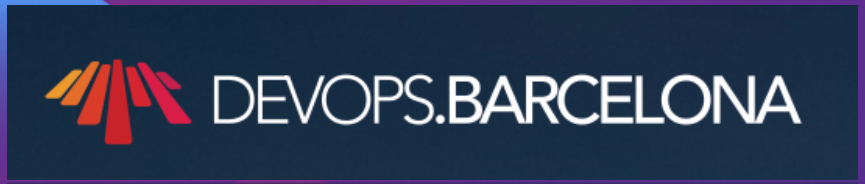
GENERATIVE AI: BLISS OR MISS?

DIANA TODEA - Senior Site Reliability Engineer



14TH OF NOVEMBER 2024

SRE focused on o11y

ML, AI, OSS enthusiast

support women in tech

Live in sunny Valencia  ... Spain

You can contact me:
LINKEDIN | **GITHUB**

01

CONCEPTS

02

OBSERVABILITY

03

AI ASSISTANTS

04

SRE WORKFLOWS

05

LESSONS LEARNED

06

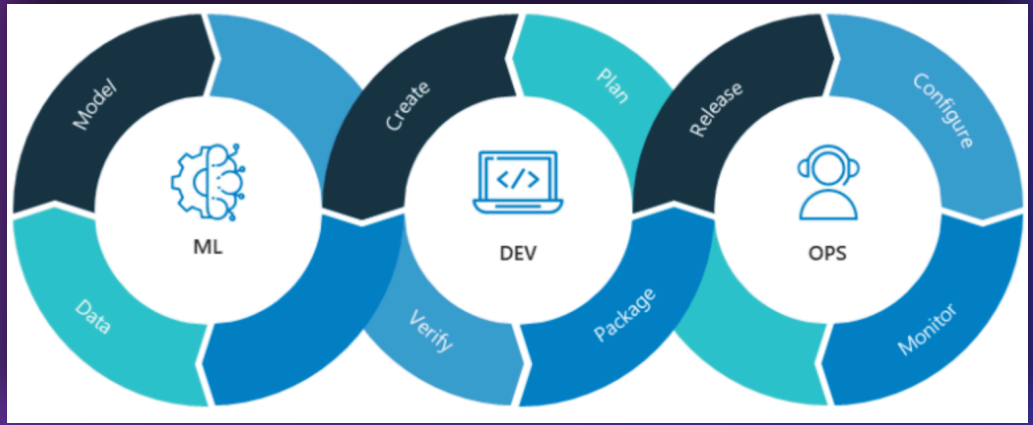
TAKEAWAYS

DISCLAIMER

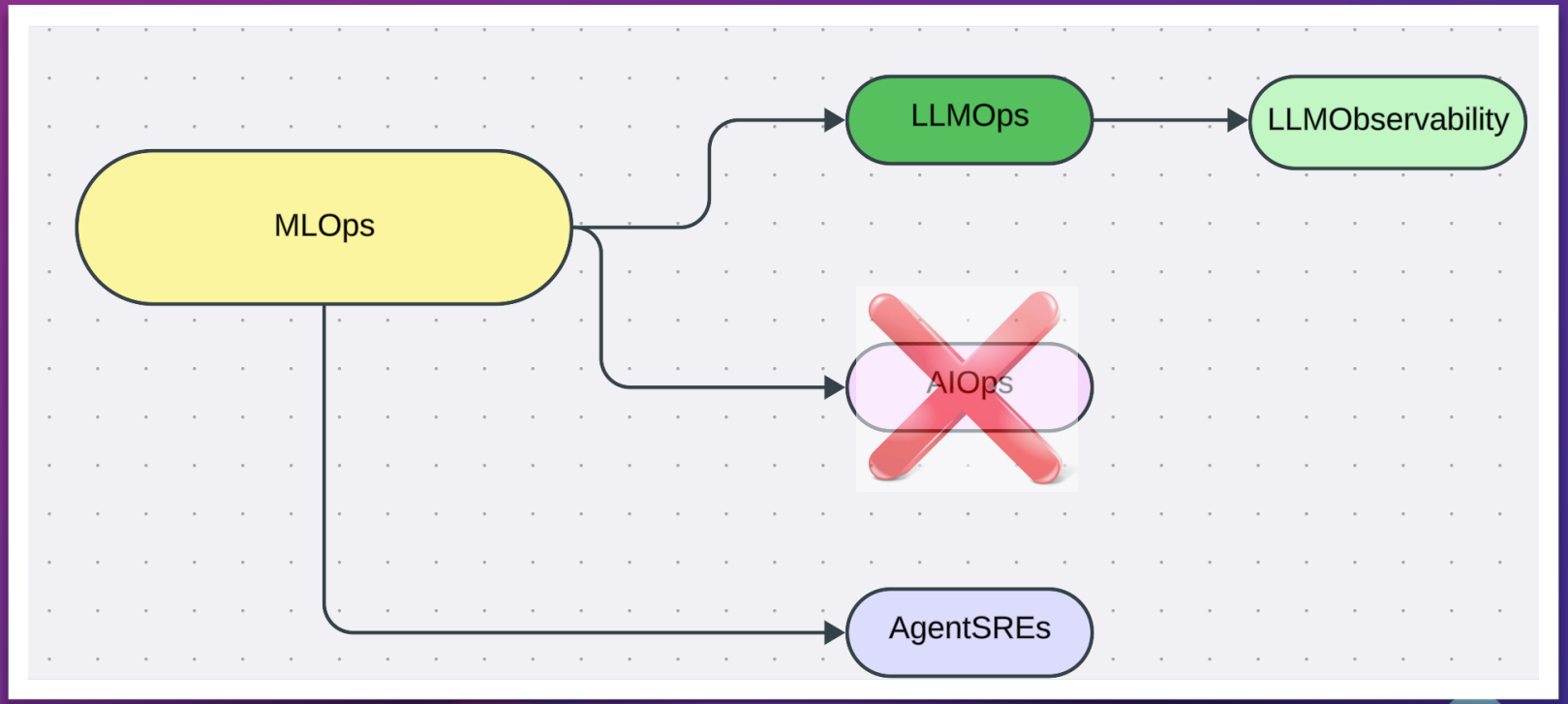
The opinions expressed in this presentation are solely my own and they do not represent my employer's.

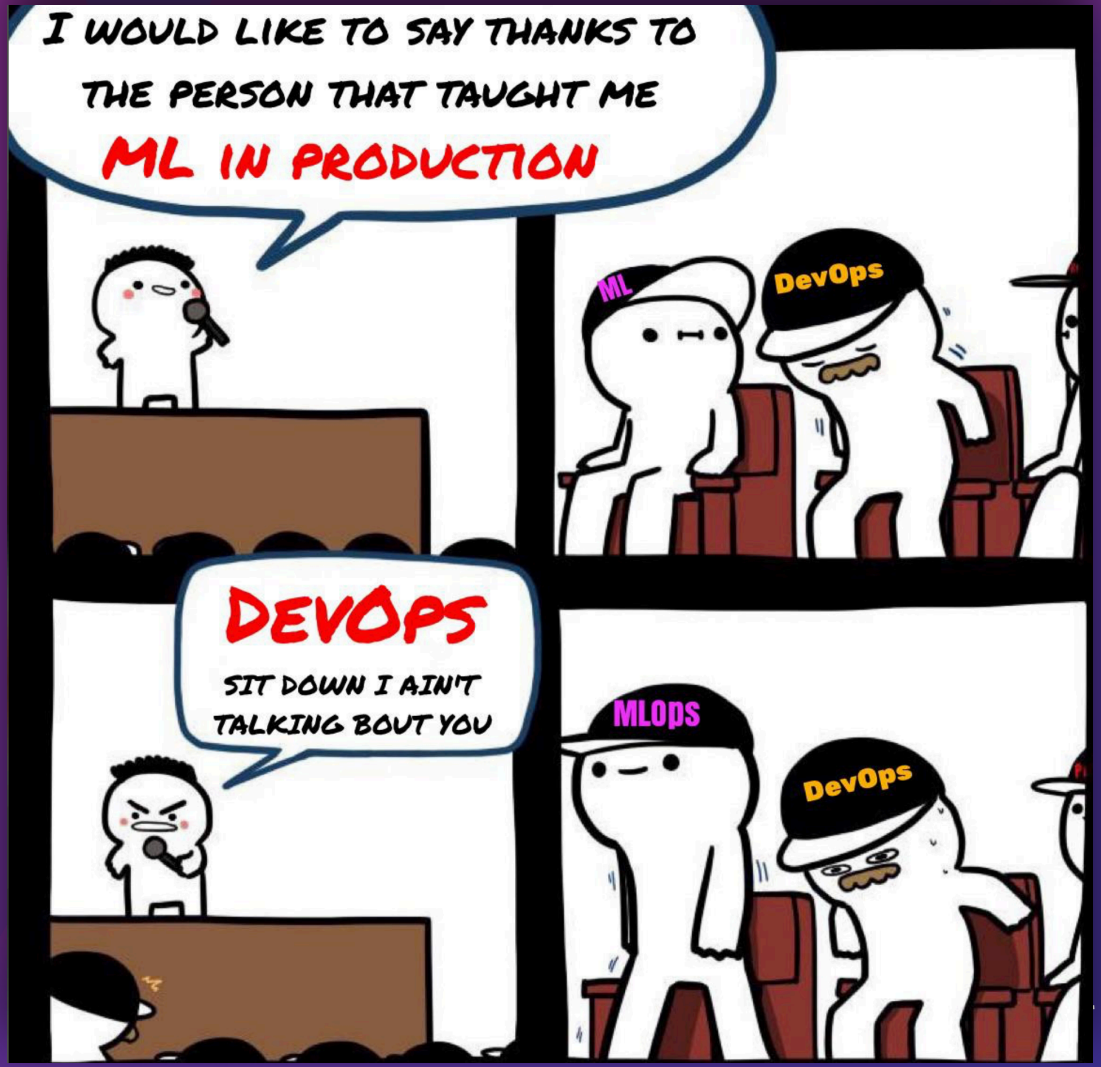
WHERE WE ARE

ML->DevOps->MLOps



Source: Nvidia


LLMOps
LLMObservability
MLOps
AIOps
AgentSREs


I WOULD LIKE TO SAY THANKS TO
THE PERSON THAT TAUGHT ME
ML IN PRODUCTION
ML
DevOps
DEVOPS
SIT DOWN I AIN'T
TALKING BOUT YOU
MLOps
DevOps

Concepts

TRANSFORMER

In machine learning, a transformer is a neural network that learns context and meaning by tracking relationships in sequential data like the words in the sentence.

ATTENTION MECHANISMS

Attention mechanisms allow a neural network to selectively weigh the importance of different input features so the model can focus on the most relevant parts of the input for a given task.

Concepts

VECTOR EMBEDDINGS

Vector embedding representations are suitable for common machine learning tasks such as clustering, recommendation and classification.

VECTOR DATABASE

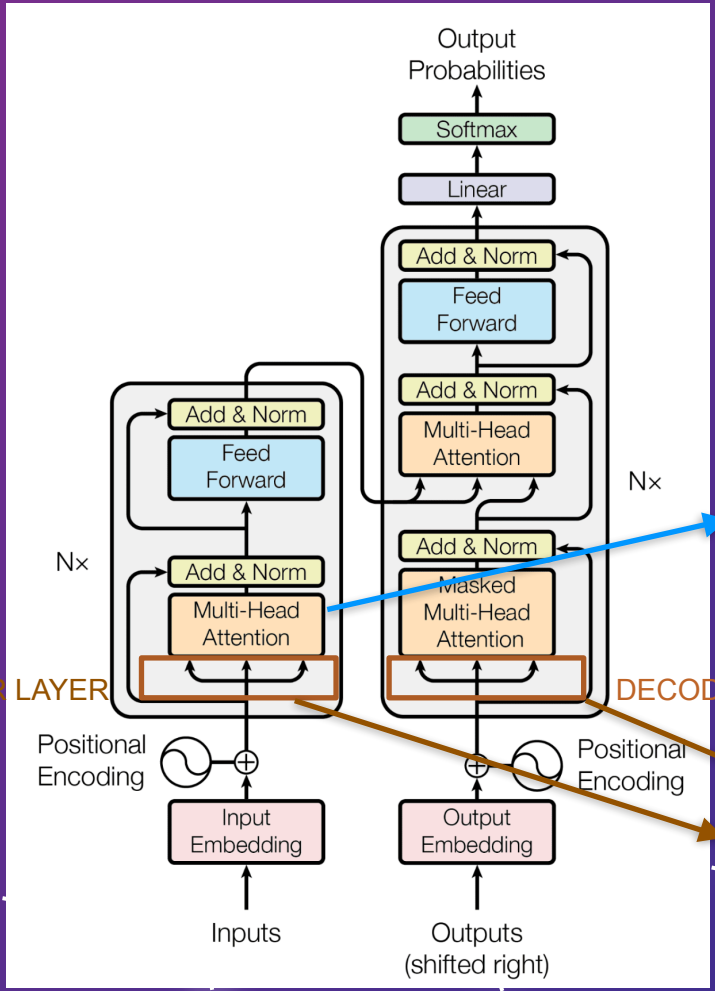
It's used for storing, updating, retrieving. We refer to retrieving set of vectors that are most similar to a query in a form of a vector that is embedded in the same Latent space.

ANN SEARCH

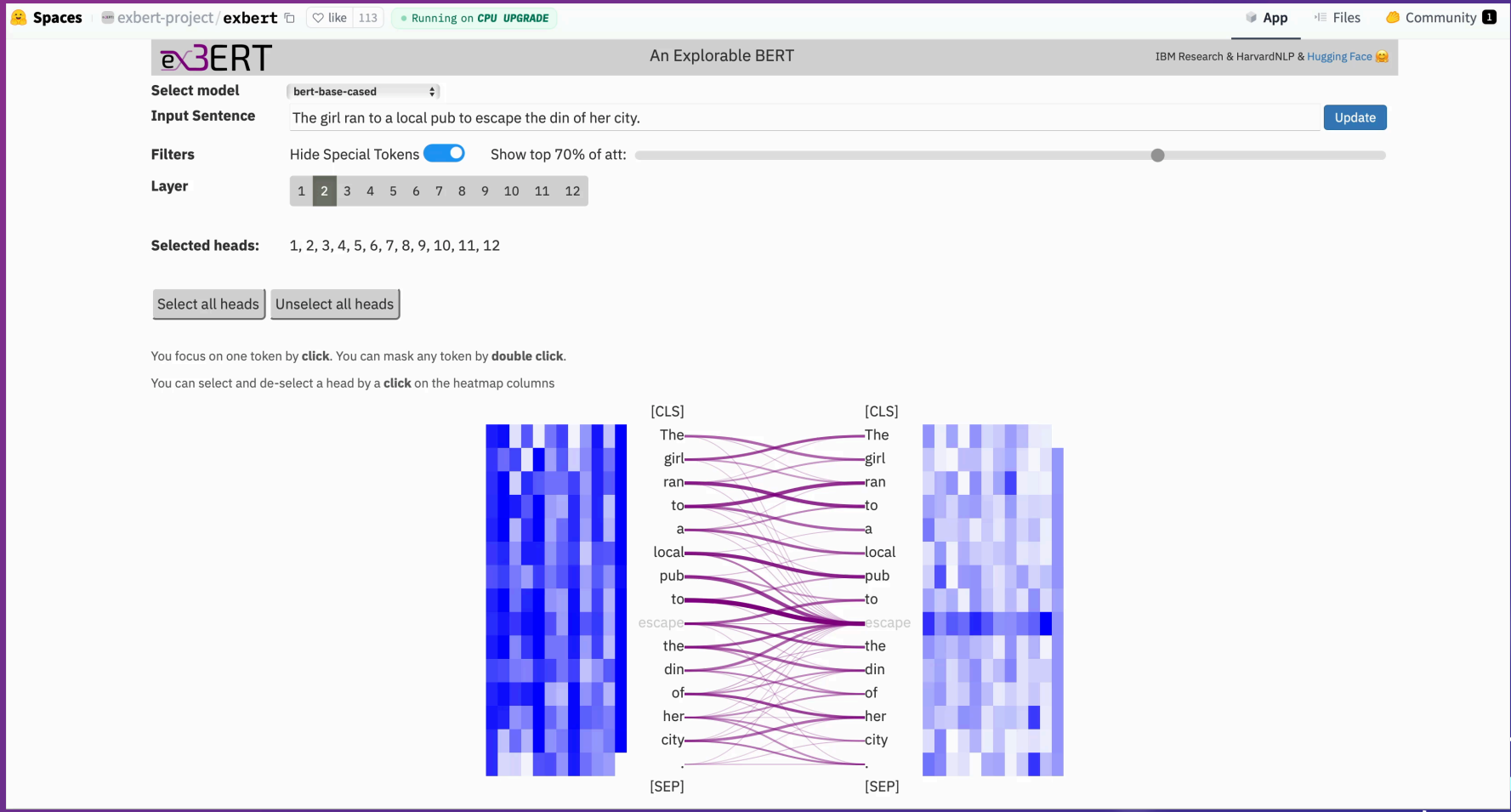
The retrieval procedure is called ANN (approximate nearest neighbor) search.

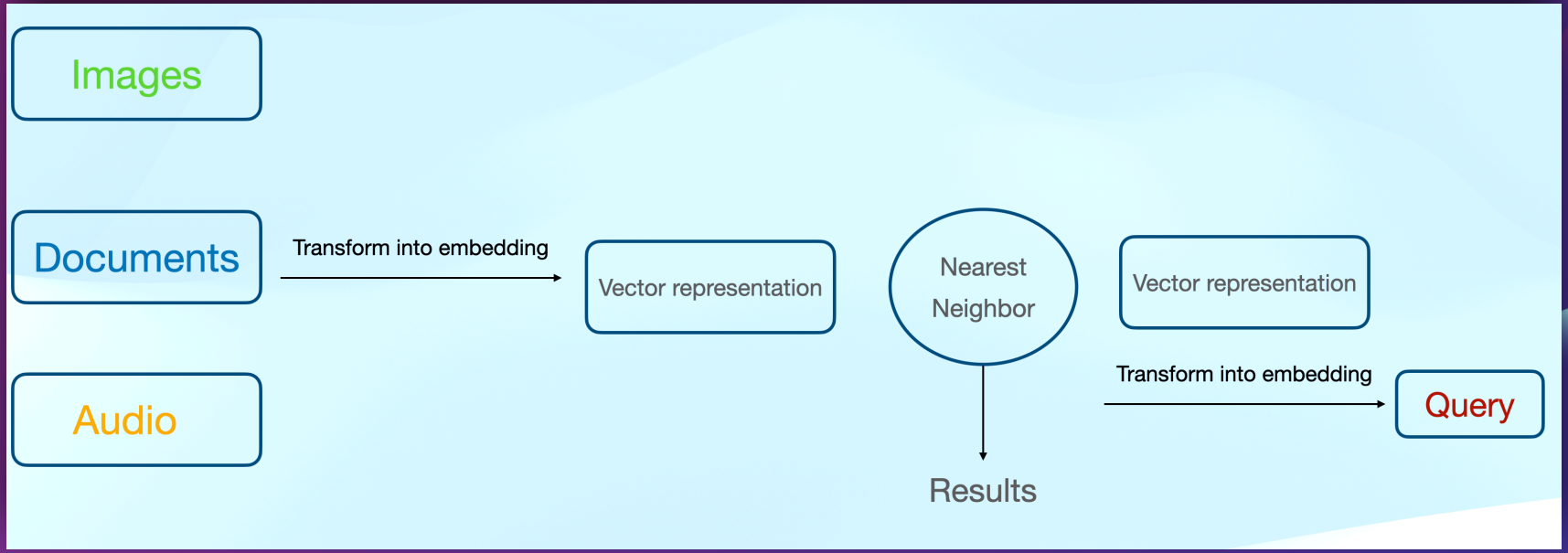
“ATTENTION IS ALL YOU NEED”

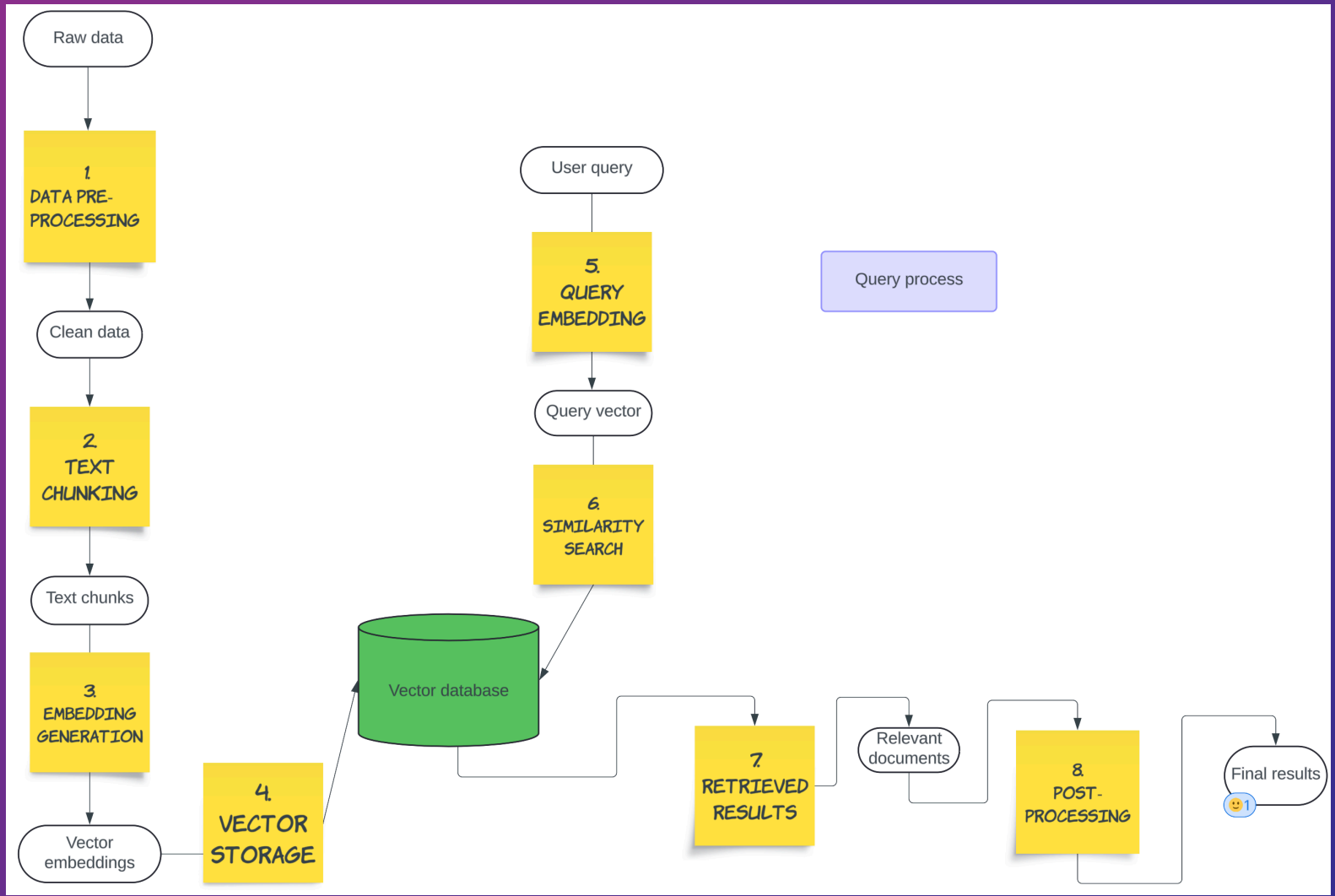
https://arxiv.org/abs/1706.03762v7, 2017



https://huggingface.co/spaces/exbert-project/exbert




Images
Documents
Audio
Transform into embedding
Vector representation
Nearest Neighbor
Results
Vector representation
Transform into embedding
Query


Raw data
1. DATA PRE-PROCESSING
Clean data
2. TEXT CHUNKING
Text chunks
3. EMBEDDING GENERATION
Vector embeddings
4. VECTOR STORAGE
Vector database
User query
5. QUERY EMBEDDING
Query vector
6. SIMILARITY SEARCH
Query process
7. RETRIEVED RESULTS
Relevant documents
8. POST-PROCESSING
Final results
1

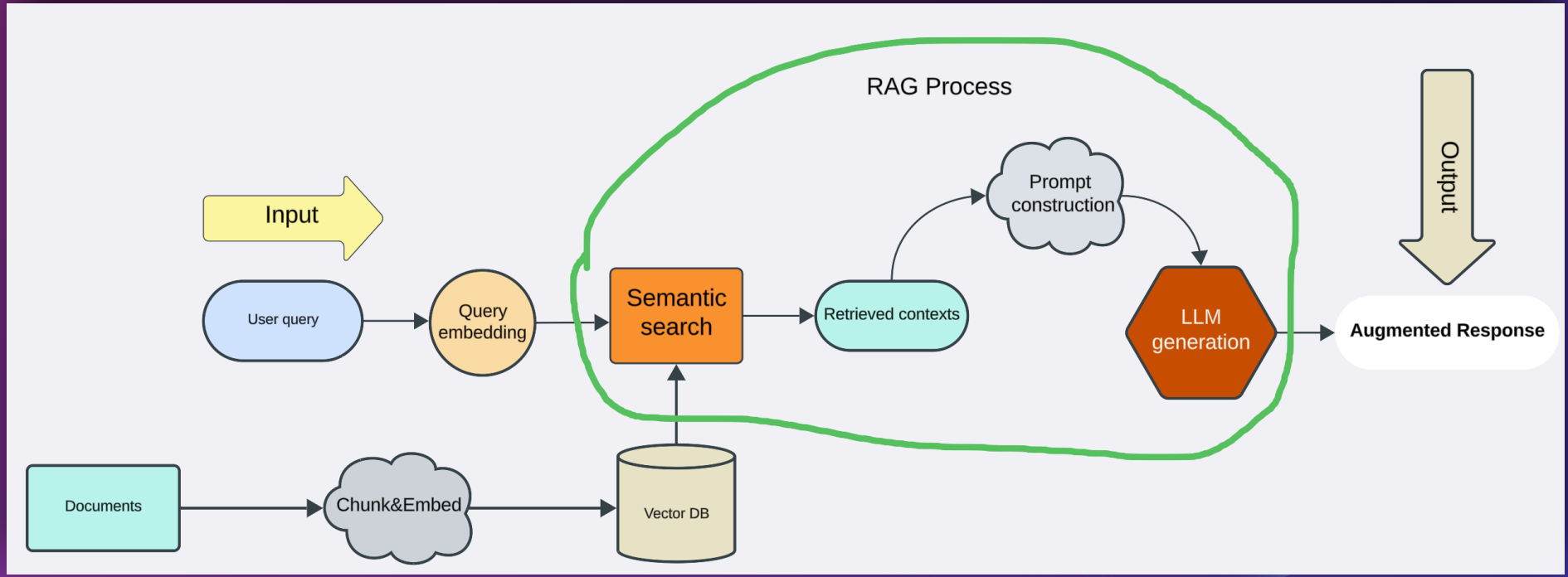
RAG vs. Fine-tuning

RAG uses external data on the fly to help generate responses.

Fine-tuning adapts an existing model to perform better on a specific task by further training it with task-specific data (no additional retrieval component).

RAG & Fine-tuning

1. Pre-train a language model on broad data.
2. Fine-tune the model with a dataset specific to your application.
3. Implement a retrieval component that fetches relevant data from an external source in real-time.
4. The fine-tuned model uses both its learned knowledge and the newly retrieved information to generate responses.


RAG Process
Input
User query
Query embedding
Semantic search
Retrieved contexts
Prompt construction
LLM generation
Augmented Response
Output
Documents
Chunk&Embed
Vector DB

Use cases

OpenAI functions on AWS Lambda: on demand NLP, text generation.

Anthropic Claude API on serverless: content generation, analysis, question-answering.

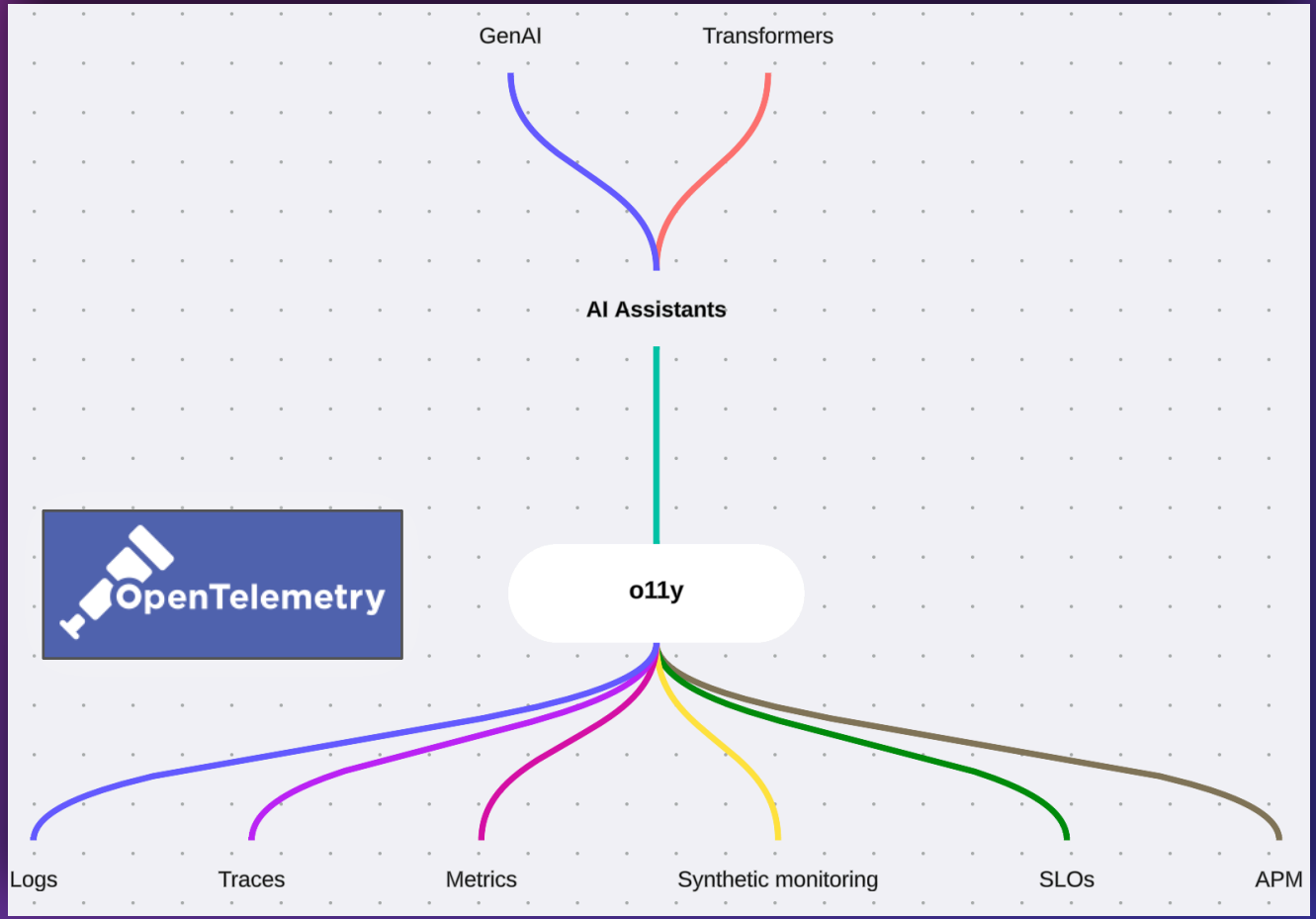
AI powered serverless chatbots: AWS Lex or Azure Bot service.

Vercel AI SDK: serverless platform offers AI SDK.

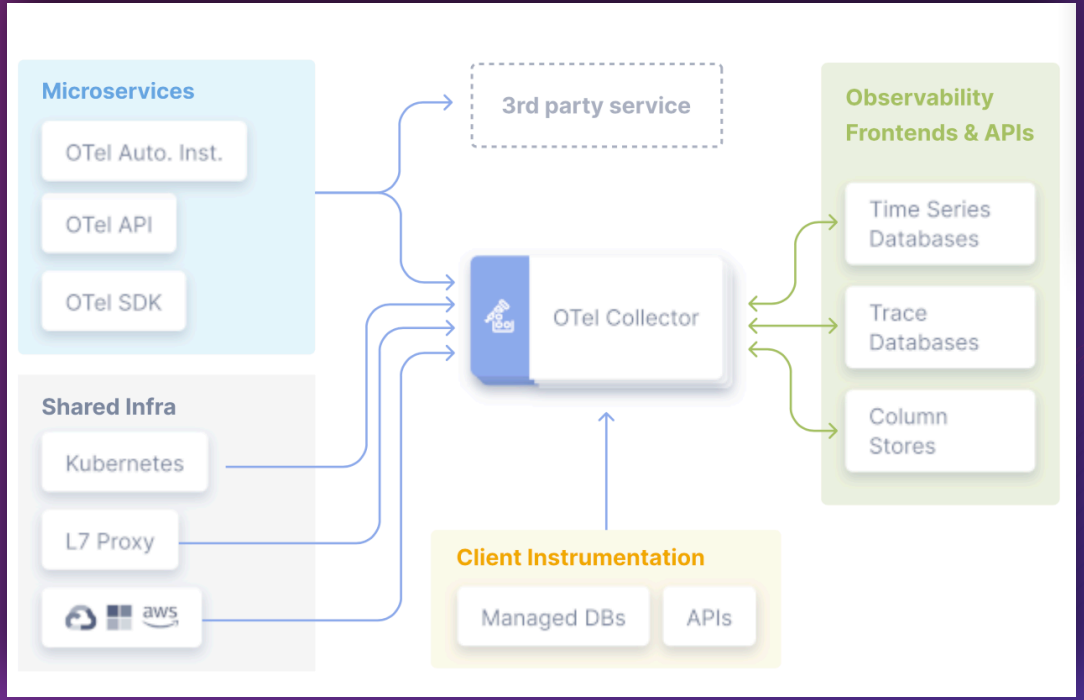
Pinecone or Weaviate offer serverless VD used for retrieval and similarity search. OSS VD: LanceDB.

Is my infrastructure ready?

Observability



OpenTelemetry



Value	Description	Stability
`anthropic`	Anthropic	experimental
`cohere`	Cohere	experimental
`openai`	OpenAI	experimental
`vertex_ai`	Vertex AI	experimental

OTel Community Day

How OpenTelemetry Helps Generative AI

Phillip Carter, Honeycomb

Prompt engineering
Custom scripts
Refine user data
Improve vector engine

Prepare my data

More or less data?

Yes

Less quality results

Pick your LLM

Use better prompts

Pick your vector database

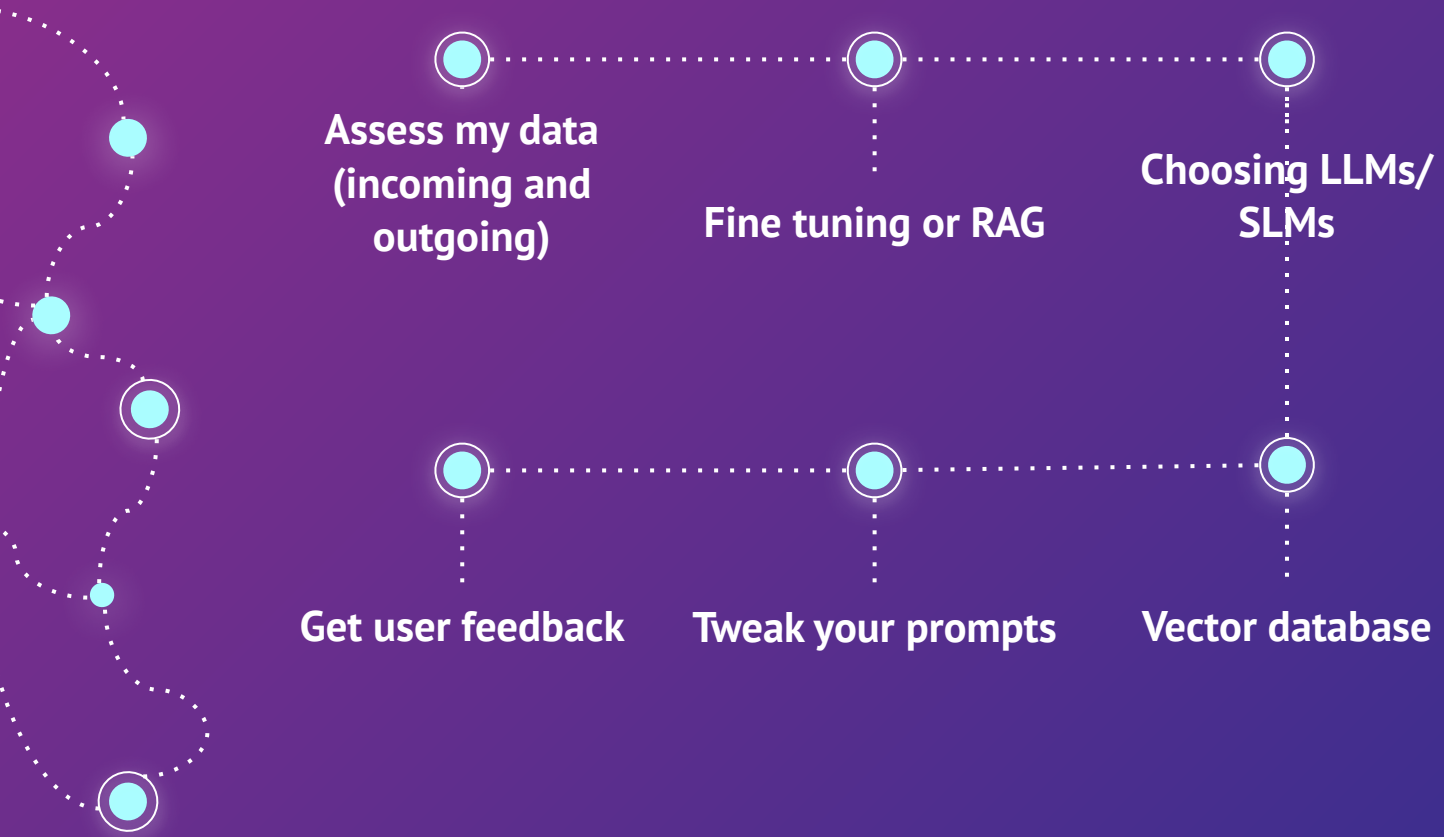
No

Better quality results

Higher speed in providing results

Not suitable for production

TAKE AWAYS


Assess my data (incoming and outgoing)
Fine tuning or RAG
Choosing LLMs/ SLMs
Get user feedback
Tweak your prompts
Vector database

Resources

https://www.youtube.com/watch?v=2IK3DFHRFfw
https://opentelemetry.io/blog/2024/llm-observability/
https://opentelemetry.io/docs/languages/js/serverless/
https://www.youtube.com/watch?v=92oGRCC8ktA
https://www.pinecone.io/learn/vector-database/#Serverless-Vector-Databases
https://foundationcapital.com/goodbye-aiops-welcome-agentsres-the-next-100b-opportunity/
https://neptune.ai/blog/llm-observability
Sebastian Raschka-“Machine Learning Q and AI”, No Starch Press, 2024

THANKS!

ANY QUESTIONS?

https://github.com/didiViking/Conferences_Talks
https://www.linkedin.com/in/diana-todea-b2a79968