



# **Mastering RAG : Build & Deploy Your GenAI Chatbot**

**Hands-on Workshop**

**April 2024**

# ➤ Meet the DataStax Team

**Kai Kühnel**  
Sales



**Sami Kaksonen**  
Solution Engineer



**Dieter Flick**  
Solution Engineer



# ➤ Meet the Jina AI Team

**Scott Martens**  
AI Evangelist



**Sofia Vasileva**  
Product Marketing



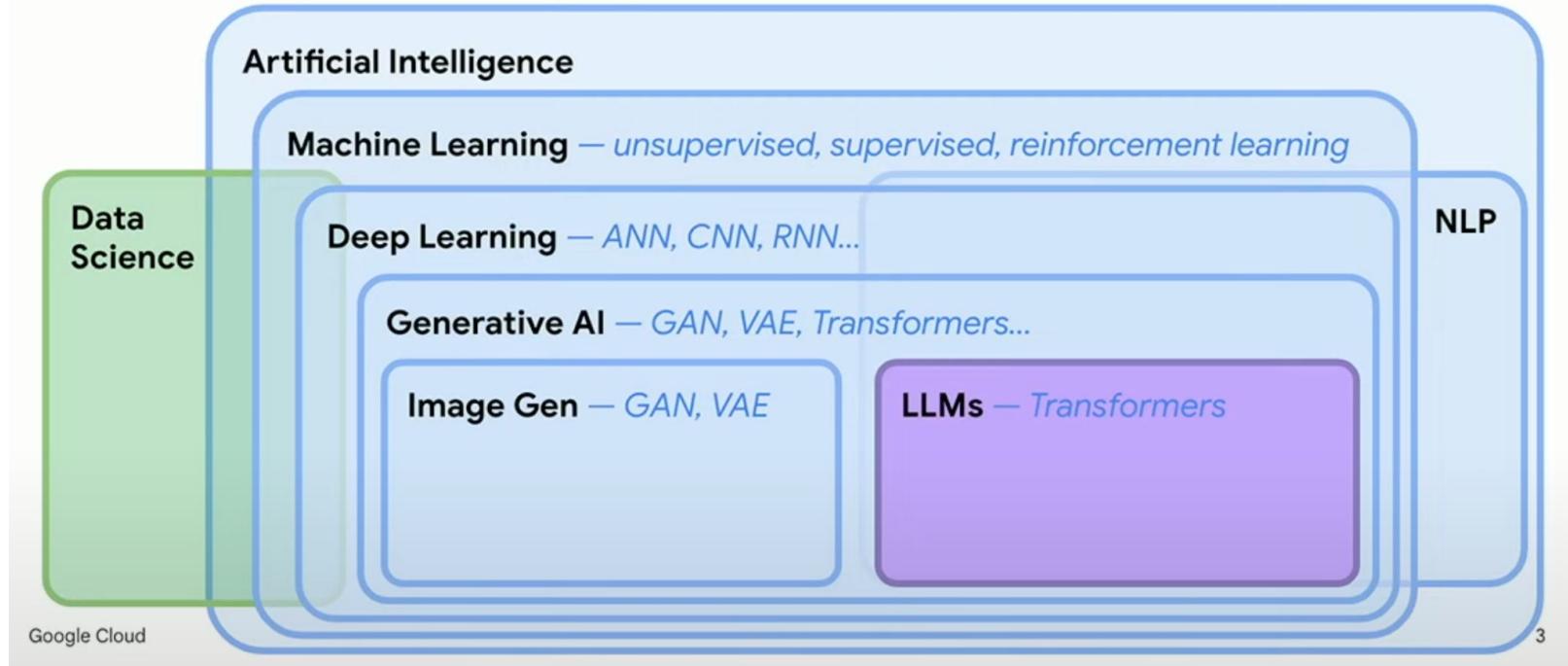
# Agenda

- 16:00 - Arrive and comfortably settle in
- 16:15 - Introduction to Generative AI
- 16:45 - Hands-on: Build Your Own RAG
- 18:30 - Networking, food and drinks
- 18:55 - Raffle
- 19:00 - We go home proudly with a working Chatbot

Let's set the  
foundation with  
some concepts



# ➤ The Big picture



# ➤ Large Language Models (LLM)

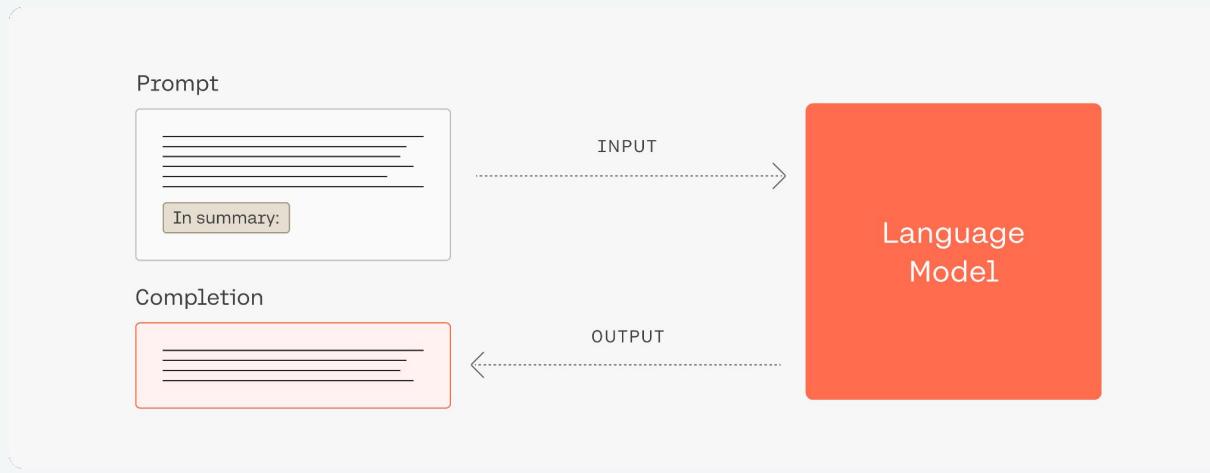


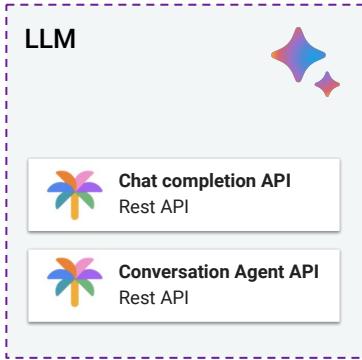
image by [cohere](#)

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.

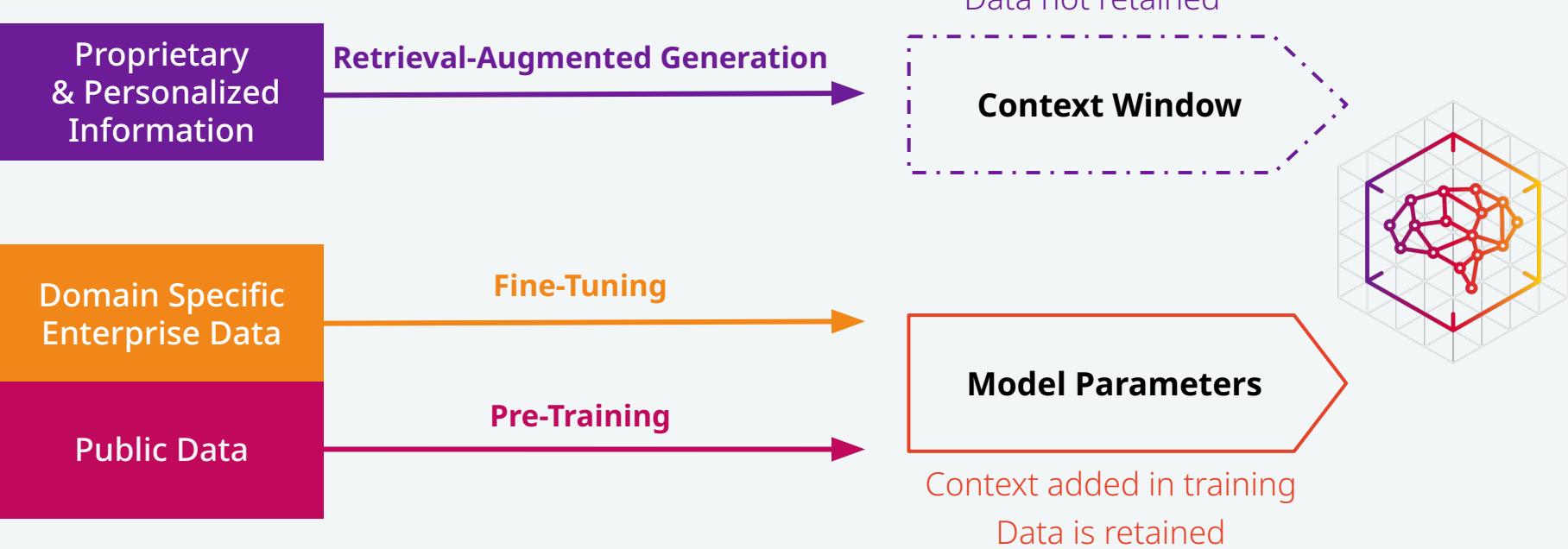
# ➤ Limitations of “LLM only” mode



- LLM .....can be outdated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!

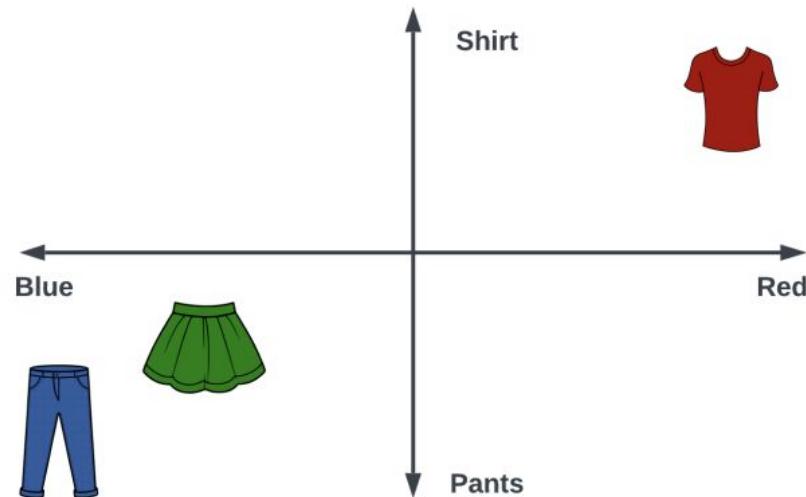
**There is no AI  
without Data**

# ➤ Adding Enterprise Context to GenAI



# ➤ What is Vector Search?

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

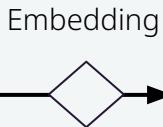


# What is a Vector?

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**

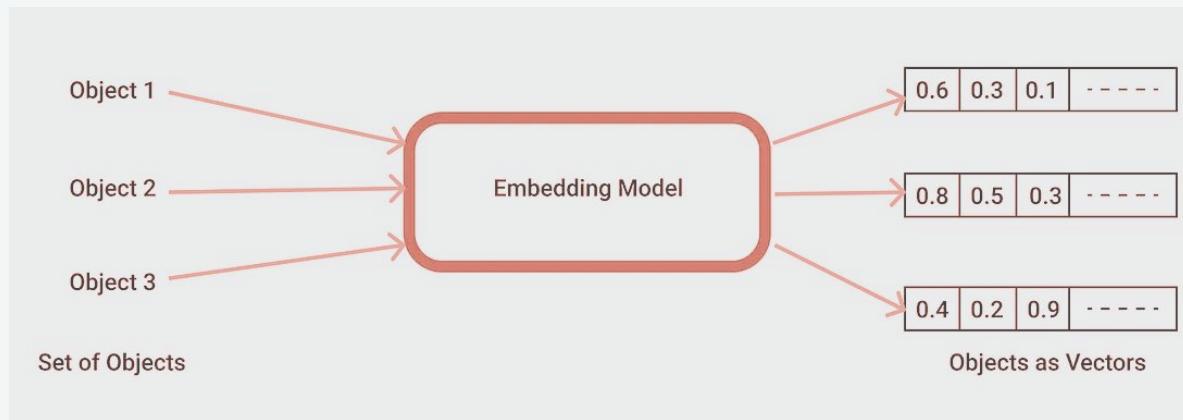


```
[-0.29334254562854767, 0.06338247656822205, 0.03711941838264465, 0.06770425289869308, 0.030722564086318016, -0.03855780512094498, 0.05715630576014519, 0.01225797235965729, 0.0320907607645987, 0.018565965816378593, 0.005725057329982519, 0.003278332995250821, 0.019661232829093933, 0.008483093231916428, 0.01011530589312315, -0.0686598121199036, -0.02742725796997547, 0.004272086545825005, 0.006464742589741945, 0.033381473273038864, -0.06456394493579865, -0.01686627672735039, -0.02538292482495308, -0.015291282730103, 0.006745325401425362, -0.09090574085712433, -0.004533072941563, -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, -0.04361098087349, 0.018591511994600296, -0.0082497153910191259, 0.03145431793928146, -0.05284067946872711, -0.0269001331803341, 0.061753395944833755, 0.034167985361814, 0.00536500802263617, -0.034285105764865875, -0.04675269049406052, 0.06229010224342346, -0.01609199591755867, -0.038237784057855606, -0.01697474904358387, 0.0023320959880948067, -0.0287341174635135, -0.07216104120016098, 0.04663623124361038, 0.0283897146806120872, -0.02821142040193081, -0.03714695945382118, -0.055613928476874496, -0.0028377221897244453, -0.0657469448785719, -0.06103818118572235, 0.06294400244951248, 0.00343078840219814, 0.07920042425394058, 0.007338271476328373, 0.06506536909097162, -0.0252226146276474, 0.027450041845440865, -0.01720043271780014, 0.046272337436676025, -0.05018896237015724, 0.015779439359903336, -0.026586400344967842, -0.0197460112306118, -0.00036689057014882565, -0.016816521063447, -0.025464840233205958, 0.000710080780211757, 0.04524853080511093, 0.001050888327816938, -0.00547241867231131, 0.0116042923224907, -0.042706481282711, -0.02004644088447094, -0.06824997076451477, -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, 0.0341712612247467, 0.034210272893476486, 0.0010842653136933222, 0.01124250236896607, 0.0379135665607452, -0.004098605364561081, 0.0120237663684845, 0.0216593053190025, 0.03850710451301957, -0.03979567810893059, 0.024909289553761482, 0.003612052416289236, 0.03026977797636988, 0.03532775491476059, 0.04048445075750351, -0.02123659290733253, 0.05895552039146423, 0.04913758486509323, -0.047305576503276825, 0.05272323951426506, 0.01215427809438705, -0.02513653226196766, -0.0105082932010293, -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, -0.0160540884263323, -0.1282315403220337, 0.005963715258985758, -0.01607099547982216...]
```

**Vector**

# What is an Embedding?

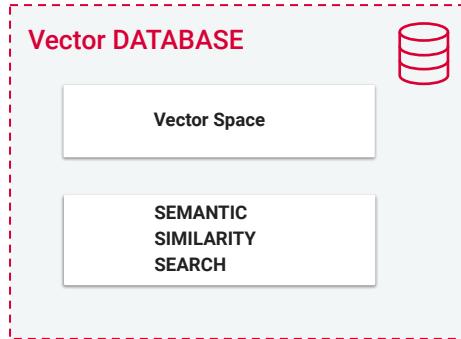
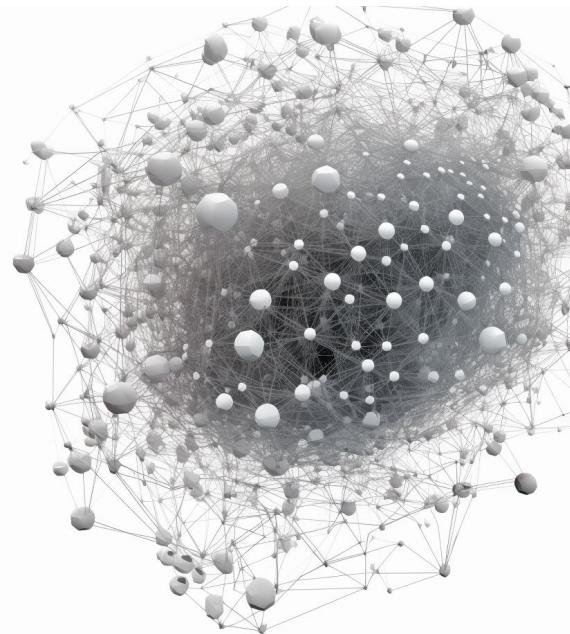
Take source objects (text, images, sound, movies) and create Vector format representation of the context.  
This allows for Similarity Search on the Database finding Semantically comparable objects.



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

# ➤ What is 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**

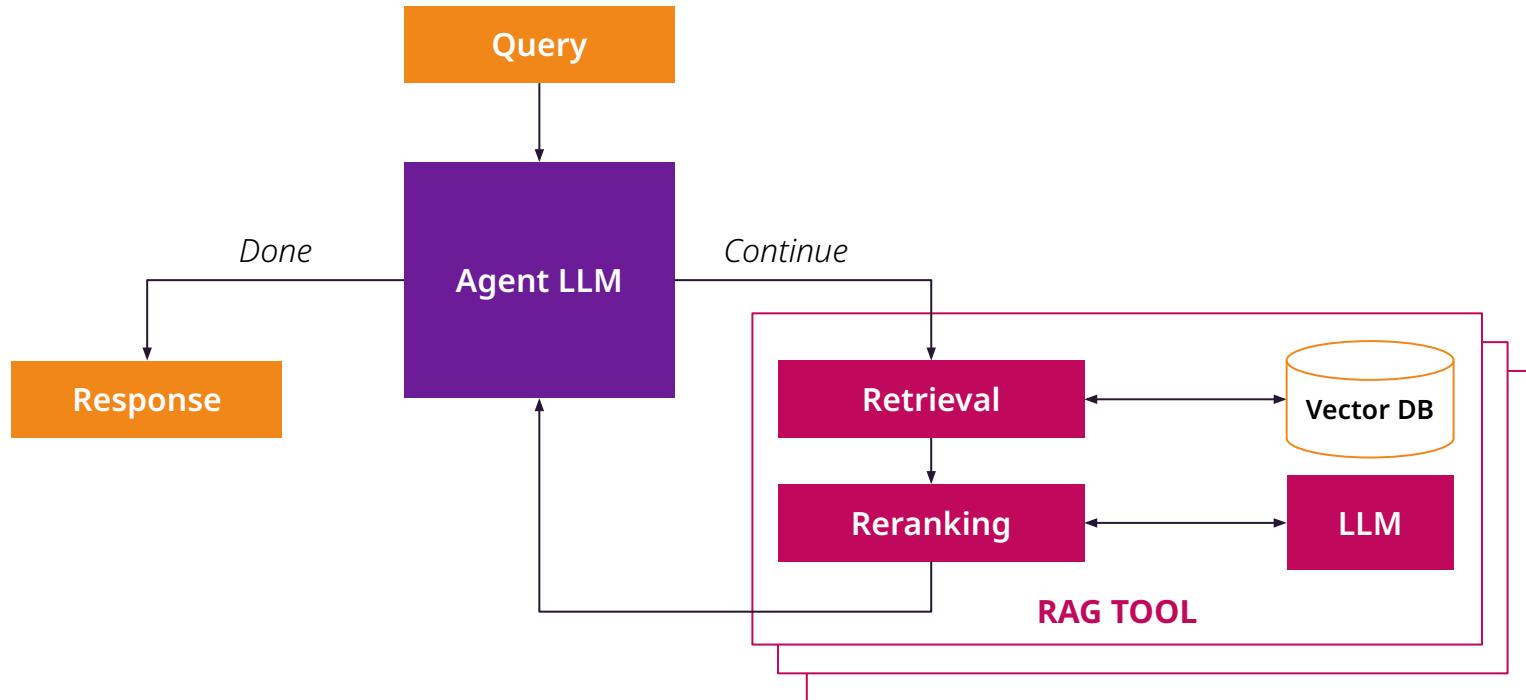
**On-Prem and Managed**

**All CSPs**

## › Enterprise challenges



# RAG is key to GenAI, yet it is **hard** for enterprises



# Why Is RAG Hard For Enterprises?



Sudden changes in behaviour,  
what caused it?



Physics Wallah



We're in a regulated vertical,  
we need supported software



Due to regulations we can not  
share sensitive data externally

**Large Dutch Retail Bank**

# Why Is RAG Hard For Enterprises?

## Gen AI apps break often

- Lots and lots of dependencies
- External APIs without strict guarantees
- External APIs release on their own schedule

## Enterprises need a proven path to success, which is NOT

- Depending on *unsupported* open source projects
- Small vendors, whom often don't want to work with them



“Every company building with generative AI right now is looking for answers about the most effective way to implement RAG within their applications,” said Harrison Chase, CEO, LangChain.

**“DataStax has recognized a pain point in the market and is working to remedy that problem with the release of RAGStack. Using top-choice technologies, like LangChain and Astra DB among others, Datastax is providing developers with a tested, reliable solution made to simplify working with LLMs.”**



See the joint webinar



**Harrison Chase**  
CEO, LangChain



# RAGStack: What does it package?



## LangChain

### Curated software components

RAGStack includes a selection of the best open-source software for implementing RAG from DataStax, LangChain and LLamaIndex, reducing the overwhelming choices that developers often face.



## LLamaIndex

### Advanced RAG techniques

RAGStack implements advanced orchestration techniques such as Chain of Thought and Multi-Query RAG on top of [Astra DB's Vector Search](#)

## Astra DB

### LangSmith and LangServe compatibility

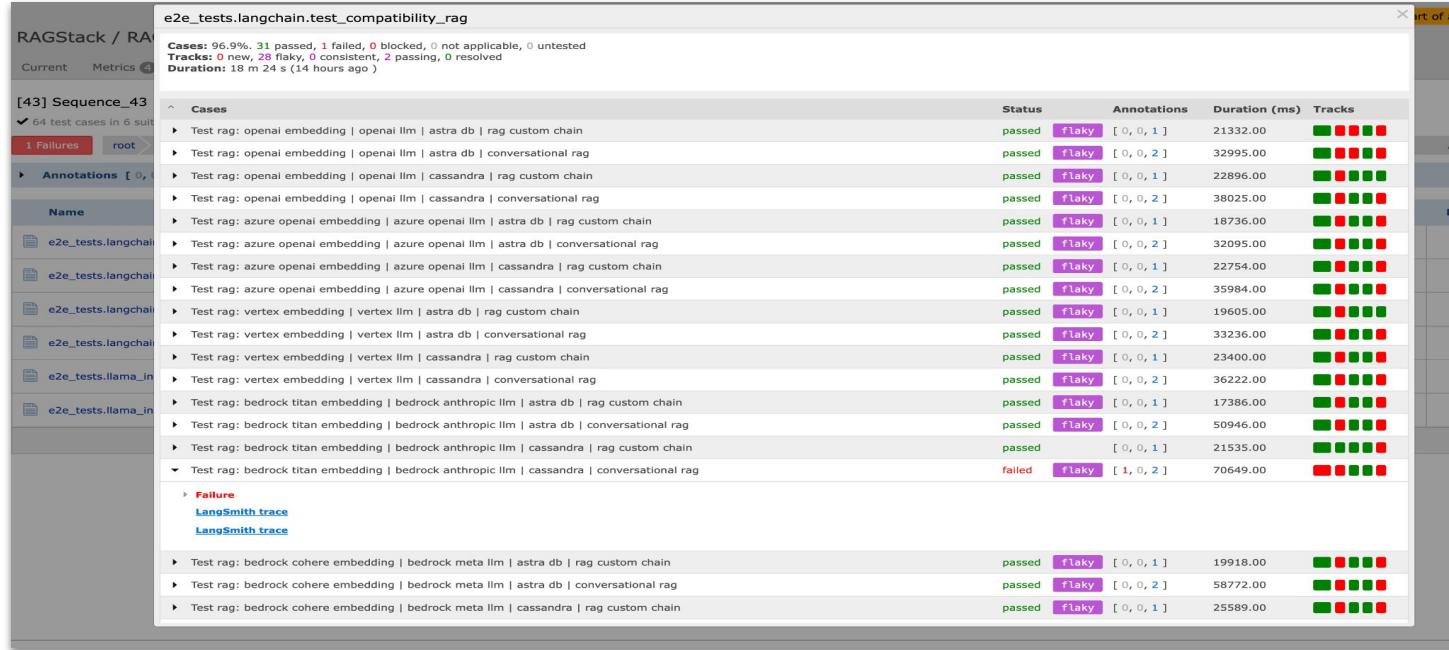
You can use RAGStack with the hosted services from LangChain to add tracing ([LangSmith](#)) and hosting (LangServe) to your production GenAI application

## Astra STREAMING

### Streaming

RAGStack packages LangStream, which combines the best of event-based architectures with the latest Gen AI technologies. With LangStream you can build streaming Gen AI applications in just 10 lines of code.

# Extensive test harness



**e2e\_tests.langchain.test\_compatibility\_rag**

Cases: 96.9% 31 passed, 1 failed, 0 blocked, 0 not applicable, 0 untested  
Tracks: 0 new, 28 flaky, 0 consistent, 2 passing, 0 resolved  
Duration: 18 m 24 s (14 hours ago)

Name	Status	Annotations	Duration (ms)	Tracks
Test rag: openai embedding   openai llm   astra db   rag custom chain	passed	flaky [ 0, 0, 1 ]	21332.00	██████████
Test rag: openai embedding   openai llm   astra db   conversational rag	passed	flaky [ 0, 0, 2 ]	32995.00	██████████
Test rag: openai embedding   openai llm   cassandra   rag custom chain	passed	flaky [ 0, 0, 1 ]	22896.00	██████████
Test rag: openai embedding   openai llm   cassandra   conversational rag	passed	flaky [ 0, 0, 2 ]	38025.00	██████████
Test rag: azure openai embedding   azure openai llm   astra db   rag custom chain	passed	flaky [ 0, 0, 1 ]	18736.00	██████████
Test rag: azure openai embedding   azure openai llm   astra db   conversational rag	passed	flaky [ 0, 0, 2 ]	32095.00	██████████
Test rag: azure openai embedding   azure openai llm   cassandra   rag custom chain	passed	flaky [ 0, 0, 1 ]	22754.00	██████████
Test rag: azure openai embedding   azure openai llm   cassandra   conversational rag	passed	flaky [ 0, 0, 2 ]	35984.00	██████████
Test rag: vertex embedding   vertex llm   astra db   rag custom chain	passed	flaky [ 0, 0, 1 ]	19605.00	██████████
Test rag: vertex embedding   vertex llm   astra db   conversational rag	passed	flaky [ 0, 0, 2 ]	33236.00	██████████
Test rag: vertex embedding   vertex llm   cassandra   rag custom chain	passed	flaky [ 0, 0, 1 ]	23400.00	██████████
Test rag: vertex embedding   vertex llm   cassandra   conversational rag	passed	flaky [ 0, 0, 2 ]	36222.00	██████████
Test rag: bedrock titan embedding   bedrock anthropic llm   astra db   rag custom chain	passed	flaky [ 0, 0, 1 ]	17386.00	██████████
Test rag: bedrock titan embedding   bedrock anthropic llm   astra db   conversational rag	passed	flaky [ 0, 0, 2 ]	50946.00	██████████
Test rag: bedrock titan embedding   bedrock anthropic llm   cassandra   rag custom chain	passed	flaky [ 0, 0, 1 ]	21535.00	██████████
Test rag: bedrock titan embedding   bedrock anthropic llm   cassandra   conversational rag	failed	flaky [ 1, 0, 2 ]	70649.00	███████
Failure				
LangSmith trace				
LangSmith trace				
Test rag: bedrock cohore embedding   bedrock meta llm   astra db   rag custom chain	passed	flaky [ 0, 0, 1 ]	19918.00	██████████
Test rag: bedrock cohore embedding   bedrock meta llm   astra db   conversational rag	passed	flaky [ 0, 0, 2 ]	58772.00	██████████
Test rag: bedrock cohore embedding   bedrock meta llm   cassandra   rag custom chain	passed	flaky [ 0, 0, 1 ]	25589.00	██████████

<https://ragstack-ai.testspace.com>



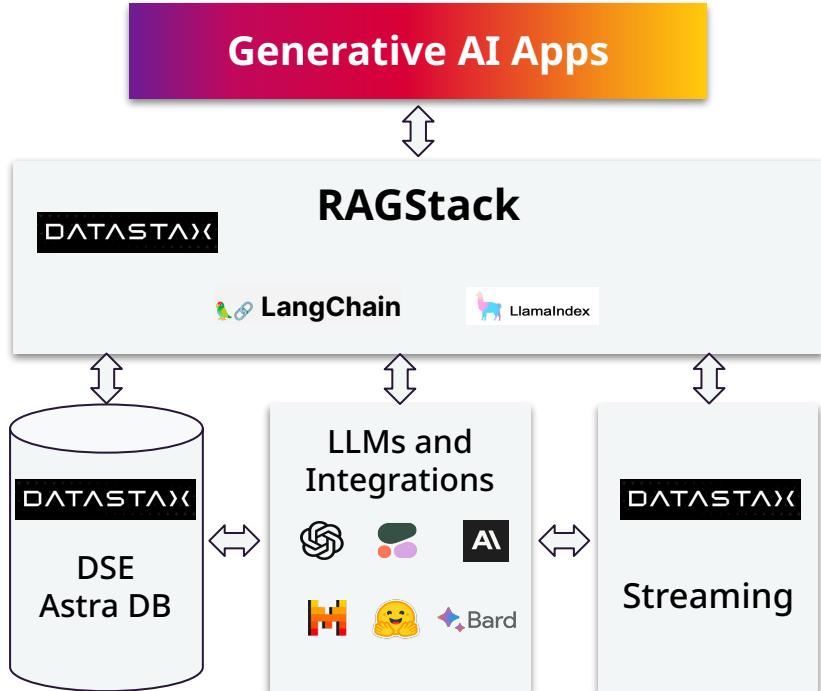
# Production GenAI

## Key Components

- LangChain/LlamaIndex for orchestration
- Astrypt/CassIO for Data Abstractions
- LangStream for Data pipelines

## Requirements

- Implement SOTA RAG techniques proven work in production
- Each LLM aggressively tested to meet enterprise requirements





# Everybody wins

## Developers

- Python
- LangChain or LlamaIndex
- Core competencies in app development
- Not open source maintainers or contributors

## Other stakeholders

- Security scans, confidence for compatibility
- Long term support releases
- 24x7x365 enterprise support



# It's a strategic alternative to...

## Build your own

Fork open source and take on tech debt and maintenance

## Closed stacks

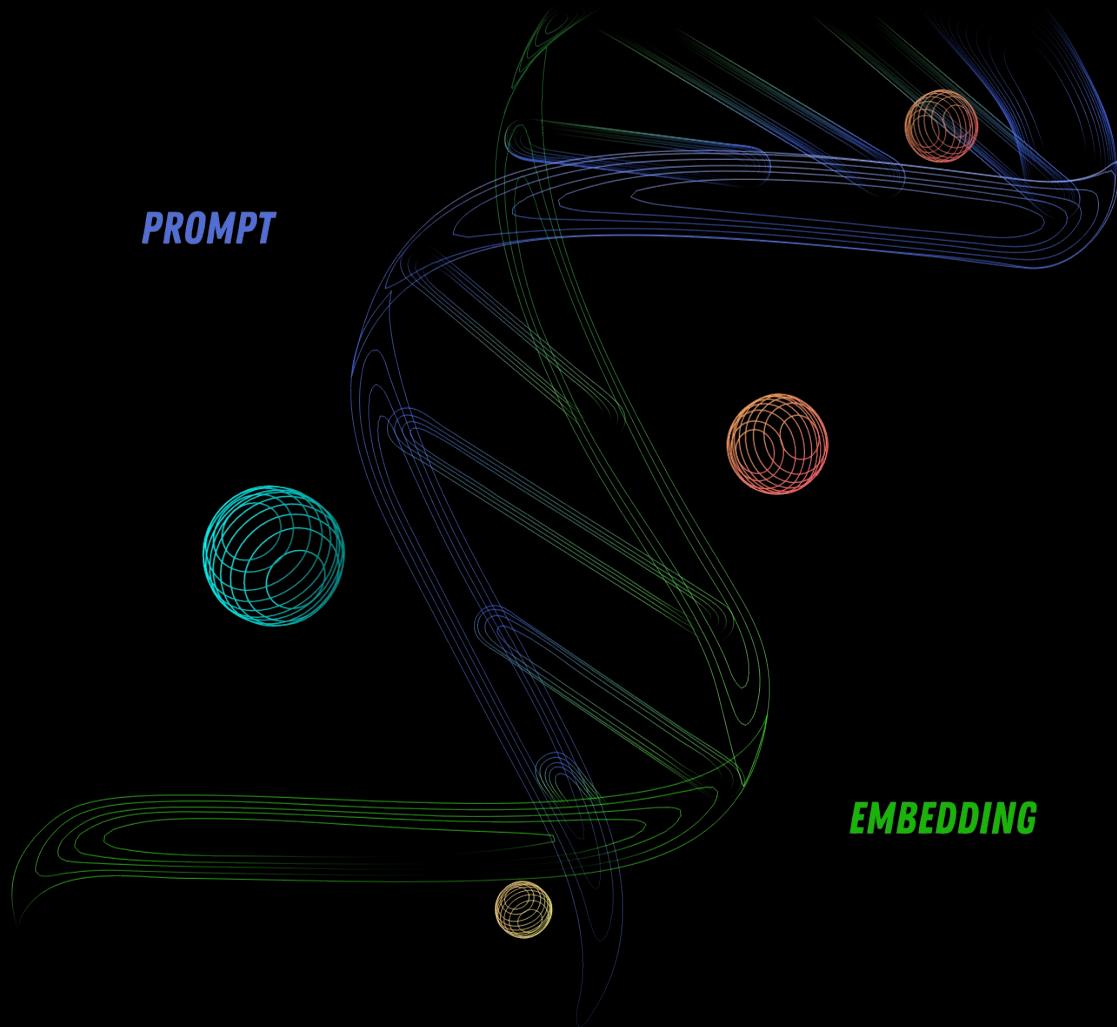
Open AI  
RAG as a service

# Our Mission

**Jina AI**'s mission is to lead the advancement of multimodal AI through **embedding models and prompt-engineering technologies**, focusing on NLP, image and video analysis, and cross-modal data interaction.

This lets us provide solutions that turn complex, multi-source data into actionable insights and applications.

**PROMPT**



**EMBEDDING**

# Jina Embeddings

- 8192 token input context
- State-of-the-Art Embedding and Reranking models
- Bilingual retrieval for:
  - German/English
  - Chinese/English
  - Spanish/English
  - More languages to come!
- Code Embeddings for 30 programming languages and frameworks and natural English

# Embeddings and Rerankers

- Scalable API – 20% the cost of OpenAI
- Native integration with vector databases and RAG frameworks  
*including DataStax, LangChain, LlamaIndex*
- 1 million token free trial (tokens at [jina.ai](https://jina.ai))



[jina.ai/embeddings](https://jina.ai/embeddings)



[jina.ai/reranker](https://jina.ai/reranker)

Now let's build  
your own  
GenAI application

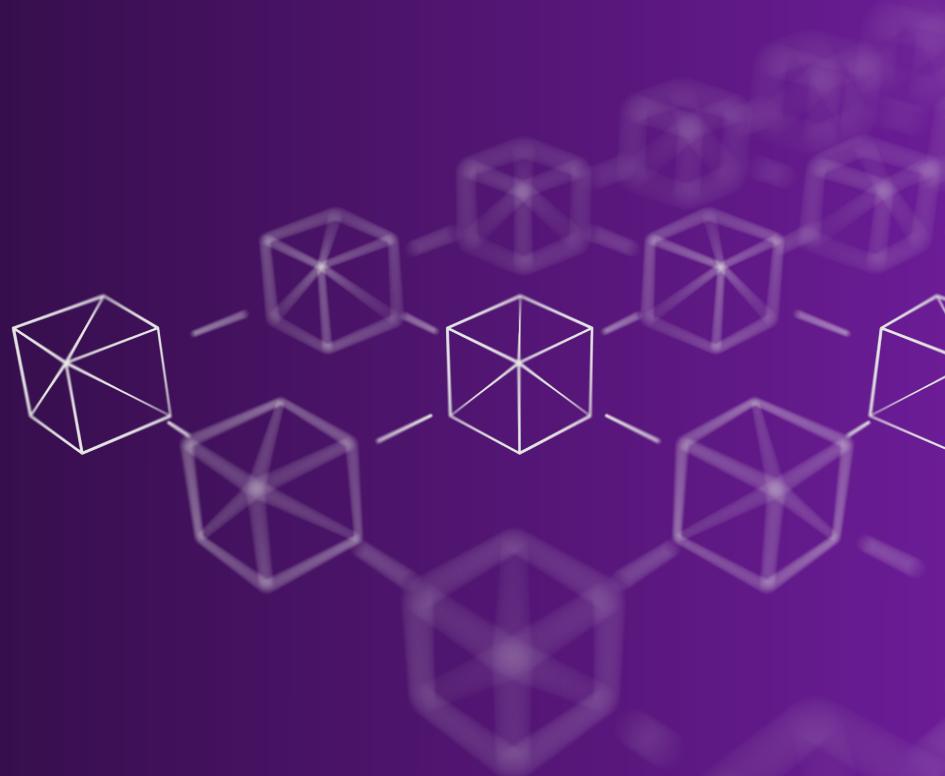


# › Getting started

- Go to  
<https://github.com/difli/build-your-own-rag-chatbot/tree/jina.ai>



› What will you  
do next?



# 1

## Enhanced Search

Semantic Search  
Document Search  
Image Search



# 2

## Enhanced Matching

Personalization  
Recommendation  
Pairing  
Fraud Detection



# 3

## Faster Analytics

Classification  
Filtering  
Segmentation



# 4

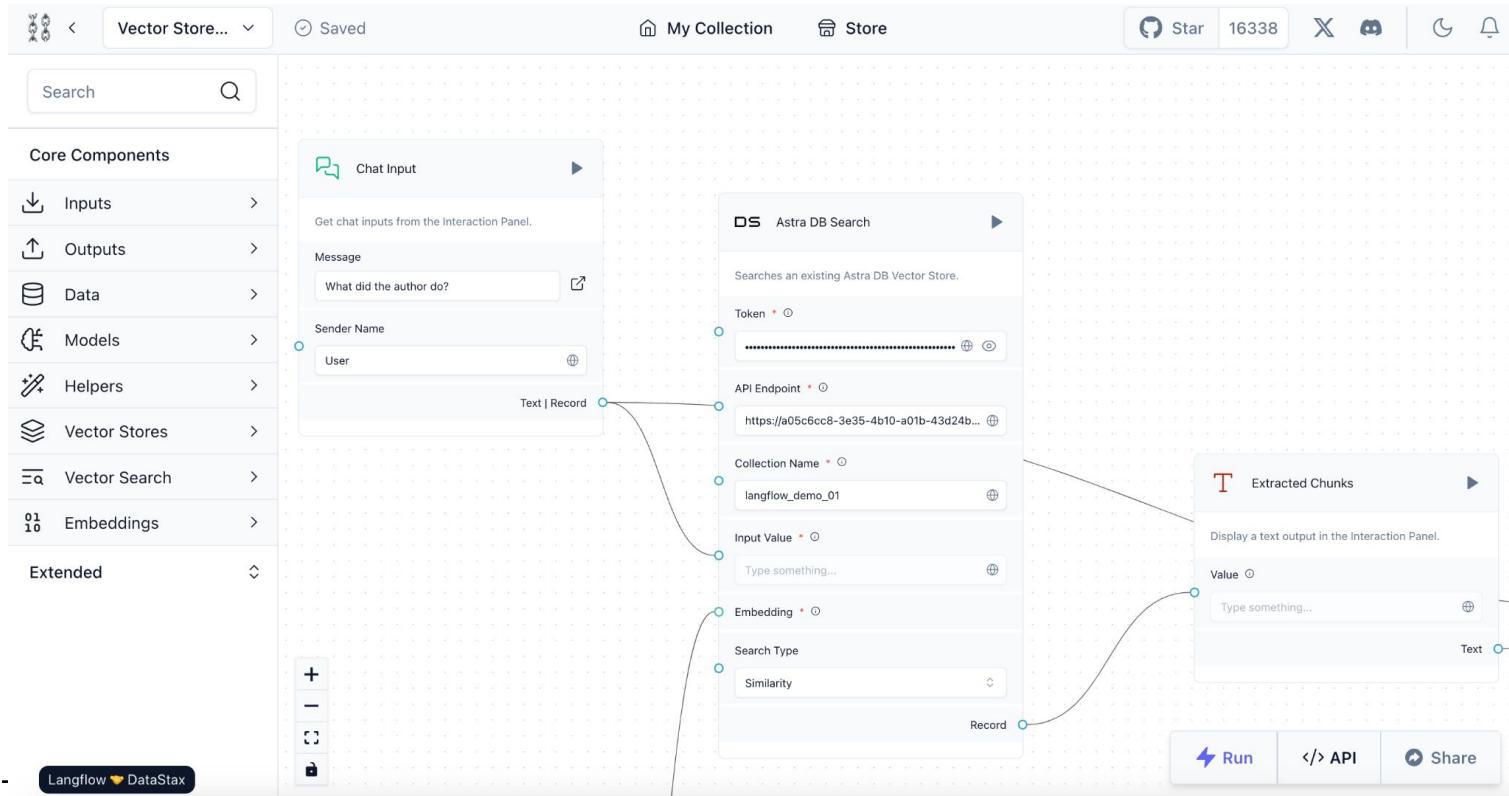
## Conversational Interfaces

Chatbots  
Voice Interfaces  
"Talk to your data"  
Bring  
Natural-Language  
Processing to Data

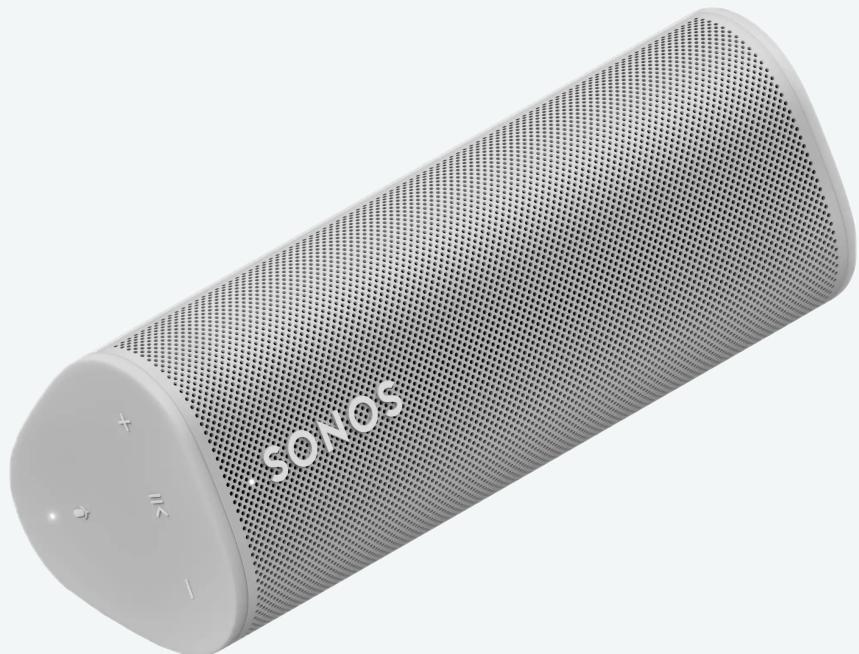


# DataStax Langflow - low code solution

<https://www.datastax.com/blog/datastax-acquires-langflow-to-accelerate-generative-ai-app-development>



# › Sonos Roam raffle time in 30 min!



**Submit to win:**

<https://bit.ly/rag-berlin>



# › Jina AI Swag: 3 Embedding Hoodies!



**Submit to win:**

<https://bit.ly/rag-berlin>





Thank  
You

DATASTAX