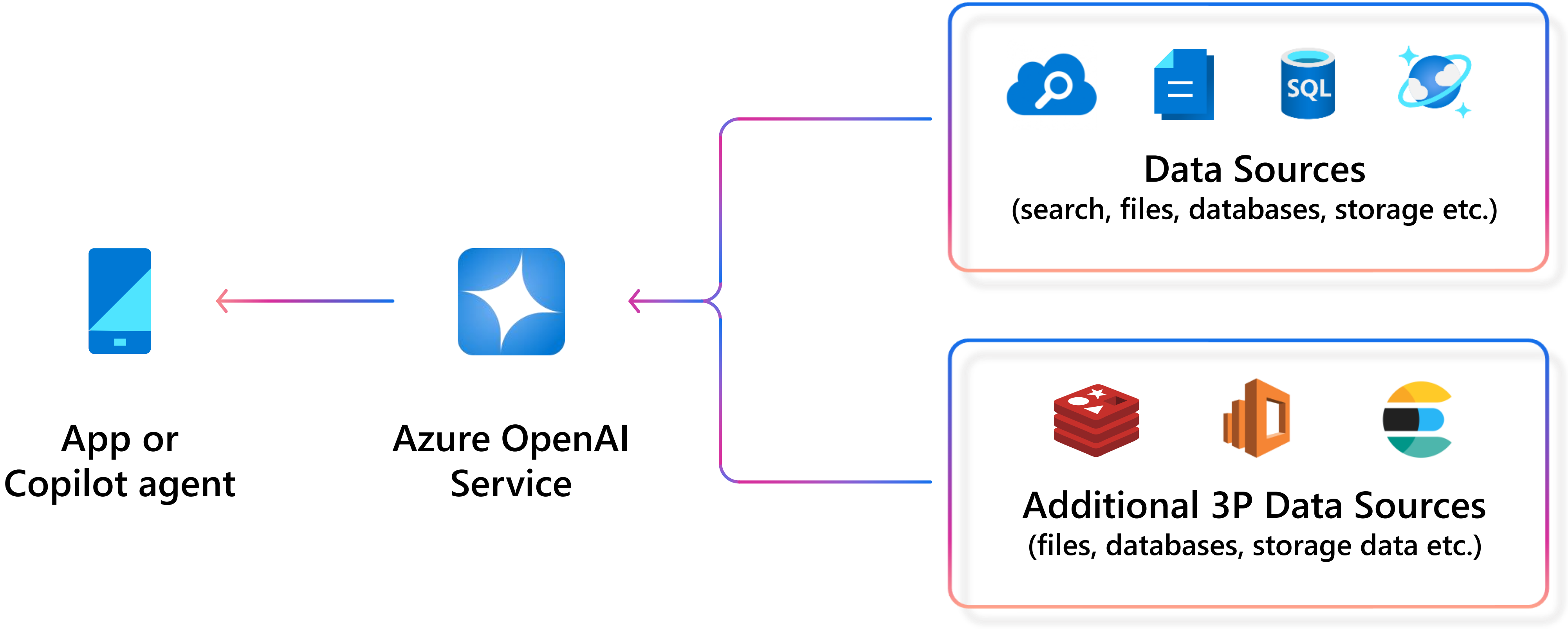
Retrieval-augmented generation (RAG)

RAG: Bring your own data to LLM



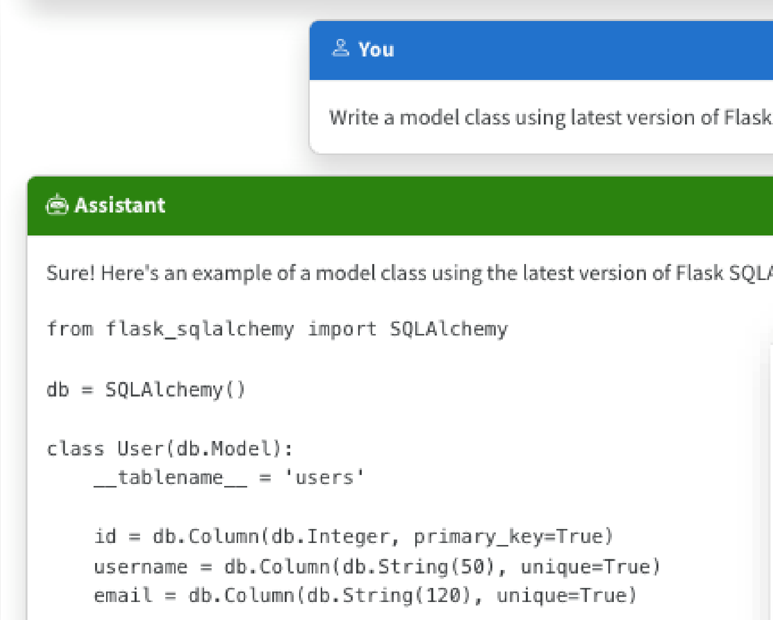
App or
Copilot agent

Azure OpenAI
Service

Data Sources
(search, files, databases, storage etc.)

Additional 3P Data Sources
(files, databases, storage data etc.)

The limitations of LLMS



Write a model class using latest version of Flask SQL Alchemy

Sure! Here's an example of a model class using the latest version of Flask SQLAlchemy:

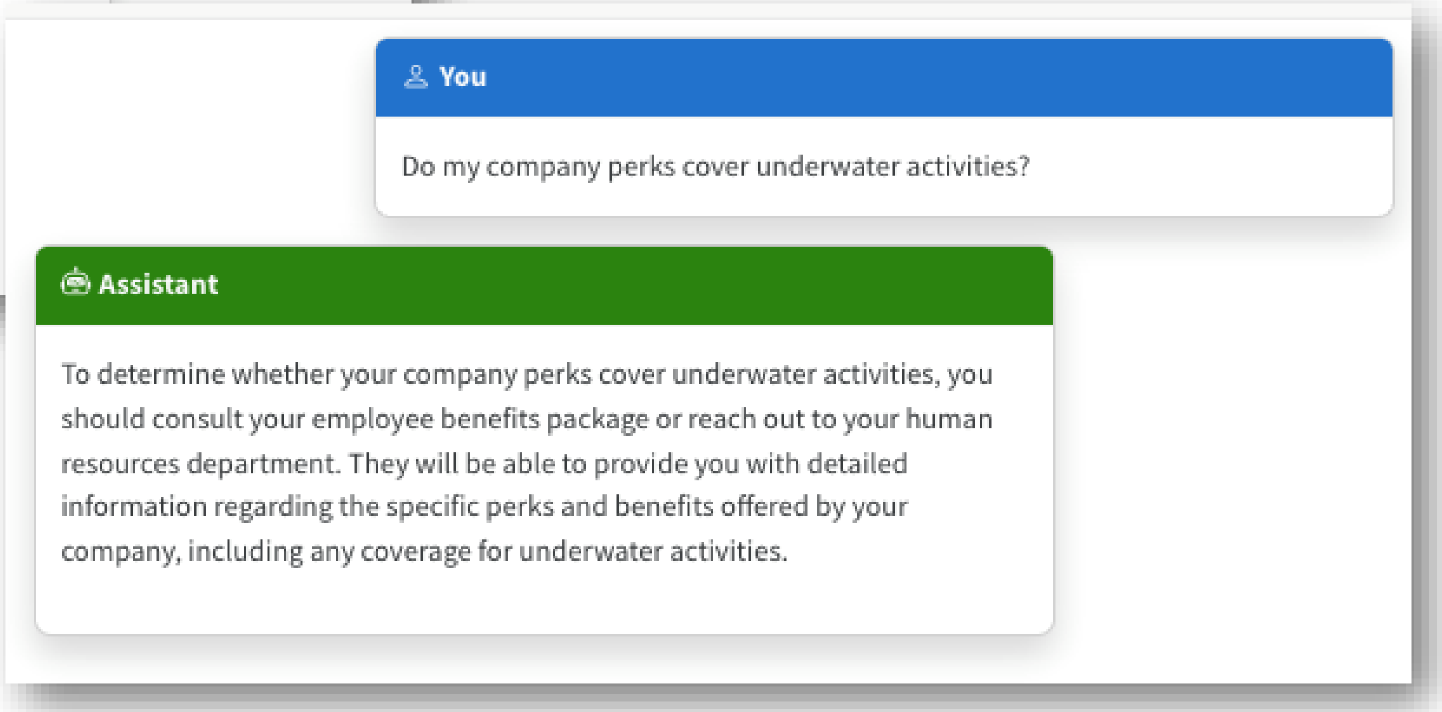
```
from flask_sqlalchemy import SQLAlchemy

db = SQLAlchemy()

class User(db.Model):
    __tablename__ = 'users'

    id = db.Column(db.Integer, primary_key=True)
    username = db.Column(db.String(50), unique=True)
    email = db.Column(db.String(120), unique=True)
```

Outdated public knowledge

Chat GPT 4o was trained on public data up to October 2024.

No internal knowledge

Companies have lots of valuable documents and data outside of the reach of Chat GPT.



Do my company perks cover underwater activities?

To determine whether your company perks cover underwater activities, you should consult your employee benefits package or reach out to your human resources department. They will be able to provide you with detailed information regarding the specific perks and benefits offered by your company, including any coverage for underwater activities.

Incorporating domain knowledge

Prompt engineering

In-context learning

Fine tuning

Learn new skills (permanently)

Retrieval augmentation

Learn new facts (temporarily)

RAG - Bring your data to the prompt

Prompt

You are an intelligent assistant helping Contoso, Inc. employees with questions about their healthcare plan as well as the employee handbook. Answer the following question using only the data provided in the sources below.

Question: Does my health plan cover annual eye exams?

Sources:
1. Northwind Health Plus offers coverage for vision exams, glasses, and contact lenses, as well as dental exams, cleanings, and fillings.
2. Northwind Standard only offers coverage for vision exams and glasses.
3. Both plans offer coverage for vision and dental services.

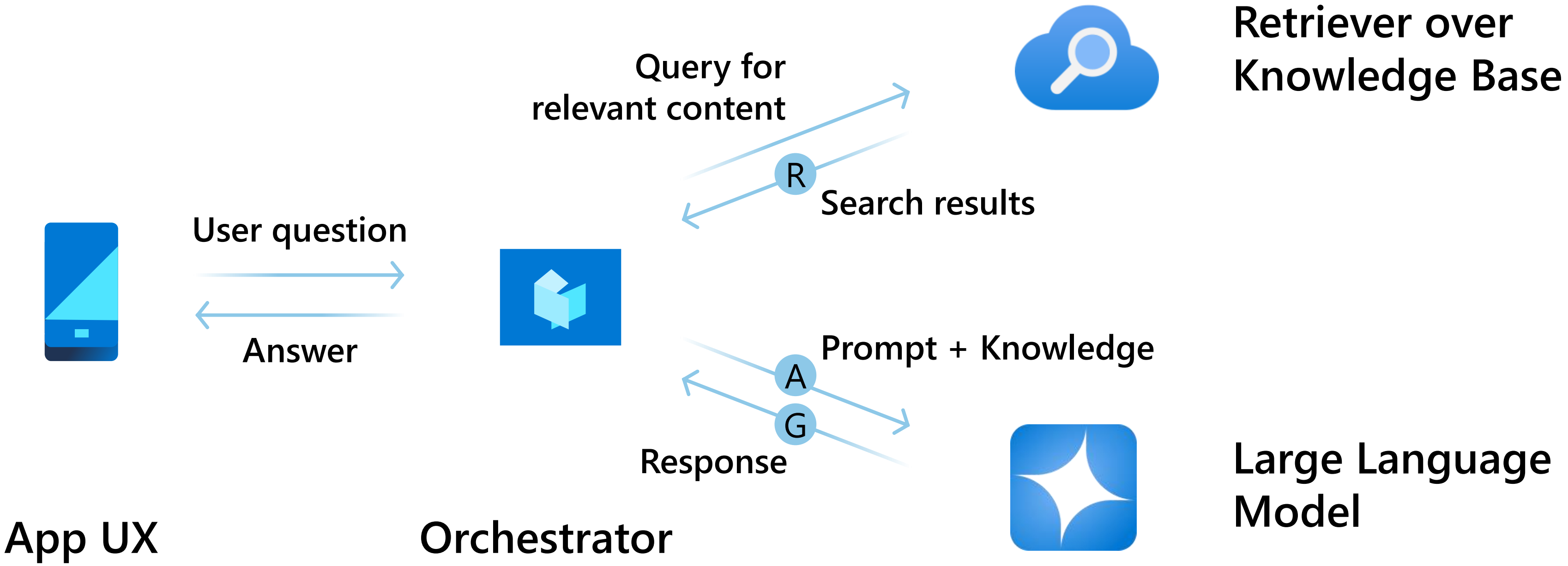
Response

Based on the provided information, it can be determined that both health plans offered by Northwind Health Plus and Northwind Standard provide coverage for vision exams. Therefore, your health plan should cover annual eye exams.

Text input that provides some framing as to how the engine should behave → User provided question that needs to be answered → Sources used to answer the question

Retrieval Augmented Generation (RAG)

Anatomy of the workflow



App UX

User question

Answer

Orchestrator

Query for
relevant content

R

Search results

Retriever over
Knowledge Base

Prompt + Knowledge

A

G

Response

Large Language
Model

What is the RAG searching?

Database rows (Structured data)

You need a way to **vectorize** & **search** target columns.

- **PostgreSQL** + pgvector
- **Azure SQL** + vector search
- **Cosmos MongoDB** + vector

Documents (Unstructured data)

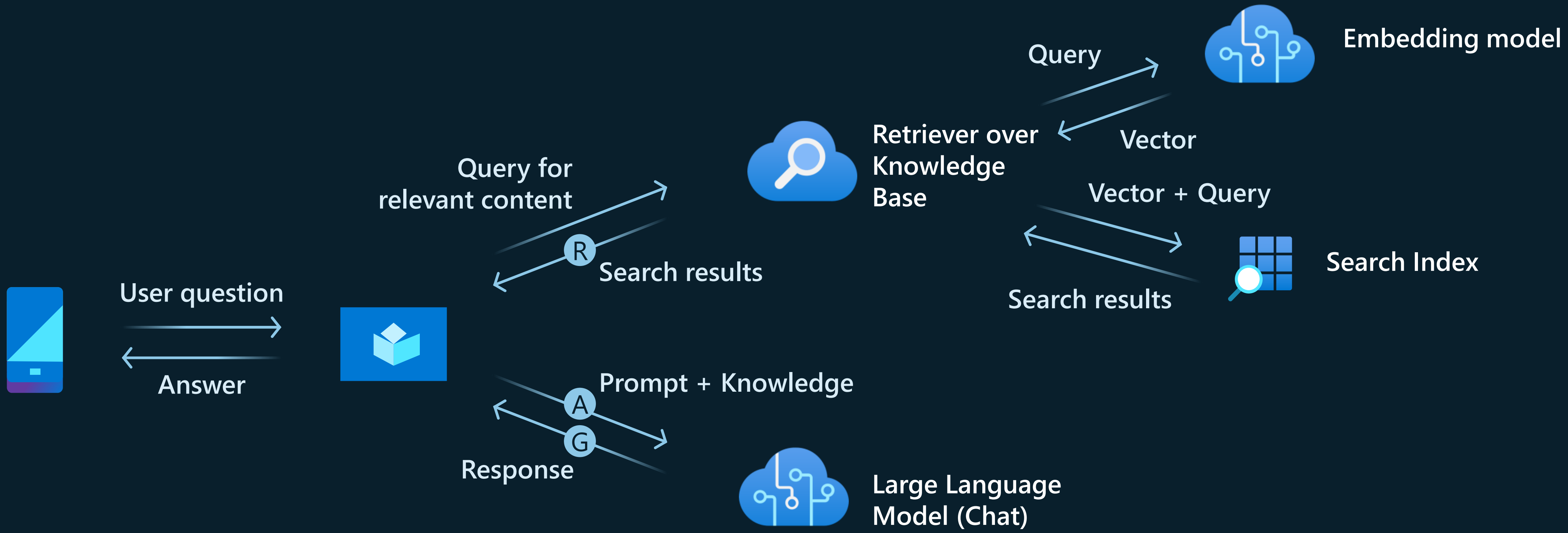
PDFs, docx, pptx, md, html, images

You need an ingestion process for **extracting**, **splitting**, **vectorizing**, and **storing** document chunks.

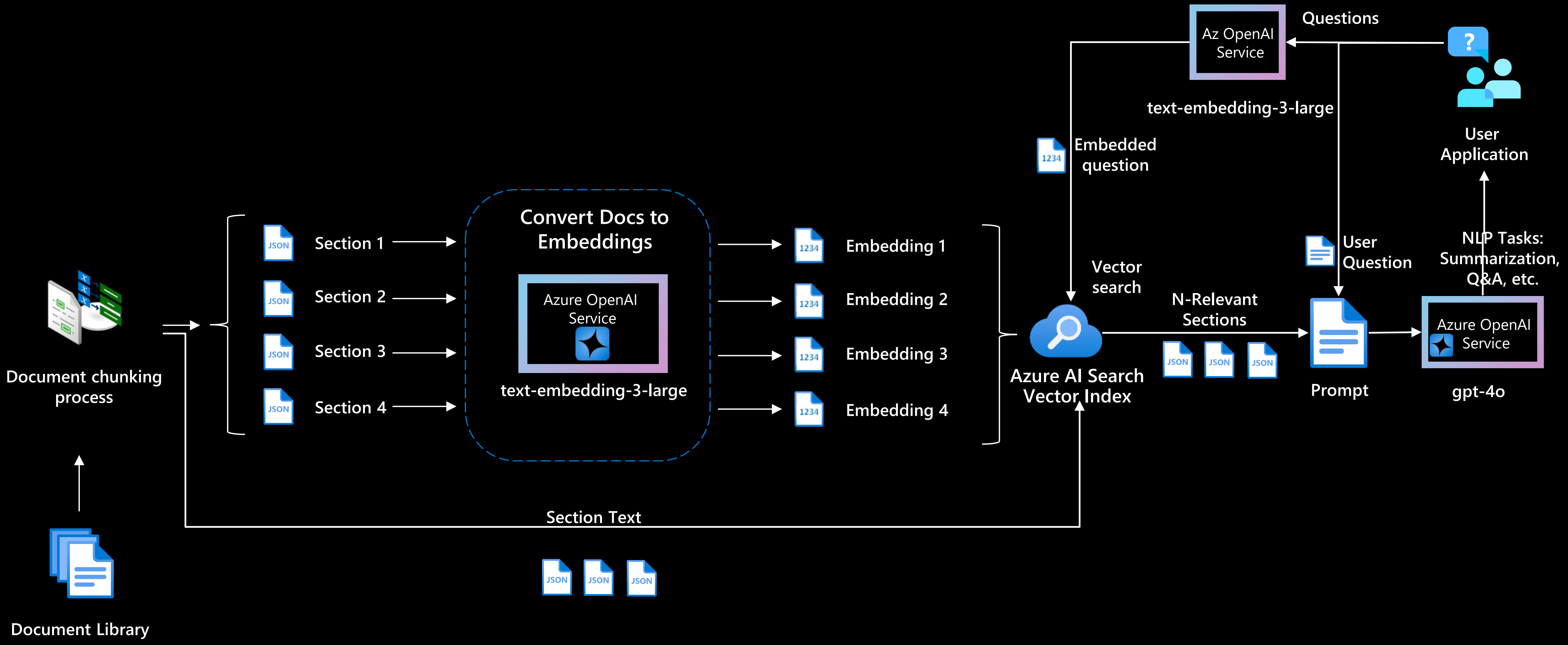
Azure AI Search +
Integrated Vectorization
Document Intelligence

Retrieval Augmented Generation (RAG)

with Embeddings



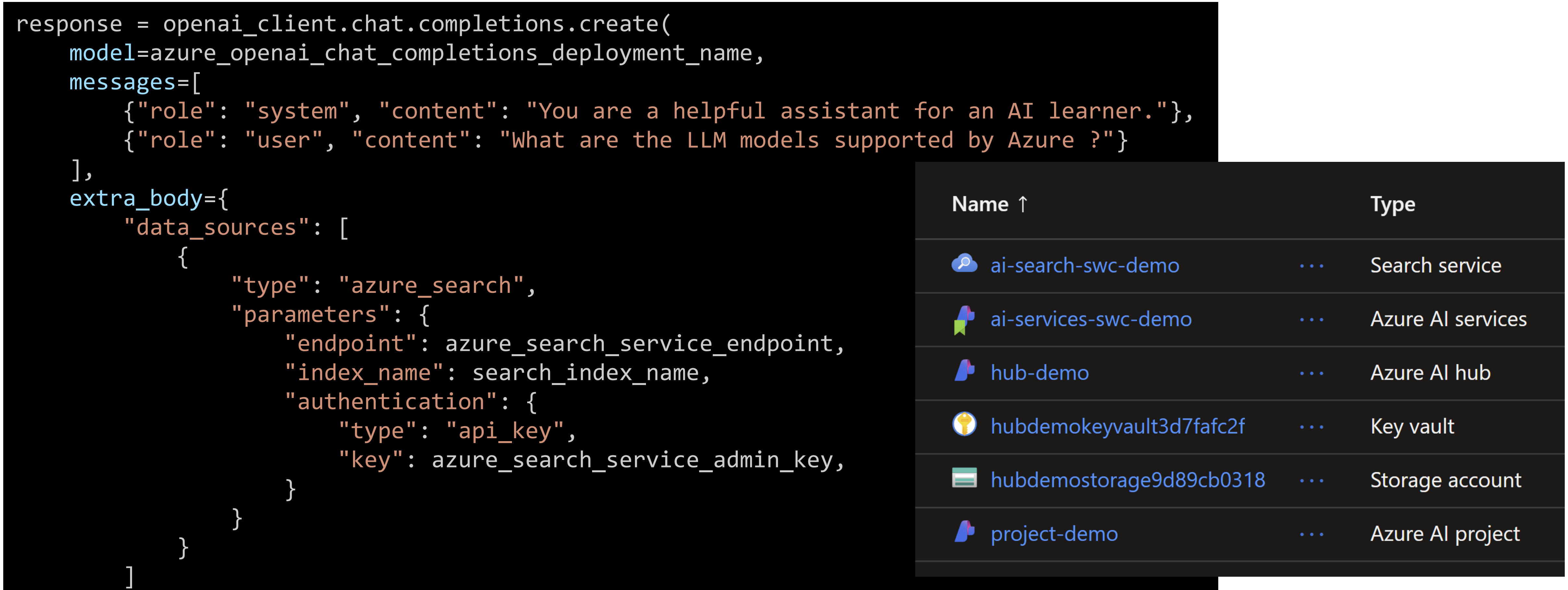
Complete RAG flow



Retrieval Augmented Generation (RAG) - Demo

RAG involves the following tools

- LLM model (ChatGPT, Mistral, Gemini…)
- Embedding model (embedding-model-3-large…)
- Document search database with support for vectorization (Azure AI Search…)

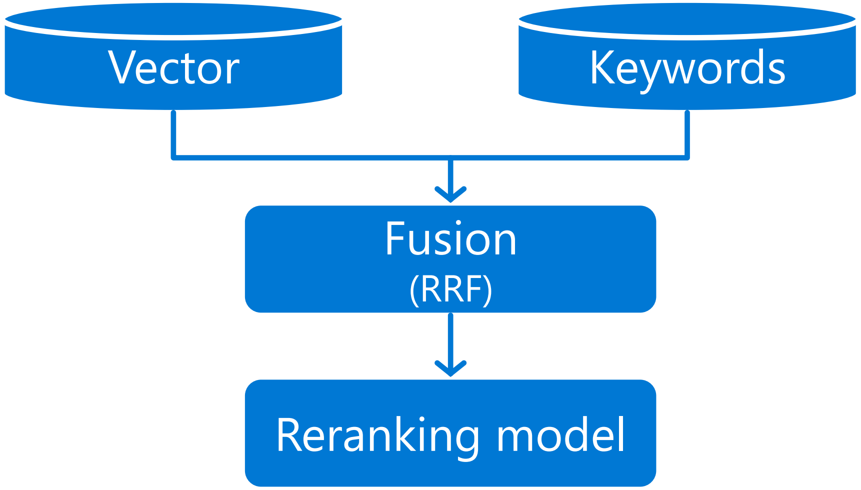


```
response = openai_client.chat.completions.create(
    model=azure_openai_chat_completions_deployment_name,
    messages=[
        {"role": "system", "content": "You are a helpful assistant for an AI learner."},
        {"role": "user", "content": "What are the LLM models supported by Azure ?"}
    ],
    extra_body={
        "data_sources": [
            {
                "type": "azure_search",
                "parameters": {
                    "endpoint": azure_search_service_endpoint,
                    "index_name": search_index_name,
                    "authentication": {
                        "type": "api_key",
                        "key": azure_search_service_admin_key,
                    }
                }
            }
        ]
```

Name ↑		Type
ai-search-swc-demo	···	Search service
ai-services-swc-demo	···	Azure AI services
hub-demo	···	Azure AI hub
hubdemokeyvault3d7fafc2f	···	Key vault
hubdemostorage9d89cb0318	···	Storage account
project-demo	···	Azure AI project

Components of a high-quality RAG

- **Powerful search functionality**
  - Vector search
  - Hybrid search
  - Semantic re-ranking
  - Filtering
- **Well prepared data**
  - Reasonably sized text
  - Meaningful vectors
- **Sophisticated LLM**
  - Adheres to instructions
  - Supports function calling


Vector
Keywords
Fusion
(RRF)
Reranking model