MICROSOFT IGNITE

NYC 2026 / COMMUNITY SUMMIT

MADE POSSIBLE BY OUR PARTNERS

THANK YOU, SPONSORS!

An extraordinary community deserves extraordinary supporters.

01

Morgan Stanley

02

Microsoft

03

SOHO DRAGON

Soho Dragon

Unlocking Corporate Intelligence:

Mastering RAG in Microsoft Foundry

Exploring Azure & Microsoft Foundry AI technologies to build scalable intelligent systems

By Sadiq Ahmed

About Me !

Sadiq Ahmed

20+ years in IT

Sr. Consultant

Love Academia & Microsoft Technologies !

Adjacent faculty of computer science dept

Trainer & Speaker

Retrieval-Augmented Generation (RAG)



RAG : What

Retrieval-Augmented Generation (RAG)

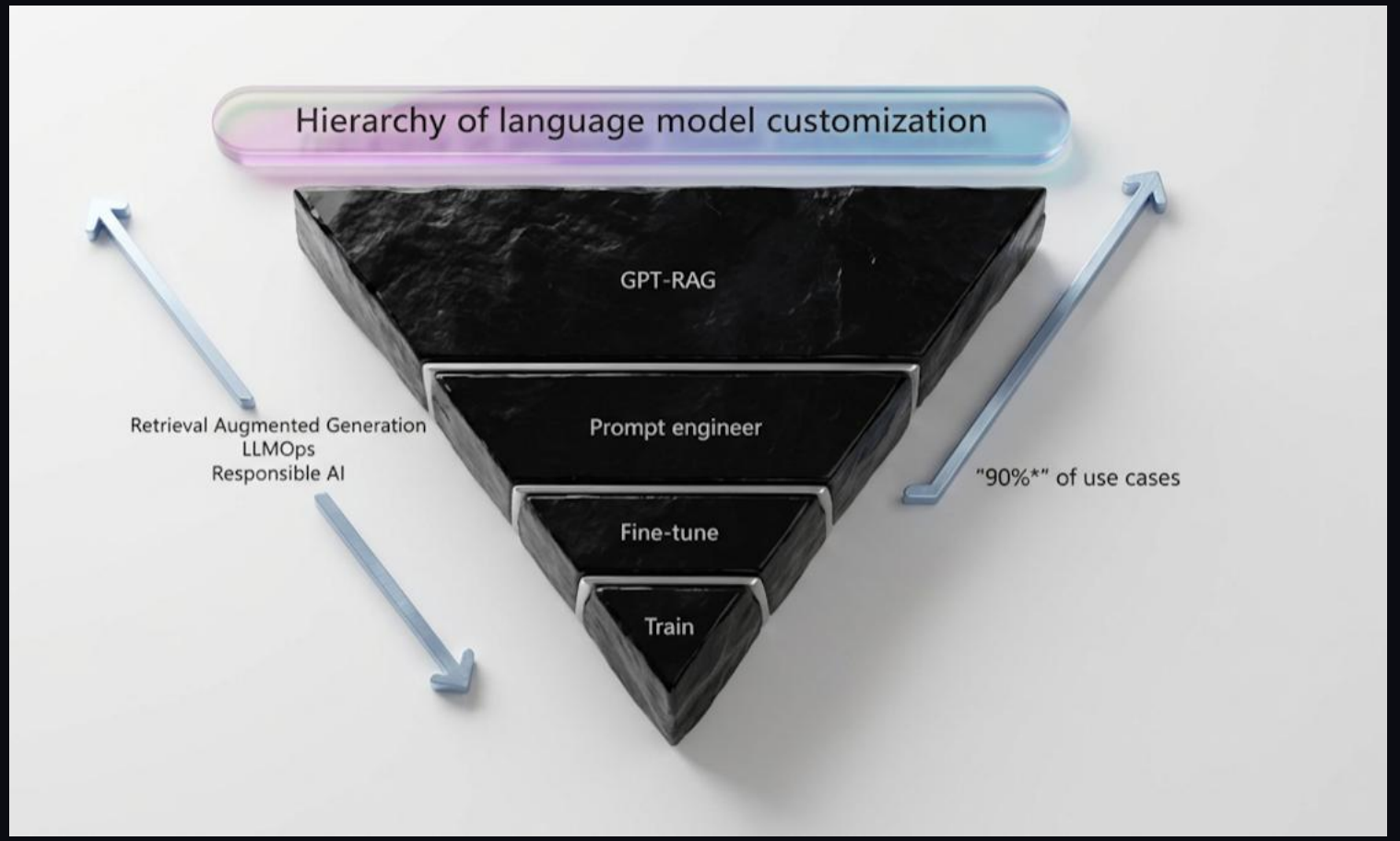
An Architectural Pattern

Paper published “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks” [2005.11401] Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks (arxiv.org)

https://arxiv.org/abs/2005.11401

RAG : Why

Generate answers based on your data



RAG Flow

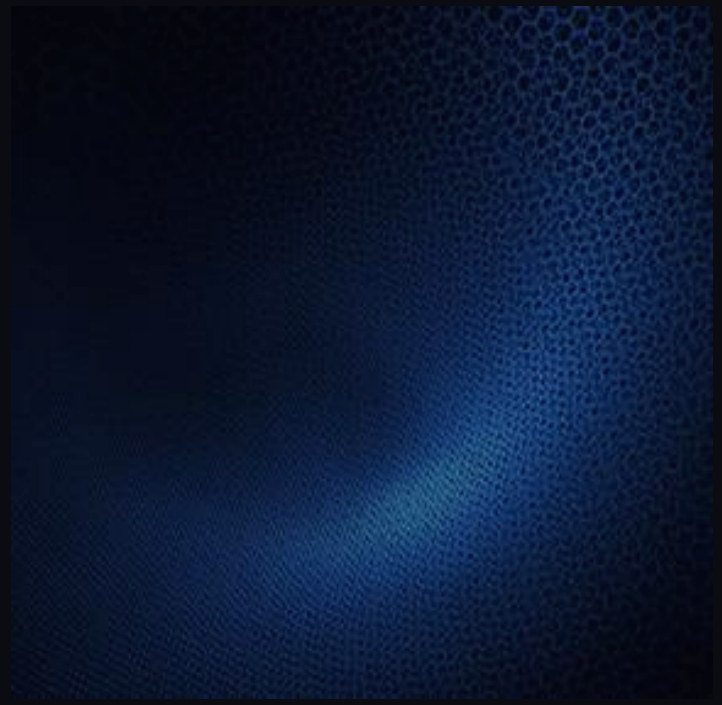
Retrieve (Data): Based on user question, application queries an index or data store to find relevant content.

Augment (Prompt): Application than combines the user's question and the retrieved content (grounding data) into a prompt.

Generate (Response): Next model receives the augmented prompt and generates a response grounded in the retrieved relevant content along with accurate citations.



RAG : Key Services on Azure Platform

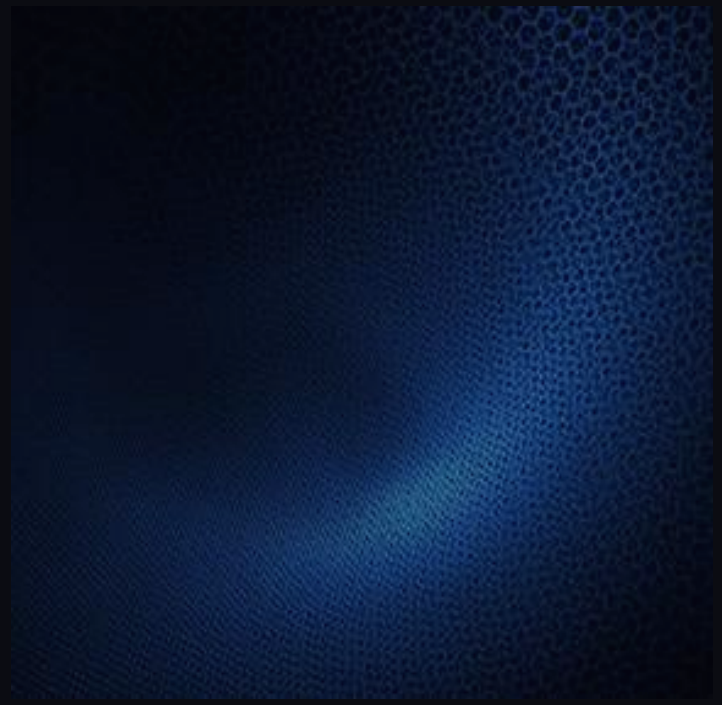


Azure AI Search (Foundry IQ)	Knowledge Source Knowledge Base
Azure Storage	Storage (Container , Cosmos DB etc)
Microsoft Foundry	Embedding LLM

RAG Types

RAG Type	Description	Enterprise Usage
Naive / Basic / Standard RAG	Retrieve documents → send to LLM → generate answer	Good for prototypes
Vector RAG	Uses embeddings and vector search	Very common
Hybrid RAG	Combines keyword + vector search	Most common in production
Graph RAG	Uses knowledge graphs and relationships	Used for complex enterprise knowledge
Multimodal RAG	Retrieves text, images, audio, video	Growing rapidly
Agentic RAG	AI agent plans multiple retrieval steps and reasoning	Fastest growing enterprise pattern

Top 3 RAG Types



1# Hybrid RAG

- Better accuracy than pure vector search
- Easier governance
- Faster implementation
- Supported natively by Azure AI Search

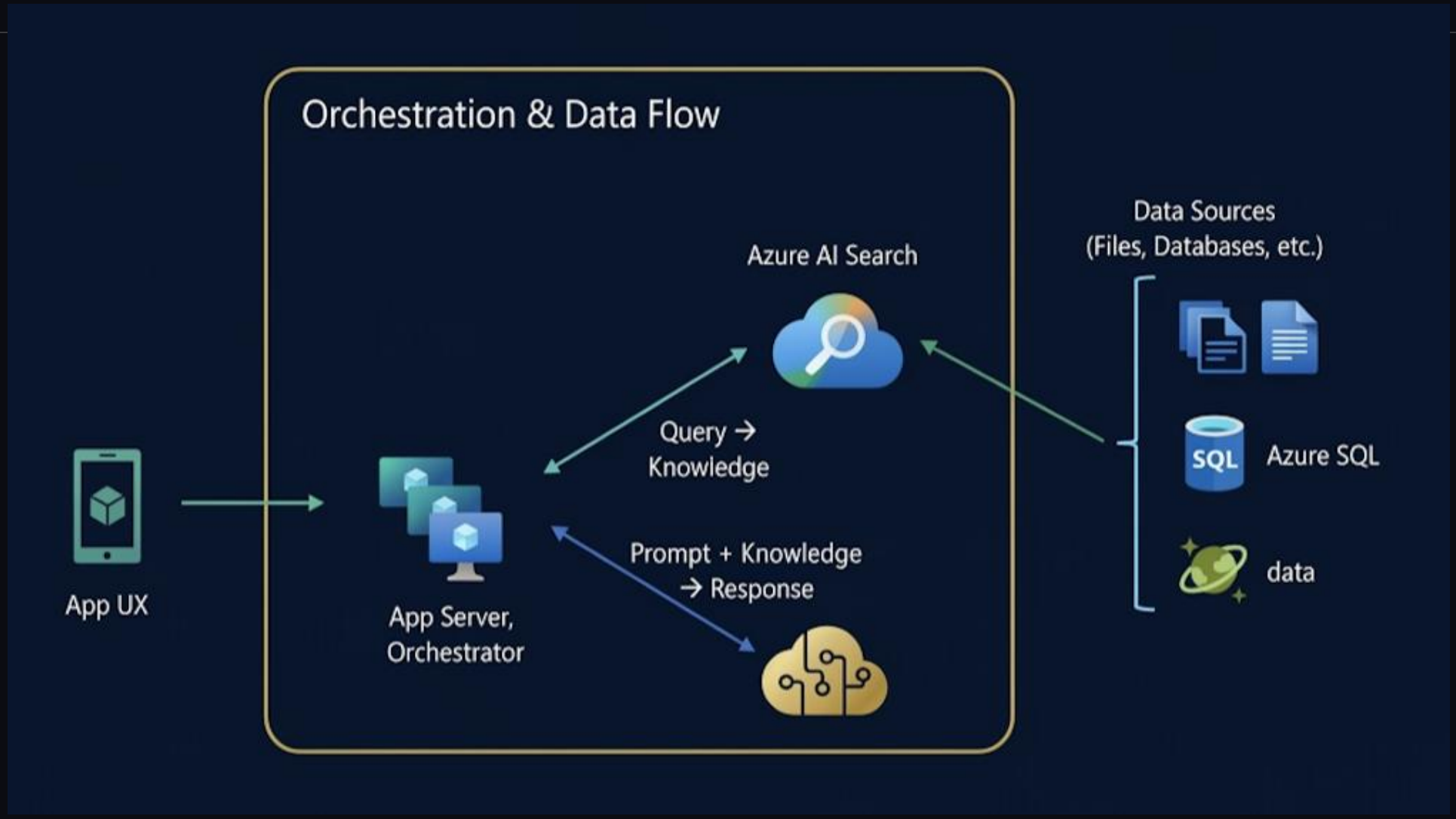
2# Agentic RAG

- Search product database
- Search support tickets
- Search recall notices
- Compare findings
- Generate answer

3# Graph RAG

- Insurance
- Healthcare
- Banking
- Supply Chain

RAG : Big Picture



Microsoft AI Foundry

Microsoft Foundry Platform

The unified platform for building, deploying, and managing AI solutions on Azure

Model Deployment

Deploy and configure generative AI models in the Foundry portal

Agent Builder

Create, test, and iterate on single-agent AI solutions

Foundry SDK

Build lightweight chat clients and integrate AI into apps

Speech Tools

Azure Speech integration for voice-enabled AI experiences

Vision & Generation

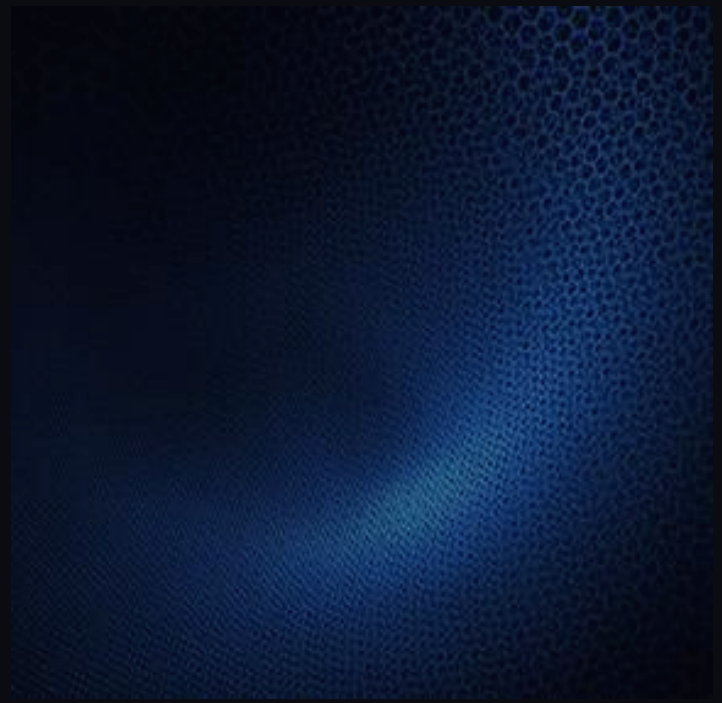
Multimodal models for visual content interpretation and creation

Content Understanding

Extract structured data from documents, images, audio, video

Azure AI Search

Information Retrieval System

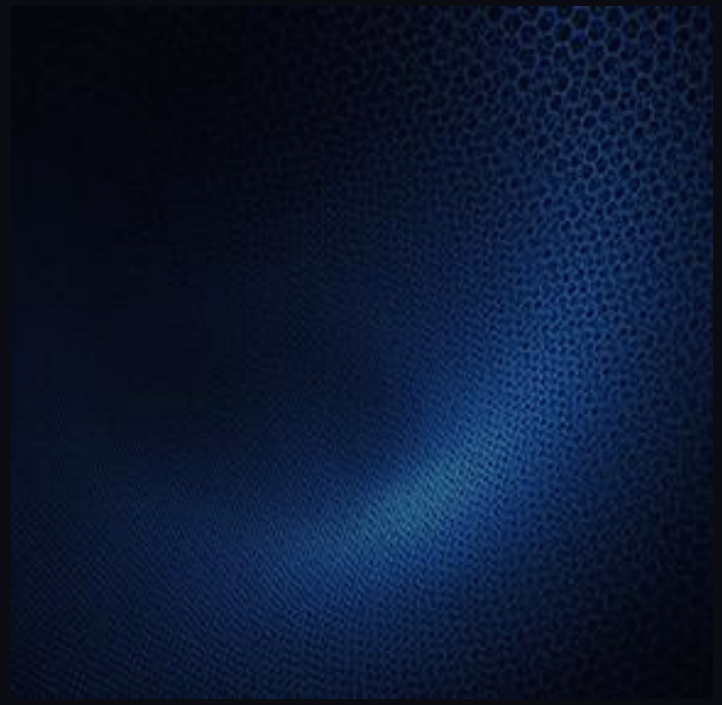


The decision about which information retrieval system to use is critical because it determines the inputs to the LLM.

Characteristics

Indexing strategies	Query capabilities and relevance tuning	Payload (query response)	Security, global reach, and reliability	Integration with embedding models for indexing

Azure AI Search : Key Points



1. Provide the Searchable Index

2. Implement Query Logic

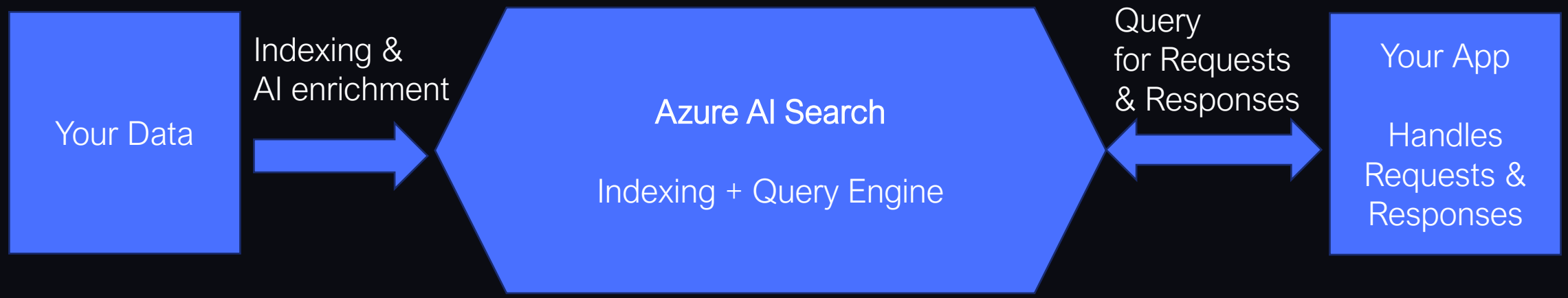
3. Generate the Payload

4. Include Vectors or Nonvector Content

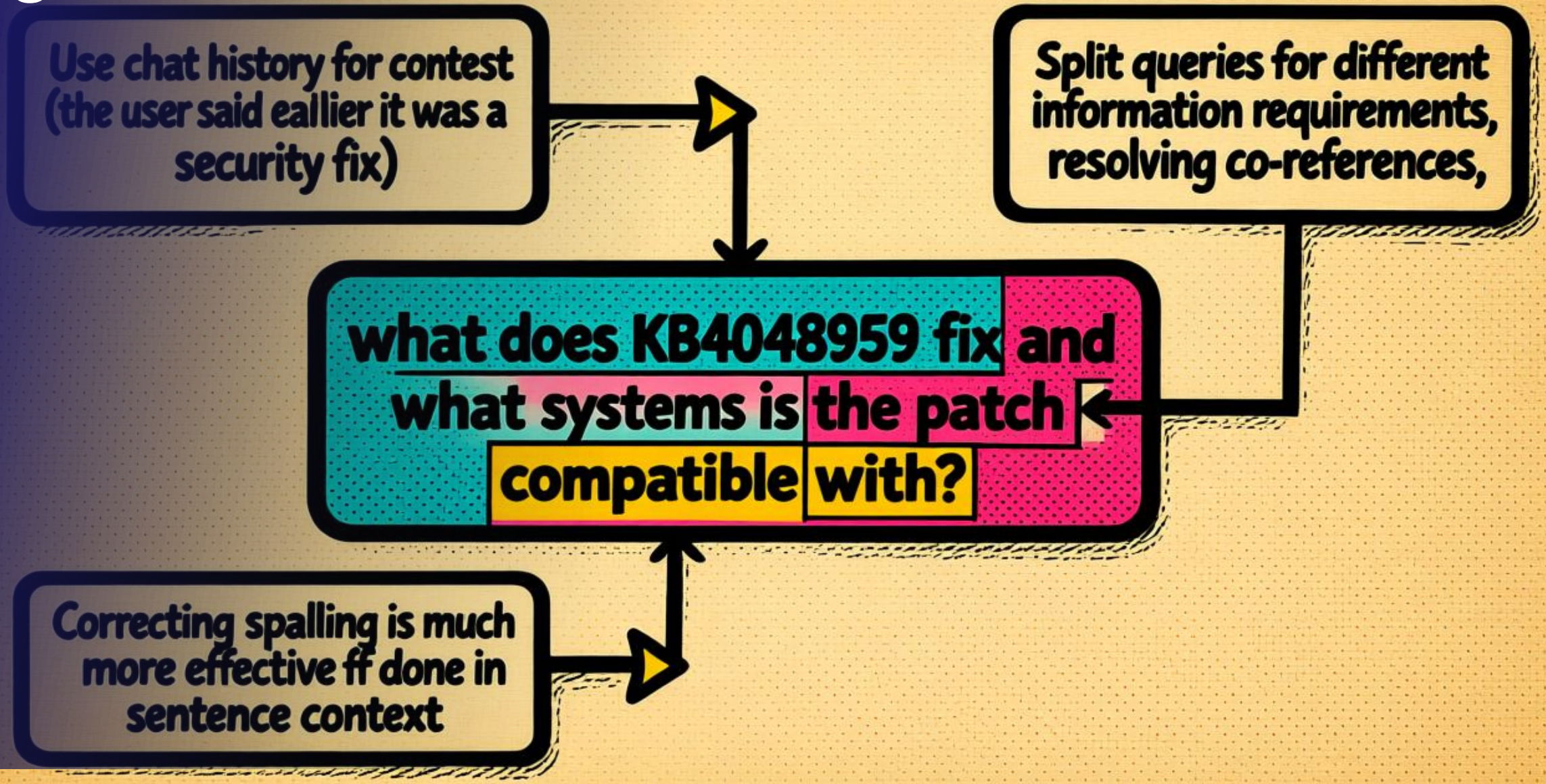
Hybrid Retrieval:
Combines keyword and vector search to provide the best results, even for semantically similar queries1.

Vector and Non-Vector Indexing:
Supports both types of indexing and queries, enhancing the search capabilities1.

Azure AI Search – Behind the scene

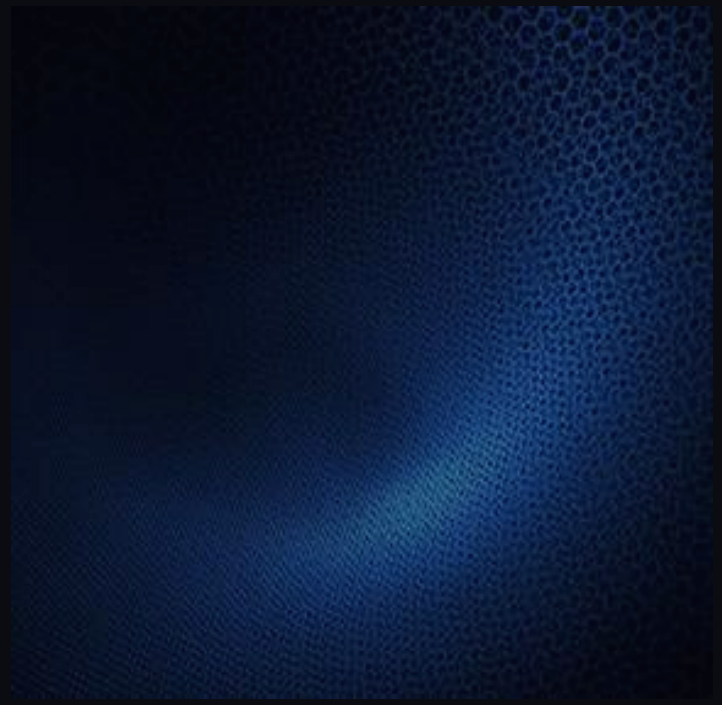


Agentic Retrieval



Embedding

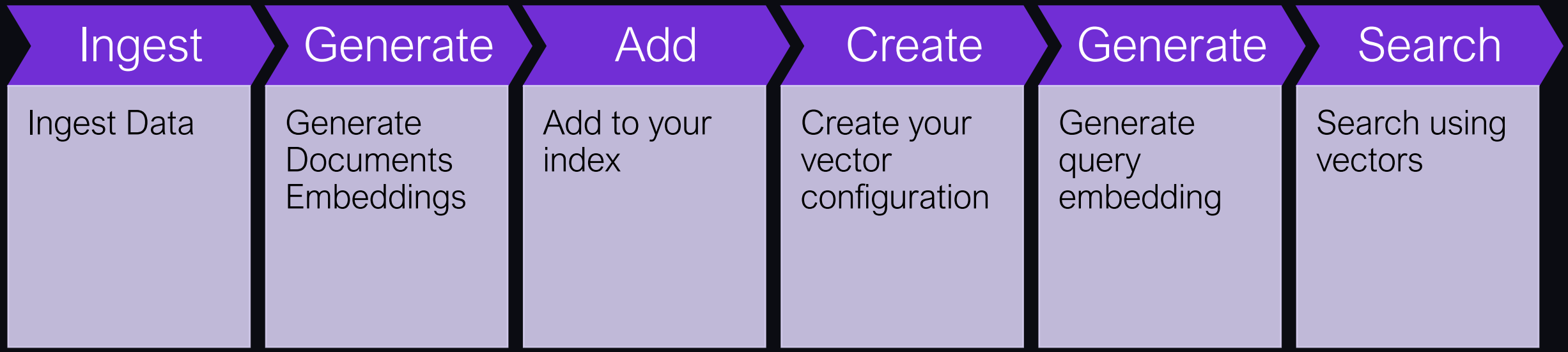
- Convert Data into Vector Representation
- Each embedding is a Vector of floating-point number.
- OpenAI Model for embedding is “Text-embedding-x-y”
- Vectors are a universal representation of data



Learn more :
https://www.datacamp.com/blog/vector-embedding

https://openai.com/index/introducing-text-and-code-embeddings/

AI enrichment in Azure AI Search : Ingest – Enrich & Index – Explore



MICROSOFT IGNITE

NYC 2026 / COMMUNITY SUMMIT

MADE POSSIBLE BY OUR PARTNERS

THANK YOU, SPONSORS!

An extraordinary community deserves extraordinary supporters.

01

Morgan Stanley

02

Microsoft

03

SOHO DRAGON

Soho Dragon