Deploying and Maintaining RAG Systems

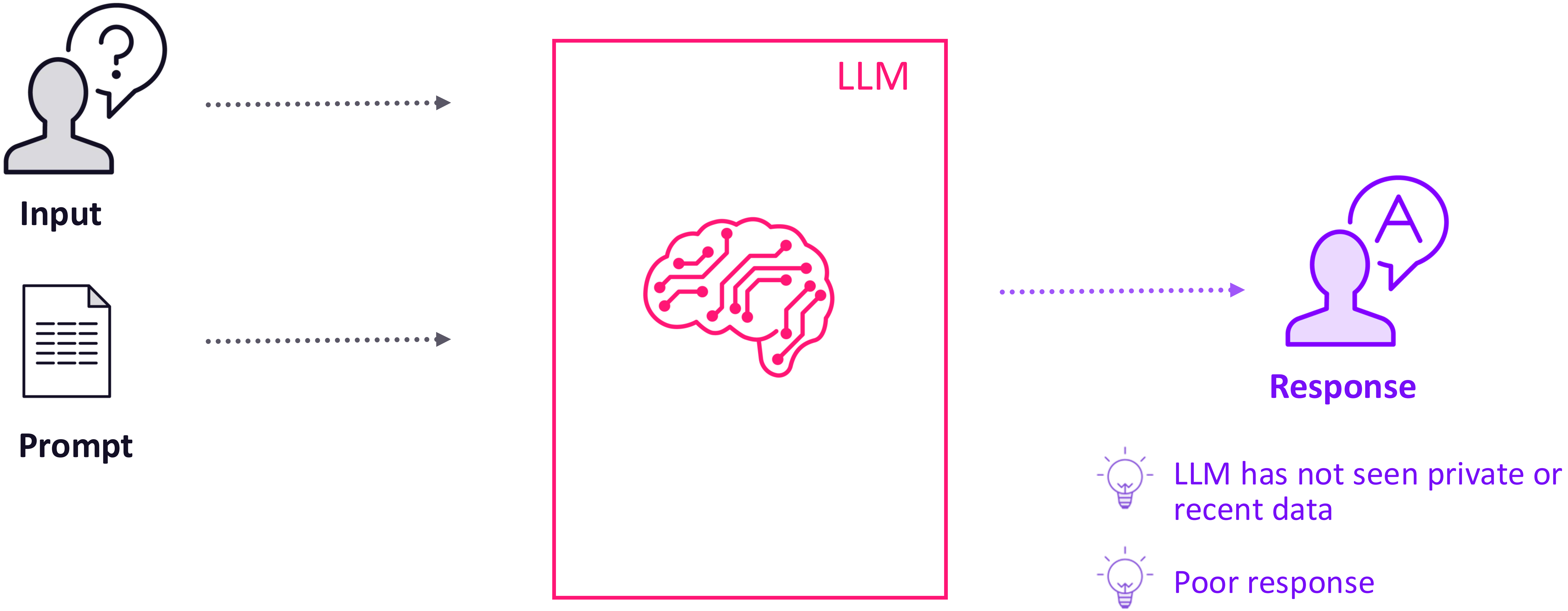
Building and Deploying RAG in Production

Abhishek Kumar

Data Scientist | Author | Speaker

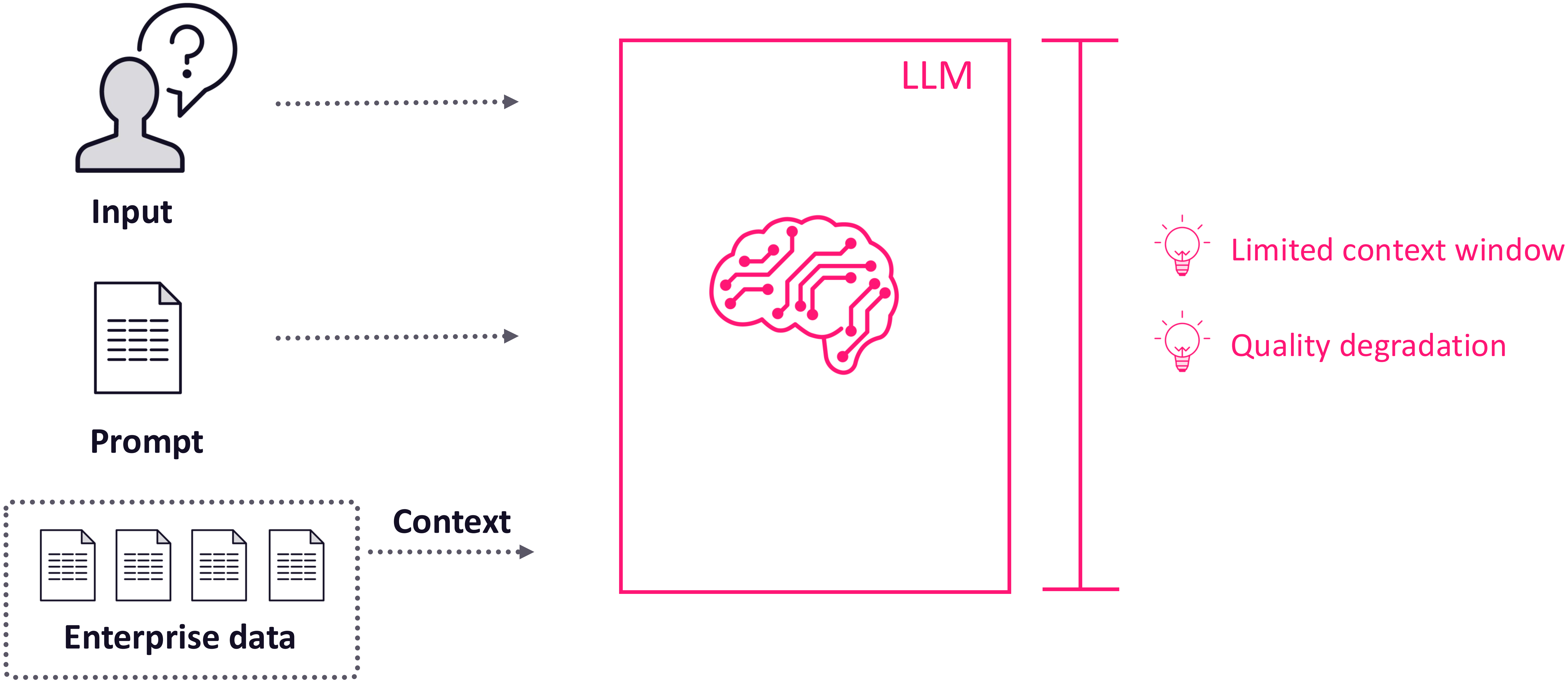
@meabhishekkumar

Prompt Engineering Alone Is Not Sufficient



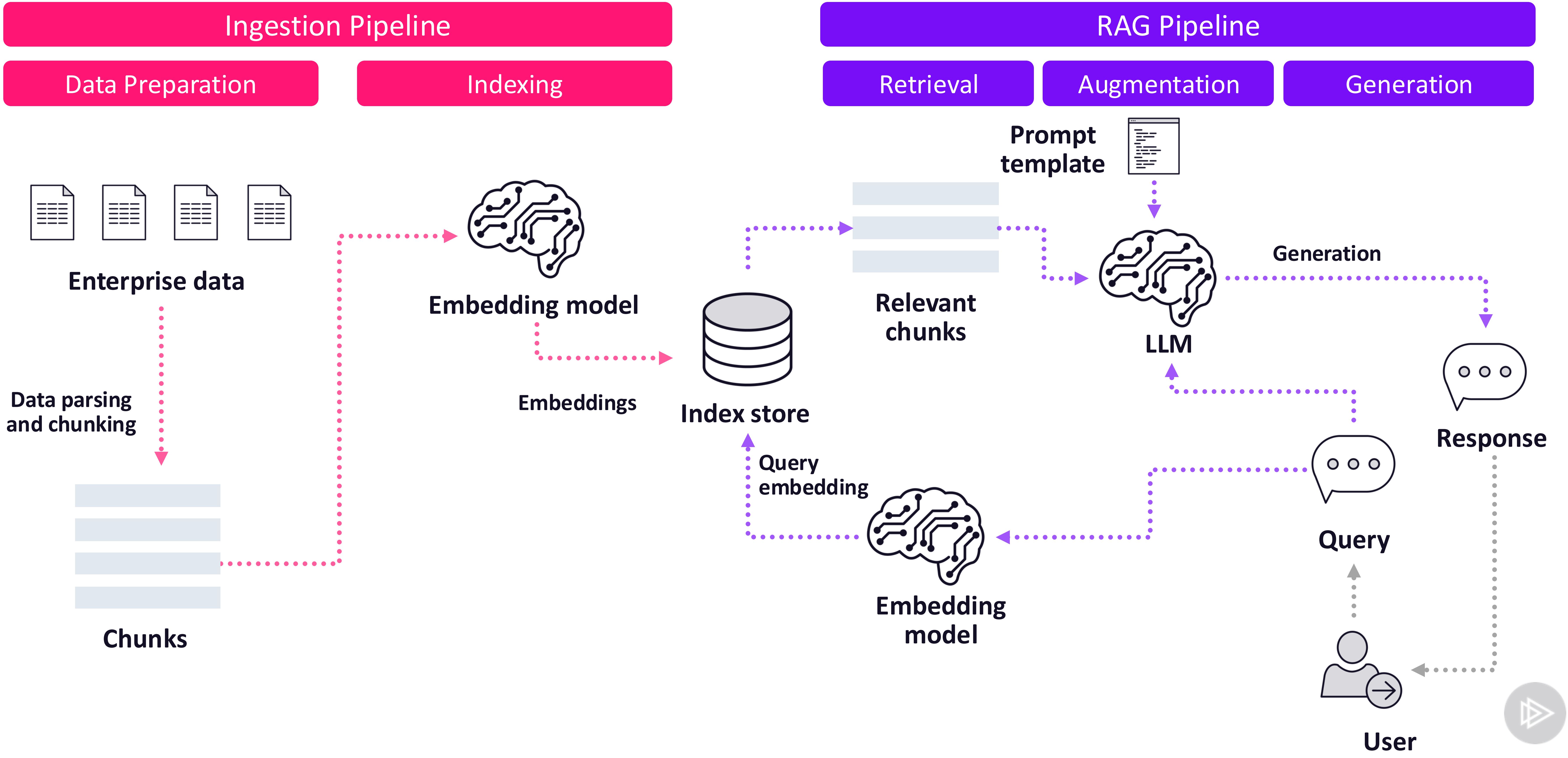
- LLM has not seen private or recent data
- Poor response

Can We Not Stuff Everything in Prompt?



Retrieval Augmented Generation (RAG) to the rescue

RAG System


Ingestion Pipeline
Data Preparation
Indexing
RAG Pipeline
Retrieval
Augmentation
Generation
Enterprise data
Data parsing and chunking
Chunks
Embedding model
Embeddings
Index store
Query embedding
Embedding model
Relevant chunks
Prompt template
LLM
Generation
Response
Query
User

Demo

Creating embeddings

Benefits of RAG Systems

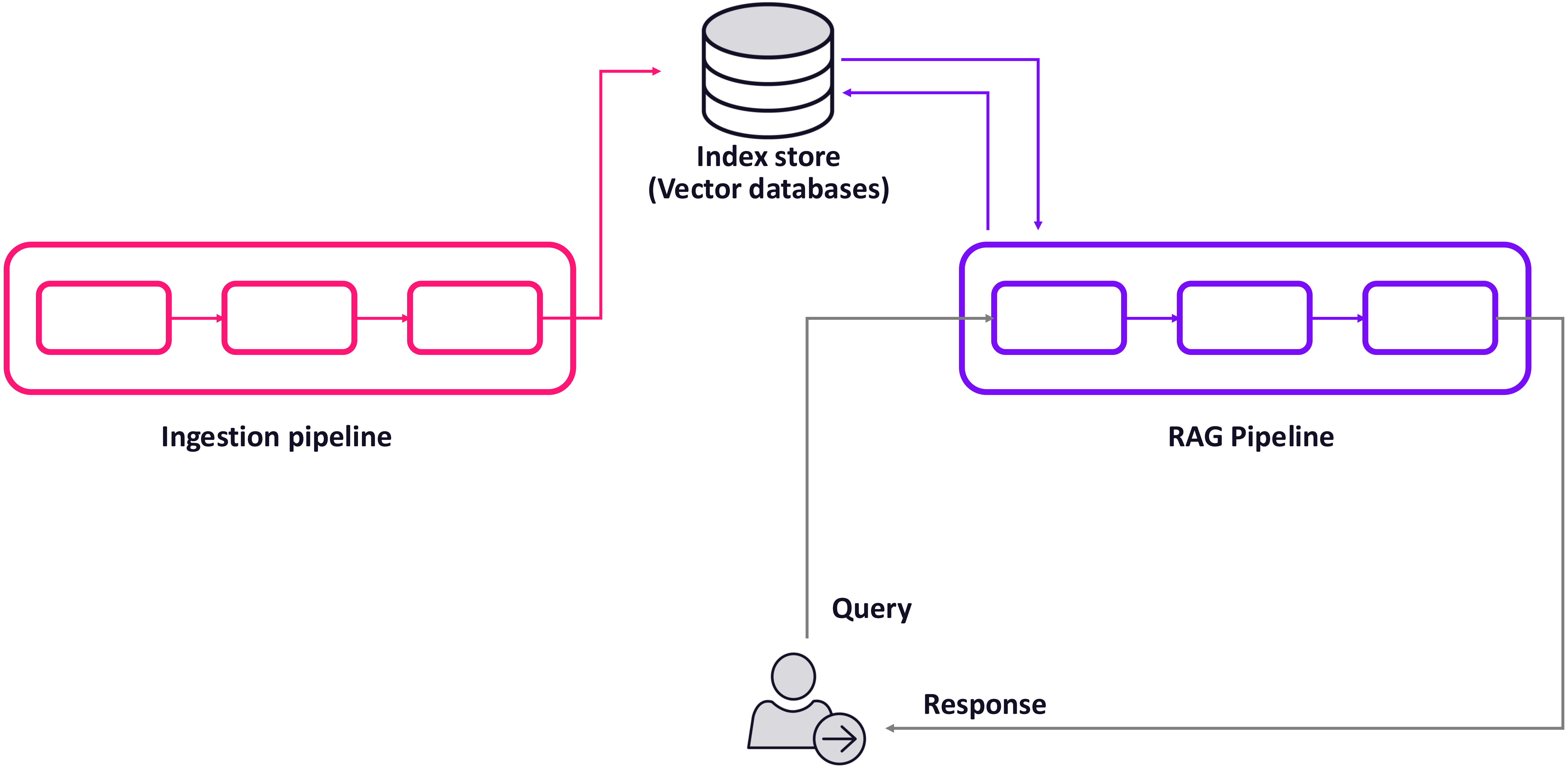
Include your private and recent data

Low LLM cost based on usage

Grounding and reducing hallucination

Source citations and attribution

RAG System


Index store
(Vector databases)
Ingestion pipeline
RAG Pipeline
Query
Response

Many Choices to Be Made

Parsing and chunking strategy

Embedding model and configuration

Retrieval techniques

Top N chunks

LLM and associated parameters

Prompt template

Building all RAG system components from scratch can be tedious.

RAG Integration Frameworks

Fasttrack RAG system development

Modular components and pre-configured chains

Adapt as per your needs

Open source

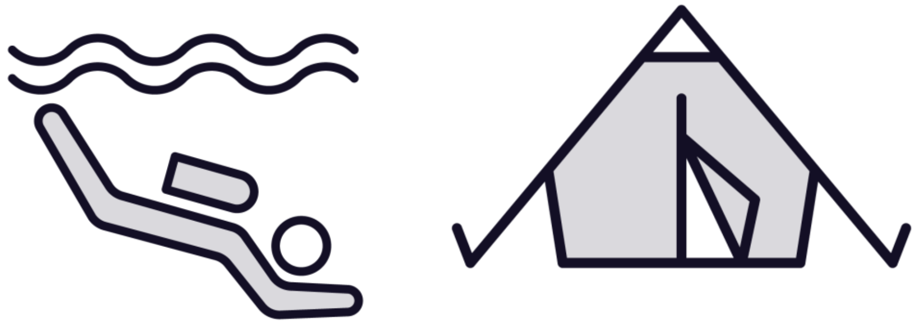
- LlamaIndex
- Langchain

Demo

Building simple RAG system

- Using LlamaIndex framework

Online Retailer

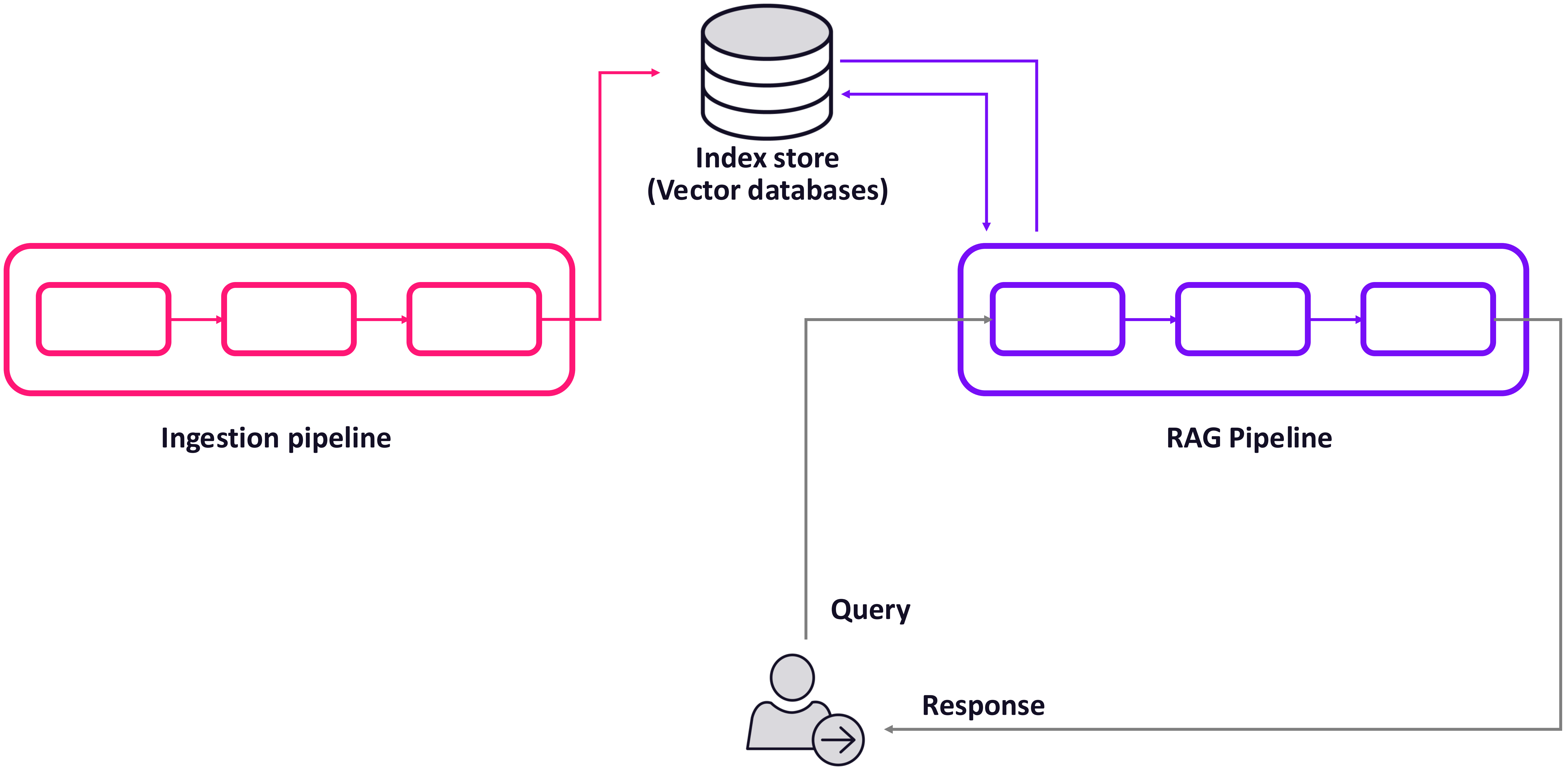


Foo
adventure gear company

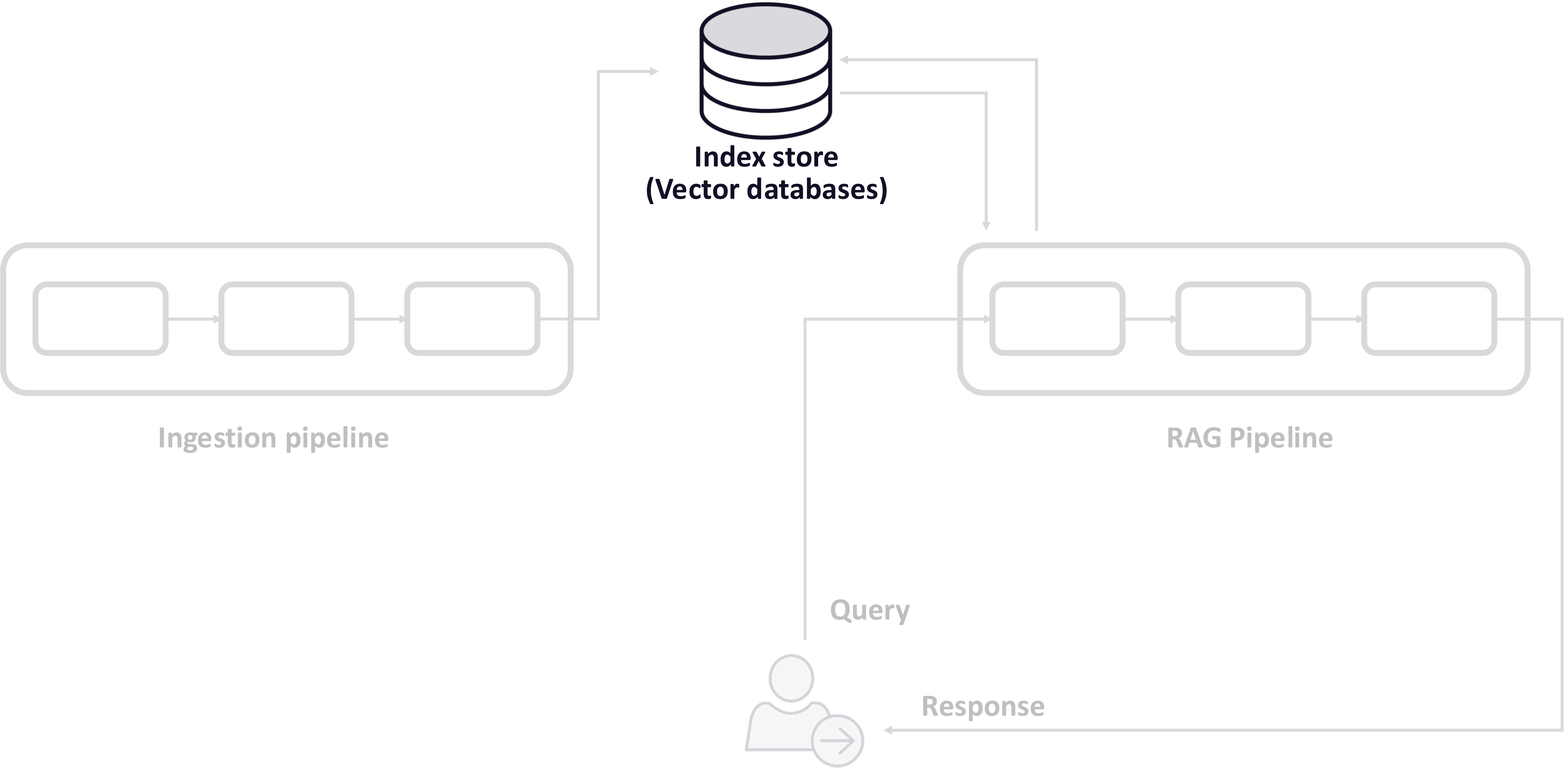
Bar

Qux

RAG System


Index store
(Vector databases)
Ingestion pipeline
RAG Pipeline
Query
Response

Deployment Blueprint - Vector Databases



Deployment Blueprint - Vector Databases

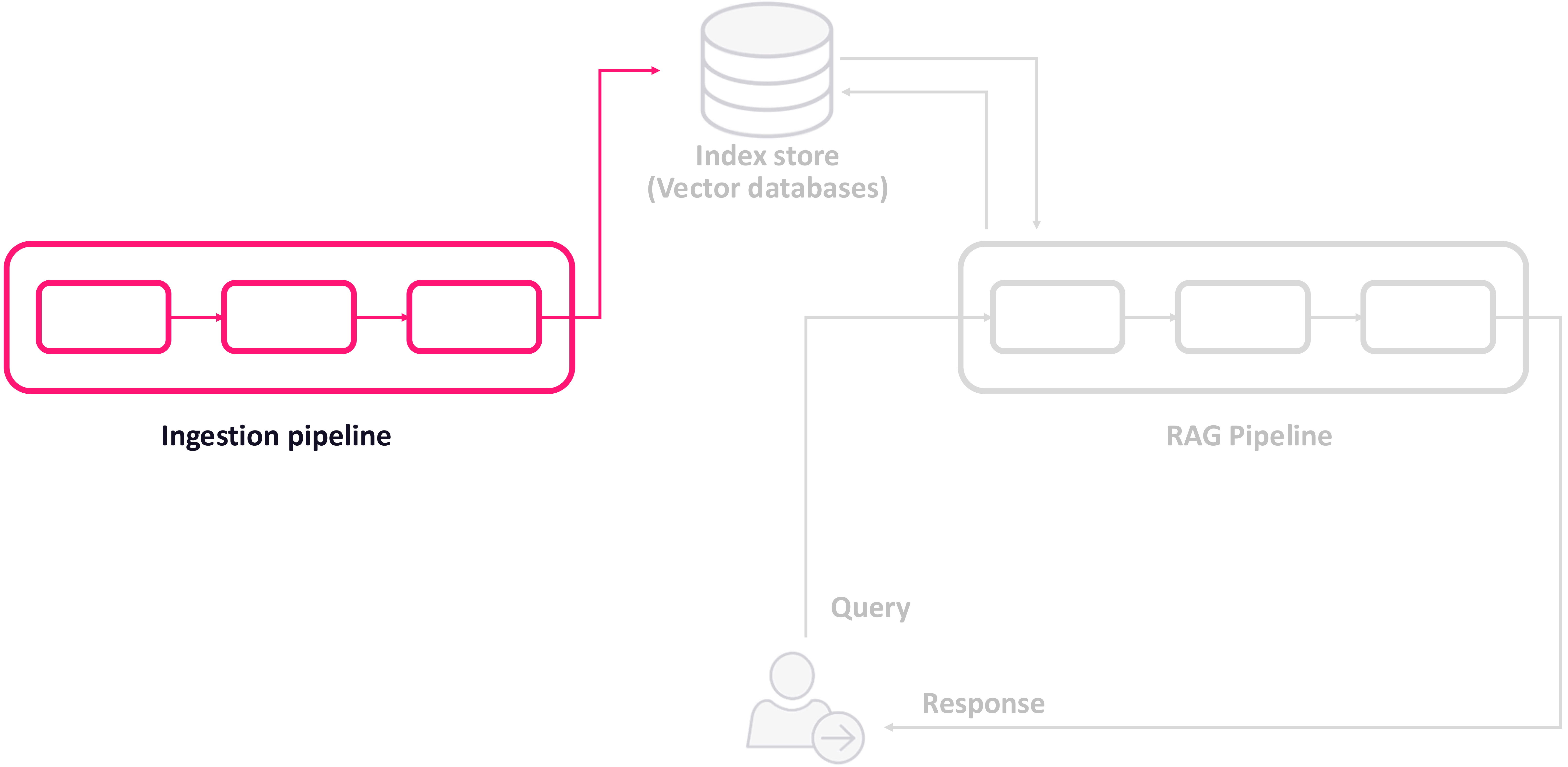
Managed databases

- Weaviate, Pinecone
- Cloud provider's offerings such as GCP vector search

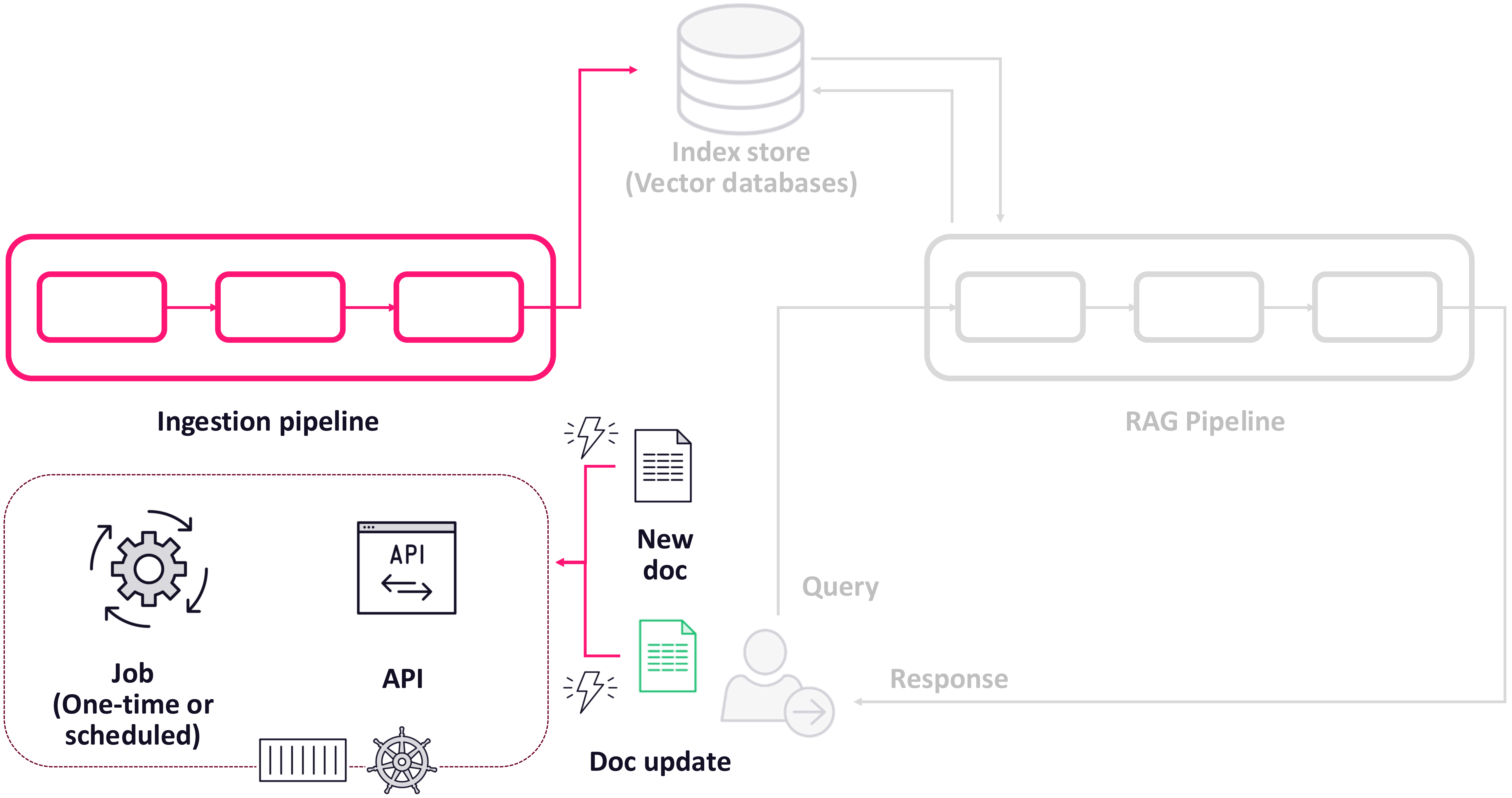
Self-hosted and managed database

- Open-source database such as Chroma, Milvus, Elasticsearch
- Containers with persistent volumes
- Container orchestration with Kubernetes

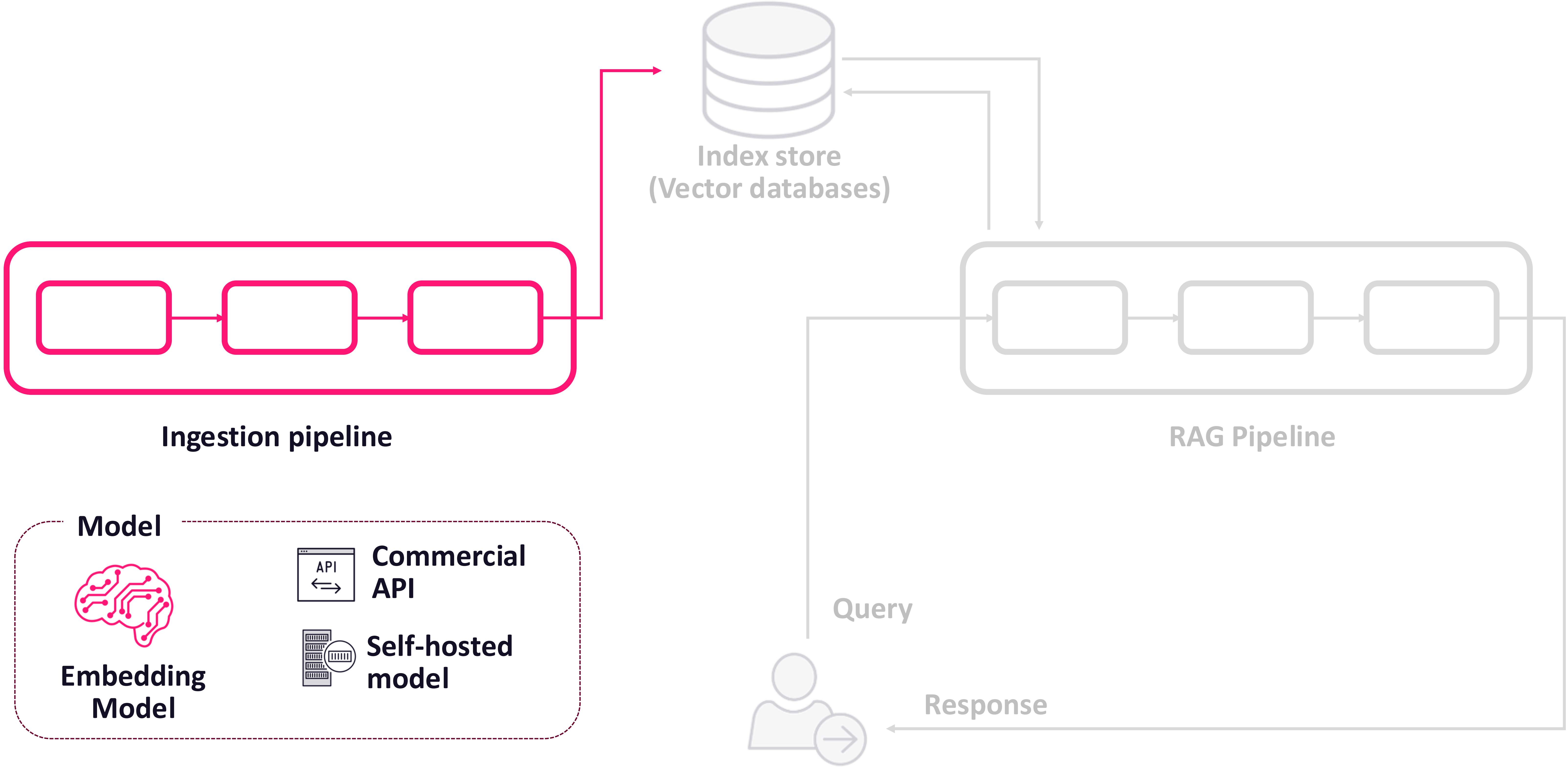
RAG System - Ingestion Pipeline



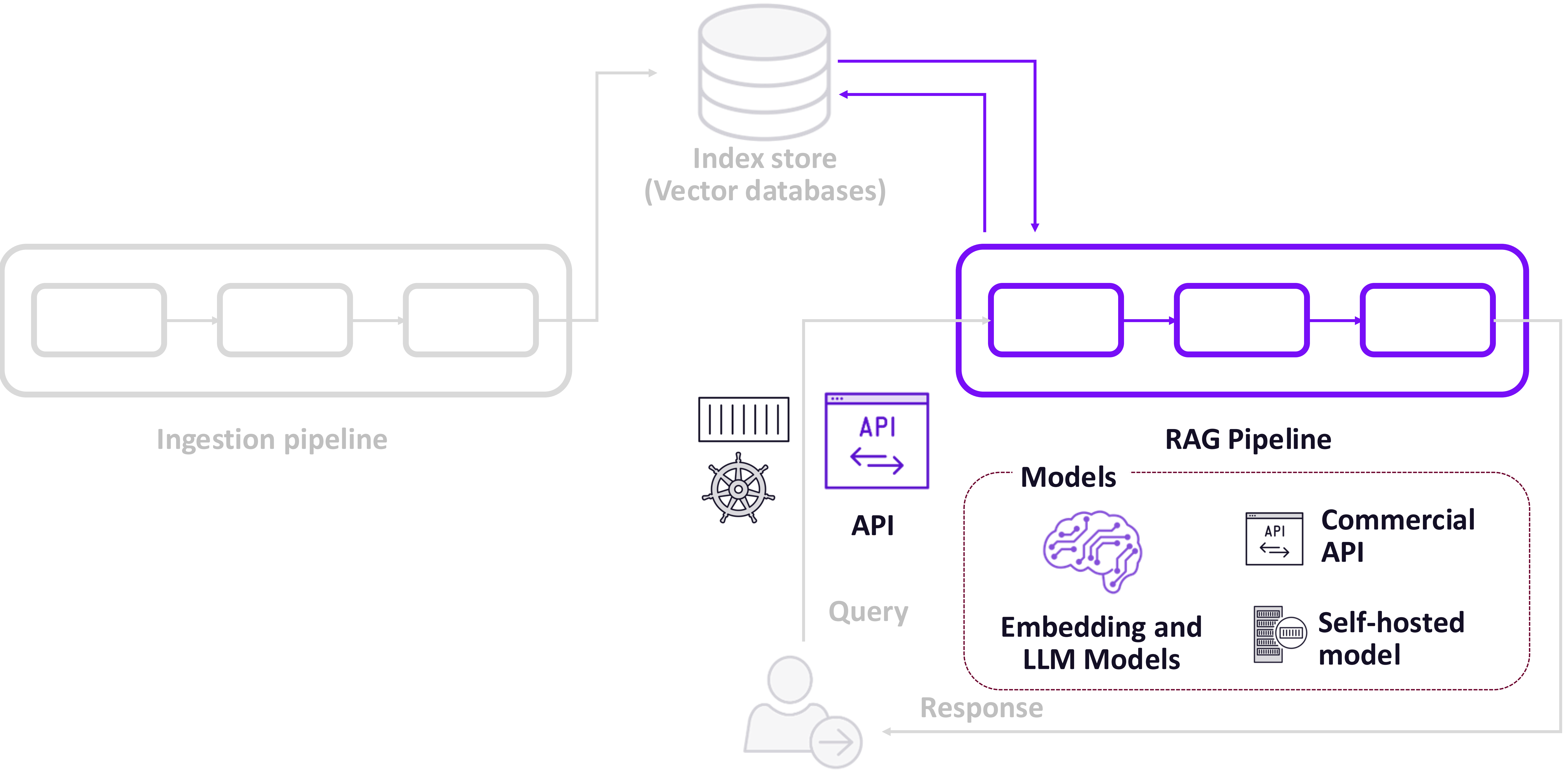
Deployment Blueprint - Ingestion Pipeline



Deployment Blueprint - Ingestion Pipeline



RAG System - RAG Pipeline



Up Next:

Managing RAG Systems