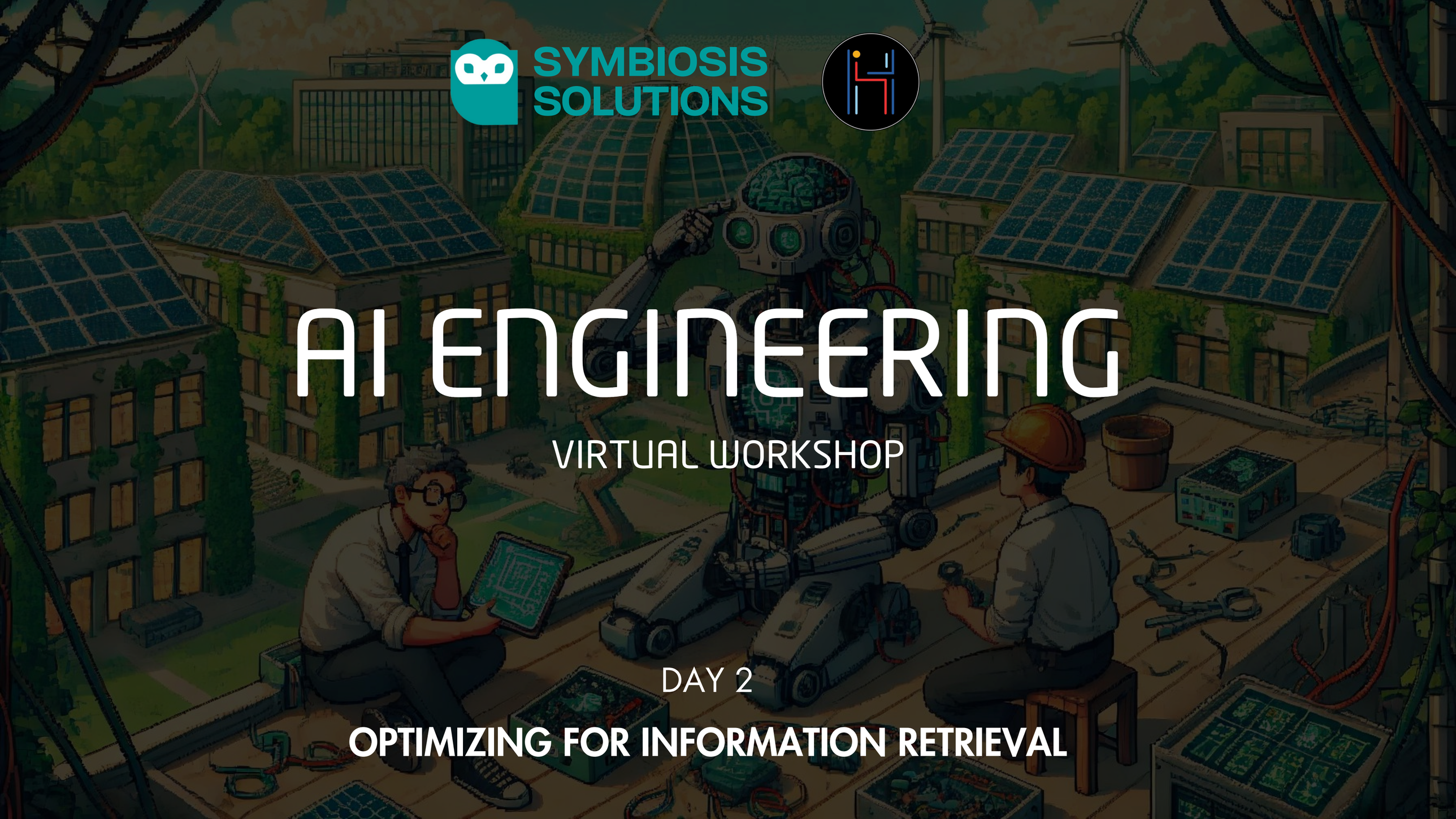


AI ENGINEERING

VIRTUAL WORKSHOP

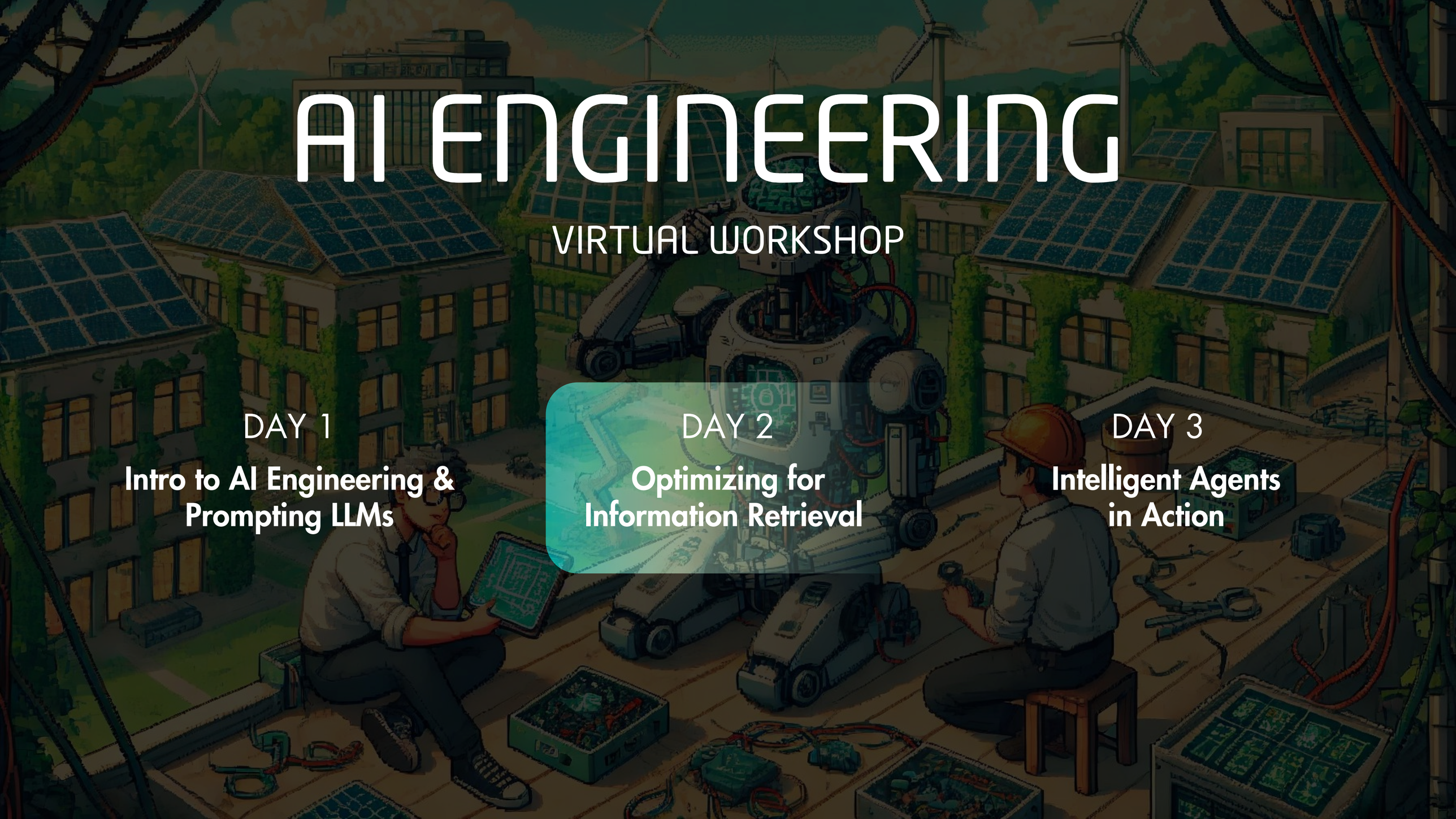
DAY 2

OPTIMIZING FOR INFORMATION RETRIEVAL

FLASH RECAP

menti.com

6566 9721

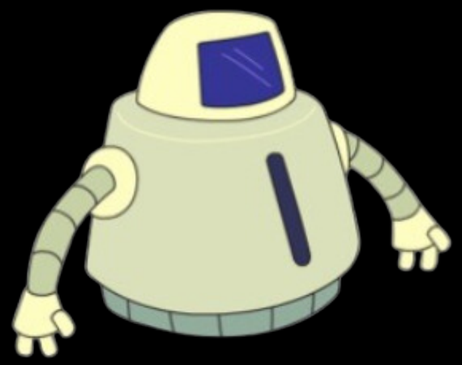

AI ENGINEERING
VIRTUAL WORKSHOP
DAY 1
Intro to AI Engineering & Prompting LLMs
DAY 2
Optimizing for Information Retrieval
DAY 3
Intelligent Agents in Action

AGENDAS

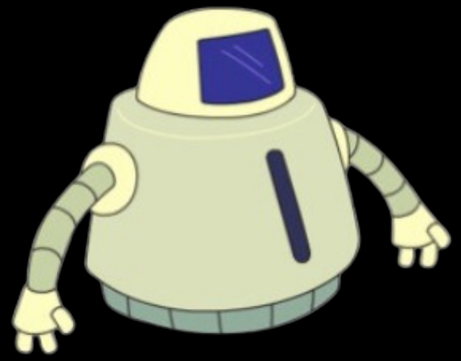
OPTIMIZING FOR INFORMATION RETRIEVAL

1. Why IR Matters
2. Evaluating RAG
3. RAG Techniques

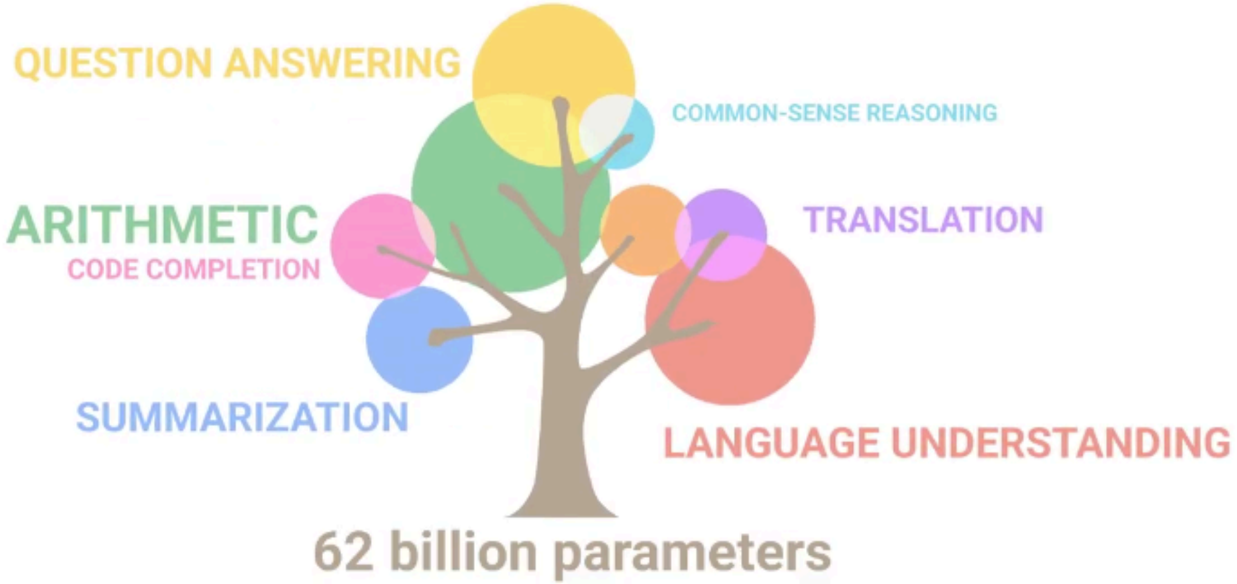
EUREKA!



EUREKA!



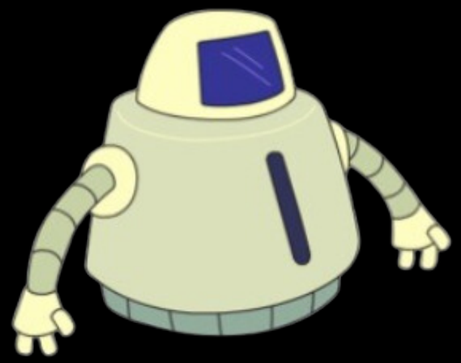
LLMs

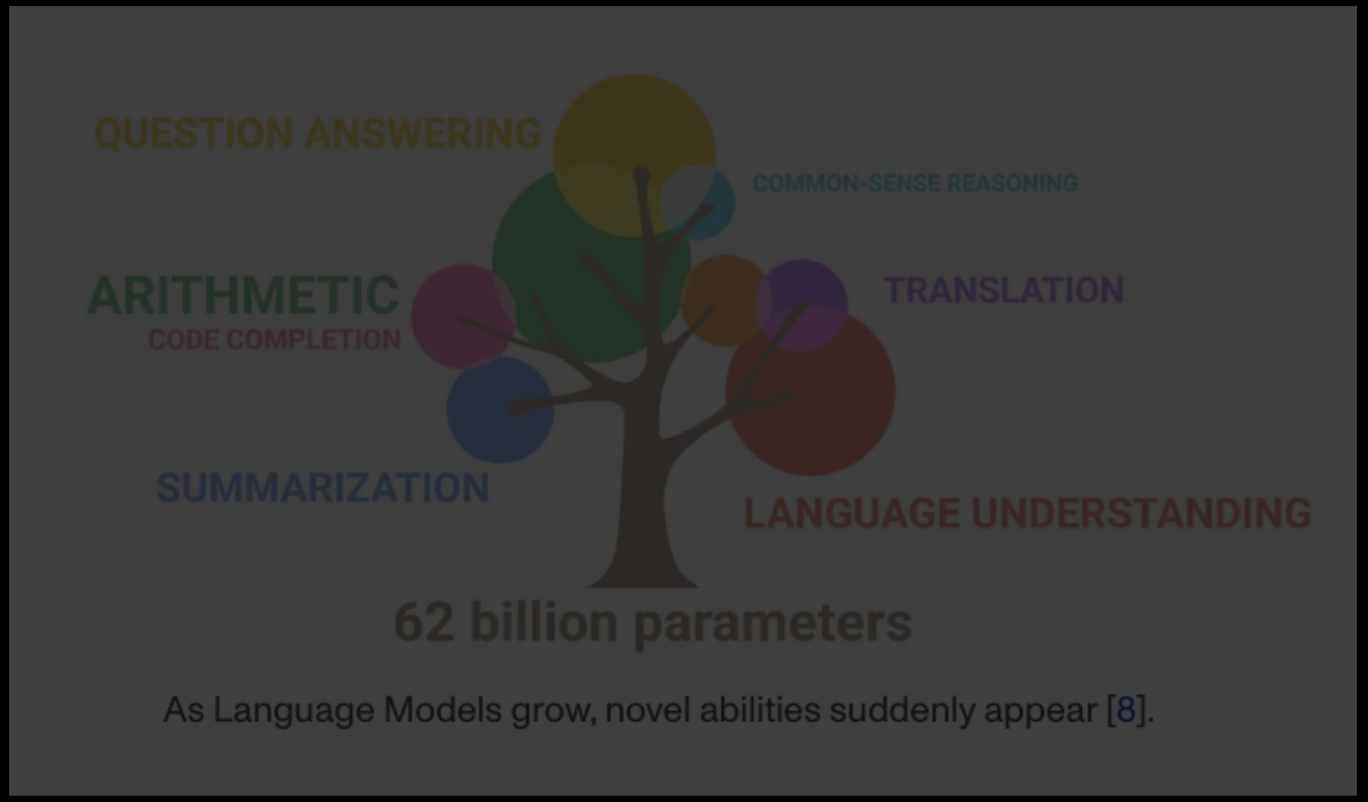


As Language Models grow, novel abilities suddenly appear [8].

BUT...

What are it's LIMITATIONS?

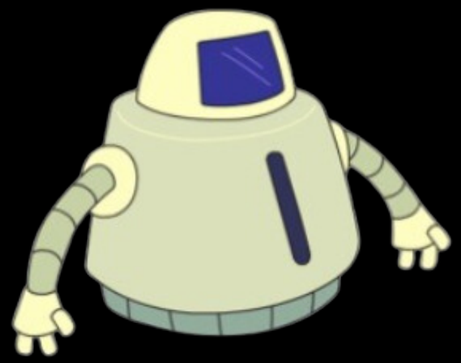




As Language Models grow, novel abilities suddenly appear [8].

BUT...

What are it's LIMITATIONS?

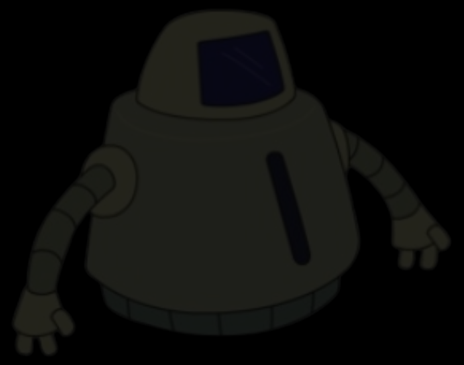


Key Limitations of LLMs

1. Hallucinations and Misinformation
2. Contextual Understanding
3. Bias and Ethical Concerns

As Language Models grow, novel abilities suddenly appear [8].

BUT... What are it's LIMITATIONS?



Key Limitations of LLMs

1. Hallucinations and Misinformation
2. Contextual Understanding
3. Bias and Ethical Concerns

SOLUTION???

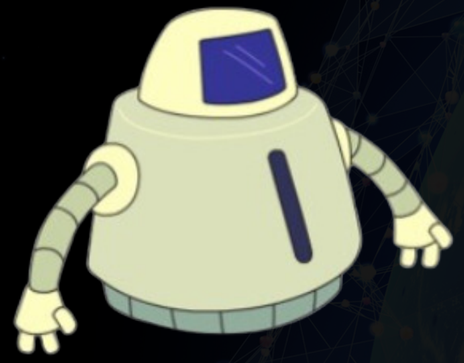
INFORMATION RETRIEVAL

INFORMATION RETRIEVAL

refers to
the techniques and processes
used to efficiently
access and retrieve
relevant information from
large datasets.

INFORMATION RETRIEVAL

refers to
the **techniques and processes**
used to **efficiently**
access and retrieve
relevant information from
large datasets.

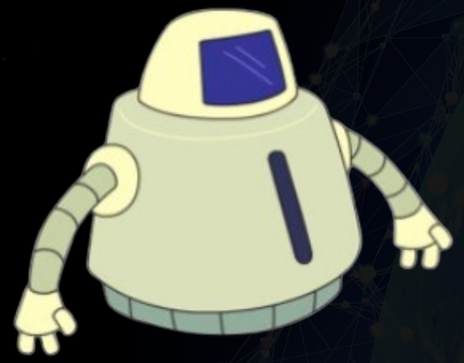


Benefits?

INFORMATION RETRIEVAL

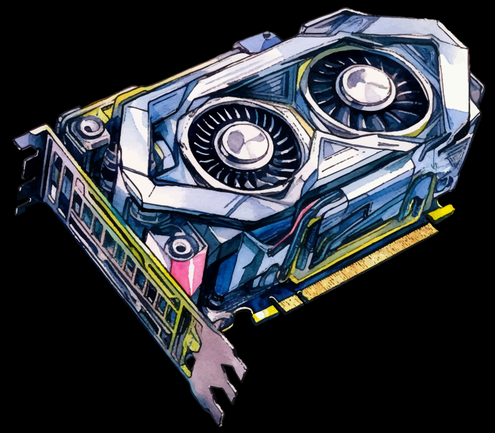
refers to
the **techniques and processes**
used to **efficiently**
access and retrieve
relevant information from
large datasets.

Key Limitations of LLMs
1. Hallucinations and Misinformation
2. Contextual Understanding

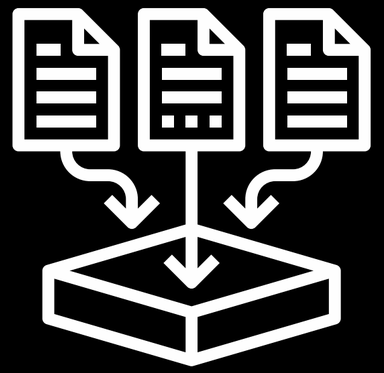


Benefits?

2 Paradigms of Inserting Knowledge into LLMs

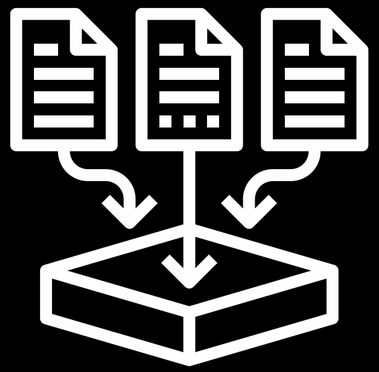


Fine-Tuning

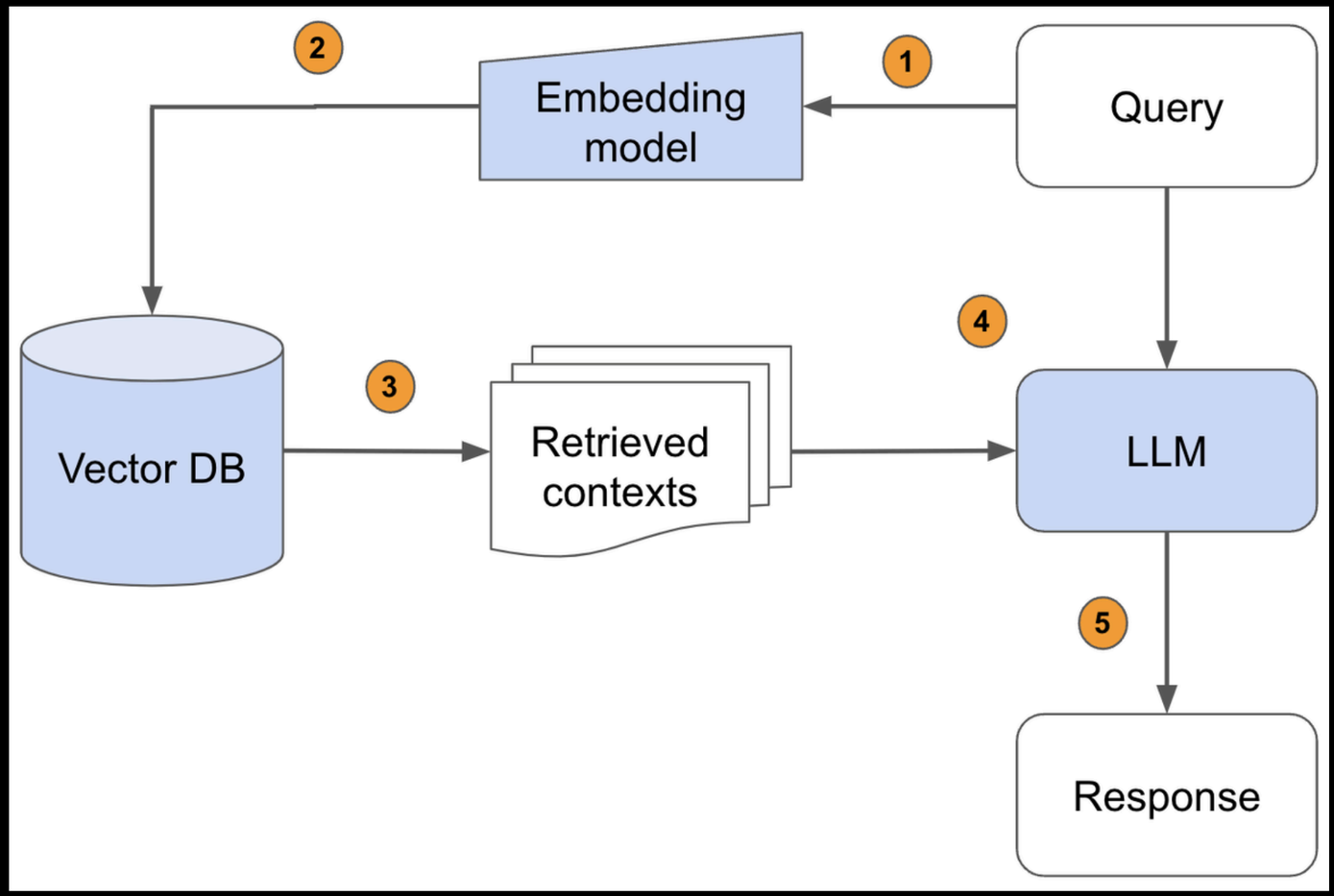


Retrieval Augmentation Generation (RAG)

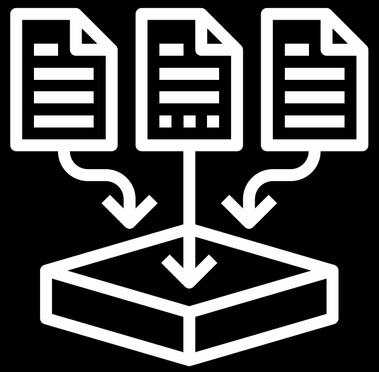
https://www.anyscale.com/blog/a-comprehensive-guide-for-building-rag-based-llm-applications-part-1



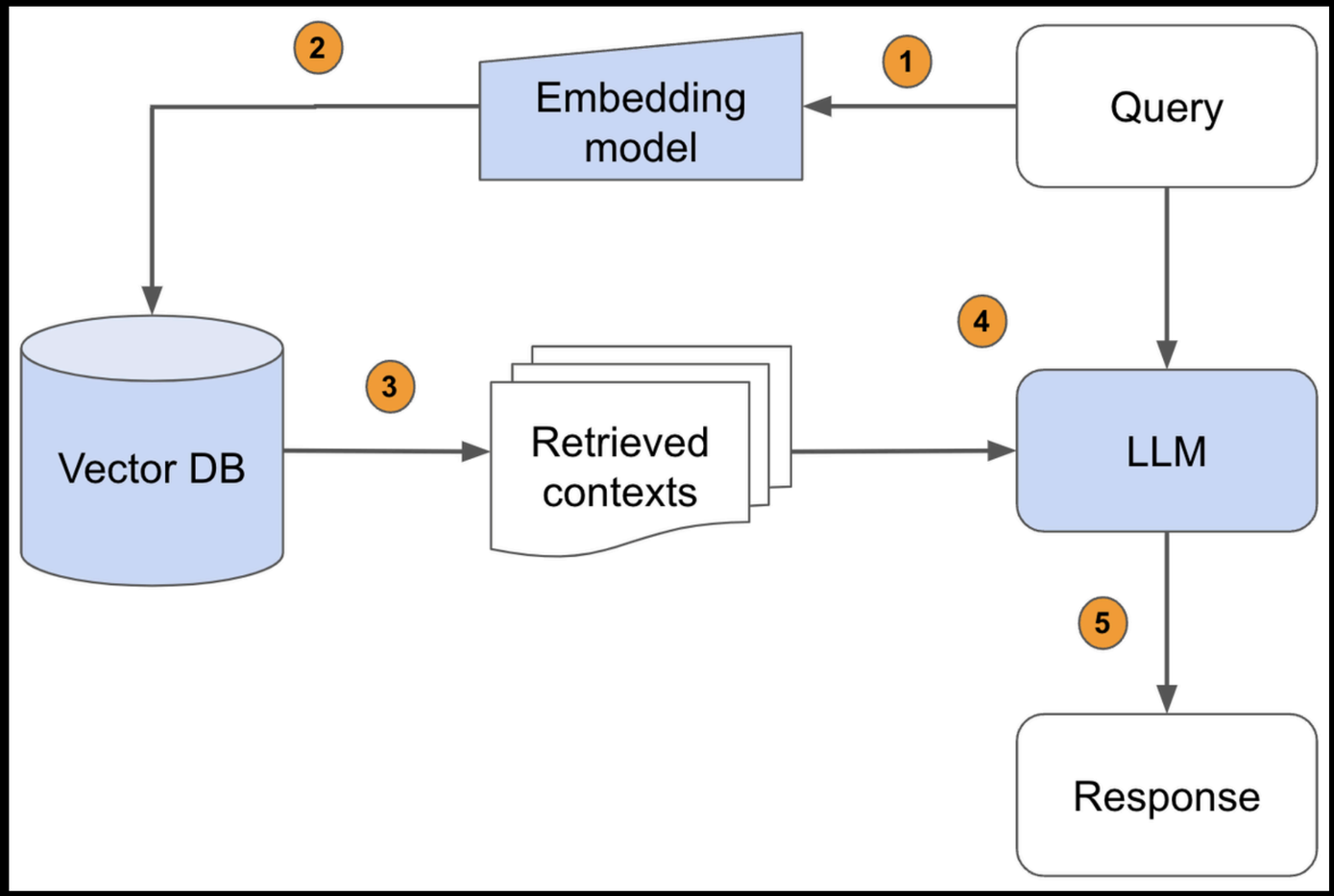
Retrieval Augmentation Generation



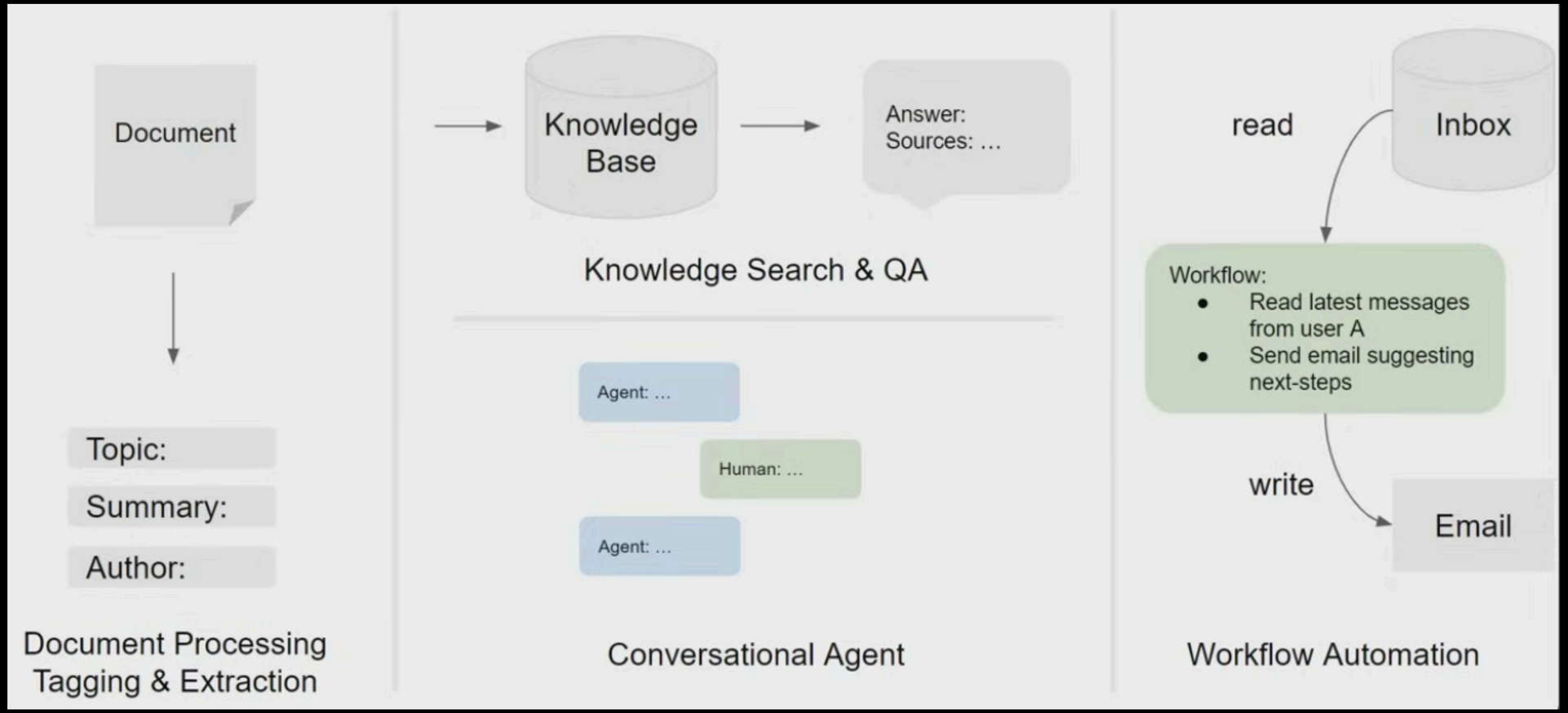
https://www.anyscale.com/blog/a-comprehensive-guide-for-building-rag-based-llm-applications-part-1

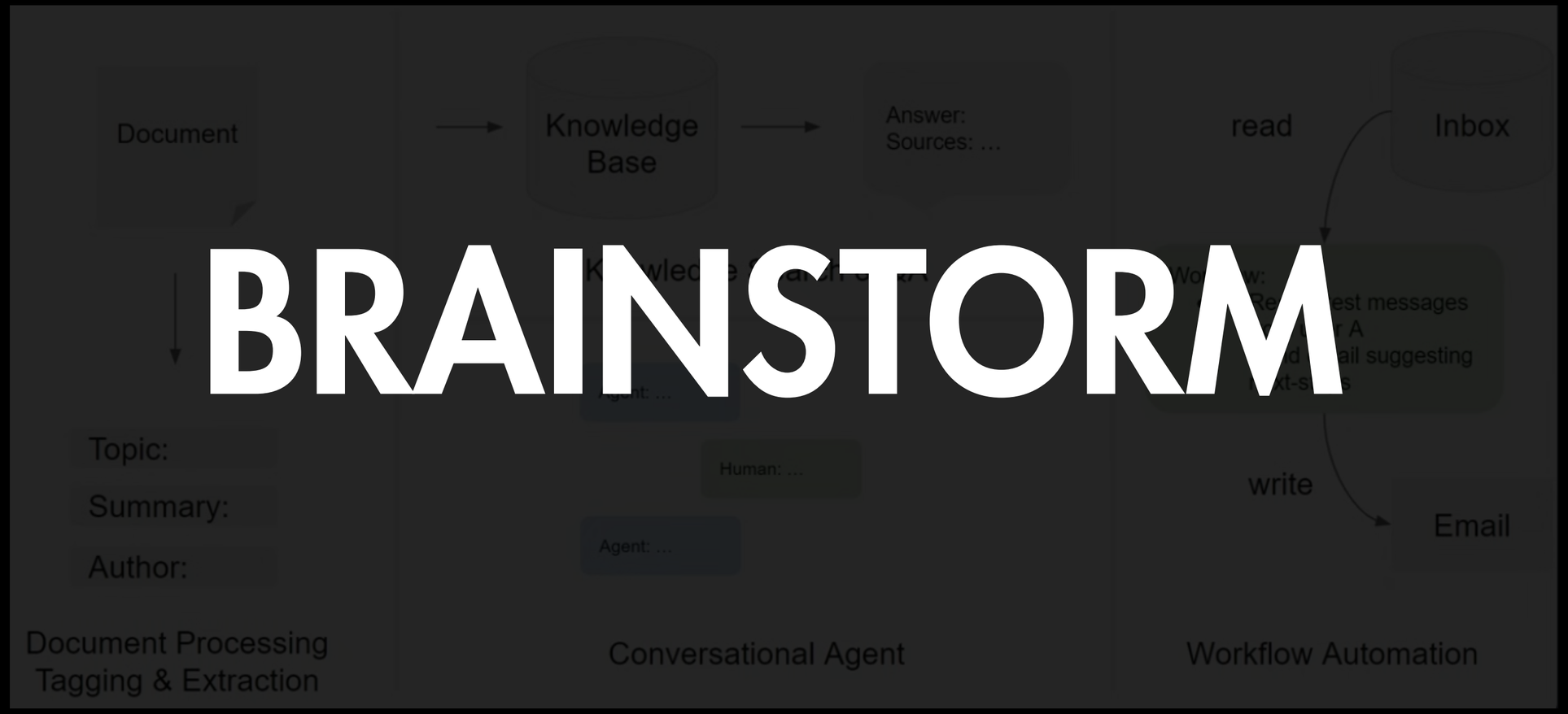


Retrieval Augmentation Generation



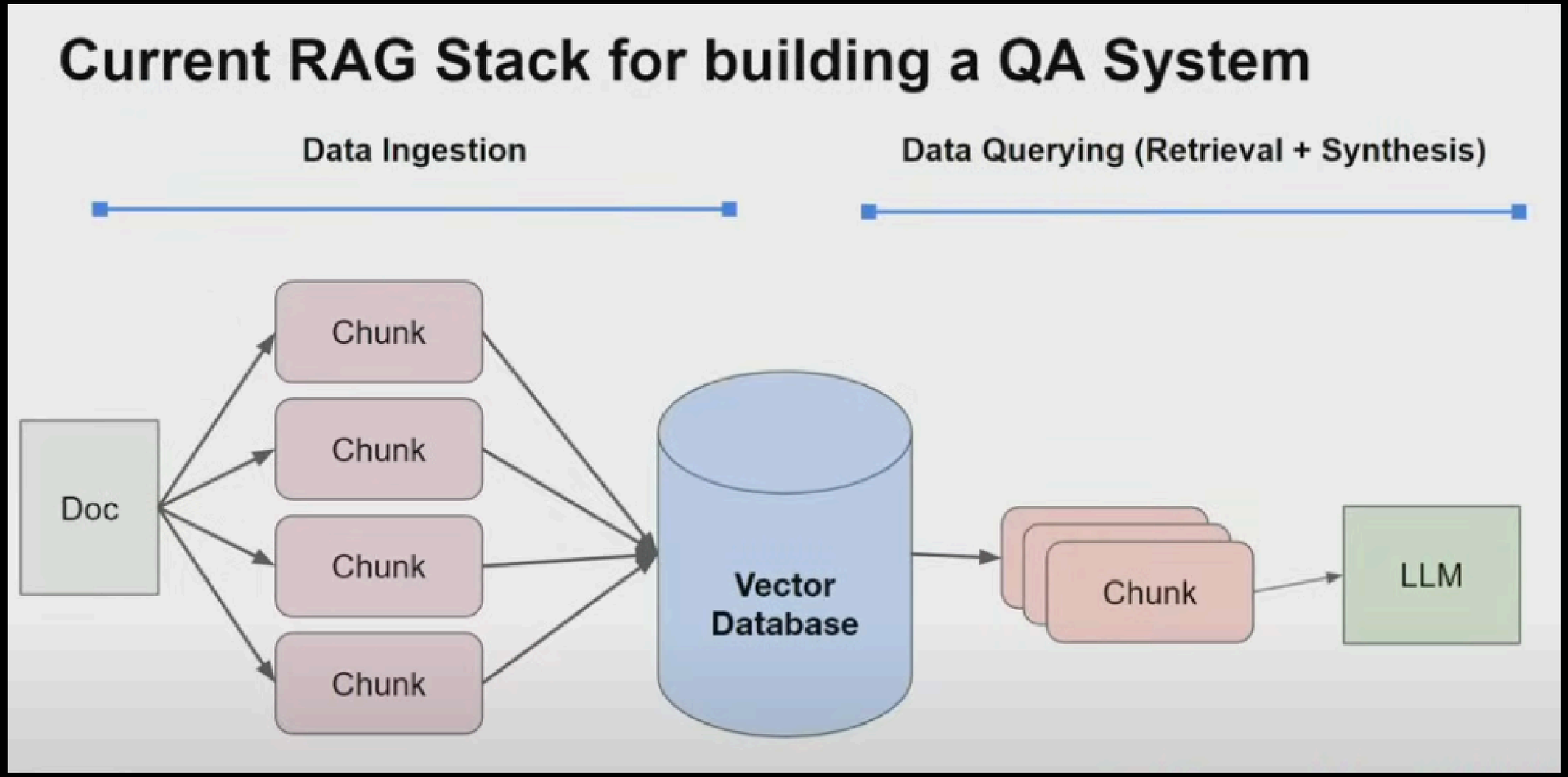
Source: https://www.youtube.com/watch?v=TRjq7t2Ms5I&ab_channel=AIEngineer



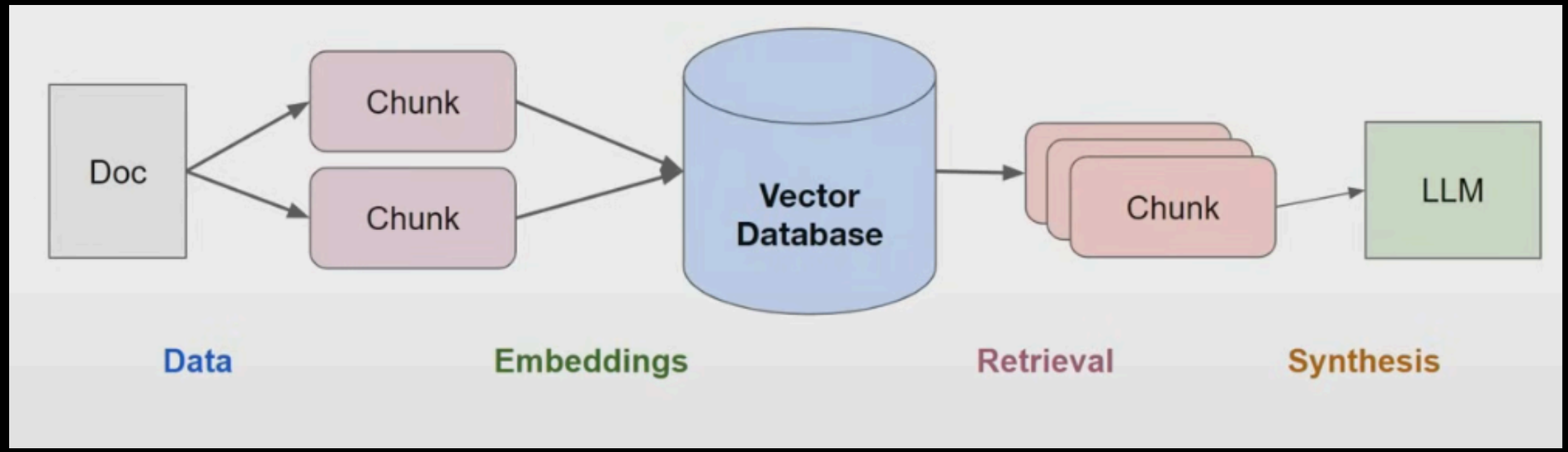

Document
Knowledge Base
Answer:
Sources: …
read
Inbox
Topic:
Summary:
Author:
Human: …
Agent: …
write
Email
Document Processing
Tagging & Extraction
Conversational Agent
Workflow Automation

BRAINSTORM

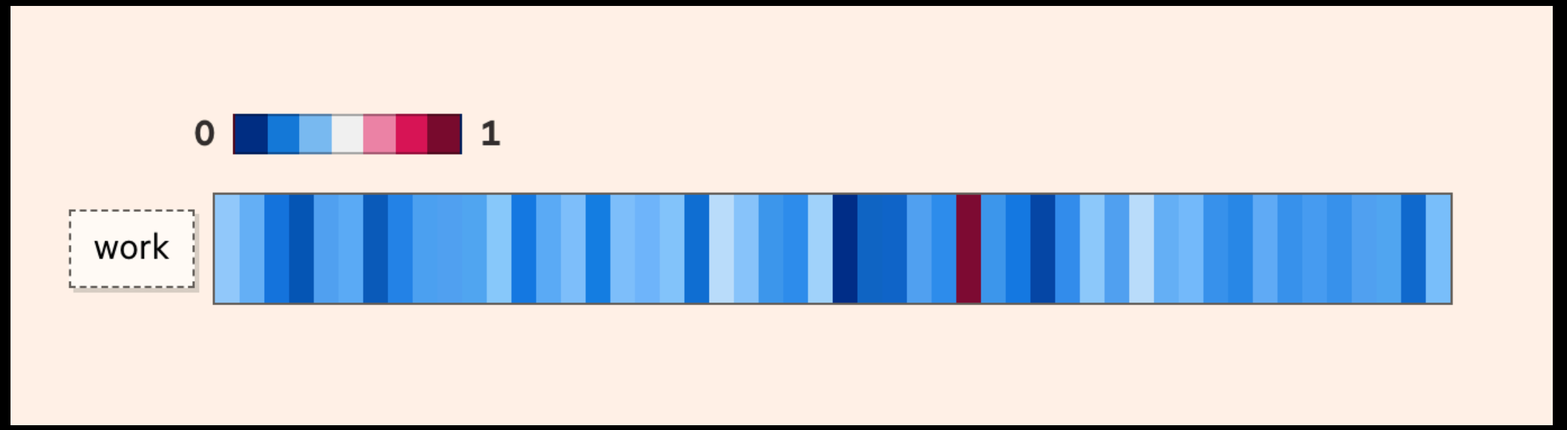
Source: https://www.youtube.com/watch?v=TRjq7t2Ms5I&ab_channel=AIEngineer



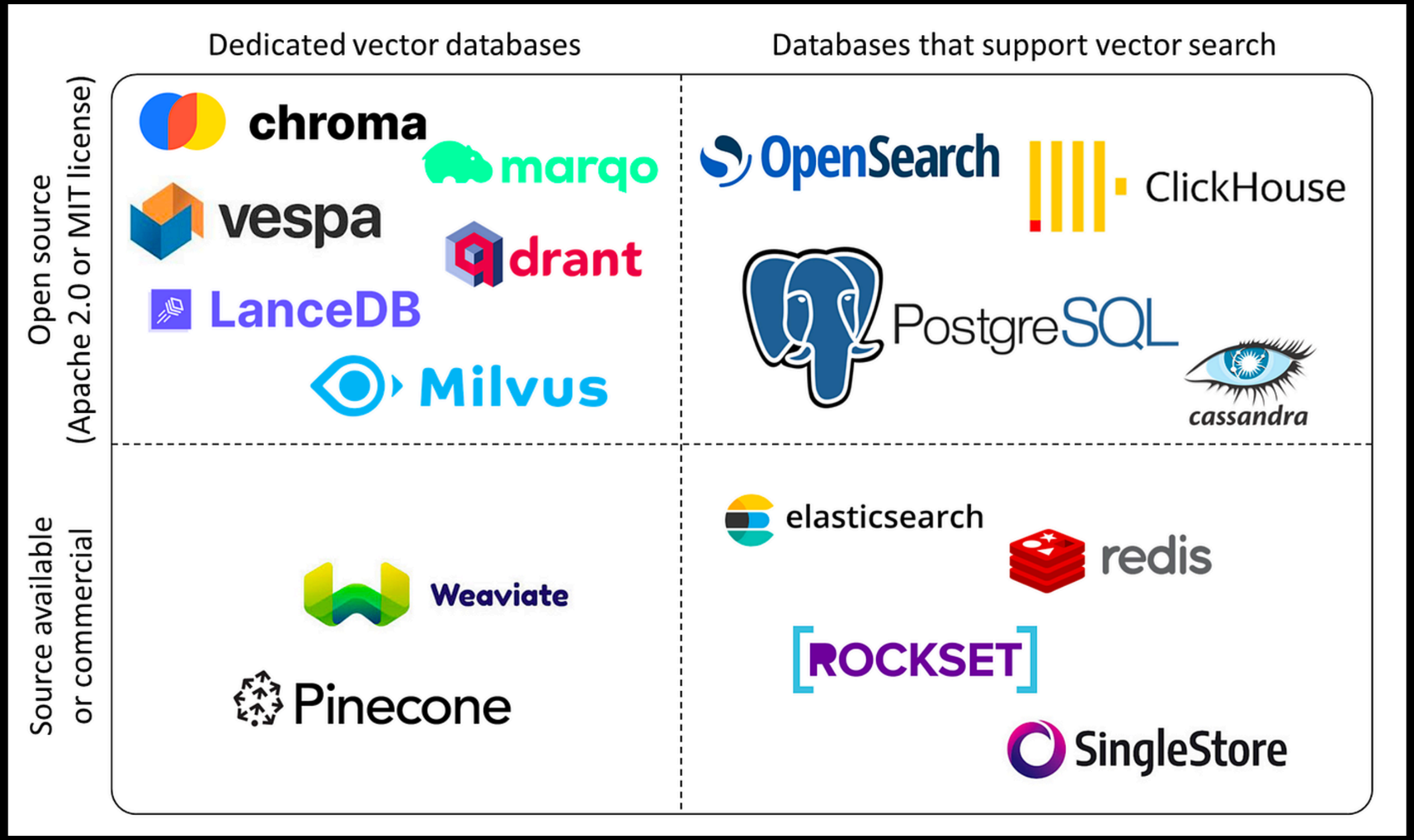
Source: https://www.youtube.com/watch?v=TRjq7t2Ms5I&ab_channel=AIEngineer



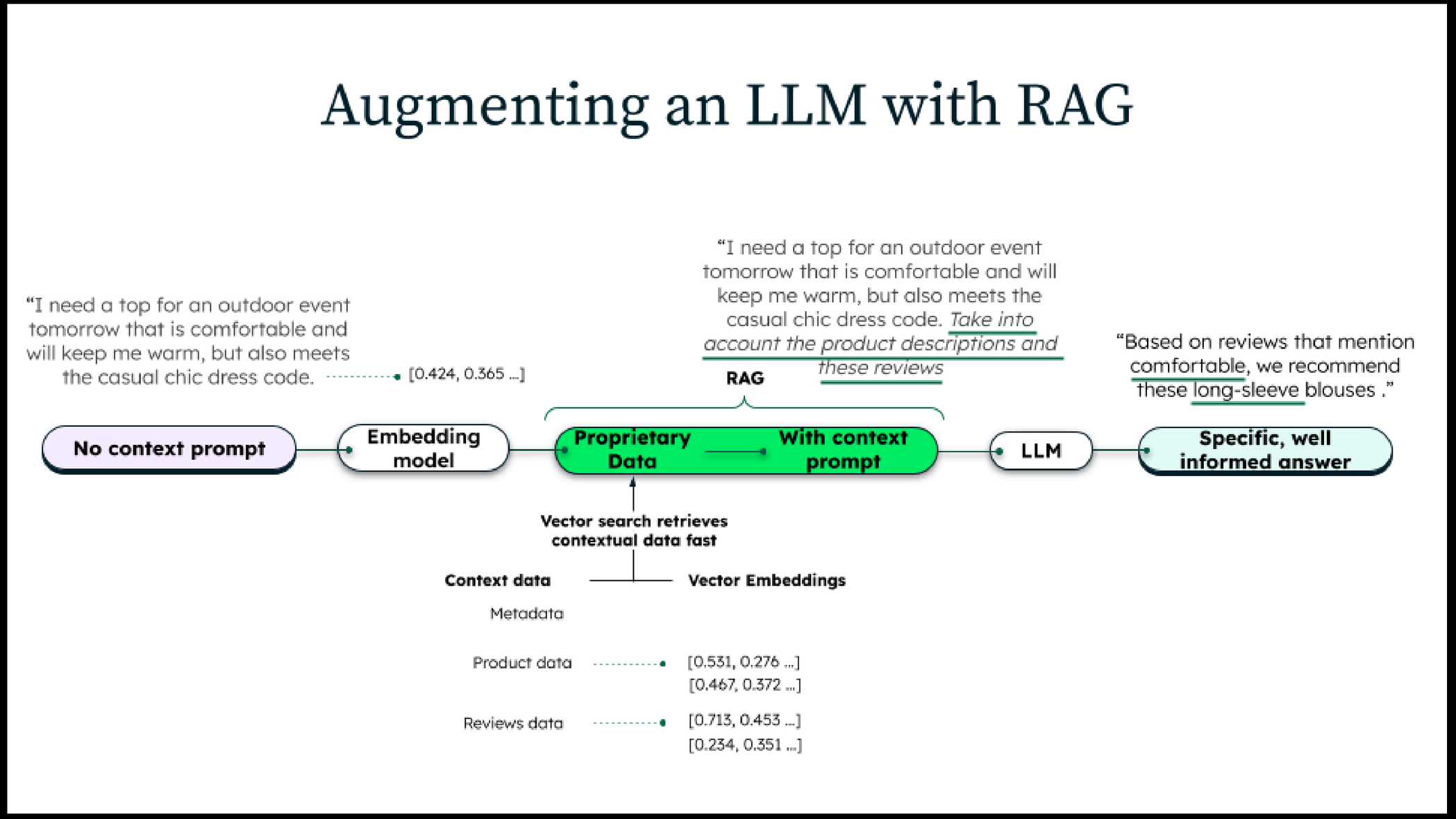
https://ig.ft.com/generative-ai/



https://www.datacamp.com/blog/the-top-5-vector-databases



https://www.mongodb.com/resources/basics/artificial-intelligence/retrieval-augmented-generation

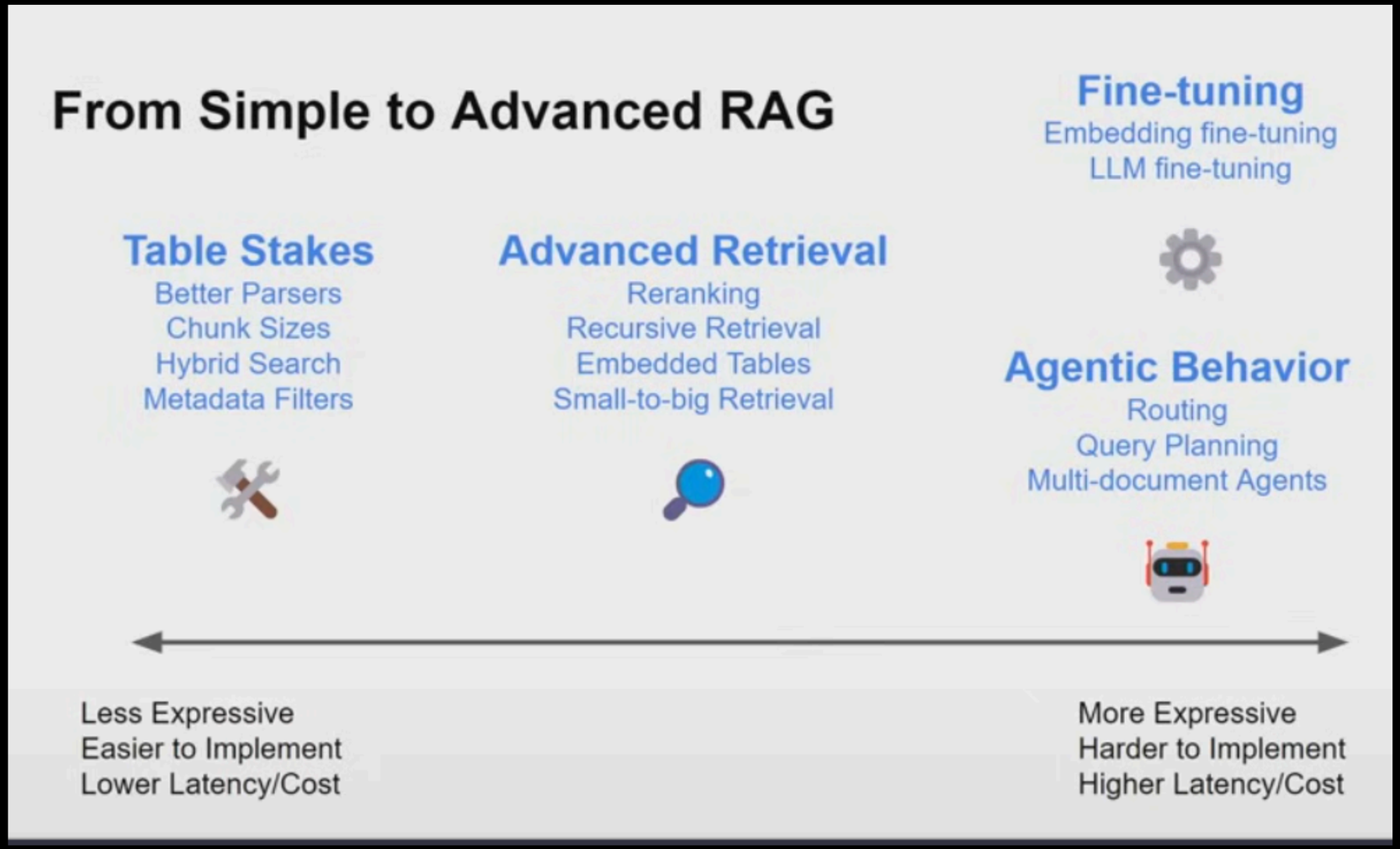


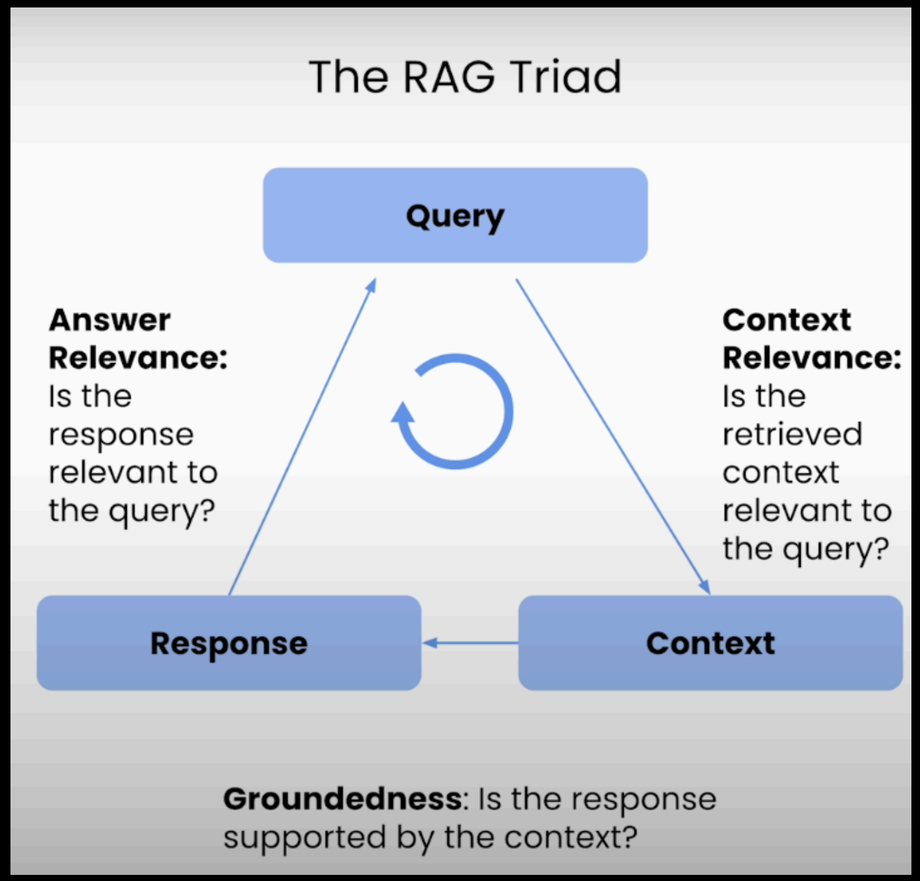
Source: https://www.youtube.com/watch?v=TRjq7t2Ms5I&ab_channel=AIEngineer

Challenges with Naive RAG (Response Quality)

- Bad Retrieval
  - **Low Precision:** Not all chunks in retrieved set are relevant
    - Hallucination + Lost in the Middle Problems
  - **Low Recall:** Now all relevant chunks are retrieved.
    - Lacks enough context for LLM to synthesize an answer
  - **Outdated information:** The data is redundant or out of date.
- Bad Response Generation
  - **Hallucination:** Model makes up an answer that isn't in the context.
  - **Irrelevance:** Model makes up an answer that doesn't answer the question.
  - **Toxicity/Bias:** Model makes up an answer that's harmful/offensive.

Source: https://www.youtube.com/watch?v=TRjq7t2Ms5I&ab_channel=AIEngineer




The RAG Triad
Query
Answer Relevance: Is the response relevant to the query?
Context Relevance: Is the retrieved context relevant to the query?
Response
Context
Groundedness: Is the response supported by the context?