

AI Tech Accelerator Switzerland 2026

Workshop #2
16 February 2026



Swiss AI startups ready to scale



Agenda

09:45 – 12:00	Database aspects for implementing AI: Strategies & Methods	Sameer Ankalgi, Sr. Solution Engineer
12:00 – 13:00	Lunch Break	
13:00 – 14:30	Hands-on Lab – RAG: Agent First Retrieval Patterns	Juan Manuel Servera Bondroit, Sr. Partner Solution Architect
14:30 – 14:45	Afternoon Break	
14:45 – 16:00	Hands-on Lab continued – RAG: Agent First Retrieval Patterns	Juan Manuel Servera Bondroit, Sr. Partner Solution Architect



Database aspects for implementing AI Strategies & Methods

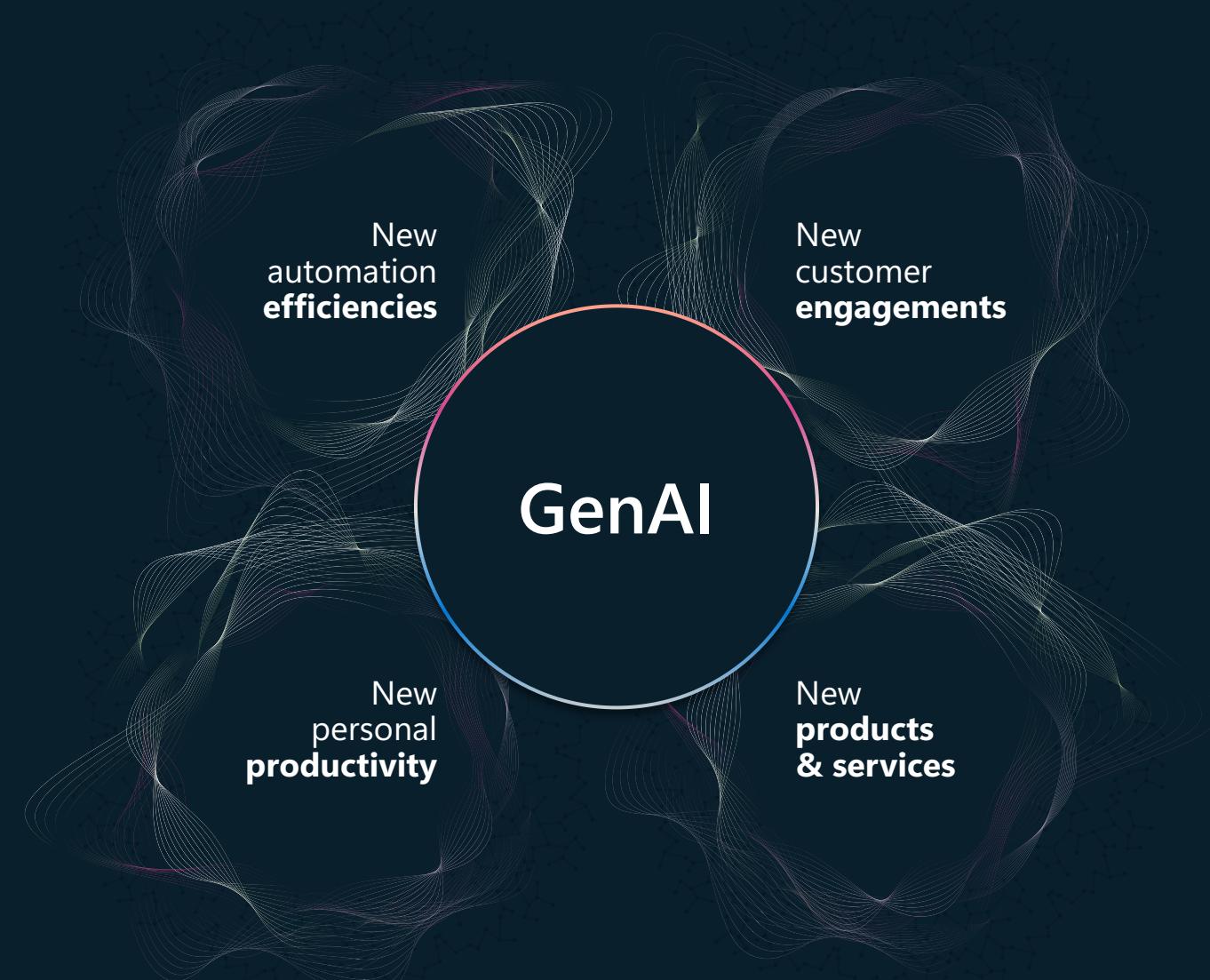
Sameer Ankalgi, Sr. Solution Engineer



FAILURES?

6	3	6	1	4	9	5	2	9	3	9	5	1	9	3	7	8	4	3	0	1	9	3	9	1	3	0	7	2	6	0	2	7	5	0	4	
7	5	2	6	8	7	4	3	9	9	2	2	7	4	9	6	7	4	7	7	9	6	7	1	5	3	6	8	2	2	5	3	0	8	4	9	
7	7	9	4	6	1	2	7	2	9	4	3	1	2	5	5	8	4	7	0	4	0	8	0	4	7	6	1	0	3	0	0	7	6	8	8	
7	8	9	4	3	6	4	1	4	0	4	1	7	5	6	8	1	6	8	0	0	4	0	7	5	6	1	4	8	3	6	8	5	6	3	1	
1	9	5	2	0	2	1	3	1	5	9	6	9	1	0	4	6	7	9	9	4	5	1	7	7	9	0	2	0	1	9	2	4	1	7	7	
3	9	1	1	4	3	7	0	7	0	7	6	8	0	6	4	4	6	4	0	4	0	2	3	1	5	1	1	5	4	4	8	3	0	7	8	
4	3	0	1	9	4	5	0	7	2	8	9	0	6	2	7	9	7	5	2	7	2	0	4	4	0	5	1	7	1	4	2	8	6	4	5	
4	1	6	8	6	0	0	3	8	5	0	3	4	2	3	1	4	4	4	7	7	5	3	2	1	2	1	6	4	6	8	6	9	0	6	3	
8	5	7	7	8	0	8	6	0	8	4	9	0	4	0	3	3	9	0	5	9	7	8	3	9	2	8	5	7	8	3	4	6	8	7	1	
1	8	9	2	3	8	4	6	5	5	2	5	8	0	0	0	0	0	0	0	0	0	4	1	0	3	0	9	0	8	4	2	6	1	2	5	
8	7	3	9	7	1	2	7	8	9	7	9	8	3	2	9	9	8	0	3	6	0	5	2	8	9	0	0	8	9	6	9	0	5	3	3	
4	3	5	3	9	5	1	3	2	0	4	7	2	5	1	9	8	7	8	2	4	4	3	4	0	4	0	9	1	4	6	2	6	2	1	3	
4	0	7	0	0	0	2	8	5	8	3	8	1	6	8	7	0	0	5	2	4	7	9	4	2	1	7	4	9	1	0	3	8	6	9	1	3
8	9	0	9	1	2	5	0	0	3	6	3	7	7	5	9	7	8	5	6	5	3	3	4	3	4	6	4	3	6	6	7	9	9	2	9	
0	8	5	3	1	8	3	2	3	2	5	9	2	4	4	2	0	7	6	4	8	8	6	2	4	9	4	5	6	9	5	3	9	2	9	3	
3	6	5	1	2	8	5	6	3	6	5	2	5	6	0	4	0	3	3	3	3	3	0	3	4	4	4	4	2	2	0	6	4	8	7	7	
2	0	7	8	1	0	7	2	1	5	7	8	0	8	9	3	3	7	1	3	4	9	7	7	9	6	1	1	2	5	6	4	2	9	2	4	
9	0	4	2	1	1	4	0	3	0	2	7	3	0	2	2	0	2	0	2	5	0	6	2	1	0	0	1	5	0	1	2	9	7	6		

Generative AI is the greatest **force multiplier** we've ever had at our fingertips



Today's
databases
need to be
ready to power
AI apps

To power intelligent apps,
databases need...



To scale to any size



Built-in AI capabilities



Ability to deliver real-time
experiences



Enterprise-grade security

No matter where you are in your cloud journey,
you can modernize and build new intelligent apps with Azure

1

Connected
smart
products

2

Transaction
processing
at scale

3

Real-time
fraud
detection

4

Service &
support
bot

5

Information
& product
discovery

6

Personalize
&
recommend

Build your own copilots

Bringing domain knowledge to LLMs



Retrieval augmentation

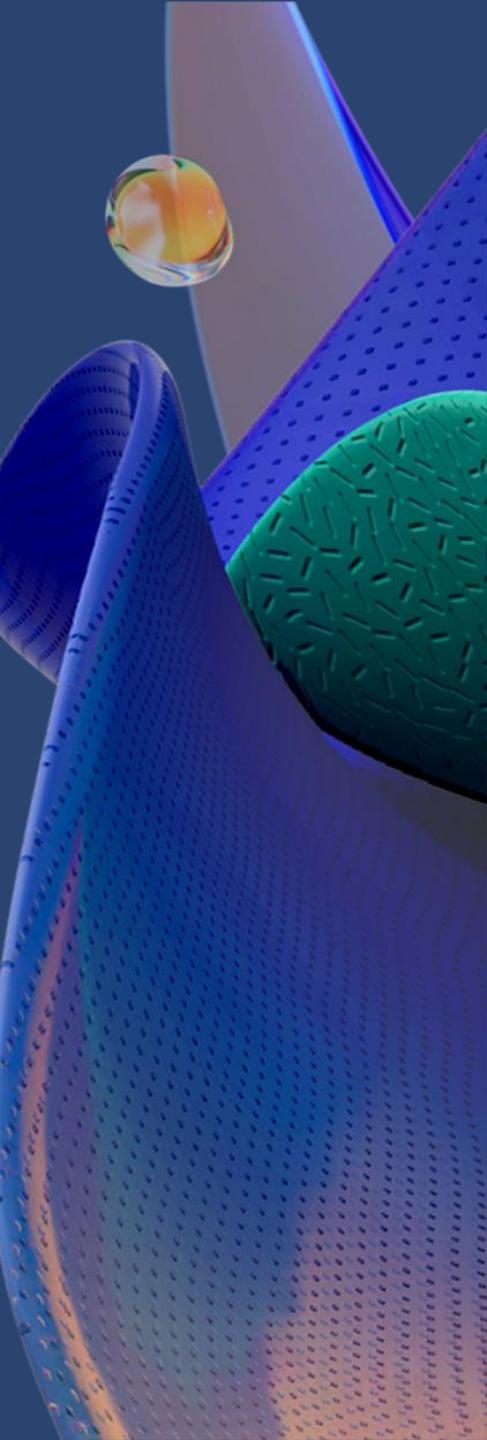
Learn new facts



**Fine
tuning**

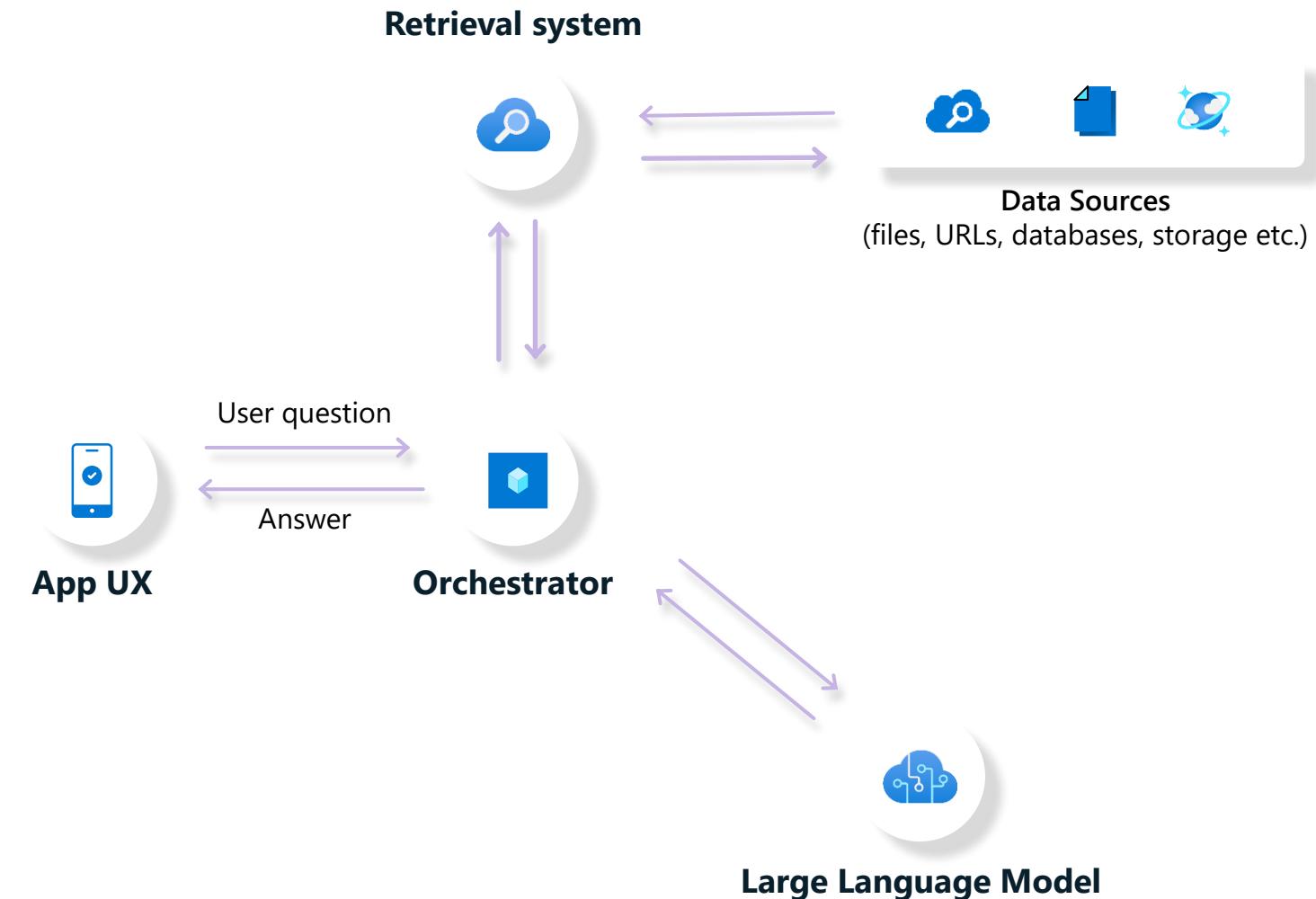
Learn new skills

RAG refresher



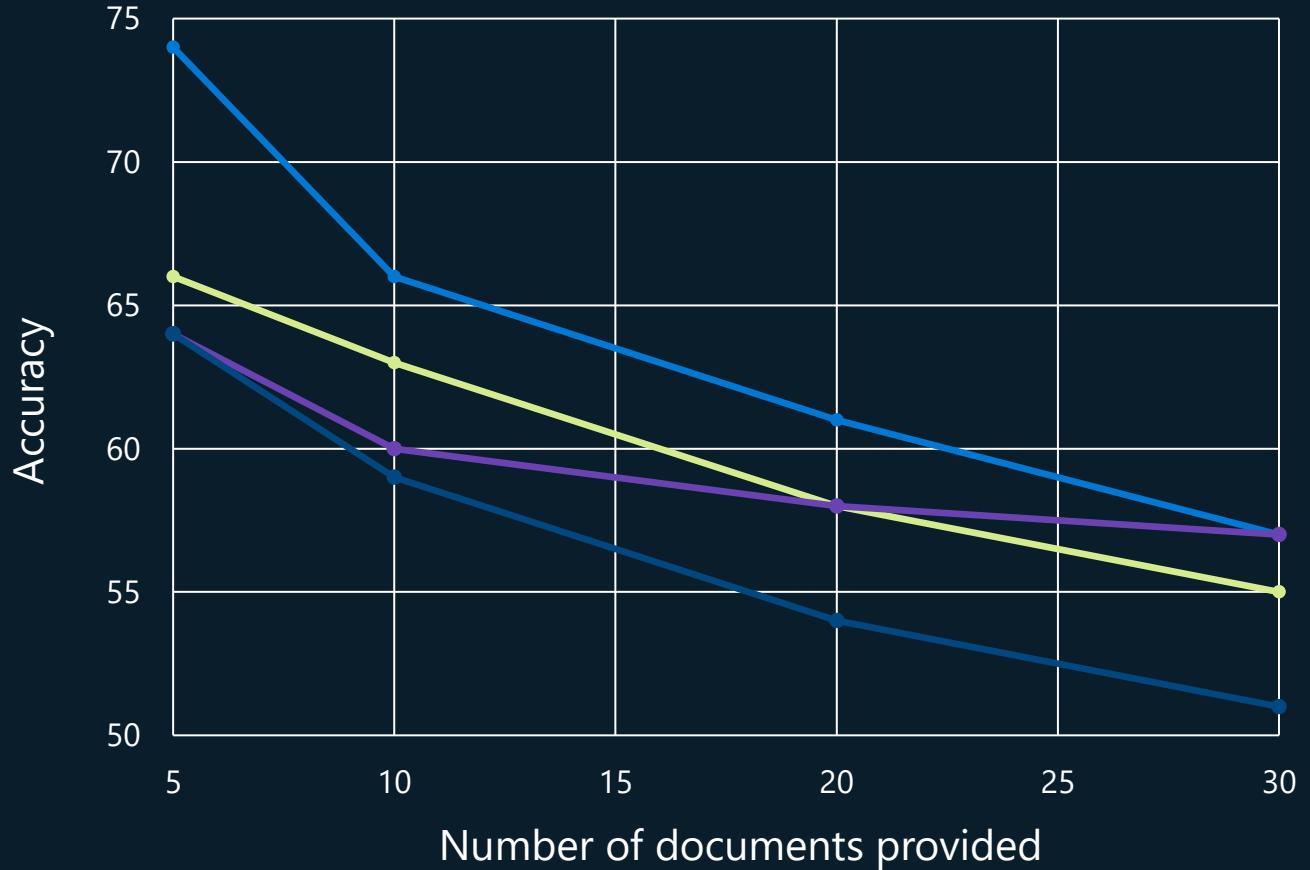
Retrieval – augmented generation

Anatomy of the workflow



Your retrieval strategy matters

More information ≠
better results



Source: Lost in the Middle: How Language Models Use Long Contexts, Liu et al. arXiv:2307.03172

Retrieval strategies



Keyword search

- **For exact, plain text matches**
- “Vocabulary gap” in Q&A systems like Copilot



Vector search

- **For conceptual similarity, or underlying meaning**
- Weak performance on exact matches (like a product ID or code)



Hybrid search

- **Best of both vectors and keywords**
- Brings more accurate responses across various scenarios



Search re-ranking

- **Scores and ranks all retrieved documents by relevance**
- Reranking runs after performing search strategy (can't retrieve information)



Anatomy of AI Design

Developer Productivity



Visual Studio



GitHub Copilot



Copilot Studio



Azure AI Studio



Azure AI Studio

AI Design solutions

AI Platform



Azure OpenAI



Model as a Service (MaaS)



Azure AI Search



Azure AI Services



Azure AI Studio

App Platform



Azure Kubernetes Service



Azure Container Apps



Azure App Service



Azure API Management



Azure Functions

Data Platform



Azure SQL Database



Azure Cosmos DB



OSS Databases

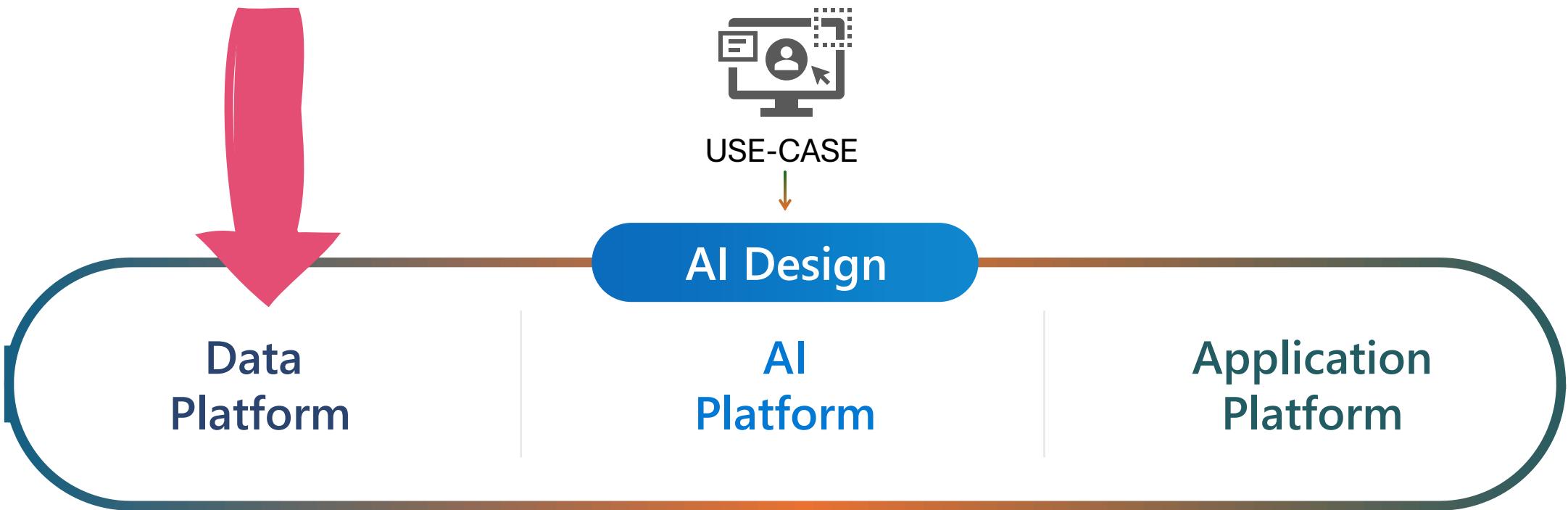


Microsoft Fabric

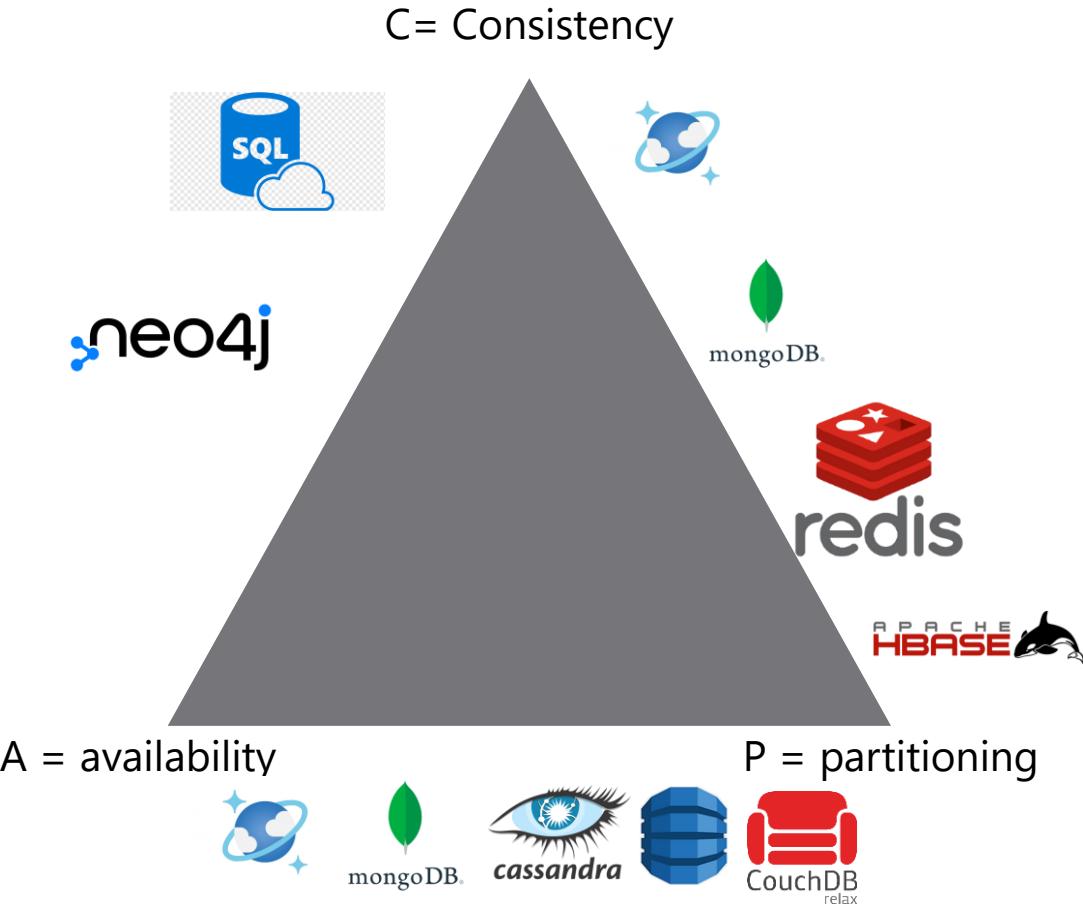


Databricks

Security, Governance, AI Safety, AI-ready infrastructure



CAP Theorem



The CAP theorem, also known as Brewer's theorem, states that in a distributed database system, it is impossible to simultaneously achieve all three of the following guarantees:

1. Consistency: Every read receives the most recent write or an error.
2. Availability: Every request receives a response, without guarantee that it contains the most recent write.
3. Partition-tolerance: The system continues to operate despite arbitrary partitioning due to network failures.

According to the CAP theorem, a distributed database can only provide two out of these three guarantees at any given time. This means that database designers must make trade-offs between consistency, availability, and partition-tolerance based on the specific requirements of their applications.

*Cosmos DB and MongoDB can be configured either way as CP or AP, depending on consistency models

Why would your intelligent app rely on a database?

Database+ Generative AI - scenarios

What

Why

When

Semantic Caching

Reduce latency
Saves on token Consumption

Not so dynamic content
FAQ, policies, ...

Chat history

Conversational context
UX improvements
LLM optimizations
Auditing

A MUST for chat sessions
Improving cost & performance

Retrieval Augmented Generation (RAG)

Personalize LLM on your data
Cheaper than fine tuning
Faster iteration on new data

Any GenAI app

Vector + operational database

No ETL
Consistent data
Reduced complexity and cost

Proximity search

Powering a new era of apps that run anywhere, at any scale



Managed

Ease your life and focus on what's important



Integrated

Modern, AI apps on an open, flexible, and common platform



Trusted

Available, scalable, safe and compliant. Unified support

Intelligent

AI-powered management and performance at any scale

Database native vector search OR/AND Azure AI search ?

Use Database native capabilities when you...

Want to have an **intelligent app** based on your **dynamic operational data**.

Augment your operational data (e.g. Real-time location-based search on inventory)

Need a **simple architecture**, you do not need anything else.

Need **always on, low latency, guaranteed throughput, geo-replication**, etc.

Cosmos DB for NoSQL



Cosmos DB for MongoDB vCore



Azure Database for PostgreSQL



Azure SQL



Use AI search if you...

Need to search across **multiple sources**, not only Cosmos DB or PostgreSQL, also pdfs, etc.

Content **does not change often**.

You need **additional capabilities** such as e.g. fuzzy, autocomplete, multi-language, meta-filtering, beyond vector search.

AI Search



Vector Search in Azure Cosmos DB for NoSQL

Store data + vectors together

Reduced complexity & cost
Transactional data & vectors
Vector indexing and search in a single service

Vector Search + Query Filters

Combine with equality, range & spatial filters
Optimize query

Flexible Indexing

Flat, quantized flat, and DiskANN indexing available

Azure Cosmos DB for NoSQL Capabilities

Serverless or provisioned throughput
Built-in multitenancy
Instant & dynamic autoscale
<10ms point-reads
Globally-replicated
Industry-leading 99.999% SLA

The DiskANN Advantage

Creates a graph-based index that stores compressed vectors in memory, and a full-fidelity graph on SSDs.
Advantages:

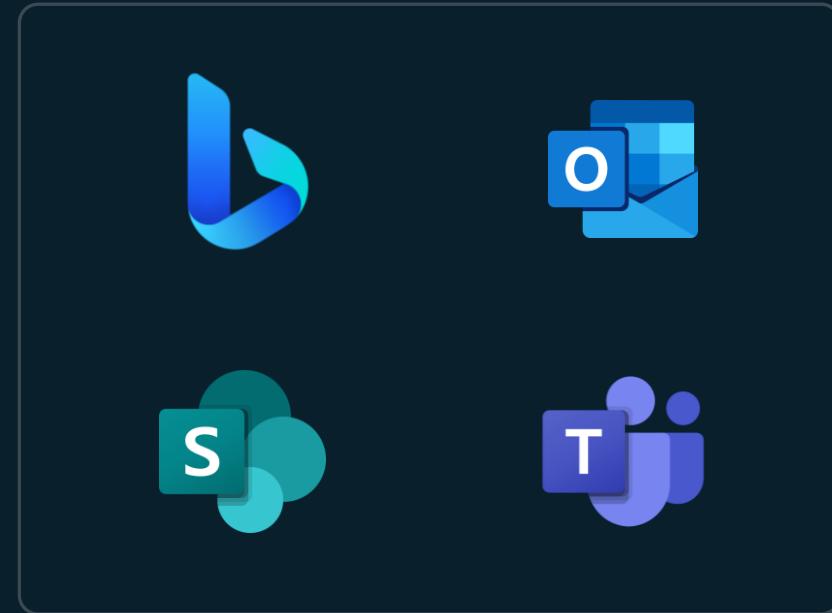
Low latency

High accuracy

Low cost

Robust to data changes

DiskANN powers semantic search across Microsoft



- Indexes with up to 400B vectors, trillions of points
- Mean query latency <5ms, >90% recall

DiskANN

Compression of vectors

Large Vectors
{ D1, D2, D3, D4, D5, ..., D99, D100 }



Quantization



Compressed Vectors
{ D1, D2 .., D10 }

Storage and graph construction

RAM

[compressed vector 1]
[compressed vector 1]
⋮



SSD

Low-diameter graph
+ full-precision vectors

Algorithms

Vamana
(vector indexing)

Pruning
(to reduce vector data
space on disk)

Search
(lookups)

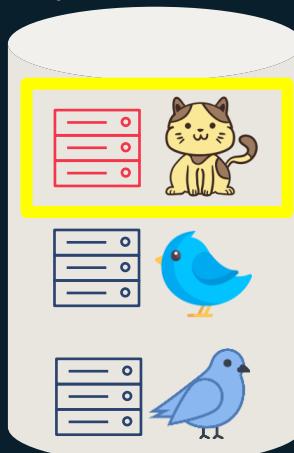
DiskANN



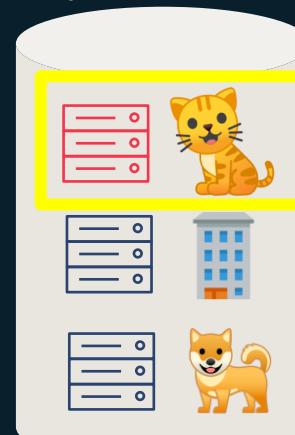
What is similar to
this?



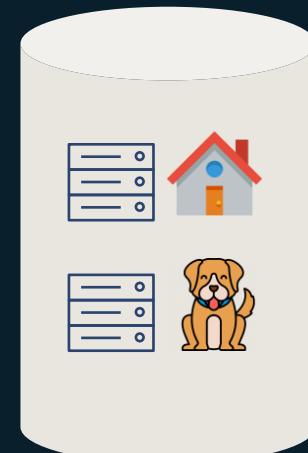
Physical Partition



Physical Partition



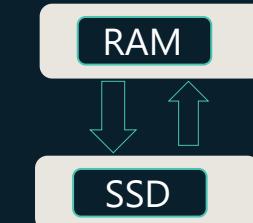
Physical Partition



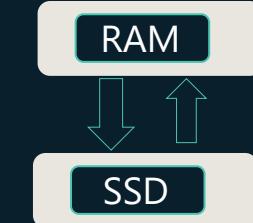
Vectors are stored as a property in the same logical unit as the data they represent.

DiskANN index information sits beside each partition.

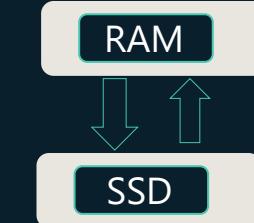
DiskANN index



DiskANN index



DiskANN index



RU consumption is very low because DiskANN creates a graph of reduced dimensions (low memory usage) which acts as a pointer to a full-precision graph on SSD to retrieve info on nearest neighbors

Document Ingestion & Processing



Solution Accelerator: <https://aka.ms/doc2cdb>

Natural language to Cosmos DB for NoSQL



Turn your natural
language data questions
in NoSQL queries



Integrated into
Data Explorer
at no charge
(Public Preview)



Copilot in Azure
capability for multi-
turn conversations
(Coming Soon)

Unlocking the power of Open AI and pgvector with Azure Database for PostgreSQL

Simplified LLMs with direct use of PostgreSQL data



Industry leading AI for building intelligent applications



[1.5, -0.4, ..., 20.2]

Native
Vectors



Azure AI
Integrated



Copilot
Powered

Industry leading AI for building intelligent applications



[1.5, -0.4, ..., 20.2]



Native Vectors

Store data + vectors
together

Reduced complexity & cost

Vector indexing and search
in a single service

Azure AI Integrated

Copilot Powered

Industry leading AI for building intelligent applications



[1.5, -0.4, ..., 20.2]



Native
Vectors

Azure AI
Integrated

Copilot
Powered

SQL Interface with Azure Open AI

Create embeddings from the DB with SQL
statements

SQL Interface with Azure AI
Language/Vision Services

Invoke ML models from within the DB

Industry leading AI for building intelligent applications



[1.5, -0.4, ..., 20.2]

Native
Vectors



Azure AI
Integrated



Copilot
Powered

Chat-based interface for
querying or debugging
(coming soon)

Native Vector Search

- **pgvector** extension
- Storing and indexing vectors alongside relational data
- Various indexing & retrieval strategies
- Combine vector queries with metadata filters

DiskANN now in public preview for better latency and accuracy!!!

[DiskANN Vector Index in Azure Database for PostgreSQL \(microsoft.com\)](#)

Generative AI apps

RAG (Retrieval Augmented Generation) apps
Retrieve private data to ground LLM model responses

Recommendation/Semantic Search

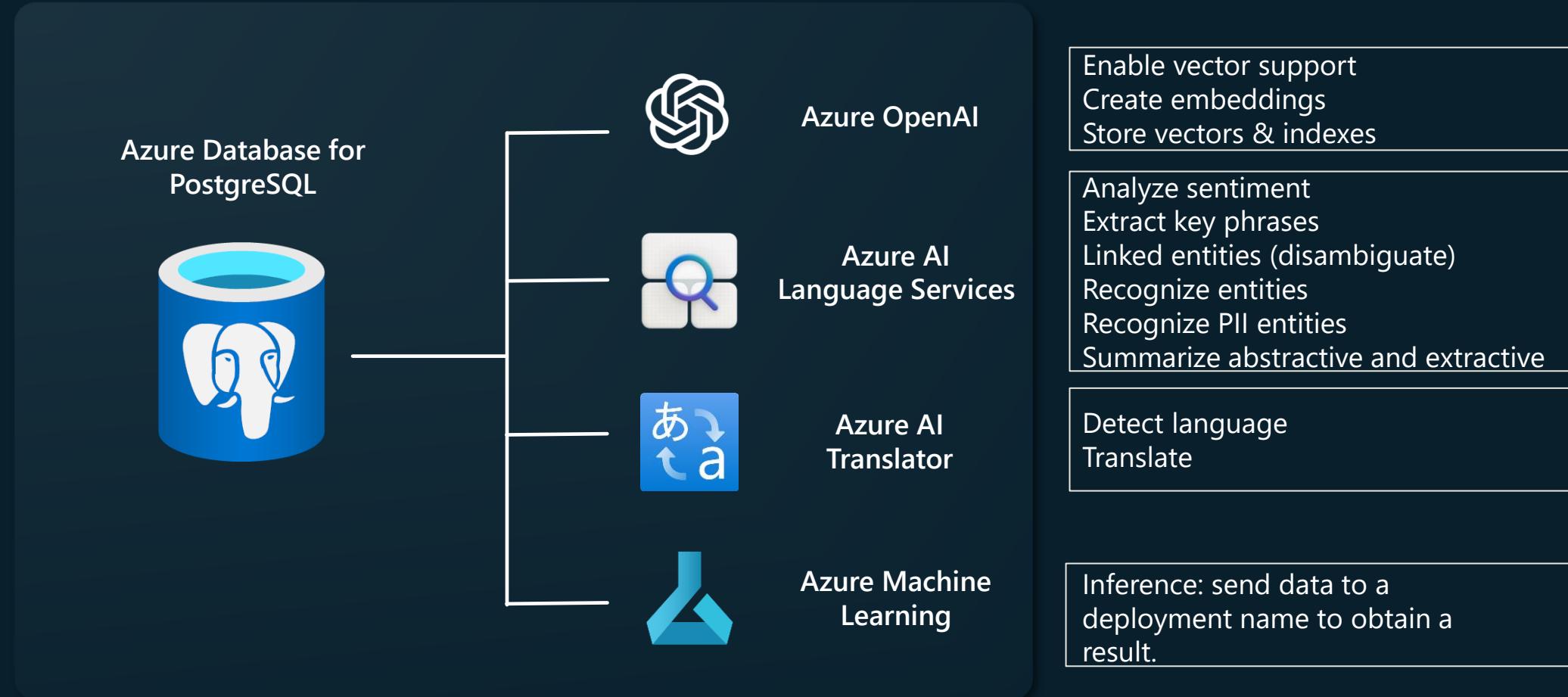
Retrieve similar documents by distance between vectors

Hybrid Search

Combine vector search, row filtering, and full-text search

AI Services integrated into Azure Postgres

Make remote calls directly from PostgreSQL with the `azure_ai` extension



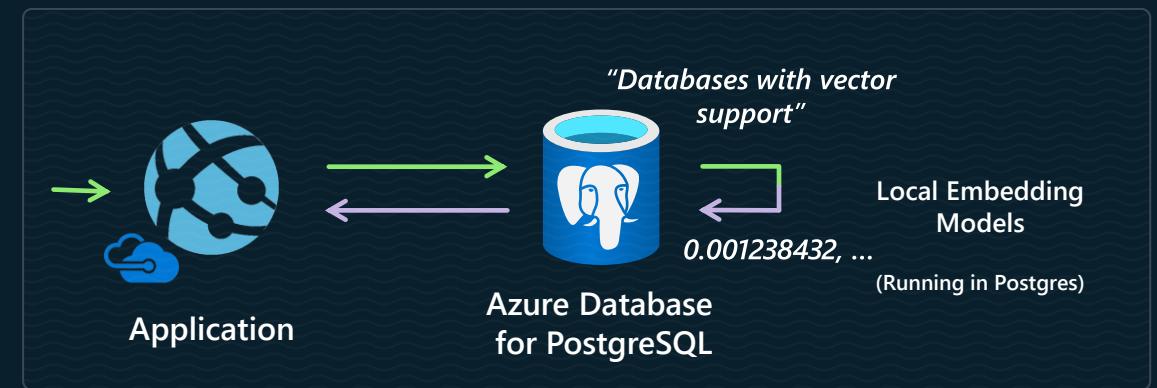
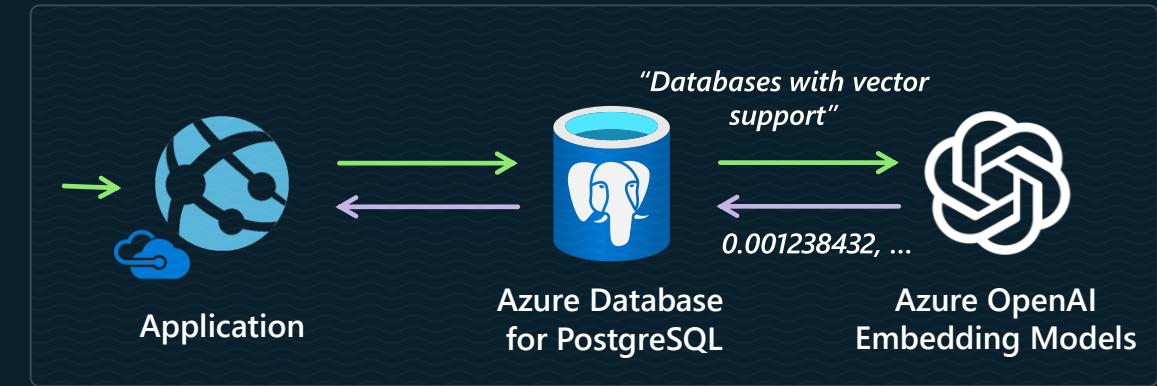
Vectors generation with SQL statement

Remote Embedding Models

```
SELECT * FROM <table>
ORDER BY
database_description <->
azure_openai.create_embeddings(
'text-embedding-ada-002',
'Databases with vector support')
```

In-Database Embedding Models (Preview, azure_local_ai extension)

```
SELECT * FROM <table>
ORDER BY
recipe_embedding <#>
azure_local_ai.create_embeddings(
'multilingual-e5-small:v1',
'Databases with vector support')
```



Data Privacy, data stays in database
90% faster and less costly (no data transmission, no additional external services used)

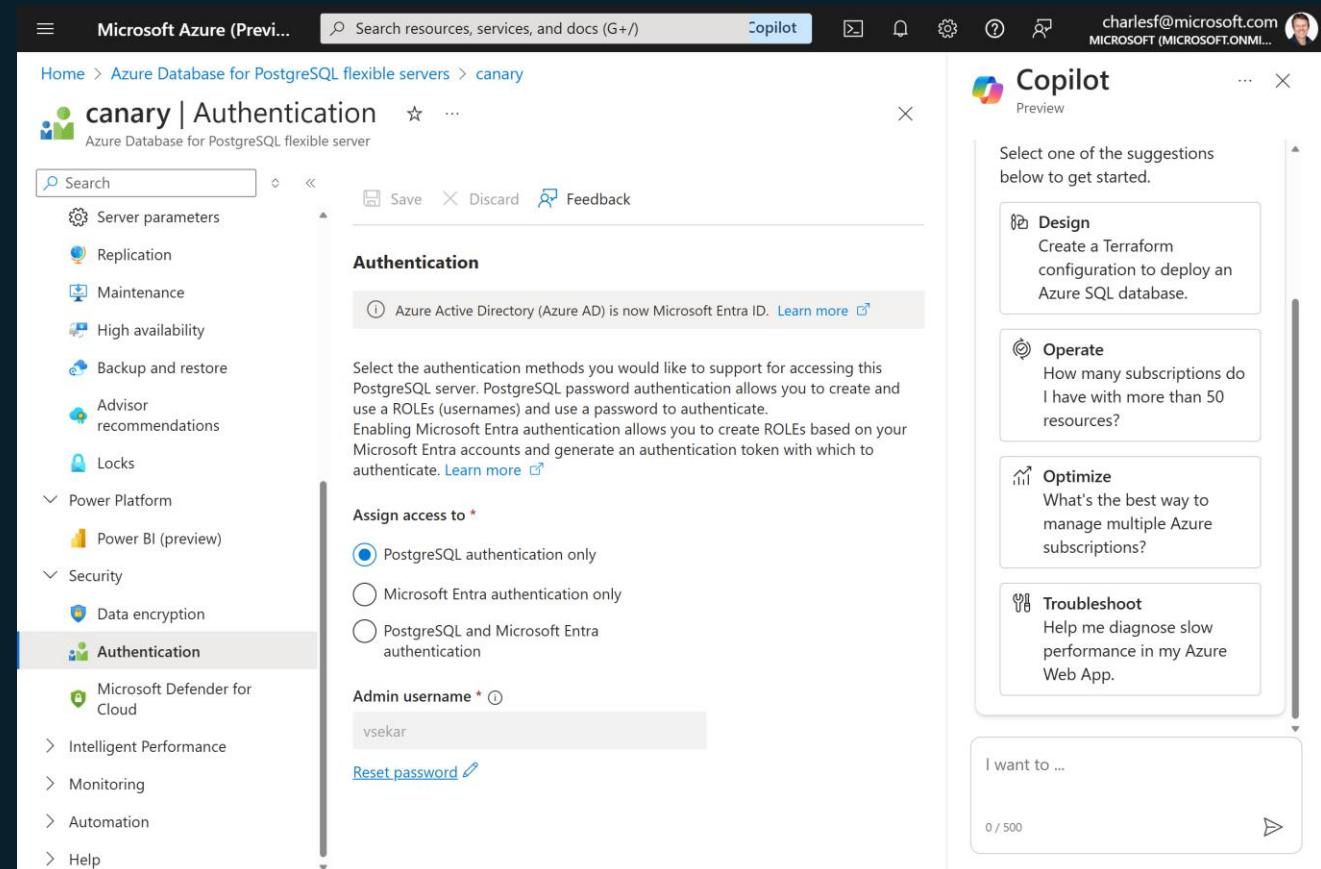
Coming soon

Azure Copilot for PostgreSQL Flexible Server

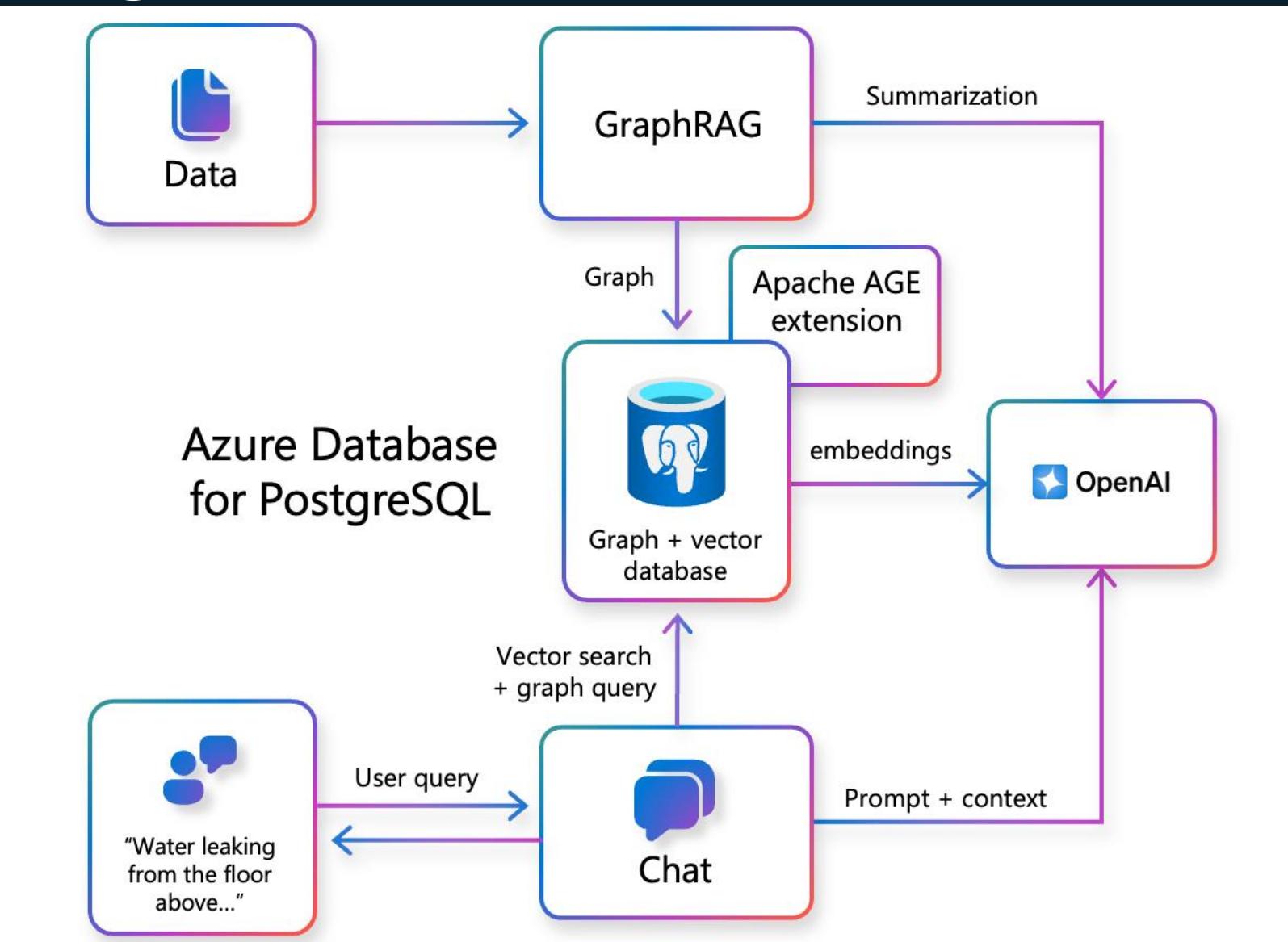
Chat-based interface for querying or debugging

Contextual awareness of the individual server provides the most relevant responses on how to troubleshoot vs. generic documentation

Read-only scenario initially, followed by actions where the Copilot can apply the recommended changes to the server itself with user approval



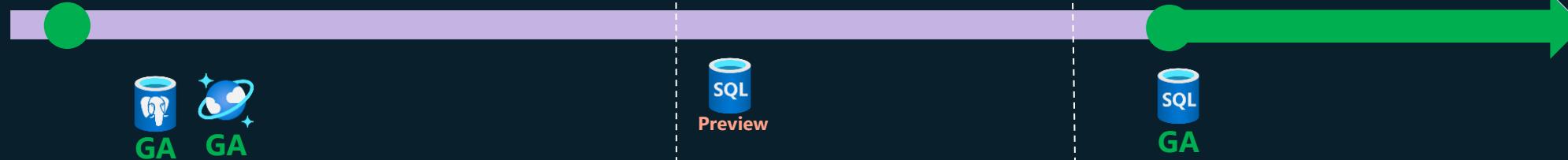
GraphRAG with PostgreSQL



Solution Accelerator:
<https://github.com/Azure-Samples/graphrag-legalcases-postgres>

Recently

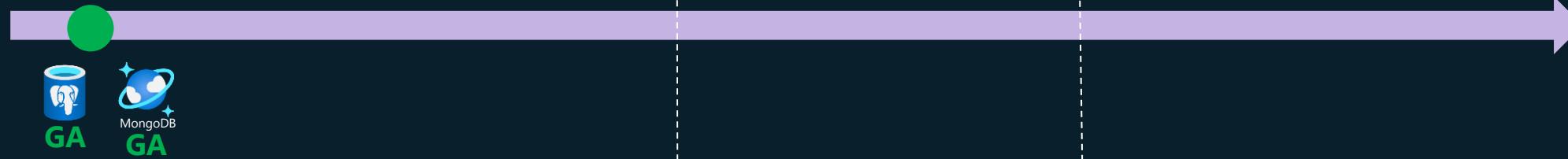
Vector Data Type and Query



Vector Index (IVFFLAT / Flat)



Vector Index with HNSW



Currently

Near Future

Recently

Vector Data Type and Query



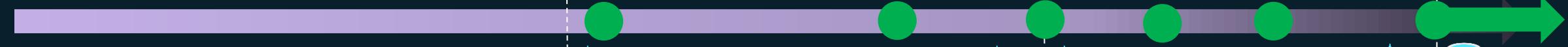
Vector Index (IVFFLAT / Flat)



Vector Index with HNSW



Vector Index with DiskANN



**Index
Feature
Parity**

Database native vector search Comparison

Cosmos DB for NoSQL



Cosmos DB for MongoDB vCore



Azure Database for PostgreSQL



Azure SQL



AI Search



Technical features

- Native vector search support with DiskANN.
- Full text search (BM25) and Hybrid search (RRF) (in preview).

- Scalable document storage with key-value and document models.
- Useful for **fast lookup operations** in metadata-heavy vector search

- Native vector search (DiskANN, HNSW, IVF)
- Filtered vector search (preview)

- Suitable for applications **already using MongoDB**, seeking Azure-managed infrastructure.

- Native support through the pgvector and pg_diskann (preview) extension.
- Semantic Ranking through azure_ai extention and Azure ML.
- GraphRAG

- Ideal for applications needing **relational capabilities** alongside vector search.
- PostgreSQL extensions (like pgvector) enable seamless vector search integration.

- Native vector type (preview)
- Vector index through DiskANN (coming soon)
- Hybrid Search (preview)

- Ideal for applications needing **relational capabilities** alongside vector search.

- Native vector search capabilities.
 - Integration with OpenAI embeddings and cognitive skills.
 - Full hybrid search support and Semantic Ranker
-
- Purpose-built for **search applications**, including hybrid search
 - Excellent for content-based recommendation systems, enterprise search, and ML-driven search interfaces.

Use-cases

Database native vector search Comparison

Cosmos DB for NoSQL



Cosmos DB for MongoDB vCore



Azure Database for PostgreSQL



Azure SQL



AI Search



Technical features

- Native vector search support with DiskANN.
- Full text search (BM25) and Hybrid search (RRF) (in preview).

- Scalable document storage with key-value and document models.
- Useful for **fast lookup operations** in metadata-heavy vector search

Pay for *provisioned throughput* (RU/s)
30K RU/s, 1TB storage
~2,400 EUR/month

- Native vector search (DiskANN, HNSW, IVF)
- Filtered vector search (preview)

Suitable for applications **already using MongoDB**, seeking Azure-managed infrastructure.

Pricing based on *vCores and storage capacity*
M200 cluster, 64vcores, 1TB
~7,200 EUR/month

- Native support through the pgvector and pg_diskann (preview) extension.
- Semantic Ranking through azure_ai extention and Azure ML.
- GraphRAG

Ideal for applications needing **relational capabilities** alongside vector search.
PostgreSQL extensions (like pgvector) enable seamless vector search integration.

Pricing based on *vCore and storage tiers* (for DiskANN).
E64bsv5, 64 vcores, 1TB SDD v2
~6,000 EUR/month

- Native vector type (preview)
- Vector index through DiskANN (coming soon)
- Hybrid Search (preview)

Ideal for applications needing **relational capabilities** alongside vector search.

Tiered pricing based on *performance (DTUs or vCores)*. Hyperscale Premium series, 64vCores
~10,600 EUR/month

- Native vector search capabilities.
- Integration with OpenAI embeddings and cognitive skills.
- Full hybrid search support and Semantic Ranker

Purpose-built for **search applications**, including hybrid search
Excellent for content-based recommendation systems, enterprise search, and ML-driven search interfaces.

Pricing based on *query volume and storage*.
S3 (5 Partitions, 5 Replicas 25 Units)
~49,000 EUR/month

Use-cases

Case: 100M vectors (1536 dimensions), 100QPS



Hands-on Lab

RAG: Agent-First Retrieval Patterns

Juan Manuel Servera Bondroit, Sr. Partner Solution Architect



Equipment manuals,
markdown files,
consolidated in OneLake

**Equipment with label P4324 not working, CTL11 light
is red, maybe power supply problem? The cord has
another label UL 817 whatever that means. Okay to
replace the part?**

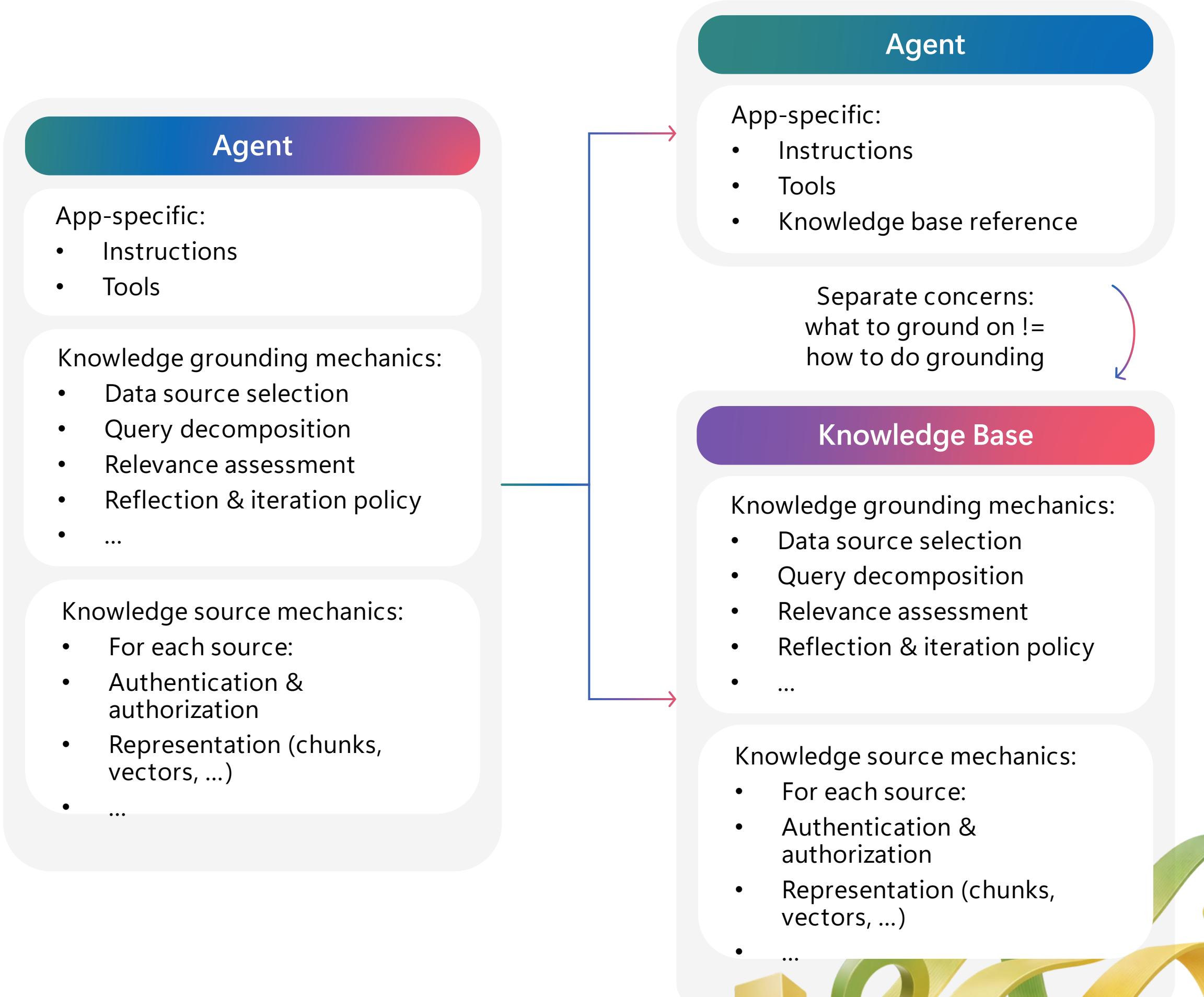
Electrical standard,
public information,
available on the Web

Company policy,
Office documents,
maintained in SharePoint

Knowledge Bases

Represent a complete domain of knowledge

Separate concerns between agents and knowledge retrieval



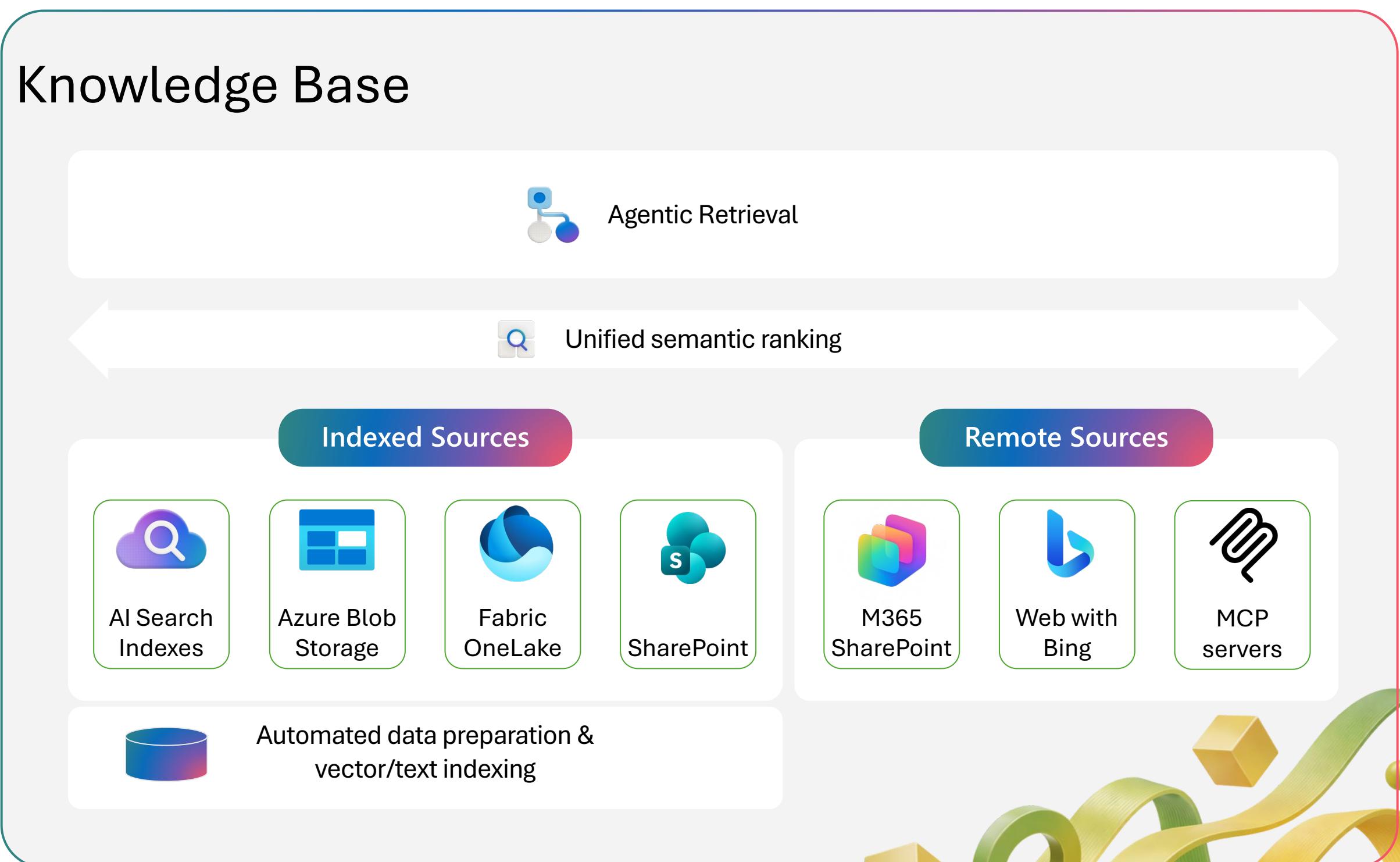
Knowledge Sources

Encapsulate a source of data

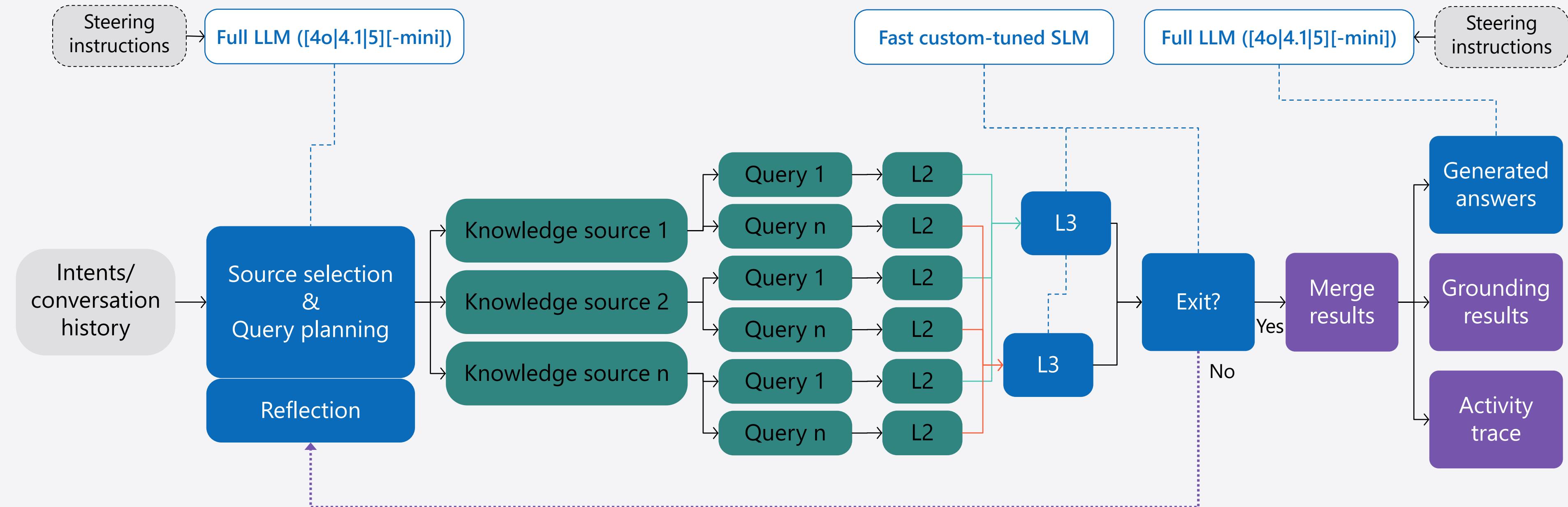
- Connection
- Authentication
- Authorization
- Instructions

Indexed or Remote

All get semantic reranking boost

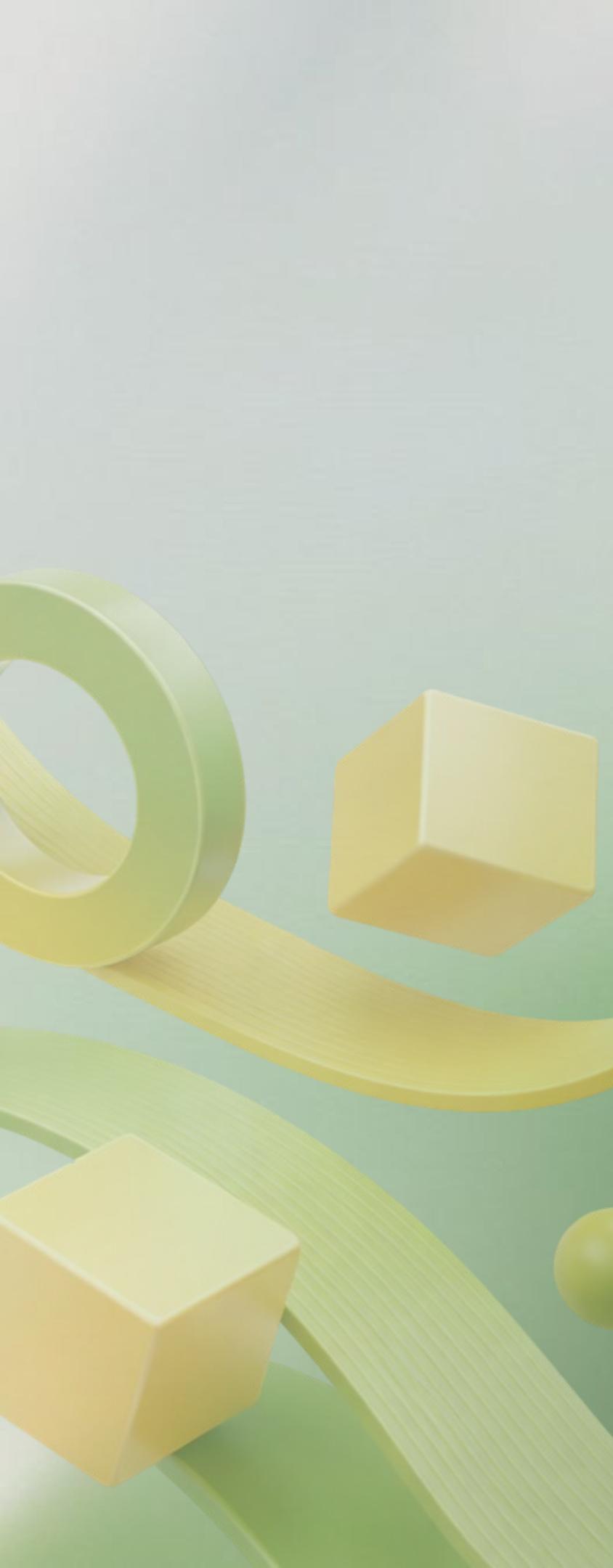


Agentic Retrieval Pipeline



Reasoning effort level →	Minimal	Low	Medium
Query planning & source selection	No	Yes	Yes
Web grounding	No	Yes*	Yes*
L2 ranking	Yes	Yes	Yes
L3 classifier	No	No	Yes
Reflection/iteration	No	No	Yes
Answer generation	No	Optional	Optional

* Requires answer gen.



LAB

<https://aka.ms/aita/lab511>

1. Fork the repo
2. Clone locally (or open in codespace)
3. Run infra/deploy-yourself/deploy.[ps1 or sh]
4. Run infra/deploy-yourself/setup-environment.[ps1 or sh]
5. Go through notebooks 1-8 (you can skip nr 3 if you don't have a sharepoint site)

Estimated cost <\$10 per day