DATAVISION
by soda

Build an AI-powered data pipeline without vector databases

Bobur Umurzokov

AGENDA

1 **Intro**

GPT Limitations

2 **Common solutions**

Solutions to use in-context learning in AI responses

3 **LLM App**

Key features, how it works

4 **Demo**

Build an LLM App without a vector database

ABOUT ME

Bobur Umurzokov - Developer Advocate

- LinkedIn: Bobur Umurzokov
- Twitter: @BoburUmurzokov
- Instagram: @boburumurzokofficial
- Website: www.iambobur.com

LARGE LANGUAGE MODEL

Structured data

Text

Voice

3D signals

Images

Training

Adaptation

Information extraction

Instruction following

Object recognition

Image captioning

Q&A

Sentiment analysis

GPT LIMITATIONS

GPT excels at answering questions, but only on topics it remembers from its training data.

UNFAMILIAR TOPICS

- Recent events after Sep 2021.
- Your non-public documents.
- Information from past conversations.
- Real-time data.
- etc.

CHALLENGES OF PRODUCTIONIZING CHATGPT

01 The ambiguity of natural languages

02 Cost and latency

03 Backward and forward compatibility

04 Offline Evaluation

LET'S TEST IT

Can you find me discounts this week for Adidas men's shoes?

GPT-3.5 GPT-4

ChatGPT PLUS

Show me a code snippet
of a website's sticky header

Write a thank-you note
to our babysitter for the last-minute help

Give me ideas
for what to do with my kids' art

Suggest some names
for my cafe-by-day, bar-by-night business

Send a message

PROMPT ENGINEERING

Given the following discounts data: {input_data} answer this query: {user_query}

GPT-3.5 GPT-4

Press F11 to exit full screen

ChatGPT PLUS

Plan a trip
to experience Seoul like a local

Help me pick
an outfit that will look good on camera

Design a database schema
for an online merch store

Show me a code snippet
of a website's sticky header

"current_price_upper": {"value": 121.7, "currency": "USD", "symbol": "$", "raw": "121.7"}, "current_price": {"value": 14.8, "currency": "USD", "symbol": "$", "raw": "14.8 - 121.7", "name": "Current Price"}, "merchant_name": "Amazon.com", "free_shipping": false, "is_prime": true, "is_map": false, "deal_id": "e1b91f60", "seller_id": "ATVPDKIKX0DER", "description": "adidas are on sale for limited time only. Valid while supplies last and when shipped & sold by Amazon.com. Discount reflected in current price.", "rating": 4.6, "ratings_total": 1448, "page": 1, "old_price": 22, "currency": "USD"}

and can you find me discounts this week for Adidas men shoes?

OpenAI API

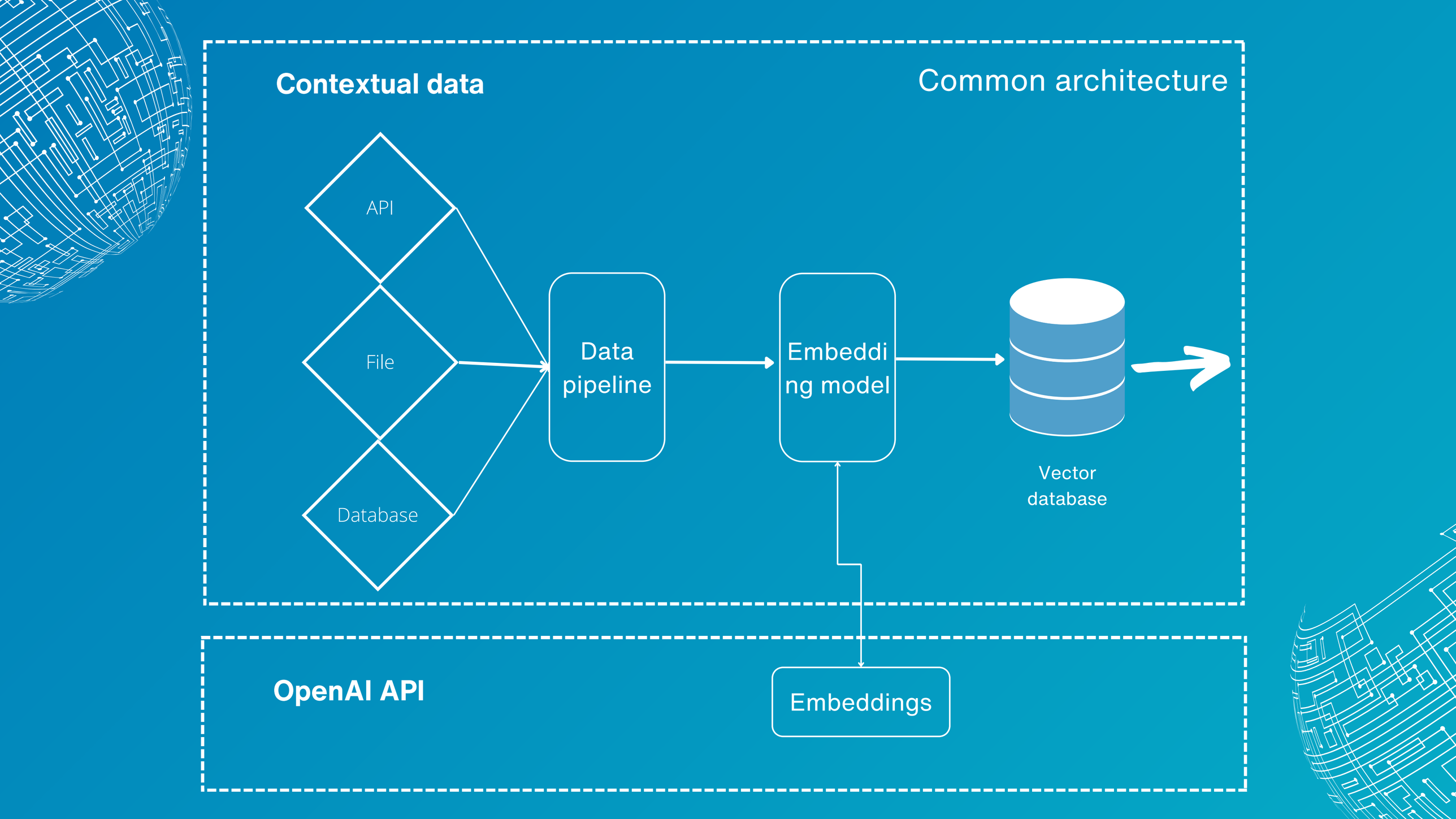
MAKES AVAILABLE POWERFUL AI MODELS FOR DEVELOPERS THROUGH API.

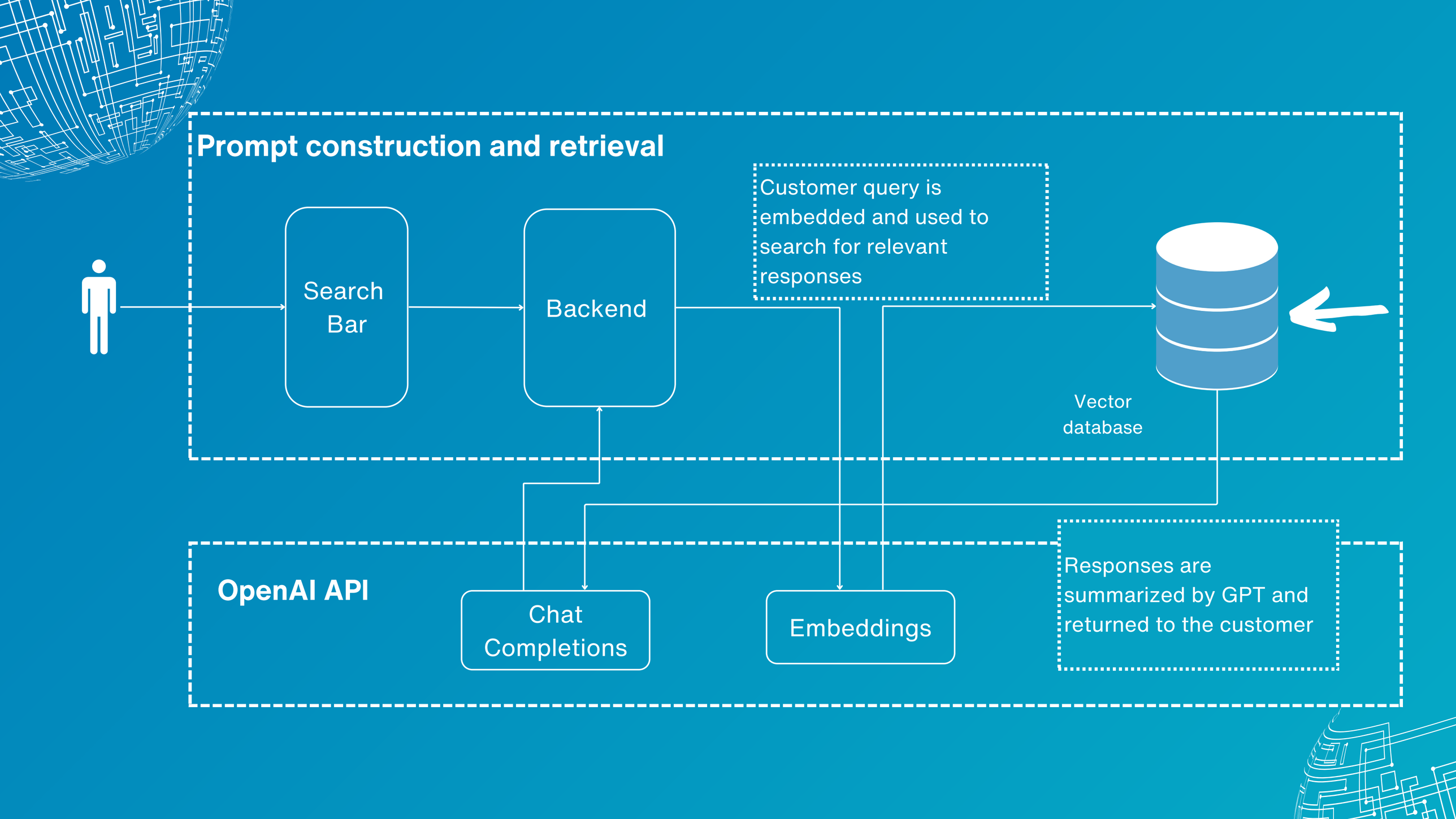
01 EMBEDDINGS

Text embeddings measure the relatedness of text strings.

02 CHAT COMPLETIONS

Take a list of messages as input and return a model-generated message as output.


Contextual data
Common architecture
API
File
Database
Data pipeline
Embeddi ng model
Vector database
OpenAI API
Embeddings


Prompt construction and retrieval
Search Bar
Backend
Customer query is embedded and used to search for relevant responses
Vector database
OpenAI API
Chat Completions
Embeddings
Responses are summarized by GPT and returned to the customer

NO VECTOR DATABASE

REAL-TIME DOCUMENT INDEXING PIPELINE

With pathway LLM App

Pathway - is an application layer in Python for batch and streaming data processing

LLM App- is a framework in Python to build AI Apps

LLM-enabled data pipeline for real-time data

UNIFIED DATA PIPELINES

Data Ingestion

Includes Data cleaning

Storage & Processing

SQL databases

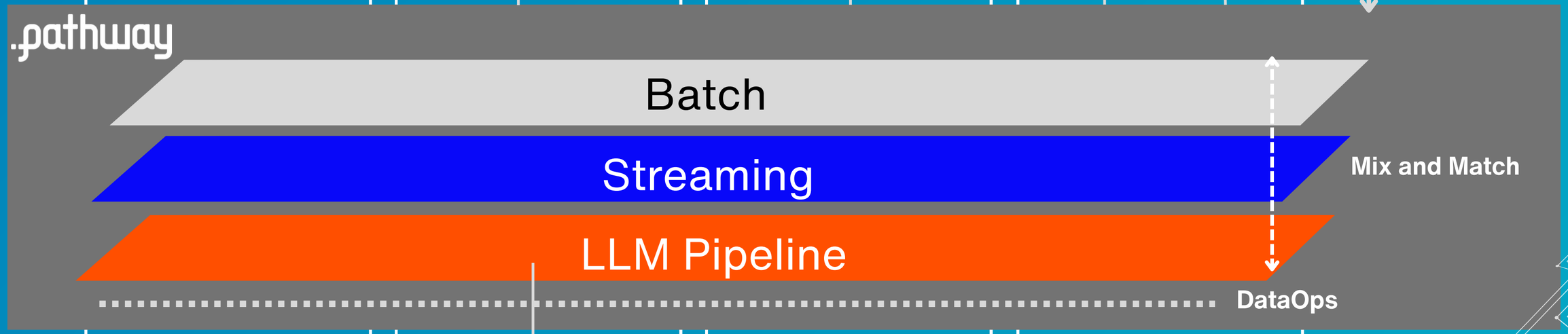
Transformation & Modeling

Includes Realtime ML

BI, Analytics & Data Products

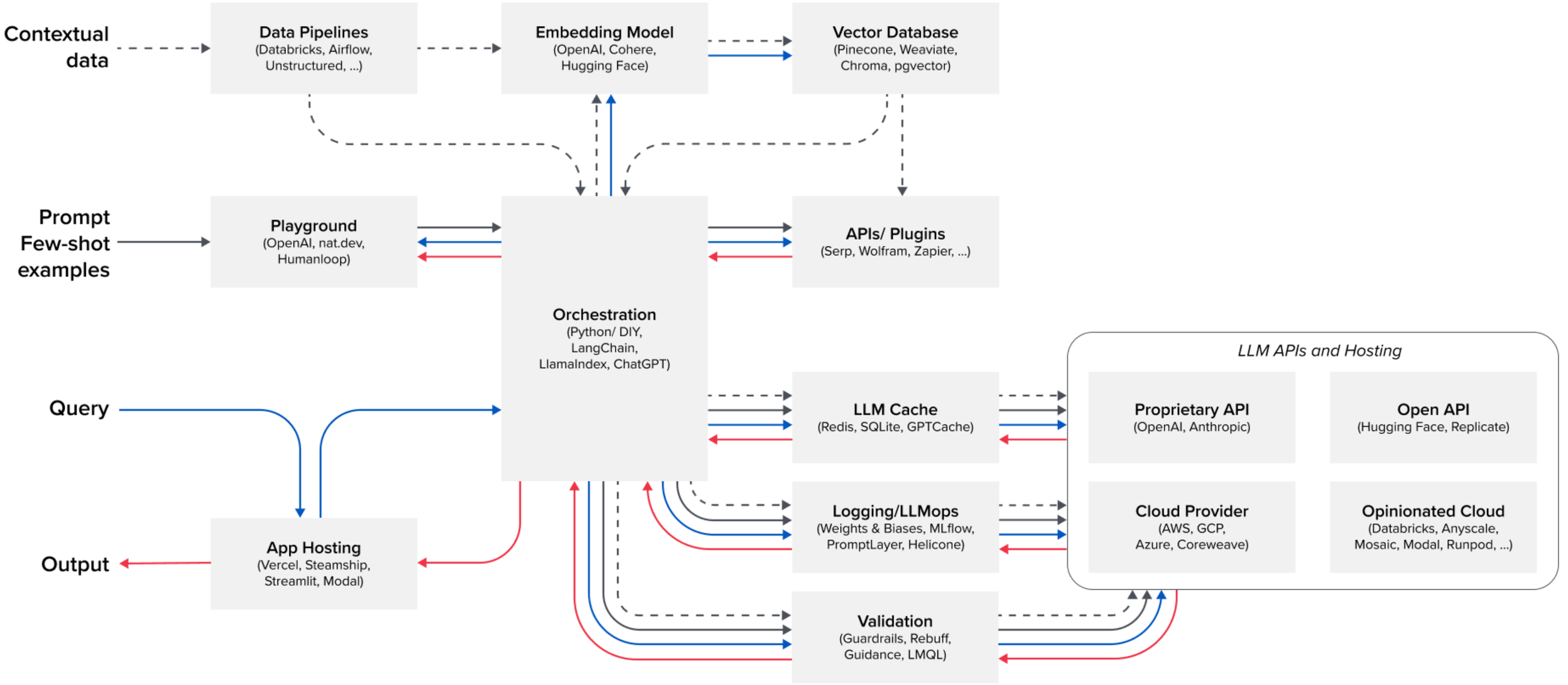
Includes Realtime ML

API



Vector Index

Emerging LLM App Stack



HOW I UNDERSTAND LLM APP

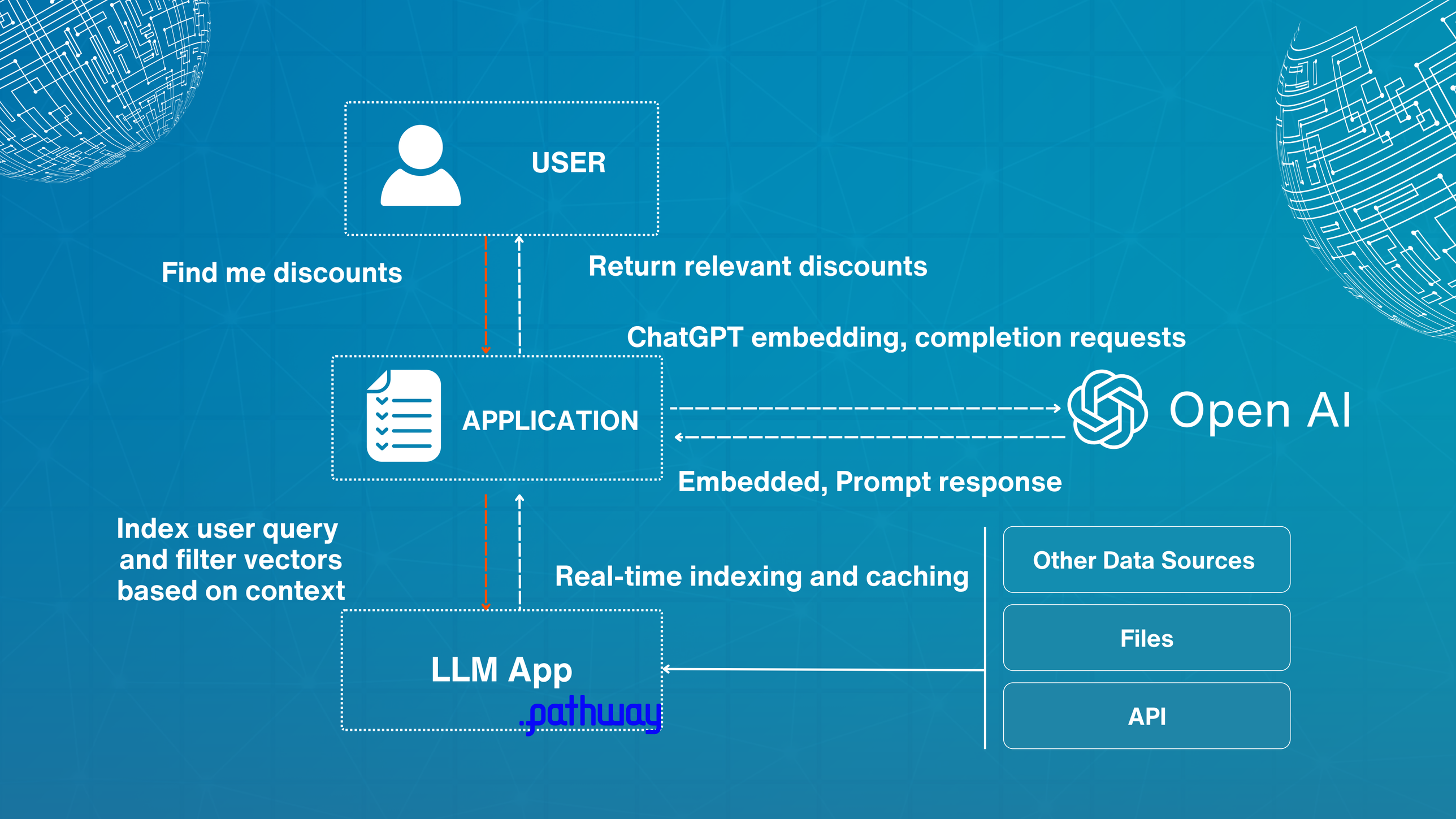
LLM App = ~~Pinecone/Weaviate~~ + ~~Langchain~~ + ~~Redis~~ + ~~FastAPI~~

Vector databases

LLM frameworks

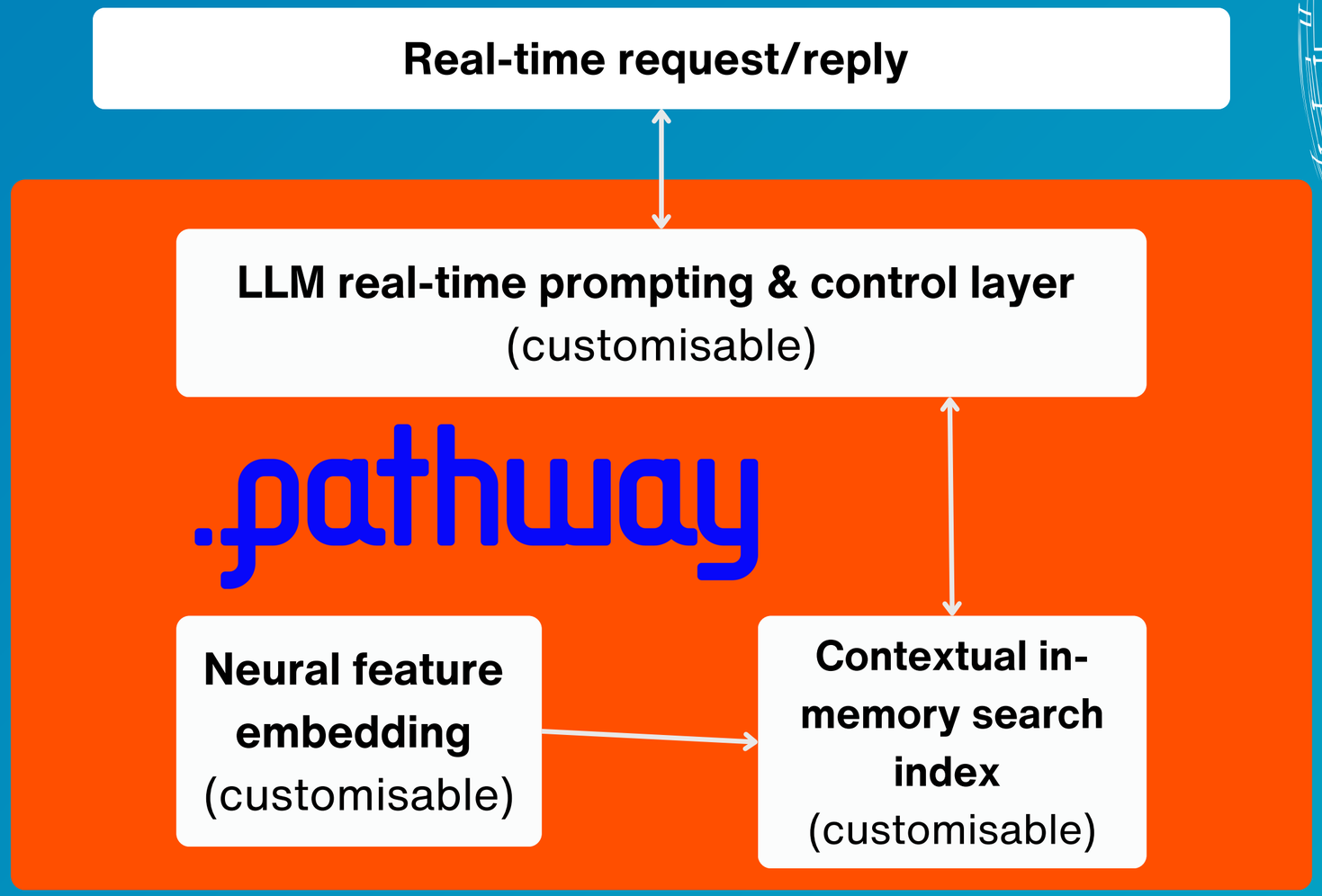
LLM Caches

API builders


USER
Find me discounts
Return relevant discounts
ChatGPT embedding, completion requests
APPLICATION
Open AI
Embedded, Prompt response
Index user query and filter vectors based on context
Real-time indexing and caching
LLM App
pathway
Other Data Sources
Files
API

VECTOR INDEXING ADVANTAGES

- **Data Privacy:** Keeps original data secure and undisturbed, minimizing data exposure risk.
- **Cost-Efficiency:** Reduces costs associated with extra storage, computing power, and licensing.
- **Scalability:** Simplifies scaling by decreasing the number of components to manage.


Real-time request/reply
LLM real-time prompting & control layer
(customisable)
.pathway
Neural feature embedding
(customisable)
Contextual in-memory search index
(customisable)

KEY FEATURES

- Built-in connectors: CSV, JSONLines, streaming APIs, and more.
- User’s role-specific response.
- Data and code reusability for offline evaluation.
- Model testing in static mode.
- Local Machine Learning models.
- Live data sources: Kafka, Debezium, Redpanda, and more.
- Join data from multiple data sources.

To learn more about advanced features see: Features for Organizations.

FIND DISCOUNTS APP GITHUB REPO

Scan me

TRY IT YOURSELF

Scan me

BUILD YOUR OWN PATHWAY-POWERED LLM APP

*Simply add **llm-app** to your Python project's dependencies.*

Scan me

TAKEAWAYS

- Automate knowledge insertion with embeddings-based search and chat completion APIs.
- Simplify infrastructure, increase performance, and remove complexity with real-time indexing.
- Use the LLM App to build an LLM-enabled data pipeline.
- Use Pathway to extend streaming architectures.

ThankYou

Bobur Umurzokov

@BoburUmurzokov

@Boburmirzo