



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

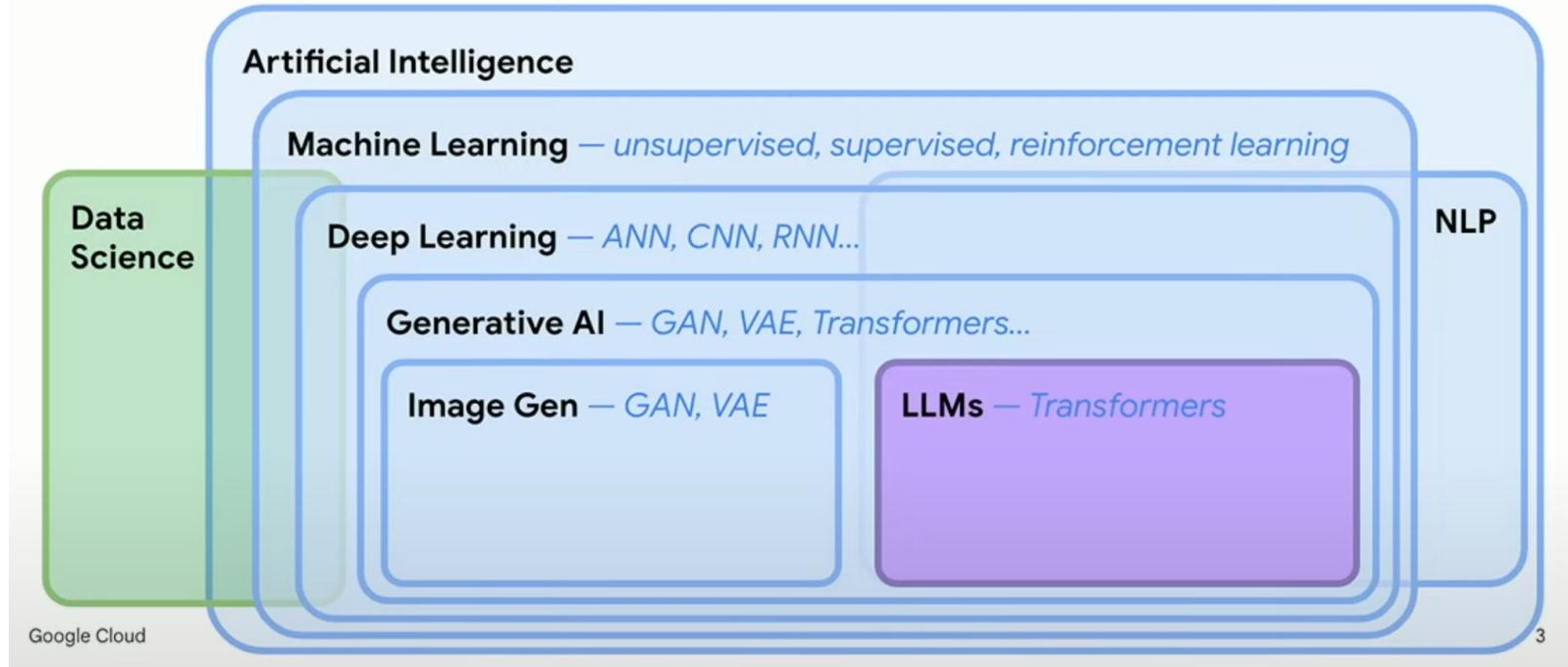
DataStax

- 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...

Let's set the
foundation with
some concepts



➤ The Big picture



➤ Large Language Models (LLM)

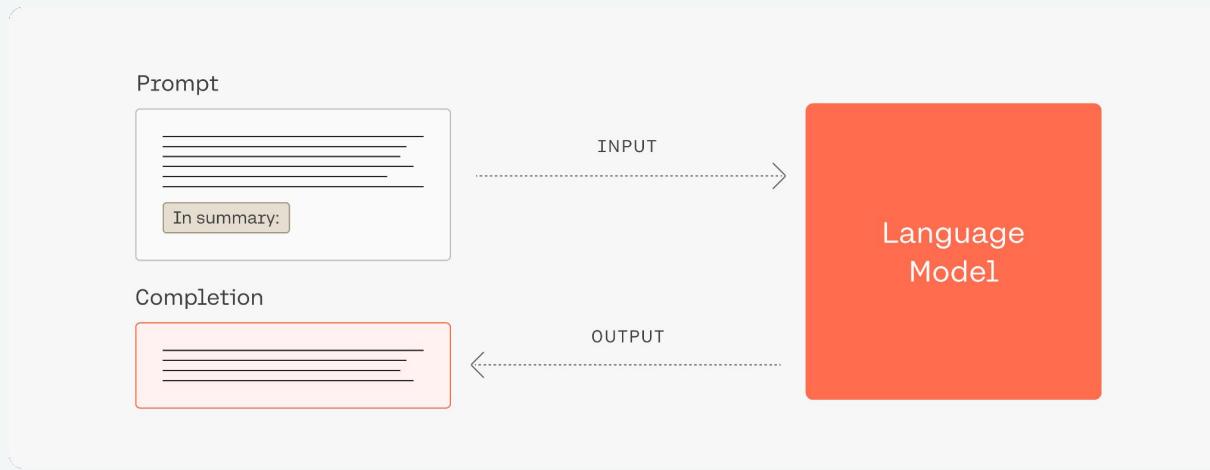


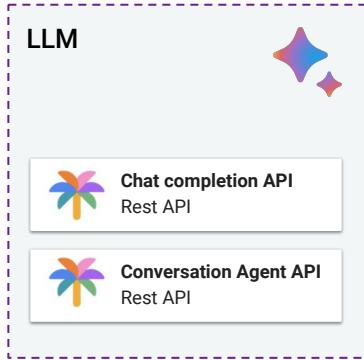
image by [cohere](#)

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.

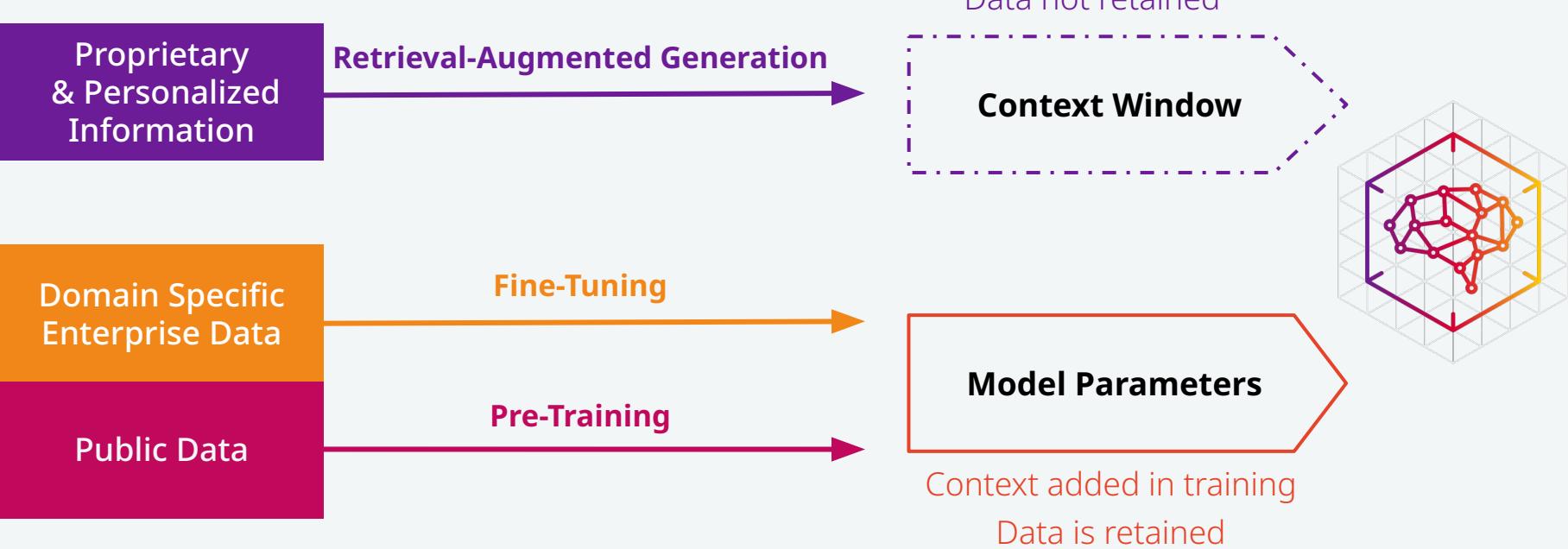
➤ Limitations of “LLM only” mode



- LLMcan be outdated
- LLMDoes not know *your* data
- LLMis not tuned = hard steerability
- LLMHallucinating if not properly prompted
- LLMworks with limited Input windows (tokens)
- LLMis not secure!

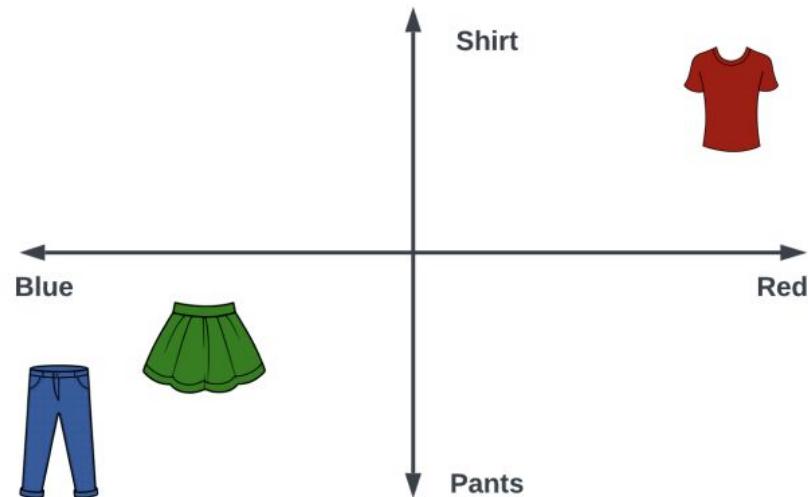
There is no AI without Data

➤ Adding Enterprise Context to GenAI



➤ 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

"To create a security token that can be used to log into a database, select Token Management from the User Management menu. Then, choose an appropriate role for the user, and click the Generate Token button. Copy the token details to a safe place, as the secret that is shown can never be reproduced in the Astra console for security reasons."

Raw Text

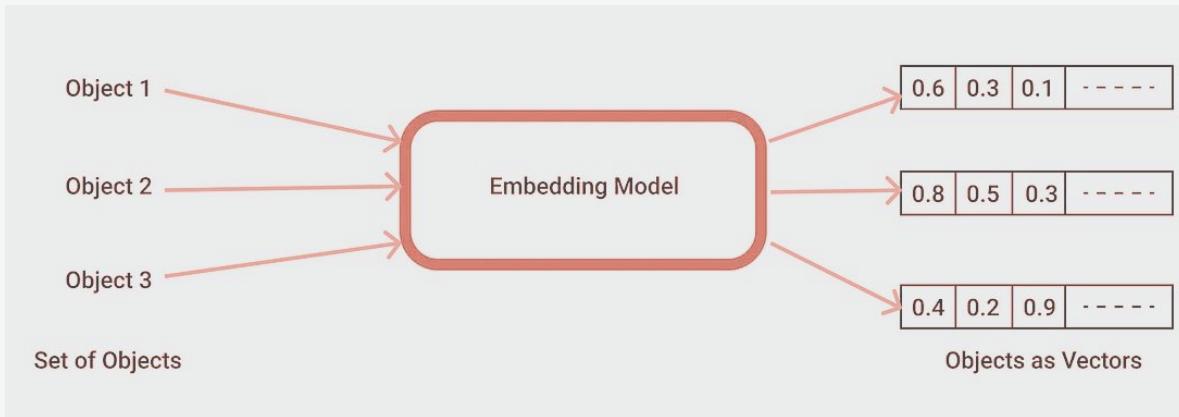


```
[ -0.29334254562854767, 0.06338247656822205, 0.03711941838264465, 0.06770425289869308, 0.030722564086318016, -0.03855780512094498, 0.05715630576014519, 0.0125797235965729, 0.0320907607645987, 0.018565965816378593, 0.005725057329982519, 0.003278332995250821, 0.019661232829093933, 0.008483093231916428, 0.01011530589312315, -0.0686598121199036, -0.02742725796997547, 0.004272086545825005, 0.006464742589741945, 0.033381473273038864, -0.06456394493579865, -0.01686627672735039, -0.02538292482495308, -0.01529121282730103, 0.006745325401245362, -0.09090574085712433, -0.004533097174914563, -0.0157530866495653, -0.01593307810218048, -0.0135655625995397568, 0.0373402051626763596, -0.013345368206501007, -0.04631896317005157, 0.01822153415060043, -0.030514687299728394, 0.06087908893823624, -0.015947293490171432, -0.004384475760161877, 0.01510275304317474, 0.0318120121958716, 0.004961538943462074, -0.00960769411764431, -0.0026698557194322348, -0.02120211534202099, -0.02445143461227417, 0.018808122724294662, -0.04526928439736366, -0.0350750833749771, 0.01936022831367491, -0.04246317967772484, -0.04538712650537491, 0.003057188587263265, -0.02645616978406906, 0.00433978391812516, -0.004989228677004576, 0.0019529943620089102, -0.015389597043395042, -0.008066490292549133, -0.0436101980807349, 0.018591511994600296, -0.0082497153910191259, 0.031459431793928146, -0.052840679486872711, -0.0269001331803341, 0.061753395944833755, 0.034167985361814, 0.005365008022636175, -0.034285105764865875, -0.04675269049406052, 0.06229010224342346, -0.01609119951755867, -0.038237784057855606, -0.01697474904358387, 0.0023320959880948067, -0.028734117463515, -0.07216104120016098, 0.04663623124361038, 0.028397146806120872, -0.02821142040193081, -0.03714695945382118, -0.05561392847684496, -0.0028377221897244453, -0.0657469448785719, -0.06103818118572235, 0.06294400244951248, 0.003430788408219184, 0.07920042425394058, 0.007338271476328373, 0.0650653690901762, -0.0252226146276474, 0.02745004184540865, -0.07200432717801804, 0.046272337436676025, -0.05018896237015724, 0.015779439359093336, -0.026586400344967842, -0.0197460112306118, -0.00036689057014882565, -0.016816521063447, -0.025464840233235958, 0.000710080780211757, 0.04524853080511093, 0.001050888327816938, -0.0054724186723224907, -0.042706481282711, -0.02004644088447094, -0.06824997067451477, -0.08084388822317123, -0.08167271316051483, 0.038480401039123535, -0.04149484634399414, 0.0621405728161335, 0.01636849343776703, -0.02775057591497898, 0.02410232089459896, 0.021344885230064392, 0.056428126990795135, 0.02979239635169506, -0.05207456275820732, 0.00429974822374946, 0.03471621247467, 0.034210722893476486, 0.00010842653136933222, 0.01124250236896607, 0.0379135665607452, -0.004098605364561081, 0.0120247663684845, 0.0216593053190025, 0.03850710451301575, -0.03979567810893059, 0.024909289553761482, 0.003612052416289236, 0.03026977797636983, 0.03532775491476059, 0.04048445075750351, -0.02123659290733253, 0.05895552039146423, 0.04913758486509323, -0.047305576503276825, 0.05272323951426506, 0.01215427089438705, -0.02513653226196766, -0.0105582932010293, -0.049685653299093246, 0.032950107008218765, -0.00743673815158606, -0.07494320720434189, -0.04471016082201004, 0.0316404938697815, -0.029877835884690285, -0.020543526858091354, 0.0253277961647129, 0.011234065517783165, 0.07374250143766403, 0.04288359731435776, 0.03435317426919937, -0.02951200306415558, -0.09385887533426285, -0.005317367161315582, 0.0170515943467617, -0.00934696663171053, 0.01293235830962658, 0.02108096517622471, 0.03006218373754822, 0.004270109347999096, -0.005795920733362436, 0.006119553931057453, -0.009726069867610931, -0.01605408842633233, -0.12823154032203377, 0.005963715258985758, -0.01607099547982216...]
```

Vector

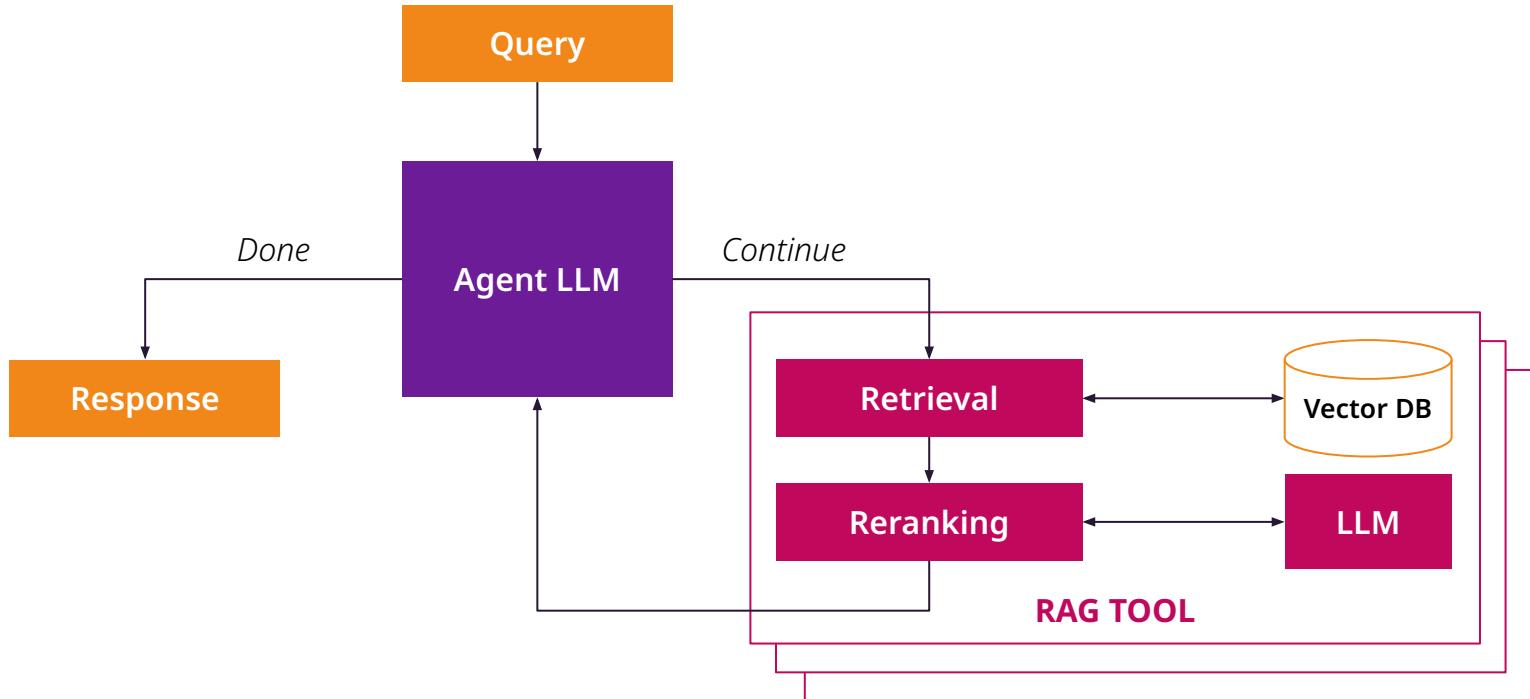
What is an Embedding?

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.



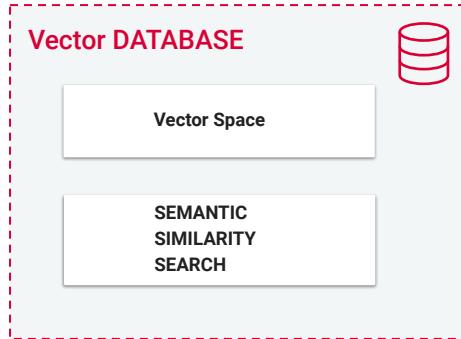
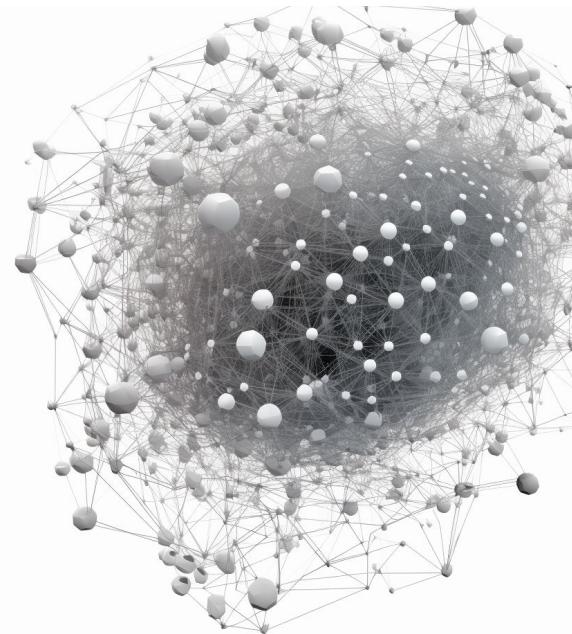
**Goal is to capture semantics and context.
Processed by a ML model**

What is Retrieval Augmented Generation (RAG)



➤ 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

74% Faster than any Vector DB

20% higher **relevance**

Ready for **actual production use**

On-Prem and Managed

All CSPs



Named a Leader.

**The ONLY Hybrid
Leader.**

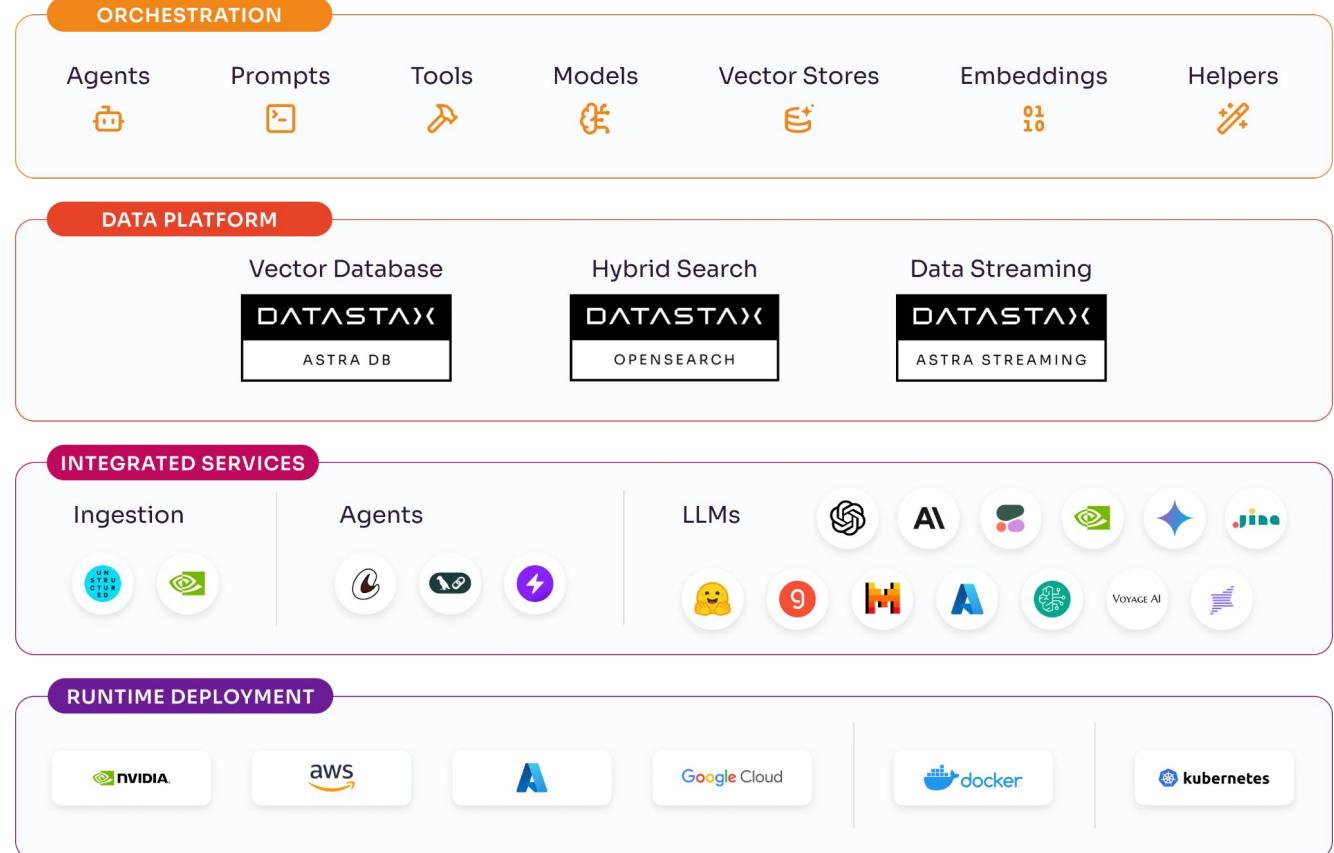
A COMPANY FIRST.

THE FORRESTER WAVE™
The Forrester Wave™: Vector Databases
Q3 2024



Langflow

AI PaaS



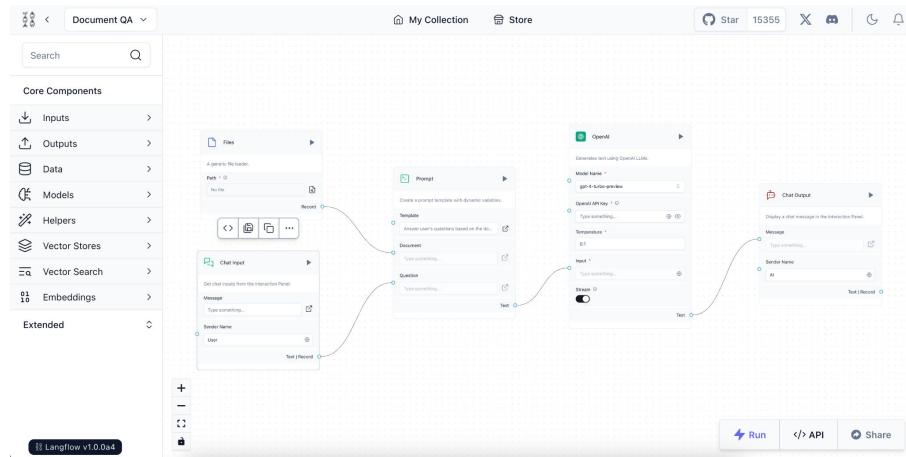
➤ Langflow AI PaaS

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-low-code-chatbot>





Thank
You

DATASTAX