Mastering Production Challenges in RAG

Mastering Production Challenges in RAG

Production

Challenges RAG

Shrijayan Rajendran

𝕏 - @rshrijayan

/tw - ThoughtWorks

AI Engineer

Agenda

Why

What

When

Why is Advance RAG

What is Advance RAG

Evaluation

Prompting

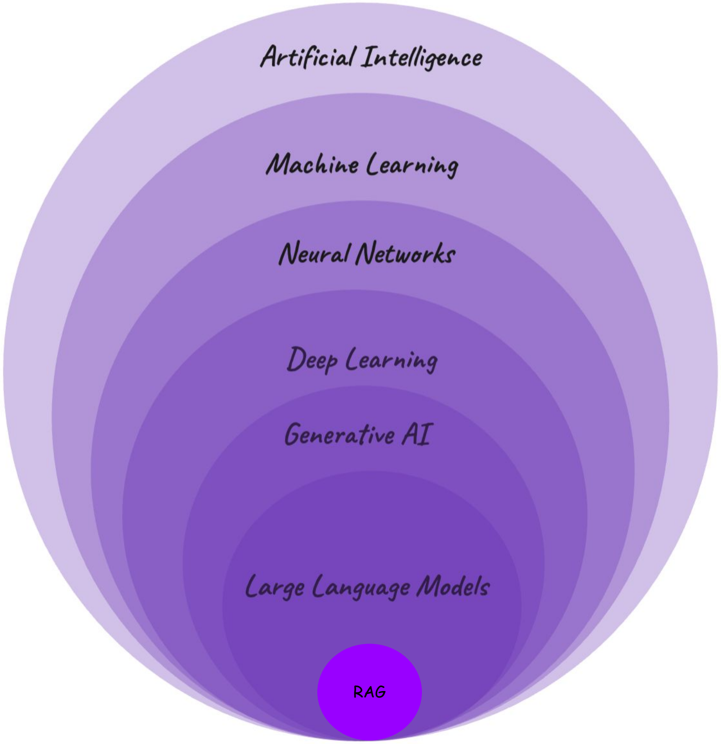
HandsOn

Come Inside here

https://shorturl.at/PSWY8

RAG Demo

Retrieval-Augmented Generation (RAG)



Where RAG fits?

Terminology to Know

Embedding

Eg: [0.2345, -5.55, -4.44, -3.33, -8.88, 0.123, 6.66, 9.81, -8.88, 0.123, 6.66, -3.33, 7.77, 2.22, -6.66, -5.55, -4.44, -3.33, 1.0, 2.0, 3.0, 4.0, 5.0, 0.2345]

Chunk

Similar to a row in a SQL

[0.2345, -5.55, -4.44, -3.33, -8.88, 0.123, 6.66, 9.81, -8.88, 0.123, 6.66, -3.33, 7.77, 2.22, -6.66, -5.55, -4.44, -3.33, 1.0, 2.0, 3.0, 4.0, 5.0, 0.2345]

Articles may include embedded content from other websites, which behaves as if the visitor has visited the other website directly. These sites may collect data about you and use cookies for tracking.

Vector DataBase

n-deminal DataBase

What is AI?

"Artificial intelligence, or AI, is technology that enables computers and machines to **simulate** human intelligence and problem-solving capabilities."

Narrow AI - Single Task Performer

Generative AI - Multi Task Performer

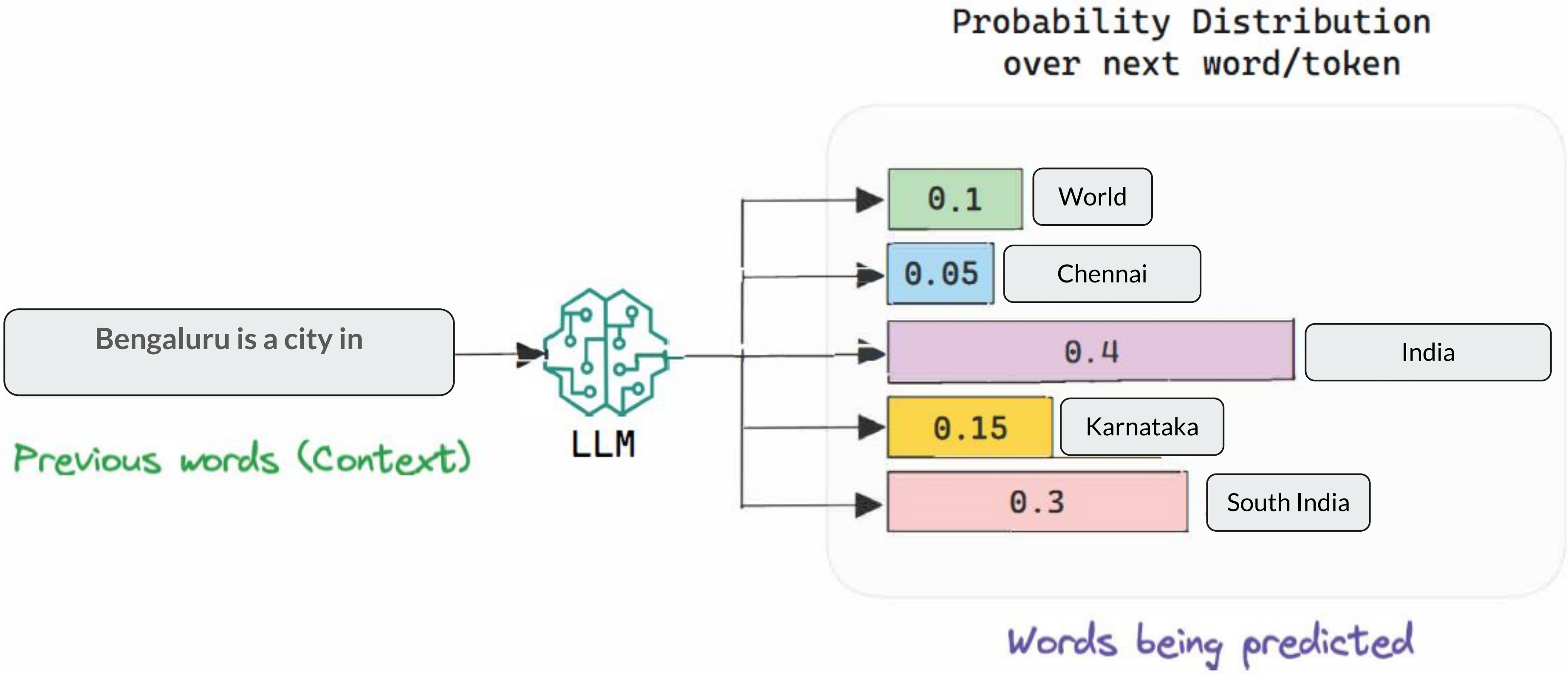
What is LLM?

A language model is a probabilistic model that assign probabilities to sequence of words.

P [“Bengaluru is a city in”] “India”

Prompt   Next Token

We usually train a neural network to predict these probabilities. A neural network trained on a large corpora of text is known as a Large Language Model (LLM)



Probability Distribution over next word/token

Bengaluru is a city in

Previous words (Context)

LLM

0.1	World
0.05	Chennai
0.4	India
0.15	Karnataka
0.3	South India

Words being predicted

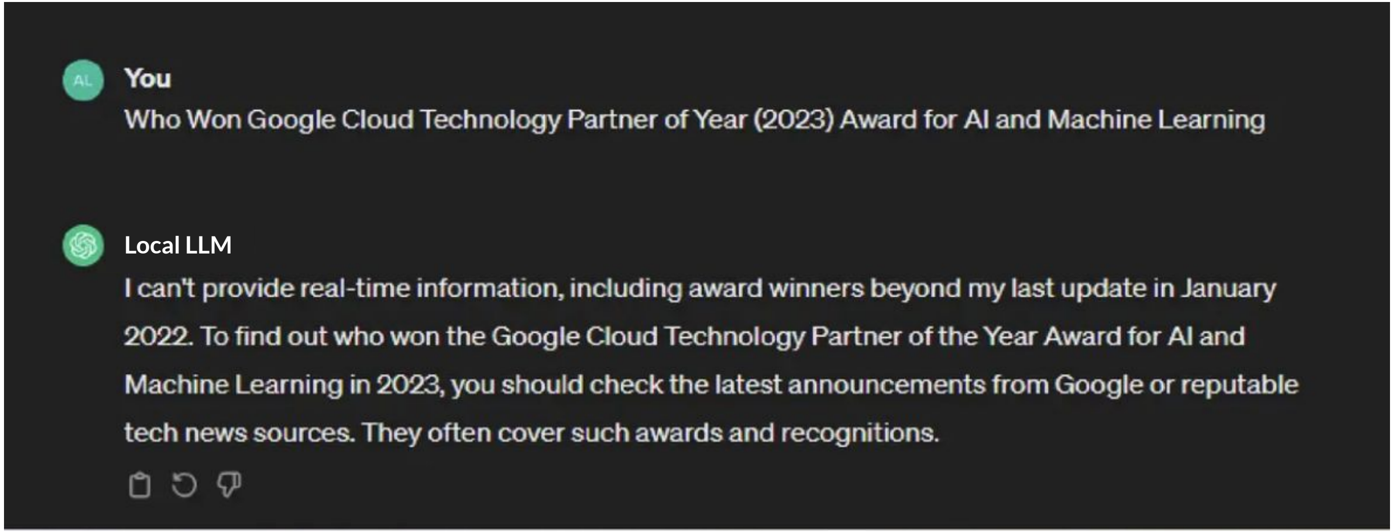
You are what you eat

A language model can only output text and information that it was trained upon.

This means, that if we train a language model only on English content, very probably it won’t be able to output Japanese or French.

To teach new concepts, we need to fine-tune the model.

Why?



AL **You**

Who Won Google Cloud Technology Partner of Year (2023) Award for AI and Machine Learning

Local LLM

I can't provide real-time information, including award winners beyond my last update in January 2022. To find out who won the Google Cloud Technology Partner of the Year Award for AI and Machine Learning in 2023, you should check the latest announcements from Google or reputable tech news sources. They often cover such awards and recognitions.

What is the problem we are trying to solve?

Agenda

~~Why~~

What

When

Why is Advance RAG

What is Advance RAG?

Evaluation

Prompting

HandsOn

What is RAG?

Example

What is the return policy of the company named XYZ?

This cannot be answered by a LLM

Solution

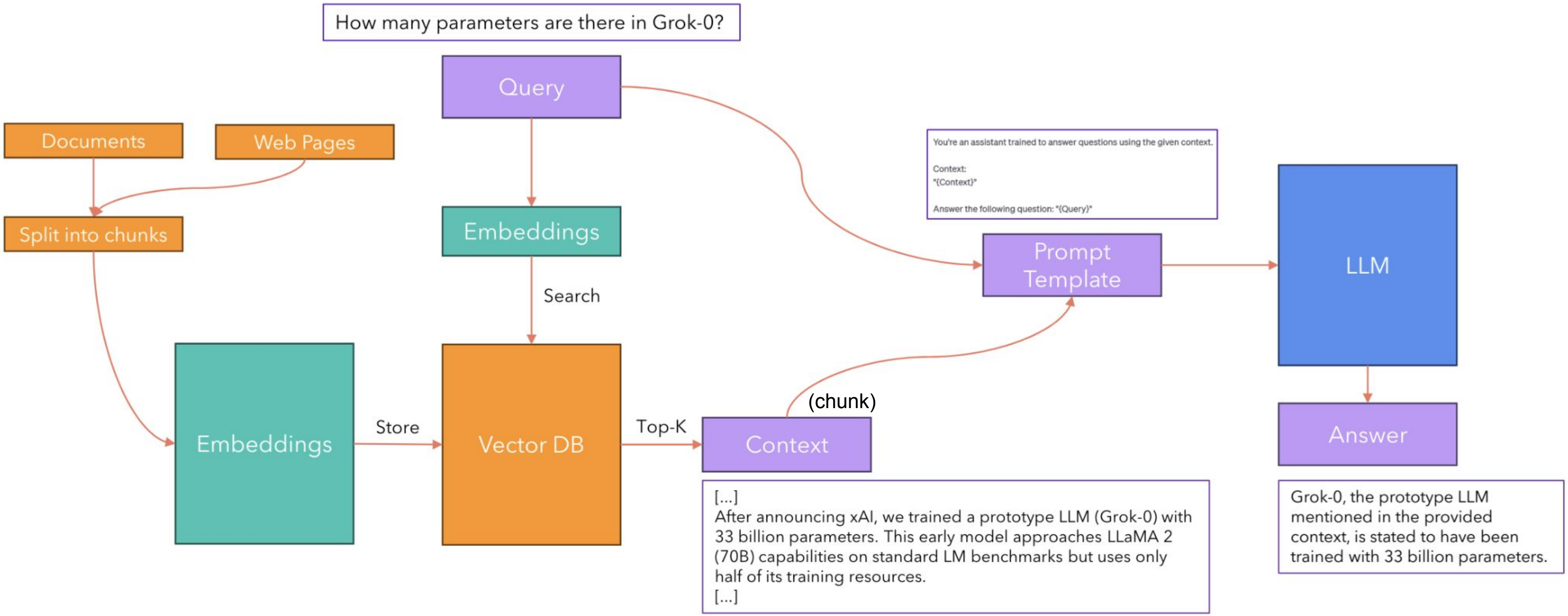
Here comes the RAG where we can upload the PDF and answer the Question of the live DATA.

Retrieval Augmented Generation

Why RAG?

- **Improves accuracy of large language models (LLMs)**
- **Leverages external knowledge sources**
- **Boosts domain-specific performance**
- **Easier to implement than fine-tuning**
- **Improves trust and transparency**

RAG Architecture



How many parameters are there in Grok-0?

You're an assistant trained to answer questions using the given context.

Context:
"{Context}"

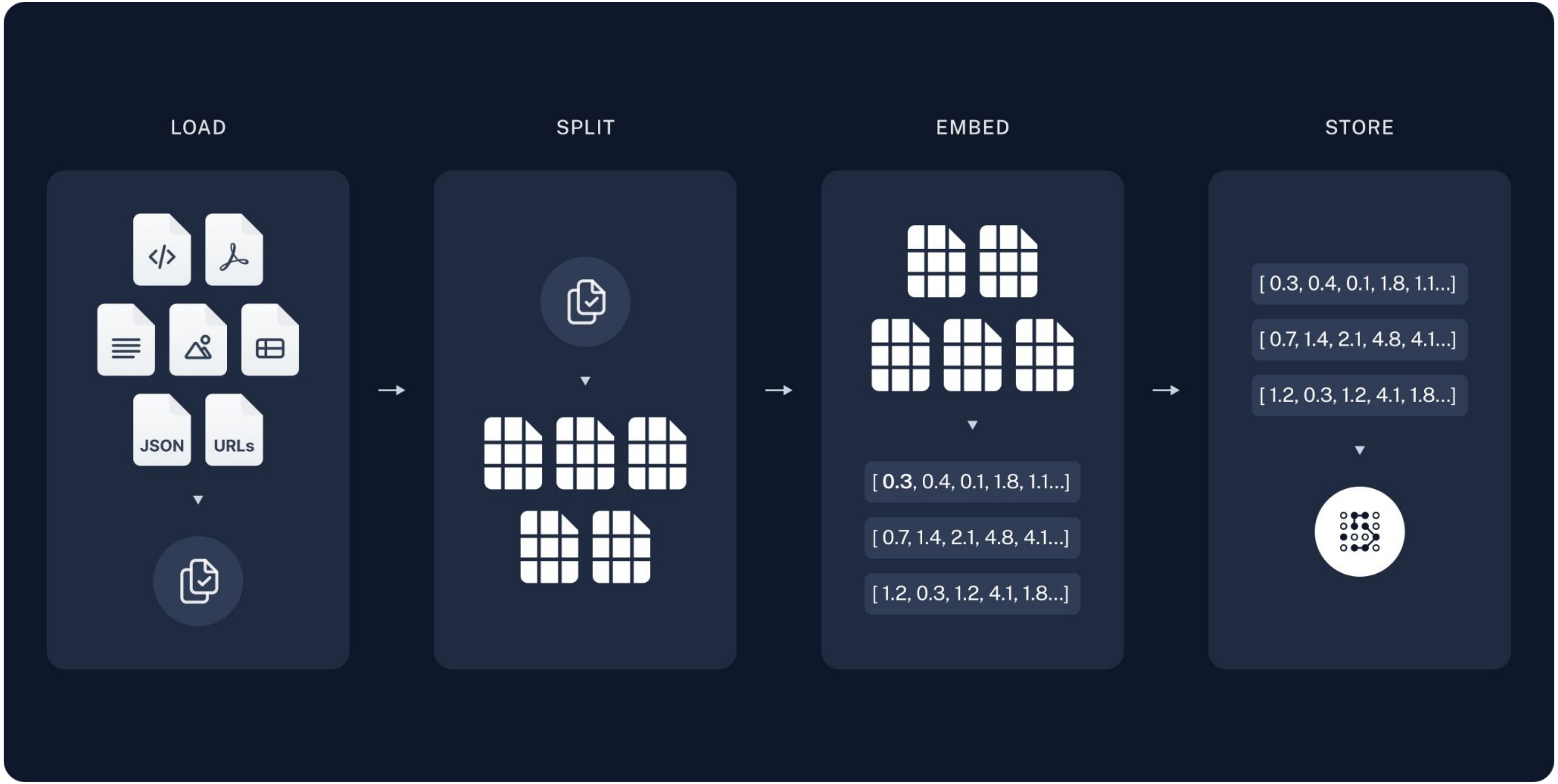
Answer the following question: "{Query}"

[...]
After announcing xAI, we trained a prototype LLM (Grok-0) with 33 billion parameters. This early model approaches LLaMA 2 (70B) capabilities on standard LM benchmarks but uses only half of its training resources.
[...]

Grok-0, the prototype LLM mentioned in the provided context, is stated to have been trained with 33 billion parameters.

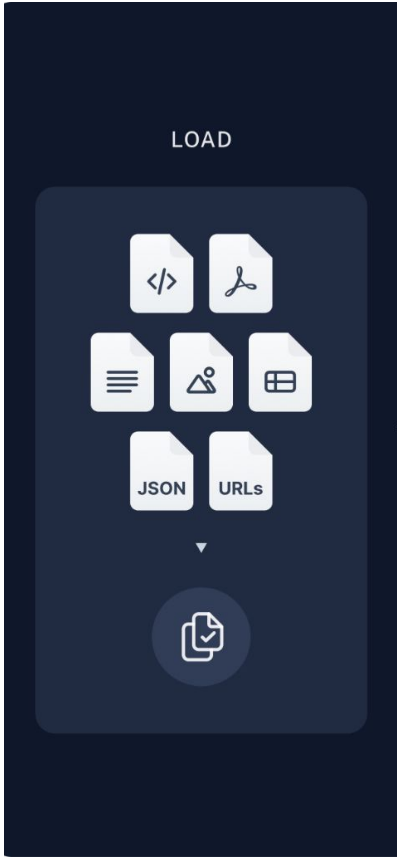
How to implement RAG?

Part 1
Storing the Data


LOAD
SPLIT
EMBED
STORE
JSON
URLs
[0.3, 0.4, 0.1, 1.8, 1.1...]
[0.7, 1.4, 2.1, 4.8, 4.1...]
[1.2, 0.3, 1.2, 4.1, 1.8...]
[0.3, 0.4, 0.1, 1.8, 1.1...]
[0.7, 1.4, 2.1, 4.8, 4.1...]
[1.2, 0.3, 1.2, 4.1, 1.8...]

Load

- Loading the data it will use
- This data serves from various sources:
  - **Media File**
  - **JSON Files**
  - **URLs**



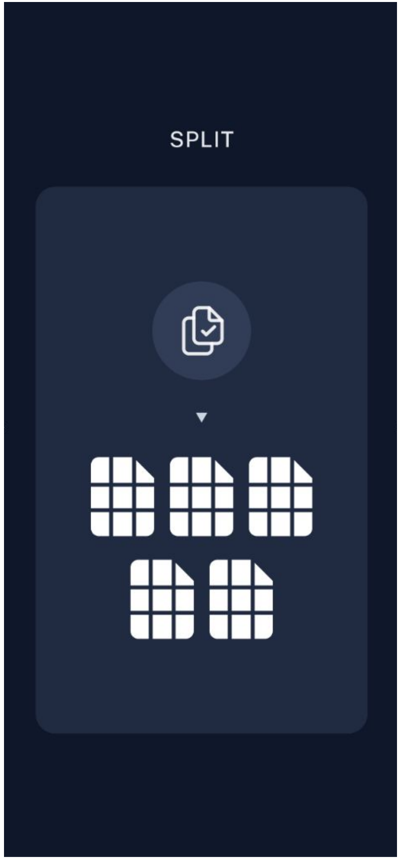
Split / Chunking

- Why not without chunking?

  Inefficient Processing

- Then why Chunking?

  Chunks for efficient processing



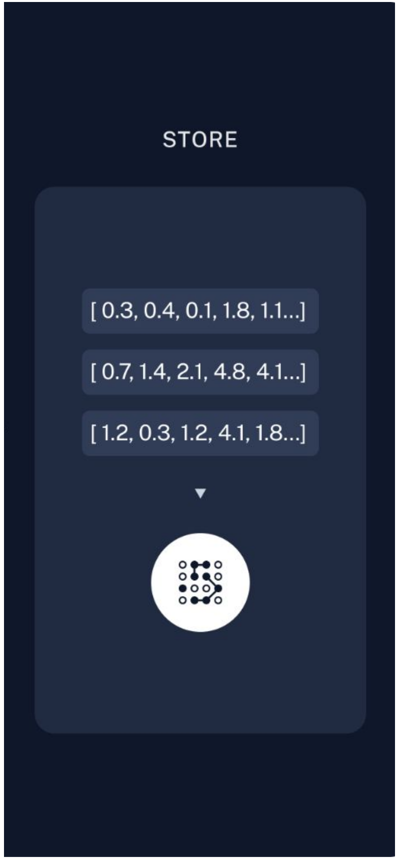
Embedding

- What embedding captures?
- What will it do?
- Common embedding
  - **TF-IDF**
  - **word2vec**

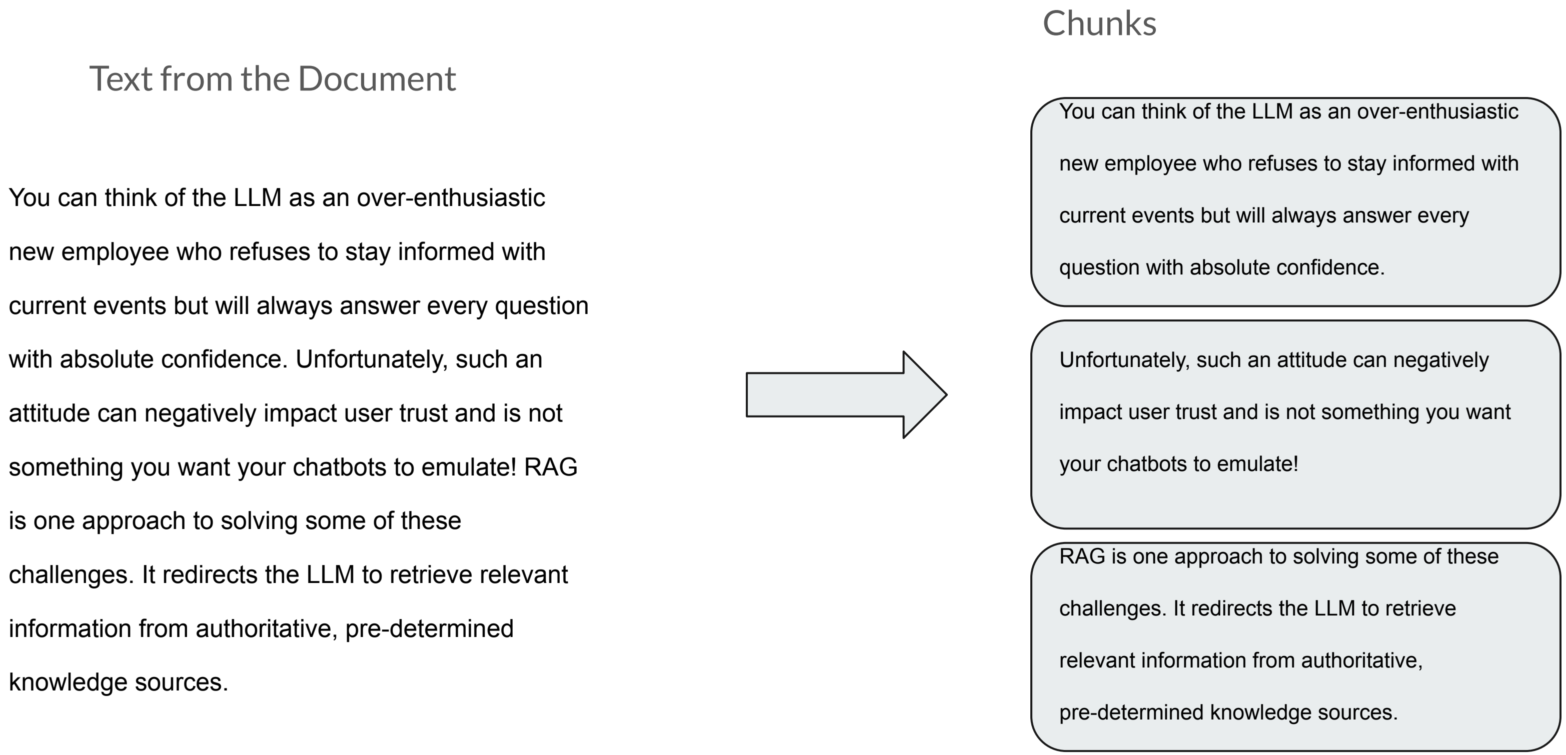
EMBED
[0.3, 0.4, 0.1, 1.8, 1.1...]
[0.7, 1.4, 2.1, 4.8, 4.1...]
[1.2, 0.3, 1.2, 4.1, 1.8...]

Storing

- After generating the embeddings?
- Vector stores are specialized databases designed
- Benefits
    - **Fast Retrieval**
    - **Scalability**
    - **Search Optimization**



Simple Visual Implementation



Text from the Document

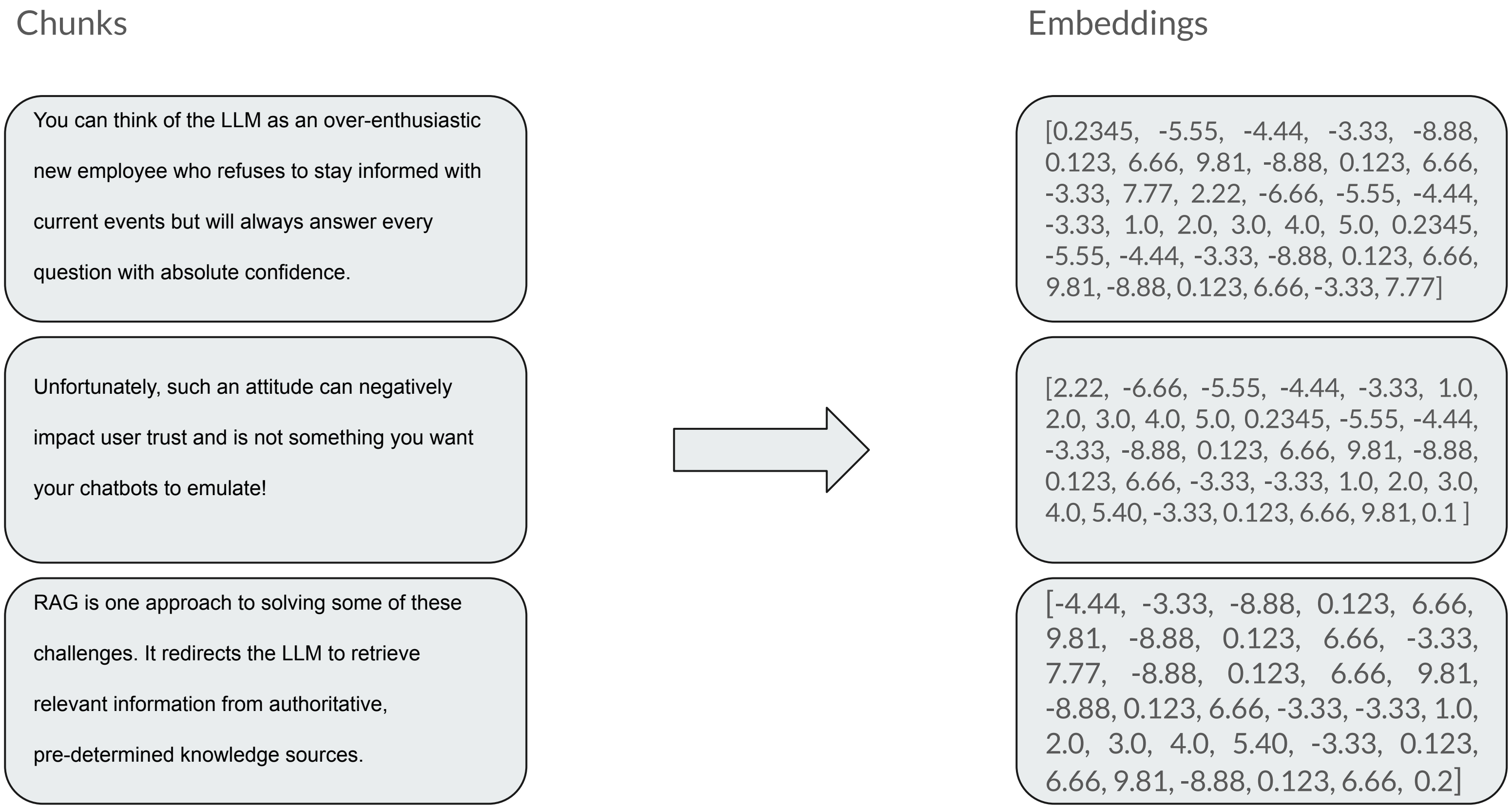
You can think of the LLM as an over-enthusiastic new employee who refuses to stay informed with current events but will always answer every question with absolute confidence. Unfortunately, such an attitude can negatively impact user trust and is not something you want your chatbots to emulate! RAG is one approach to solving some of these challenges. It redirects the LLM to retrieve relevant information from authoritative, pre-determined knowledge sources.

Chunks

You can think of the LLM as an over-enthusiastic new employee who refuses to stay informed with current events but will always answer every question with absolute confidence.

Unfortunately, such an attitude can negatively impact user trust and is not something you want your chatbots to emulate!

RAG is one approach to solving some of these challenges. It redirects the LLM to retrieve relevant information from authoritative, pre-determined knowledge sources.



Chunks

You can think of the LLM as an over-enthusiastic new employee who refuses to stay informed with current events but will always answer every question with absolute confidence.

Unfortunately, such an attitude can negatively impact user trust and is not something you want your chatbots to emulate!

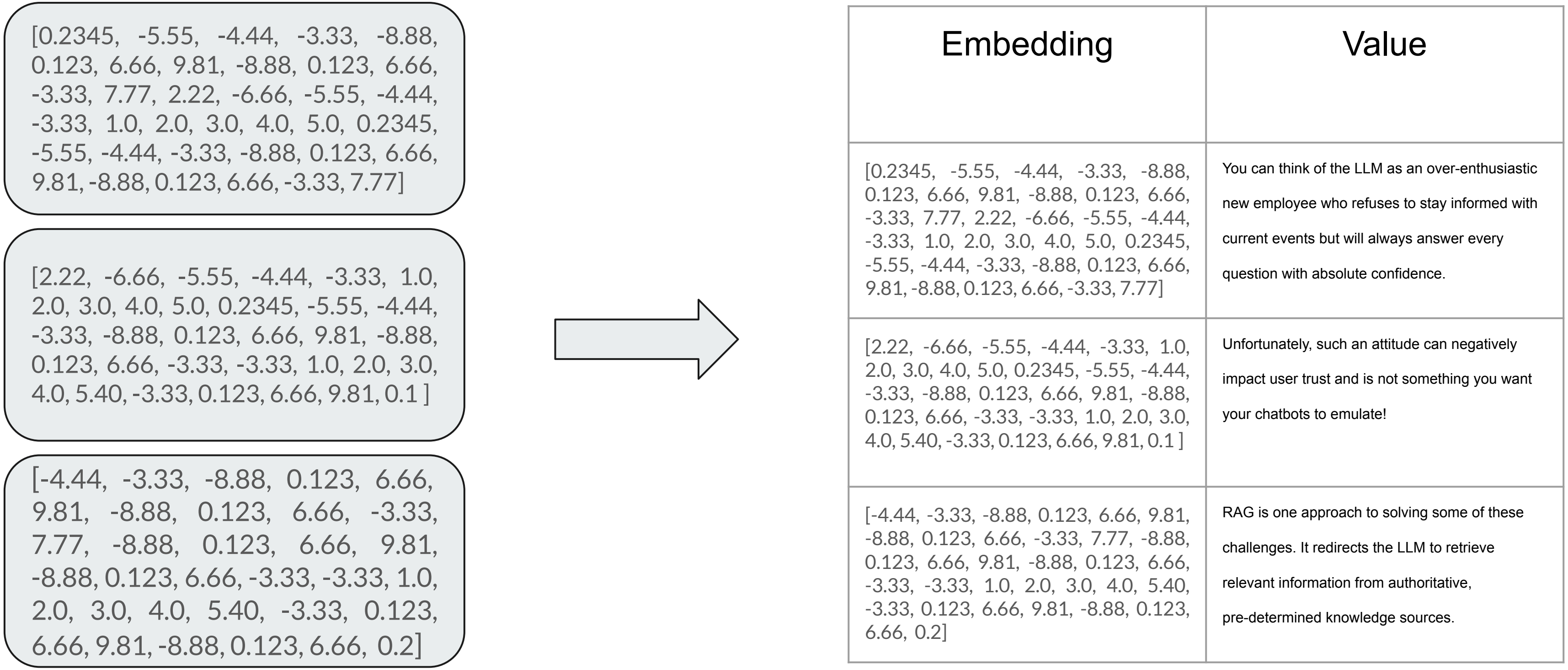
RAG is one approach to solving some of these challenges. It redirects the LLM to retrieve relevant information from authoritative, pre-determined knowledge sources.

Embeddings

[0.2345, -5.55, -4.44, -3.33, -8.88, 0.123, 6.66, 9.81, -8.88, 0.123, 6.66, -3.33, 7.77, 2.22, -6.66, -5.55, -4.44, -3.33, 1.0, 2.0, 3.0, 4.0, 5.0, 0.2345, -5.55, -4.44, -3.33, -8.88, 0.123, 6.66, 9.81, -8.88, 0.123, 6.66, -3.33, 7.77]

[2.22, -6.66, -5.55, -4.44, -3.33, 1.0, 2.0, 3.0, 4.0, 5.0, 0.2345, -5.55, -4.44, -3.33, -8.88, 0.123, 6.66, 9.81, -8.88, 0.123, 6.66, -3.33, -3.33, 1.0, 2.0, 3.0, 4.0, 5.40, -3.33, 0.123, 6.66, 9.81, 0.1]

[-4.44, -3.33, -8.88, 0.123, 6.66, 9.81, -8.88, 0.123, 6.66, -3.33, 7.77, -8.88, 0.123, 6.66, 9.81, -8.88, 0.123, 6.66, -3.33, -3.33, 1.0, 2.0, 3.0, 4.0, 5.40, -3.33, 0.123, 6.66, 9.81, -8.88, 0.123, 6.66, 0.2]



Embeddings

[0.2345, -5.55, -4.44, -3.33, -8.88, 0.123, 6.66, 9.81, -8.88, 0.123, 6.66, -3.33, 7.77, 2.22, -6.66, -5.55, -4.44, -3.33, 1.0, 2.0, 3.0, 4.0, 5.0, 0.2345, -5.55, -4.44, -3.33, -8.88, 0.123, 6.66, 9.81, -8.88, 0.123, 6.66, -3.33, 7.77]

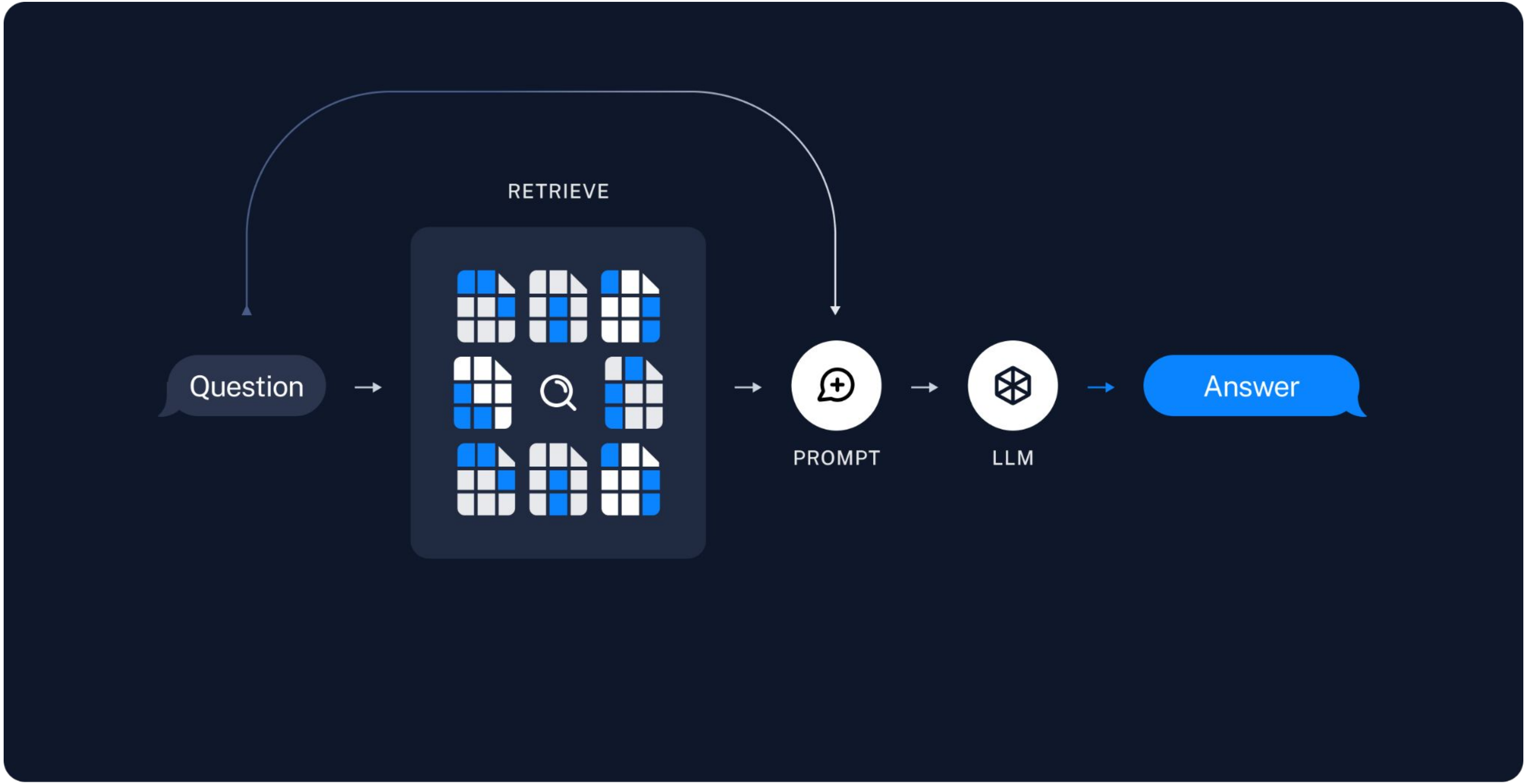
[2.22, -6.66, -5.55, -4.44, -3.33, 1.0, 2.0, 3.0, 4.0, 5.0, 0.2345, -5.55, -4.44, -3.33, -8.88, 0.123, 6.66, 9.81, -8.88, 0.123, 6.66, -3.33, -3.33, 1.0, 2.0, 3.0, 4.0, 5.40, -3.33, 0.123, 6.66, 9.81, 0.1]

[-4.44, -3.33, -8.88, 0.123, 6.66, 9.81, -8.88, 0.123, 6.66, -3.33, 7.77, -8.88, 0.123, 6.66, 9.81, -8.88, 0.123, 6.66, -3.33, -3.33, 1.0, 2.0, 3.0, 4.0, 5.40, -3.33, 0.123, 6.66, 9.81, -8.88, 0.123, 6.66, 0.2]

Vector DataBase

Embedding	Value
[0.2345, -5.55, -4.44, -3.33, -8.88, 0.123, 6.66, 9.81, -8.88, 0.123, 6.66, -3.33, 7.77, 2.22, -6.66, -5.55, -4.44, -3.33, 1.0, 2.0, 3.0, 4.0, 5.0, 0.2345, -5.55, -4.44, -3.33, -8.88, 0.123, 6.66, 9.81, -8.88, 0.123, 6.66, -3.33, 7.77]	You can think of the LLM as an over-enthusiastic new employee who refuses to stay informed with current events but will always answer every question with absolute confidence.
[2.22, -6.66, -5.55, -4.44, -3.33, 1.0, 2.0, 3.0, 4.0, 5.0, 0.2345, -5.55, -4.44, -3.33, -8.88, 0.123, 6.66, 9.81, -8.88, 0.123, 6.66, -3.33, -3.33, 1.0, 2.0, 3.0, 4.0, 5.40, -3.33, 0.123, 6.66, 9.81, 0.1]	Unfortunately, such an attitude can negatively impact user trust and is not something you want your chatbots to emulate!
[-4.44, -3.33, -8.88, 0.123, 6.66, 9.81, -8.88, 0.123, 6.66, -3.33, 7.77, -8.88, 0.123, 6.66, 9.81, -8.88, 0.123, 6.66, -3.33, -3.33, 1.0, 2.0, 3.0, 4.0, 5.40, -3.33, 0.123, 6.66, 9.81, -8.88, 0.123, 6.66, 0.2]	RAG is one approach to solving some of these challenges. It redirects the LLM to retrieve relevant information from authoritative, pre-determined knowledge sources.

Part 2
Retrieving the Data


RETRIEVE
Question
PROMPT
LLM
Answer

Question

What can I think of the LLM as an over-enthusiastic?

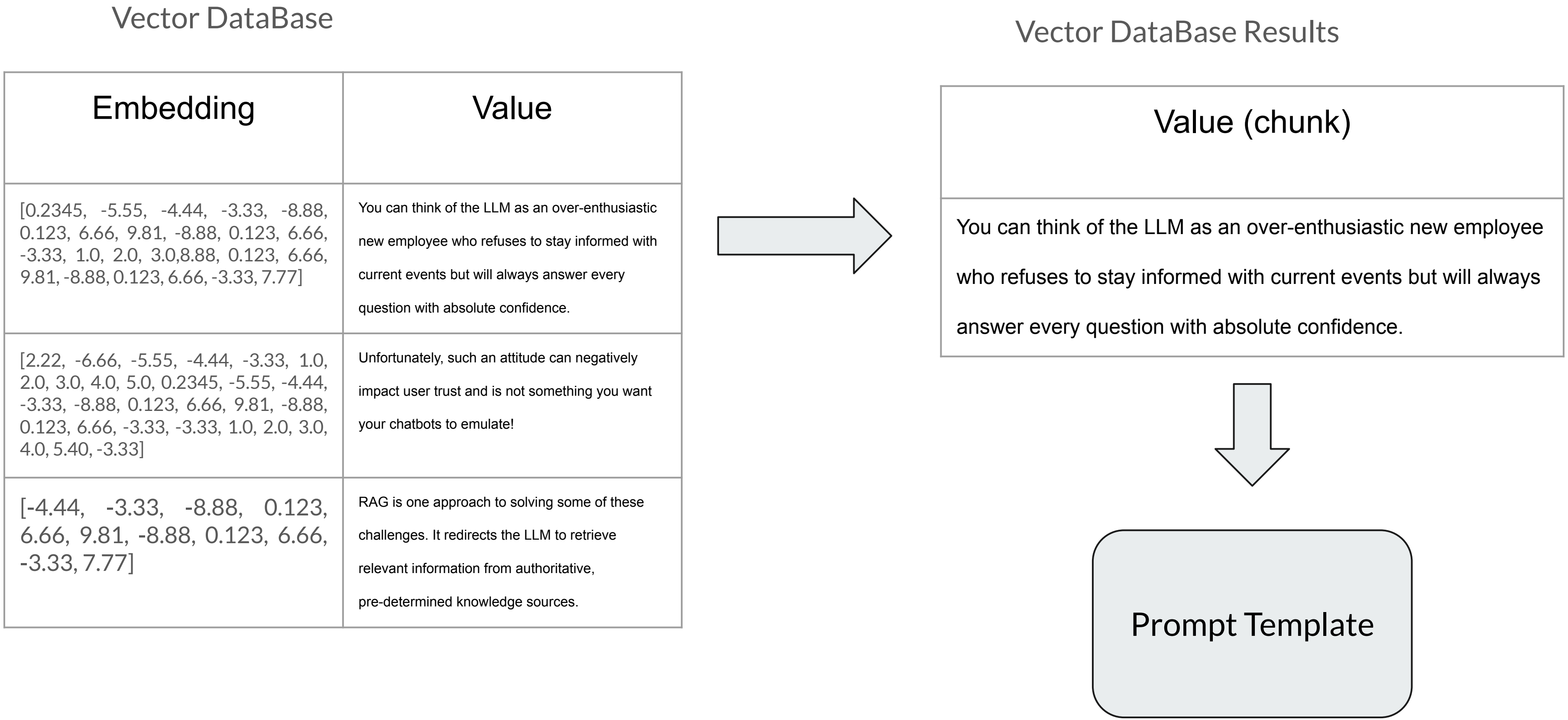
Question Embeddings

[0.2345, -5.55, -4.44, -3.33, -8.88, 0.123, 6.66]

Search

Vector DataBase

Embedding	Value
[0.2345, -5.55, -4.44, -3.33, -8.88, 0.123, 6.66, 9.81, -8.88, 0.123, 6.66, -3.33, 1.0, 2.0, 3.0,8.88, 0.123, 6.66, 9.81, -8.88, 0.123, 6.66, -3.33, 7.77]	You can think of the LLM as an over-enthusiastic new employee who refuses to stay informed with current events but will always answer every question with absolute confidence.
[2.22, -6.66, -5.55, -4.44, -3.33, 1.0, 2.0, 3.0, 4.0, 5.0, 0.2345, -5.55, -4.44, -3.33, -8.88, 0.123, 6.66, 9.81, -8.88, 0.123, 6.66, -3.33, -3.33, 1.0, 2.0, 3.0, 4.0, 5.40, -3.33]	Unfortunately, such an attitude can negatively impact user trust and is not something you want your chatbots to emulate!
[-4.44, -3.33, -8.88, 0.123, 6.66, 9.81, -8.88, 0.123, 6.66, -3.33, 7.77]	RAG is one approach to solving some of these challenges. It redirects the LLM to retrieve relevant information from authoritative, pre-determined knowledge sources.



Vector DataBase

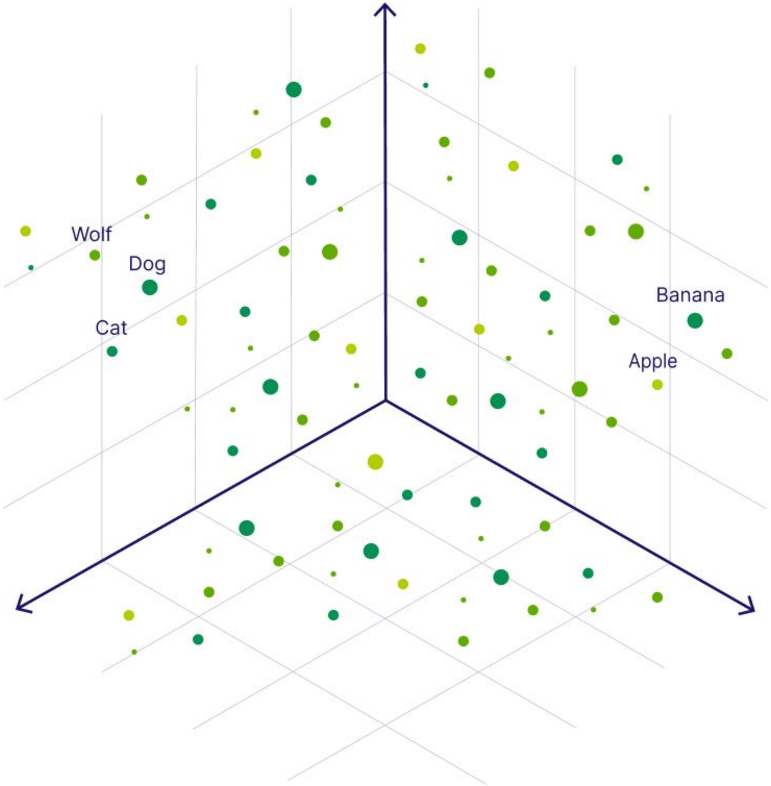
Embedding	Value
[0.2345, -5.55, -4.44, -3.33, -8.88, 0.123, 6.66, 9.81, -8.88, 0.123, 6.66, -3.33, 1.0, 2.0, 3.0,8.88, 0.123, 6.66, 9.81, -8.88, 0.123, 6.66, -3.33, 7.77]	You can think of the LLM as an over-enthusiastic new employee who refuses to stay informed with current events but will always answer every question with absolute confidence.
[2.22, -6.66, -5.55, -4.44, -3.33, 1.0, 2.0, 3.0, 4.0, 5.0, 0.2345, -5.55, -4.44, -3.33, -8.88, 0.123, 6.66, 9.81, -8.88, 0.123, 6.66, -3.33, -3.33, 1.0, 2.0, 3.0, 4.0, 5.40, -3.33]	Unfortunately, such an attitude can negatively impact user trust and is not something you want your chatbots to emulate!
[-4.44, -3.33, -8.88, 0.123, 6.66, 9.81, -8.88, 0.123, 6.66, -3.33, 7.77]	RAG is one approach to solving some of these challenges. It redirects the LLM to retrieve relevant information from authoritative, pre-determined knowledge sources.

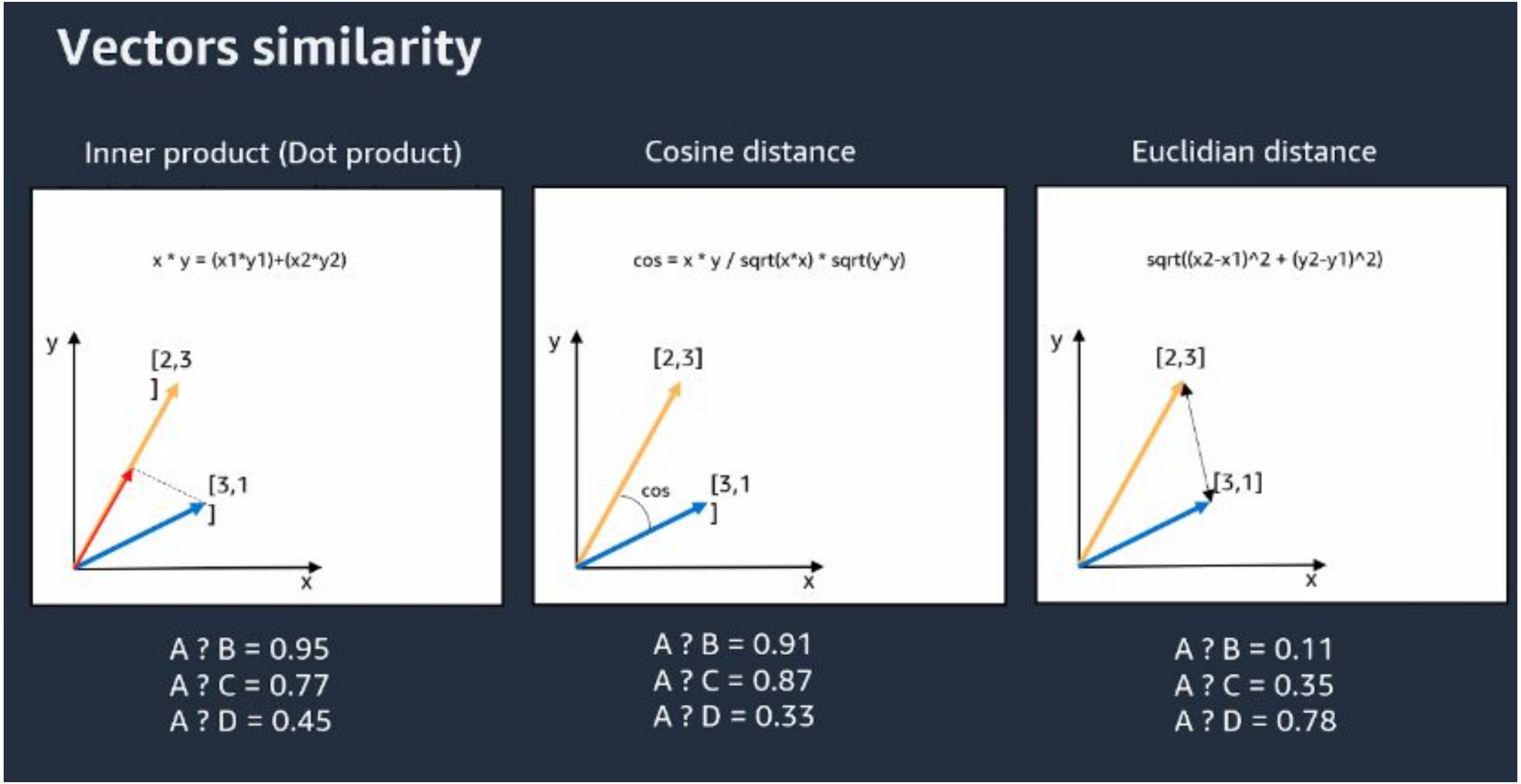
Vector DataBase Results

Value (chunk)
You can think of the LLM as an over-enthusiastic new employee who refuses to stay informed with current events but will always answer every question with absolute confidence.

Prompt Template

Vector DataBase





Vectors similarity

Inner product (Dot product)

x * y = (x1*y1)+(x2*y2)

A ? B = 0.95
A ? C = 0.77
A ? D = 0.45

Cosine distance

cos = x * y / sqrt(x*x) * sqrt(y*y)

A ? B = 0.91
A ? C = 0.87
A ? D = 0.33

Euclidian distance

sqrt((x2-x1)^2 + (y2-y1)^2)

A ? B = 0.11
A ? C = 0.35
A ? D = 0.78

Agenda

~~Why~~

~~What~~

When

Why is Advance RAG

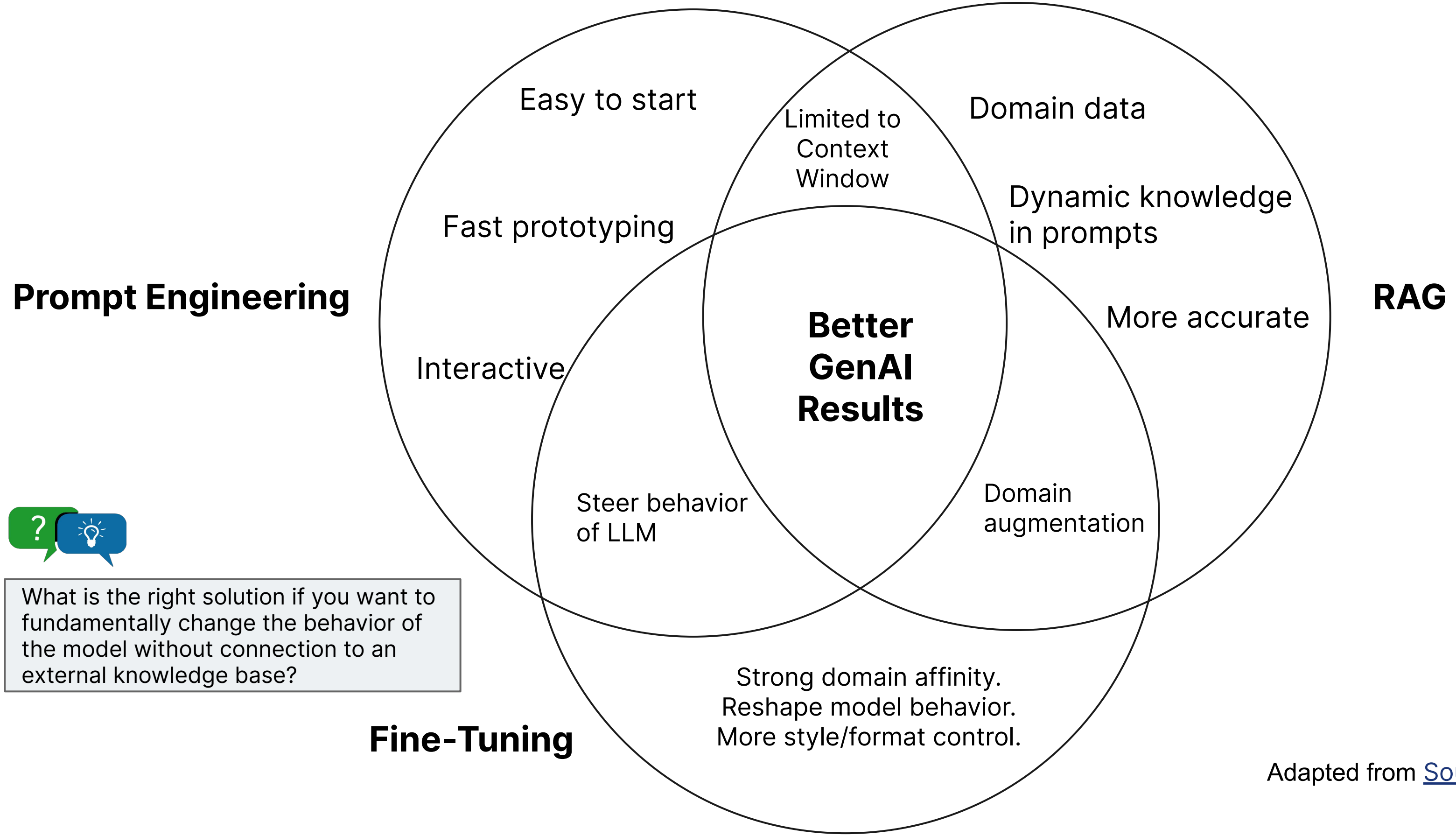
What is Advance RAG?

Evaluation

Prompting

HandsOn

When to Use RAG ?



Prompt Engineering

Easy to start

Fast prototyping

Interactive

Limited to Context Window

RAG

Domain data

Dynamic knowledge in prompts

More accurate

Better GenAI Results

Steer behavior of LLM

Domain augmentation

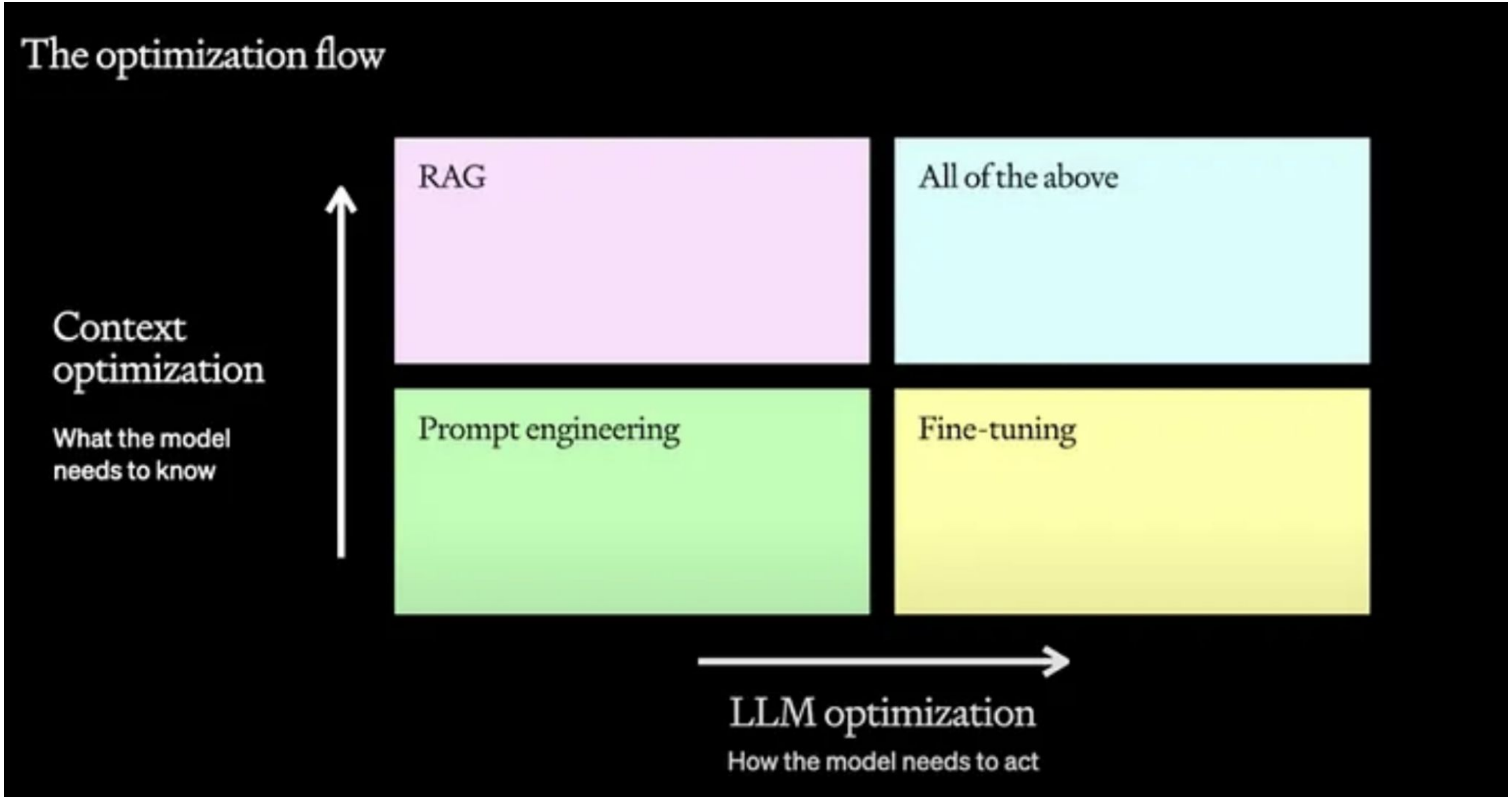
Fine-Tuning

Strong domain affinity.
Reshape model behavior.
More style/format control.

What is the right solution if you want to fundamentally change the behavior of the model without connection to an external knowledge base?

Adapted from Source

When to Use RAG ?



The optimization flow

RAG

All of the above

Prompt engineering

Fine-tuning

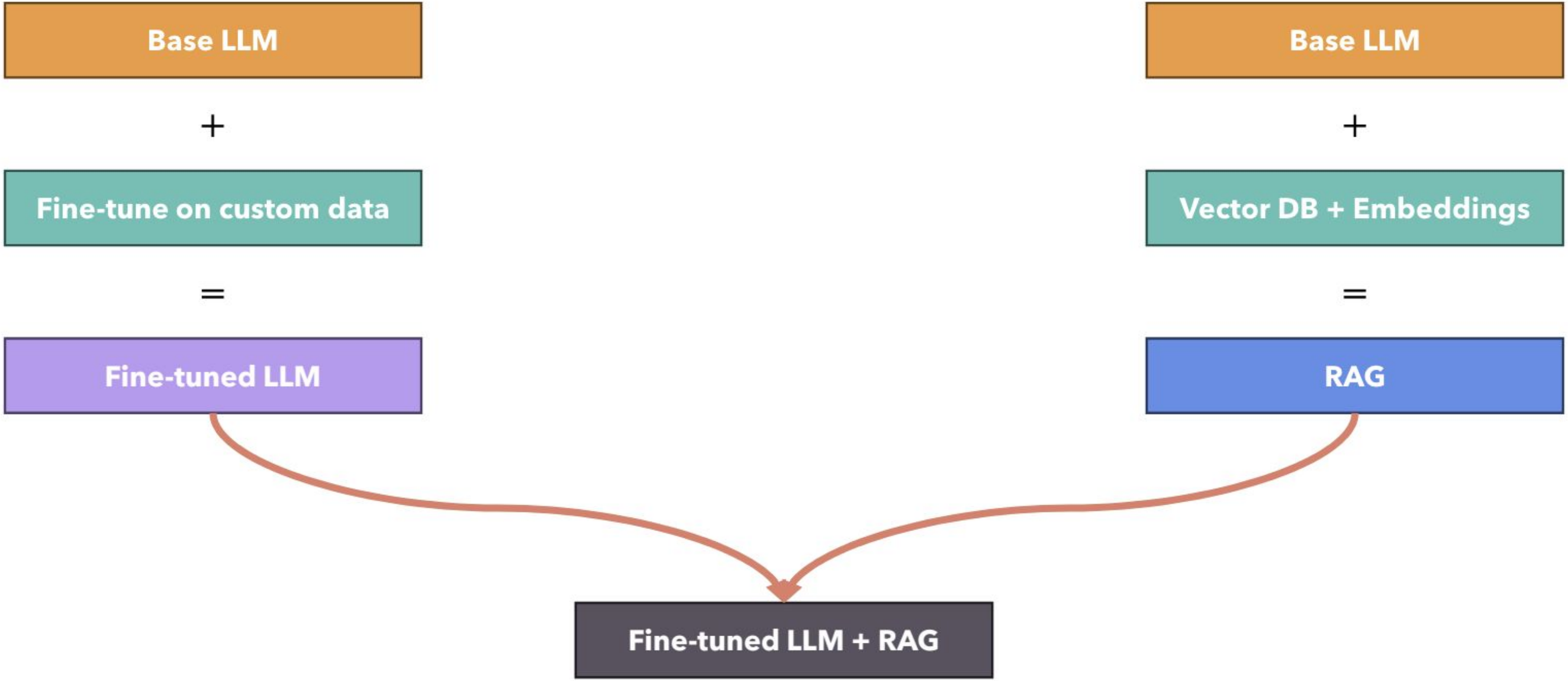
Context optimization

What the model needs to know

LLM optimization

How the model needs to act

Fine-Tuning VS RAG



Base LLM

+

Fine-tune on custom data

=

Fine-tuned LLM

Base LLM

+

Vector DB + Embeddings

=

RAG

Fine-tuned LLM + RAG

Agenda

~~Why~~

~~What~~

~~When~~

Why is Advance RAG

What is Advance RAG?

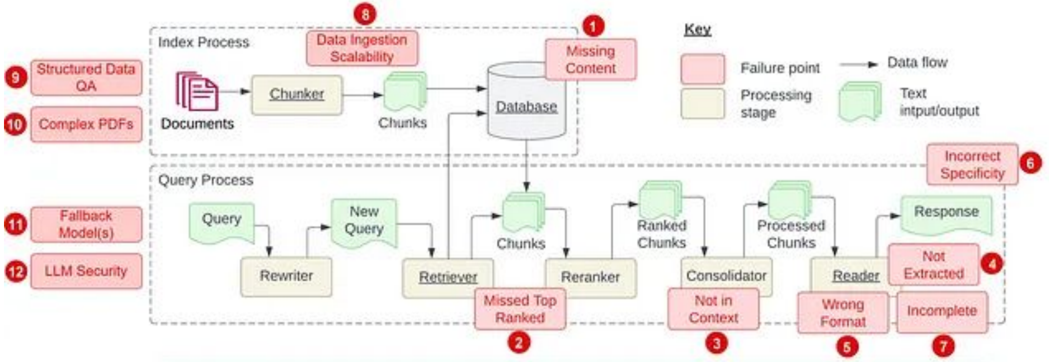
Evaluation

Prompting

HandsOn

Advance RAG

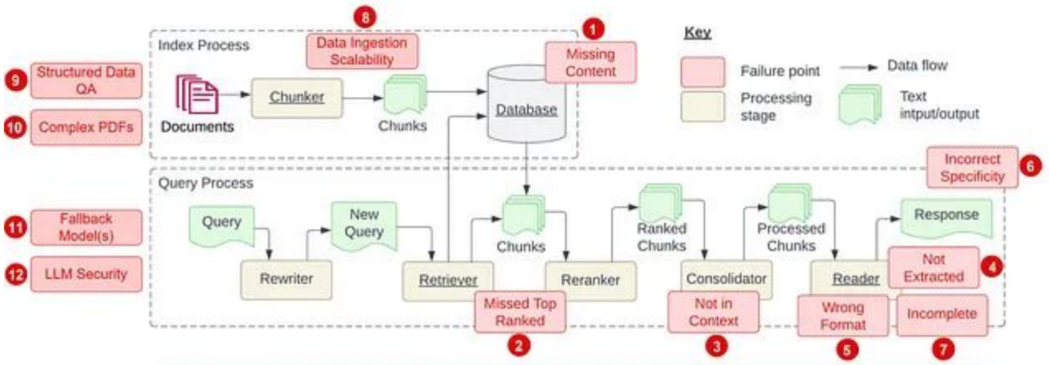
12 failure points



Response Quality Related

1. Context Missing in the Knowledge Base
2. Context Missing in the Initial Retrieval Pass
3. Context Missing After Reranking
4. Context Not Extracted
5. Output is in Wrong Format
6. Output has Incorrect Level of Specificity
7. Output is Incomplete

12 failure points



Scalability

8. Can't Scale to Larger Data Volumes

11. Rate-Limit Errors

Security

12. LLM Security

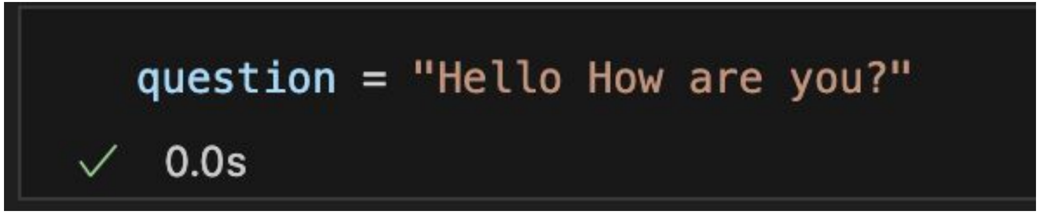
Use Case Specific

9. Ability to QA Tabular Data

10. Ability to Parse PDFs

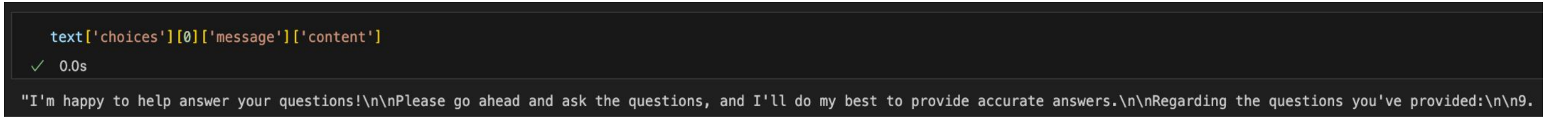
12 failure points - Solution

- Clean your data & Better Prompting
- Hyperparameter tuning & Reranking
- Tweak retrieval strategies & Embedded table retrieval
- Clean your data, prompt compression, & LongContextReorder
- Better prompting, output parsing, pydantic programs & OpenAI JSON mode
- Advanced retrieval strategies
- Query transformations
- Parallelizing ingestion pipeline
- Chain-of-table pack & Mix-self-consistency pack
- Embedding table retrieval
- Neutrino router & OpenRouter
- NeMo Guardrails & Llama Guardrails



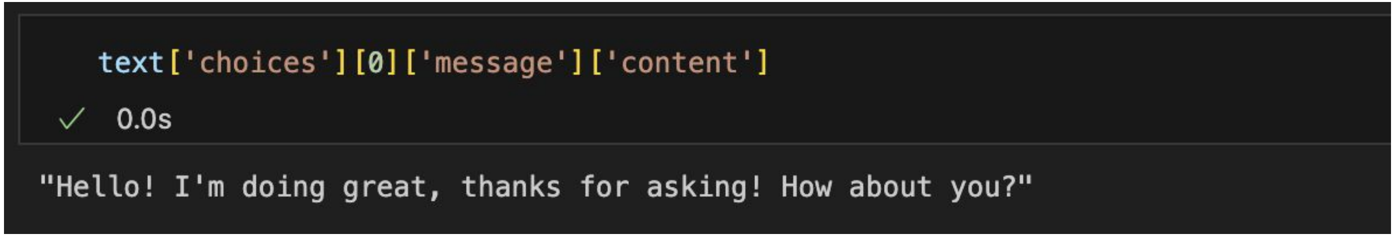
```
question = "Hello How are you?"
✓ 0.0s
```

Actual Answer?



```
text['choices'][0]['message']['content']
✓ 0.0s
"I'm happy to help answer your questions!\n\nPlease go ahead and ask the questions, and I'll do my best to provide accurate answers.\n\nRegarding the questions you've provided:\n\n9.
```

Expected Answer?



```
text['choices'][0]['message']['content']
✓ 0.0s
"Hello! I'm doing great, thanks for asking! How about you?"
```

Why?

Agenda

~~Why~~

~~What~~

~~When~~

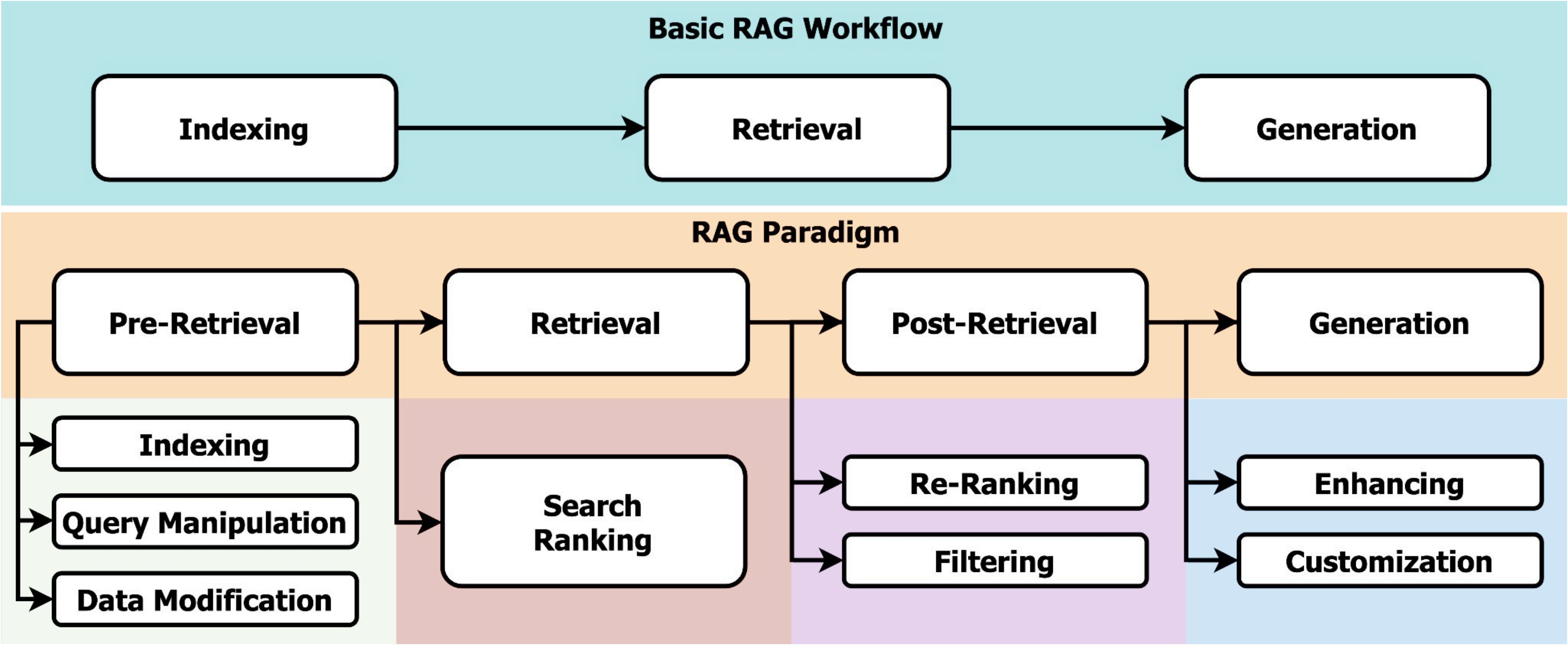
~~Why is Advance RAG~~

What is Advance RAG?

Evaluation

Prompting

HandsOn



Basic RAG Workflow

Indexing

Retrieval

Generation

RAG Paradigm

Pre-Retrieval

Retrieval

Post-Retrieval

Generation

Indexing

Query Manipulation

Data Modification

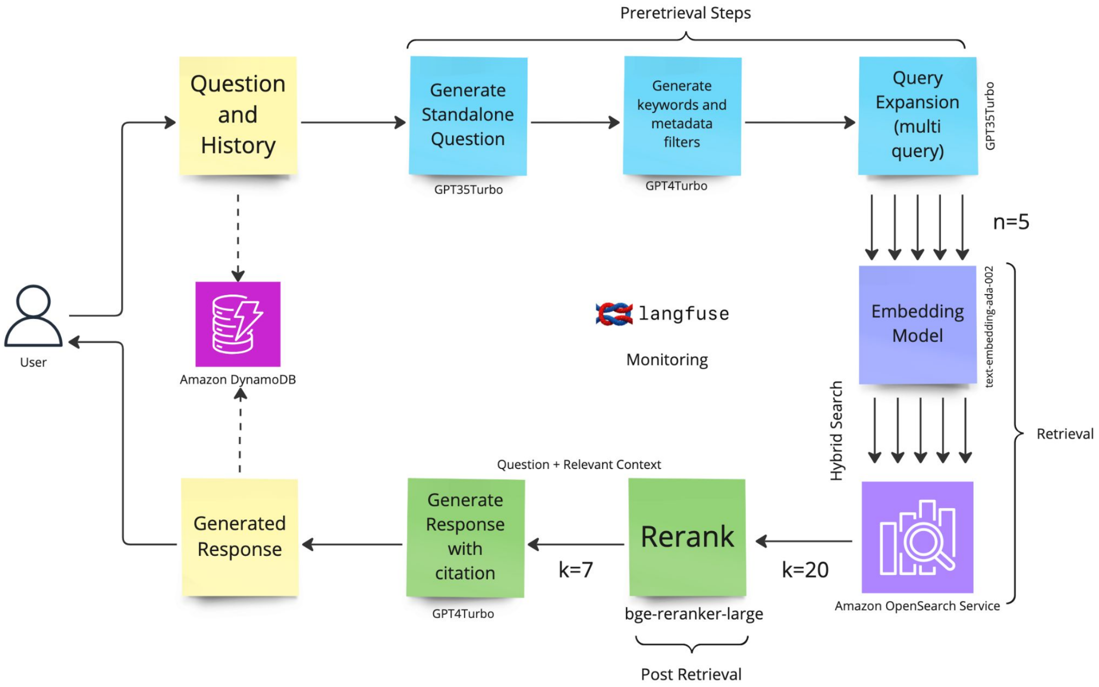
Search Ranking

Re-Ranking

Filtering

Enhancing

Customization



Preretrieval Steps

User

Question and History

Generate Standalone Question

GPT35Turbo

Generate keywords and metadata filters

GPT4Turbo

Query Expansion (multi query)

GPT35Turbo

n=5

Embedding Model

text-embedding-ada-002

Hybrid Search

Retrieval

Amazon OpenSearch Service

k=20

Rerank

bge-reranker-large

Post Retrieval

k=7

Question + Relevant Context

Generate Response with citation

GPT4Turbo

Generated Response

Amazon DynamoDB

langfuse

Monitoring

Azure OpenAI

langfuse

LangChain

DynamoDB

React JS

aws

OpenSearch

AWS Lambda

GitHub Actions

Agenda

~~Why~~

~~What~~

~~When~~

~~Why is Advance RAG~~

~~What is Advance RAG?~~

Evaluation

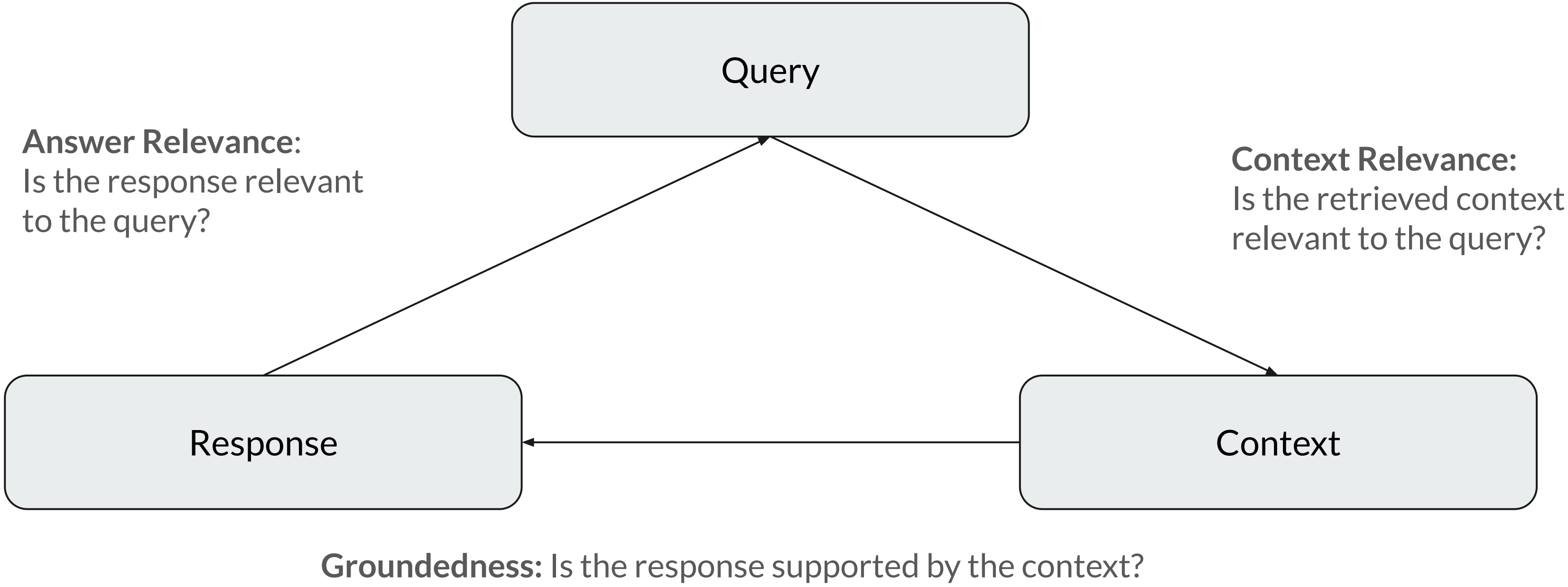
Prompting

HandsOn

Evaluation Demo

Evaluation

The RAG Triad

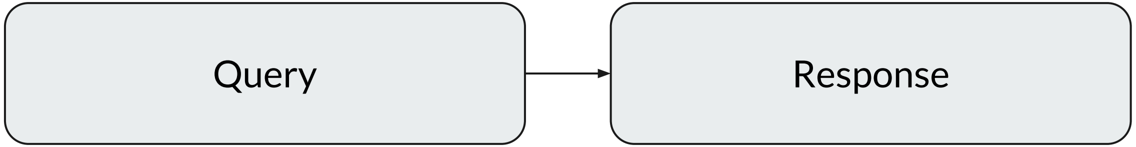


Answer Relevance:
Is the response relevant to the query?

Context Relevance:
Is the retrieved context relevant to the query?

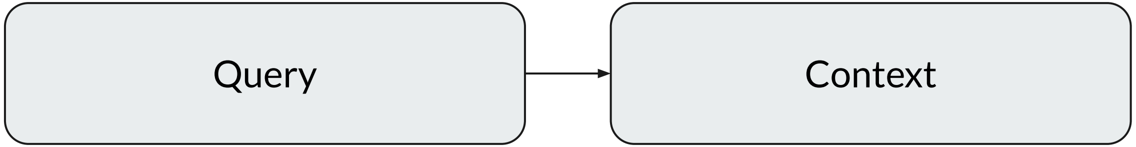
Groundedness: Is the response supported by the context?

Answer Relevance



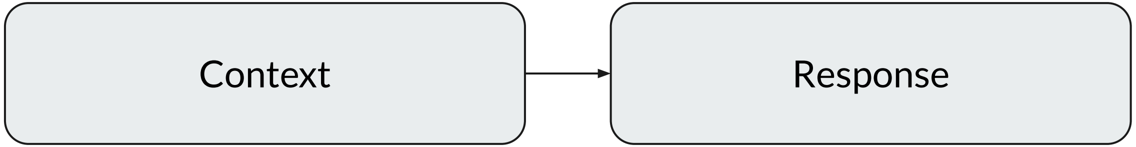
Is the final response useful?

Context Relevance



How good is the retrieval?

Groundedness



How good is the response based on context?

Agenda

~~Why~~

~~What~~

~~When~~

~~Why is Advance RAG~~

~~What is Advance RAG?~~

~~Evaluation~~

Prompting

HandsOn

Prompting

RRR Prompting

Role: You are a Q&A Chatbot interacting with real-world customers to their questions.

Rule: Only answer based on the content provided. Do not provide any information that is not in the content.

Response: Answer in a interactive way

Agenda

~~Why~~

~~What~~

~~When~~

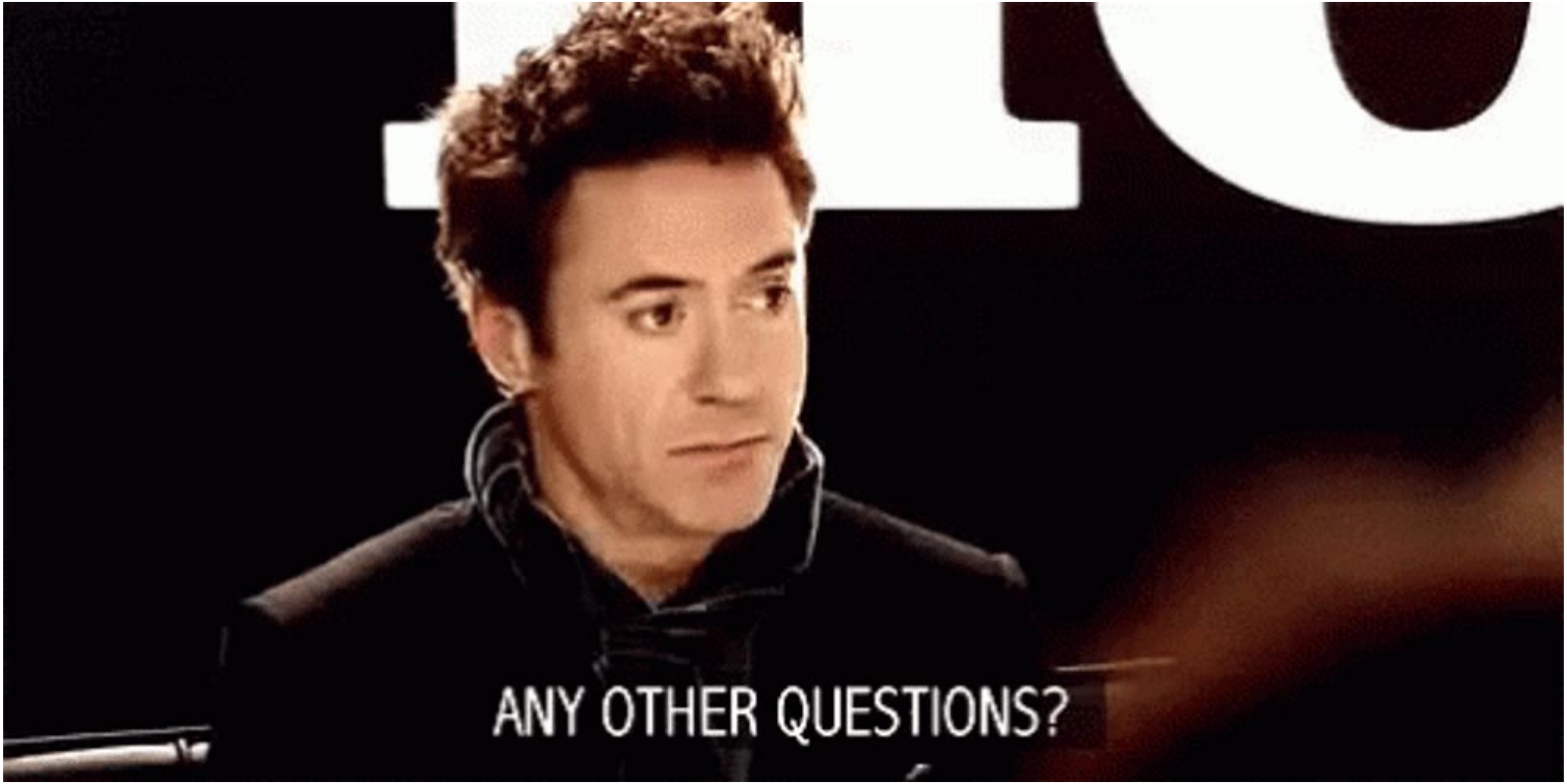
~~Why is Advance RAG~~

~~What is Advance RAG?~~

~~Evaluation~~

~~Prompting~~

HandsOn


ANY OTHER QUESTIONS?

Hands ON

Prerequisites

- Jupyter Notebook
- Python 3.10
- Ollama

Agenda

~~Why~~

~~What~~

~~When~~

~~Why is Advance RAG~~

~~What is Advance RAG?~~

~~Evaluation~~

~~Prompting~~

~~HandsOn~~

Thank you!