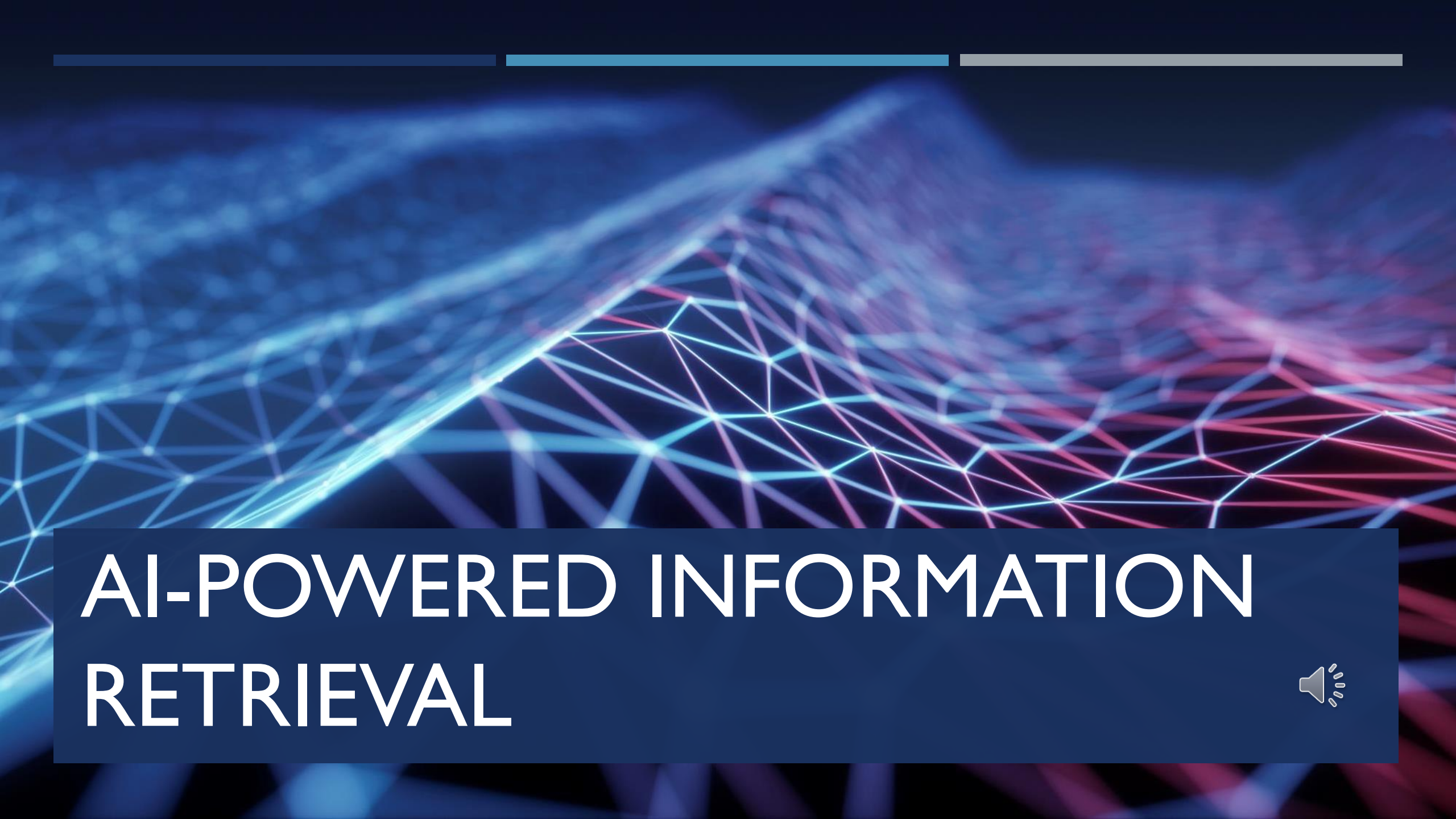

AI-POWERED INFORMATION
RETRIEVAL

CONTENT

- Understanding the Basics of LLMs
- Working with Embeddings and Vectors
- Storing and Searching Vectorized Data
- Retrieval-Augmented Generation (RAG)
- Using LangChain for AI Workflows
- Optimizing AI Model Performance

WHAT ARE LLMs ?

- Advanced AI systems
- Generate human-like text
- Trained on vast textual data

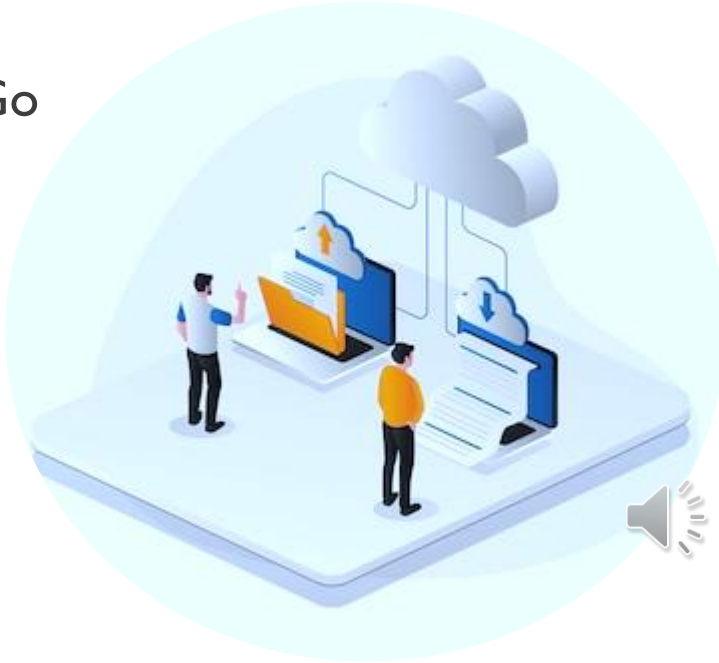
LOCAL VS. CLOUD LLM

Local LLM

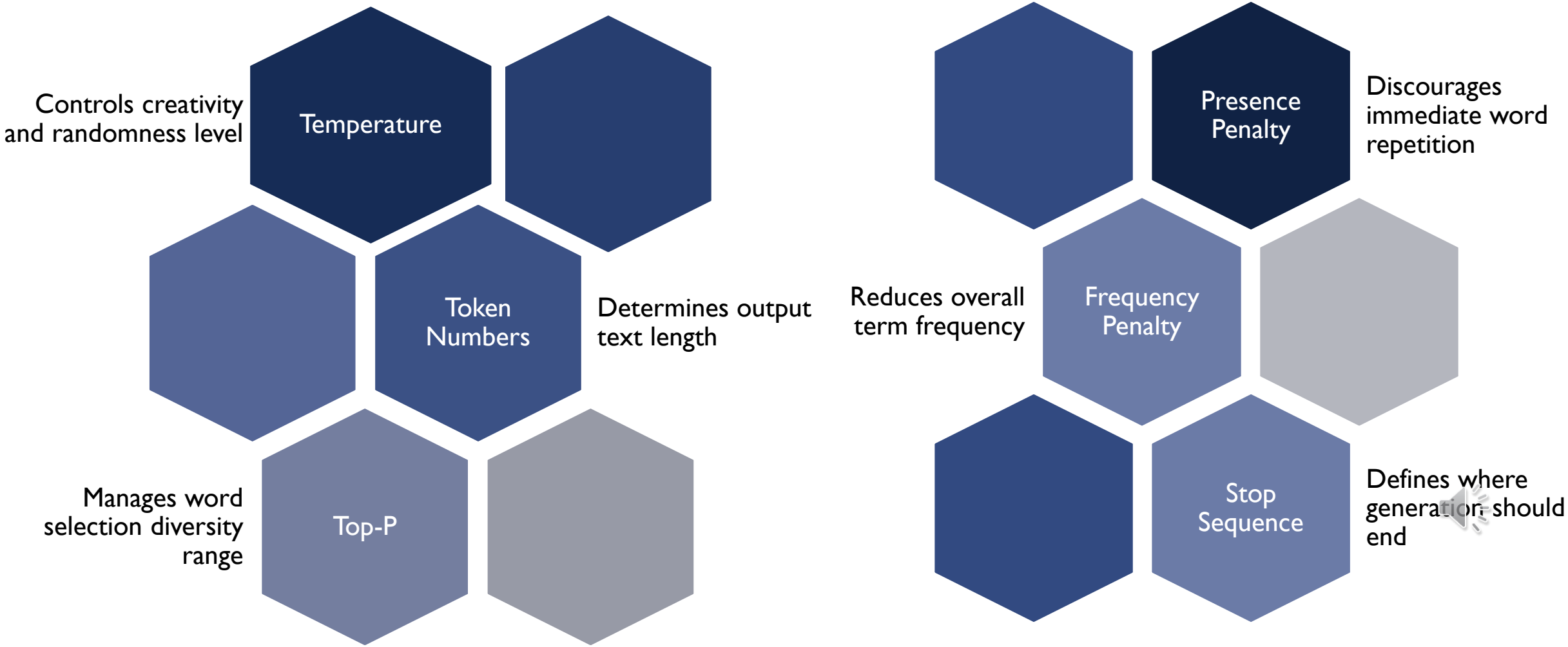
- High Privacy
- Full Control
- Higher Costs

Cloud LLM

- Easy Access
- Scalable
- Pay-As-You-Go



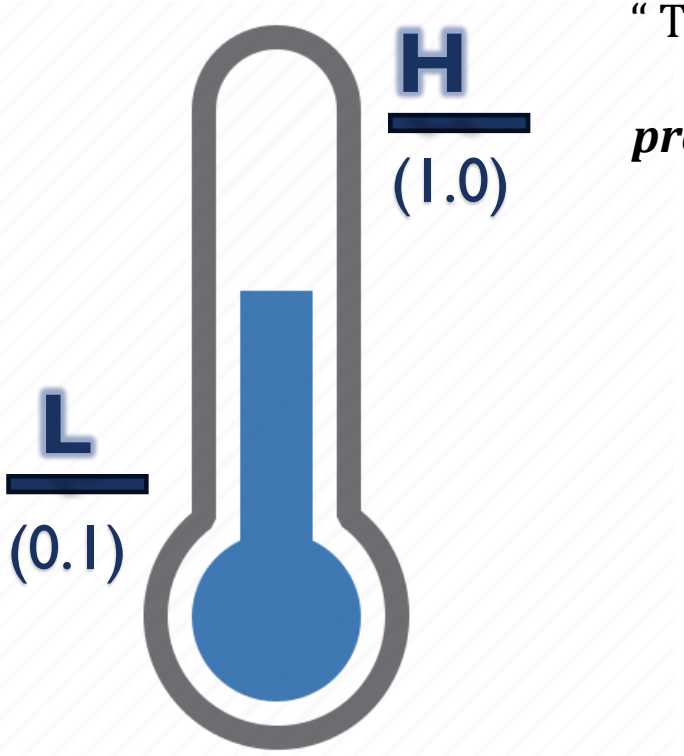
LLM PARAMETERS



TEMPERATURE IN LLM

- **The best way to learn coding is?**

" The best way to learn to code is ***to practice a lot and follow online tutorials.*** "



" The best way to learn coding is ***to go back in time and meet the programming language inventors.*** "

CONTEXT WINDOW AND TOKEN LIMITS

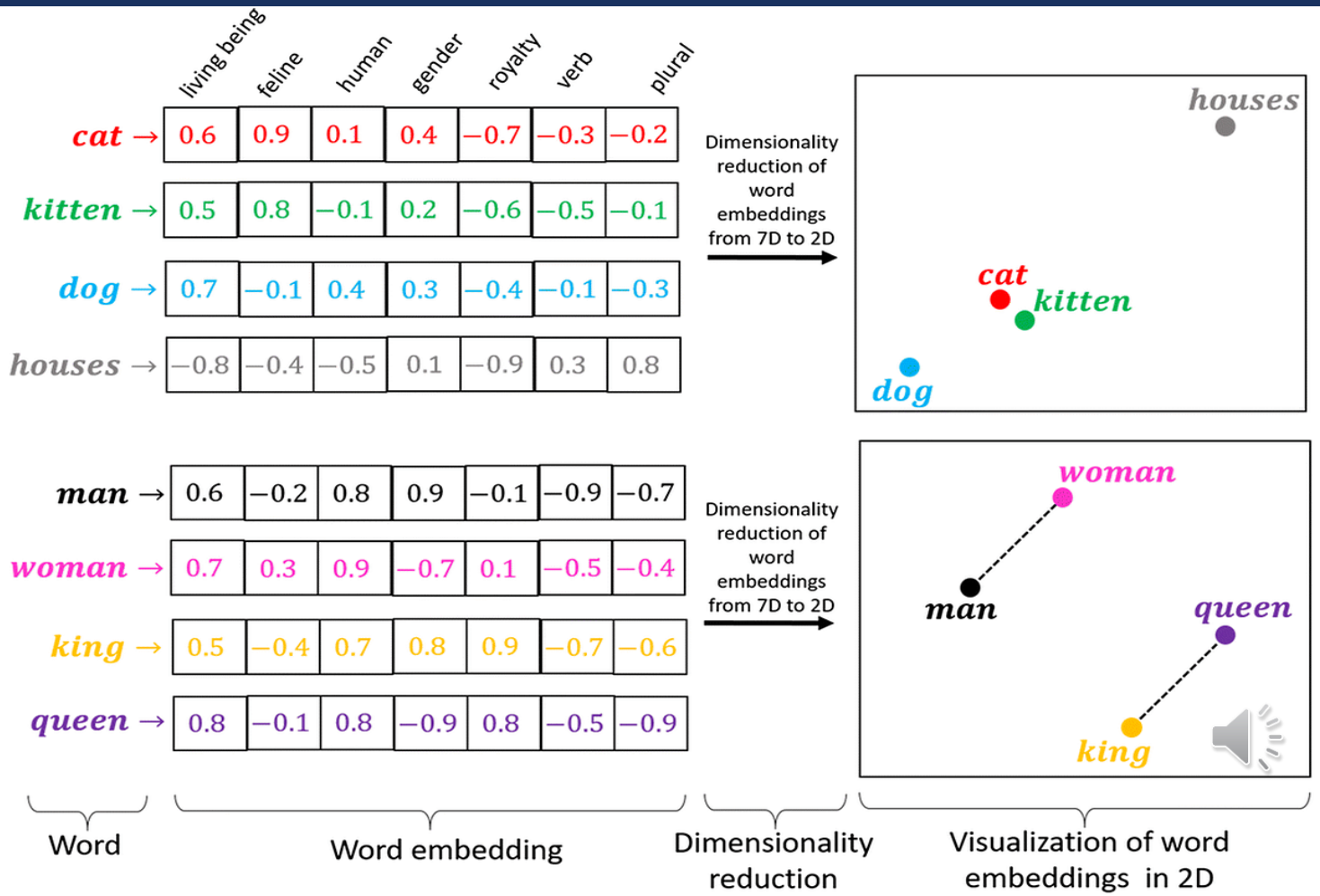
- Token Limit
- Input Token Limit
- Output Token Limit

CONTENT

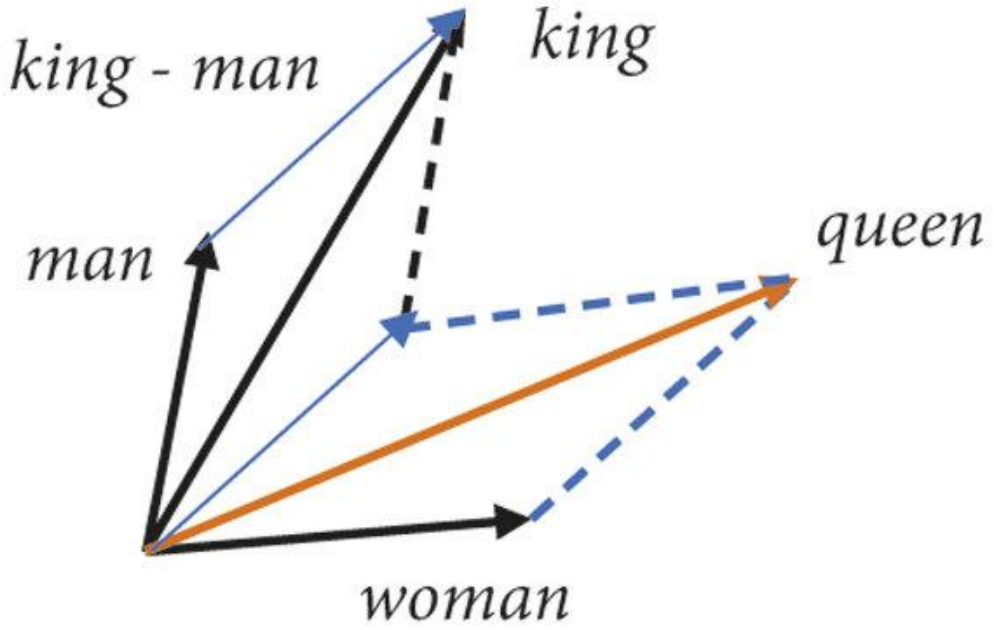
- Understanding the Basics of LLMs
- Working with Embeddings and Vectors
- Storing and Searching Vectorized Data
- Retrieval-Augmented Generation (RAG)
- Using LangChain for AI Workflows
- Optimizing AI Model Performance

VECTOR EMBEDDING

- Converting data into mathematical space



HOW DOES VECTOR EMBEDDING WORK?



$$king - man + woman \approx queen$$

Representation | Contextual Meaning | Training | Dimensionality Reduction | Usage in Models | Beyond Words

DIFFERENT TYPES OF EMBEDDING

Word Embeddings

Contextual Word Embeddings

Document Embeddings

Transformer-Based Embeddings

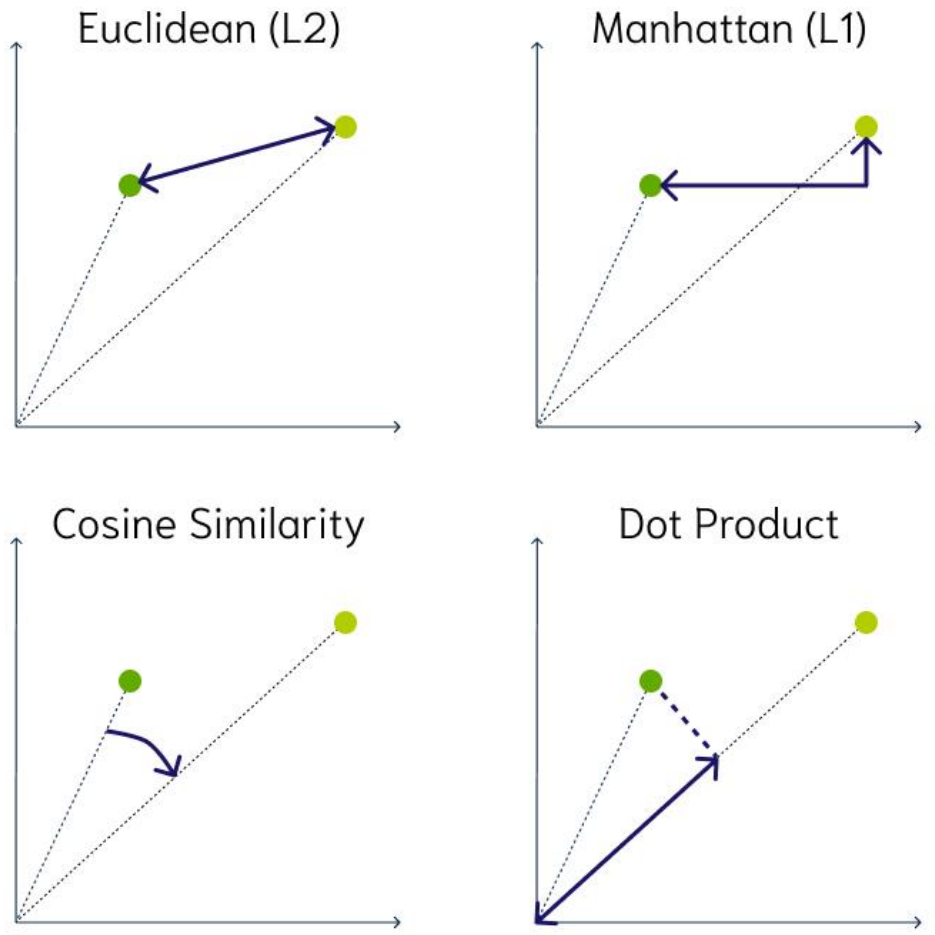
Graph Embeddings

Image Embeddings

Knowledge Graph Embeddings

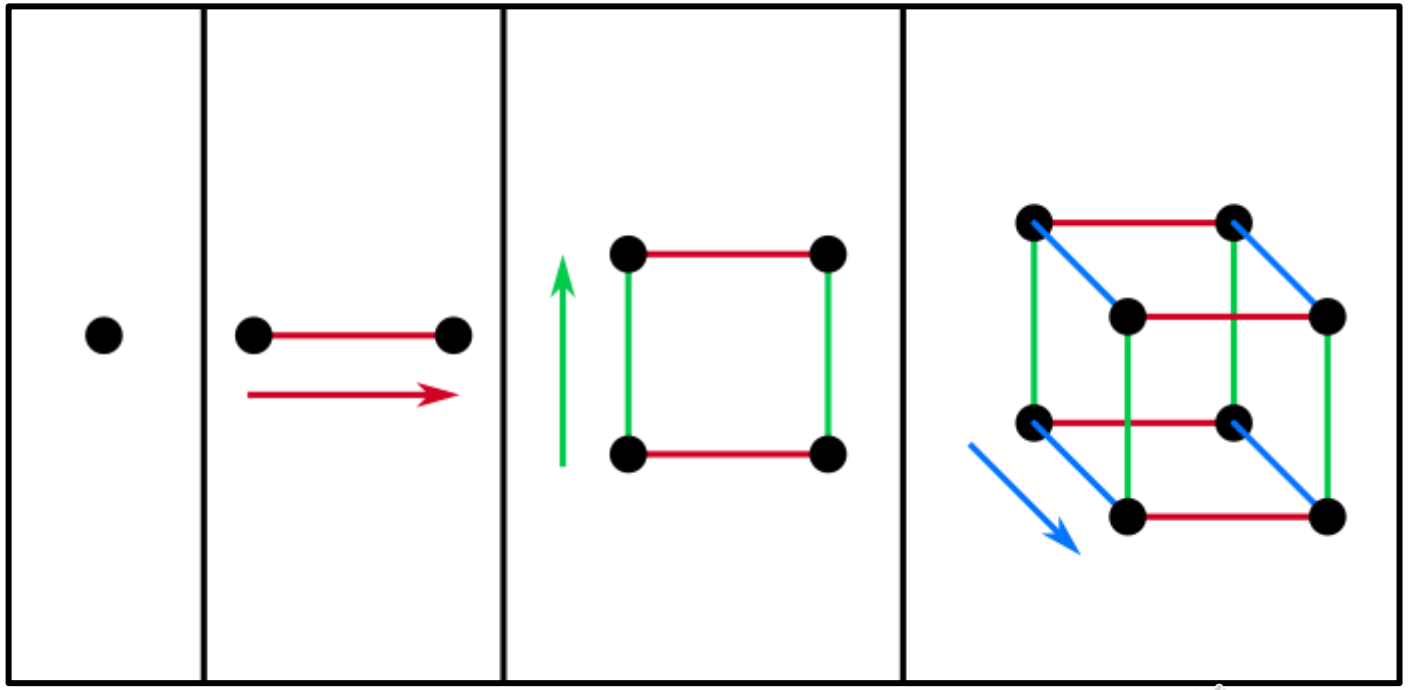
VECTOR SIMILARITY

- Dot Product
- Cosine Similarity
- Manhattan
- Euclidean Distance



VECTOR DIMENSIONALITY

- Each vector has N components
- Example: [age, height, weight] = 3D
- Higher dimensions = More features

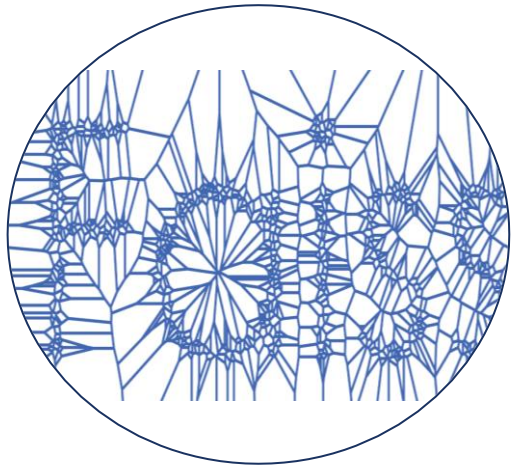


CONTENT

- Understanding the Basics of LLMs
- Working with Embeddings and Vectors
- Storing and Searching Vectorized Data
- Retrieval-Augmented Generation (RAG)
- Using LangChain for AI Workflows
- Optimizing AI Model Performance

VECTOR DATABASES

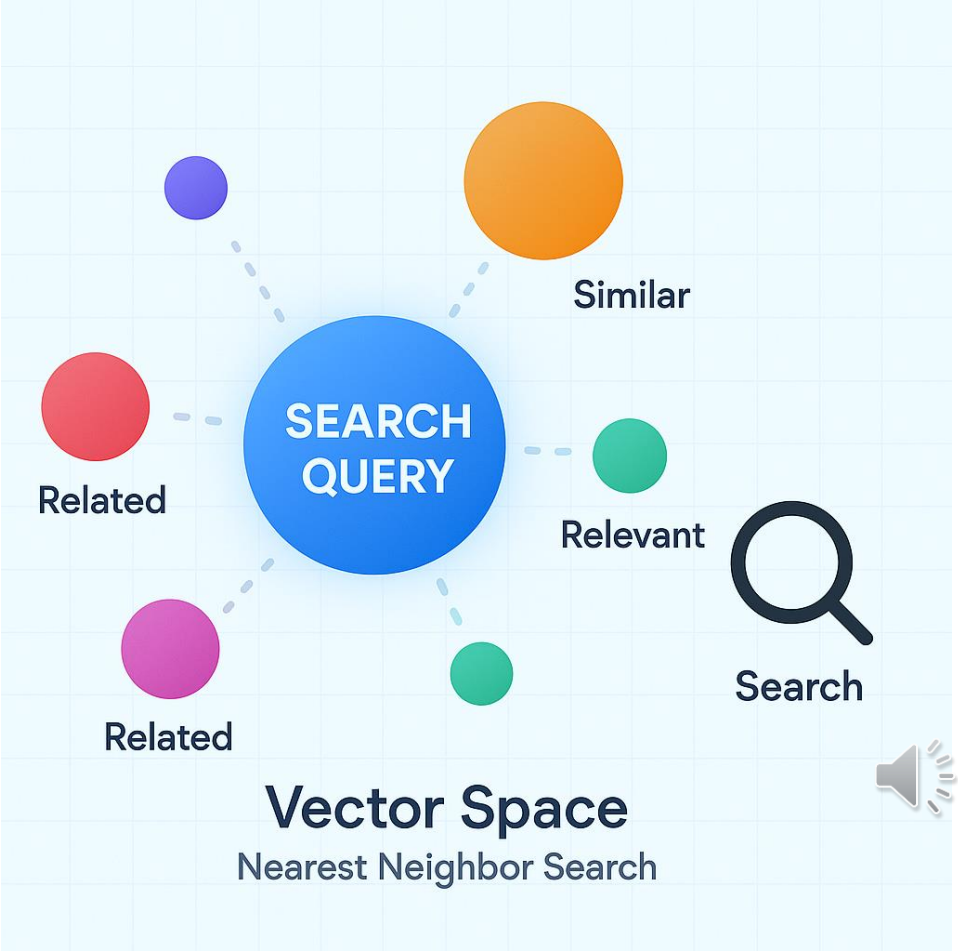
- Popular choices:
  - FAISS
  - Chroma DB
  - Pinecone



VECTOR SEARCH

- Converts data to numeric vectors
- Finds nearest neighbors
- Uses ANN for efficiency

- Real-world examples:
  - Netflix: "Similar shows"
  - E-commerce: "You may like"
  - Search: Understanding intent



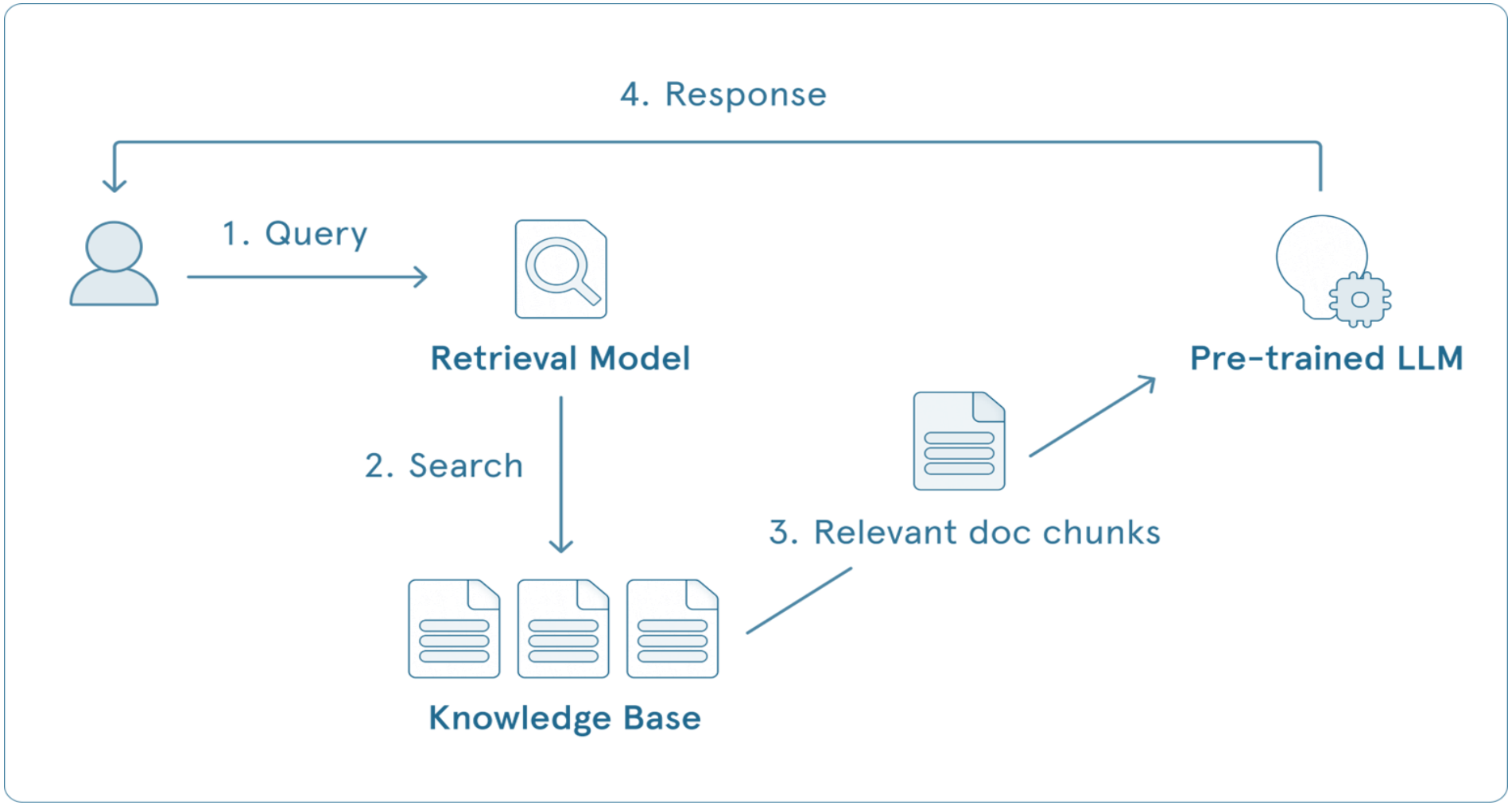
CONTENT

- Understanding the Basics of LLMs
- Working with Embeddings and Vectors
- Storing and Searching Vectorized Data
- Retrieval-Augmented Generation (RAG)
- Using LangChain for AI Workflows
- Optimizing AI Model Performance

RAG

- To create more accurate and informed AI responses RAG combines the power of:
  - Information Retrieval
  - Text Generation
- Key Benefits
  - Real-time information access
  - Up-to-date responses
  - Factual accuracy
  - Customizable knowledge sources

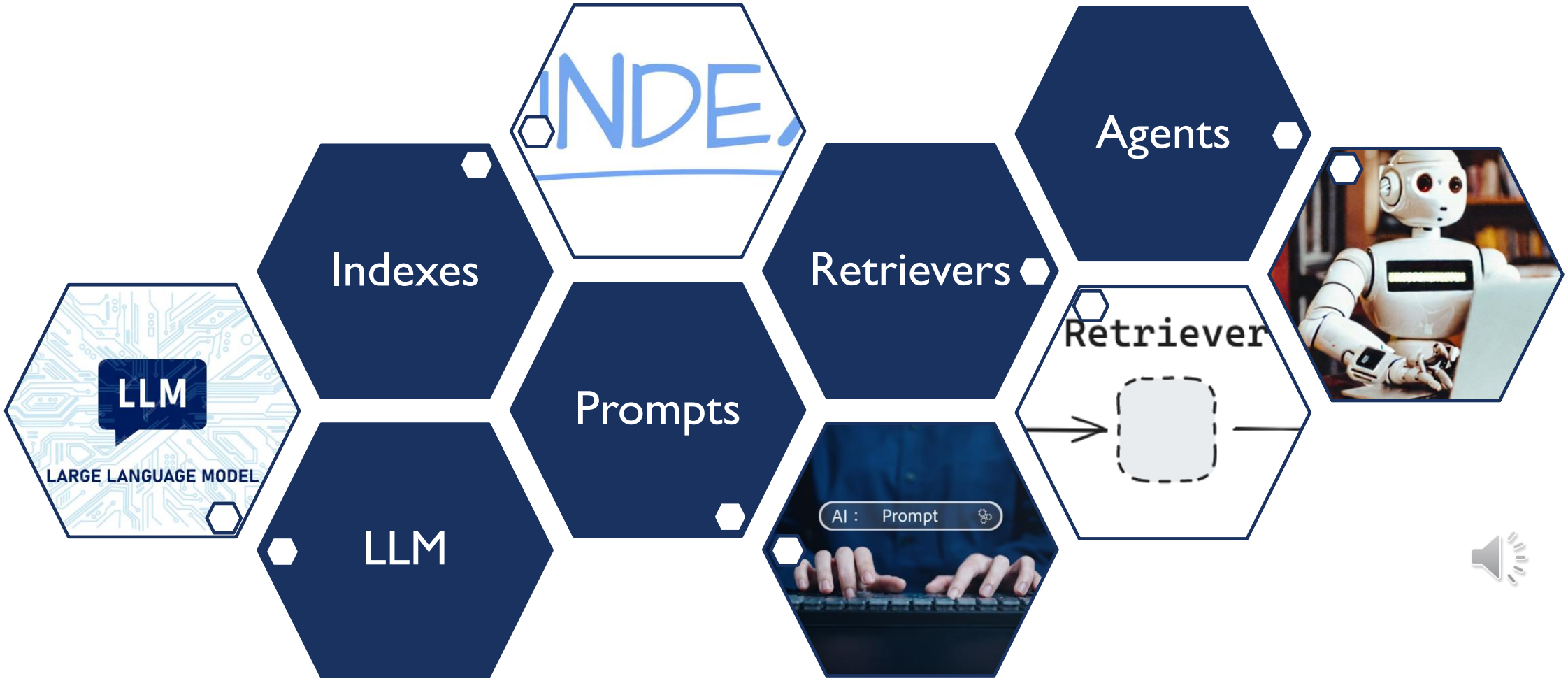
RAG


4. Response
1. Query
Retrieval Model
Pre-trained LLM
2. Search
3. Relevant doc chunks
Knowledge Base

CONTENT

- Understanding the Basics of LLMs
- Working with Embeddings and Vectors
- Storing and Searching Vectorized Data
- Retrieval-Augmented Generation (RAG)
- **Using LangChain for AI Workflows**
- Optimizing AI Model Performance

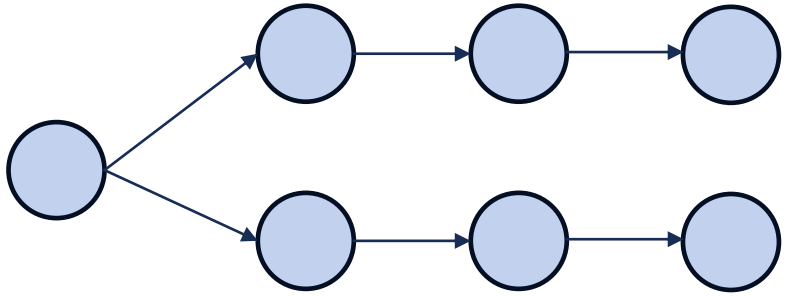
LANGCHAIN COMPONENTS



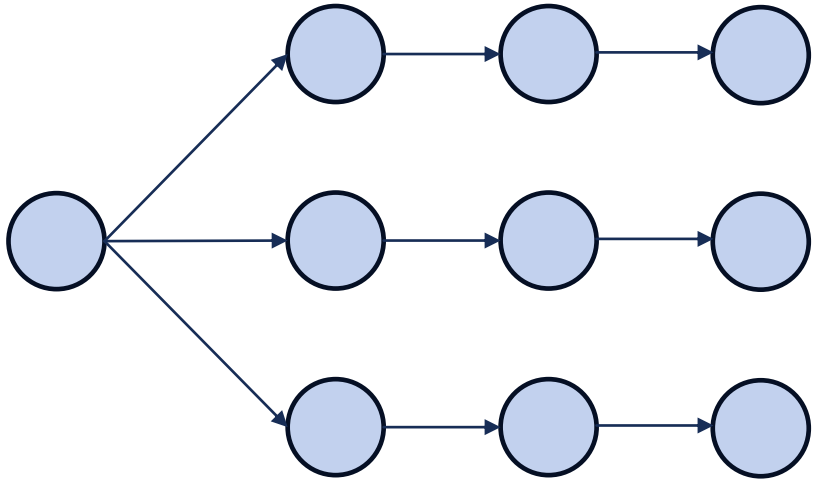
CHAIN TYPES

- Extended

- Parallel



- Branching



CONTENT

Understanding the Basics of LLMs

Working with Embeddings and Vectors

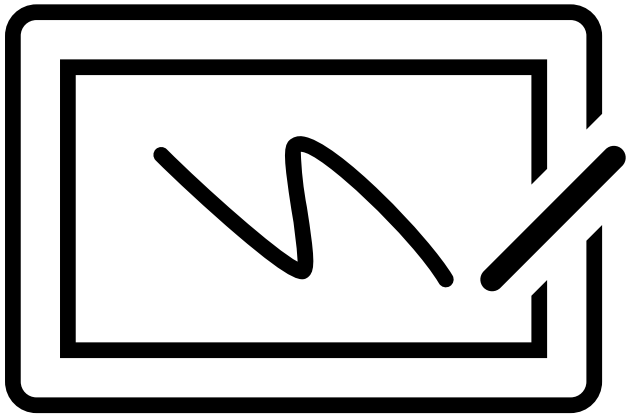
Storing and Searching Vectorized Data

Retrieval-Augmented Generation (RAG)

Using LangChain for AI Workflows

Optimizing AI Model Performance

PROMPT ENGINEERING



- Key aspects:
  - Clear instructions
  - Context setting
  - Constraints