Tiger Analytics

Azure OpenAI Build -RAG

Azure OpenAI Workshop

Confidential
Date: 20-March-2024
Author: Tiger Analytics

Tiger Analytics' coaches introduction - Malaysia AI build

Srikanth Sripada

Lead Data Scientist

5.5 years of experience in leading multiple teams and providing analytical solutions to business problems in QSR, Retail CPG, IoT, and Insurance industries

Pushvinder Rohtagi

Senior Data Scientist

Python developer with knowledge of machine learning ,Deep Learning and NLP. Splunk User certified

Abin Joseph

Senior Machine Learning Engineer

Python developer with 3 years of experience AWS, Microservices and ETL. Developed scalable cloud based solutions for advanced data analytics. Strong working knowledge in Spark, Docker and Linux.

Vinay Kumar Soni

Senior Analyst

Data Science specialist with 3+ years of experience with interests in NLP & Python.

Rahul Mehta

Tech Presales Manager

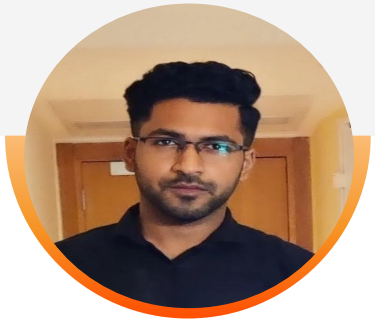
Data Engineering, Microsoft Infrastructure and Cloud professional with over 10+ years of experience. Passionate about delivering successful projects, fostering client relationships, and driving business growth.

Tiger Analytics' coaches introduction - Indonesia AI build

Ankita Malarya

Lead Data Scientist

Data Scientist with ~5 years of industry experience in leading projects, providing end to end scalable analytical solutions on business and research problem and meeting needs of the client across sectors including insurance, CPG & retail.



Venkatesh Das

Senior Data Scientist

5+ years of experience of working on cutting-edge Generative AI solutions using state-of-the-art Transformer architectures to address critical business challenges with significant impact.

Abdul Shamgarhwala

Lead Machine Learning Engineer

Technology Enthusiast with 8+ years of software development experience. Expertise in applied machine learning, model deployment, MLOps, LLMOps and ML related data engineering roles.

Manikandan Kannibabu

Associate Director • Partnership & Alliances

17 years of experience in Sales and Account Management of IT Solutions and services and Public Cloud Services. Focus on owning the enterprise customer conversation while building channels to drive run-rate business.

Deepak Sharma

Senior Analyst

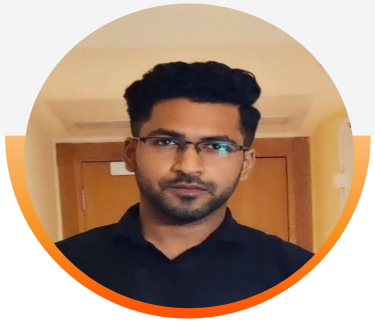
Data Science specialist with 4+ years of experience with interests in NLP & Python.

Tiger Analytics' coaches introduction - Thailand AI build

Ankita Malarya

Lead Data Scientist

Data Scientist with ~5 years of industry experience in leading projects, providing end to end scalable analytical solutions on business and research problem and meeting needs of the client across sectors including insurance, CPG & retail.



Venkatesh Das

Senior Data Scientist

5+ years of experience of working on cutting-edge Generative AI solutions using state-of-the-art Transformer architectures to address critical business challenges with significant impact.

Abdul Shamgarhwala

Lead Machine Learning Engineer

Technology Enthusiast with 8+ years of software development experience. Expertise in applied machine learning, model deployment, MLOps, LLMOps and ML related data engineering roles.

Manikandan Kannibabu

Associate Director • Partnership & Alliances

17 years of experience in Sales and Account Management of IT Solutions and services and Public Cloud Services. Focus on owning the enterprise customer conversation while building channels to drive run-rate business.

Deepak Sharma

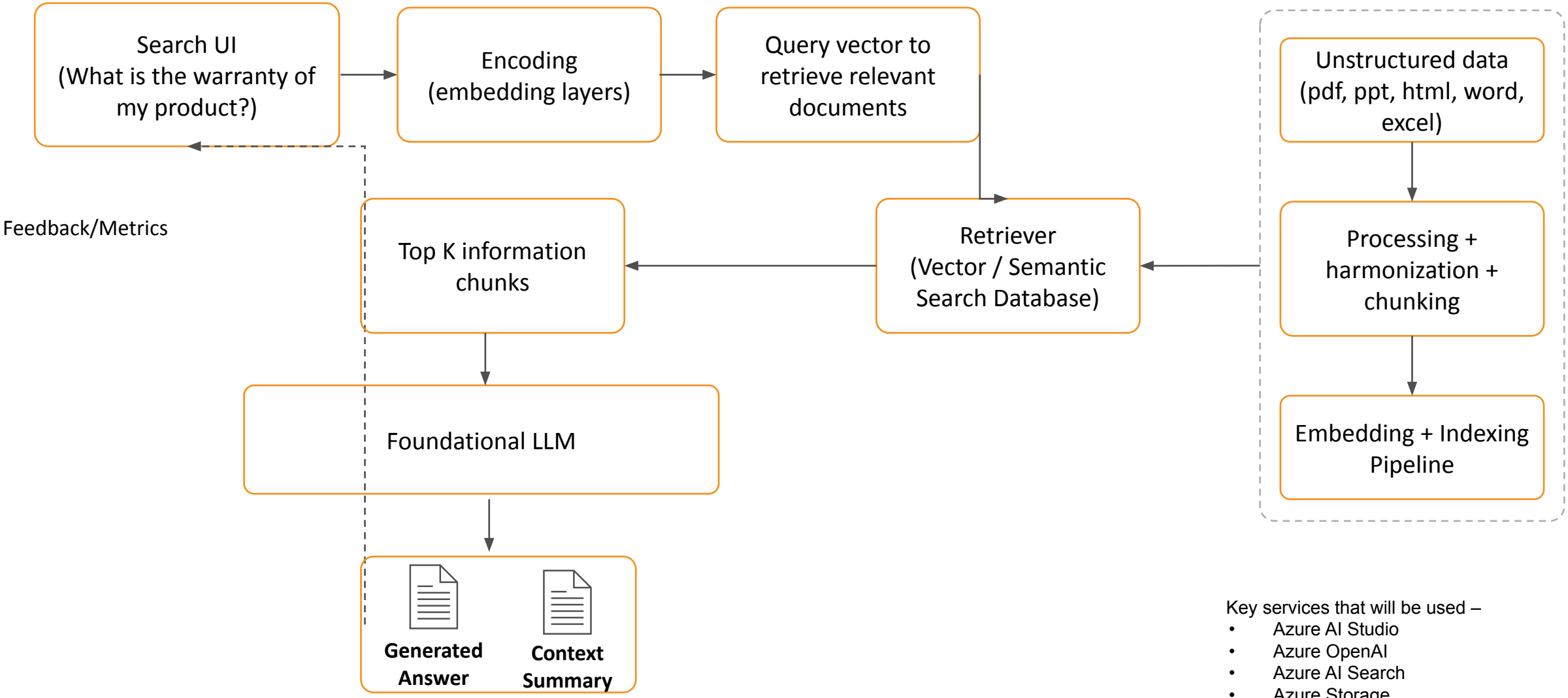
Senior Analyst

Data Science specialist with 4+ years of experience with interests in NLP & Python.

AI Build Use Case Introduction

Search Summarization – RAG – Chat on your own data
Key Components – Prompt Engineering, Chunking, Embedding & Indexing

RAG – Solution Approach



Key services that will be used –

- Azure AI Studio
- Azure OpenAI
- Azure AI Search
- Azure Storage

Prompt Engineering | Structure

System Role : You are an AI assistant that helps people find information. → Persona Definition

User: \nProvide crisp answer in 1 to 7 bullet points for the question based on the given context truthfully. Please use neutral and professional language. If you are unable to, please say 'I don't know' as one bullet. → Clear Instruction, Desired Output Format

\n\nquestion: <user_query> → User Query

\n\ncontext: <context> → Business Context

\nFollow the below mentioned guidelines while generating response:
\nGuideline 1:
\nGuideline 2: → Additional Guidelines

\n\nanswer:\n-" → Answer Key

Prompt Examples

- Text Summarization
- Question Answering
- Content Generation

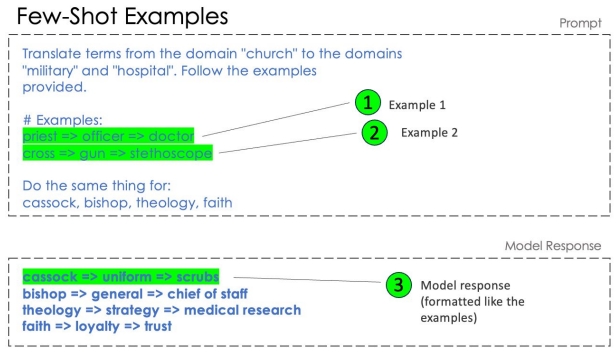
Prompt Engineering | Techniques

Zero Shot Prompting

Simplest which generates response without any examples

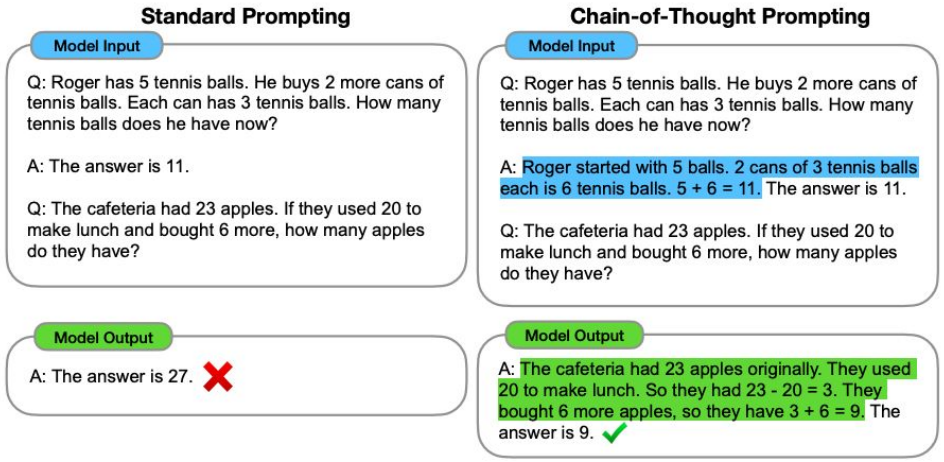
Few Shot Prompting

Provide examples to generate response



Chain of Thought

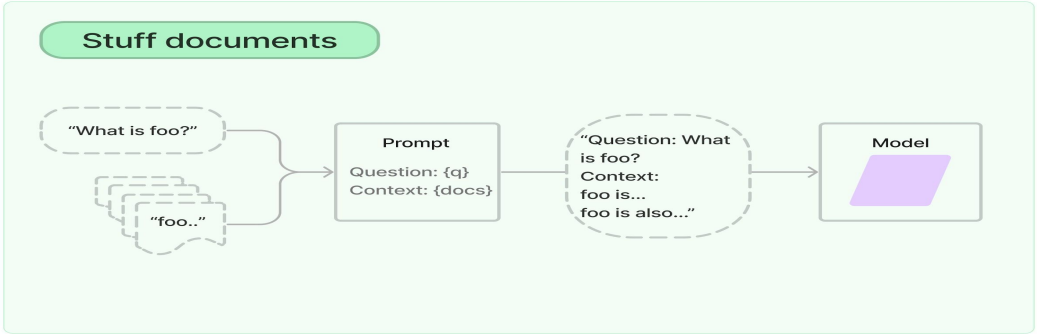
Uses multi-step reasoning ability by thinking step by step using the provided instruction or examples



Prompt Engineering | Techniques

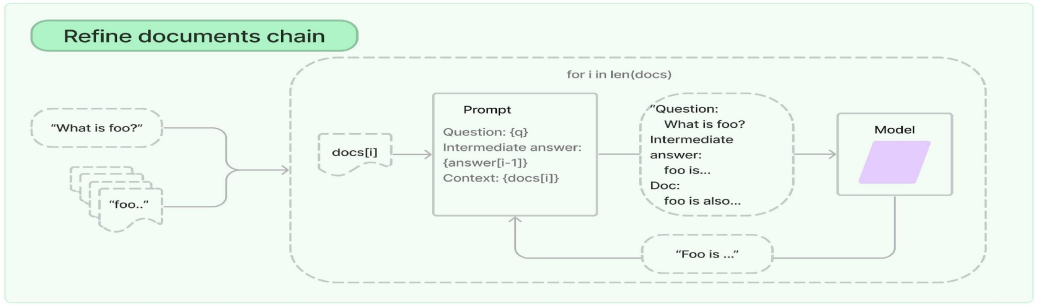
Stuff Documents Chain

Simplest combines all context.



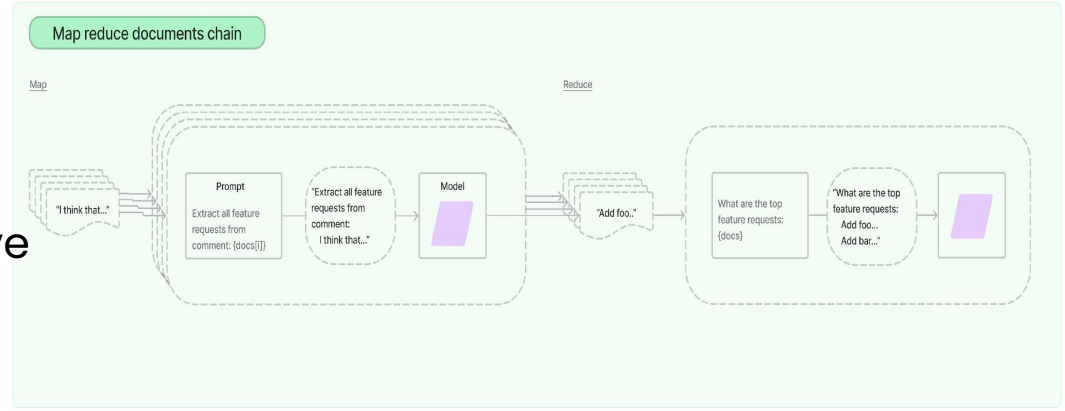
Refine Documents Chain

Uses second prompt to refine the first one.

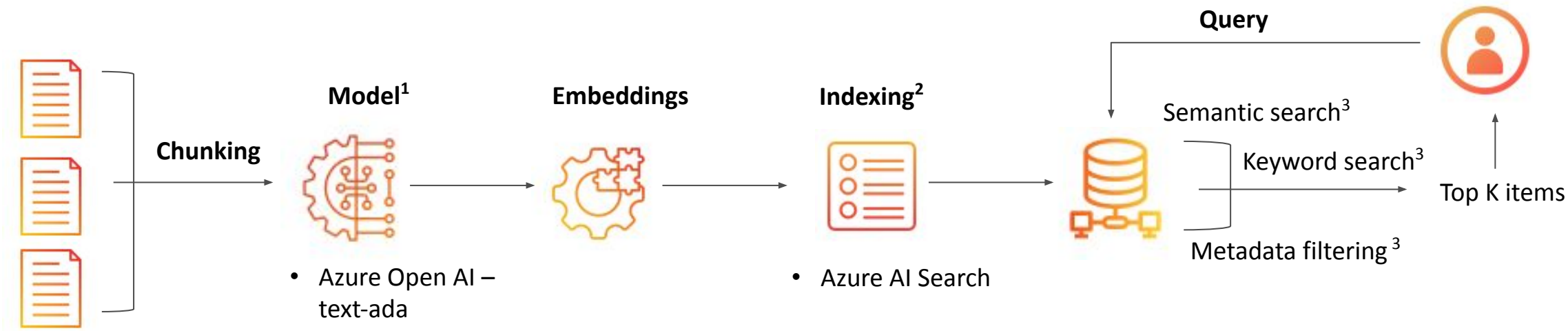


Map reduce Documents Chain

Uses multiple prompts to combine and reduce to effective context.



Vector Search



[1] Some of the vector databases offer generating vector embeddings. But it is generally better to generate the embeddings separately to avoid overloading the database.

[2] Calculating the distance between the query and each of the documents is computationally intensive. A variety of tree, hashing and graph-based indexing/search strategies exist to solve this Approximate Nearest Neighbor (ANN) problem.

[3] Hybrid search (combination of semantic search and keyword search) and metadata filtering are allowed in most of the vector databases to improve the relevance of the search results to the user.

Setup Checklist

- Wifi Connection
- Laptop Charge Connection
- Github Link Check
- Username & Login Check – Received and Able to login
- Subscription Check
- Azure OpenAI Check

AI Studio

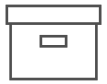
Azure OpenAI Hackathon

Hands-on Exercise

Appendix

Search Summarization | Generative AI

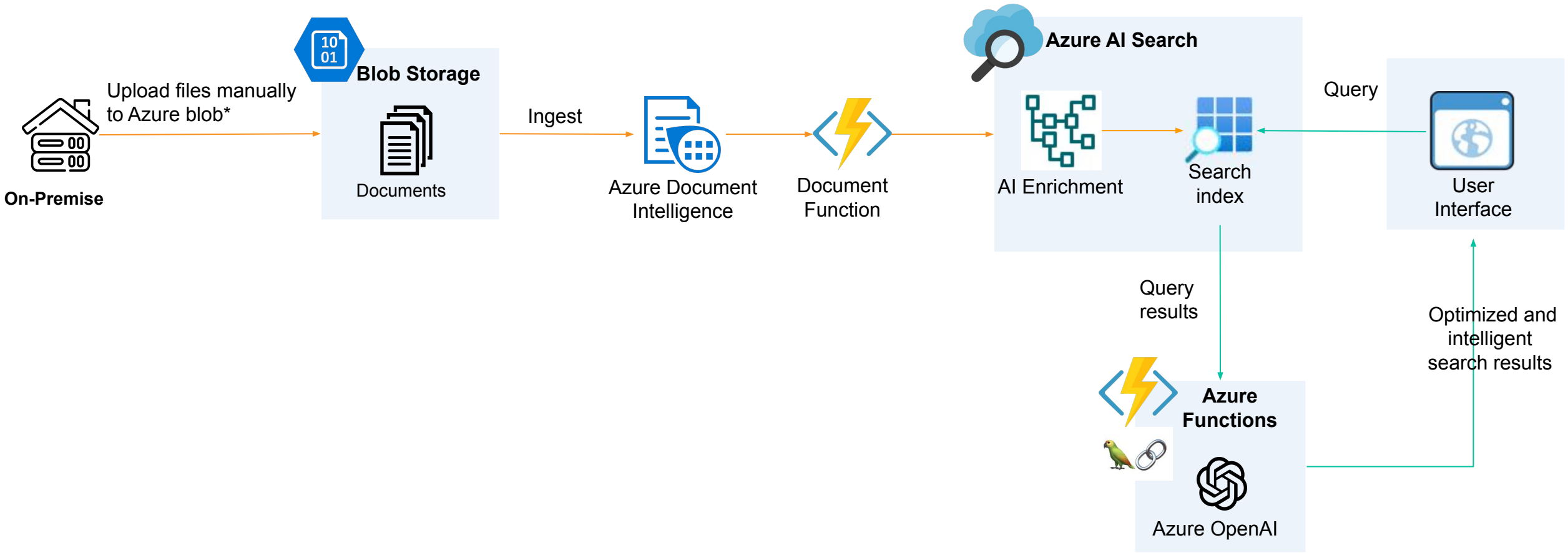
An illustrative pipeline with steps involved to solve a search and summarization problem.


Document Parsing	Document Store	Retriever	Question Answer	Summarization
• **Azure Form Recognizer**	• **Azure Cognitive Search**	• **text-embedding-ada-002**	• BERT	• **GPT-4**
• LayoutLM	• Pinecone	• BM25	• RoBERTa	• **GPT-3.5**
• PDFPlumber	• ElasticSearch	• TF-IDF	• ALBERT	• T5
• Camelot	• In Memory	• Dense Passage Retrieval	• DistilBERT	• BART
• Tabula	• Open Search	• Sentence Transformer	• **LLM Models**	• Pegasus
• PyPDF2	• Graph Knowledge Base	• LLM Embeddings		• GPT-3
	• FIASS Knowledge Base			• BLOOM

** LLM Models – Enterprise Models (GPT3.5, GPT4, Claude, Titan, PaLM), Open Source Models (Dolly, LLaMa, Falcon) etc.*

Azure – Reference Architecture Diagram



* *file transfer from on-premise to blob can be automated in the future stages.*
Not a final architecture, subject to change.

PDF Parsing

Most of the documents are a mix of text and tables with tables being responsible for a majority of the answers

Challenges with parsing PDFs

PDFs contain data in varying presentation formats that need to parsed accurately

Integration issues when using multiple packages

Text

- General text
- Multi-column text
- Headings
- Bullets points
- Headers / footers

Tables

- Multiple lines of text
- Merged cells
- Borderless tables / missing borders
- Images
- Tick marks

Approach to resolve issues

Evaluation of multiple packages

Text

- Document Intelligence (formerly Azure Form Recognizer)
- Pymupdfloader
- Pypdfloader
- Unstructuredpdfloader
- Pdfplumberloader

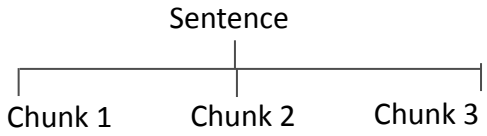
Tables

- Document Intelligence (formerly Azure Form Recognizer)
- Camelot: Lattice
- Camelot: Stream
- Tabula

Chunking

What is Chunking?

The process of breaking the text up in smaller and more mangeable pieces



Need for Chunking

- Large parsed outputs not fitting LLM's token limit
- Allows greater summarization
- Allows for logical recall points

Chunking Size & Overlapping

- Amount of business info required to answer the query
- Amount of context required from previous chunks
- Token limit of LLM used

Chunking Methods - Tools

Tools	How text is split
Recursive TextSplitter	By list of characters
Tiktoken	By character passed in
NLTK	By NLTK tokenizer
Spacy TextSplitter	By SpaCy tokenizer
Hugging Face Tokenizer	By character passed in

Custom Chunking

- Based on business objective & data (if you need custom changes)
- Example – If you don't want to break the slide content while creating chunks for PPT

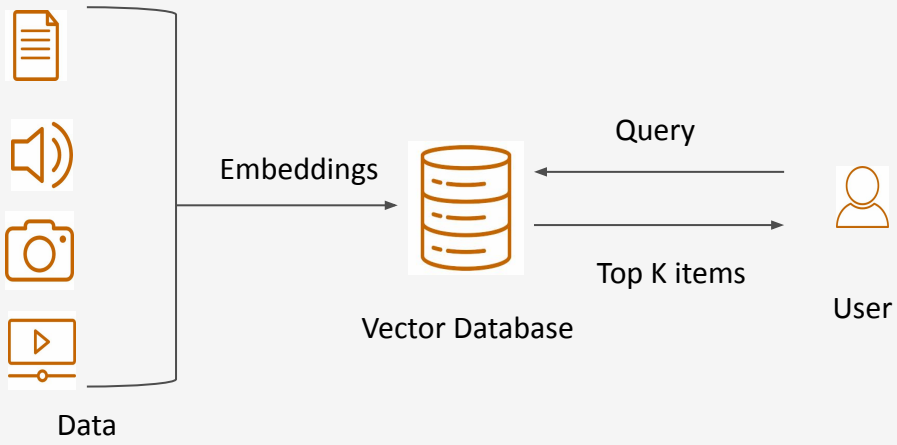
Vector Database

Introduction

Specialized storage systems that store data in the form of vector embeddings

Enable efficient semantic search and retrieval of data based on vector distance

The data can be text, image, audio, video etc.



Usecases

Document Retrieval

- Better results compared to traditional keyword search

Generative AI

- Provide context to the LLM to improve response
- Retrieve relevant chat history to improve long-term memory of LLMs

Product Recommendation

- Improves discoverability of products

Vector Database: Landscape

There are a lot of vector databases in the market today and is an evolving space. They can be categorized based on:

Source code availability

- Majority are open source
- Cost - open source vs cloud services

Hosting methods

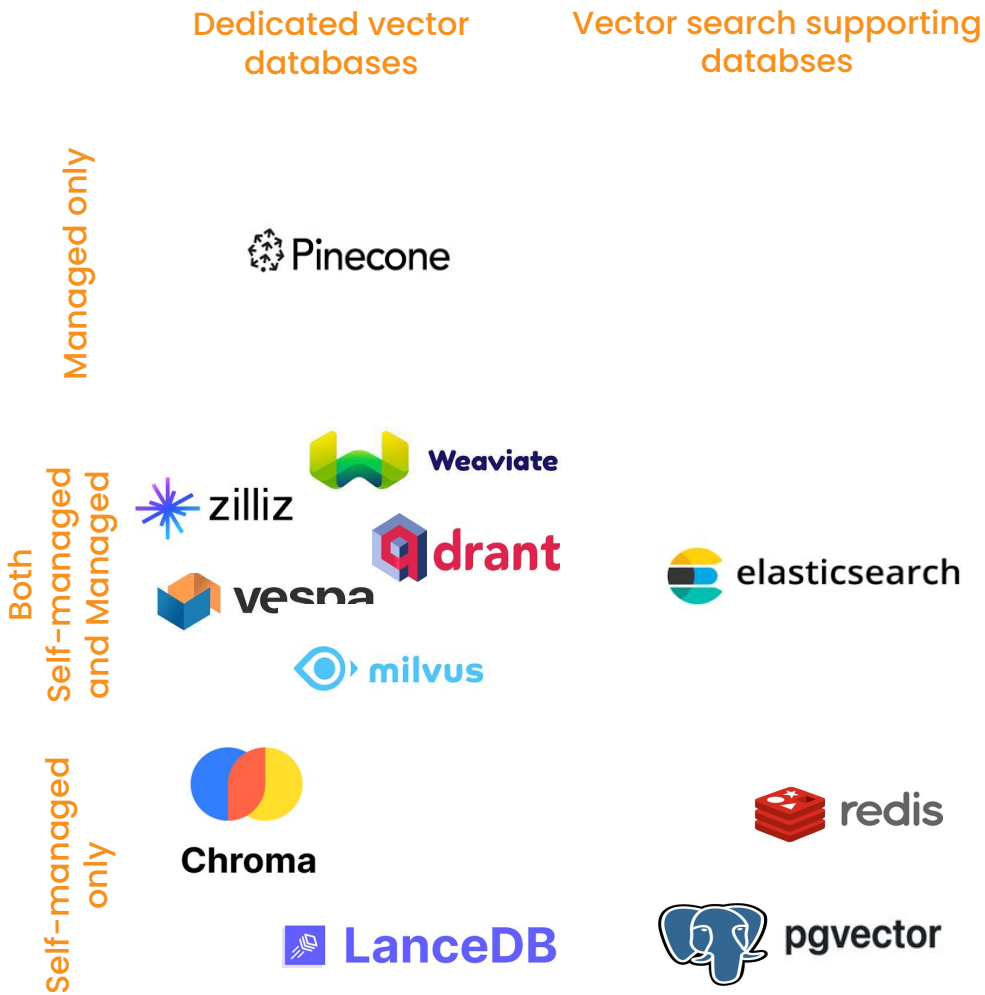
- Majority of the vendors offer both self-managed and managed versions
- Pinecone offers only managed version

Dedicated or supporting databases

- Cloud Marketplace availability for seamless deployment
- Multi-tenancy
- Multi-Authentication
- Data security
- Latencies
- Standalone vs Cluster Deployment

Support

- Embeddings:OpenAI Sentence Transformers, Cohere (multi-lingual), custom embedding, Aleph Alpha, Bert, huggingface
- Indexes:Flat index, HNSW, Locality Sensitive Hashing, Scalar quantizer, IVF
- Metrices: Cosine Similarity Dot Product Euclidean Distance, manhattan, IP (Inner product)
- Langchain/Llama Index implementation



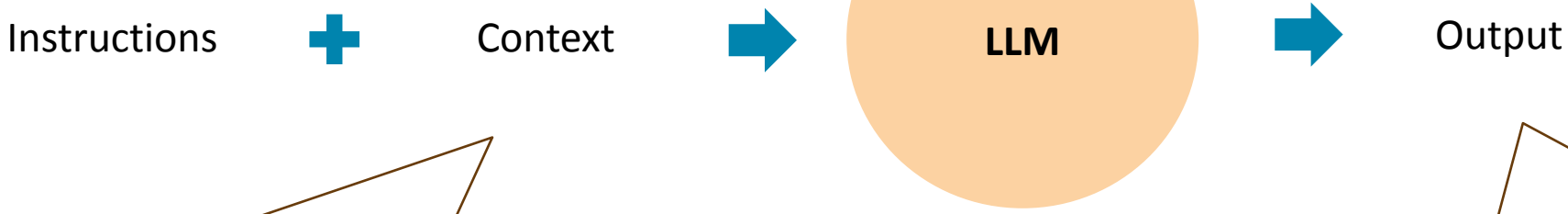
Source: https://thedataquarry.com/posts/vector-db-1/

Prompt Engineering | Guidelines

- Define a persona for the model
- Specify the instructions clearly
- Follow a structured format
- Provide step by step instructions for complex tasks

Broad guidelines

- Start with a broad prompt and gradually provide specific instructions
- Use external tools where applicable
- Test changes systematically



- Provide relevant business context
- Provide any specific context for answering the question
- For conversational applications, pass previous (input, output) as context

- Specify required style, length or format of the output
- For classification tasks, present the model with a set of options to choose from

Parameters to Track

Hyperparameters

- Top k
- Top p
- Temperature
- Chunk Size and Overlap
- Completion Tokens
- Chain Type
- Prompts

Latencies

- Parsing Time
- Embedding, Indexing & Persisting Time
- Retrieval Time
- LLM Response Time
- UI Latency
- GPU / CPU Utilization

User UI

- Query
- LLM Response
- Retrieved Documents
- Retrieved Doc Relevancy Scores
- Metadata Filters
- User Feedback

Performance

- Rouge
- BleU
- Semantic Similarity
- Bleurt
- Toxicity
- Harmfulness
- Bias
- Recall@k
- MRR

Workflow - Guidelines

- Development
  - Test changes systematically
  - Create ground truth examples to evaluate prompt efficacy
  - Define high level KPIs for evaluation
  - Review model output for each individual ground truth example to understand specific issues
  - Keep track of token usage

- Deployment
  - Gather human/user feedback on actual user queries
  - Monitor similarity/dissimilarity of actual user queries compared to ground truth examples
  - Monitor application accuracy (overall, by segments of user queries, by similarity to ground truth examples)
  - Update ground truth examples if needed
  - Refine the prompt