REPUBLIC POLYTECHNIC

botpress

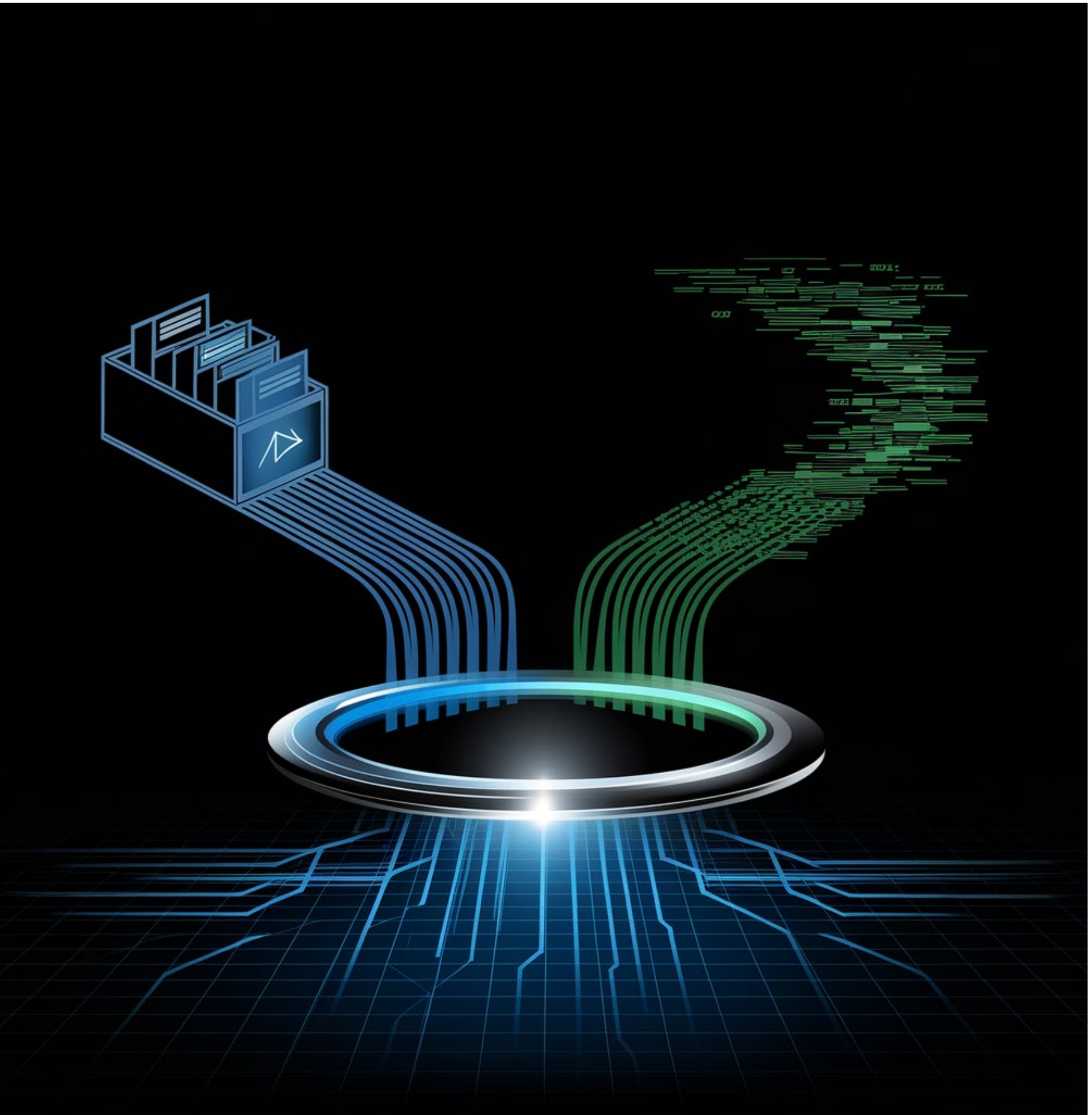
Master the Future: Build Smarter RAG Bots with Botpress!

Unleash AI Brilliance, One Bot at a Time

RAG

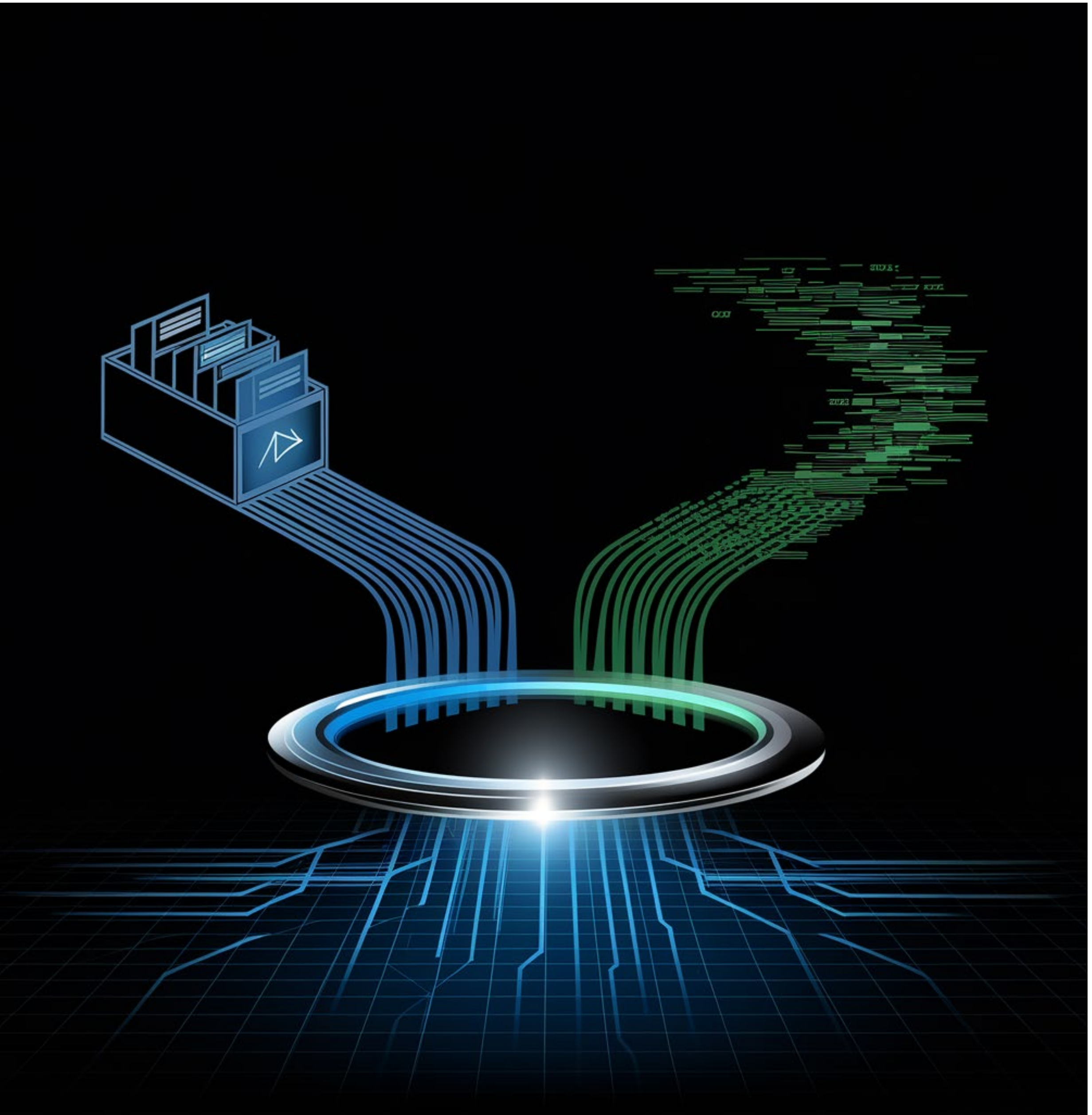
Retrieval Augmented Generation

Retrieval-Augmented Generation (RAG)



- **What It Is:** A hybrid AI approach combining retrieval and generation for smarter, context-aware responses.
- **Retrieval Step:** Pulls relevant info from a data source or other sources using a query.
- **Generation Step:** Uses a language model to create a coherent, tailored answer based on retrieved data.
- **Why It's Powerful:** Boosts accuracy and relevance by grounding AI outputs in real, up-to-date information.
- **Use Case:** Think chatbots that fetch facts from documents before replying—no more guessing!

Retrieval-Augmented Generation (RAG)



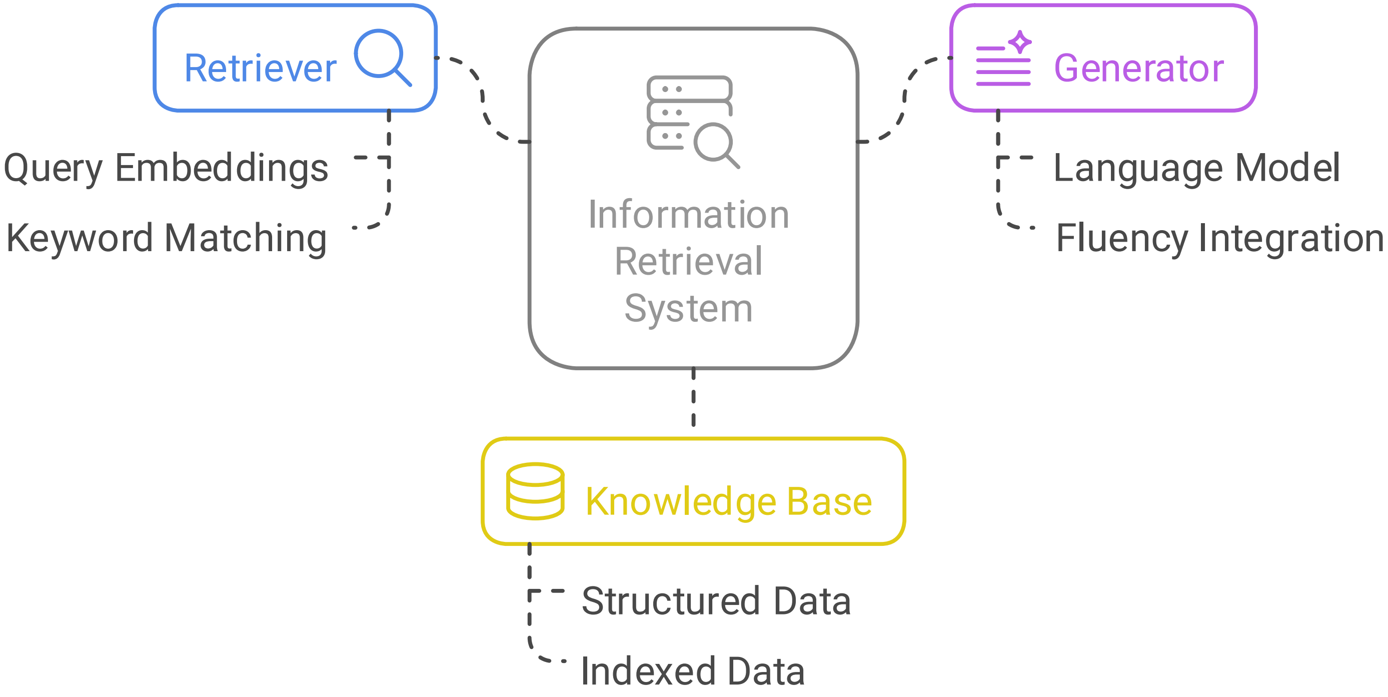
YouTube

But it is not all Sunshine and Lollipops

Pros	VS	Cons
Enhances accuracy		Slower process
Reduces hallucinations		Quality reliance
Scalable		Complex implementation
Contextual		Higher resource use
Flexible integration		Risk of irrelevant data

- **Pros:**
  - Enhances accuracy: up-to-date, dynamic information.
  - Reduces hallucinations by using real data.
  - Scalable: adapts easily to new datasets.
  - Contextual: tailors responses to each query.
  - Flexible: integrates with existing knowledge bases.
- **Cons:**
  - Slower due to the retrieval process.
  - Relies on the quality of external data.
  - Complex to implement and optimize.
  - Higher resource usage (compute, memory).
  - Risk of irrelevant or noisy retrieved data.

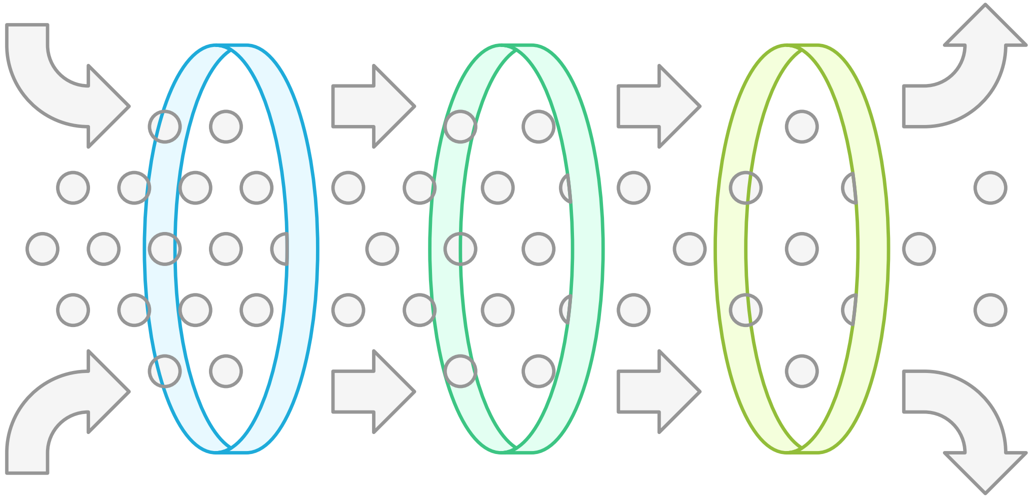
Key Components



- **Retriever:**
  Searches for relevant info using query embeddings or keywords. Matches queries to documents in the knowledge base via vector **similarity** search
- **Generator:**
  A language model that takes retrieved documents plus the query to produce a coherent answer. Blends retrieved info with its pre-trained knowledge for fluency.
- **Knowledge Base:**
  The external data repository (e.g., documents, databases, web). Must be well-structured and indexed for efficient retrieval, often using embeddings for semantic search.

Vector Database

Query Processing Funnel



Generate Embedding

Convert query to vector

Match Q&A Pair

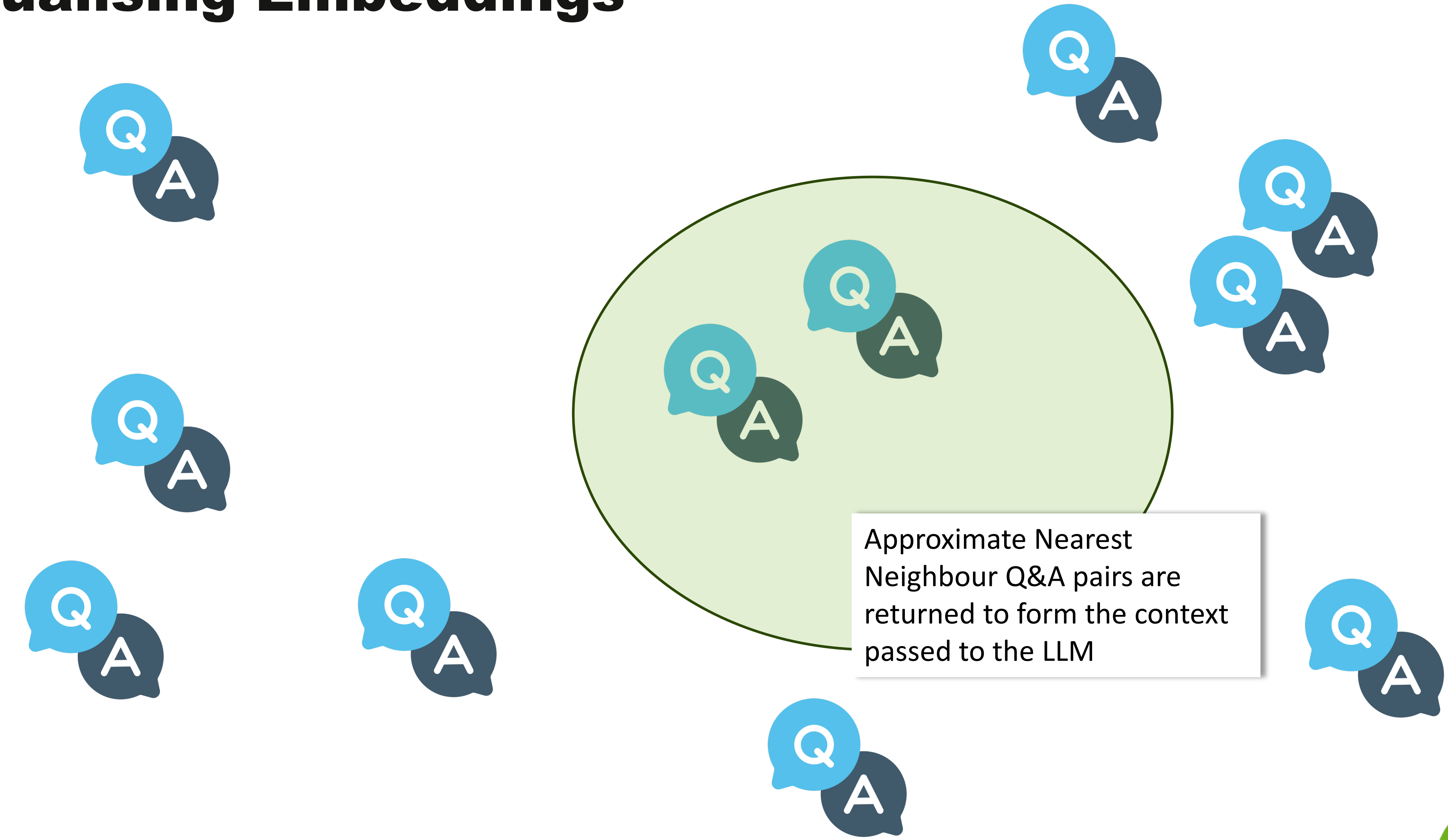
Find closest match in AstraDB

Pass to Generator

Provide matched data for generation

- **What They Are:** Databases storing text as vectors (big numbers) for semantic search.
- **Role in RAG:** Retriever uses them to fetch relevant Q&A pairs from a knowledge base. Closest few answer vectors to the question vector.
- **Practical Setup:** For a Q&A document, chunk each Q&A pair. Convert each Q&A into embeddings using a model like `SentenceTransformers`. Store in vector DB.
- **How It Works:** Query → embedding query → match closest Q&A pairs → pass closest QnA pairs to the model as context.
- **Why It Helps:** Ensures precise retrieval of specific Q&A chunks, improving answer relevance.

Visualising Embeddings



Approximate Nearest Neighbour Q&A pairs are returned to form the context passed to the LLM

Terms

- **Grounded:** Responses are based on real, retrieved data, not just the model’s guesses.
- **Retrieval:** The process of finding relevant info from a knowledge base using a query.
- **Generation:** Creating a coherent answer using a language model and retrieved data.
- **Embedding:** Numerical vector representing text’s **meaning** for similarity search.
- **Vector Database:** Stores embeddings for fast, semantic retrieval (e.g., **Weaviate**, AstraDB, Pinecone).
- **Knowledge Base:** External data source (e.g., documents, Q&A) from which RAG retrieves.
- **Retriever:** Component that searches the knowledge base for relevant info.
- **Generator:** Language model that produces the final answer from the context and the user's question.
- **Semantic Search:** Finding data based on **meaning**, not just keywords, using embeddings.
- **Hallucination:** When a model makes up incorrect info. RAG reduces hallucinations.

botpress

Streamline AI development with Botpress’s powerful, low-code conversational AI platform.

Introducing Botpress

- **What is Botpress?**
  - Open-source conversational AI platform.
  - Purpose-built for creating chatbots and virtual agents.
- **Key Features:**
  - Visual drag-and-drop flow editor.
  - Built-in knowledge bases with retrieval-augmented generation (RAG).
  - Connects easily to APIs, databases, and tools.
  - Fast prototyping and cloud-hosted deployment options.
- **Why Use Botpress?**
  - Simpler and faster to build chatbots compared to heavier frameworks.
  - Ideal for AI applications needing real, document-grounded responses.

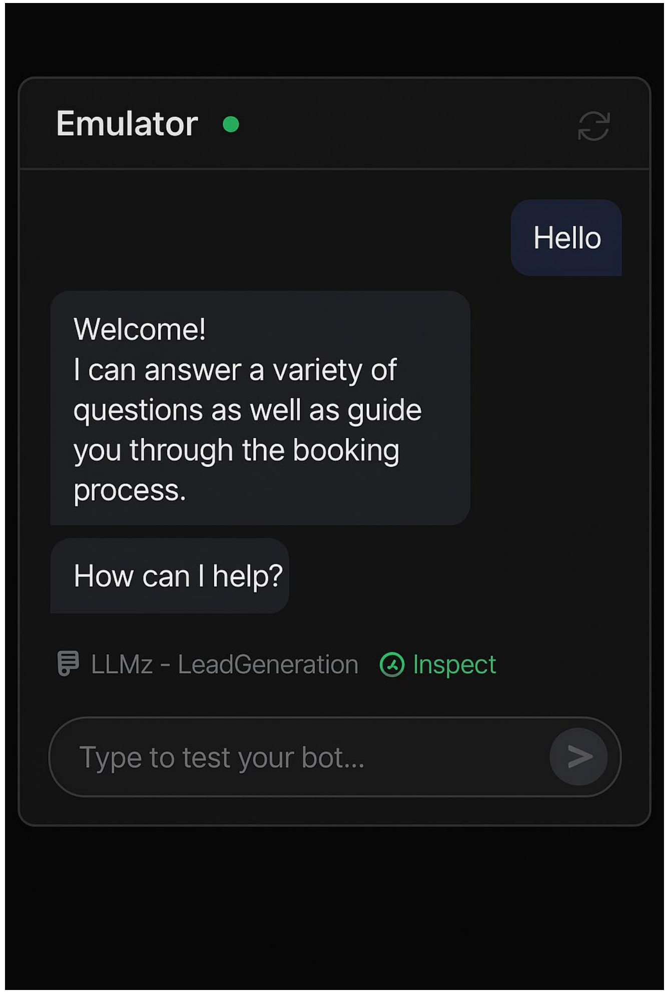
Botpress Knowledge Base & Vector Search

- Built-in Knowledge Base:
  - Upload documents, text, or URLs to create a searchable repository.
  - Enables bots to provide accurate, document-grounded responses.
- Semantic Search with Vector Embeddings:
  - Transforms content into vector embeddings to capture semantic meaning.
  - Facilitates retrieval of relevant information based on user queries.
- Vector Database Integration:
  - Utilizes **Weaviate**, an open-source vector database, to store and manage embeddings.
  - Supports efficient semantic search and retrieval-augmented generation (RAG).
- Scalable and Efficient:
  - Handles large datasets with optimized search capabilities.
  - Enhances the bot's ability to provide precise and contextually relevant answers.

Botpress Studio & Visual Flow Editor

Slide 3: Botpress Studio & Visual Flow Editor

- **Botpress Studio:**
  - Centralized environment for building, testing, and deploying AI agents.
  - Integrates tools for managing knowledge bases, flows, and integrations.
- **Visual Flow Editor:**
  - Drag-and-drop interface for designing conversation flows.
  - Utilize nodes and cards to define dialogue logic and actions.
  - Supports modular workflows for complex conversation structures.
- **Built-in Emulator:**
  - Test and debug conversations in real-time within the Studio.
  - Inspect LLM decisions and iterations for accurate responses.



Emulator ●

Hello

Welcome!
I can answer a variety of questions as well as guide you through the booking process.

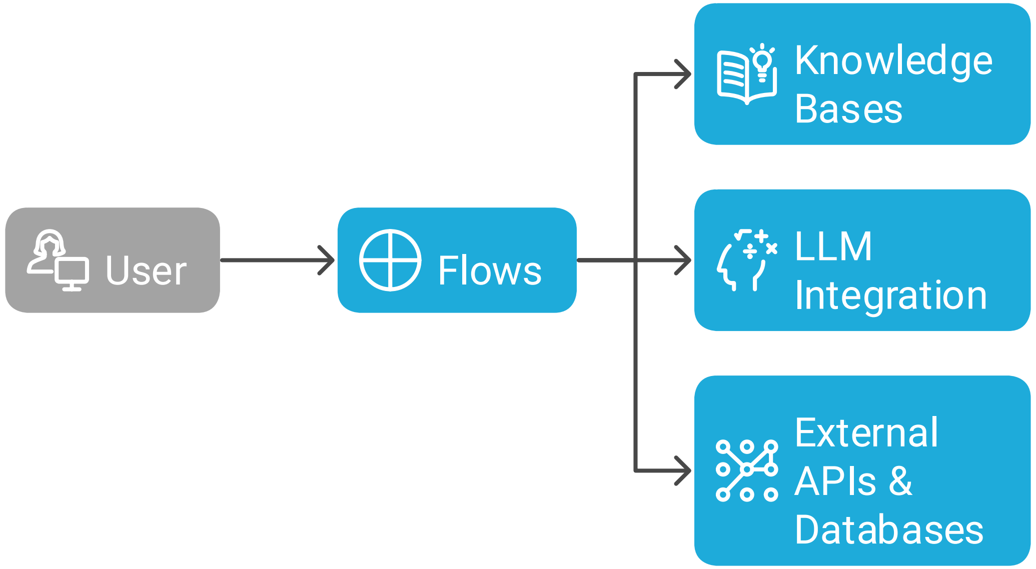
How can I help?

LLMz - LeadGeneration Inspect

Type to test your bot...

Botpress: How it all connects!

- **User:** Interacts with the chatbot via messaging interface.
- **Flows:** Handle dialogue structure, control conversation logic, context switching, and actions.
- **Knowledge Bases:** Power Retrieval-Augmented Generation (RAG) respond using real documents and data.
- **LLM Integration:** Under the hood, Botpress can call large language models (e.g., OpenAI) to enhance responses.
- **External APIs & Databases:** Optional connections to pull live data or trigger business workflows.



Activity

Ai

Kahoot!

Botpress

Activity

Activity: Create a RAG Chatbot

- Open the document called: **Building a RAG System in Botpress**
- Individually, follow the timing given in the document
- Post your results to MST.

Lesson 09

Summary of Material Covered

Lesson 09 Review

- ❖ **Retrieval-Augmented Generation (RAG):** Combines retrieval of real-world data with AI generation for accurate, grounded responses.
- ❖ **Key Components:** Retriever, Generator, Knowledge Base, Vector Database (e.g., Weaviate).
- ❖ **RAG Pros & Cons:** Boosts accuracy and context but adds complexity and resource demands.
- ❖ **Botpress Introduction:** Open-source, low-code platform for creating RAG-powered conversational AI.
- ❖ **Botpress Features:** Built-in knowledge bases, semantic search with vector embeddings, drag-and-drop visual flow editor, LLM integration.

Thank you

School of Infocomm

C240 AI Essentials and Innovations

© Republic Polytechnic 2025: All Rights Reserved