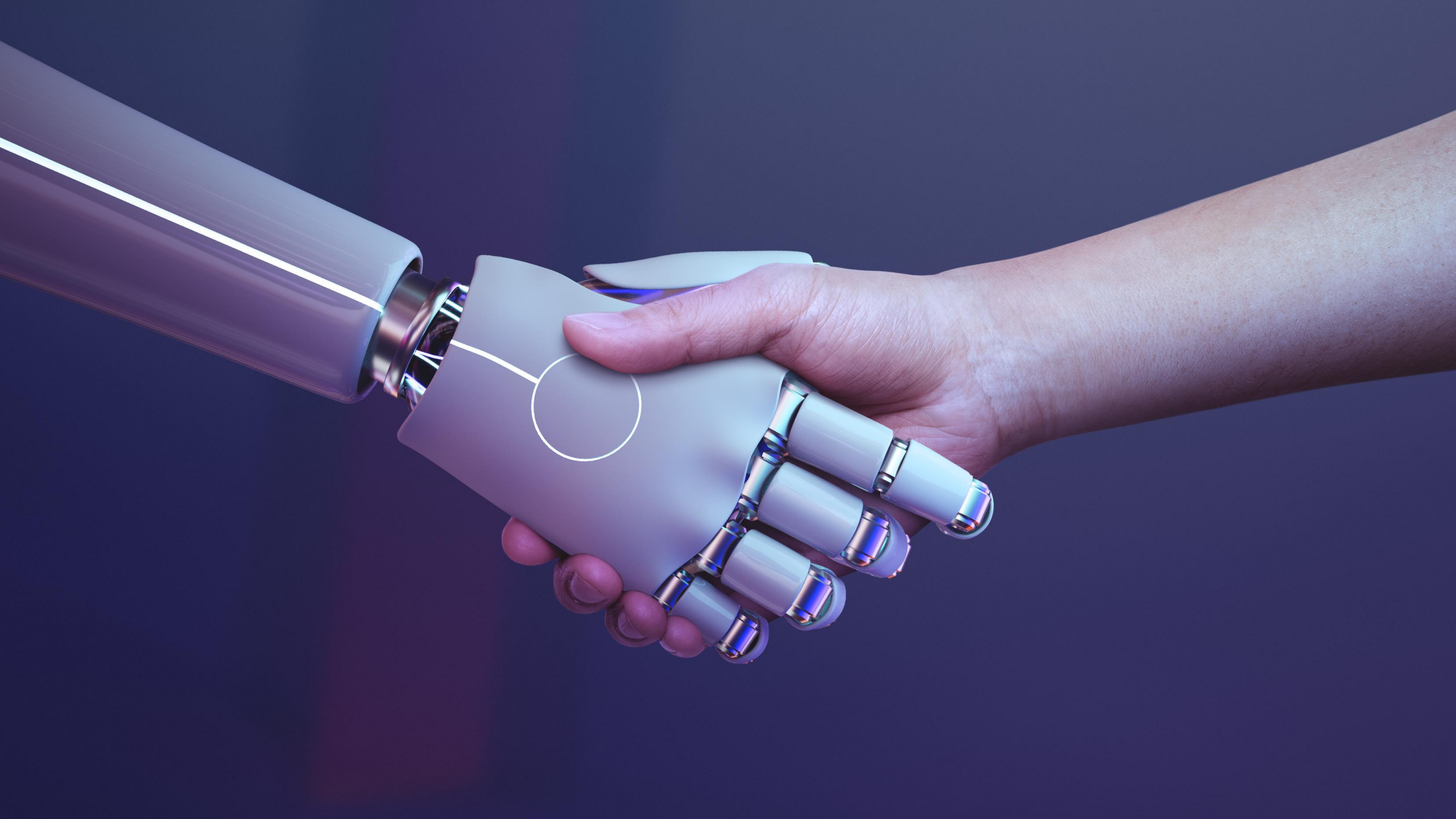
- Prompt Engineering
- What is RAG and how to use?
- Crew AI
- Q&A

Prompt Engineering RAG Crew AI

Ryan Heida

AI.Mastery.Student


Prompt Engineering
RAG
Crew AI



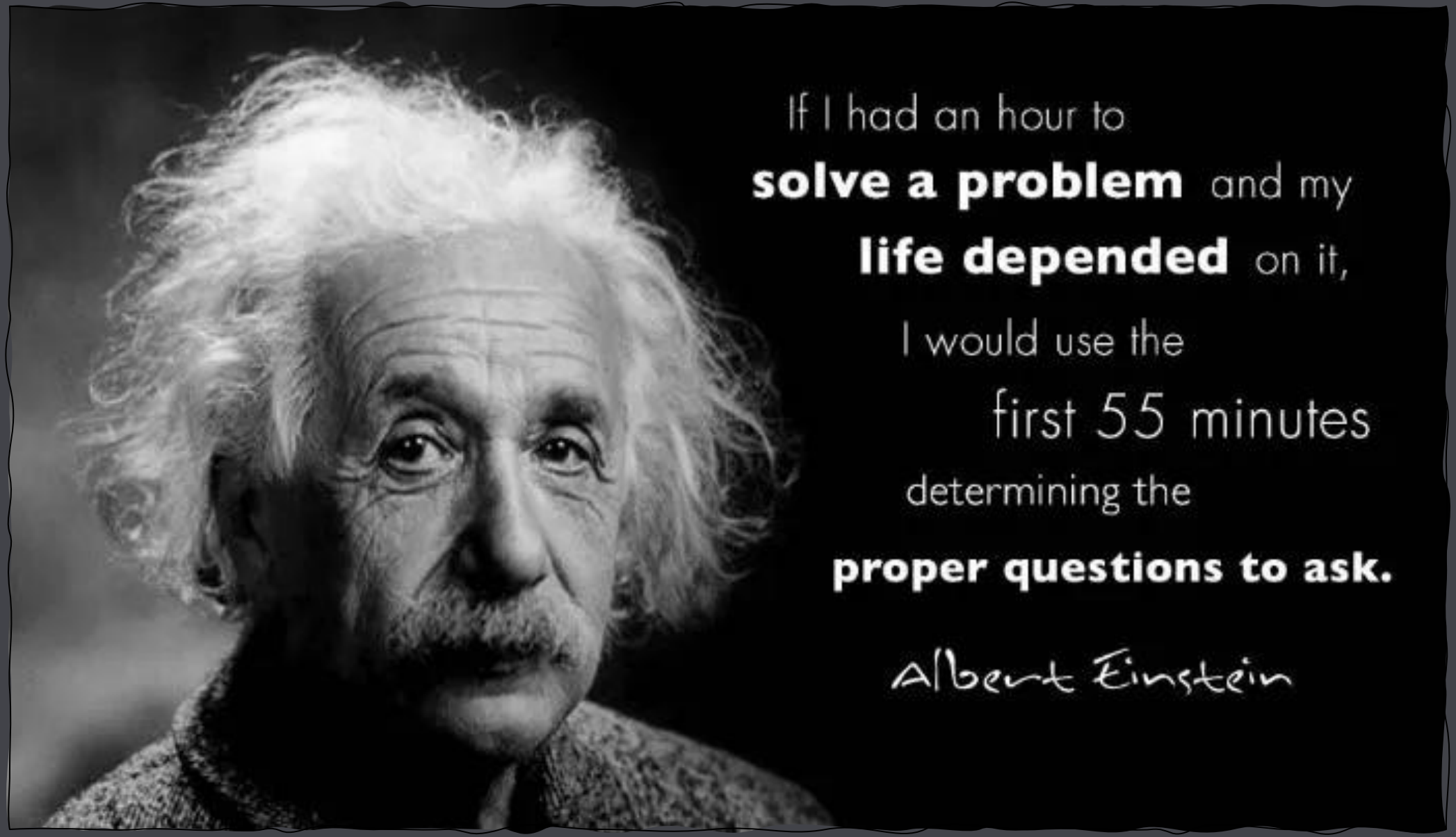
دورهمی هوش مصنوعی - مدلهای زبانی بزرگ (LLMs)

قسمت اول - ارائه‌ی اصلی

رایان هیدا

Review of Previous Session (LLMs)

Review of Previous Session (LLMs)


If I had an hour to
solve a problem and my
life depended on it,
I would use the
first 55 minutes
determining the
proper questions to ask.
Albert Einstein

What is Prompt Engineering ?

Sam Altman
@sama

writing a really great prompt for a chatbot persona is an amazingly high-leverage skill and an early example of programming in a little bit of natural language

1:23 am · 21 Feb 2023 · **782.8K** Views

476 Retweets **111** Quote Tweets **4,917** Likes

The Art of Asking

> “The art and science of asking questions is the source of all knowledge” — Thomas Berger

What is Prompt Engineering?

What is Prompt Engineering?

The process of

What is Prompt Engineering?

The process of

Designing and optimizing prompts

What is Prompt Engineering?

The process of

Designing and optimizing prompts

To effectively communicate with AI models.

What is Prompt Engineering?

The process of

Designing and optimizing prompts

To effectively communicate with AI models.

Programming language of speaking to AI

Why is it Important?

ENHANCES MODEL PERFORMANCE

Instead of asking "**Tell me about AI**“,
Ask "**Explain how artificial intelligence is used in healthcare to improve patient outcomes.**"

Why is it Important?

IMPROVES TASK-SPECIFIC

A prompt like "**Analyze this sales data and summarize key trends for the last quarter**" helps the **AI focus** specifically on **identifying trends** in the data rather than providing **irrelevant information**.

Types of Prompts

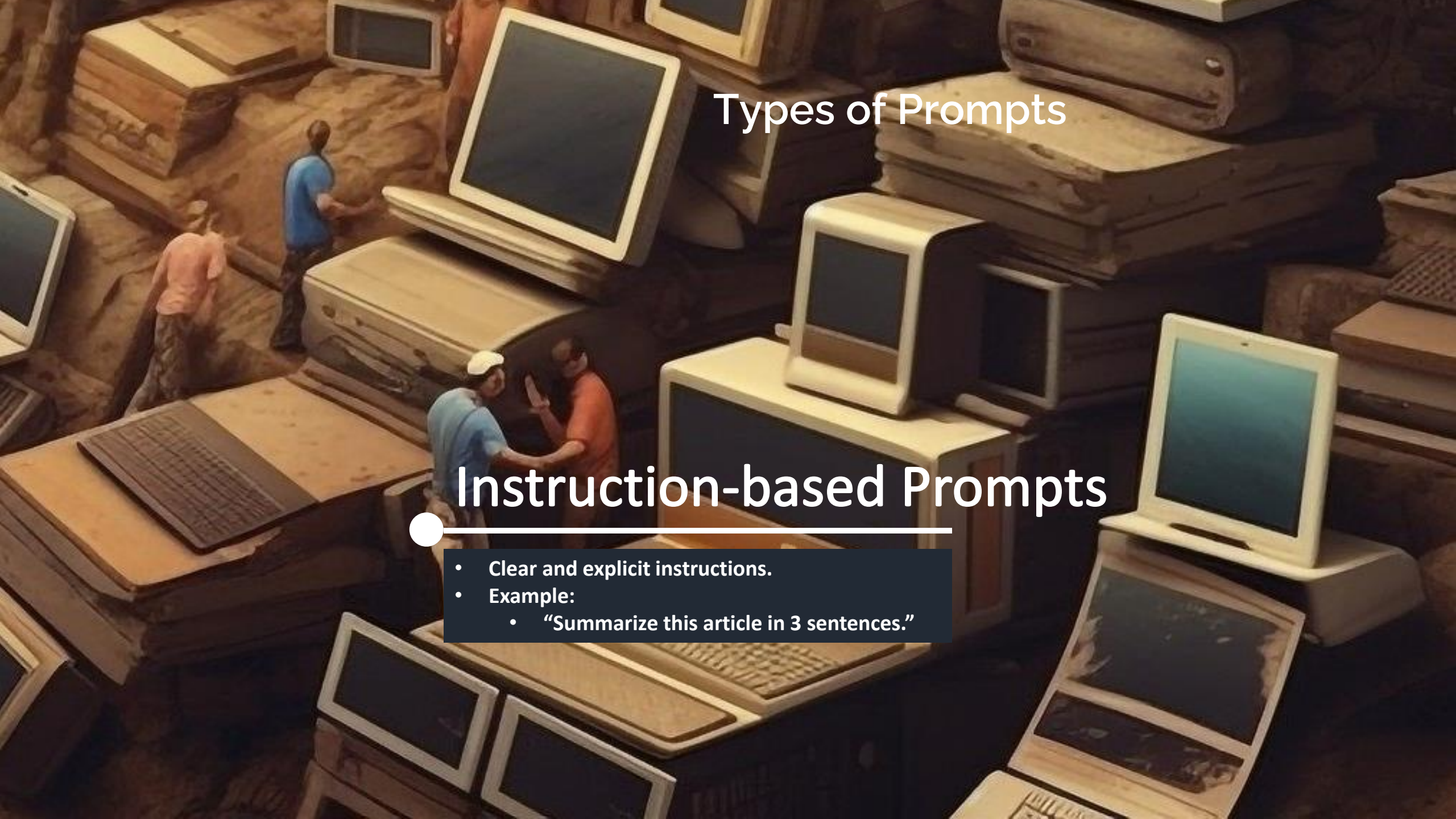
Zero-shot Prompts

- **No examples provided; relies on general knowledge.**
- **Example: “What are the benefits of renewable energy?”**

Types of Prompts

Few-shot Prompts

- Includes examples to guide the response.
- Example:
  - “Translate the following:
    English: Hello
    Spanish: Hola
    English: Thank you
    Spanish:”



Types of Prompts

Instruction-based Prompts

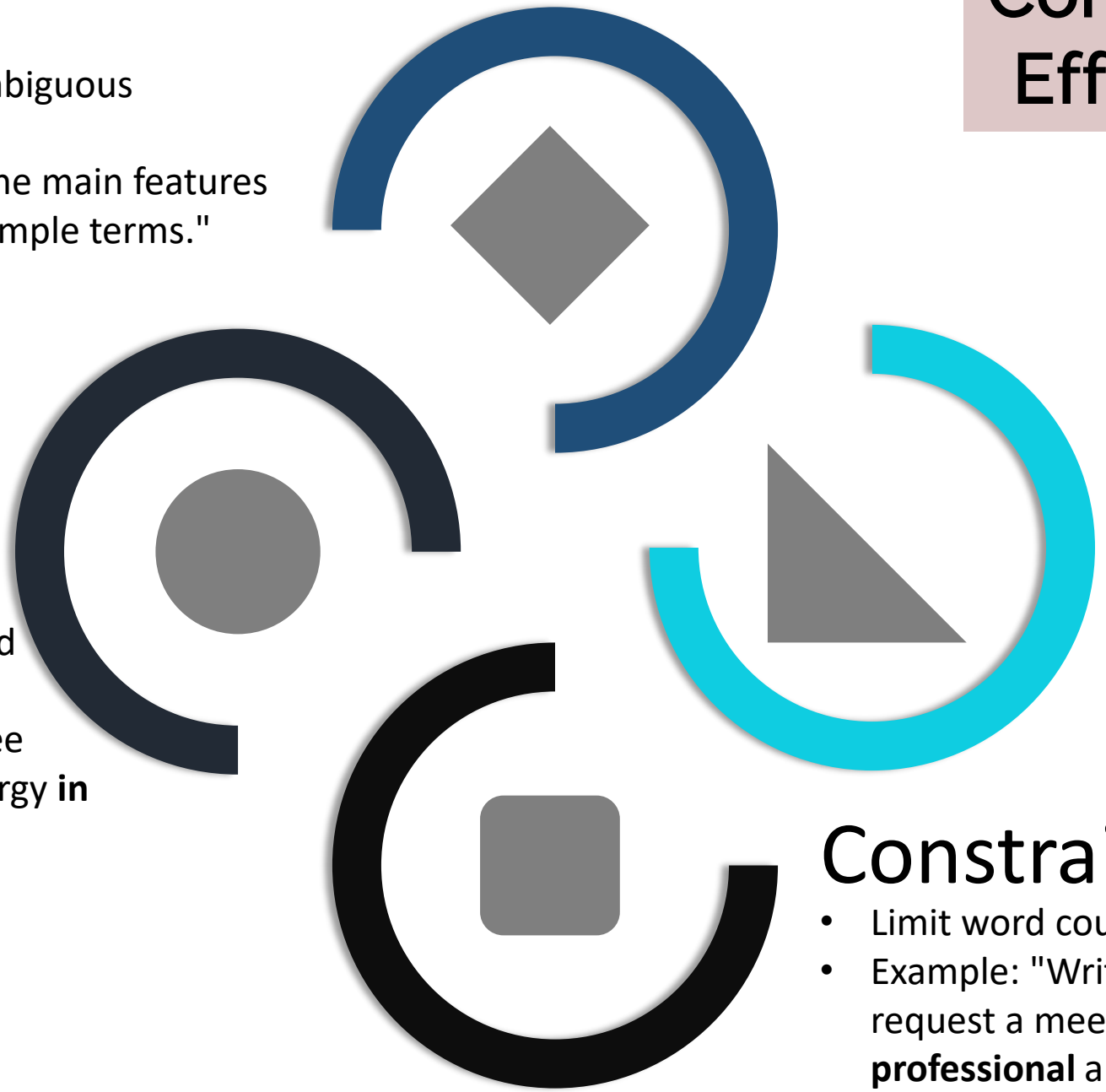
- Clear and explicit instructions.
- Example:
  - “Summarize this article in 3 sentences.”

Components of an Effective Prompt

Components of an Effective Prompt

Clarity

- Use simple and unambiguous language.
- Example: "Describe the main features of an electric car in simple terms."



Context

- Provide relevant background or examples.
- Example: "Explain the concept of photosynthesis **assuming the audience has a basic understanding of biology.**"

Specificity

- Define the task and desired format.
- Example: "List the top three benefits of renewable energy **in bullet points.**"

Constraints

- Limit word count, tone, or style.
- Example: "Write a **formal email** to request a meeting, ke**eping the tone professional** and **under 100 words**."

Common Techniques

Common Techniques

Chaining

1

Break complex tasks into smaller, manageable steps.
Example: "Step 1: Outline the essay. Step 2: Write the introduction."

Common Techniques

Iterative Refinement

1 2

Continuously tweak the prompt to improve results.

3

Common Techniques

Use of Tokens

Tokens can define sections or highlight important elements
Like “A” or `A` or “””A””” or [A] even space.

Retrieval

Augmented

Generation

Remember LLMs Challenges...

What Can't LLMs Do?

LACK OF TRUE UNDERSTANDING

LLMs don't "understand" text the way humans do.
They rely on **statistical patterns** in data rather than reasoning or comprehension.

We want data that is…

We want data that is…

Accurate

Up to date

We want data that is…

Accurate | Up to date

2021

We want data that is…

Accurate | Up to date

2025

2021

LLM Model

We want data that is…

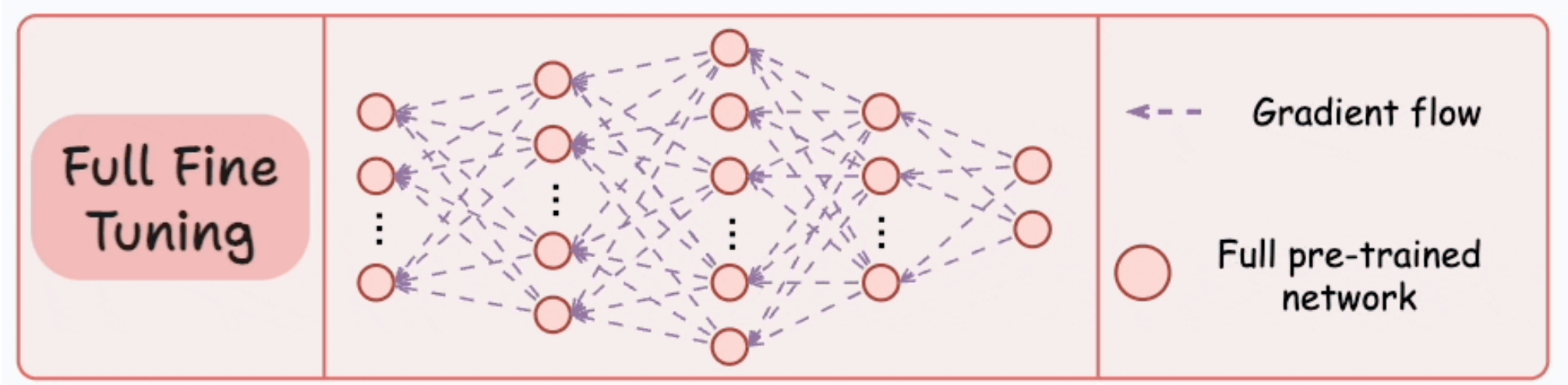
Accurate | Up to date

2021

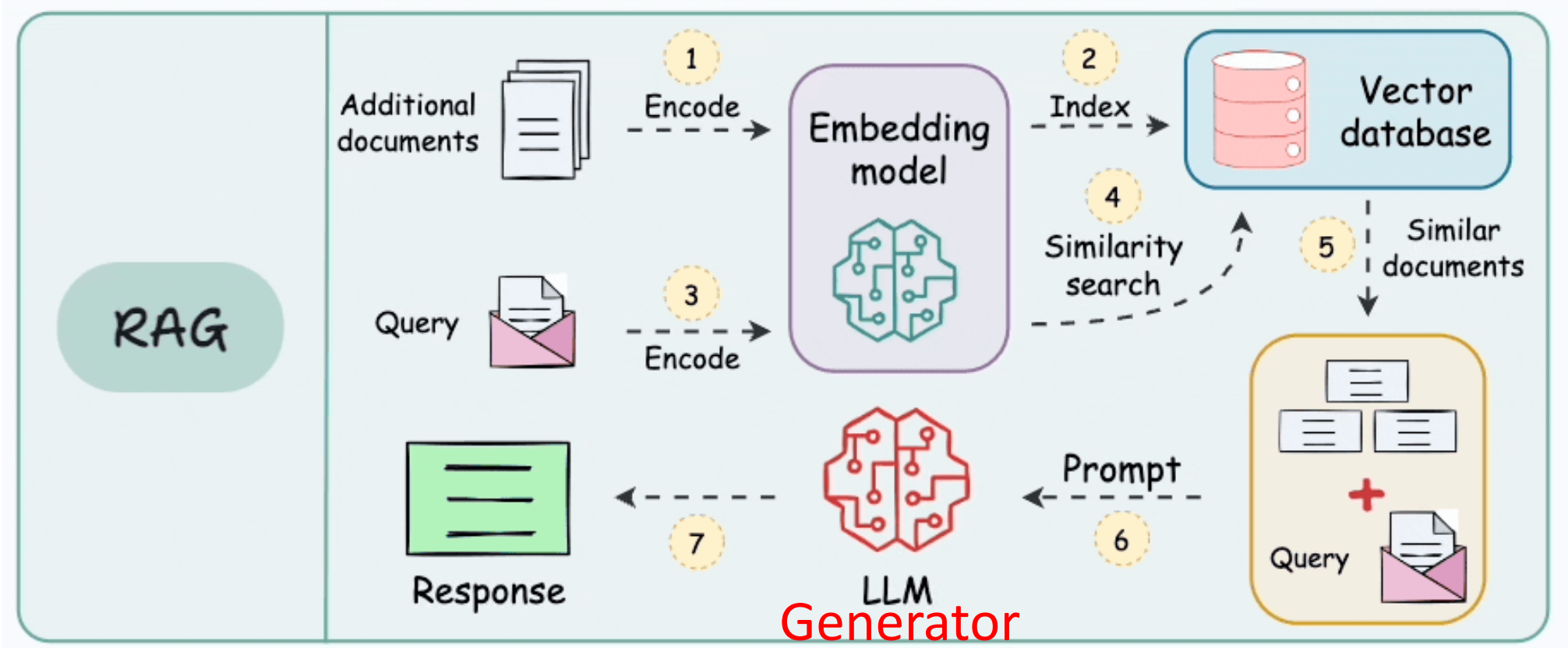
2025

LLM Model

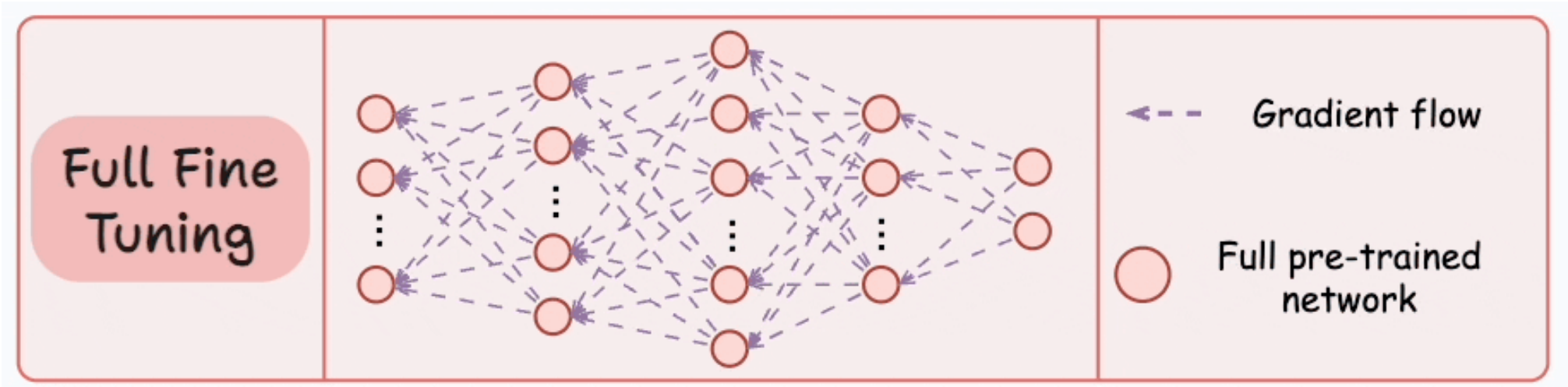
1


Full Fine Tuning
Gradient flow
Full pre-trained network

2

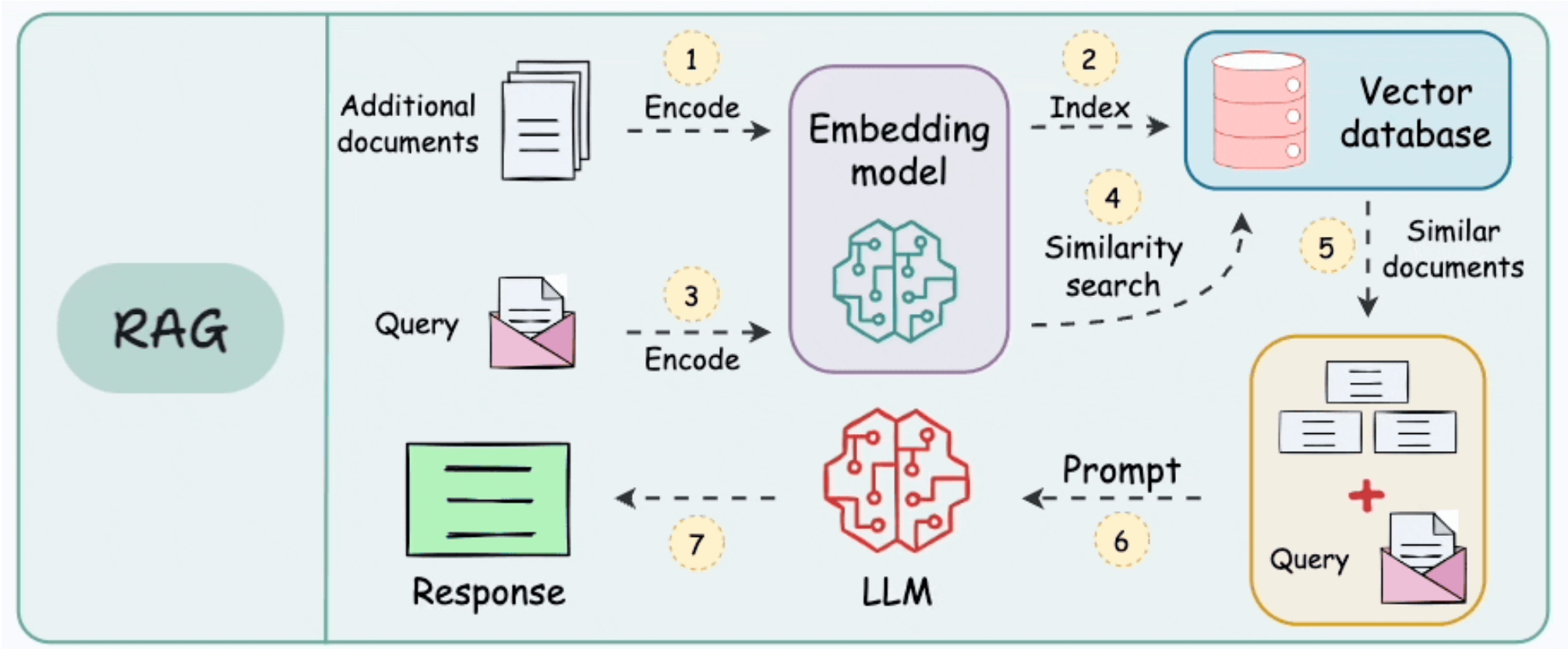

RAG
Additional documents
1
Encode
Embedding model
2
Index
Vector database
4
Similarity search
5
Similar documents
Query
3
Encode
Prompt
6
Query
7
Response
LLM
Retriever
Generator

1

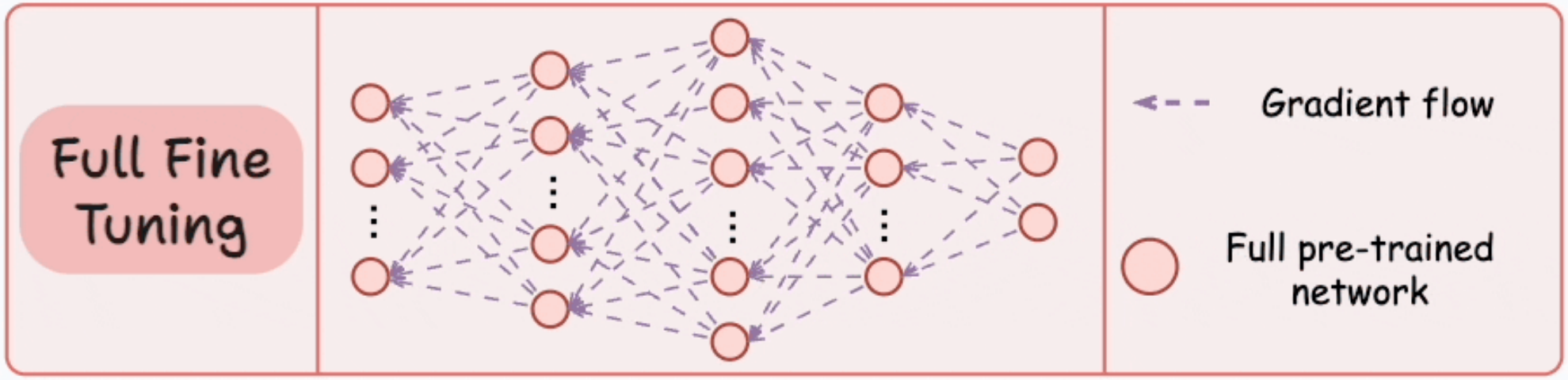

Full Fine Tuning
Gradient flow
Full pre-trained network

+ Data

2


RAG
Additional documents
1
Encode
Embedding model
2
Index
Vector database
4
Similarity search
5
Similar documents
Query
3
Encode
Prompt
6
Query
7
Response
LLM

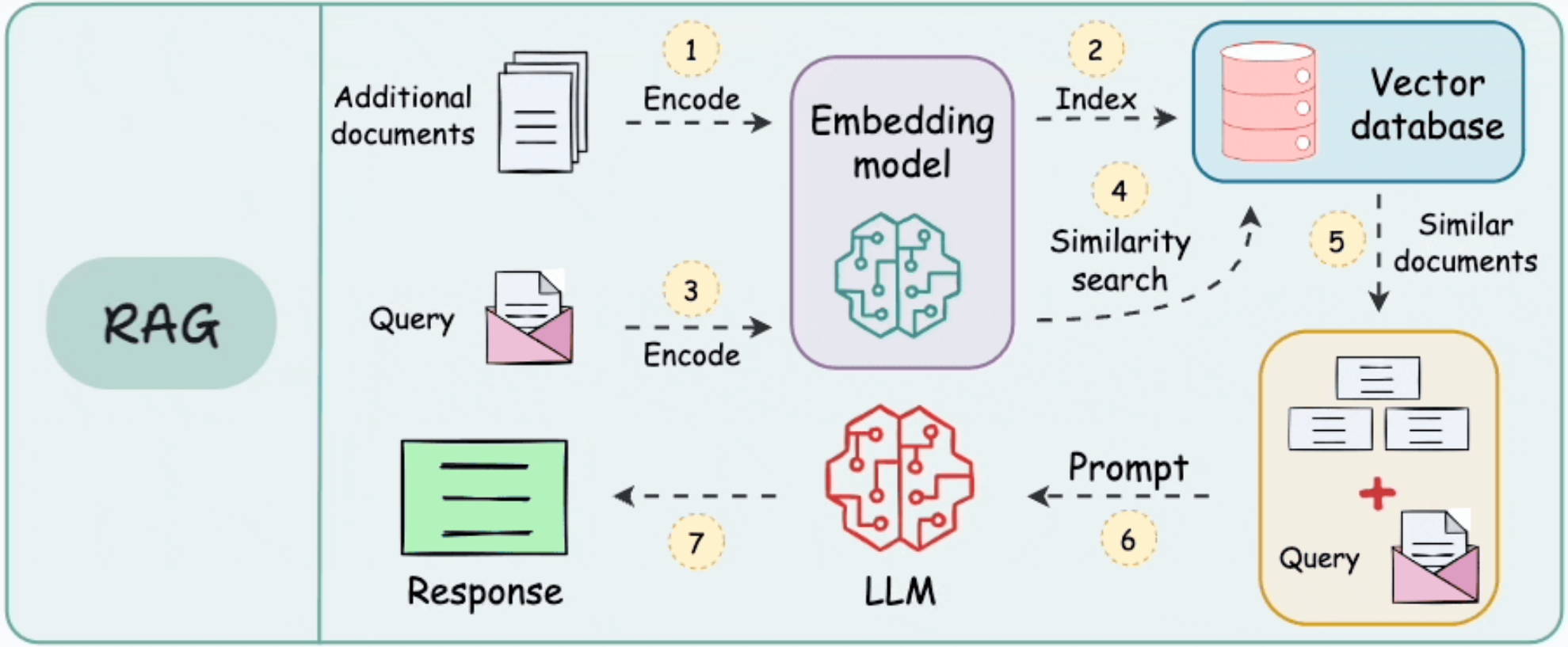
1



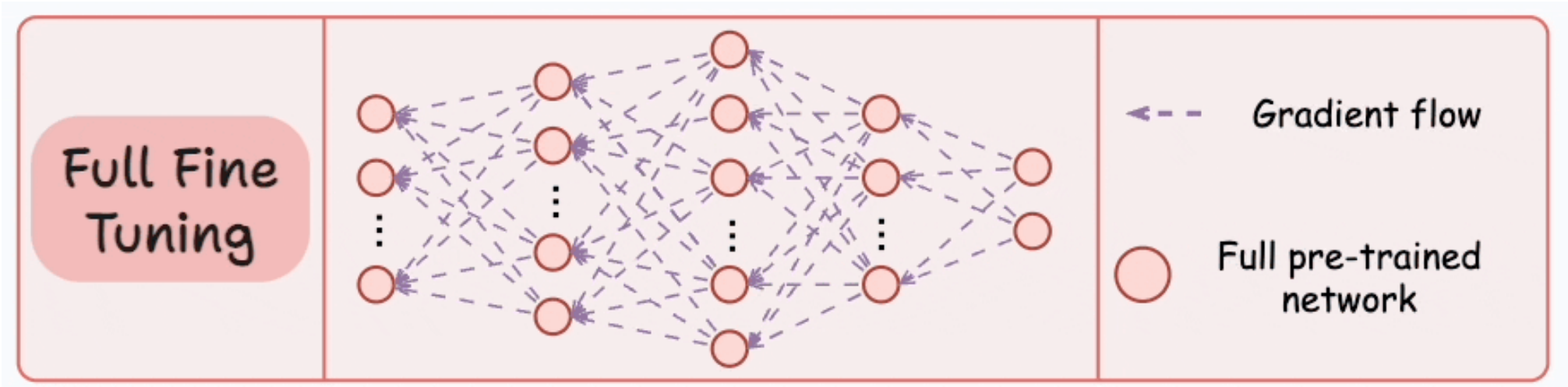
+ Data

+ Source

2



1

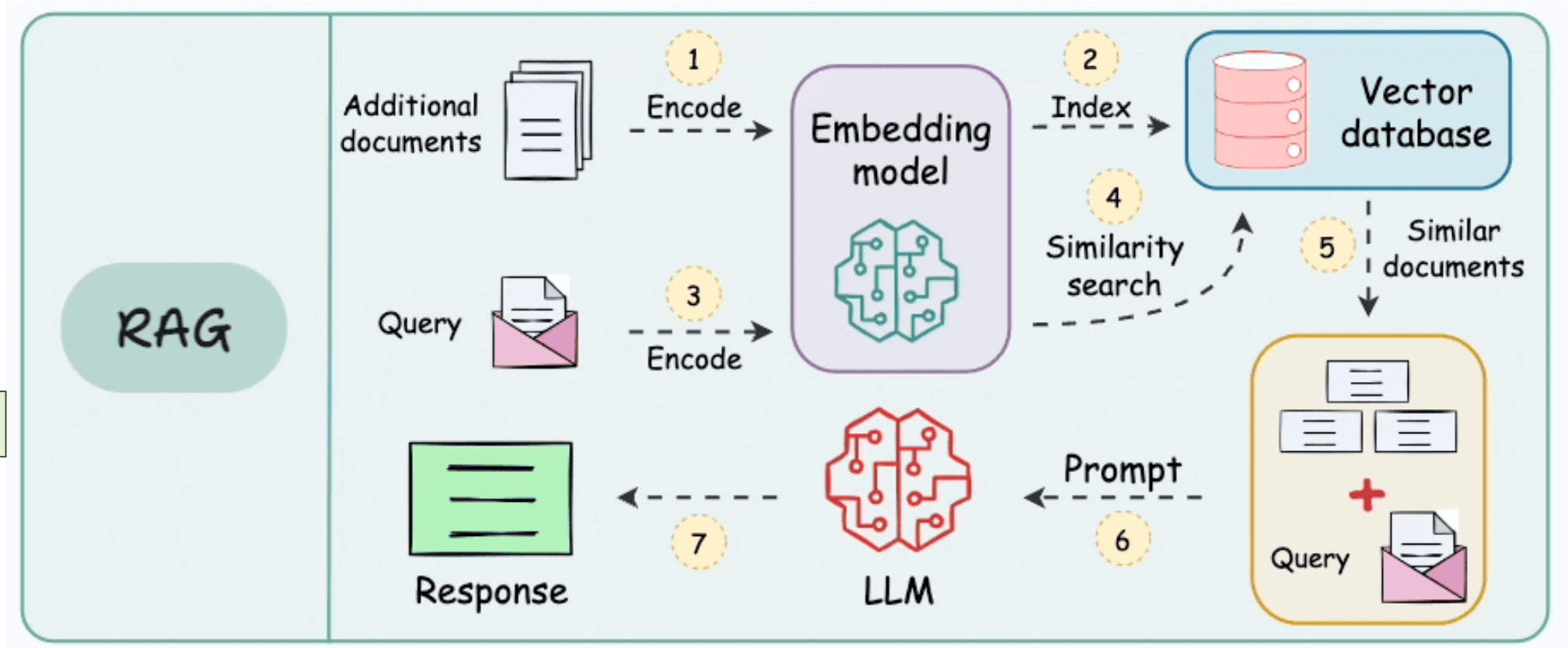


+ Data

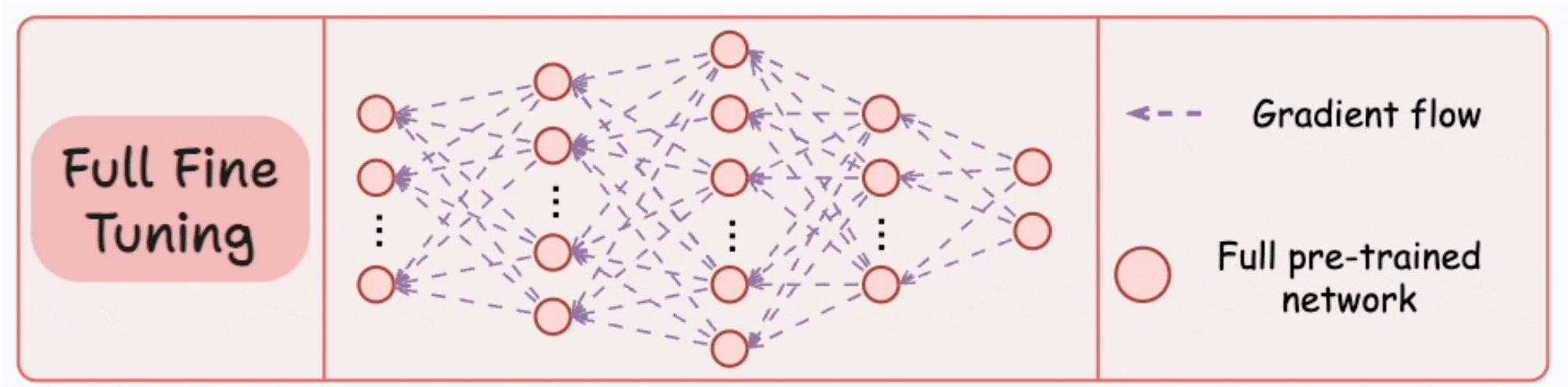
+ Source

2

+ Transparency



1



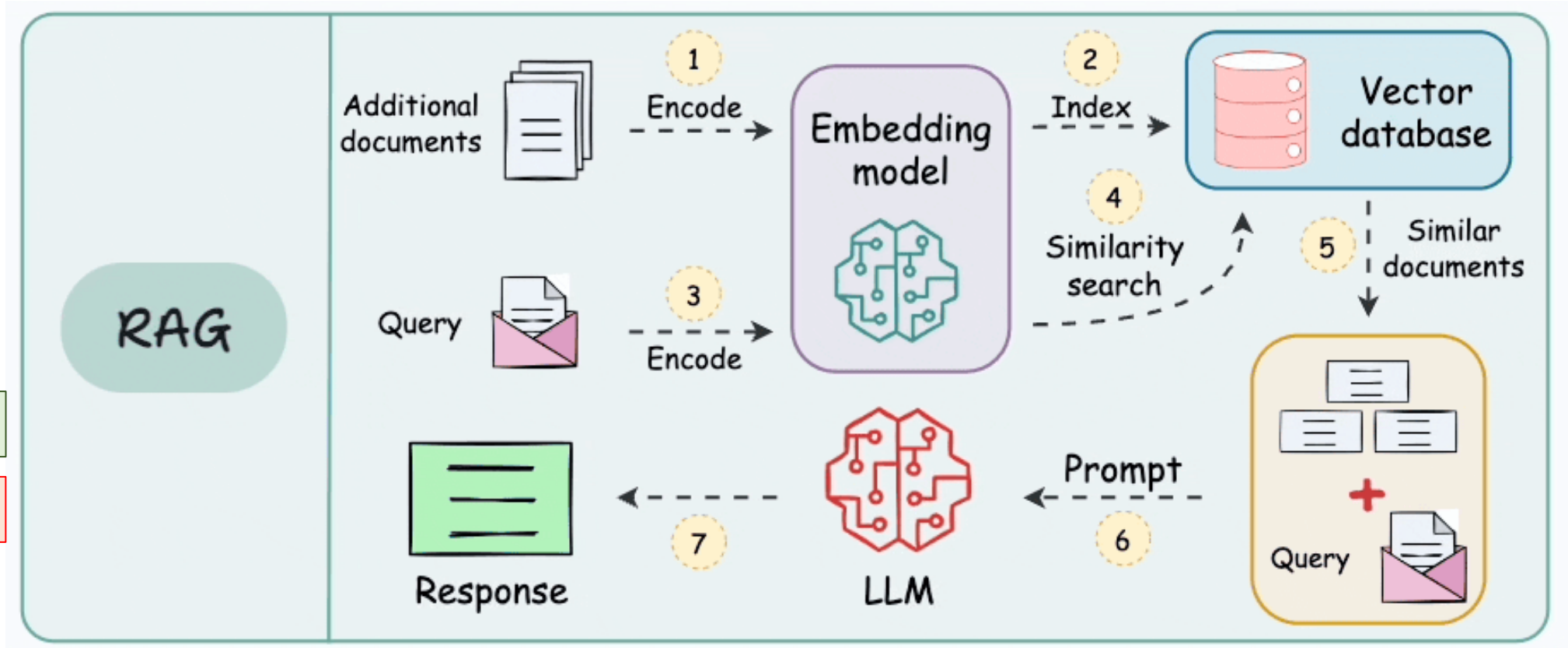
+ Data

+ Source

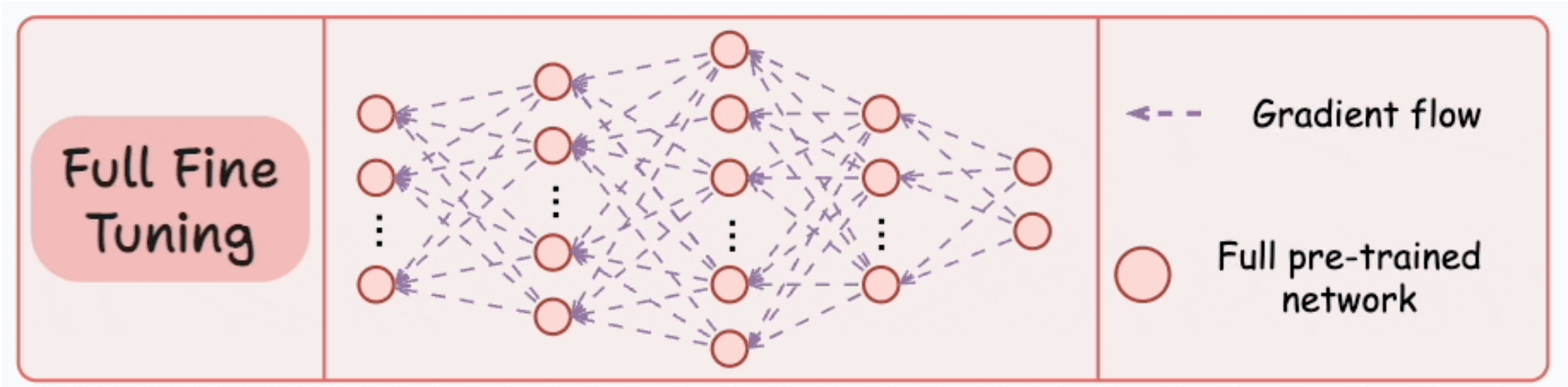
2

+ Transparency

- Hallucination



1



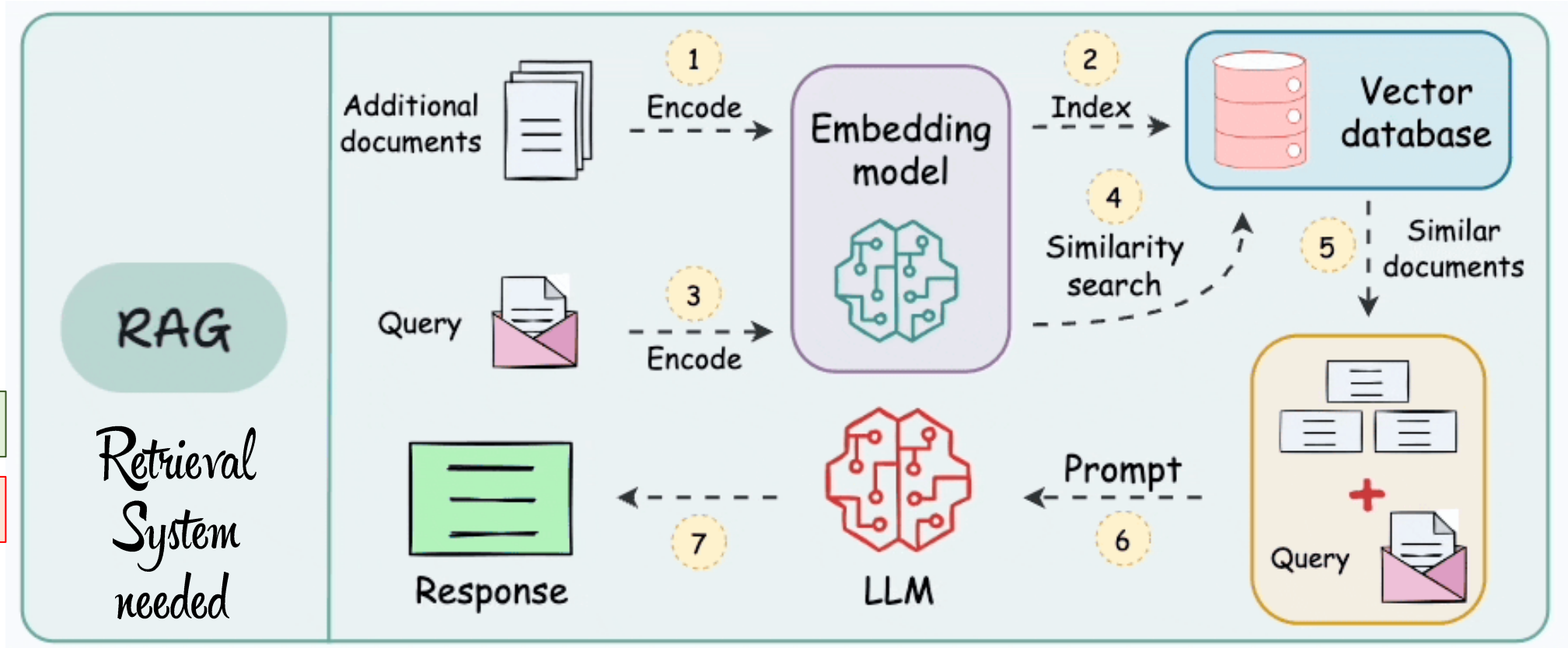
+ Data

+ Source

2

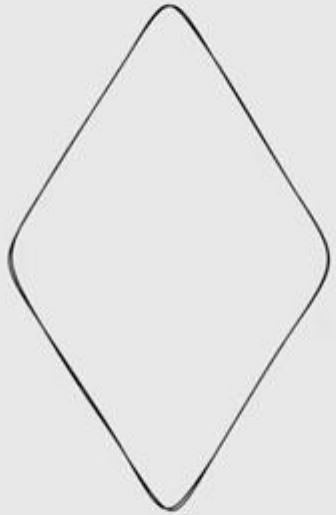
+ Transparency

- Hallucination

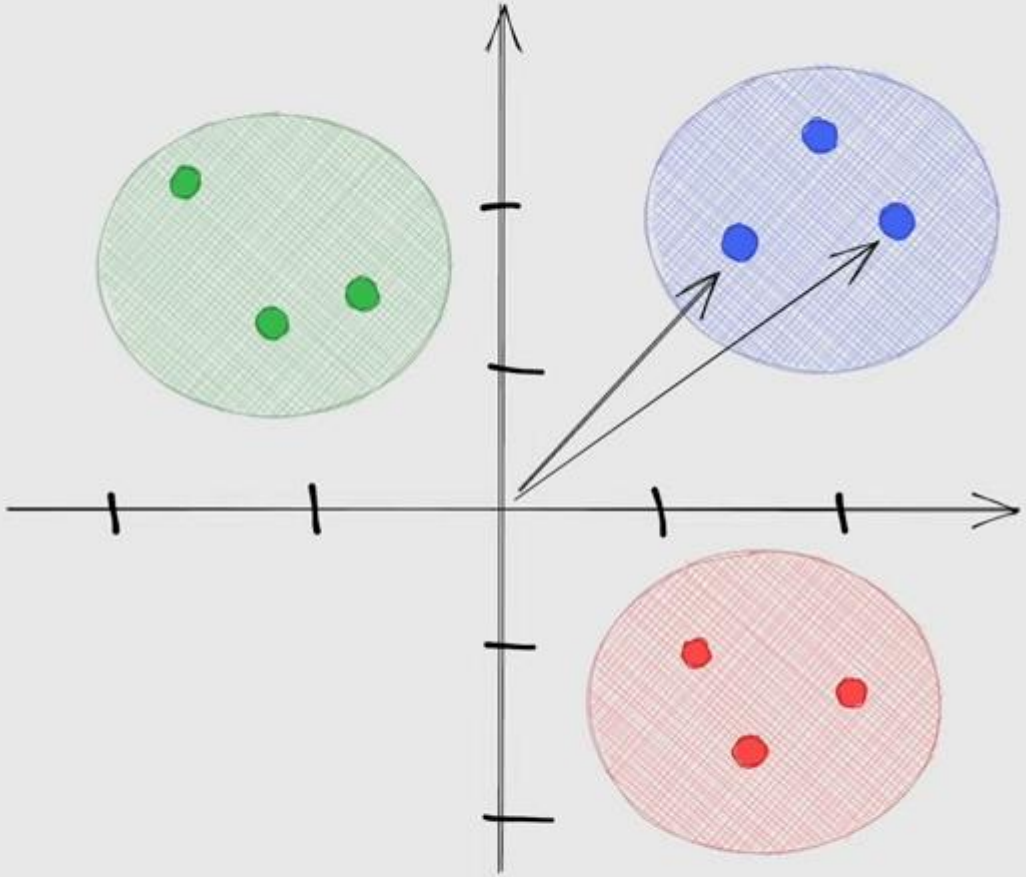


vector embeddings (2D example)

king
man
woman
apple
banana
orange
football
golf
tennis

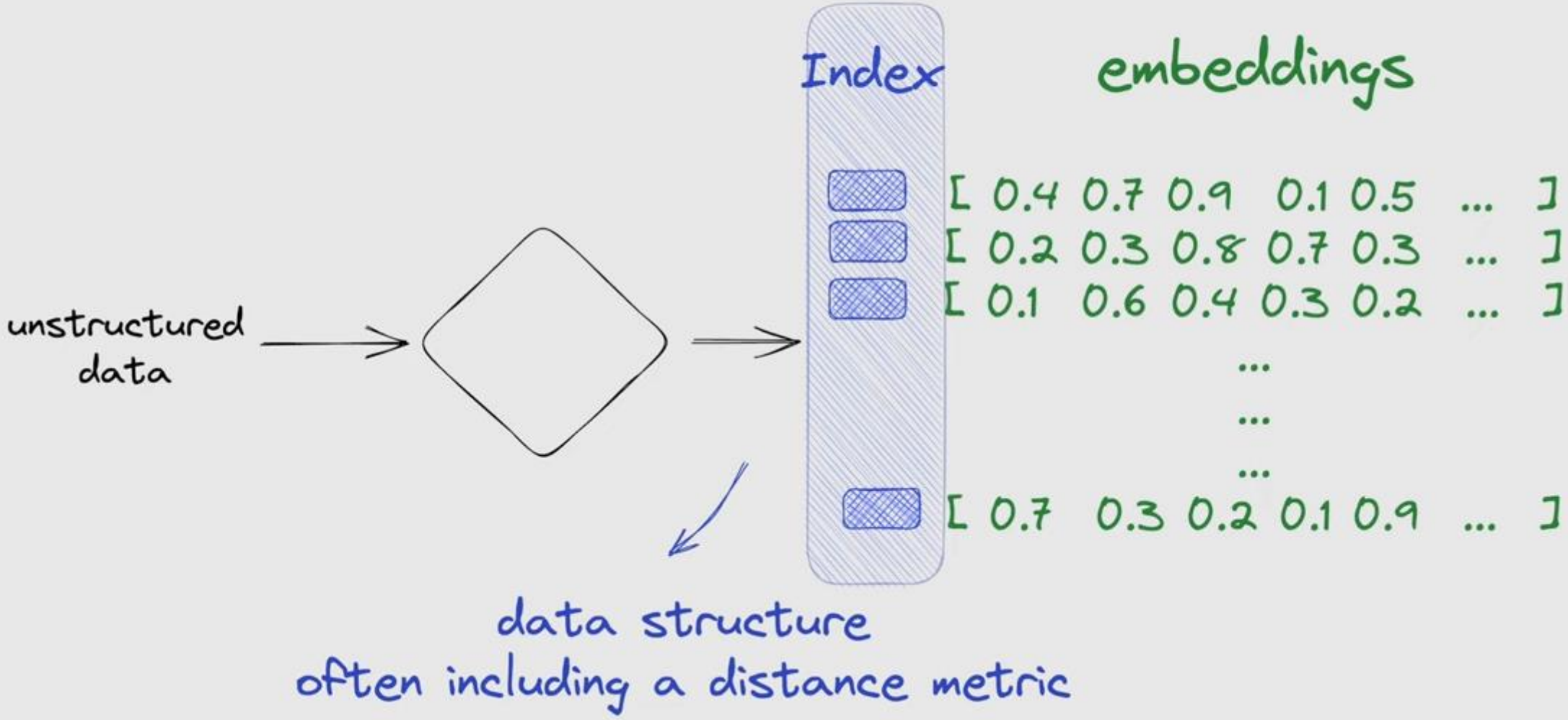


[4 5]
[3 3.5]
[5 3]
[2.5 -2]
[2.5 -3]
[4 -2.5]
[-3 4]
[-1.5 3]
[-2 2]



$$d = \sqrt{(x2 - x1)^2 + (y2 - y1)^2}$$

vector indexing



use cases

1. long-term memory for LLMs
2. semantic search: search based on the meaning or context
3. similarity search for text, images, audio, or video data
4. recommendation engine

Vector Databases

Pinecone

Qdrant

Weaviate

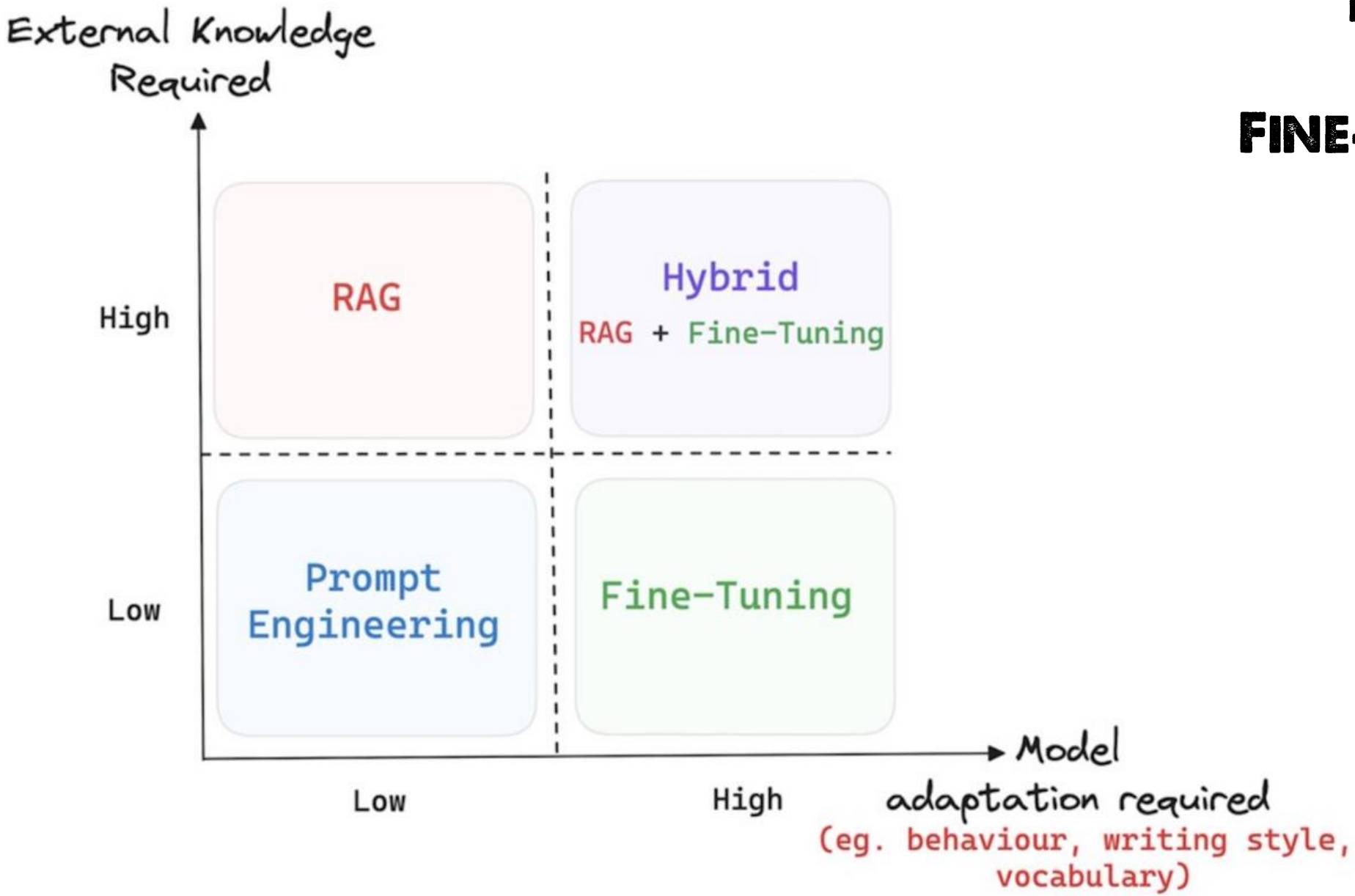
Milvus

Chroma

Vespa

Redis

RAG Vs Fine-Tuning



RAG Vs Fine-Tuning

Product Documentation Chatbot

Sources

Industries

Capability	RAG	Fine Tuning
External Knowledge Needed	Yes	No
Desire to Minimize Hallucination	Yes	No
Dynamic Data Source	Yes	No
Source Tracing Interpretability Required	Yes	No
Desire to override Core LLM Behavior	No	Yes
Is Training Data Available for fine-tuning?	No	Yes

AI Agents

Crew AI

Hands on code...

RAG

References

Vector Databases simply explained! (Embeddings & Indexes)

RAG vs. Fine Tuning

You can have presentations on these topics…

Vector Databases

Fine-Tuning

Retriever System in RAGs

AI Agents (CrewAI) Tutorial

Multi AI Agent Systems with crewAI

My Recommendations...

News (10 min)

...

links.ryanheida.com

@RYANHEIDA

Thank You For Watching