DATASTAX

Building Your Generative AI Application

Hands-on Workshop

September 2024

Meet Sami and Dieter

Sami Kaksonen
Solution Engineer

Dieter Flick
Solution Engineer

Agenda

- The Big Picture: Generative AI
- What is Retrieval Augmented Generation (RAG)
- Hands-on: Build Your Own RAG based Chatbot
- Langflow for a 100X easier AI
- We are proud to have developed a functional chatbot

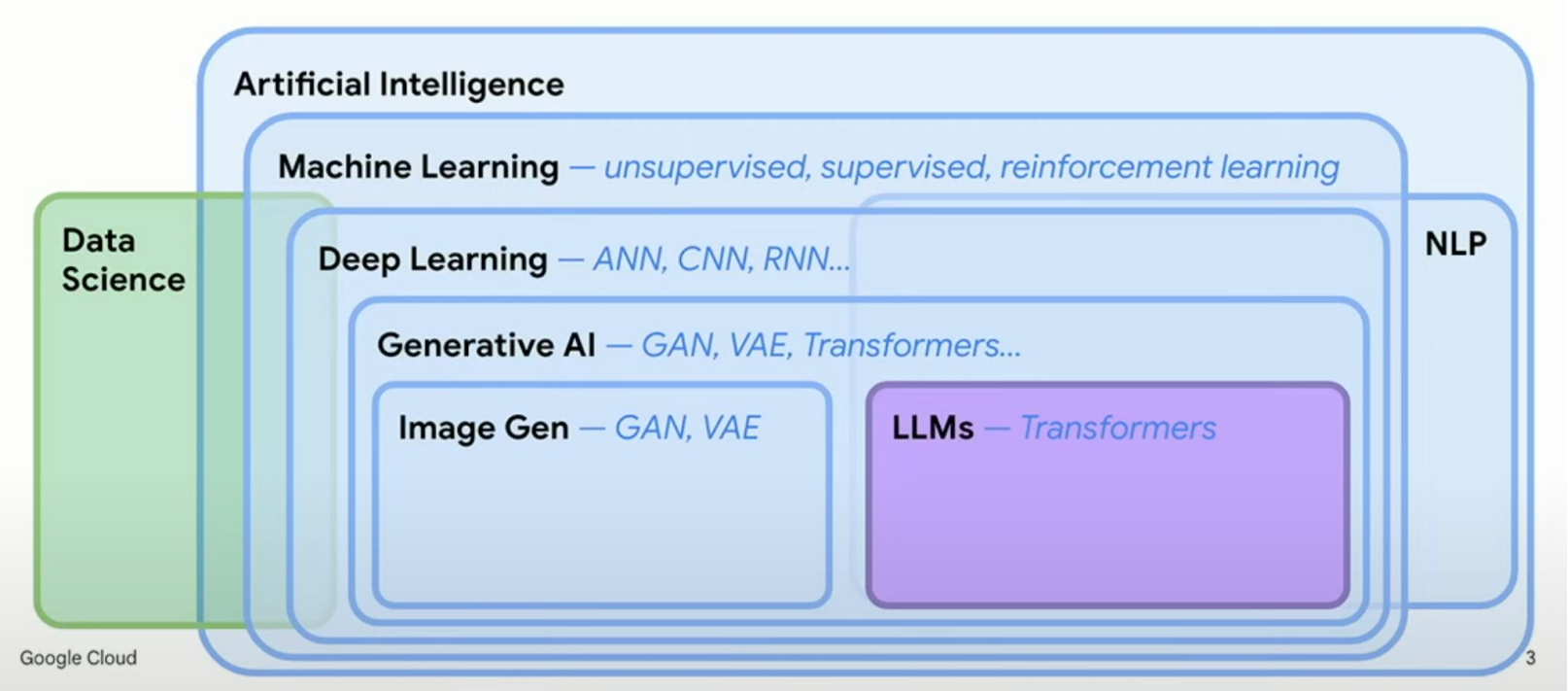
DataStax

- DataStax is the real-time data company for Gen AI applications.
- **Data = AI**
- Astra DB
    - Get real-time vector data and RAG capabilities that Gen AI apps need, with seamless integration with developers' stacks of choice (RAG, Langchain, LlamaIndex, OpenAI, GCP, AWS, Azure, etc)
    - Get accurate AI applications into production, fast…

DATASTAX

Let's set the foundation with some concepts

The Big picture



Large Language Models (LLM)

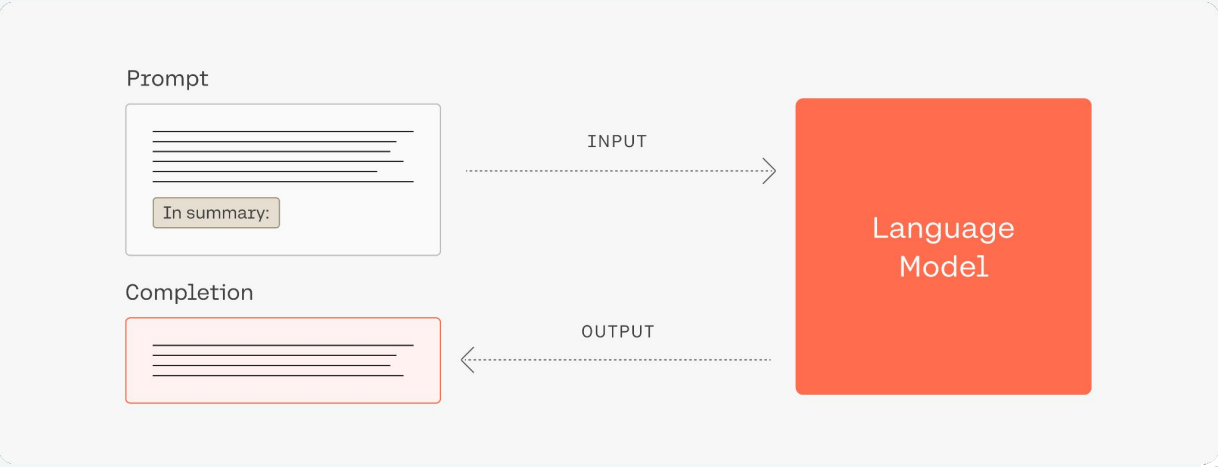


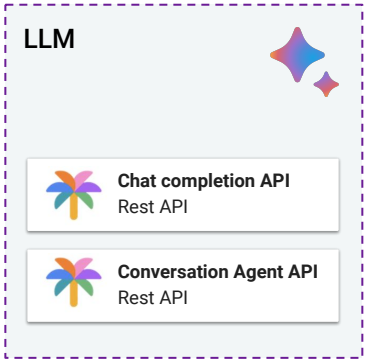
image by cohere

Large language models are trained on massive amounts of text data.

Designed to process and understand natural language, such as human speech and text.

This allows the model to learn patterns and relationships between words, phrases, and sentences, enabling it to generate coherent and meaningful language output.

Limitations of “LLM only” mode



- LLM .....can be outdated
- LLM .....Does not know *your* data
- LLM .....is not tuned = hard steerability
- LLM .....Hallucinating if not properly prompted
- LLM .....works with limited Input windows (tokens)
- LLM .....is not secure!

There is no AI without Data

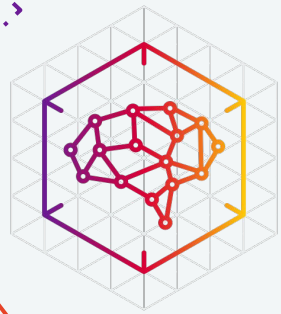
Adding Enterprise Context to GenAI

Context added at run-time
Data not retained

Proprietary & Personalized Information

Retrieval-Augmented Generation

Context Window



Domain Specific Enterprise Data

Fine-Tuning

Public Data

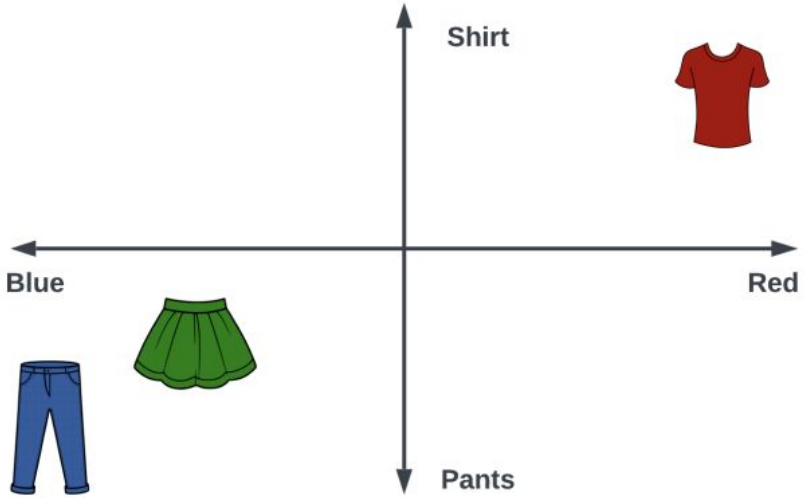
Pre-Training

Model Parameters

Context added in training
Data is retained

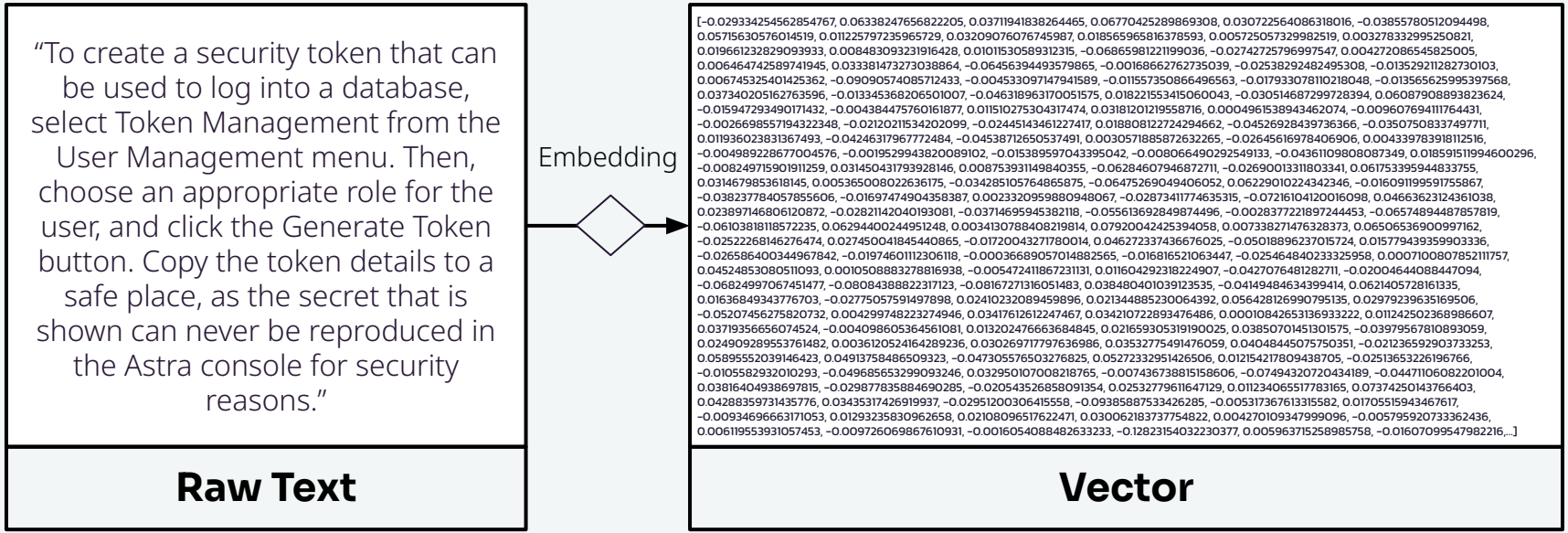
What is Vector Search?

- Vector search finds objects that have similar meaning
- Vector search understands MEANING
- Vectors created from EXISTING data through EMBEDDINGS



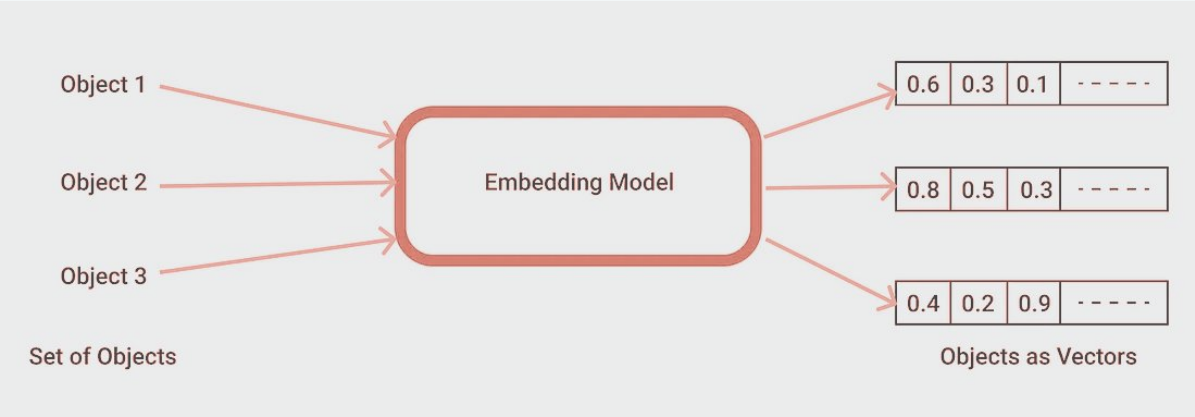
What is a Vector?

A Vector is a multi-dimensional numerical representation of text/image/video



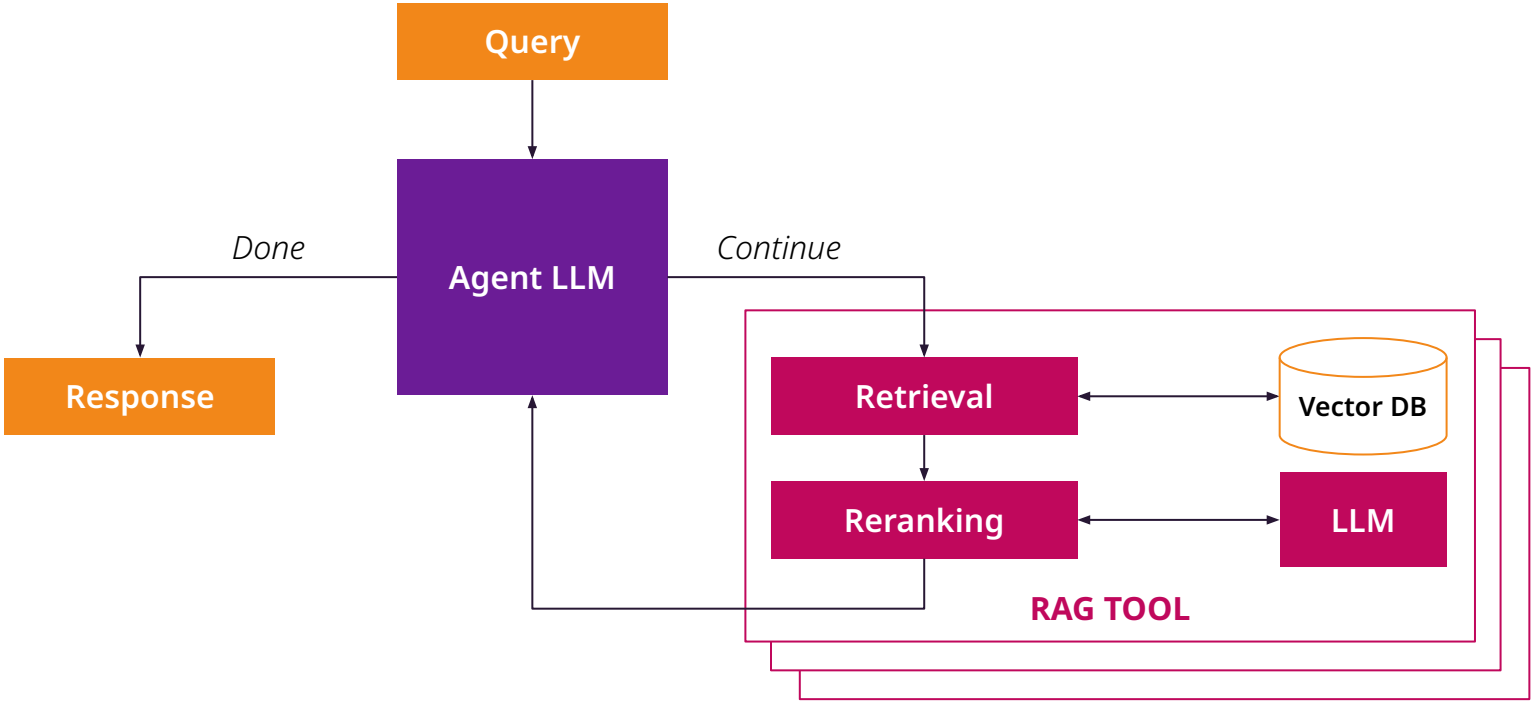
What is an Embedding?

Take source objects (text, images, sound, movies) and create Vector format representation of the context.
This allows for Similarity Search on the Database finding Semantically comparable objects.

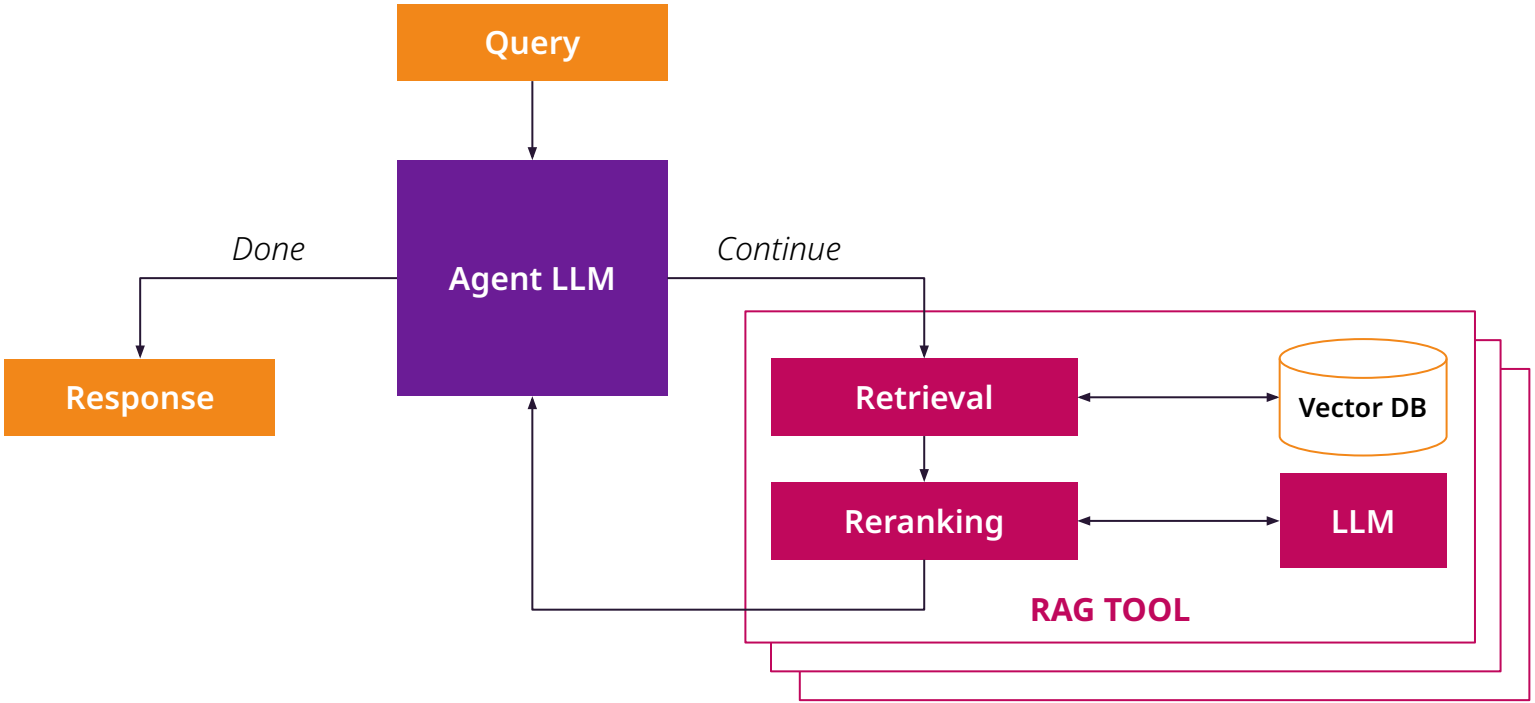


Goal is to capture semantics and context.
Processed by a ML model

What is Retrieval Augmented Generation (RAG)

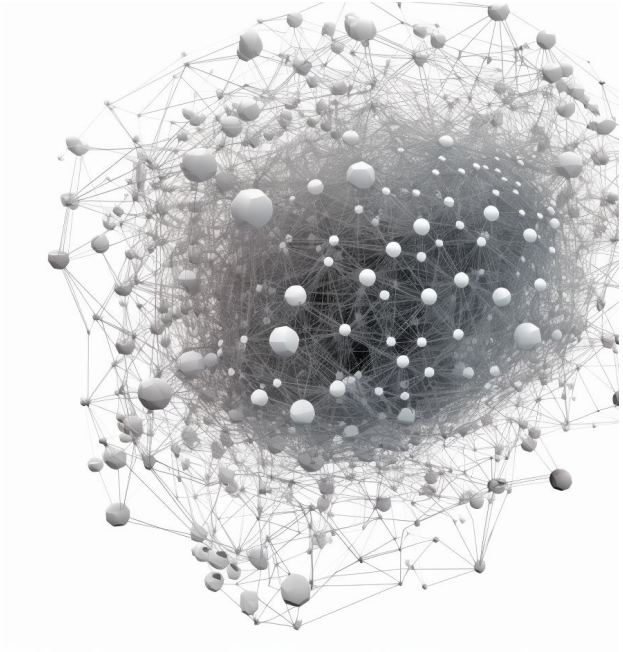


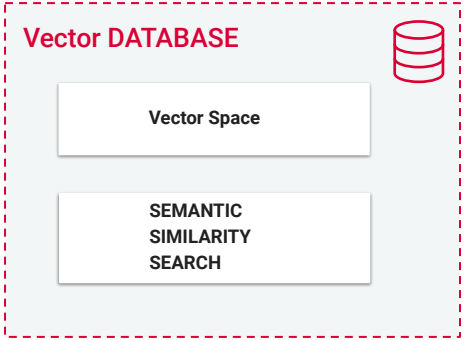
Chatbot



What is a Vector Database?

- Handling **A LOT** of vector
- Effective Search Algorithms
- Performant, Resilient
- Dynamic (vector sizes)
- Meta Data Filtering
- Keyword search
- Semantic Caching
- Chat History
- Key Value cache





Best Vector Search in the World

Challenge? Solved!

Scale
Cost Efficiency
Security
Reliability
Low latency
Performance

DATASTAX

74% Faster than any Vector DB

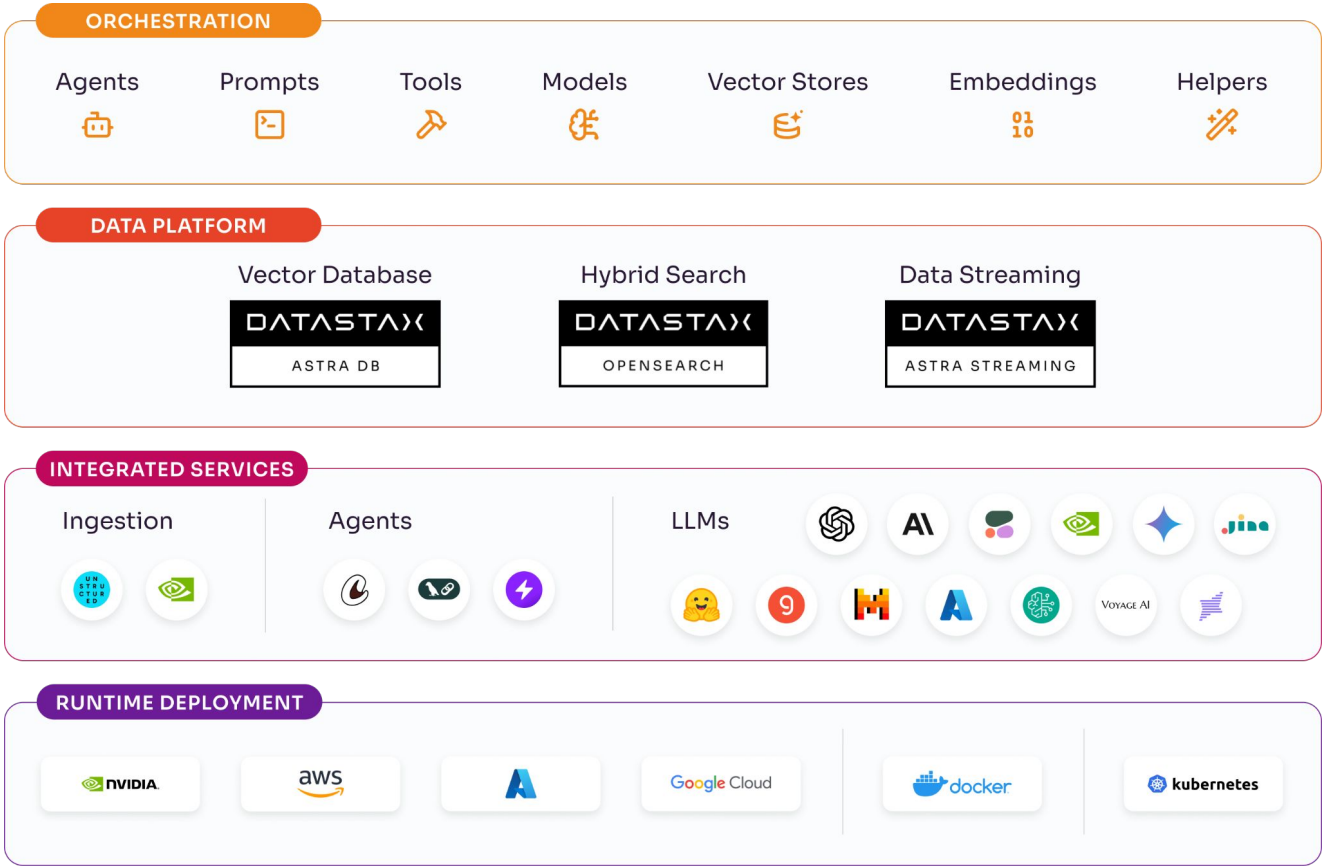
20% higher **relevance**

Ready for **actual production use**

On-Prem and Managed

All CSPs

Langflow AI PaaS



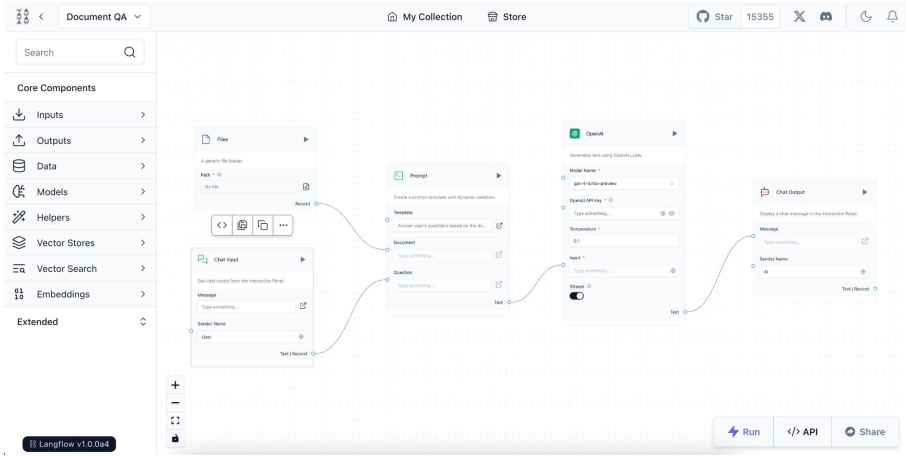
Langflow AI PaaS

What is Langflow?

Langflow is a new way to build AI and LLM-powered applications. It includes a visual builder and a workflow engine for Python.

What value does it deliver?

Langflow makes is 100x faster for developers to prototype, iterate and ship AI applications to production.



https://docs.langflow.org/

Now let's build your own GenAI application

Astra DB sign up

Getting started

Go to https://github.com/difli/build-your-own-rag-chatbot

Thank You


DATASTAX