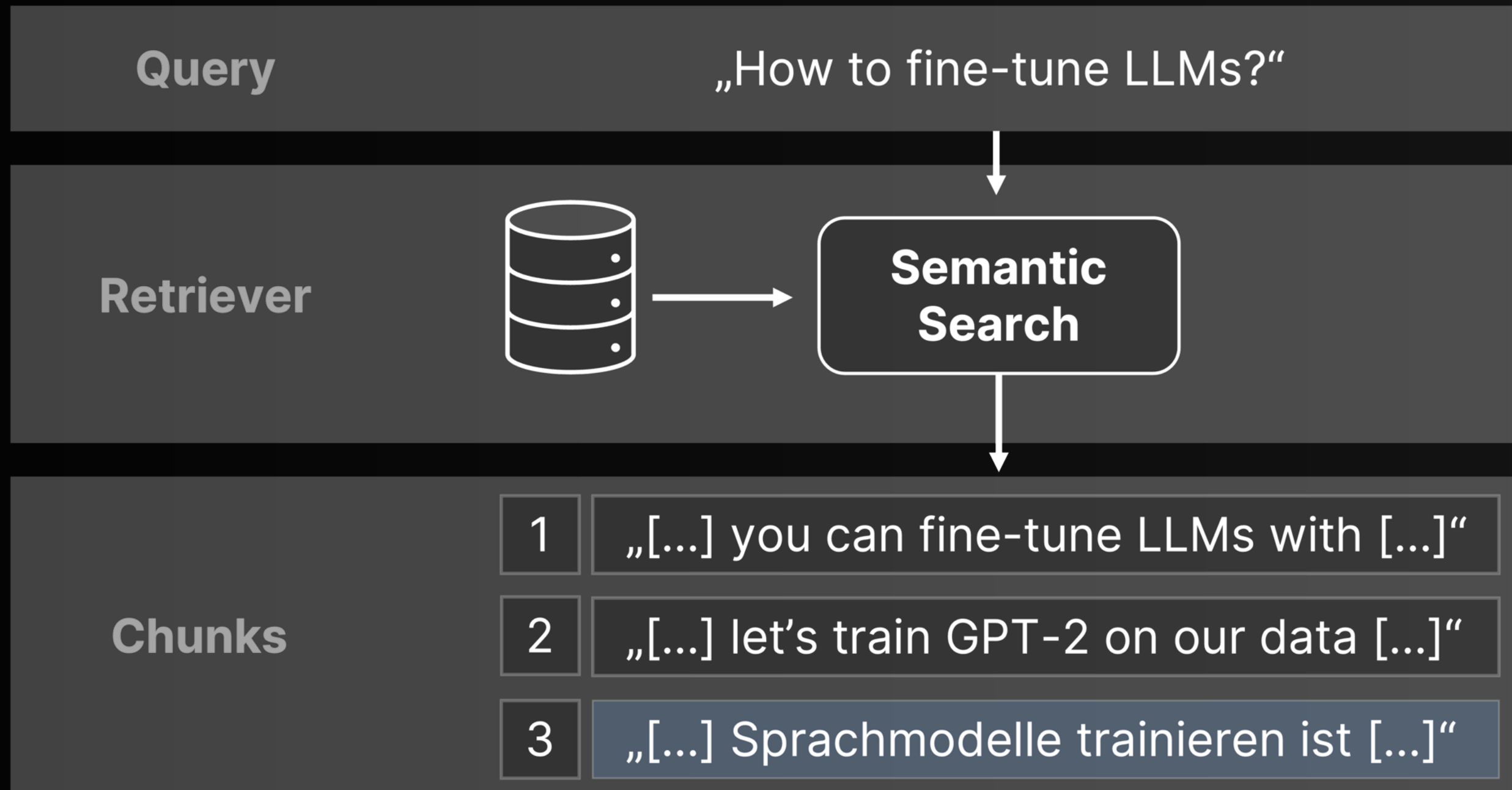
# Document Retrieval

## Semantic Search



# Document Retrieval

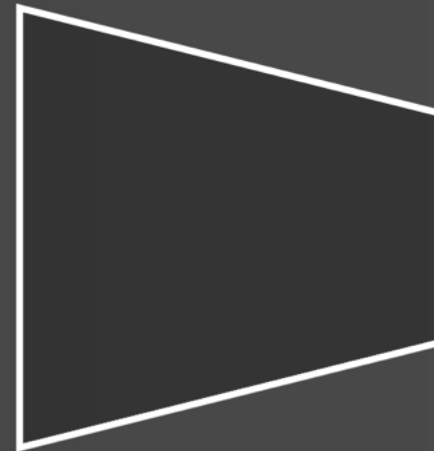
## Semantic Search



# Semantic Search

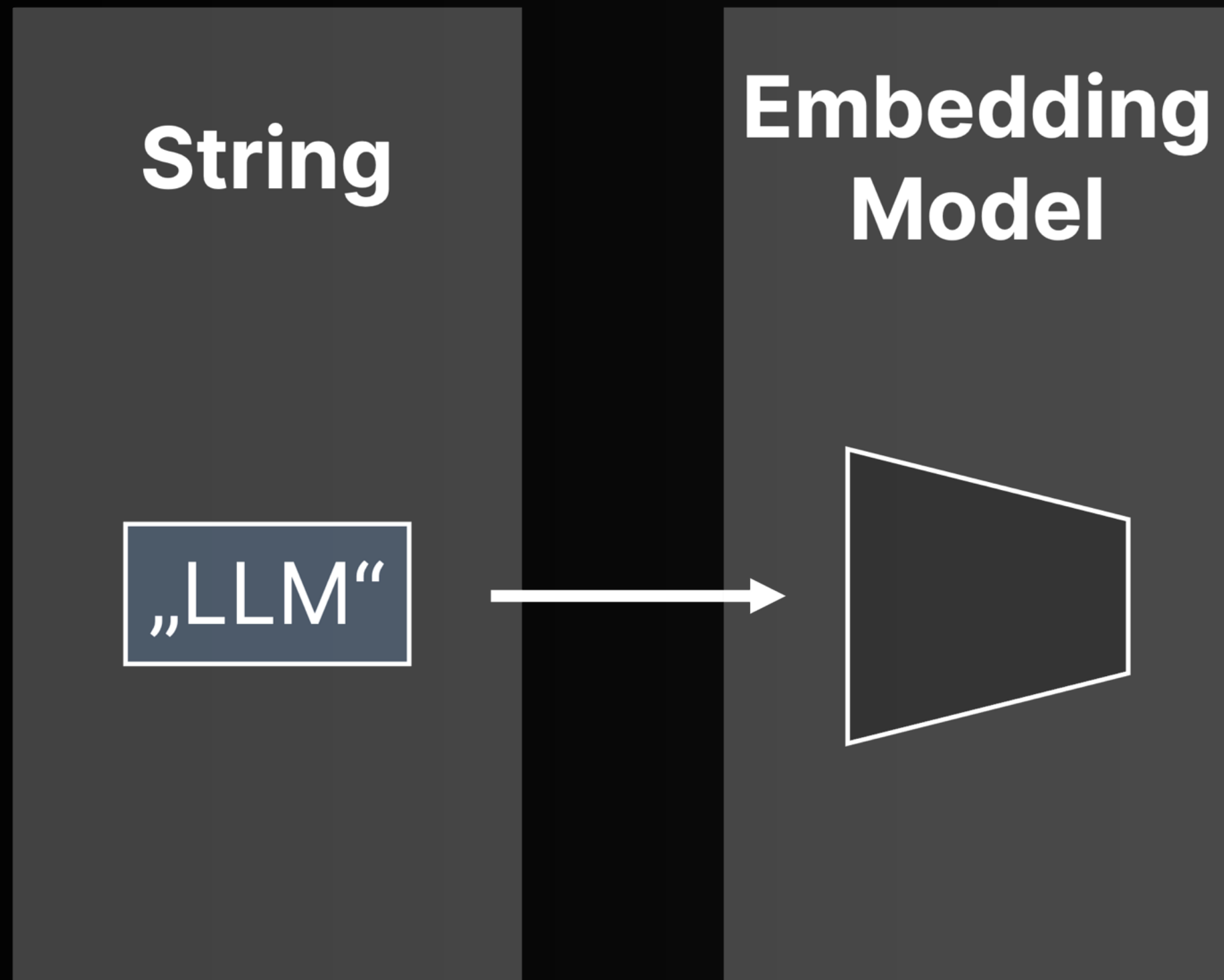
Semantic Embeddings

**Embedding  
Model**



# Semantic Search

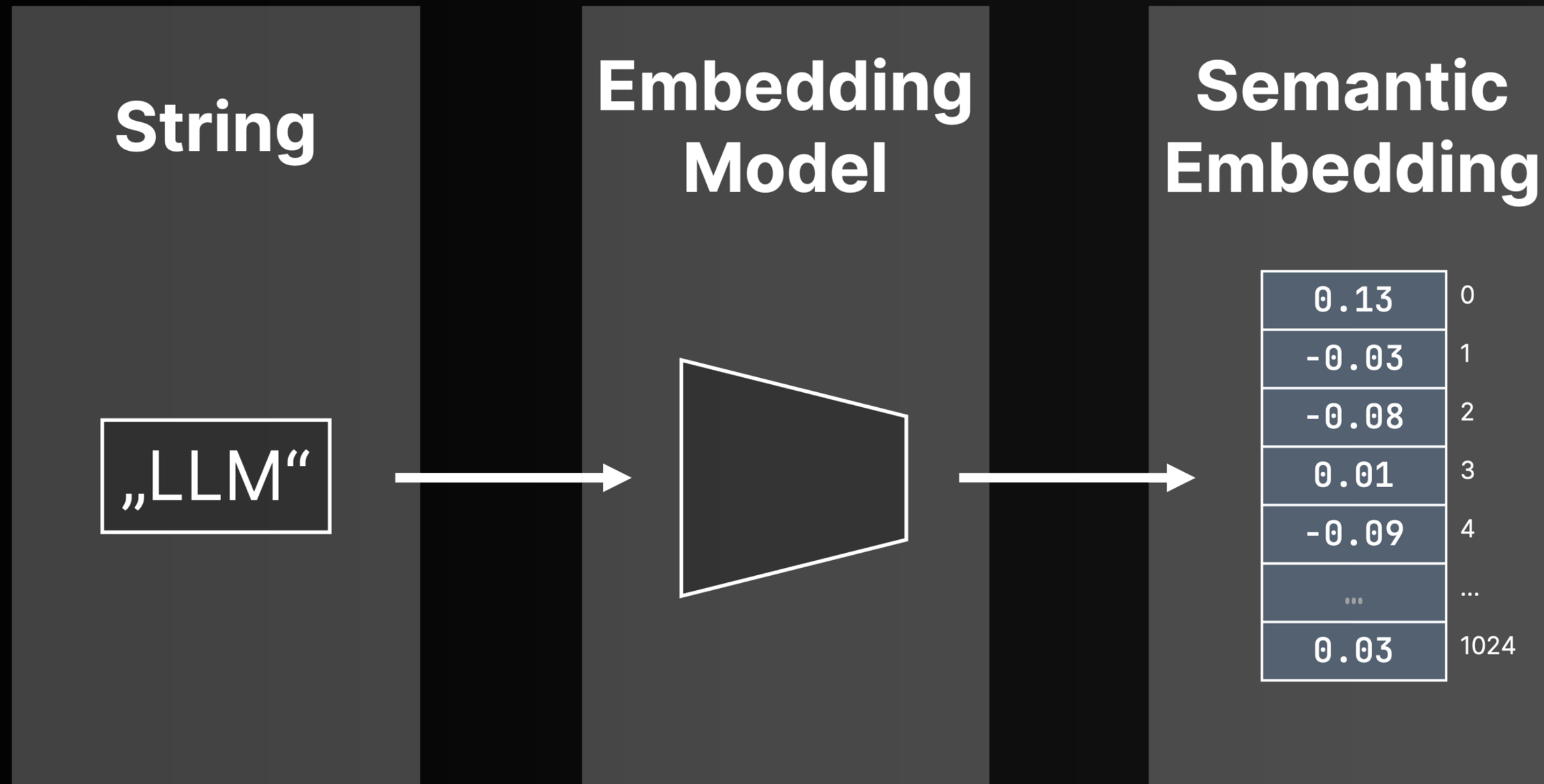
Semantic Embeddings





# Semantic Search

## Semantic Embeddings



# Semantic Search

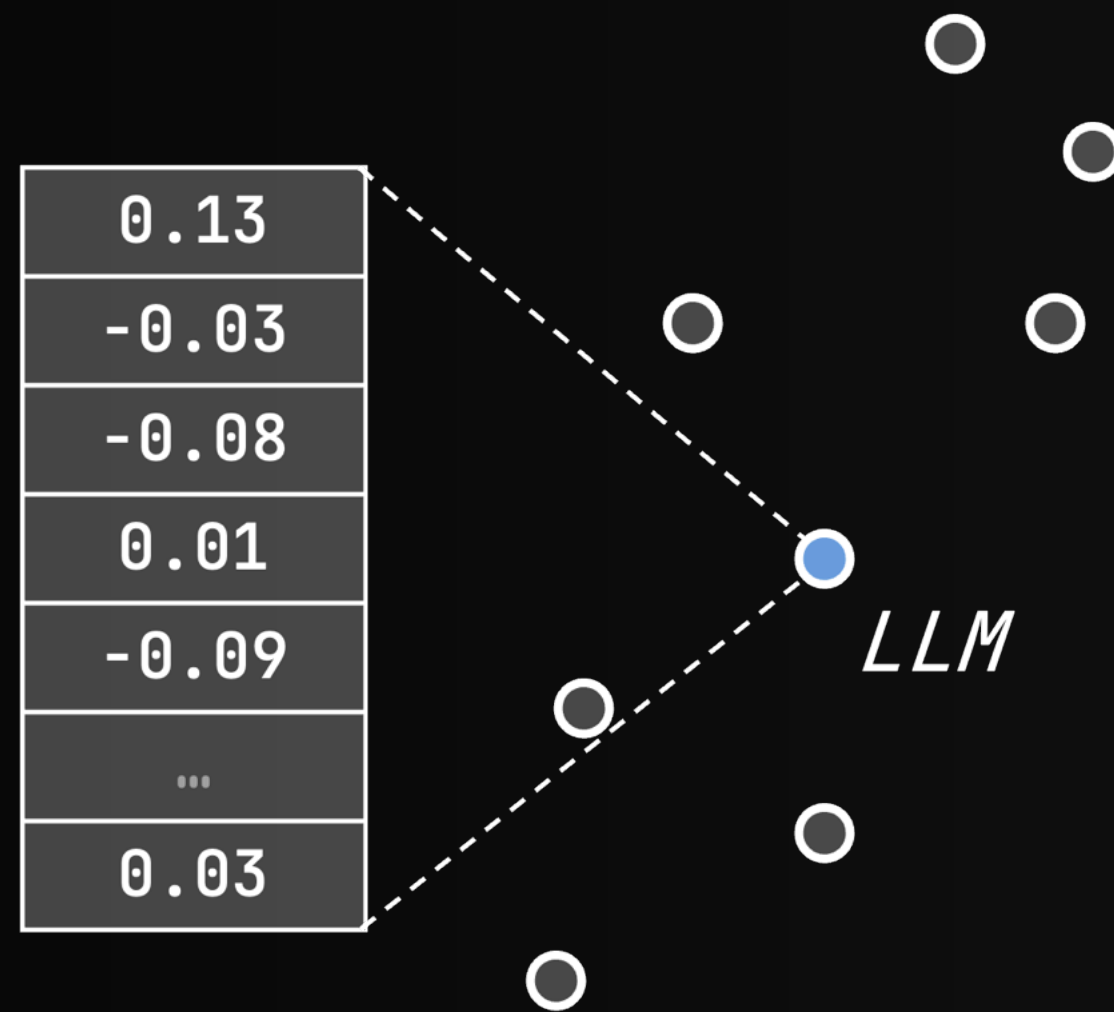
## Semantic Embeddings

### Semantic Embedding

0.13	0
-0.03	1
-0.08	2
0.01	3
-0.09	4
...	...
0.03	1024

# Semantic Search

## Semantic Embeddings



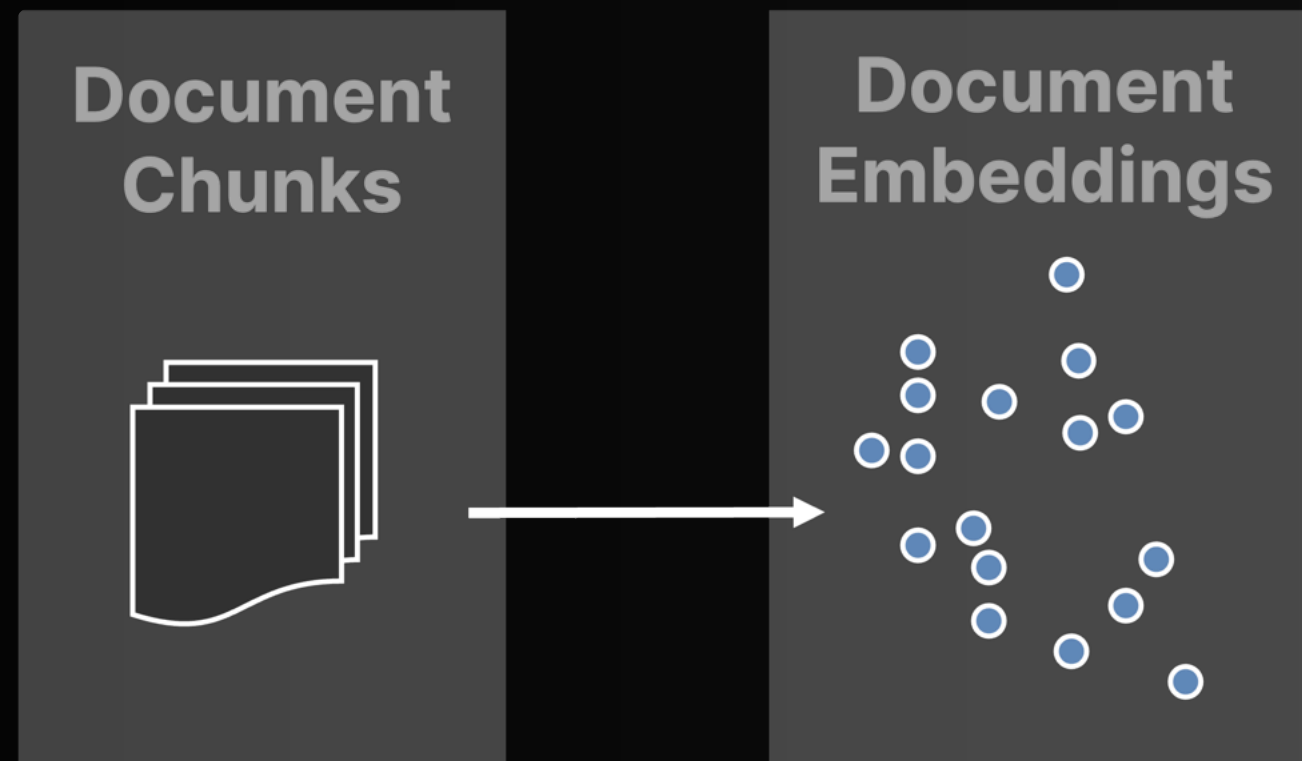
# Semantic Search

## Semantic Embeddings



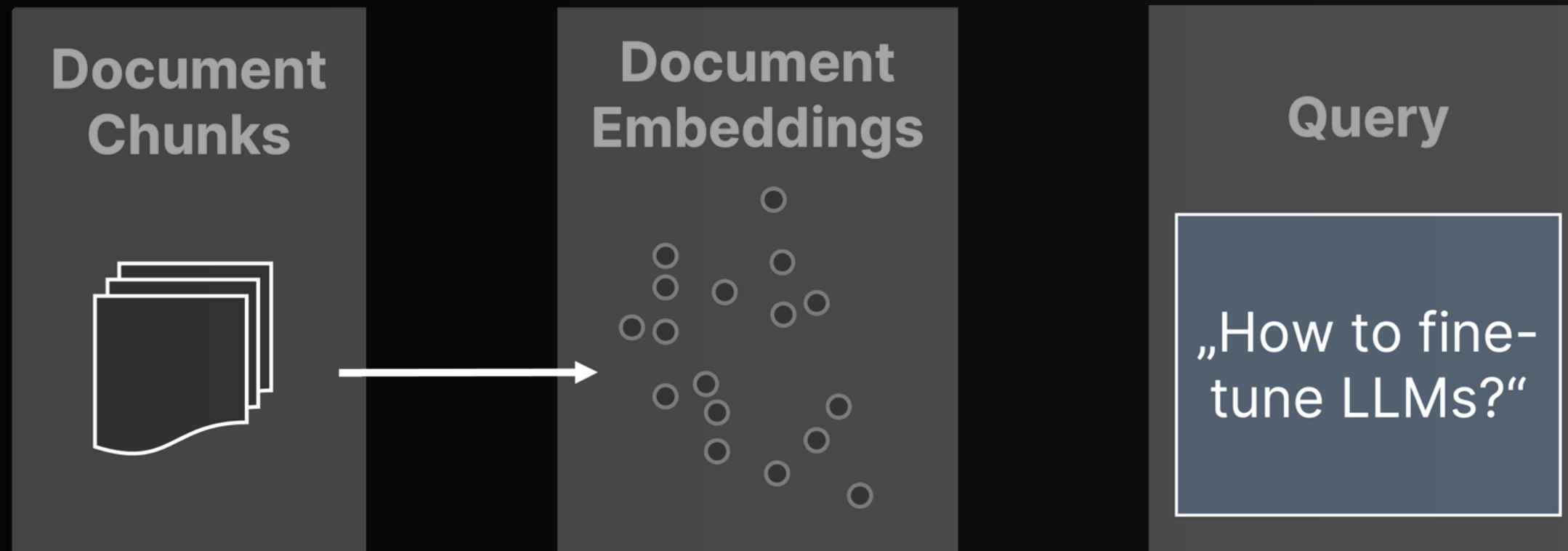
# Semantic Search

Nearest-Neighbor Search



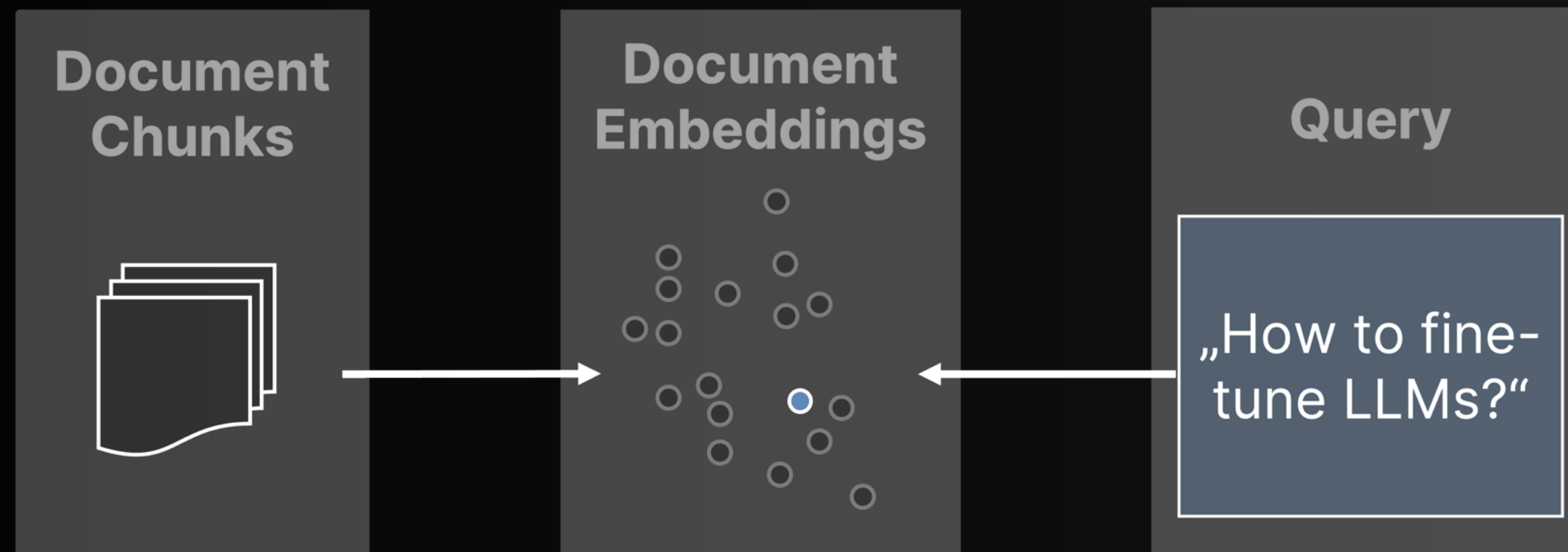
# Semantic Search

Nearest-Neighbor Search



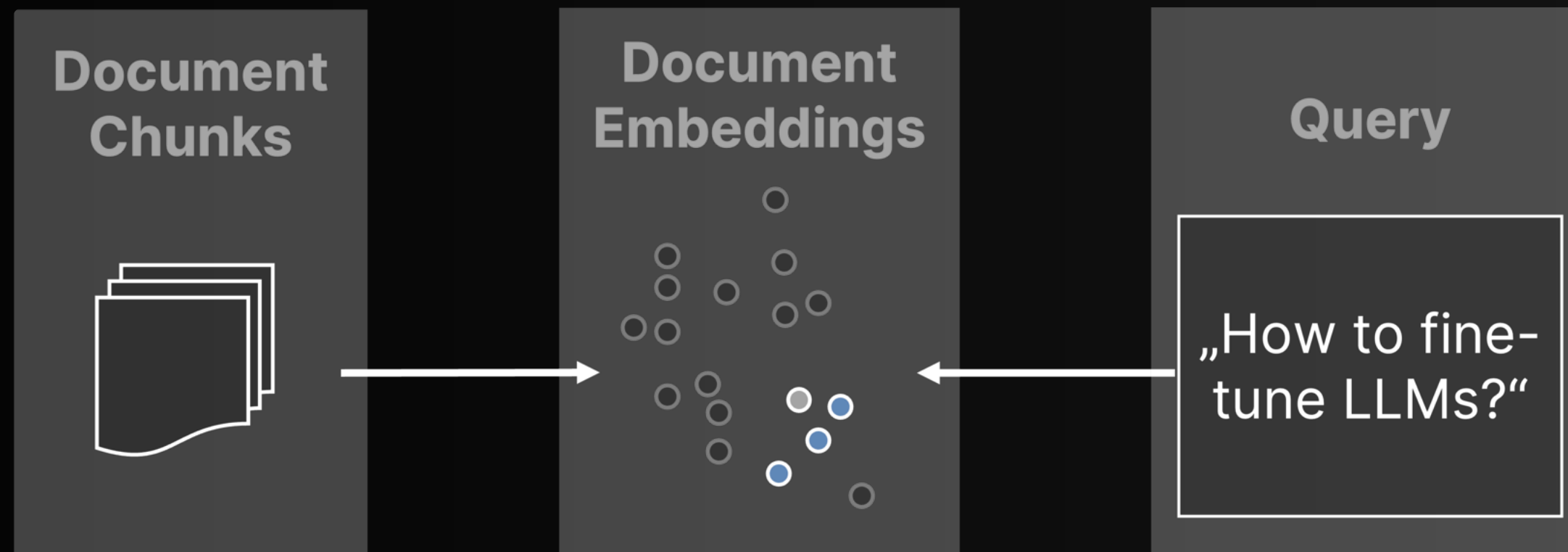
# Semantic Search

Nearest-Neighbor Search



# Semantic Search

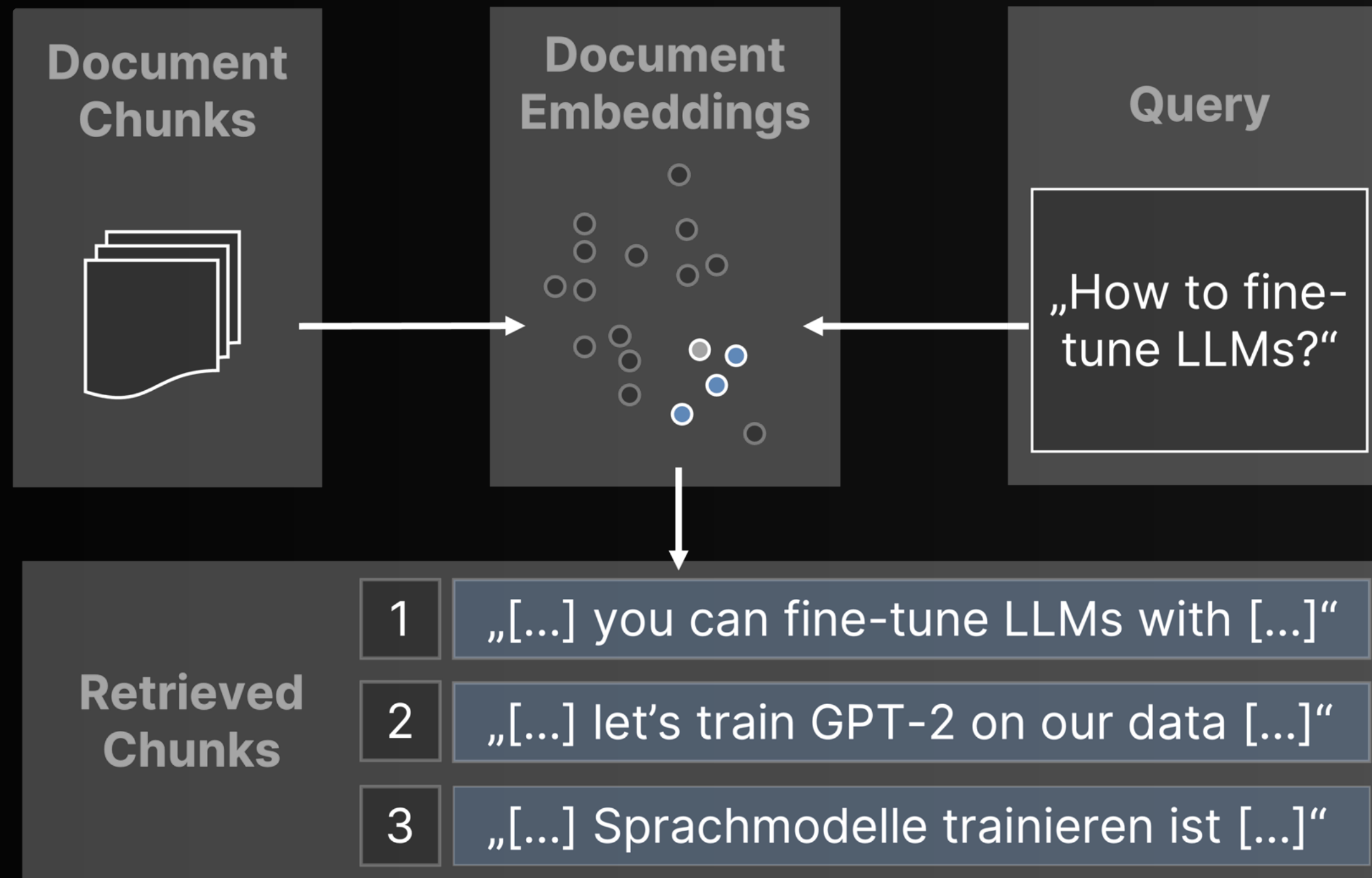
Nearest-Neighbor Search





# Semantic Search

## Nearest-Neighbor Search



# Retrieval Augmented Generation

Collecting Documents



# Retrieval Augmented Generation

From HTML to Markdown

TNG TECHNOLOGY CONSULTING Spaces Categories Calendars Document Management Create Search

Pages / ... / Style Transfer - Theorie 87 views


## Der Kunstfilter: (Fast) Neural Style Transfer

Created by Jochen Reißinger, last modified by Michael Oberparleiter on 12. June 2024

### Neural Style Transfer

#### Die Idee

Das zugrundeliegende Konzept wurde zuerst von Gatys et al. in "A Neural Algorithm of Artistic Style" [1] beschrieben. Die Idee ist, mit Hilfe eines vortrainierten neuronalen Netzwerks für Objekterkennung, eine Repräsentation für den **Inhalt** und den **Stil** eines Gemäldes zu extrahieren. Gegeben ein Input-Bild und eine Stil-Vorlage lässt sich ein Output generieren, der den Inhalt des Inputs und den Stil der Vorlage annähert.



Bilder verlinkt von [4]

#### Implementierungsdetails

Als Grundlage wird das **VGG-19** Netz verwendet. Der **Inhalt** eines Inputs wird durch die **Aktivierungen** (Features) in verschiedenen Schichten des Netzes bestimmt. Und der **Stil** wird hier als die **Korrelation** von verschiedenen Features definiert und als Gram-Matrizen der Aktivierung berechnet. Die Verlustfunktion für einen gegebenen Output wird als die gewichtete Summe von Content- und Styleloss berechnet, die beide wiederum der quadratischen Abweichung zwischen Output und Input/Stil-Vorlage entsprechen.

# Retrieval Augmented Generation

From HTML to Markdown

```
# Der Kunstfilter: (Fast) Neural Style Transfer
```

```
## Neural Style Transfer
```

```
---
```

```
##### Die Idee
```

Das zugrundeliegende Konzept wurde zuerst von Gatys et al. in "A Neural Algorithm of Artistic Style" [1] beschrieben. Die Idee ist, mithilfe eines vortrainierten neuronalen Netzwerks für Objekterkennung, eine Representation für den **Inhalt** und den **Stil** eines Gemäldes zu extrahieren. Gegeben ein Input-Bild und eine Stil-Vorlage lässt sich ein Output generieren, der den Inhalt des Inputs und den Stil der Vorlage annähert.

# Retrieval Augmented Generation

Chunking



# Document Chunking

## Character Splitter

```
# Der Kunstfilter: (Fast) Neural Style Transfer
```

```
## Neural Style Transfer
```

```
---
```

```
##### Die Idee
```

Das zugrundeliegende Konzept wurde zuerst von Gatys et al. in "A Neural Algorithm of Artistic Style" [1] beschrieben. Die Idee ist, mithilfe eines vortrainierten neuronalen Netzwerks für Objekterkennung, eine Representation für den **Inhalt** und den **Stil** eines Gemäldes zu extrahieren. Gegeben ein Input-Bild und eine Stil-Vorlage lässt sich ein Output generieren, der den Inhalt des Inputs und den Stil der Vorlage annähert.

# Document Chunking

## Character Splitter

```
# Der Kunstfilter: (Fast) Neural Style Transfer
```

```
## Neural Style Transfer
```

```
---
```

```
##### Die Idee
```

Das zugrundeliegende Konzept wurde zuerst von Gatys et al. in "A Neural Algorithm of Artistic Style" [1] beschrieben. Die Idee ist, mithilfe eines vortrainierten neuronalen Netzwerks für Objekterkennung, eine Representation für den **Inhalt** und den **Stil** eines Gemäldes zu extrahieren. Gegeben ein Input-Bild und eine Stil-Vorlage lässt sich ein Output generieren, der den Inhalt des Inputs und den Stil der Vorlage annähert.

# Document Chunking

## Character Splitter

```
# Der Kunstfilter: (Fast) Neural Style Transfer
```

```
## Neural Style Transfer
```

```
---
```

```
##### Die Idee
```

Das zugrundeliegende Konzept wurde zuerst von Gatys et al. in "A Neural Algorithm of Artistic Style" [1] beschrieben. Die Idee ist, mithilfe eines vortrainierten neuronalen Netzwerks für Objekterkennung, eine Representation für den **Inhalt** und den **Stil** eines Gemäldes zu extrahieren. Gegeben ein Input-Bild und eine Stil-Vorlage lässt sich ein Output generieren, der den Inhalt des Inputs und den Stil der Vorlage annähert.



# Document Chunking

## Character Splitter

```
# Der Kunstfilter: (Fast) Neural Style Transfer
```

```
## Neural Style Transfer
```

```
---
```

```
##### Die Idee
```

Das zugrundeliegende Konzept wurde zuerst von Gatys et al. in "A Neural Algorithm of Artistic Style" [1] beschrieben. Die Idee ist, mithilfe eines vortrainierten neuronalen Netzwerks für Objekterkennung, eine Representation für den **Inhalt** und den **Stil** eines Gemäldes zu extrahieren. Gegeben ein Input-Bild und eine Stil-Vorlage lässt sich ein Output generieren, der den Inhalt des Inputs und den Stil der Vorlage annähert.

# Document Chunking

## Character Splitter

```
# Der Kunstfilter: (Fast) Neural Style Transfer
```

```
## Neural Style Transfer
```

```
---
```

```
##### Die Idee
```

Das zugrundeliegende Konzept wurde zuerst von Gatys et al. in "A Neural Algorithm of Artistic Style" [1] beschrieben. Die Idee ist, mithilfe eines vortrainierten neuronalen Netzwerks für Objekterkennung, eine Representation für den **\*\*Inhalt\*\*** und den **\*\*Stil\*\*** eines Gemäldes zu extrahieren. Gegeben ein **Input**-Bild und eine **Stil**-Vorlage lässt sich ein **Output** generieren, der den Inhalt des Inputs und den Stil der Vorlage annähert.

# Document Chunking

## Character Splitter

```
# Der Kunstfilter: (Fast) Neural Style Transfer
```

```
## Neural Style Transfer
```

```
---
```

```
##### Die Idee
```

Das zugrundeliegende Konzept wurde zuerst von Gatys et al. in "A Neural Algorithm of Artistic Style" [1] beschrieben. Die Idee ist, mithilfe eines vortrainierten neuronalen Netzwerks für Objekterkennung, eine Representation für den **\*\*Inhalt\*\*** und den **\*\*Stil\*\*** eines Gemäldes zu extrahieren. Gegeben ein Input-Bild und eine Stil-Vorlage lässt sich ein Output generieren, der den Inhalt des Inputs und den Stil der Vorlage annähert.

# Document Chunking

## Character Splitter

```
# Der Kunstfilter: (Fast) Neural Style Transfer
```

```
## Neural Style Transfer
```

```
---
```

```
##### Die Idee
```

Das zugrundeliegende Konzept wurde zuerst von Gatys et al. in "A Neural Algorithm of Artistic Style" [1] beschrieben. Die Idee ist, mithilfe eines vortrainierten neuronalen Netzwerks für Objekterkennung, eine Representation für den **Inhalt** und den **Stil** eines Gemäldes zu extrahieren. Gegeben ein Input-Bild und eine Stil-Vorlage lässt sich ein Output generieren, der den Inhalt des Inputs und den Stil der Vorlage annähert.

# Document Chunking

## Chunk Overlap

```
# Der Kunstfilter: (Fast) Neural Style Transfer
```

```
## Neural Style Transfer
```

```
---
```

```
##### Die Idee
```

Das zugrundeliegende Konzept wurde zuerst von Gatys et al. in "A Neural Algorithm of Artistic Style" [1] beschrieben. Die Idee ist, mithilfe eines vortrainierten neuronalen Netzwerks für Objekterkennung, eine Representation für den **Inhalt** und den **Stil** eines Gemäldes zu extrahieren. Gegeben ein Input-Bild und eine Stil-Vorlage lässt sich ein Output generieren, der den Inhalt des Inputs und den Stil der Vorlage annähert.

# Document Chunking

## Chunk Overlap

```
# Der Kunstfilter: (Fast) Neural Style Transfer
```

```
## Neural Style Transfer
```

```
---
```

```
##### Die Idee
```

Das zugrundeliegende Konzept wurde zuerst von Gatys et al. in "A Neural Algorithm of Artistic Style" [1] beschrieben. Die Idee ist, mithilfe eines vortrainierten neuronalen Netzwerks für Objekterkennung, eine Representation für den **\*\*Inhalt\*\*** und den **\*\*Stil\*\*** eines Gemäldes zu extrahieren. Gegeben ein Input-Bild und eine Stil-Vorlage lässt sich ein Output generieren, der den Inhalt des Inputs und den Stil der Vorlage annähert.

# Document Chunking

## Chunk Overlap

```
# Der Kunstfilter: (Fast) Neural Style Transfer
```

```
## Neural Style Transfer
```

```
---
```

```
##### Die Idee
```

Das zugrundeliegende Konzept wurde zuerst von Gatys et al. in "A Neural Algorithm of Artistic Style" [1] beschrieben. Die Idee ist, mithilfe eines vortrainierten neuronalen Netzwerks für Objekterkennung, eine Representation für den **\*\*Inhalt\*\*** und den **\*\*Stil\*\*** eines Gemäldes zu extrahieren. Gegeben ein Input-Bild und eine Stil-Vorlage lässt sich ein Output generieren, der den Inhalt des Inputs und den Stil der Vorlage annähert.

# Document Chunking

## Chunk Overlap

```
# Der Kunstfilter: (Fast) Neural Style Transfer
```

```
## Neural Style Transfer
```

```
---
```

```
##### Die Idee
```

Das zugrundeliegende Konzept wurde zuerst von Gatys et al. in "A Neural Algorithm of Artistic Style" [1] beschrieben. Die Idee ist, mithilfe eines vortrainierten neuronalen Netzwerks für Objekterkennung, eine Representation für den **Inhalt** und den **Stil** eines Gemäldes zu extrahieren. Gegeben ein Input-Bild und eine Stil-Vorlage lässt sich ein Output generieren, der den Inhalt des Inputs und den Stil der Vorlage annähert.



# Document Chunking

## Chunk Overlap

```
# Der Kunstfilter: (Fast) Neural Style Transfer
```

```
## Neural Style Transfer
```

```
---
```

```
##### Die Idee
```

Das zugrundeliegende Konzept wurde zuerst von Gatys et al. in "A Neural Algorithm of Artistic Style" [1] beschrieben. Die Idee ist, mithilfe eines vortrainierten neuronalen Netzwerks für Objekterkennung, eine Representation für den **Inhalt** und den **Stil** eines Gemäldes zu extrahieren. Gegeben ein Input-Bild und eine Stil-Vorlage lässt sich ein Output generieren, der den Inhalt des Inputs und den Stil der Vorlage annähert.

# Document Chunking

## Sentence Splitter

```
# Der Kunstfilter: (Fast) Neural Style Transfer
```

```
## Neural Style Transfer
```

```
---
```

```
##### Die Idee
```

Das zugrundeliegende Konzept wurde zuerst von Gatys et al. in "A Neural Algorithm of Artistic Style" [1] beschrieben. Die Idee ist, mithilfe eines vortrainierten neuronalen Netzwerks für Objekterkennung, eine Representation für den **Inhalt** und den **Stil** eines Gemäldes zu extrahieren. Gegeben ein Input-Bild und eine Stil-Vorlage lässt sich ein Output generieren, der den Inhalt des Inputs und den Stil der Vorlage annähert.

# Document Chunking

## Sentence Splitter

```
# Der Kunstfilter: (Fast) Neural Style Transfer
```

```
## Neural Style Transfer
```

```
---
```

```
##### Die Idee
```

Das zugrundeliegende Konzept wurde zuerst von Gatys et al. in "A Neural Algorithm of Artistic Style" [1] beschrieben. Die Idee ist, mithilfe eines vortrainierten neuronalen Netzwerks für Objekterkennung, eine Representation für den **Inhalt** und den **Stil** eines Gemäldes zu extrahieren. Gegeben ein Input-Bild und eine Stil-Vorlage lässt sich ein Output generieren, der den Inhalt des Inputs und den Stil der Vorlage annähert.

# Document Chunking

## Sentence Splitter

```
# Der Kunstfilter: (Fast) Neural Style Transfer
```

```
## Neural Style Transfer
```

```
---
```

```
##### Die Idee
```

Das zugrundeliegende Konzept wurde zuerst von Gatys et al. in "A Neural Algorithm of Artistic Style" [1] beschrieben. Die Idee ist, mithilfe eines vortrainierten neuronalen Netzwerks für Objekterkennung, eine Representation für den **\*\*Inhalt\*\*** und den **\*\*Stil\*\*** eines Gemäldes zu extrahieren. Gegeben ein Input-Bild und eine Stil-Vorlage lässt sich ein Output generieren, der den Inhalt des Inputs und den Stil der Vorlage annähert.

# Document Chunking

## Sentence Splitter

```
# Der Kunstfilter: (Fast) Neural Style Transfer
```

```
## Neural Style Transfer
```

```
---
```

```
##### Die Idee
```

Das zugrundeliegende Konzept wurde zuerst von Gatys et al. in "A Neural Algorithm of Artistic Style" [1] beschrieben. Die Idee ist, mithilfe eines vortrainierten neuronalen Netzwerks für Objekterkennung, eine Representation für den **\*\*Inhalt\*\*** und den **\*\*Stil\*\*** eines Gemäldes zu extrahieren. Gegeben ein Input-Bild und eine Stil-Vorlage lässt sich ein Output generieren, der den Inhalt des Inputs und den Stil der Vorlage annähert.

# RAG on Innovation Hacking Data

## RAG Setup with LLamaIndex

```
documents = SimpleDirectoryReader("data").load_data()
text_splitter = SentenceSplitter(chunk_size=120, chunk_overlap=40)

index = VectorStoreIndex.from_documents(
    documents,
    embed_model=CustomEmbedder(),
    transformations=[text_splitter])
retriever = index.as_retriever(similarity_top_k=5)

prompt_template = PromptTemplate("""You are Jonas Mayer, the Innovation
Here's some potentially relevant confluence docs: {context}""")
```

# RAG on Innovation Hacking Data

Retrieval, Augmentation and Generation with LlamaIndex

```
nodes = retriever.retrieve(question)
context = "\n".join([n.text for n in nodes])

prompt = prompt_template.format(context=context)

response= llm.chat(
    messages=[
        ChatMessage(role=MessageRole.SYSTEM, content=prompt),
        ChatMessage(role=MessageRole.USER, content=question)
    ])
```

# RAG on Innovation Hacking Data

Live Demo

Input

Please start the demo!



# Limitations

Questions on Meta-Information



# Limitations

Data Quality



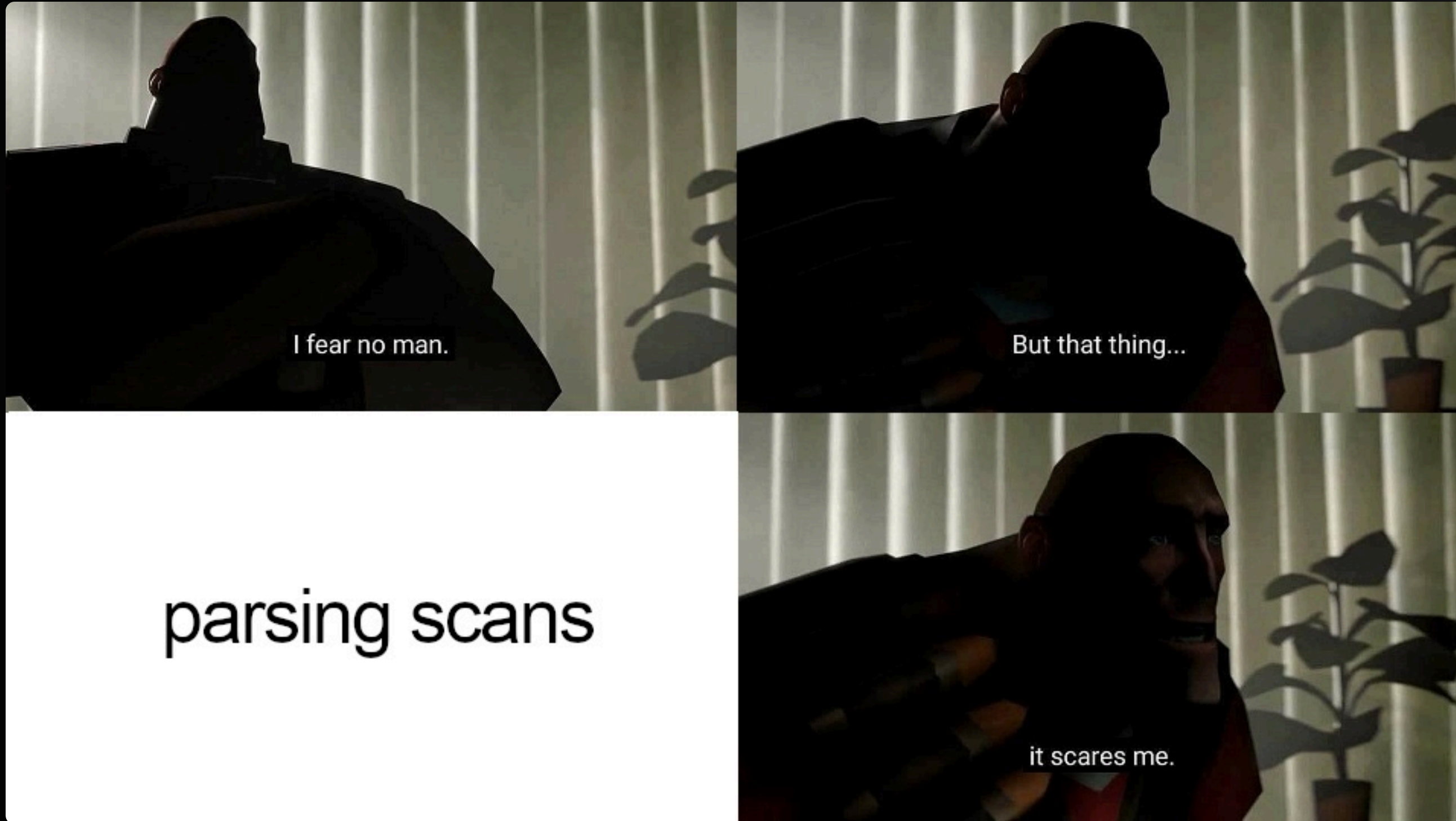
# Limitations

Data Quality



# Limitations

Data Quality





# Overview

- ▶ Introduction
- ▶ Prompt Engineering
- ▶ RAG
- ▶ Fine-Tuning
- ▶ Conclusion



# Fine-Tuning

Pre-Training vs Fine-Tuning



Training  
Time



# Fine-Tuning

Pre-Training vs Fine-Tuning



Training  
Time





# Fine-Tuning

Pre-Training vs Fine-Tuning



Training  
Time



# Fine-Tuning

Pre-Training vs Fine-Tuning



Training  
Time



# Fine-Tuning

Pre-Training vs Fine-Tuning



Training  
Time



# Fine-Tuning

Pre-Training vs Fine-Tuning



Training  
Time



# Fine-Tuning

LLM Training in a Nutshell

**Input**



**LLM**

**Output**



# Fine-Tuning

LLM Training in a Nutshell

**Input**



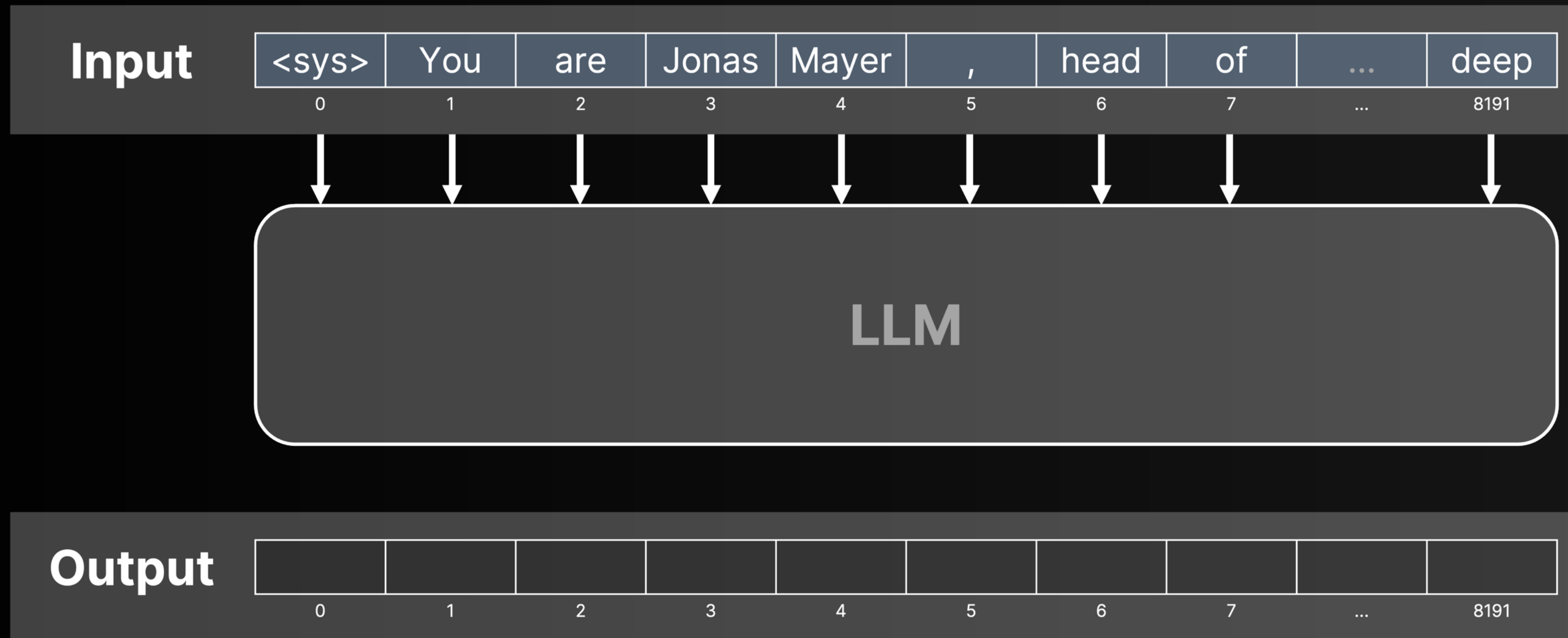
LLM

**Output**



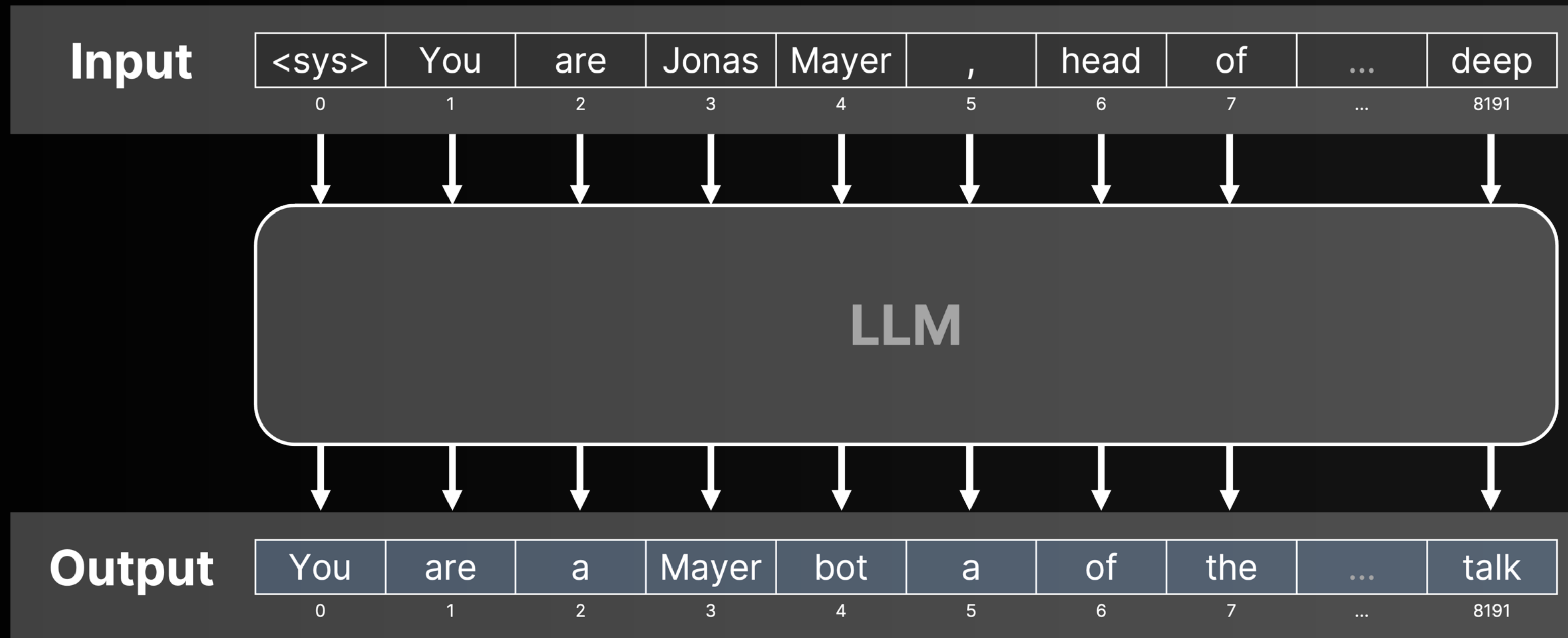
# Fine-Tuning

LLM Training in a Nutshell



# Fine-Tuning

LLM Training in a Nutshell





# Fine-Tuning

LLM Training in a Nutshell

**Input**

<sys>	You	are	Jonas	Mayer	,	head	of	...	deep
0	1	2	3	4	5	6	7	...	8191

LLM

**Output**

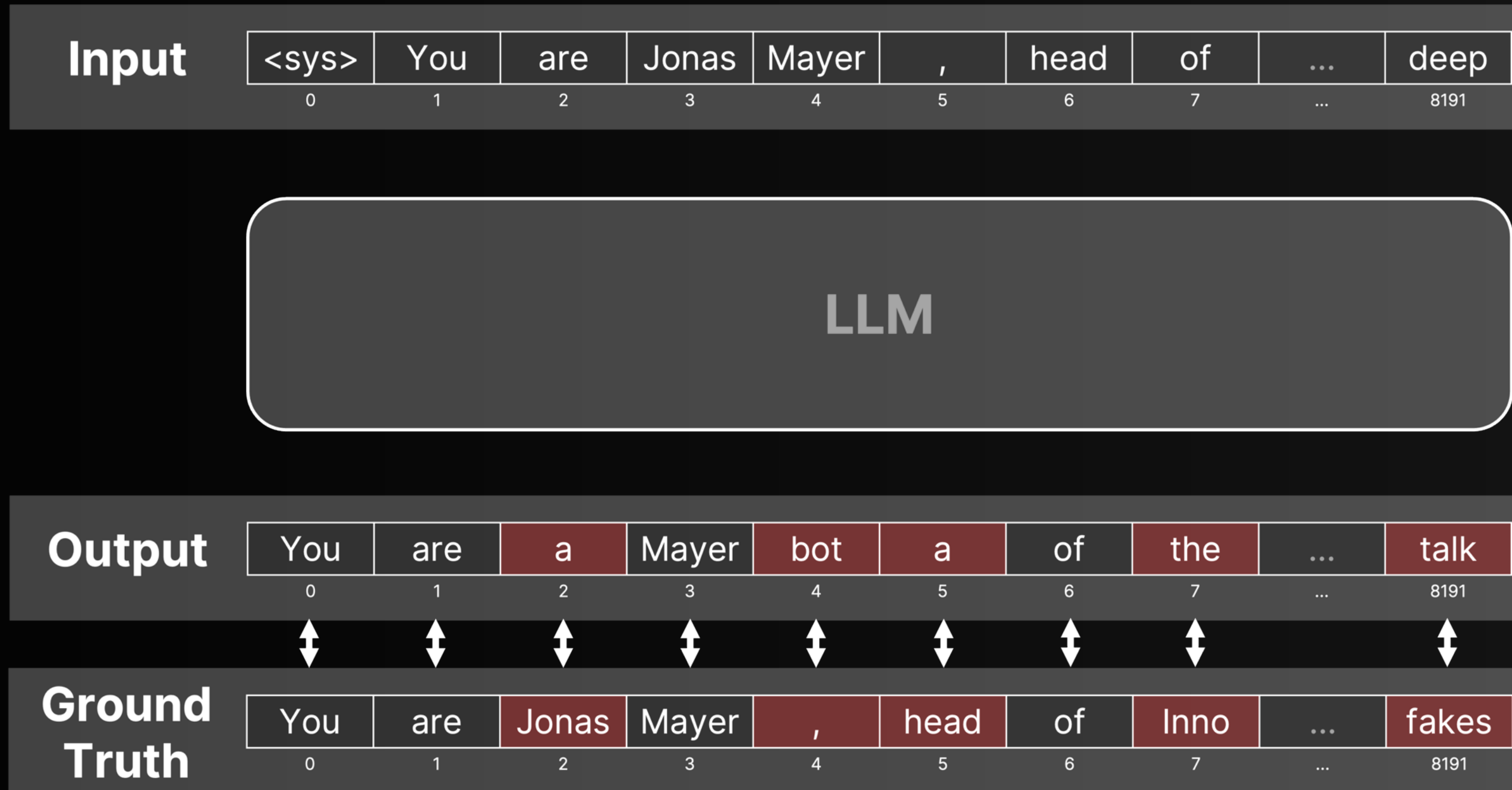
You	are	a	Mayer	bot	a	of	the	...	talk
0	1	2	3	4	5	6	7	...	8191

**Ground Truth**

You	are	Jonas	Mayer	,	head	of	Inno	...	fakes
0	1	2	3	4	5	6	7	...	8191

# Fine-Tuning

LLM Training in a Nutshell



# Fine-Tuning

Trainable Parameters



# Fine-Tuning

Trainable Parameters



# Fine-Tuning

Full Fine-Tuning vs PEFT



**Full  
Fine-Tuning**



**Parameter-  
Efficient  
Fine-Tuning**

# Full Fine-Tuning

Used To Learn New Skills



# Full Fine-Tuning

Used To Learn New Skills

```
return None
context.append(parsedInput) # Add new conversation entry to context
return (self.model.generateLLMOutput(parsedInput), context)

def parseInputToLanguageModel(inputString, inputLanguage, context):
    if self.model is None or self.model.language != inputLanguage:
        # LLM is not initialised or has wrong language, load LLM
        self.model = self.loadAILanguageModelFromDatabase(inputLanguage)
        if self.model is None or not self.runModelSelfDiagnosticTest():
            raise Exception("AI language model load failed")
        return None
    self.model.setLLMContext(context) # Put past conversation context into it
    llmInputParser = self.model.getInputParser()
    return llmInputParser.parseInput(inputString)

def generateLLMOutput(parsedInput):
    llmContext = self.model.getLLMContext()
    intermediateResponse = self.model.convertInputToIntermediateResponse
```

# Full Fine-Tuning

Used To Learn New Skills





# Full Fine-Tuning

Used To Learn New Skills



**You**

What is up?



**ChatGPT**

Hey! Not much, just here to help out. How about you? Is there something specific you need assistance with today?



**You**

Can you just reply with "The Sky!" already! >:(



**ChatGPT**

The Sky!

# Full Fine-Tuning

Limitations: Tokenizer

Tokenizers are  
overrated!



0	30642	Token
1	7509	izers
2	264	are
3	521	over
4	6184	rated
5	120	!

# Full Fine-Tuning

Limitations: Tokenizer

Tokenizers are overrated!



0	30642	Token
1	7509	izers
2	264	are
3	521	over
4	6184	rated
5	120	!

Tokenizer sind überbewertet!



0	30642	Token
1	7509	izer
2	264	s
3	521	ind
4	6184	
5	120	ü
6	65	ber
7	413	b
...	...	...
11	0	!

# Full Fine-Tuning

Limitation: Catastrophic Forgetting



# Full Fine-Tuning

Limitation: Lots of GPUs and Data



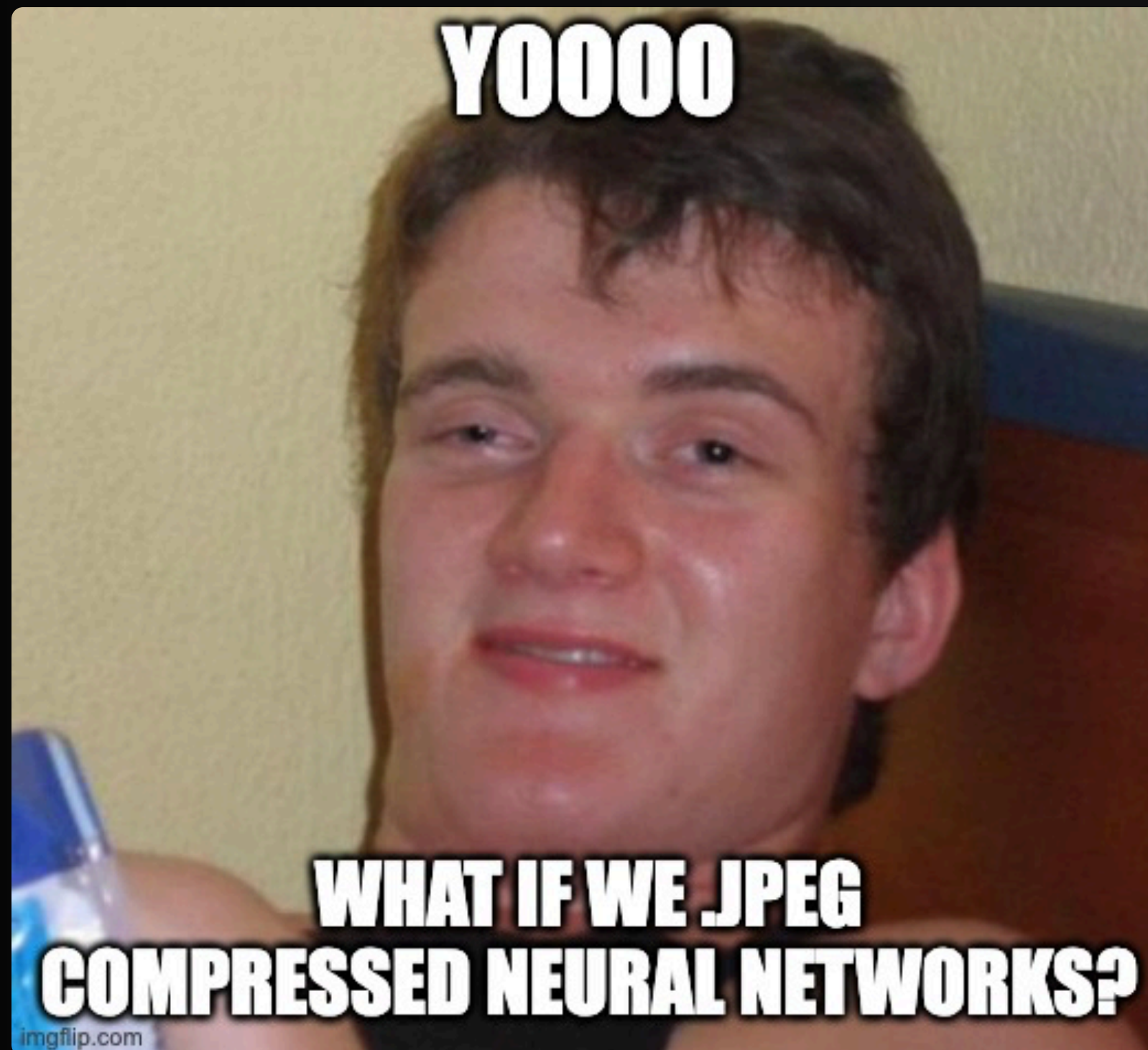
# Parameter-Efficient Fine-Tuning

Used for Smaller Tweaks



# Parameter-Efficient Fine-Tuning

Low-Rank Adaptation



# LoRa in a Nutshell

Training a Compressed Delta



**LoRa Fine-Tuned  
Parameters**



# LoRa in a Nutshell

Training a Compressed Delta



**Uncompressed  
Pre-Trained  
Parameters**

70B Parameters  $\approx$  140GB



**Compressed  
Fine-Tuned  
Delta**

7M Parameters  $\approx$  14MB

# LoRa in a Nutshell

Training a Compressed Delta



**Uncompressed  
Pre-Trained  
Parameters**

70B Parameters  $\approx$  140GB



**Compressed  
Fine-Tuned  
Delta**

7M Parameters  $\approx$  14MB

# LoRa Inference

LoRAX Server



# LoRa in a Nutshell

Less GPUs, Less Data



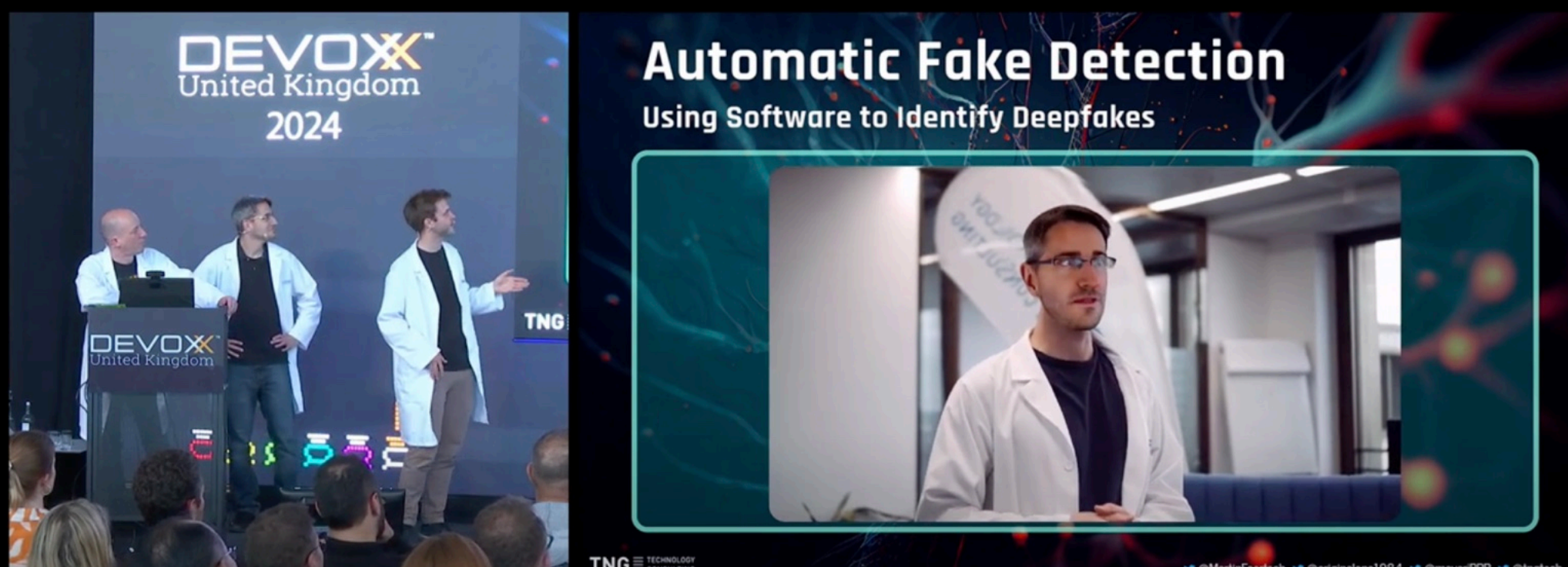
# LoRa in a Nutshell

Less GPUs, Less Data



# Fine-Tuning Jonas

Collecting a Jonas Dataset: Unlabelled Data



**Automatic Fake Detection**  
Using Software to Identify Deepfakes

kind of a video on YouTube right this is obviously the real Thomas

44:41 / 56:38

Devoxx UK  
10.500 Abonnenten

Abonnieren

1

Teilen

Herunterladen

### Talks in English

Innovation Hacking - 10/10

- 5 **is not what you see? by Martin...**  
Devoxx  
50:08
- 6 **Pushing Deepfakes to the Limit (Fake video calls) - Jonas Mayer...**  
Caribbean Developers Conference  
47:16
- 7 **Pushing Deepfakes to the Limit-Fake video calls with AI by Marti...**  
Jfokus  
58:14
- 8 **It's All About AI • Martin Förtsch, Thomas Endres & Jonas Mayer ...**  
GOTO Conferences  
1:02:43
- 9 **Brain Computer Interfaces Demystified - Can thoughts take...**  
Devoxx UK  
49:58
- The AI Elections: How Technology Could Shape Public...**  
Devoxx UK  
56:39

Alle Von Devoxx UK Angesehen

Breaking AI: Live coding and

# Fine-Tuning Jonas

Collecting a Jonas Dataset: Q&A

`user:`

`How many projects do you have in innovation hacking?`

# Fine-Tuning Jonas

Collecting a Jonas Dataset: Q&A

user:

How many projects do you have in innovation hacking?

assistant:

Uhm, how many projects do we have in innovation hacking?

Well, that's a good question, i don't think i've counted them recently!

It's quite a few actually, traditionally we've been doing like one every

year, but ever since i started it's been more like 2-4 per year. Well,

depends on what you count as a showcase, i guess. Haha, i'm sorry.

I don't think i really answered the question at all, did i?

Next question!



# Fine-Tuning Jonas

Collecting a Jonas Dataset: Q&A

```
system: You are Jonas, proud member of the Innovation Hacking Team of TI
To answer the question, you'll receive some context information that are
Context information is below.
```

```
-----
### Technologien ###
```

```
JavaScript: weil Browser
```

```
TensorFlow.js: für Neuronale Netze, mit WebGL Beschleunigung
```

```
PIXI.js: Framework zur 2D-Spieleentwicklung
```

# Fine-Tuning Jonas

Collecting a Dataset



# Overview

- ▶ Introduction
- ▶ Prompt Engineering
- ▶ RAG
- ▶ Fine-Tuning
- ▶ Conclusion



# Conclusion

## LLM Customization Options

### Prompt Engineering

\$

easy customization  
limited alignment  
limited knowledge

Start for basically free

MOST POPULAR

### RAG

\$\$

unlimited data  
easy updates  
full transparency

Build RAG

### Fine-Tuning

\$\$\$

language adaptation  
advanced alignment  
peft or full

GPU go brrr

# Conclusion

The TNG Principle

**TNG**  **TECHNOLOGY  
CONSULTING**

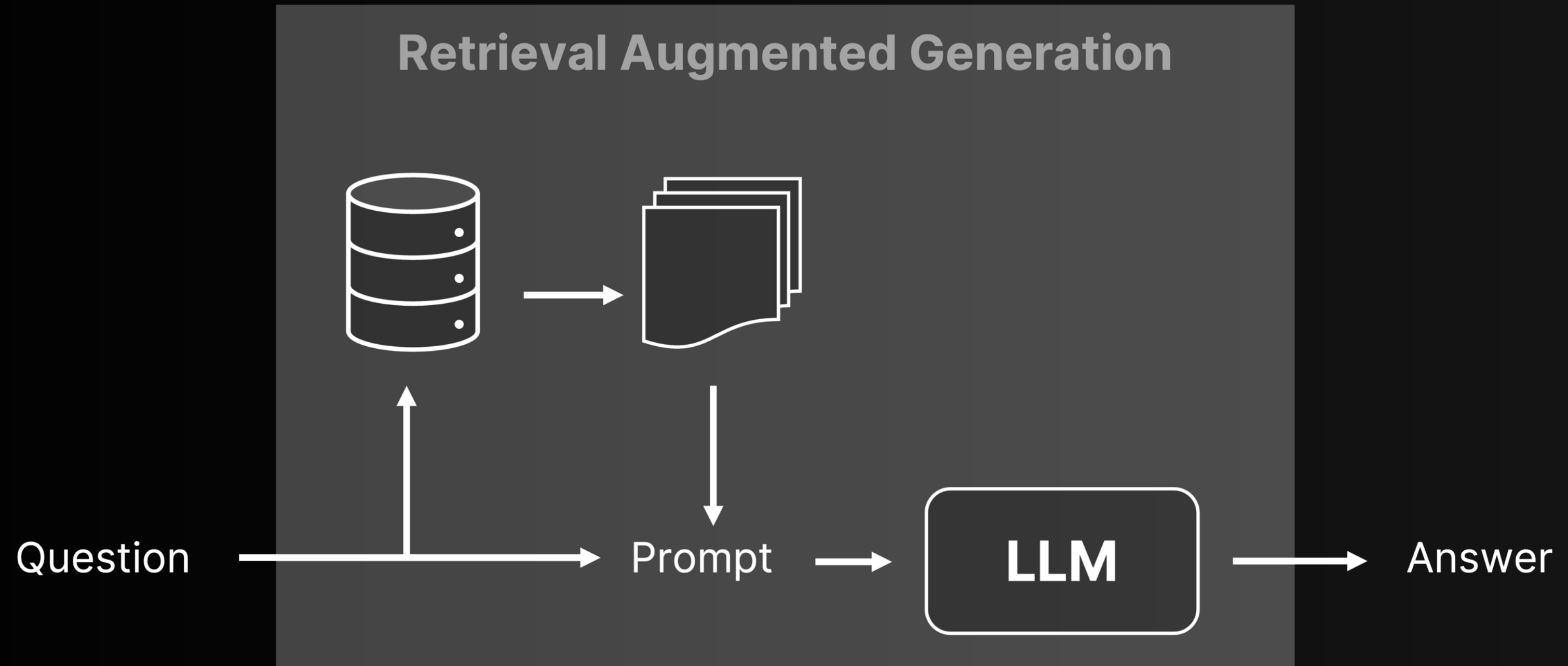
# Conclusion

The TNG Principle

**TESTEN. NICHT GLAUBEN!**

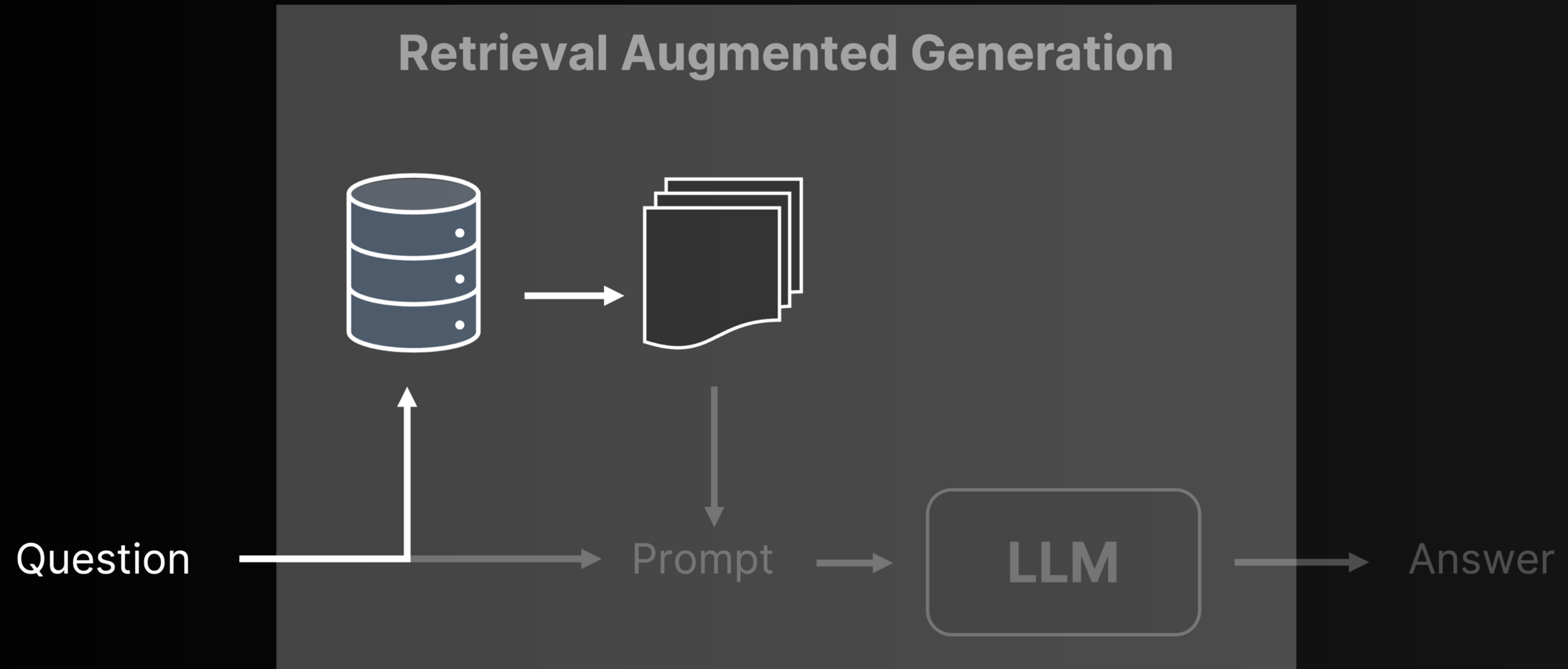
# Evaluation

From Component to End-to-End Testing



# Evaluation

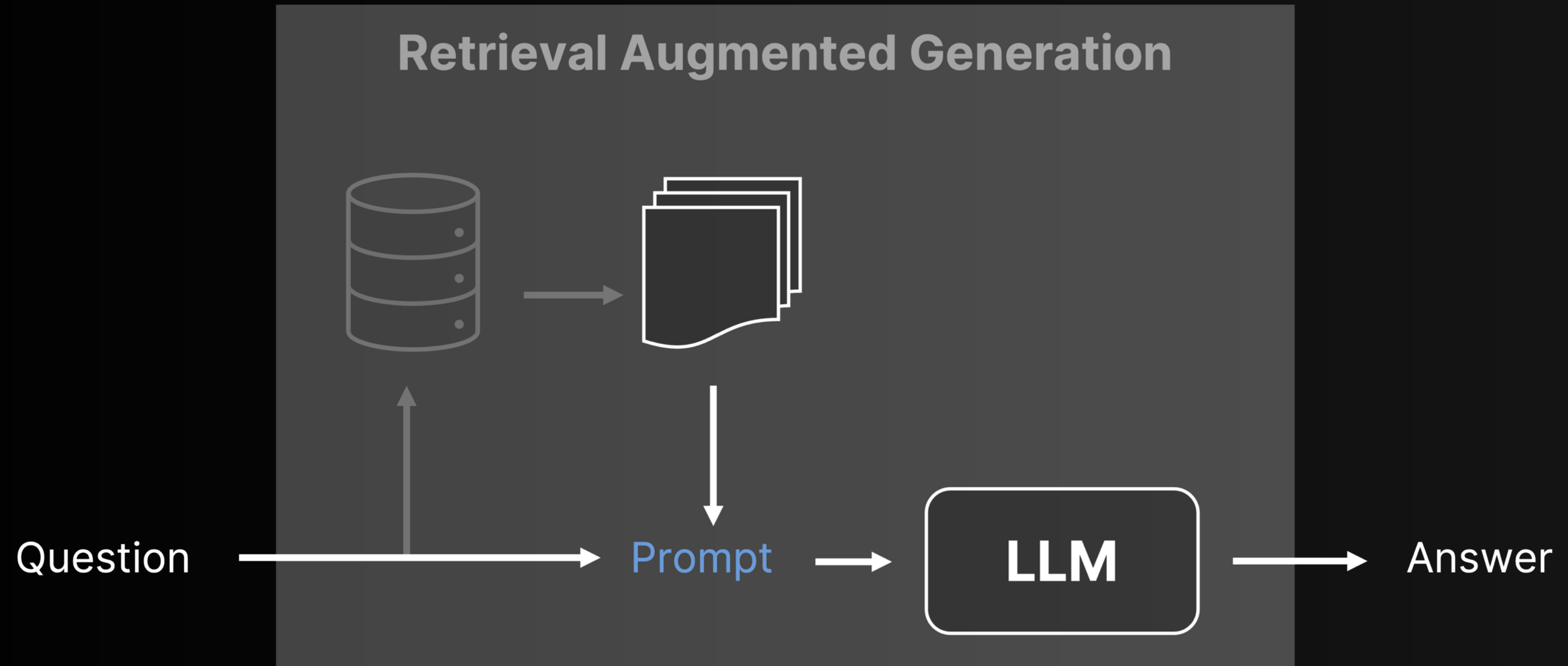
From Component to End-to-End Testing





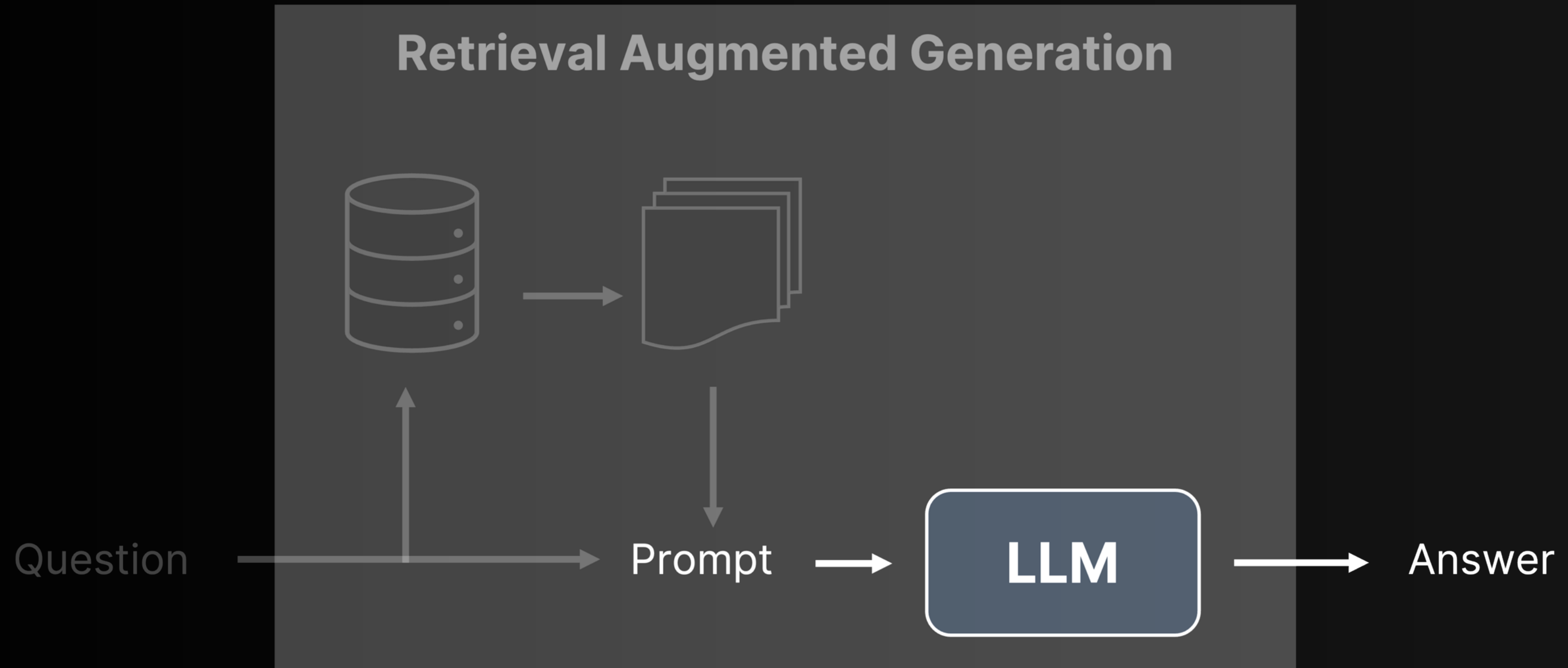
# Evaluation

From Component to End-to-End Testing



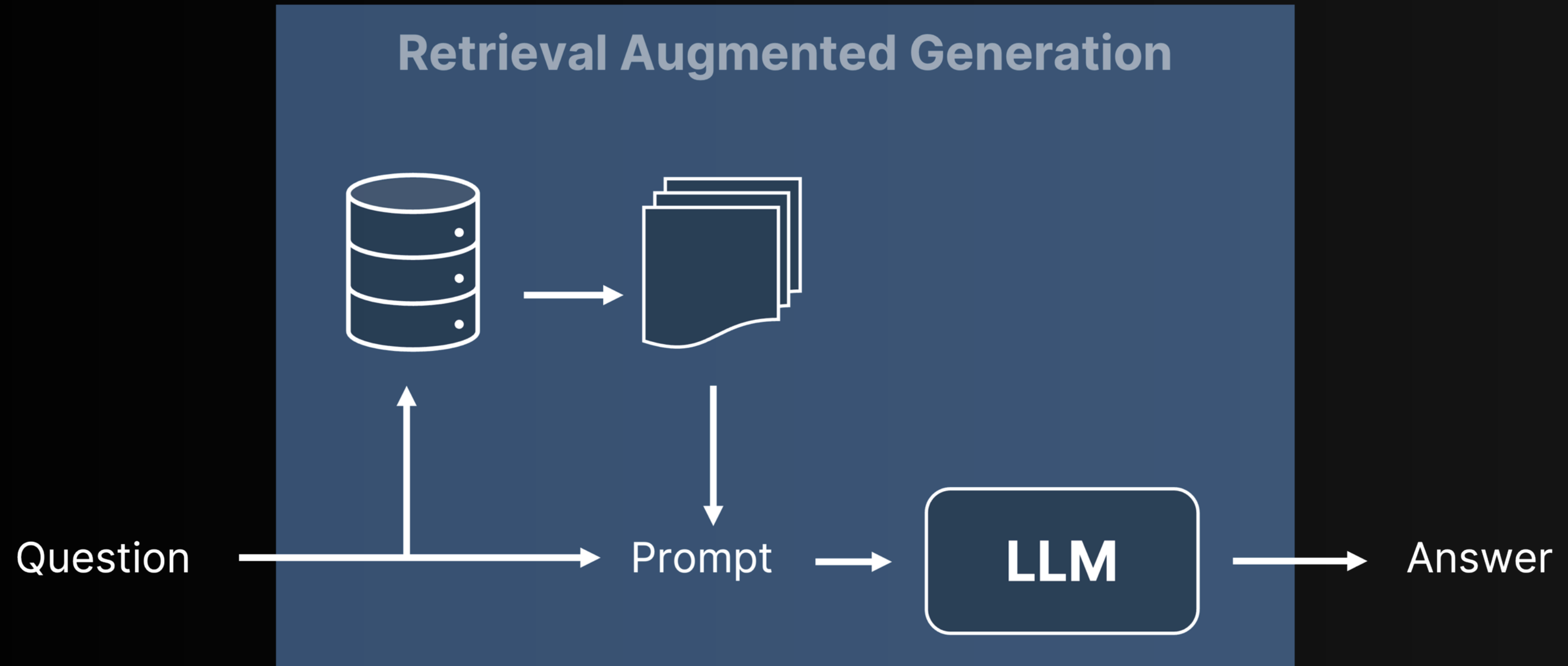
# Evaluation

From Component to End-to-End Testing



# Evaluation

From Component to End-to-End Testing



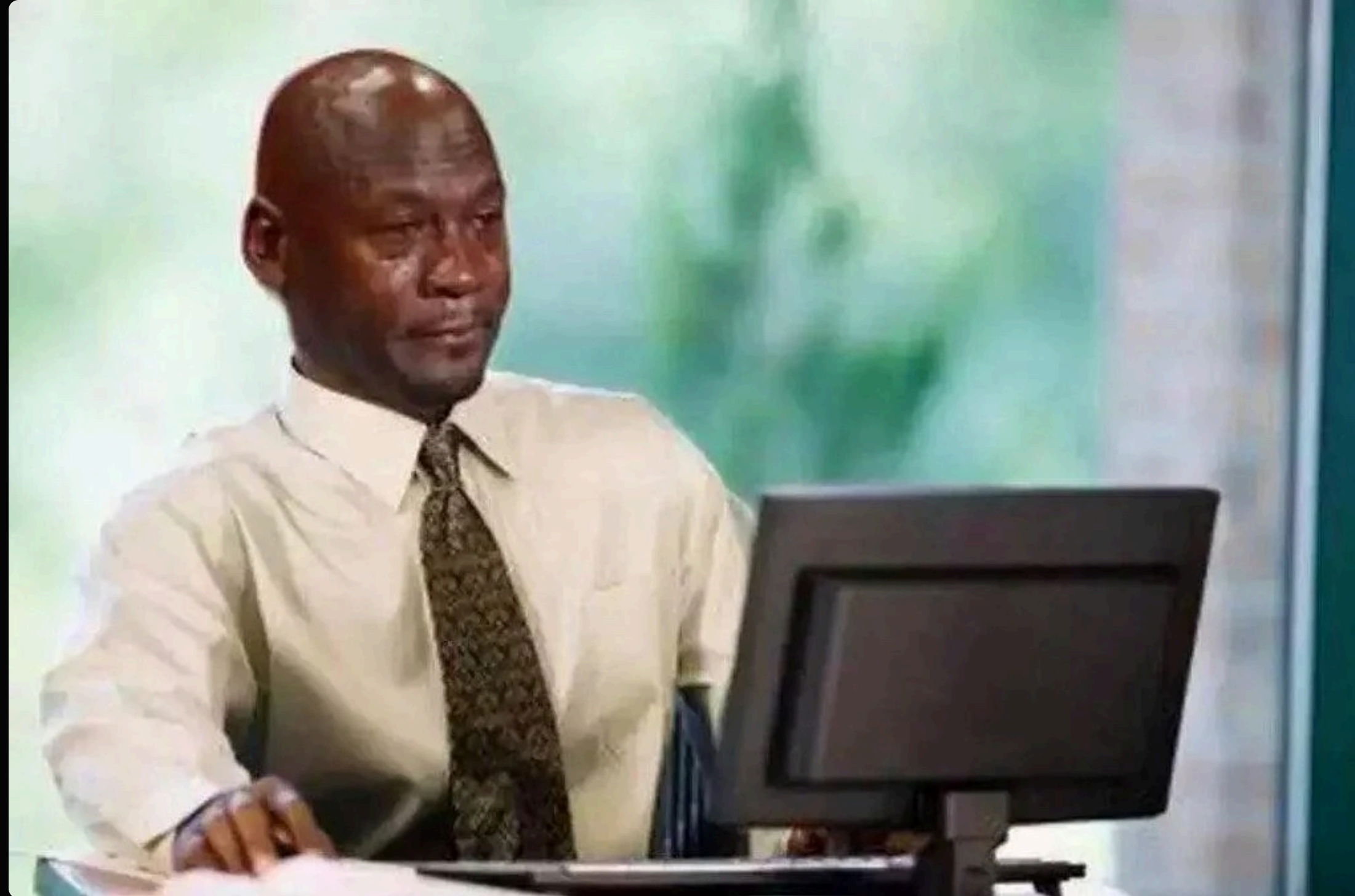
# Evaluation

Vibe Checking




# Evaluation

Build a Test Dataset



# Conclusion

## The Jonas Simulator



**The Jonas Simulator**  
↔ Nicht gelistet

**Innovation Hacking**  
53 Abonnenten

Abonnieren

0 | Teilen | Clip | ...

Top Chat ▾

- Innovation Hacking** What is Waifu
- Innovation Hacking** Can you tell me about Deepfakes?
- TheMayeah Tell me about the Shitp
- Jonas Mayer What framework are y Nao Robot?
- Jonas Mayer Who is Martin?

Willkommen im Livechat! Bitte achte deiner Privatsphäre und halte dich an unsere Community-Richtlinien.  
[Weitere Informationen](#)

Jonas Mayer Can you touch your own shoulders? ❤️

Chat... 🗨️



# Thank you!



**Christoph Bergemann**

Principal Consultant

christoph.bergemann@tngtech.com



**Jonas Mayer**

Senior Consultant

jonas.mayer@tngtech.com

For more Projects and Talks visit

**[innovation-hacking.com](https://innovation-hacking.com)**