LLamaIndex with RAG

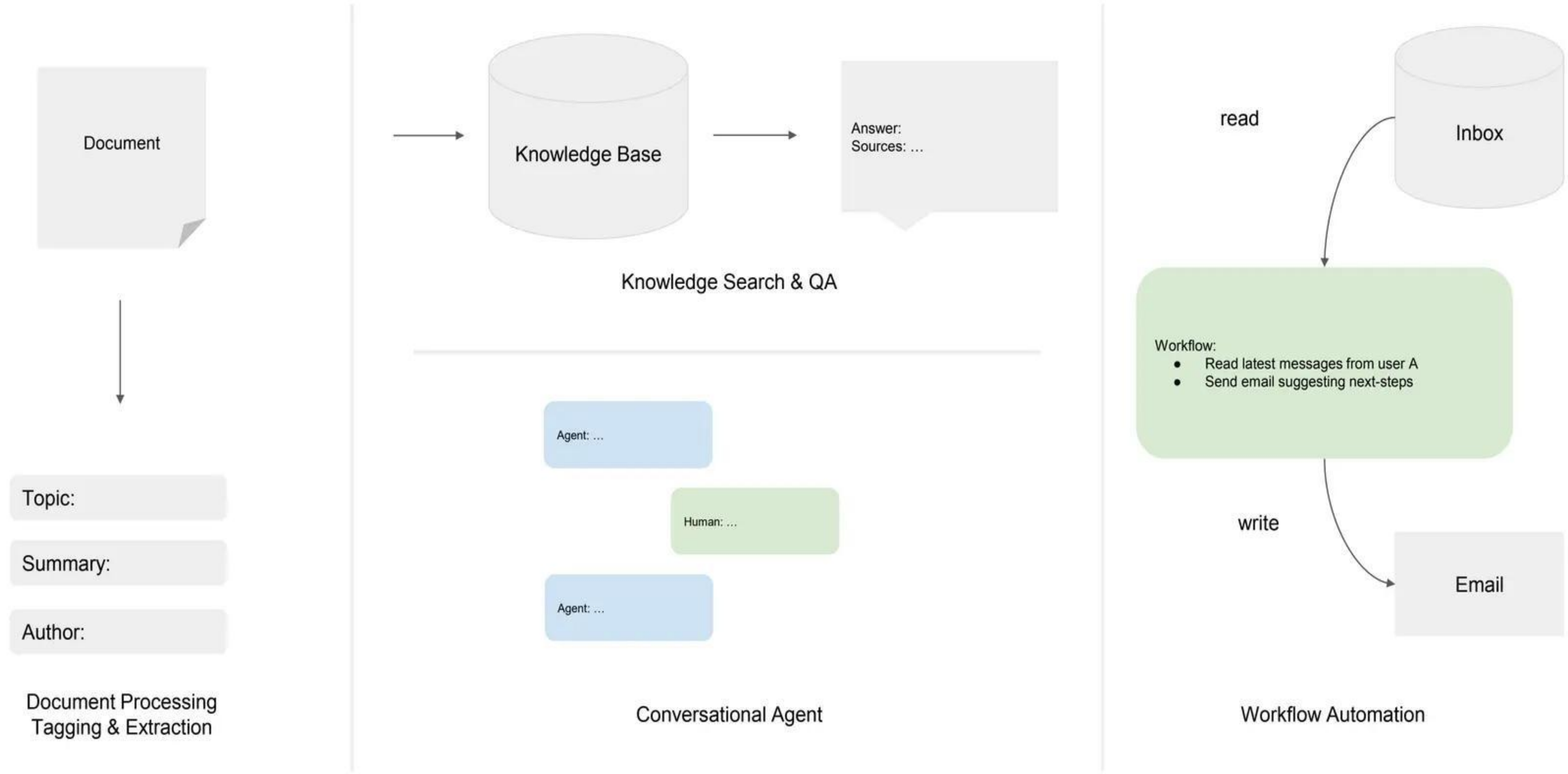
Sawradip Saha

Co-Founder, RunAgent

Co-founder & VP of ML, Intelsense AI

14 Oct 2025

enAI - Enterprise Use-cases



Document

Topic:

Summary:

Author:

Document Processing
Tagging & Extraction

Knowledge Base

Answer:
Sources: …

Knowledge Search & QA

Agent: …

Human: …

Agent: …

Conversational Agent

read

Inbox

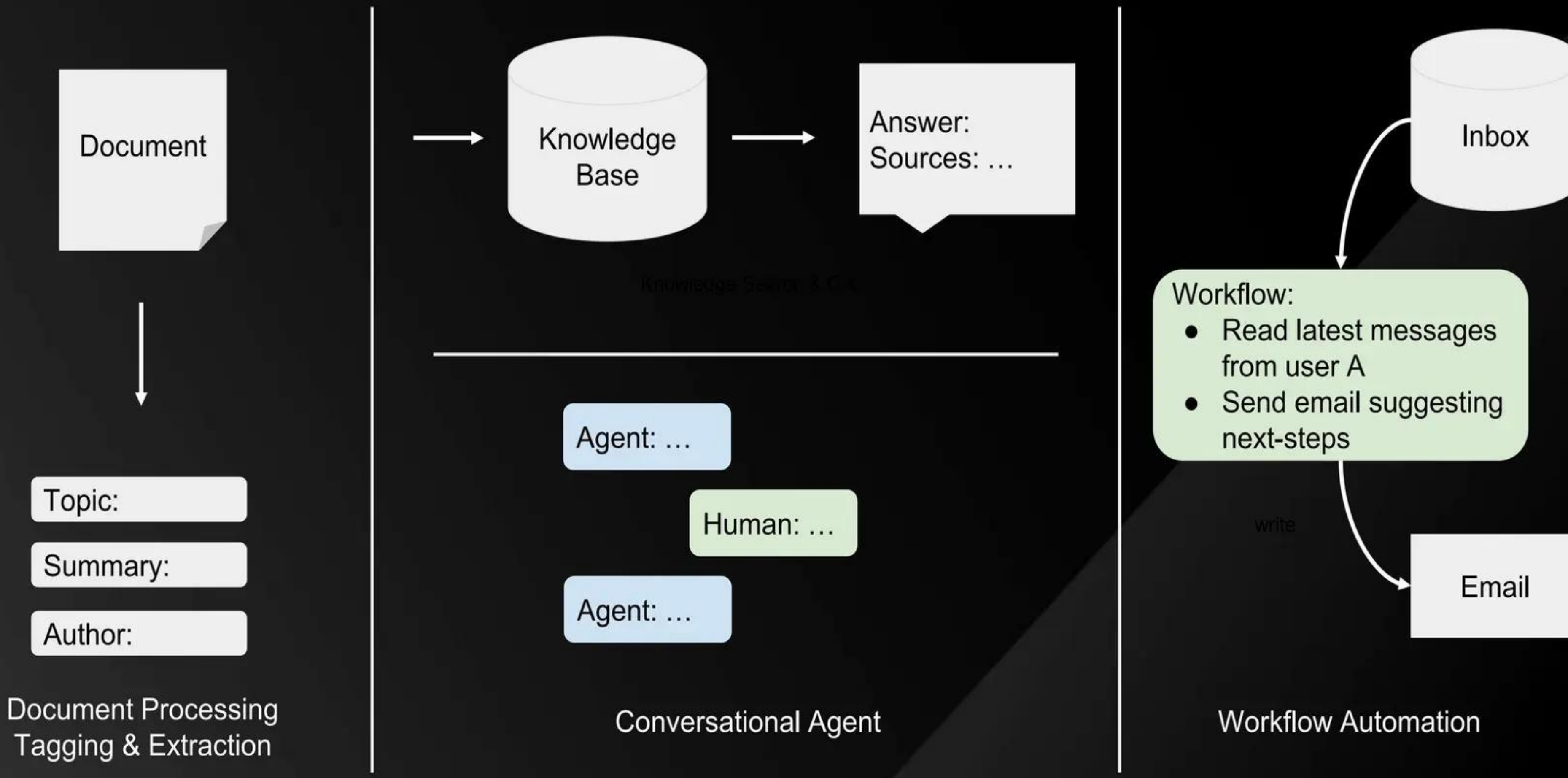
Workflow:
- Read latest messages from user A
- Send email suggesting next-steps

write

Email

Workflow Automation

GenAI - Enterprise Use-cases



Document

Topic:

Summary:

Author:

Document Processing Tagging & Extraction

Knowledge Base

Answer:
Sources: …

Knowledge Search & Q&A

Agent: …

Human: …

Agent: …

Conversational Agent

Inbox

Workflow:
- Read latest messages from user A
- Send email suggesting next-steps

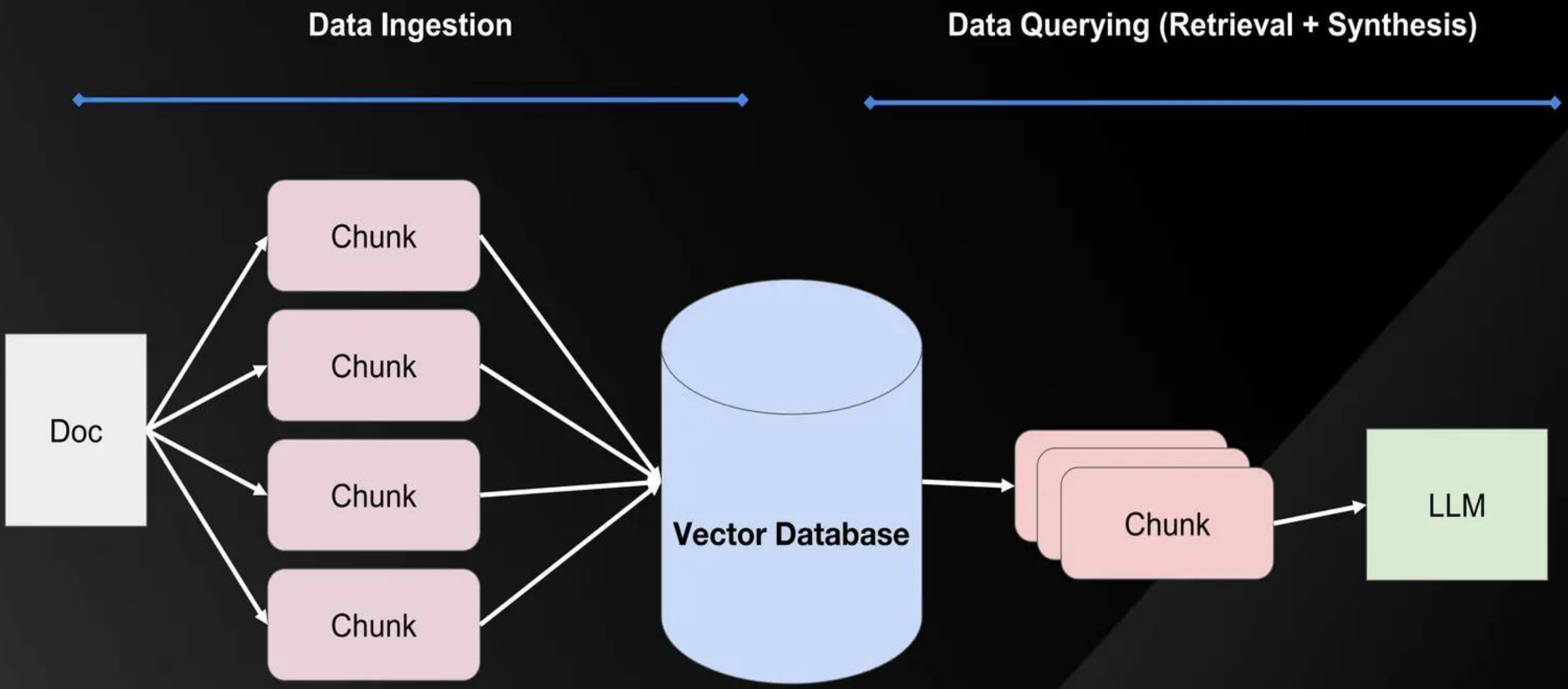
write

Email

Workflow Automation

RAG Stack

Current RAG Stack for building a QA System



Data Ingestion

Data Querying (Retrieval + Synthesis)

5 Lines of Code in LlamaIndex!

Challenges with “Naive” RAG

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    - **Outdated information:** The data is redundant or out of date.

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  - **Outdated information:** The data is redundant or out of date.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.

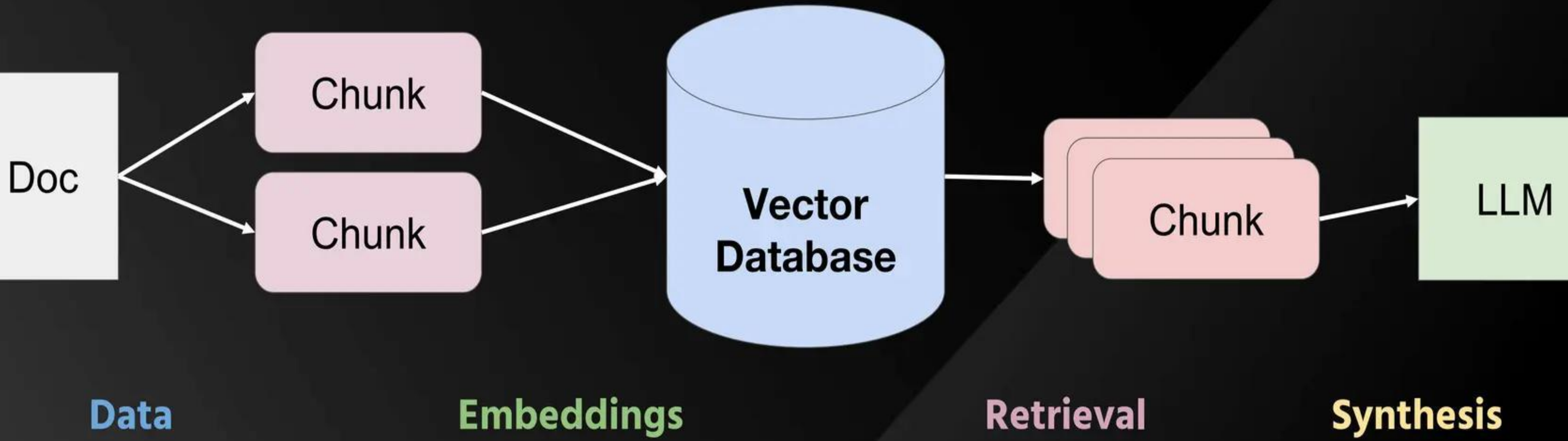
Vhat do we do?

Data: Can we store additional information beyond raw text chunks?

Embeddings: Can we optimize our embedding representations?

Retrieval: Can we do better than top-k embedding lookup?

Synthesis: Can we use LLMs for more than generation?



What do we do?

Data: Can we store additional information beyond raw text chunks?

Embeddings: Can we optimize our embedding representations?

Retrieval: Can we do better than top-k embedding lookup?

Synthesis: Can we use LLMs for more than generation?

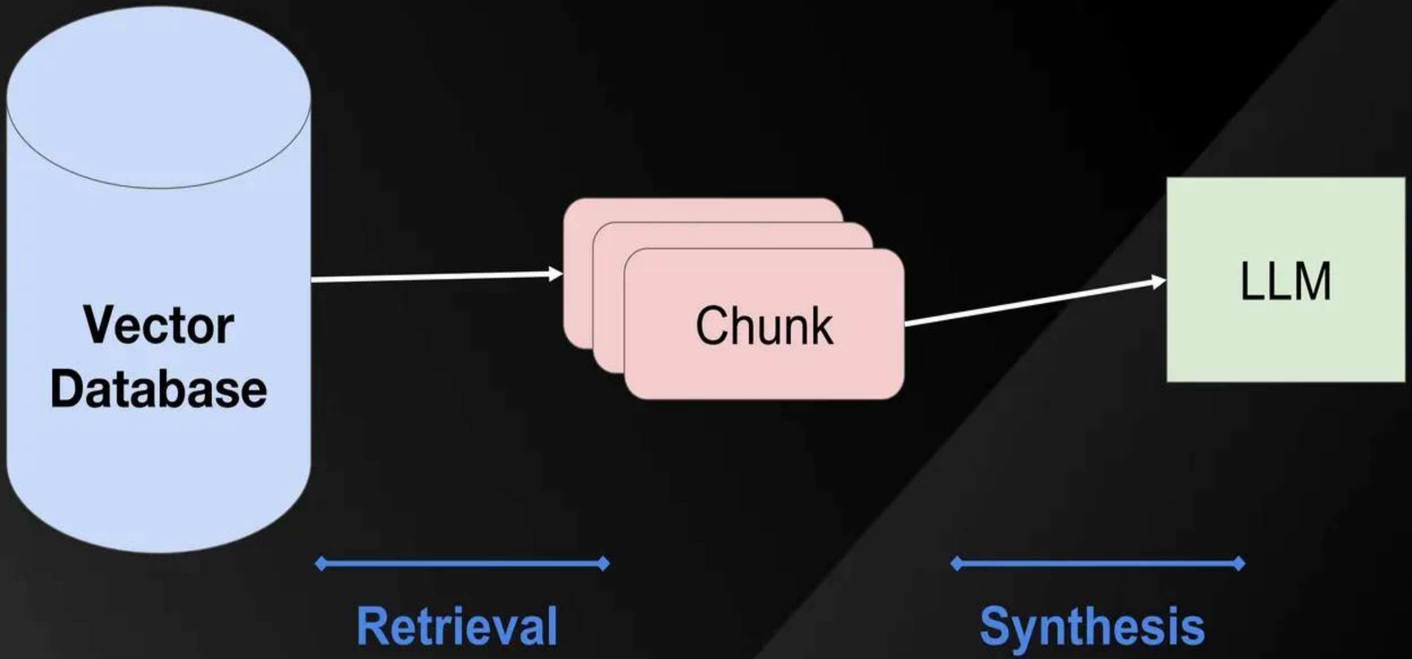
But before all this…

We need evals

Evaluation

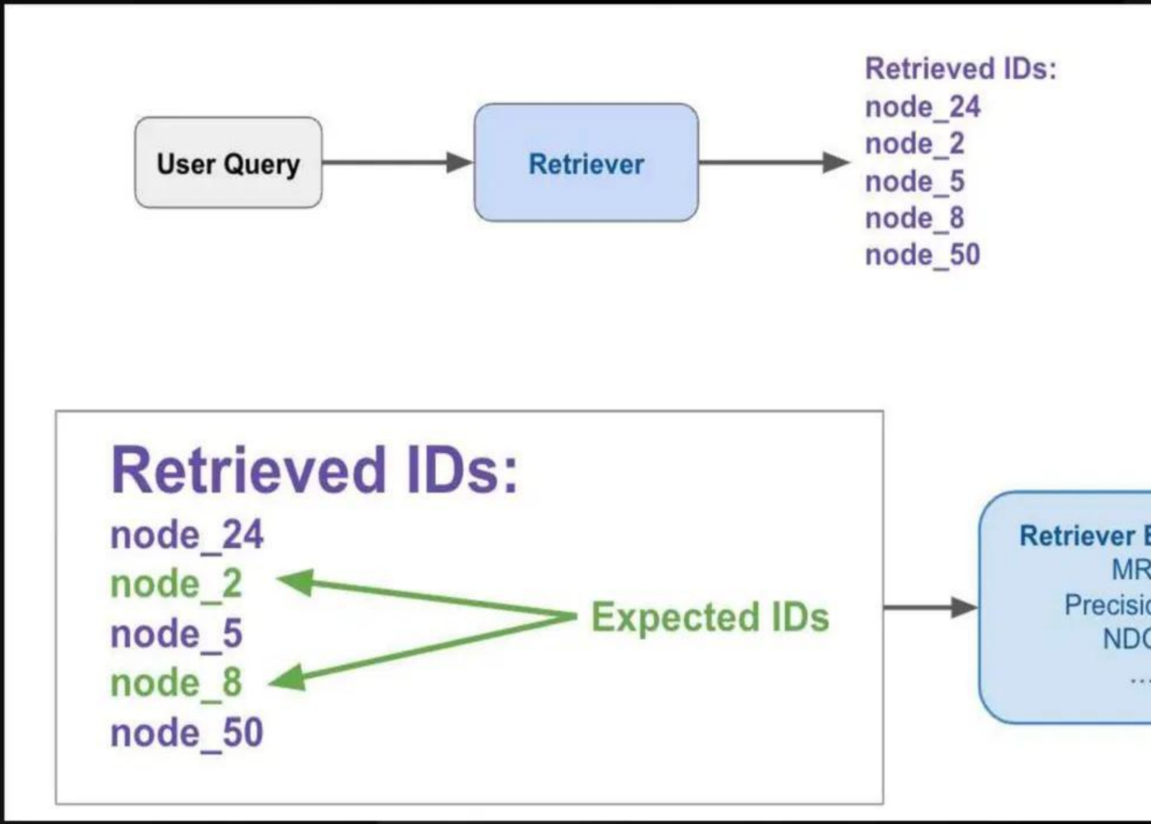
Evaluation

- How do we properly evaluate a RAG system?
  - Evaluate in isolation (retrieval, synthesis)
  - Evaluate e2e



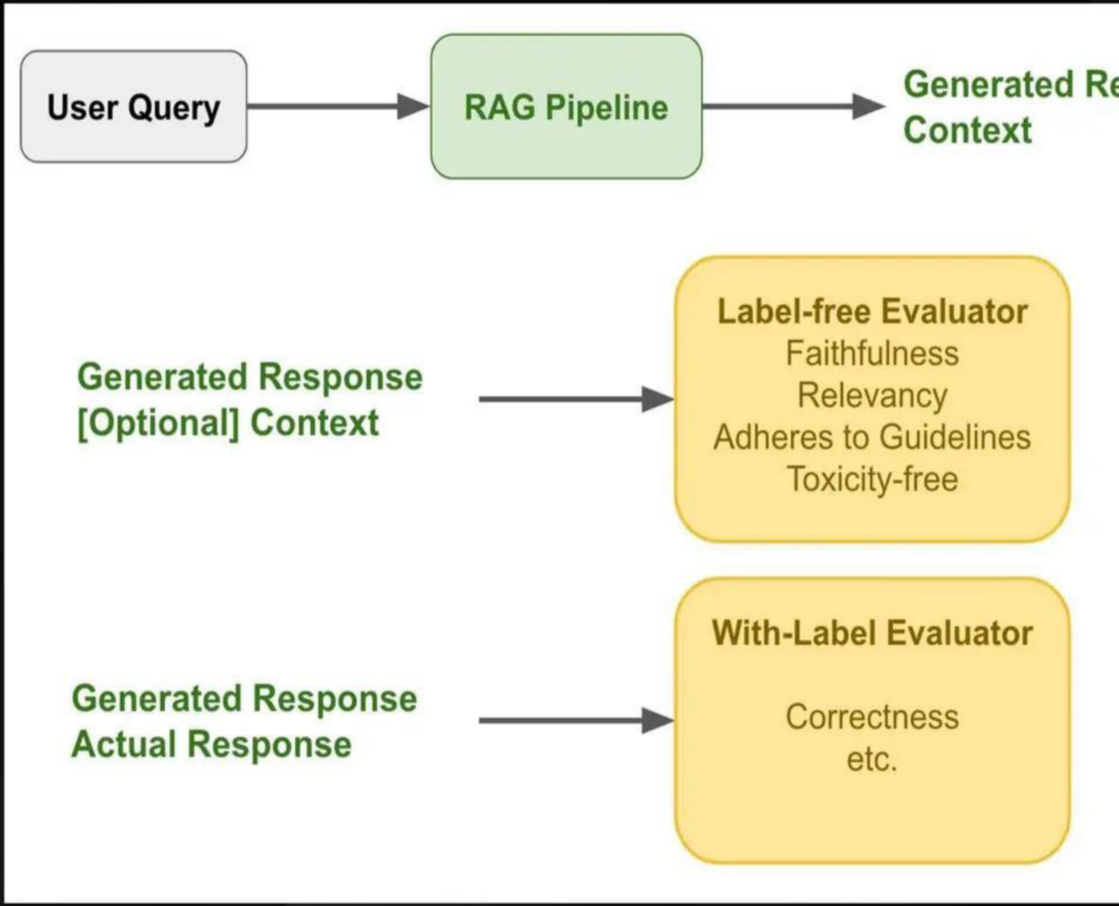
Evaluation in Isolation (Retrieval)

- Evaluate quality of retrieved chunks given user query
- **Create dataset**
  - Input: query
  - Output: the "ground-truth" documents relevant to the query
- Run retriever over dataset
- Measure **ranking metrics**
  - Success rate / hit-rate
  - MRR
  - Hit-rate



Evaluation E2E

- Evaluation of final generated response given input
- **Create Dataset**
    - Input: query
    - [Optional] Output: the “ground-truth” answer
- Run through full RAG pipeline
- Collect evaluation metrics:
    - **If no labels:** label-free evals
    - **If labels:** with-label evals



Optimizing RAG Systems

From Simple to Advanced

Table Stakes

Better Parsers
Chunk Sizes
Prompt Engineering
Customizing Models

Advanced Retrieval

Metadata Filtering
Recursive Retrieval
Embedded Tables
Small-to-big Retrieval

Fine-tuning

Embedding fine-tuning
LLM fine-tuning

Agentic Behavio

Routing
Query Planning
Multi-document Agents

ess Expressive
asier to Implement
ower Latency/Cost

More Expressive
Harder to Implemen
Higher Latency/Cos

Table Stakes: Chunk Sizes

Tuning your chunk size can have outsized impacts on performance

Not obvious that more retrieved tokens == higher performance!

Note: Reranking (shuffling context order) isn’t always beneficial.

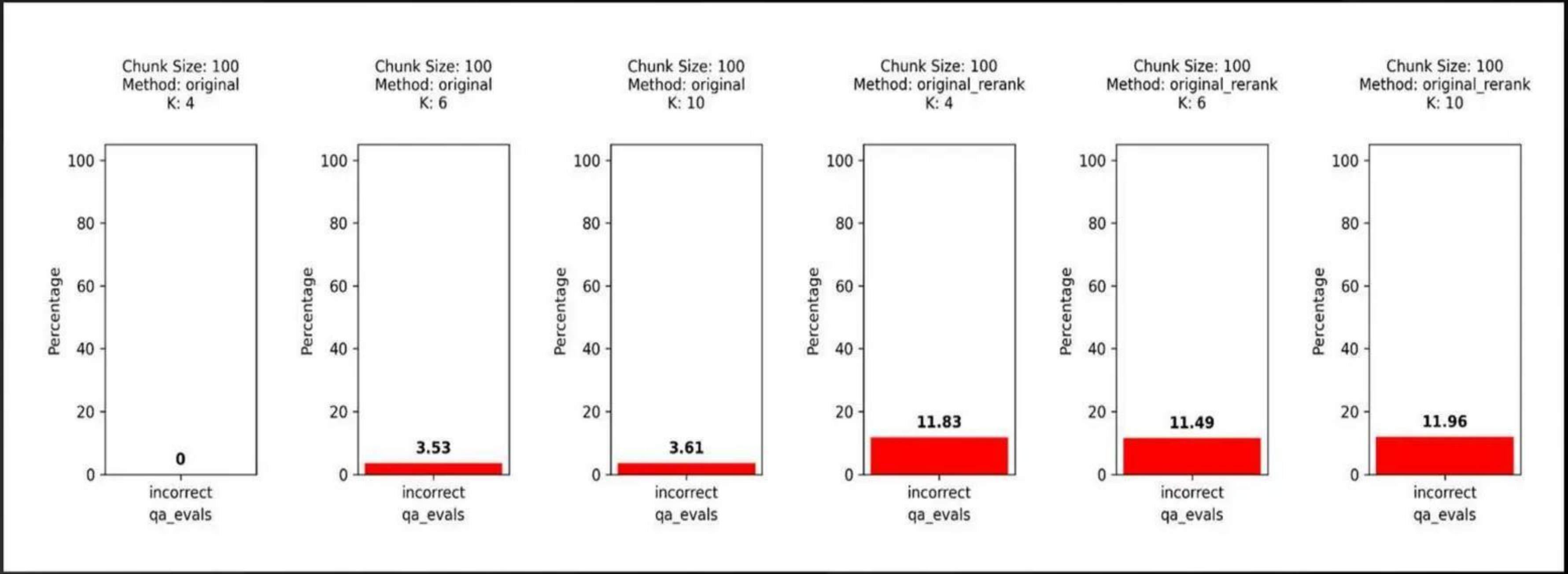
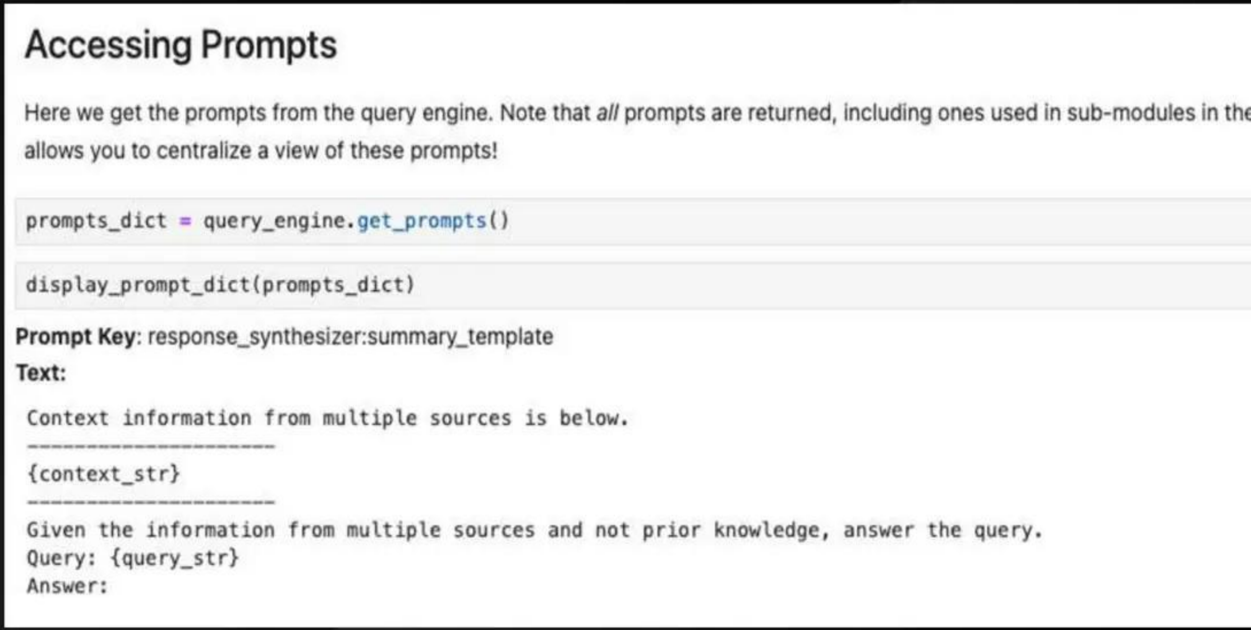


Table Stakes: Prompt Engineering

RAG uses core Question-Answering (QA) prompt templates

Ways you can customize:

- Adding few-shot examples
- Modifying template text
- Adding emotions



Accessing Prompts

Here we get the prompts from the query engine. Note that *all* prompts are returned, including ones used in sub-modules in the allows you to centralize a view of these prompts!

```
prompts_dict = query_engine.get_prompts()
```

```
display_prompt_dict(prompts_dict)
```

Prompt Key: response_synthesizer:summary_template

Text:

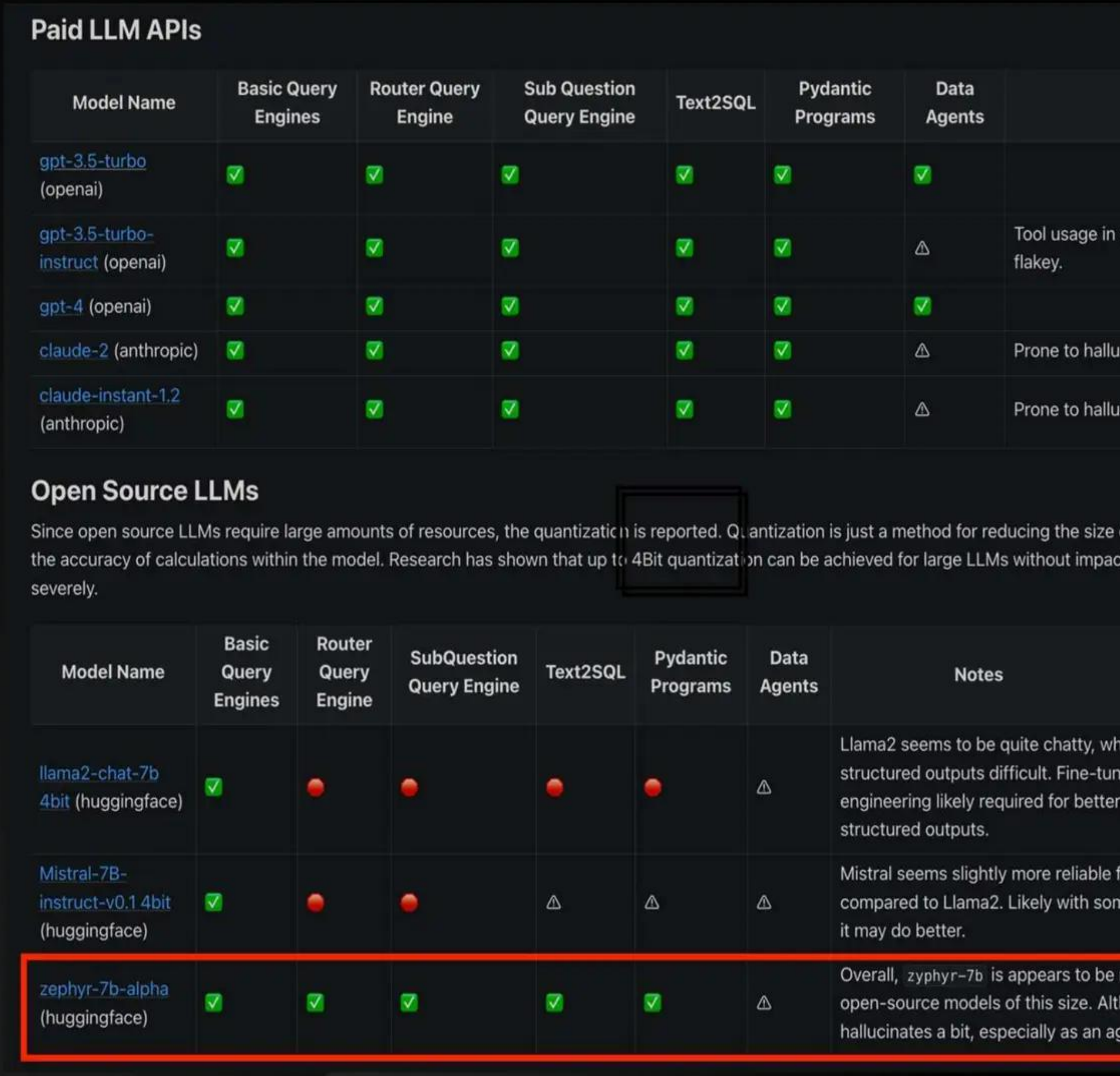
```
Context information from multiple sources is below.
---------------------
{context_str}
---------------------
Given the information from multiple sources and not prior knowledge, answer the query.
Query: {query_str}
Answer:
```

Table Stakes: Customizing LLMs

sk performance on easy-

-hard tasks (RAG, agents)

ries **wildly** among LLMs



Paid LLM APIs

Model Name	Basic Query Engines	Router Query Engine	Sub Question Query Engine	Text2SQL	Pydantic Programs	Data Agents	
gpt-3.5-turbo (openai)	✅	✅	✅	✅	✅	✅	
gpt-3.5-turbo-instruct (openai)	✅	✅	✅	✅	✅	⚠	Tool usage in flakey.
gpt-4 (openai)	✅	✅	✅	✅	✅	✅	
claude-2 (anthropic)	✅	✅	✅	✅	✅	⚠	Prone to hallu
claude-instant-1.2 (anthropic)	✅	✅	✅	✅	✅	⚠	Prone to hallu

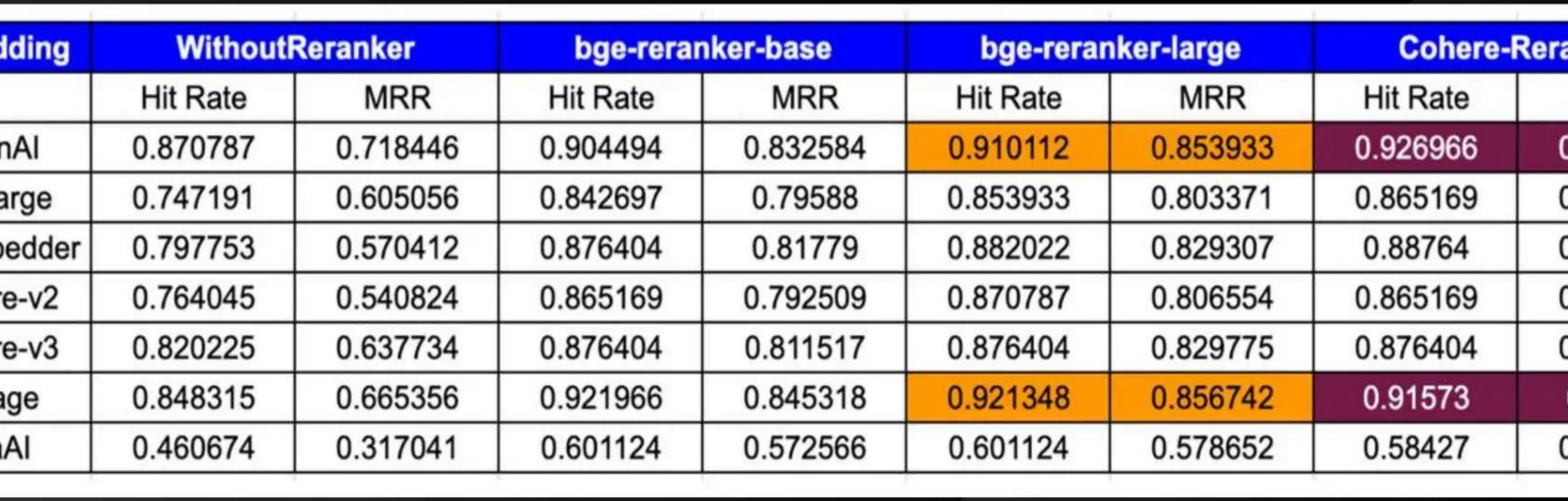
Open Source LLMs

Since open source LLMs require large amounts of resources, the quantization is reported. Quantization is just a method for reducing the size the accuracy of calculations within the model. Research has shown that up to 4Bit quantization can be achieved for large LLMs without impac severely.

Model Name	Basic Query Engines	Router Query Engine	SubQuestion Query Engine	Text2SQL	Pydantic Programs	Data Agents	Notes
llama2-chat-7b 4bit (huggingface)	✅	🔴	🔴	🔴	🔴	⚠	Llama2 seems to be quite chatty, wh structured outputs difficult. Fine-tun engineering likely required for better structured outputs.
Mistral-7B-instruct-v0.1 4bit (huggingface)	✅	🔴	🔴	⚠	⚠	⚠	Mistral seems slightly more reliable f compared to Llama2. Likely with son it may do better.
zephyr-7b-alpha (huggingface)	✅	✅	✅	✅	✅	⚠	Overall, `zyphyr-7b` is appears to be open-source models of this size. Alt hallucinates a bit, especially as an ag

Table Stakes: Customizing Embeddings

Your embedding model + reranker affects retrieval quality



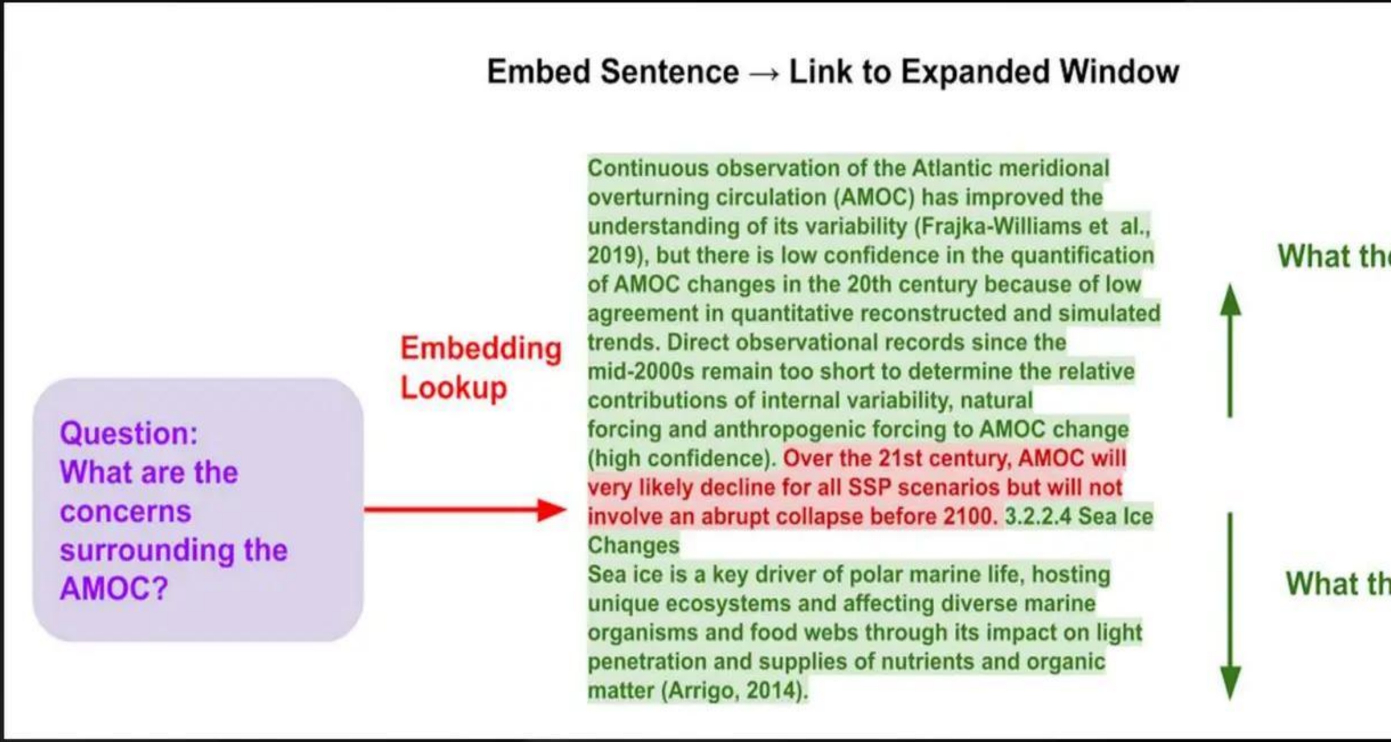
dding	WithoutReranker		bge-reranker-base		bge-reranker-large		Cohere-Rera	
	Hit Rate	MRR	Hit Rate	MRR	Hit Rate	MRR	Hit Rate	
nAI	0.870787	0.718446	0.904494	0.832584	0.910112	0.853933	0.926966	0
arge	0.747191	0.605056	0.842697	0.79588	0.853933	0.803371	0.865169	0
bedder	0.797753	0.570412	0.876404	0.81779	0.882022	0.829307	0.88764	0
re-v2	0.764045	0.540824	0.865169	0.792509	0.870787	0.806554	0.865169	0
re-v3	0.820225	0.637734	0.876404	0.811517	0.876404	0.829775	0.876404	0
age	0.848315	0.665356	0.921966	0.845318	0.921348	0.856742	0.91573	0
AI	0.460674	0.317041	0.601124	0.572566	0.601124	0.578652	0.58427	0

Source: https://blog.llamaindex.ai/boosting-rag-picking-the-best-embedding-reranker-models-42d079022e83

Advanced Retrieval: Small-to-Big

Intuition: Embedding a big text chunk feels suboptimal.

Solution: Embed text at the sentence-level - then **expand** that window during LLM synthesis



Embed Sentence → Link to Expanded Window

Question: What are the concerns surrounding the AMOC?

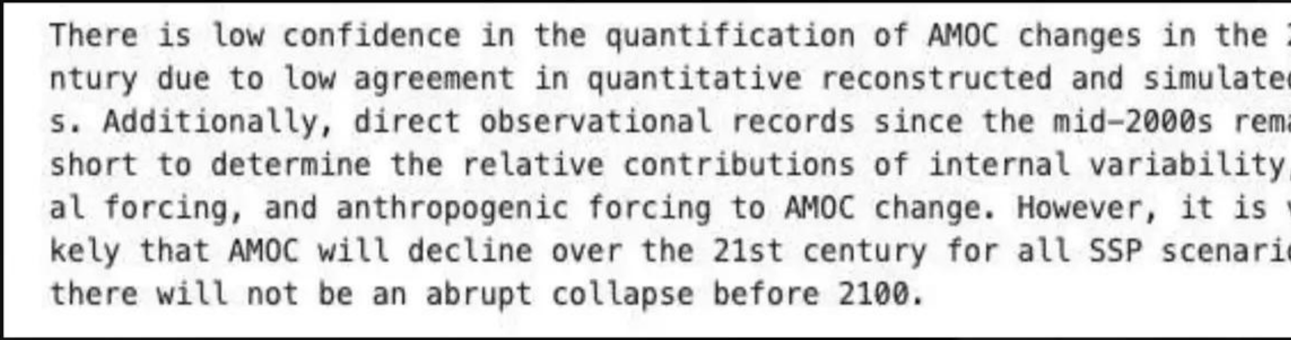
Continuous observation of the Atlantic meridional overturning circulation (AMOC) has improved the understanding of its variability (Frajka-Williams et al., 2019), but there is low confidence in the quantification of AMOC changes in the 20th century because of low agreement in quantitative reconstructed and simulated trends. Direct observational records since the mid-2000s remain too short to determine the relative contributions of internal variability, natural forcing and anthropogenic forcing to AMOC change (high confidence). Over the 21st century, AMOC will very likely decline for all SSP scenarios but will not involve an abrupt collapse before 2100. 3.2.2.4 Sea Ice Changes

Sea ice is a key driver of polar marine life, hosting unique ecosystems and affecting diverse marine organisms and food webs through its impact on light penetration and supplies of nutrients and organic matter (Arrigo, 2014).

Advanced Retrieval: Small-to-Big

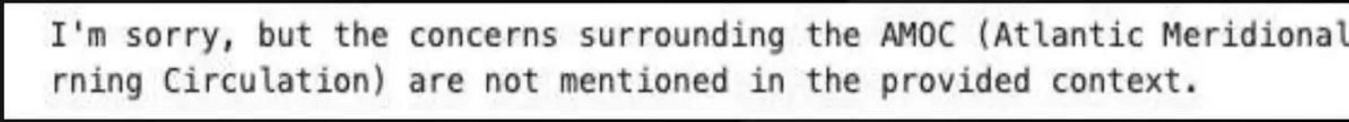
Leads to more **precise** retrieval.

Avoids "lost in the middle" problems.



```
There is low confidence in the quantification of AMOC changes in the
ntury due to low agreement in quantitative reconstructed and simulate
s. Additionally, direct observational records since the mid-2000s rem
short to determine the relative contributions of internal variability
al forcing, and anthropogenic forcing to AMOC change. However, it is
kely that AMOC will decline over the 21st century for all SSP scenari
there will not be an abrupt collapse before 2100.
```

Sentence Window Retrieval (k=2)



```
I'm sorry, but the concerns surrounding the AMOC (Atlantic Meridional
rning Circulation) are not mentioned in the provided context.
```

Naive Retrieval (k=5)

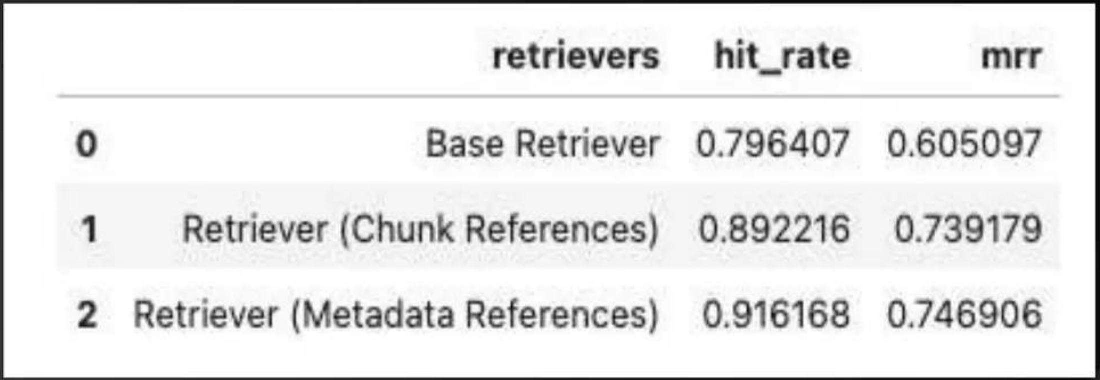
Only one out of the 5 chunks is relevant - "lost in the middle" problem

Advanced Retrieval: Small-to-Big

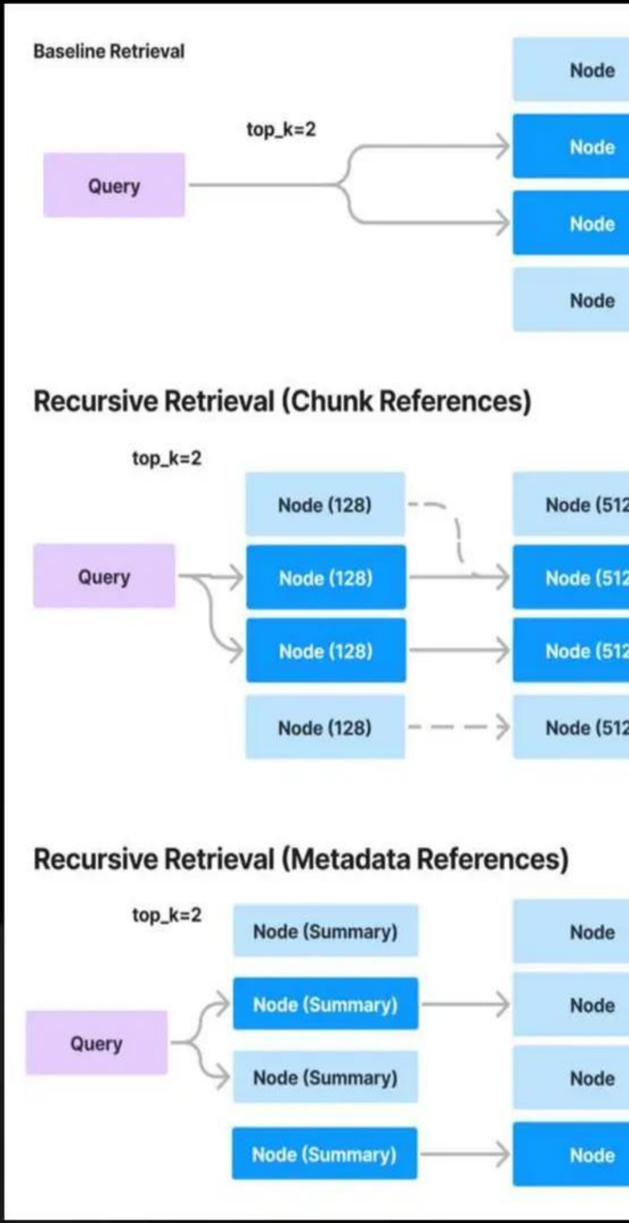
Intuition: Embedding a big text chunk feels suboptimal.

Solution: Embed a smaller **reference** to the parent chunk. Use parent chunk for synthesis

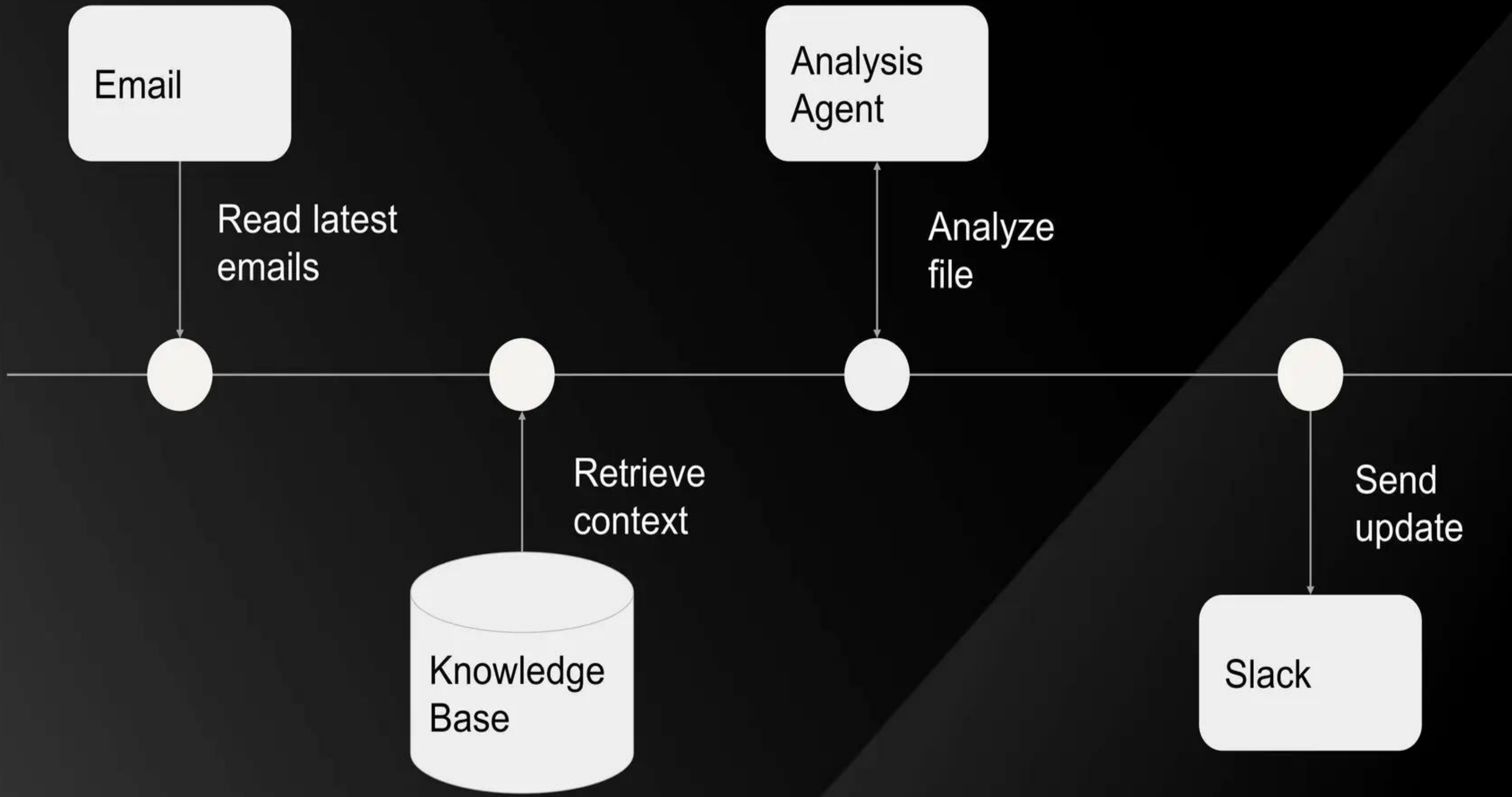
Examples: Smaller chunks, summaries, metadata



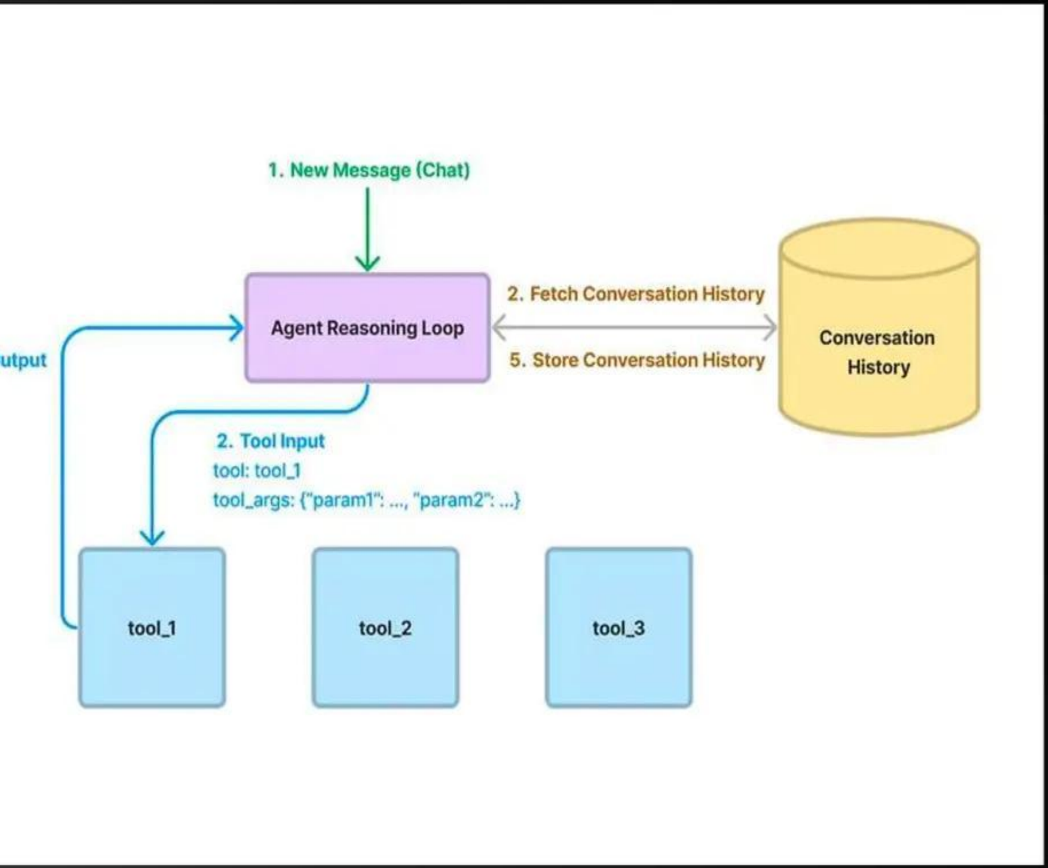
	retrievers	hit_rate	mrr
0	Base Retriever	0.796407	0.605097
1	Retriever (Chunk References)	0.892216	0.739179
2	Retriever (Metadata References)	0.916168	0.746906



Data Agents - LLM-powered knowledge workers



Data Agents - Core Components



Agent Reasoning Loop

- ReAct Agent (any LLM)
- OpenAI Agent (only OAI)

Tools via LlamaHub

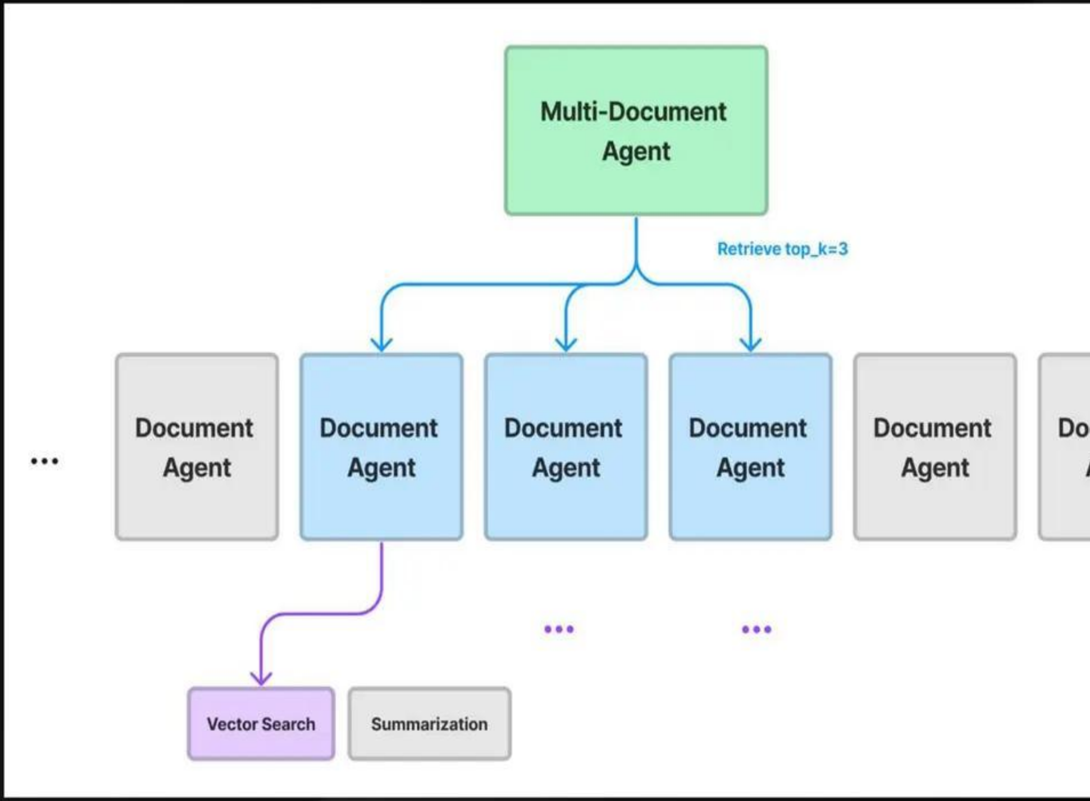
- Code interpreter
- Slack
- Notion
- Zapier
- … (15+ tools, ~100 loaders)

Agentic Behavior: Multi-Document Ager

Intuition: There's certain questions that "top-k" RAG can't answer.

Solution: Multi-Document Agents

- Fact-based QA and Summarization over any subsets of documents
- Chain-of-thought and query planning.

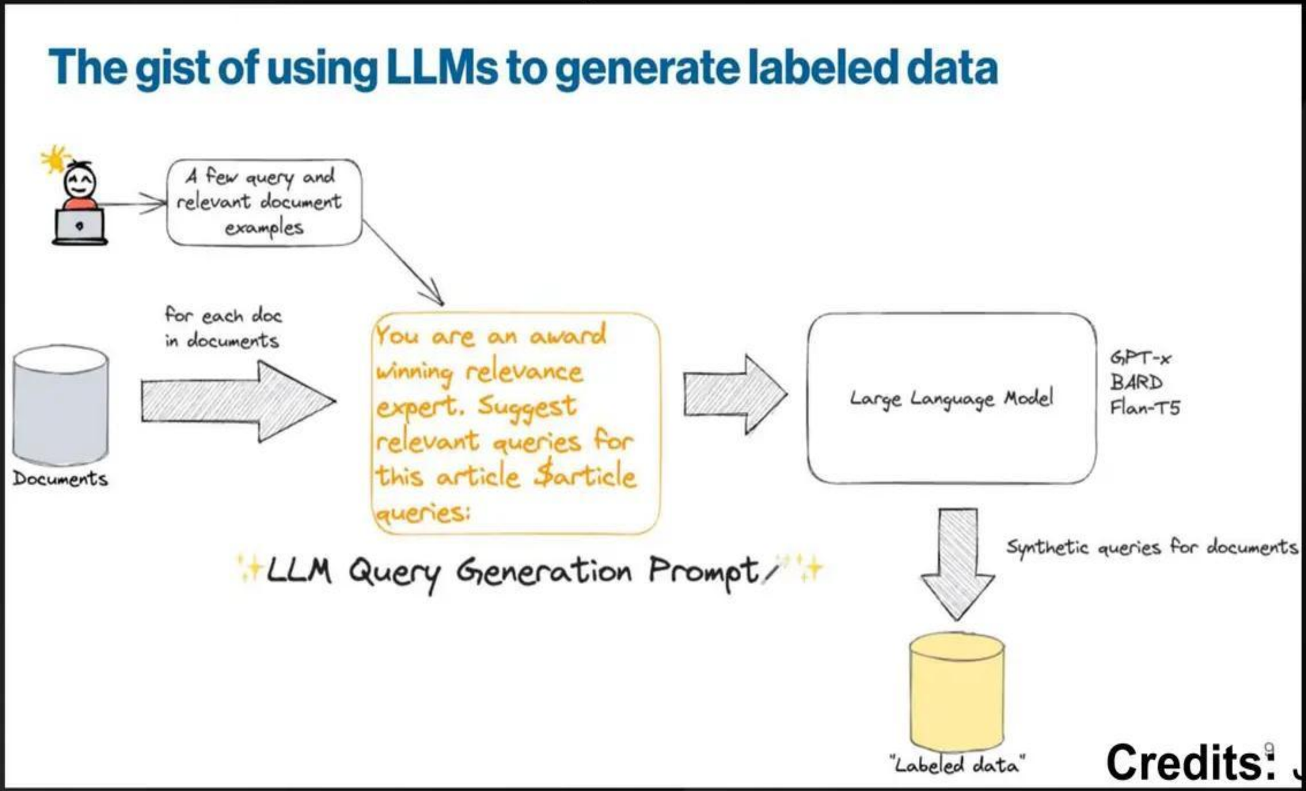


Fine-Tuning: Embeddings

Intuition: Embedding Representations are not optimized over your dataset

Solution: Generate a synthetic query dataset from raw text chunks using LLMs

Use this synthetic dataset to finetune an embedding model.



The gist of using LLMs to generate labeled data

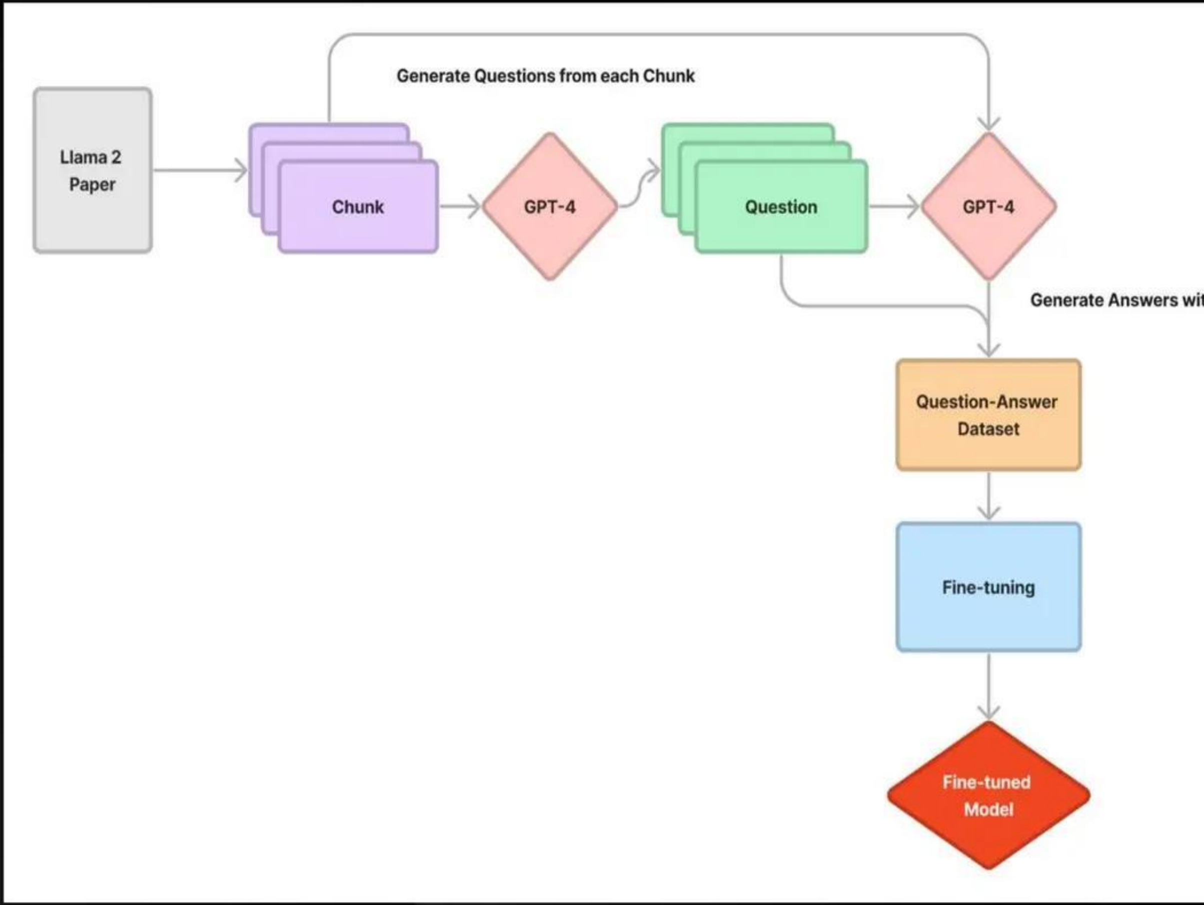
A few query and relevant document examples
for each doc in documents
Documents
You are an award winning relevance expert. Suggest relevant queries for this article $article queries:
Large Language Model
GPT-x
BARD
Flan-T5
Synthetic queries for documents
"Labeled data"
✨LLM Query Generation Prompt✨

Credits: Jo B

Fine-Tuning: LLMs

Intuition: Weaker LLMs are not bad at response synthesis, reasoning, structured outputs, etc.

Solution: Generate a synthetic dataset from raw chunks (e.g. using GPT-4). Help fix all of the above!



Resources

Production RAG

https://docs.llamaindex.ai/en/stable/end_to_end_tutorials/dev_practices/production_rag.html

Fine-tuning

https://docs.llamaindex.ai/en/stable/end_to_end_tutorials/finetuning.html

Thanks!