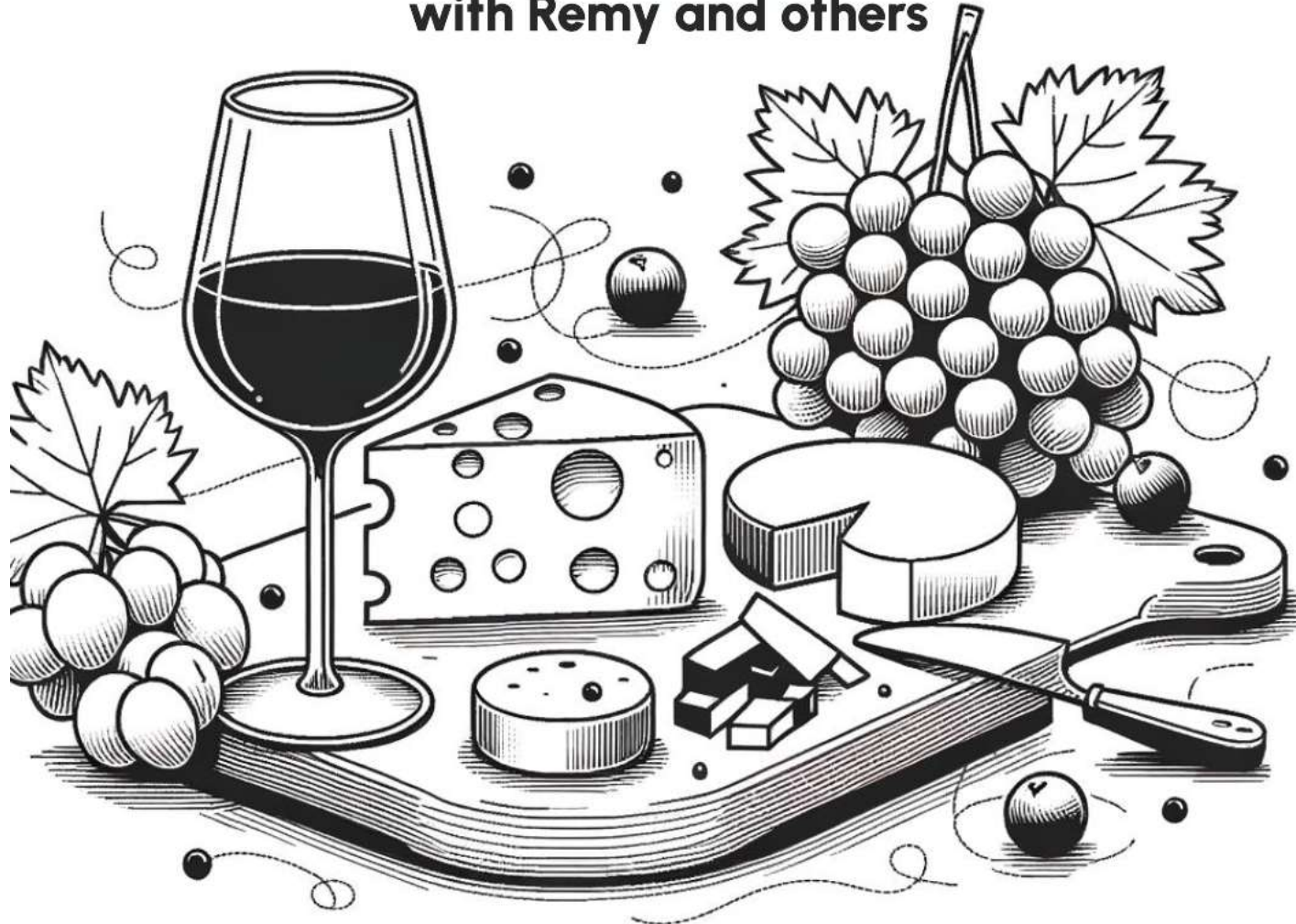


RAGATOUILLE: A CULINARY ADVENTURE

— Understanding RAG —
with Remy and others





Soham Dasgupta

Partner Solution Architect
Startups & Scaleups
App Innovation & AI
Microsoft




sohamda



iamsoham



dasguptasoham

- 
- Food without spices
 - Prep work – chopping and assembling
 - Following a recipe
 - Anton's approval
 - Mastering your masterpiece
 - Gusteau's secret spice blend
 - What should we cook tonight?

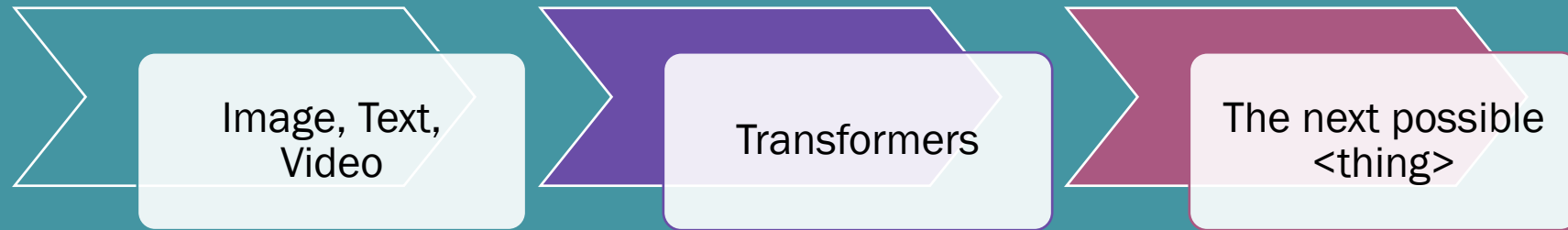
**Food without
(proper) spices**

```
graph LR; A[Is Good Friday a holiday?] --> B[LLM (GPT 4, 3.5, Llama, Claude etc.)]; B --> C[Yes, Good Friday is considered a holiday in many countries. In the UK, for example, BLA..BLA..BLA];
```

Is Good Friday a holiday?

LLM
(GPT 4, 3.5, Llama, Claude
etc.)

Yes, Good Friday is
considered a holiday in many
countries. In the UK, for
example, BLA..BLA..BLA



Whatever this transformer/model is
trained on is “what” it will give you back
as “thing”

Is Good Friday a holiday?

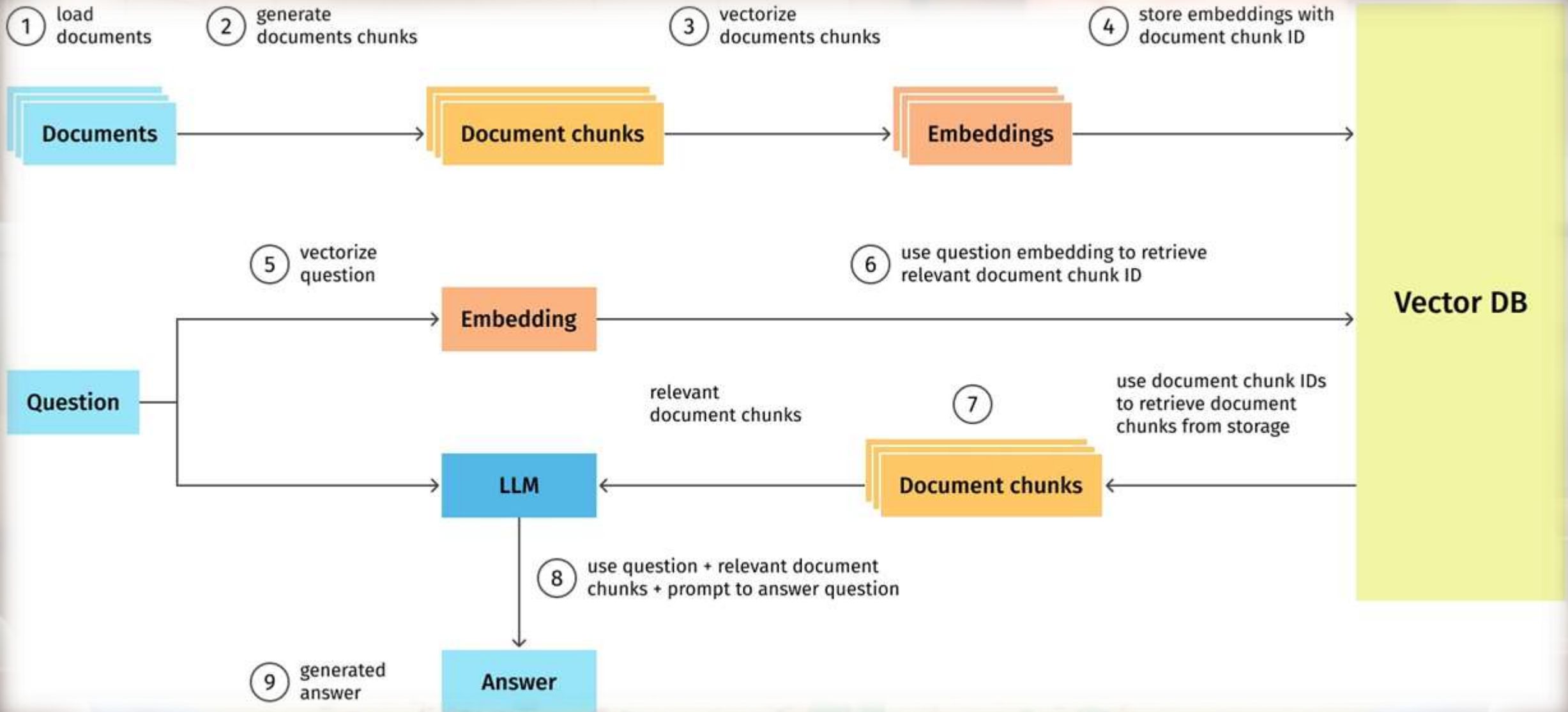
Search HR docs for holidays in 2024

Is Good Friday a holiday? + 2024 holiday list

LLM (GPT 4, 3.5, Llama, Claude etc.)

Yes, Good Friday is considered as a paid holiday for all employees

RAG process flow



Chopping and Assembling





Why Chunking?

[ChunkViz \(railway.app\)](https://railway.app/ChunkViz)



GPT-3.5 & GPT-4

GPT-3 (Legacy)

I am trying to explain why chunking is necessary to a group of very talented people. And tokens came up.

Clear

Show example

Tokens

23

Characters

104

I am trying to explain why chunking is necessary to a group of very talented people. And tokens came up.

Chunking Strategies – LlamaIndex & Langchain

- Code, Hierarchical, JSON, HTML, Markdown
- Text
- Token
- Semantic Splitter & Chunker
- Sentence Splitter
- Sentence Window

Geef het door als gegevens niet kloppen

Klopt uw adres niet? Geef uw juiste adres door aan de gemeente waar u woont. Wij nemen uw adres namelijk over uit de Basisregistratie Personen van de gemeente. Ontbreekt uw partner op het overzicht? Meld uw partner dan aan via uw persoonlijke omgeving Mijn Pensioencijfers.

Kloppen de gegevens van uw werk niet? Zaken als uw deeltijdfactor of uw voltijd salaris? Neem dan contact op met de afdeling HR of Personeelszaken van uw werkgever.

Risico nabestaandenpensioen

Het nabestaandenpensioen uit de risico nabestaandenpensioenregeling is meegenomen bij het onderdeel premieovereenkomst.

Bedragen als het mee- of tegenzit anders berekend

U ziet een inschatting van uw pensioen als we te maken krijgen met mee- of tegenvallers. We berekenen dit nu met de meest actuele gegevens. Hierdoor kunnen ze anders zijn dan vorig jaar. Na een interne controle is gebleken dat er voor uw premieovereenkomst (Module 2) verkeerde bedragen vermeld stonden op de UPO's van 2020, 2021 en 2022. In dit UPO 2023 worden in de premieovereenkomst (Module 2) de juiste actuele bedragen weergegeven.

Factor A: voor uw belastingaangifte van volgend jaar

Belastingaangifte doet u altijd achteraf. Hebt u daarbij de factor A nodig, dan is dat de factor A van het voorgaande jaar. De factor A op dit UPO hebt u dus nodig bij uw belastingaangifte van volgend jaar.

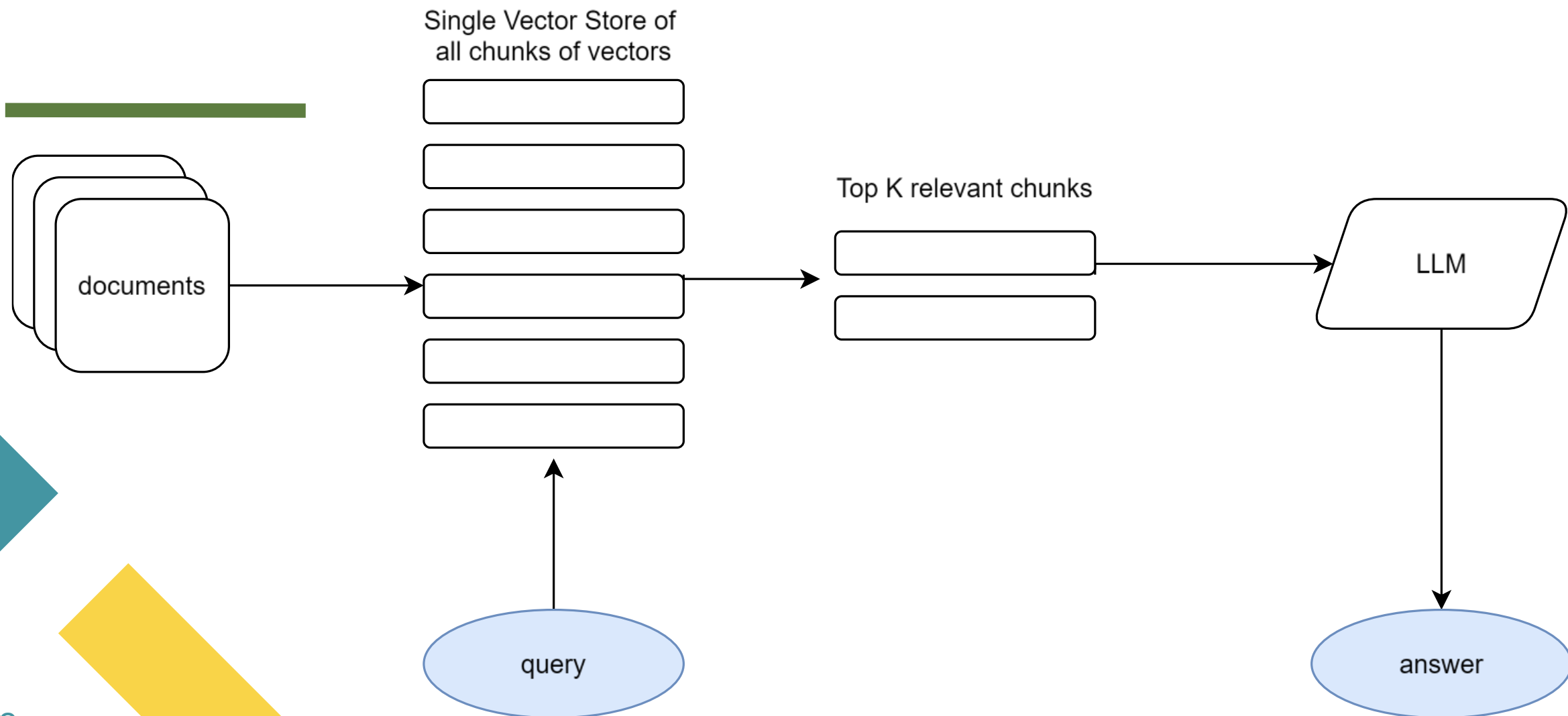
Uw privacy

Stichting Pensioenfonds Capgemini Nederland en zijn administrateur (AZL) verzamelen en verwerken persoonlijke gegevens van u. Dit doen wij om uw pensioen zo goed mogelijk te kunnen uitvoeren en administreren. Wij verzamelen alleen persoonsgegevens die wij daarvoor nodig hebben. Lees hierover meer in de privacyverklaring op www.pensioenfondscg.nl.

Chunk 1

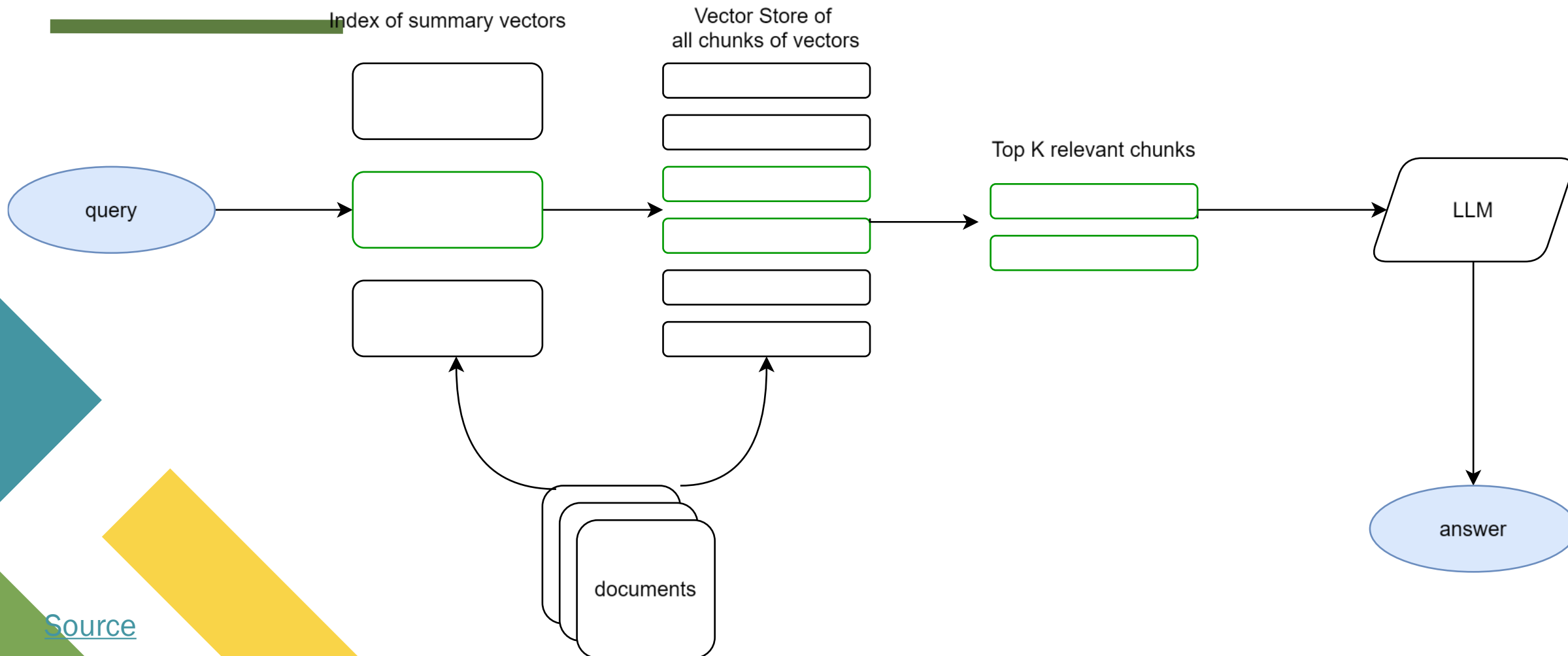
Chunk 2

Chunking (Retrieval) Strategies



Chunking (Retrieval) Strategies

Hierarchical Index Retrieval



Chunking (Retrieval) Strategies

Sentence Window Retrieval

What is preconfigured
in AKS automatic

Azure Kubernetes Service (AKS) Automatic offers an experience that makes the most common tasks on Kubernetes fast and frictionless, while preserving the flexibility, extensibility, and consistency of Kubernetes. Azure takes care of your cluster setup, including node management, scaling, security, and preconfigured settings that follow AKS well-architected recommendations. Automatic clusters dynamically allocate compute resources based on your specific workload requirements and are tuned for running production applications.

- Production ready by default:** Clusters are preconfigured for optimal production use, suitable for most applications. They offer fully managed node pools that automatically allocate, and scale resources based on your workload needs. Pods are bin packed efficiently, to maximize resource utilization.

- Built-in best practices and safeguards:** AKS Automatic clusters have a hardened default configuration, with many cluster, application, and networking security settings enabled by default. AKS automatically patches your nodes and cluster components while adhering to any planned maintenance schedules.

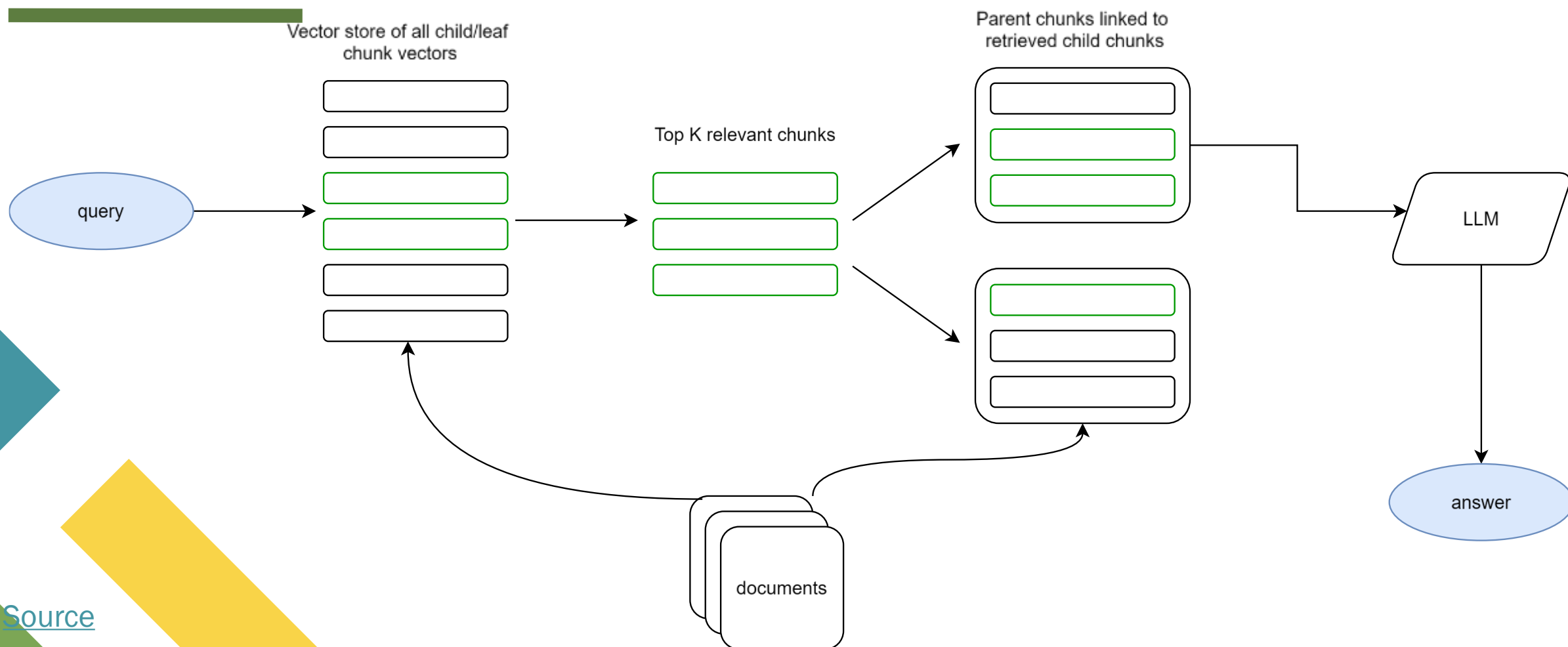
- Code to Kubernetes in minutes:** Go from a container image to a deployed application that adheres to best practices patterns within minutes, with access to the comprehensive capabilities of the Kubernetes API and its rich ecosystem.

Extended context to
LLM

LLM

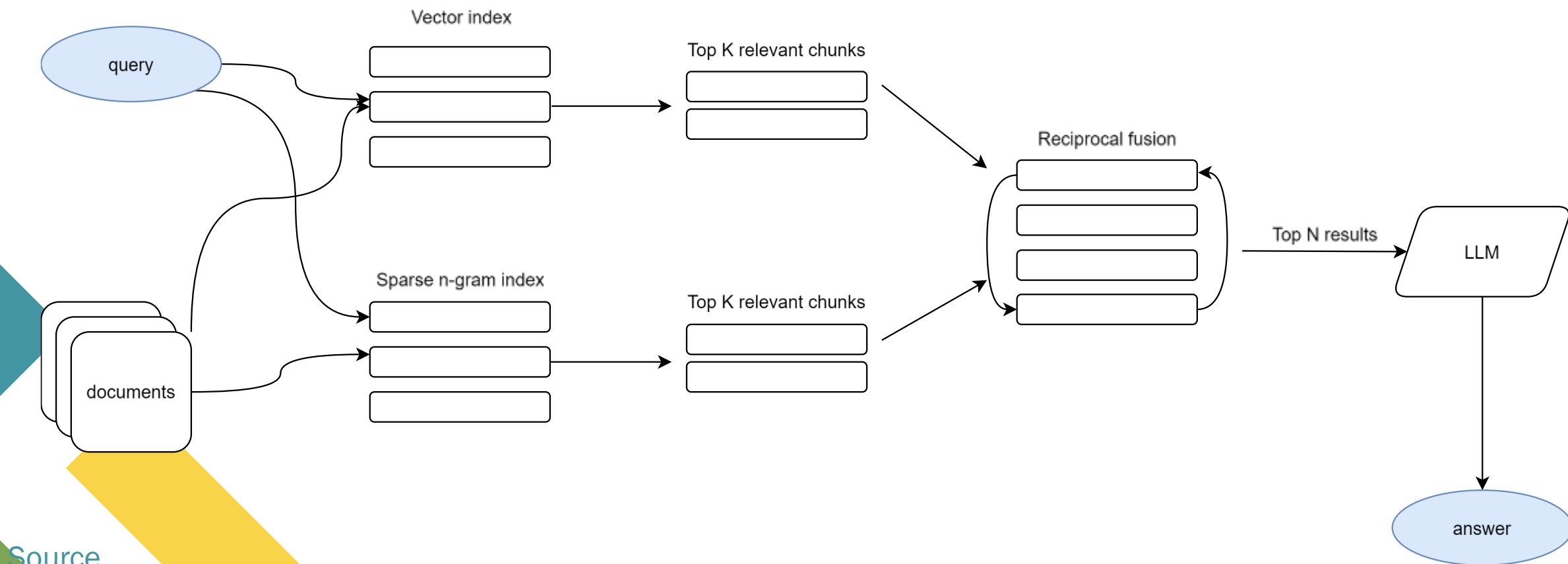
Chunking (Retrieval) Strategies

Parent-Child chunks Retrieval



Chunking (Retrieval) Strategies

Fusion Retrieval / Hybrid Search



Tokenizer/Chunking Regex

- *“Jina CEO Xiao Han shared a snippet of code on GitHub, which is the core participle implementation used in Jina tokenizer. This regular expression code fragment takes just over 50 lines, but it can efficiently process text content of various complexities for chunking. It was surprisingly powerful, parsing the entire Alice in Wonderland book in just two milliseconds, producing 1,204 text blocks.”*
- [Regex for chunking by using all semantic cues \(github.com\)](#)

Late chunking

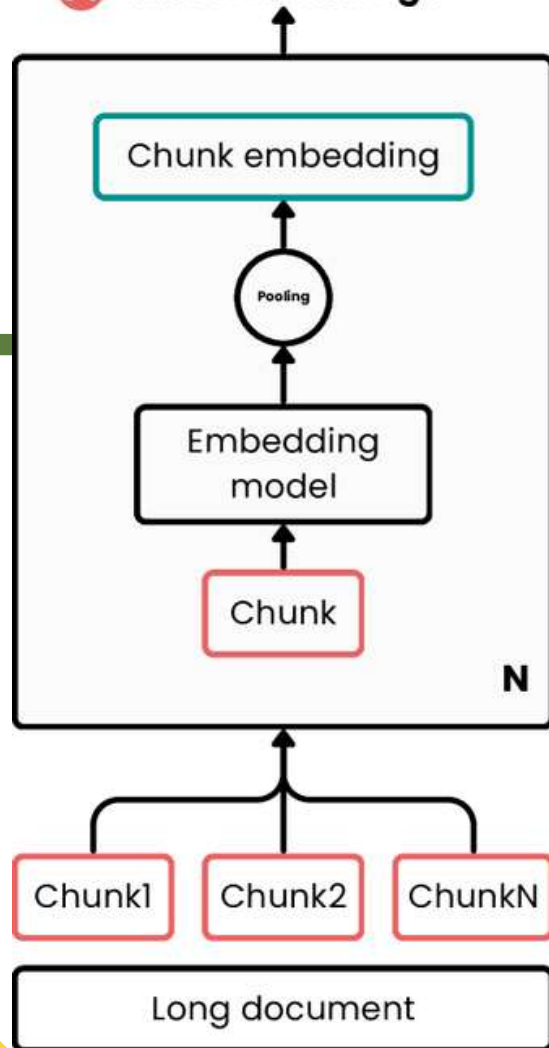
Berlin[a] is the capital and largest city of Germany, both by area and by population. [11] **I**ts more than 3.85 million inhabitants[12] make it the European Union's most populous city, as measured by population within city limits.[13] **T**he city is also one of the states of Germany, and is the third smallest state in the country in terms of area. Berlin is surrounded by the state of Brandenburg, and Brandenburg's capital Potsdam is nearby. The urban area of Berlin has a population of over 4.5 million and is therefore the most populous urban area in Germany.[5][14] The Berlin-Brandenburg capital region has around 6.2 million inhabitants and is Germany's second-largest metropolitan region after the Rhine-Ruhr region, and the sixth-biggest metropolitan region by GDP in the European Union.[15]

Berlin[a] is the capital and largest city of Germany, both by area and by population. [11].

Its more than 3.85 million inhabitants[12] make it the European Union's most populous city, as measured by population within city limits.[13].

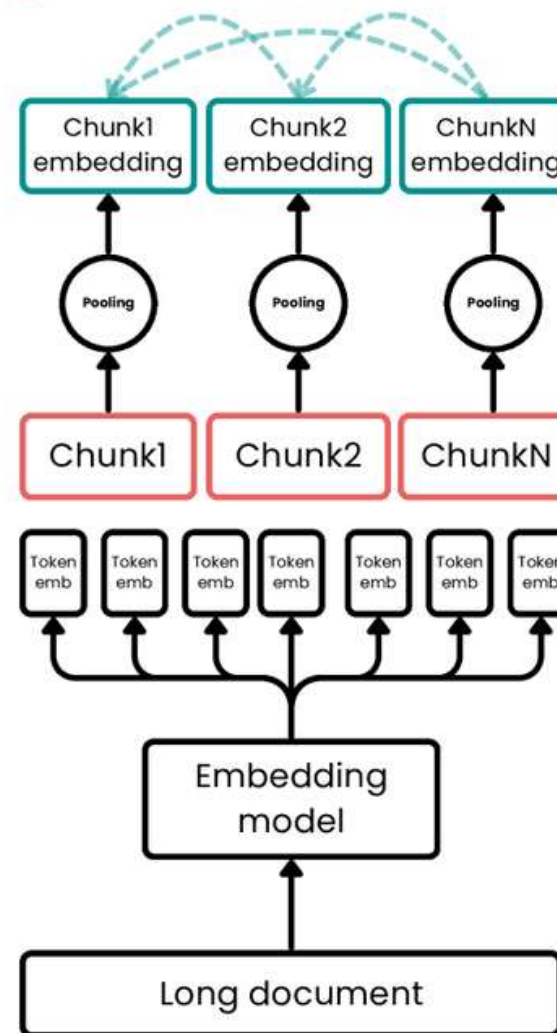
The city is also one of the states of Germany, and is the third smallest state in the country in terms of area.

❌ i.i.d. embeddings



NAIVE CHUNKING

✅ conditional embeddings



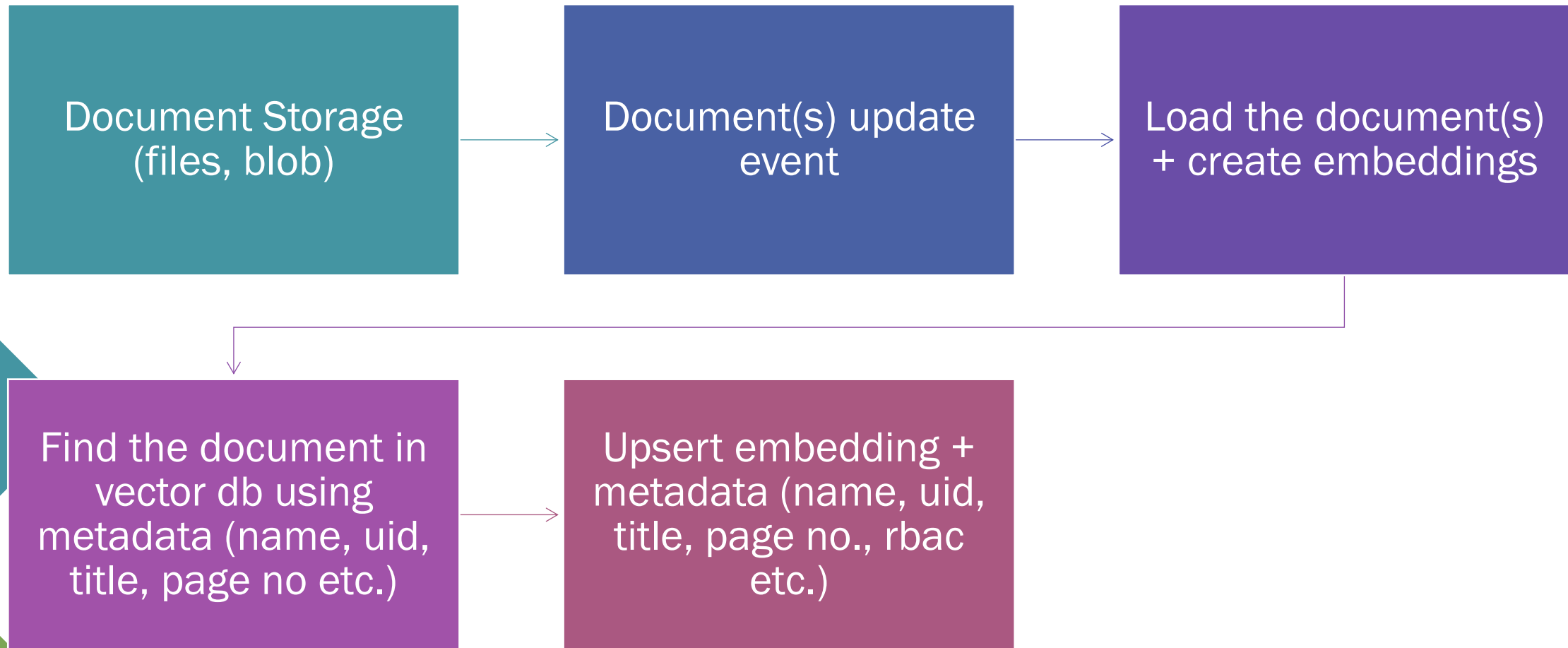
LATE CHUNKING

- [Late Chunking in Long-Context Embedding Models \(jina.ai\)](https://jina.ai)

Following a recipe

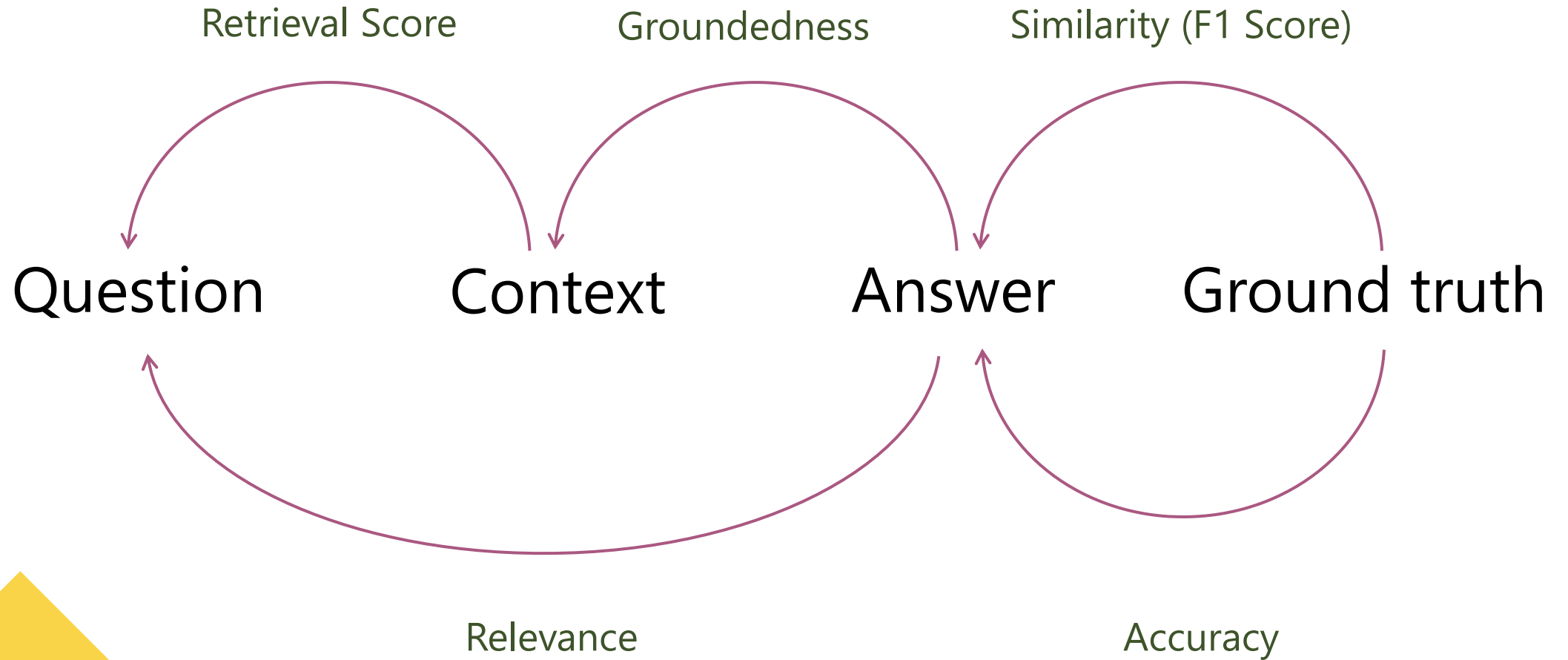


How does a workflow look like?



Anton's approval

RAG Evaluations



question	contexts	answer	ground_truths	context_precision	context_recall	faithfulness	answer_relevancy
0 What did the president say about Justice Breyer?	['Tonight, I'd like to honor someone who has dedicated his life to serve this country: Justice Breyer. And I did that 4 days ago, when I nominated Circuit Court of Appeals Judge Ketanji Brown Jackson. A former top litigator in private practice. A former federal public defender. And from a family of public servants.']	The president thanked Justice Breyer for his service and referred to him as an Army veteran, Constitutional scholar, and retiring Justice of the United States Supreme Court. The president also mentioned that he nominated Circuit Court of Appeals Judge Ketanji Brown Jackson, who will continue Justice Breyer's legacy of excellence.	['The president said that Justice Breyer has dedicated his life to serve the country and thanked him for his service.']	0.50	1.00	1.00	0.85
1 What did the president say about Intel's CEO?	['But that's just the beginning. Intel's CEO, Pat Gelsinger, who is here tonight, told us this is where Intel, the American company that helped build Silicon Valley, is going to go. For the past 40 years we were told that if we gave tax breaks to those at the very top,']	The president did not mention Intel's CEO specifically in the given context.	['The president said that Pat Gelsinger is ready to increase Intel's investment to \$100 billion.']	0.00	1.00	0.50	0.83
2 What did the president say about gun violence?	['And I ask Congress to pass proven measures to reduce gun violence. Pass universal background checks. As I said last year, especially to our younger transgender Americans, I will always have your back. Let's stop seeing each other as enemies, and start seeing each other for who we really are.']	The president called for Congress to pass measures to reduce gun violence, including universal background checks and a ban on assault weapons and high-capacity magazines. He also mentioned the need to repeal the liability shield for gun manufacturers.	['The president asked Congress to pass proven measures to reduce gun violence.']	1.00	1.00	1.00	0.91

- Context Relevancy – how much noise is there?
- Context Recall – do we have all the info needed retrieved?
- Faithfulness – what is the accuracy of the generated answer?
- Answer Relevancy – how relevant is the answer to the query?

Let me ask a LLM how is LLM doing !!

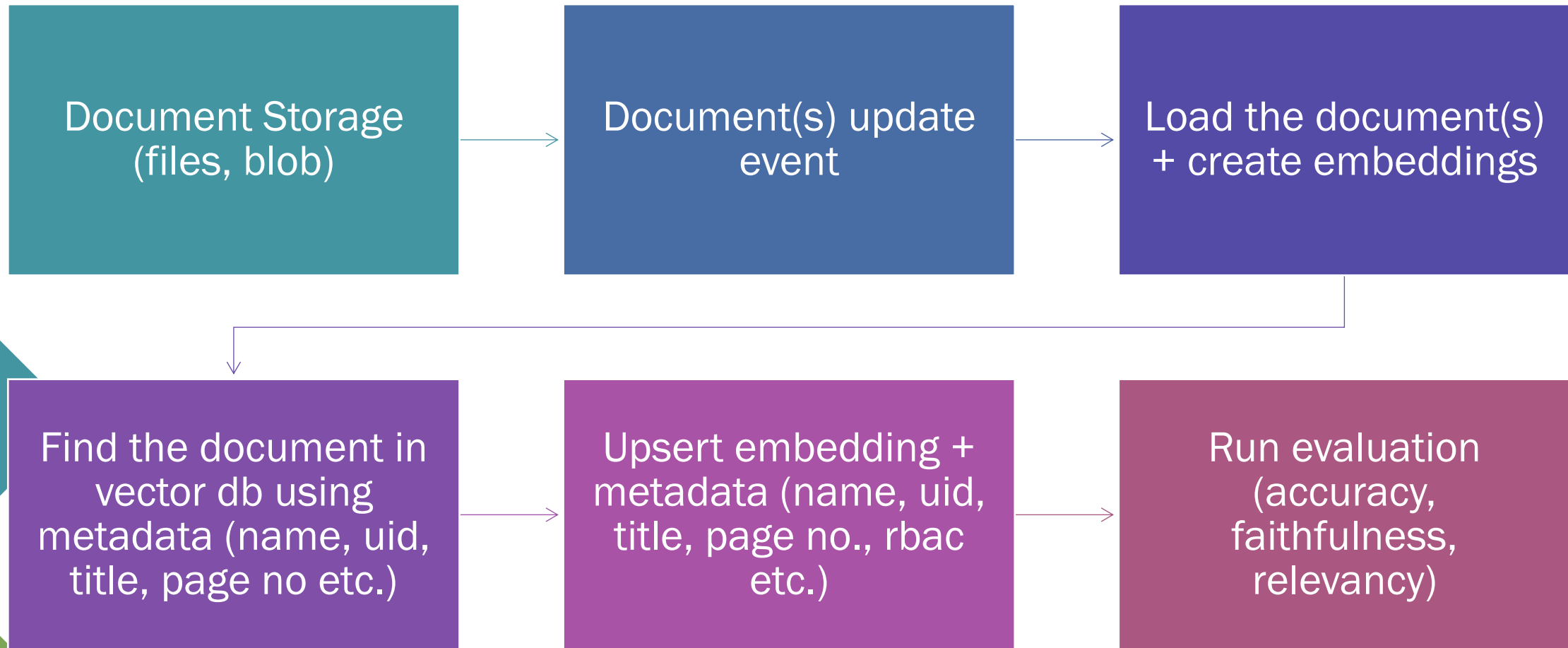


Query, Ground
truth, Answer,
Context

Based on the GT
and Q, rank the
ANS & CXT on a
scale of 0 to 1 how
accurate they are

LLM answers on a
scale of 0 to 1

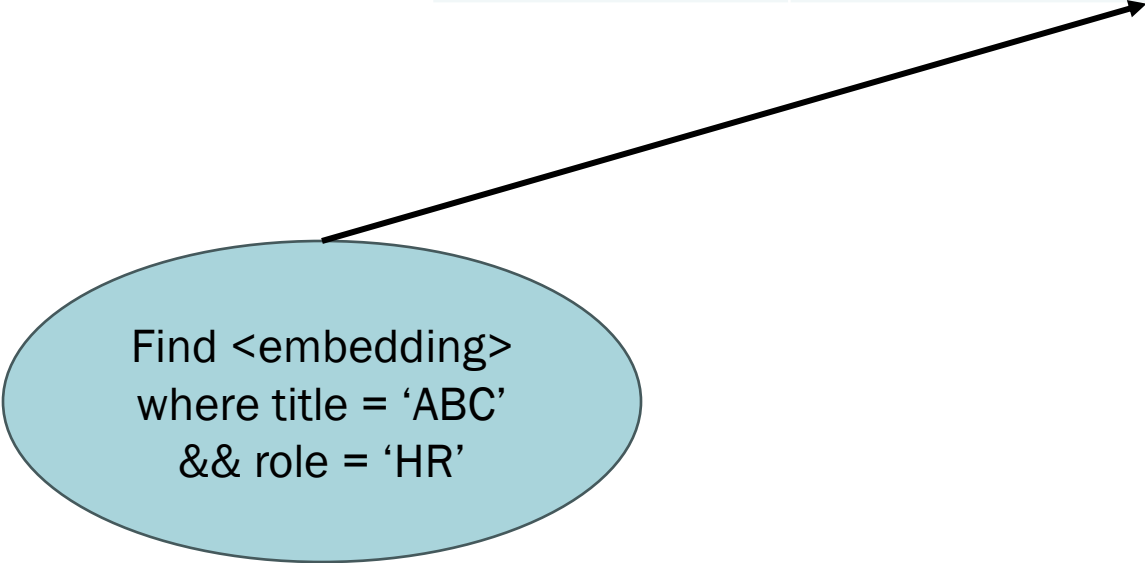
How does a workflow look like?



Mastering your masterpiece

Adding details to Retrieval (Metadata)

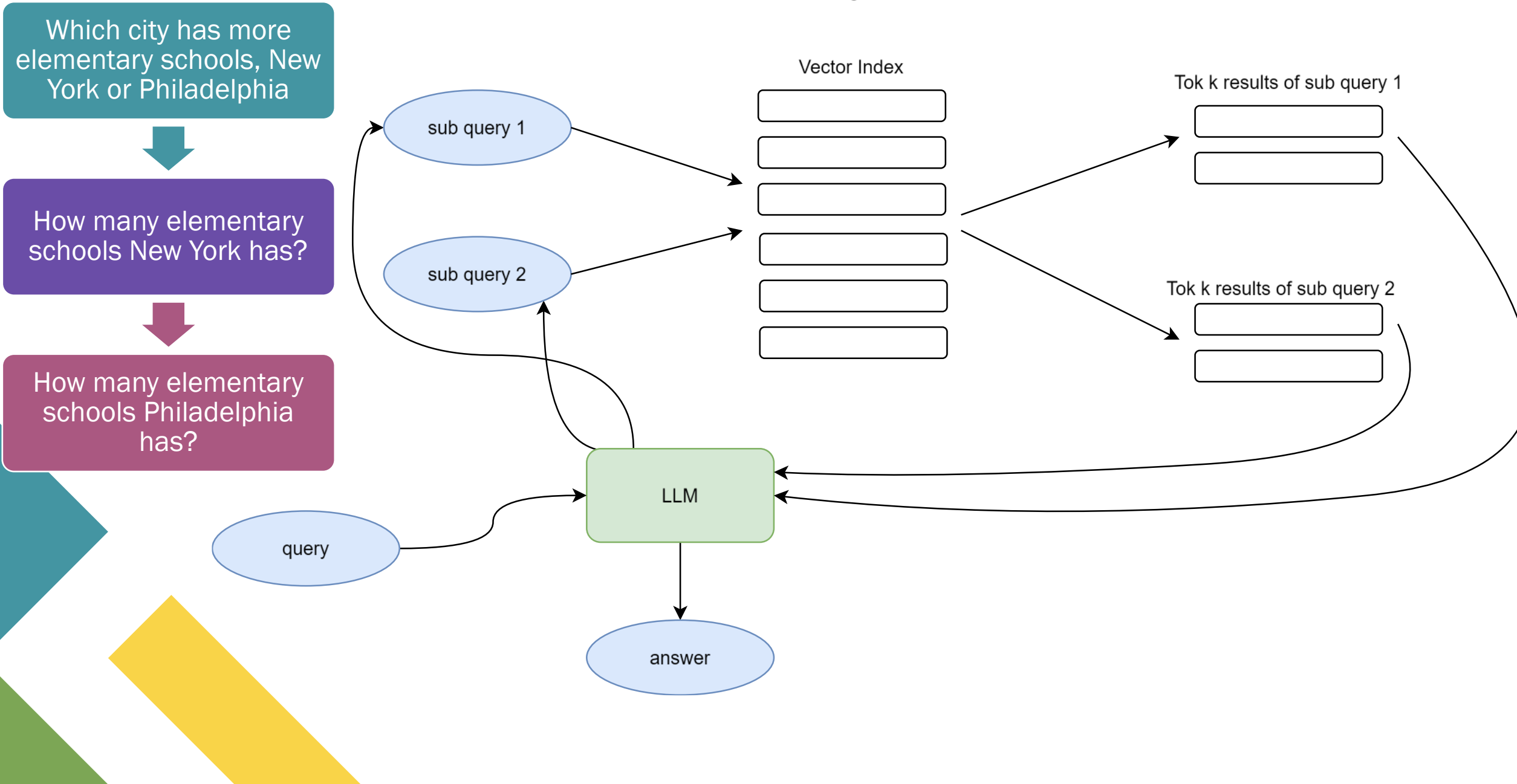
Title	Page	Role	ContentVector



Find <embedding>
where title = 'ABC'
&& role = 'HR'

Gusteau's secret spice blend

Query Transformation



The Agentic Property

A system's adaptability to achieve goals with limited direct supervision in a complex environment

Goal Complexity

How challenging are the goal(s) for a human?

How many goals does the system achieve?

Environment Complexity

How complex is the environment?

(Extent of cross-domain, multi-stakeholder, over time, external tools)

Adaptability

How much is rule-based behavior?

How much adaptability is possible for unexpected circumstances?

Independence

The level of human involvement needed

While various frameworks are still developing consensus on what an agent is, agentic as a property can be commonly understood.

Low Agenticness

- Email Filtering on Predefined Rules
- Template-based content generation
- Music recommendation based on past behavior
- Chatbots for basic customer service

High Agenticness

- Self-driving cars
- Realtime AI-powered cybersecurity Defense
- Adaptive Learning Platforms
- Autonomous financial trading bots

Agentic Reasoning Evolution

Ask a question on a topic?

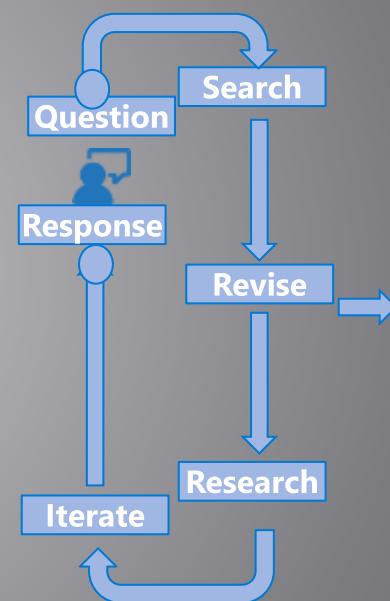
You get a response in one go.

Chat (non-agentic) workflow

Ask a question on a topic?

Do web search?
First draft response.
Need more research?
Do revision on response.
Iterate for more details?
Revise, act and respond.

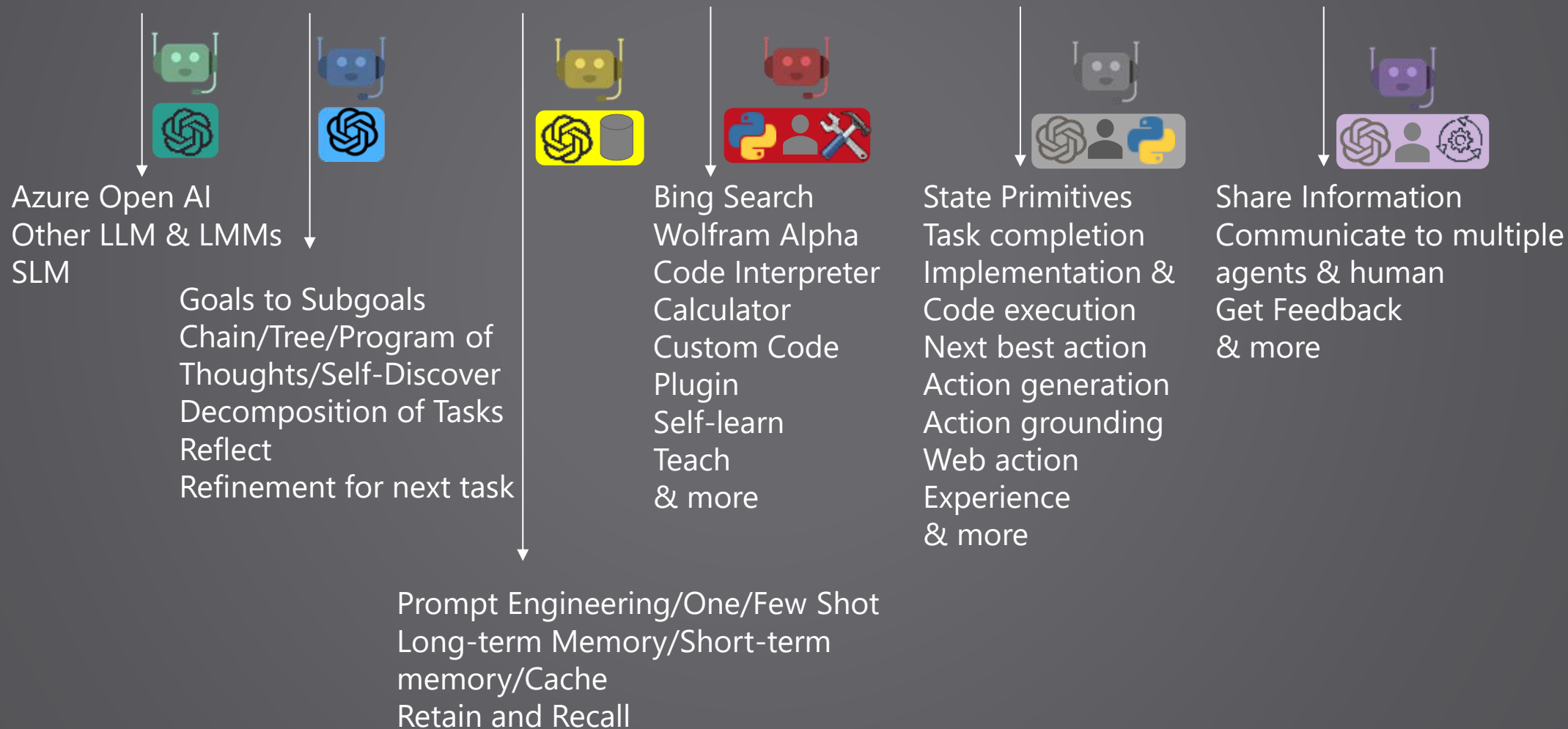
The need to combine many LLM tasks evolved into Agentic patterns and frameworks



Agentic Patterns

Microsoft Observe, Adapt, Plan, Reason, Reflect, Act & Communicate

Agent = LM + Planner + Memory + Tools / Skills + State/Action + Share/Communicate

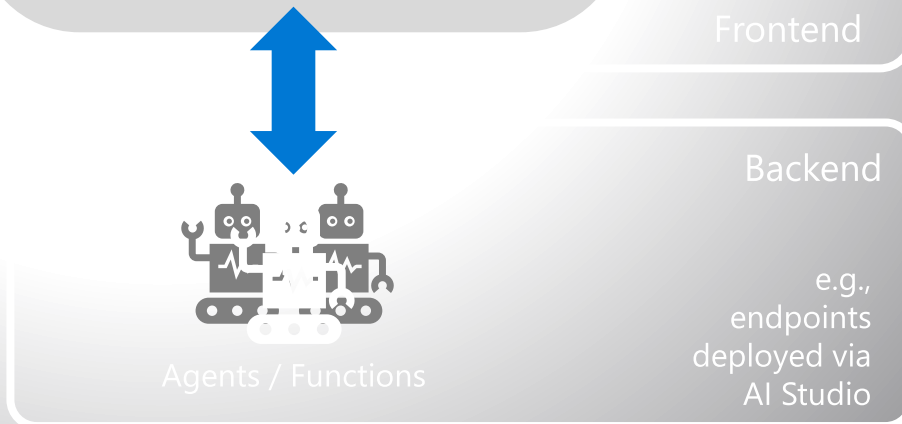
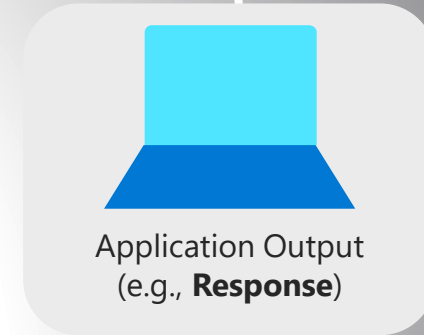
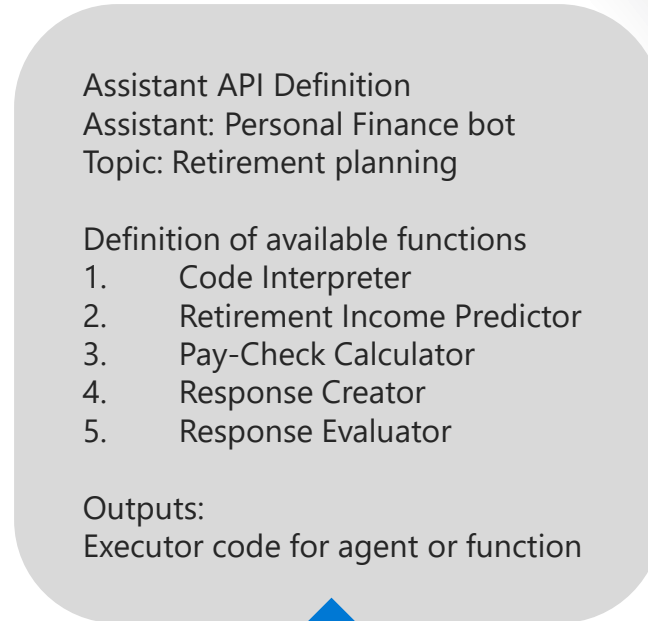
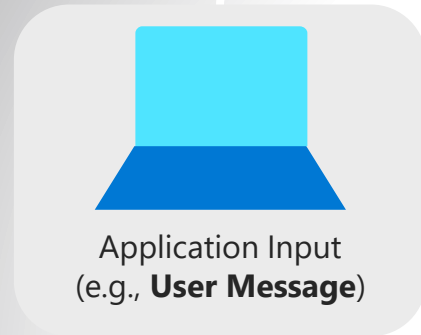


Assistants API

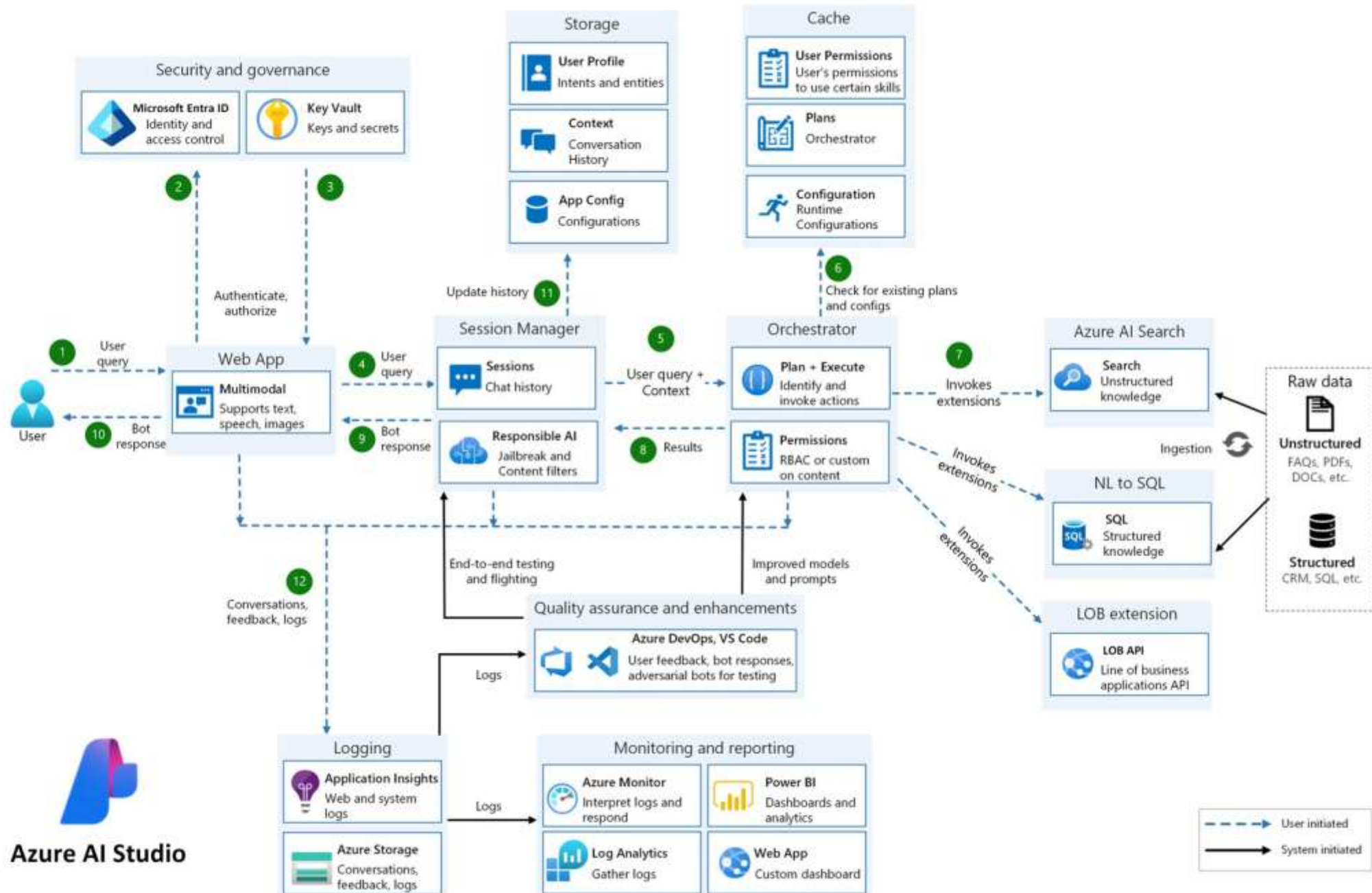
Agentic Orchestration Highlight

Build Sophisticated stateful AI assistants faster, perform complex computations and data analysis, safely act on the user's behalf, augment your copilot to access multiple APIs.

[Assistants API on Azure](#)



Agent architecture



Notable Agentic Frameworks

[AutoDev](#)

[AutoGen](#)

[MemGPT](#)

[GraphRAG](#)

[Promptflow](#)

[Semantic Kernel](#)

[Taskweaver](#)

[AgentVerse](#)

[Assistants API](#)

[AutoGPT + P](#)

[BabyAGI](#)

[BigAGI](#)

[CrewAI](#)

[DyLAN](#)

[Langchain](#)

[LATS](#)

[MetaGPT](#)

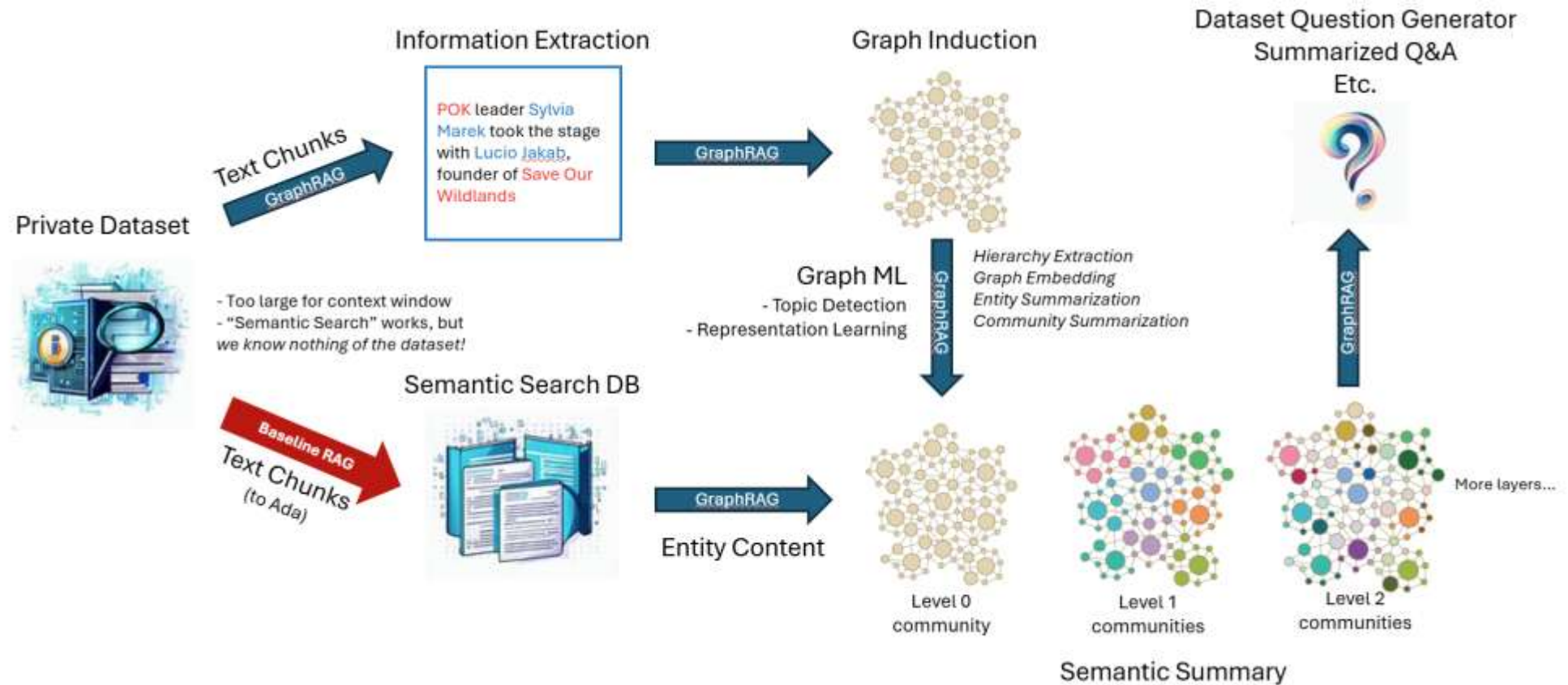
[ReAct](#)

[RAISE](#)

[Reflexion](#)

GraphRAG

Agentic Orchestration Highlight



[GraphRAG](#)

Perfect for use cases that require insights of the whole or large portions of the dataset rather than searching for specific elements within a dataset.

Graph RAG – The Longer Context Challenge

Insights required

- Global sense making over many documents
- Information discovery and analysis
- Questions are more abstract or thematic than underlying data can answer
- Critical reasoning is expected

Corpus of text data

- Is noisy
- Is mixed with mis and/or dis-information
- *Domain-specific topic identification in indexing
- *Entity rich

Users are already trained on responsible analytics

** most effective*


Comparing Approaches

	No Augmentation	Search RAG	Graph RAG
Context Awareness	None	Some	Deep
“Direct hit” search results	✗	✓	✓
Topically relevant connections	✗	✗	✓
Deep and sparse connections	✗	✗	✓
“Question behind the question”	✗	✗	partial
Hallucination Risk	High	Low	Very Low
Supports Hallucination Checking	✗	✓	✓
Provides source references	✗	✓	✓
Answers direct questions	✗	Usually	✓
Answers nuanced questions	✗	✗	✓
Explores adjacent topics and multiple perspectives	✗	✗	partial
Reasons across many documents	✗	~10	✓
Time to response	very fast (~5s)	fast (~5-10s)	medium (~10-30s)

**What should
we cook
tonight?**

Where to apply RAG

(some examples..)



Supply Chain

- Compliance Checks
- B2B sales
- Reporting
- Procurement

Retail

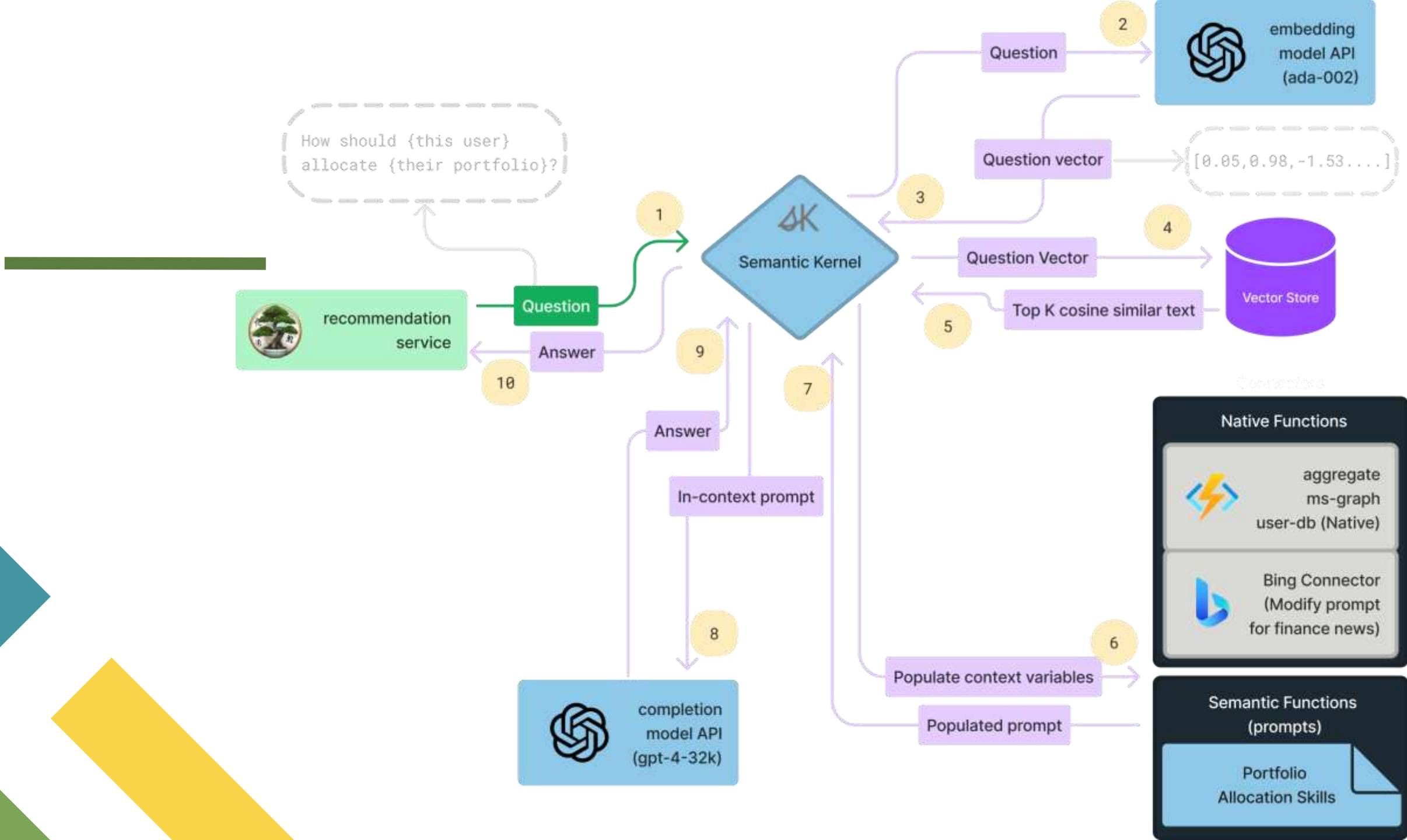
- Customer Support
- Product Recommender
- Feedback analytics
- Marketing

Finance & Insurance

- Customer Advisory
- Claim Processing
- Fin. Reporting
- Portfolio Mgmt

Banking

- Fraud detection
- Transaction Analytics
- Trend forecasting
- Document processing



How to learn more?

- [A first intro to Complex RAG \(Retrieval Augmented Generation\) | by Chia Jeng Yang | WhyHow.AI | Medium](#)
- [Advanced RAG Techniques: an Illustrated Overview | Towards AI](#)
- [Advanced RAG 01: Small-to-Big Retrieval | by Sophia Yang, Ph.D. | Towards Data Science](#)
- [Azure-Samples/graphrag-accelerator: One-click deploy of a Knowledge Graph powered RAG \(GraphRAG\) in Azure \(github.com\)](#)
- [Azure-Samples/azure-search-openai-demo-java: This repo is the Java version of Microsoft's sample app for ChatGPT + Enterprise data. \(github.com\)](#)

Thank you

