GenerativeAI and Observability in the serverless world

EQS CREATING TRUSTED COMPANIES »

www.eqs.com

DIANA TODEA, Senior SRE

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

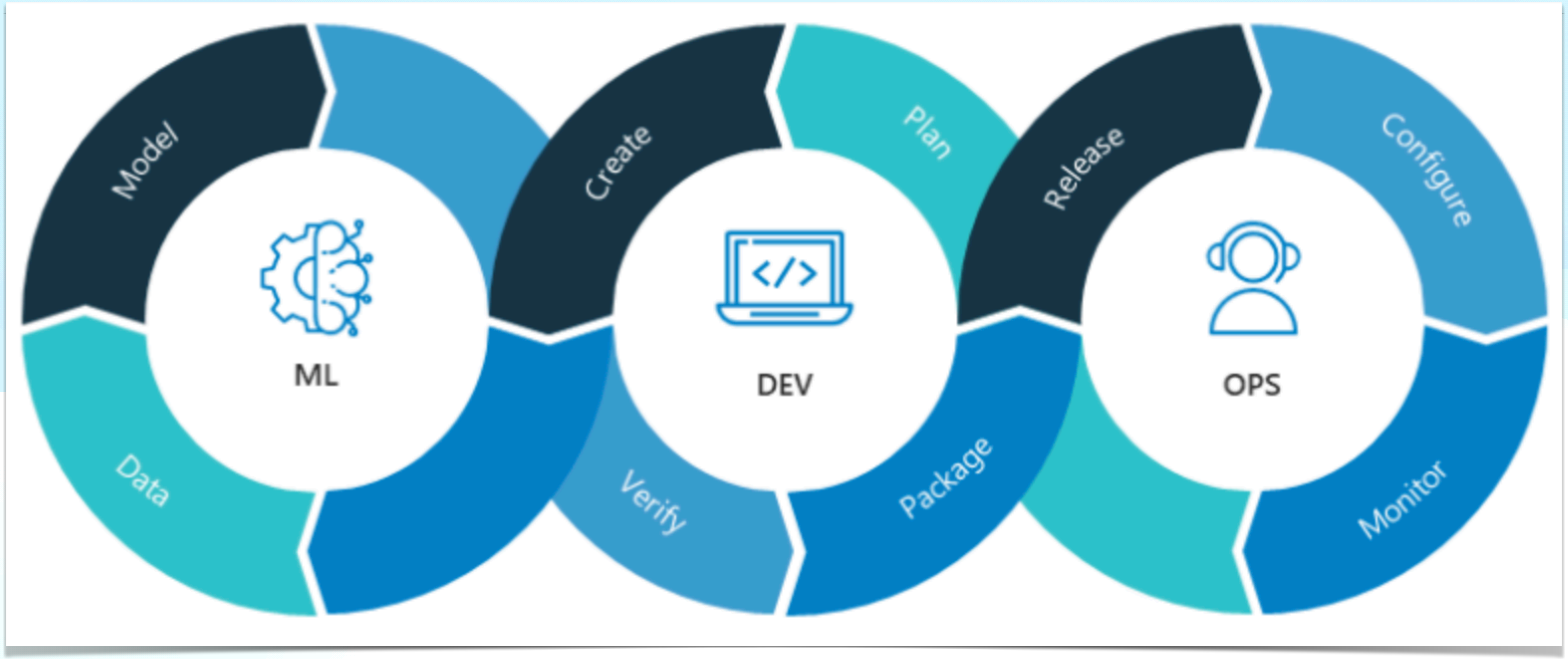
O11y Site Reliability Engineer
Passionate about AI, open source and support women in tech
Love learning foreign languages, traveling and doing sports

DISCLAIMER

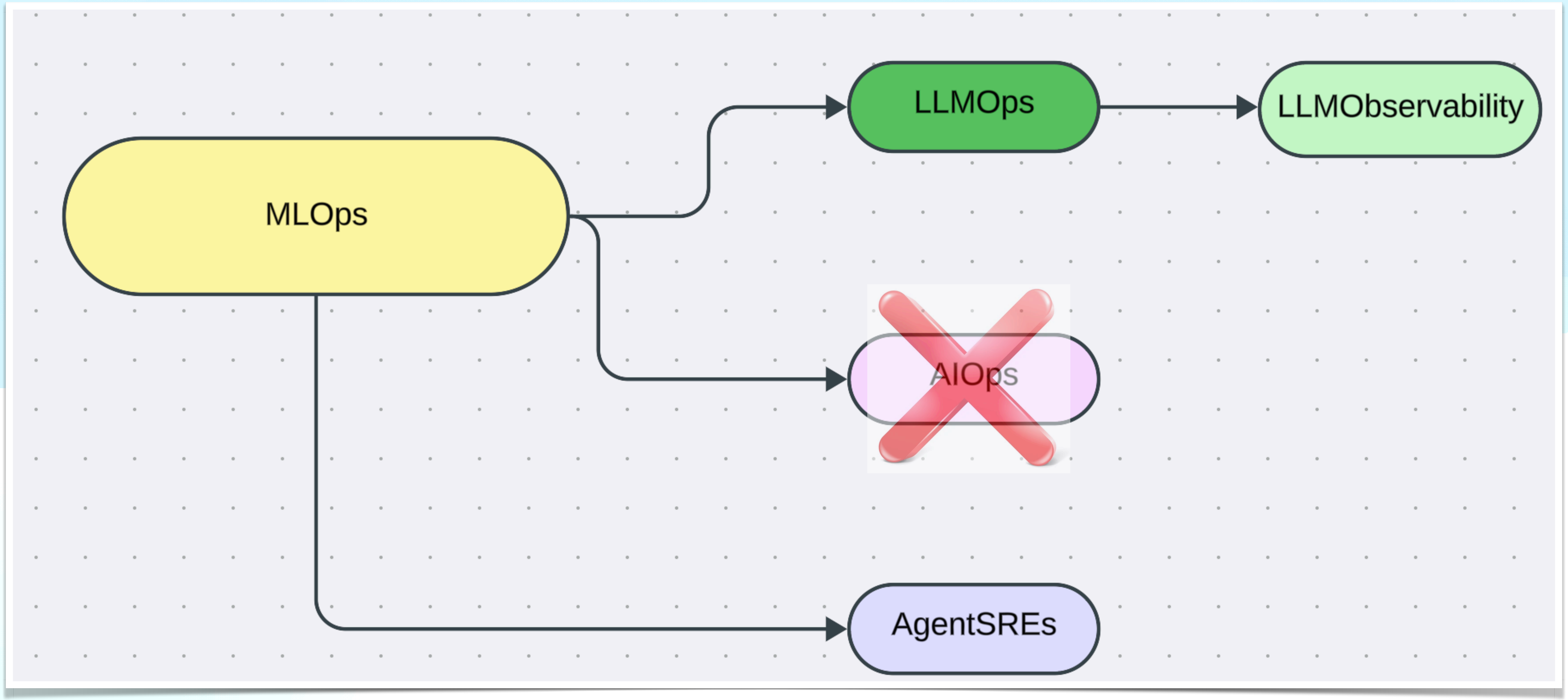
The opinions expressed in this presentation are solely my own and they do not represent my employer’s.

Agenda

- Concepts
- o11y
- AI Assistants
- Lessons learned



Source: Nvidia

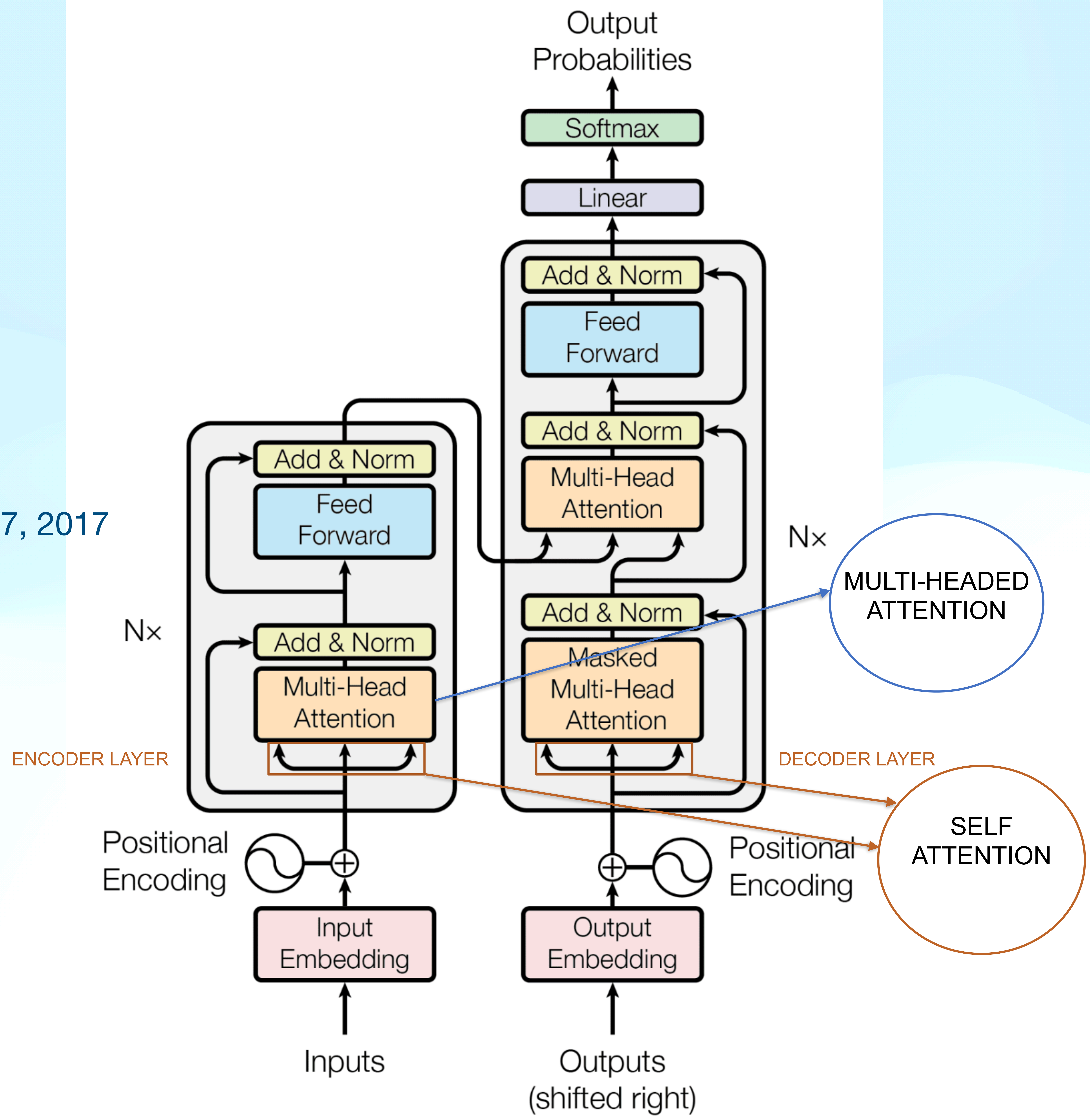

MLOps
LLMOps
LLMObservability
AIOps
AgentSREs

What’s a 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 IS ALL YOU NEED”

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



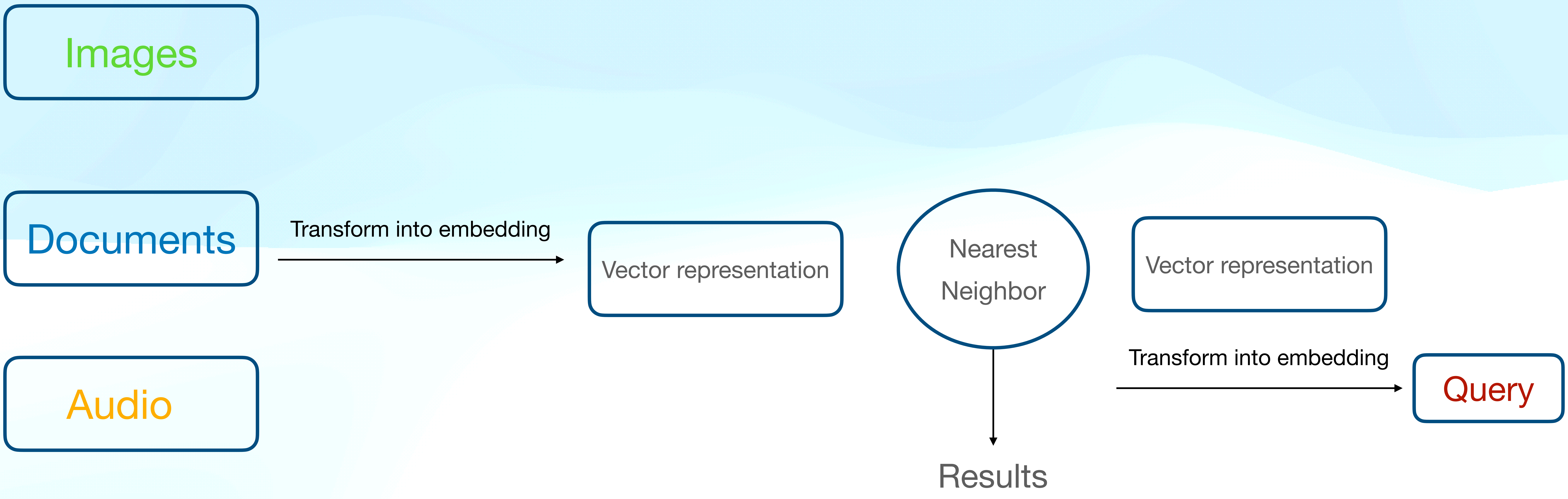
ENCODER LAYER

DECODER LAYER

MULTI-HEADED ATTENTION

SELF ATTENTION

GenAI Architecture



Images

Documents

Audio

Transform into embedding

Vector representation

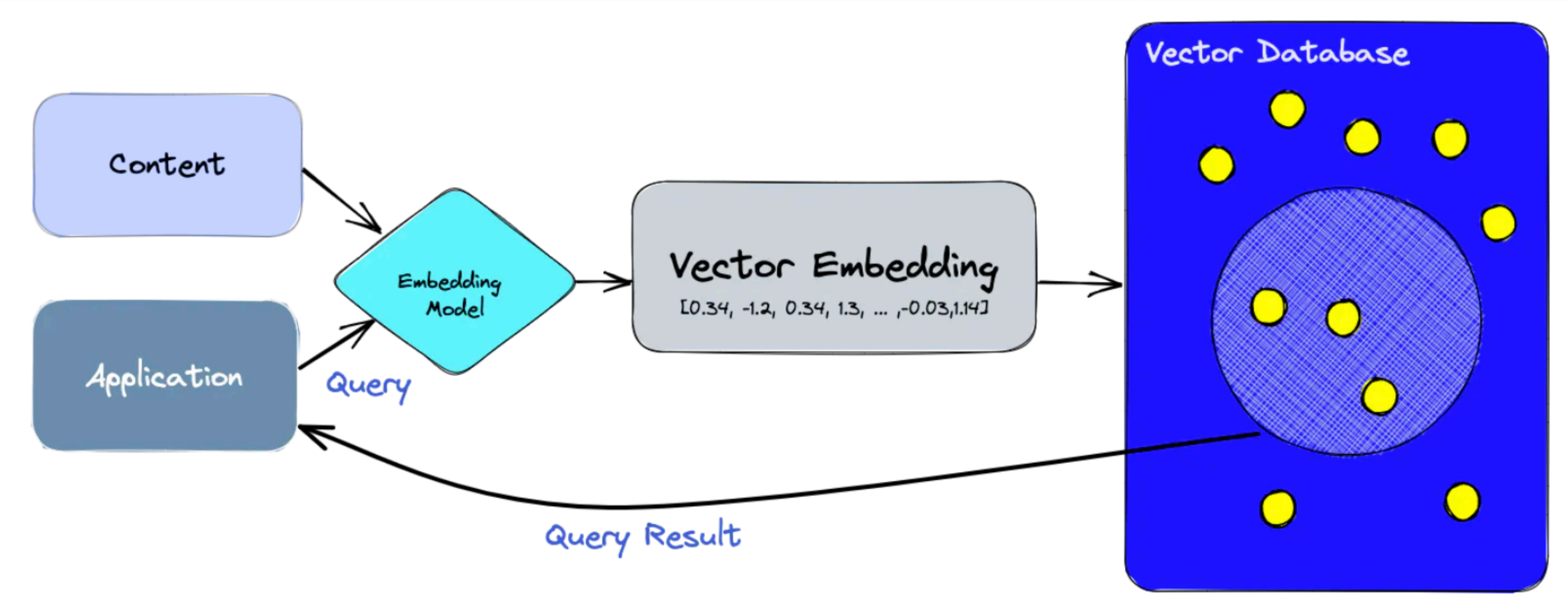
Nearest Neighbor

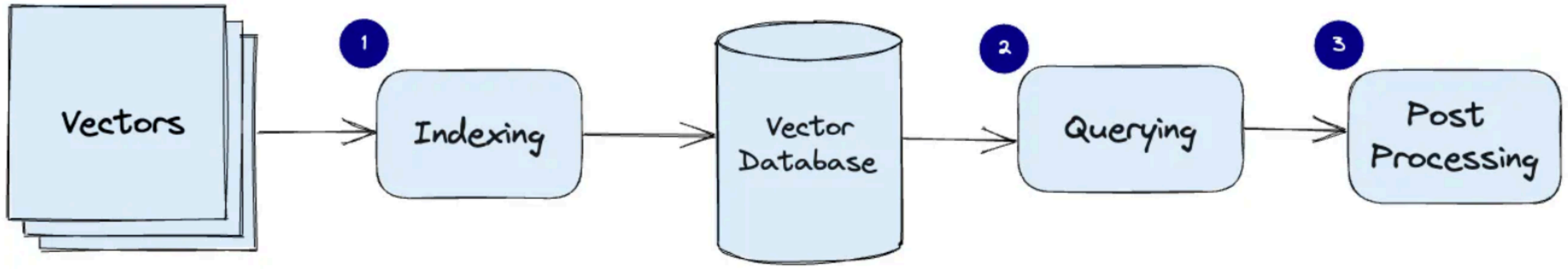
Vector representation

Transform into embedding

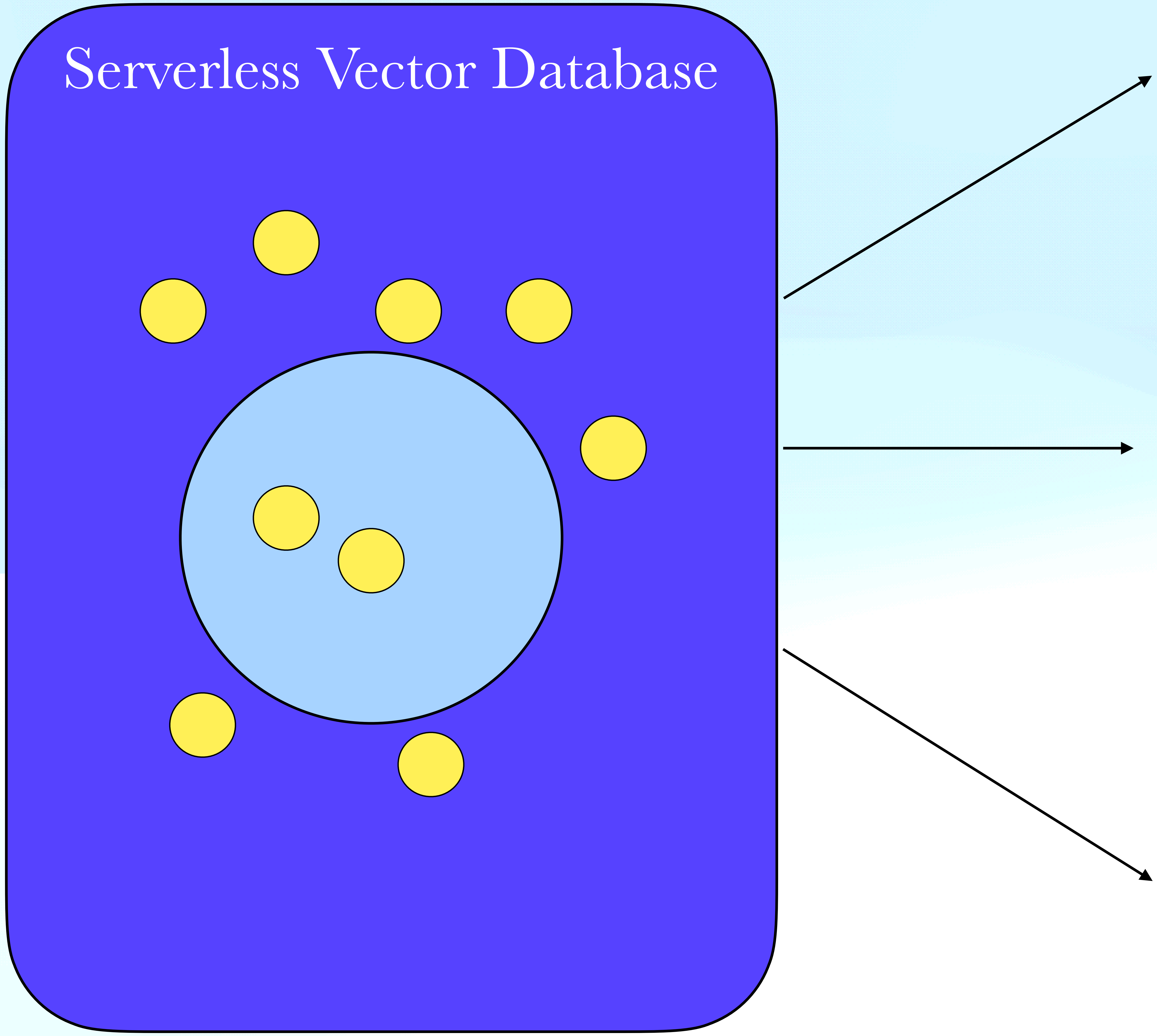
Query

Results





Source: https://www.pinecone.io



Solves: **separation of storage from compute**

HOW: highly sophisticated geometric partitioning algorithms

Solves: **freshness**

HOW: Adds a freshness layer/temporary “cache”

Solves: **multi tenancy**

HOW: identifies users with similar usage and colocates them while keeping full separation between them, which is done based on user usage metrics and automatic allocation of hot/cold infrastructure based on usage.

Use cases

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

- Azure OpenAI service with Azure Functions: serverless AI-powered applications for content generation, summarization and translation.

- Vercel AI SDK: serverless platform offers AI SDK which simplifies the AI models integration into serverless functions, helps building serverless applications.

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

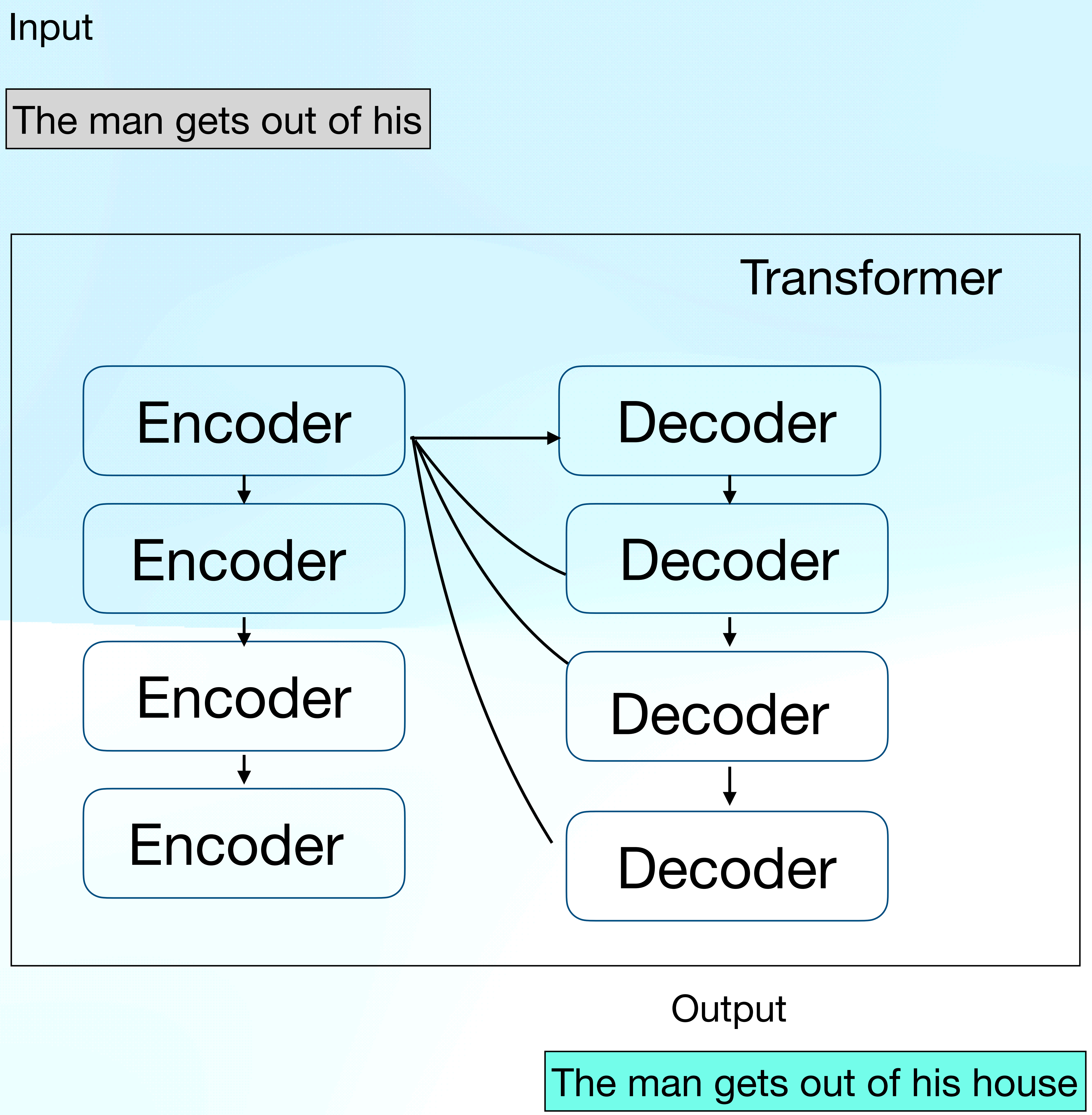
- Cloudflare Workers AI: text generation, image classification.

- HF on Google Cloud functions: serverless NLP, sentiment analysis and other ML tasks.

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

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

- Serverless AI inference platforms: AWS SageMaker serverless inference allows deploying ML models on serverless.



Input

The man gets out of his

Transformer

Encoder

Encoder

Encoder

Encoder

Decoder

Decoder

Decoder

Decoder

Output

The man gets out of his house

Retrieval augmented generation (RAG) is a technique that supplements text generation with information from private or proprietary data sources.

1. RAG starts with an input query.
2. The retrieval model grabs the relevant information from databases or external sources.
3. The retrieved information is converted into vectors in a high-dimensional space, which are then stored into a vector database.
4. The retrieval model ranks the retrieved information based on its relevance to the input query. The documents with the highest scores get selected for further processing.

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.

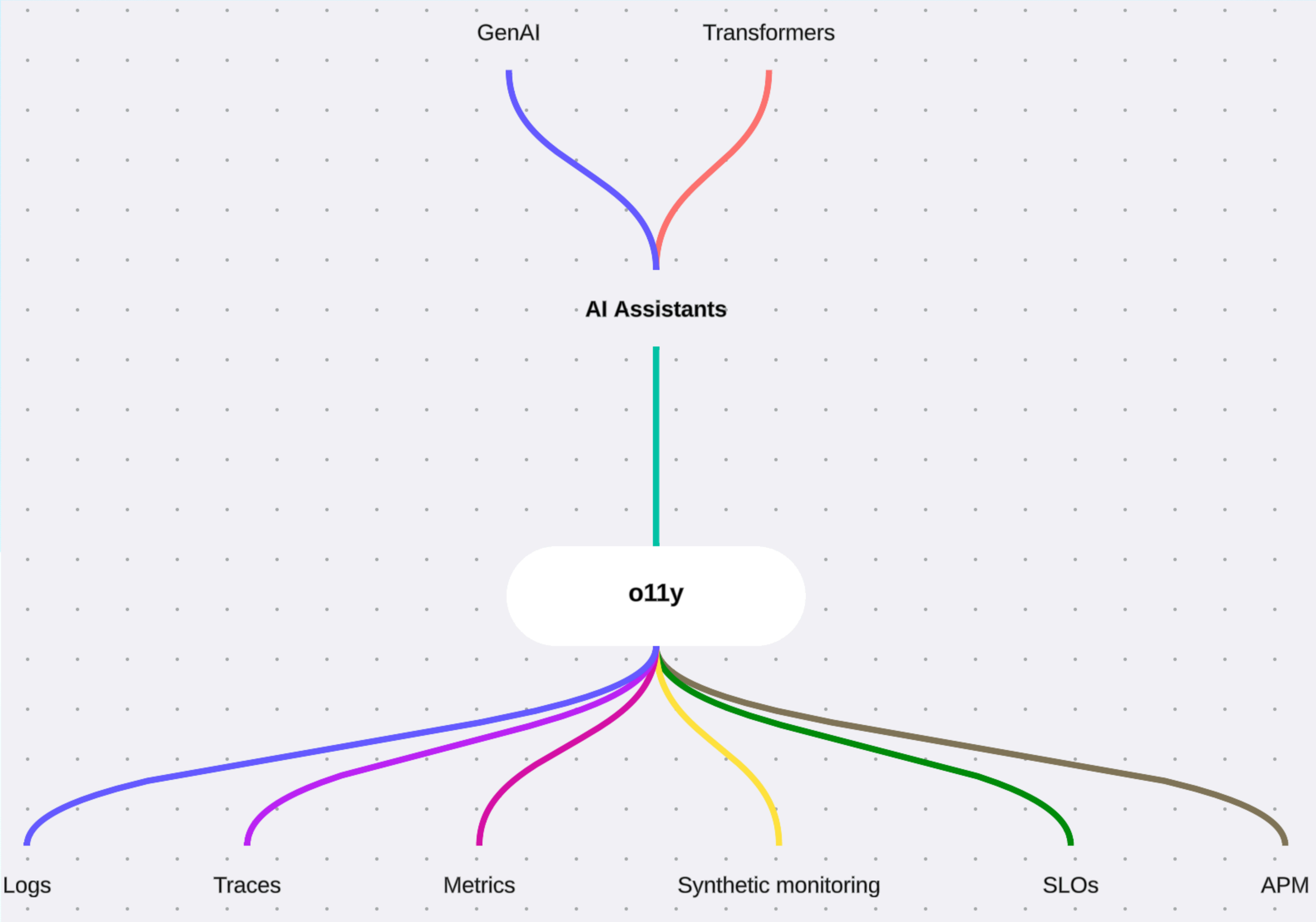
RAG & Fine-tuning

Pre-train a language model on broad data.

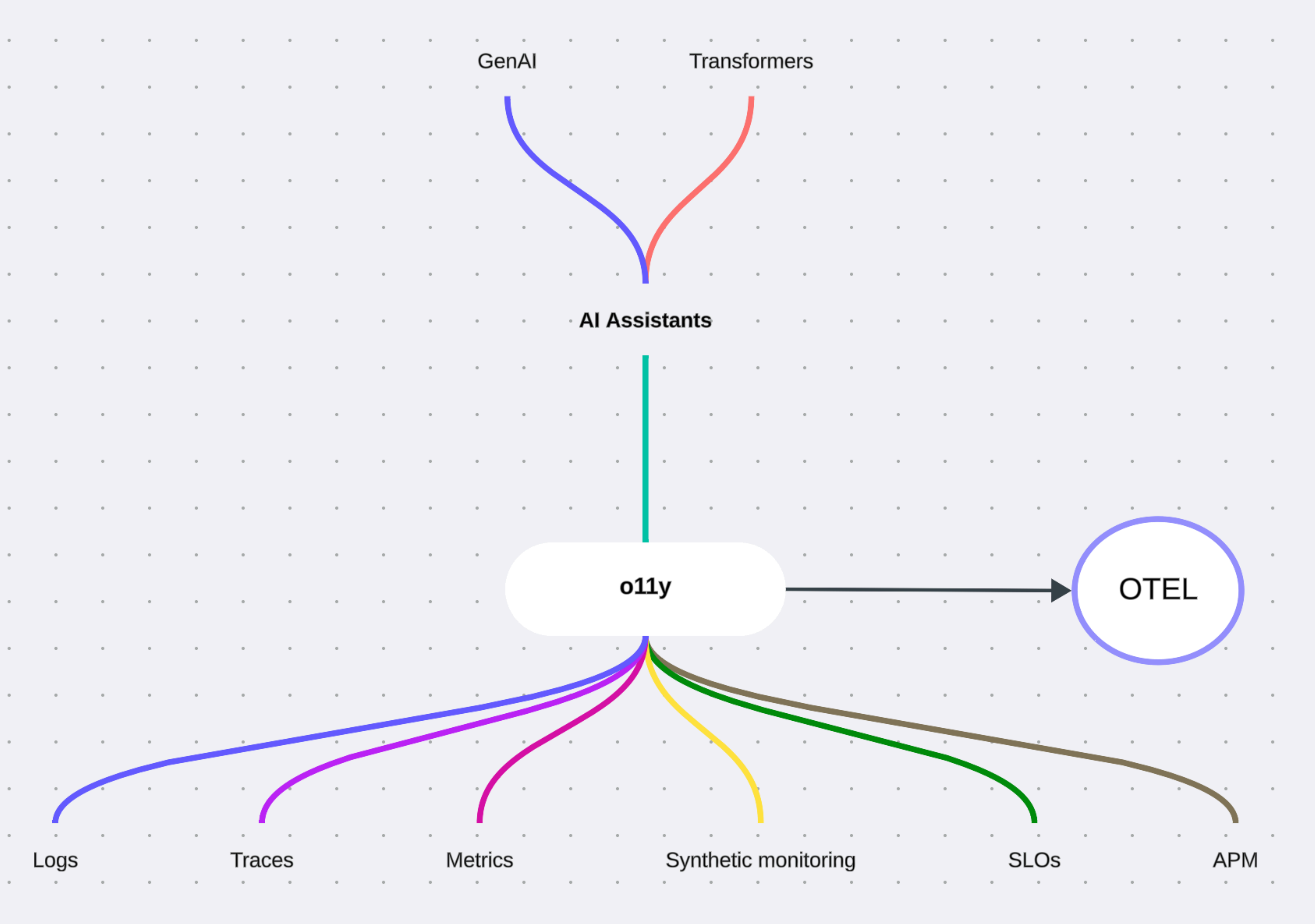
Fine-tune the model with a dataset specific to your application.

Implement a retrieval component that fetches relevant data from an external source in real-time.

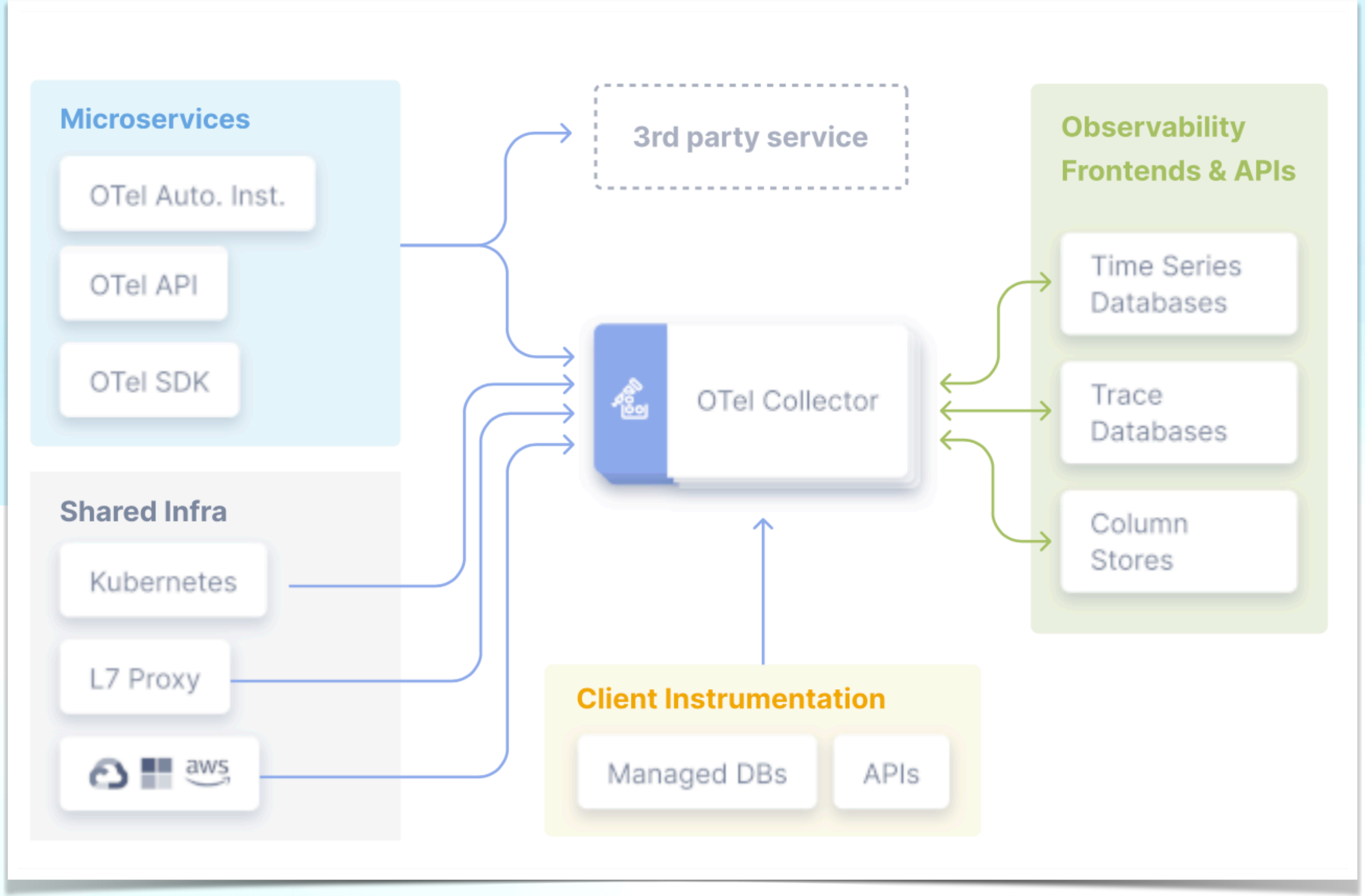
The fine-tuned model uses both its learned knowledge and the newly retrieved information to generate responses.


GenAI
Transformers
AI Assistants
o11y
Logs
Traces
Metrics
Synthetic monitoring
SLOs
APM

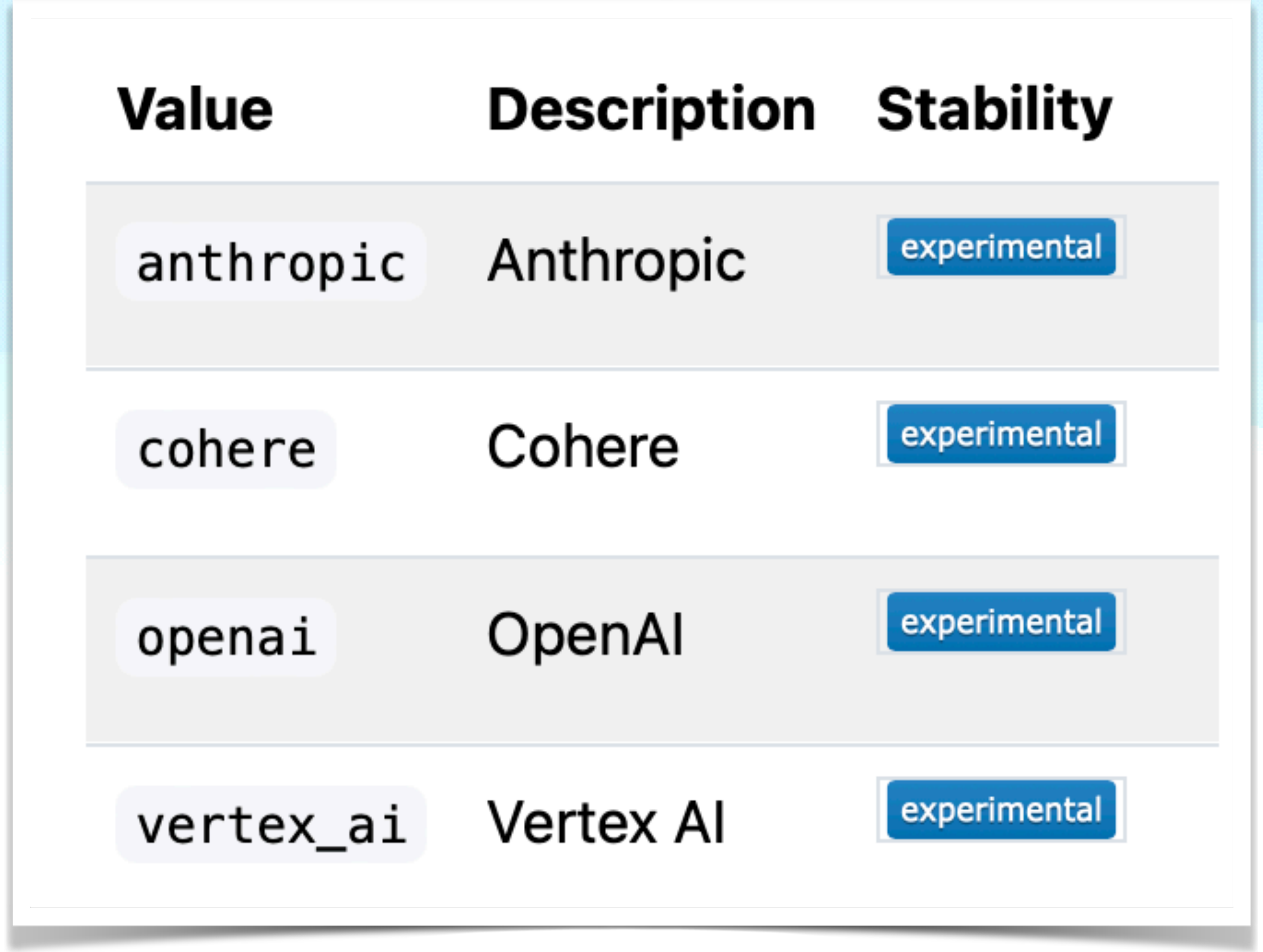
● O11y


GenAI
Transformers
AI Assistants
o11y
OTEL
Logs
Traces
Metrics
Synthetic monitoring
SLOs
APM

OpenTelemetry



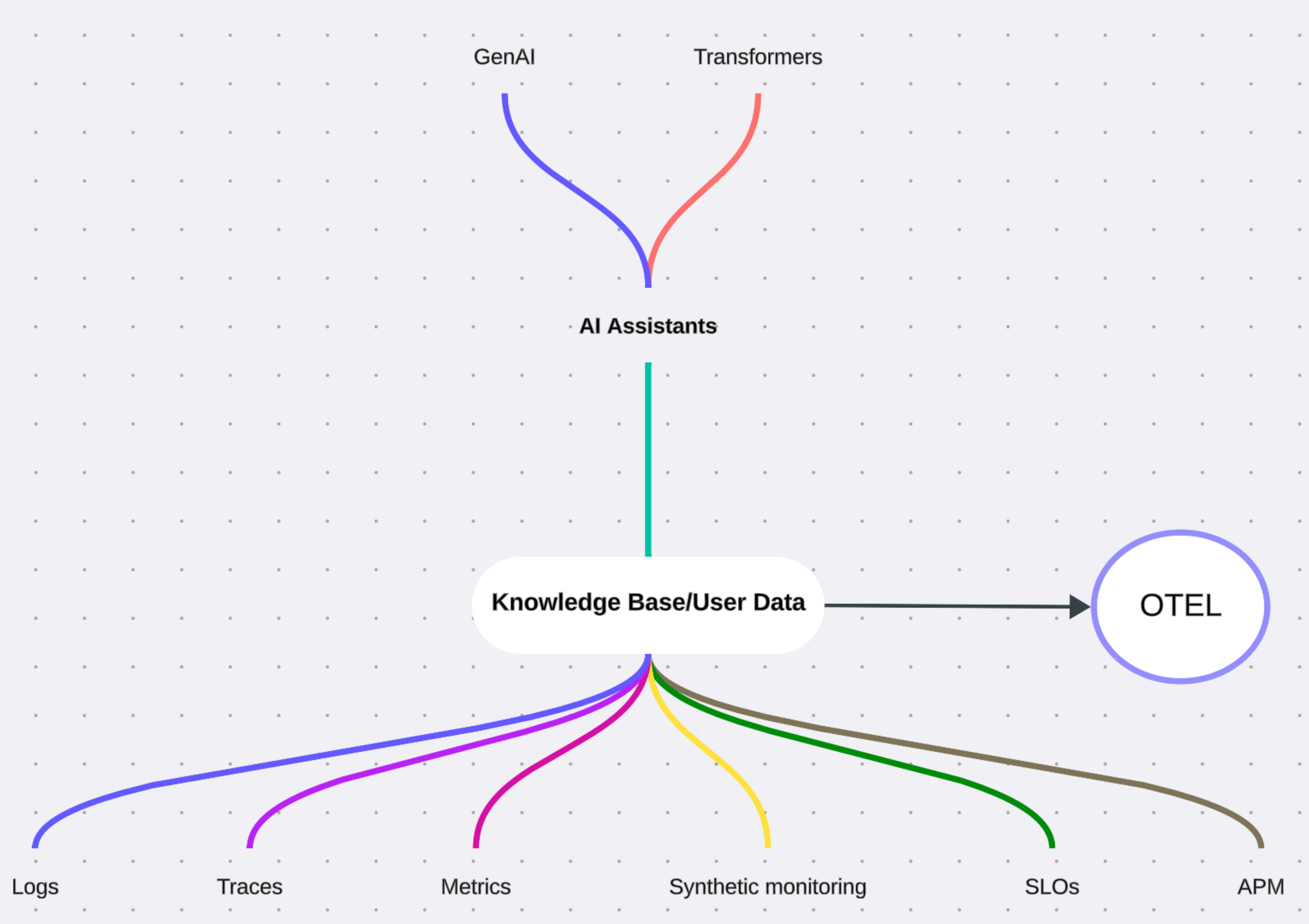
GenAI

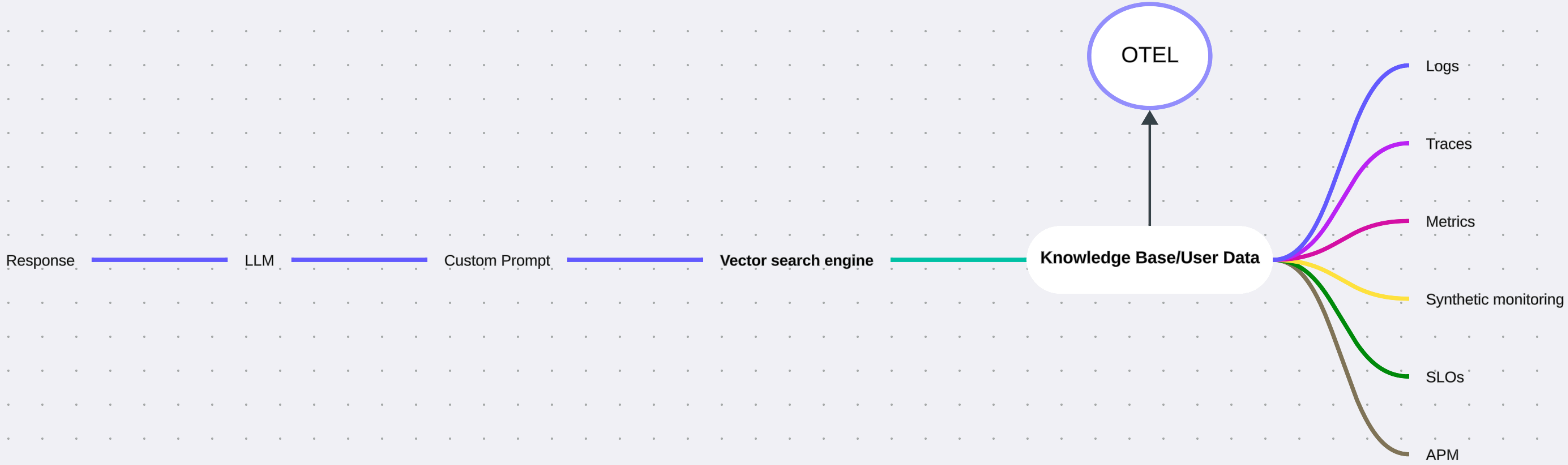


Value	Description	Stability
`anthropic`	Anthropic	experimental
`cohere`	Cohere	experimental
`openai`	OpenAI	experimental
`vertex_ai`	Vertex AI	experimental

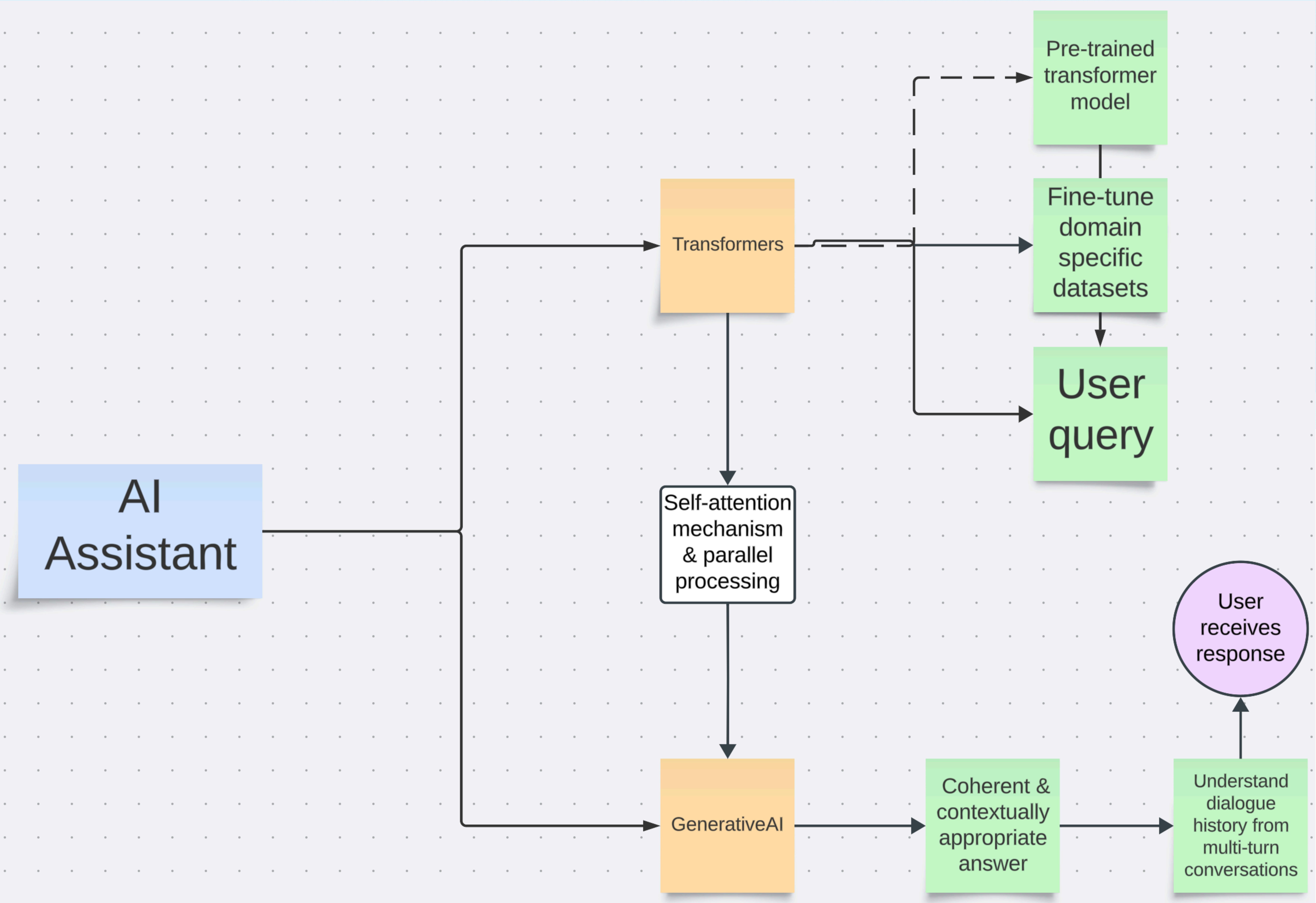
https://opentelemetry.io/docs/specs/semconv/gen-ai/

https://opentelemetry.io/docs/specs/semconv/gen-ai/gen-ai-metrics/


GenAI
Transformers
AI Assistants
Knowledge Base/User Data
OTEL
Logs
Traces
Metrics
Synthetic monitoring
SLOs
APM



AI Assistants



AI Assistant

Transformers

Pre-trained transformer model

Fine-tune domain specific datasets

User query

Self-attention mechanism & parallel processing

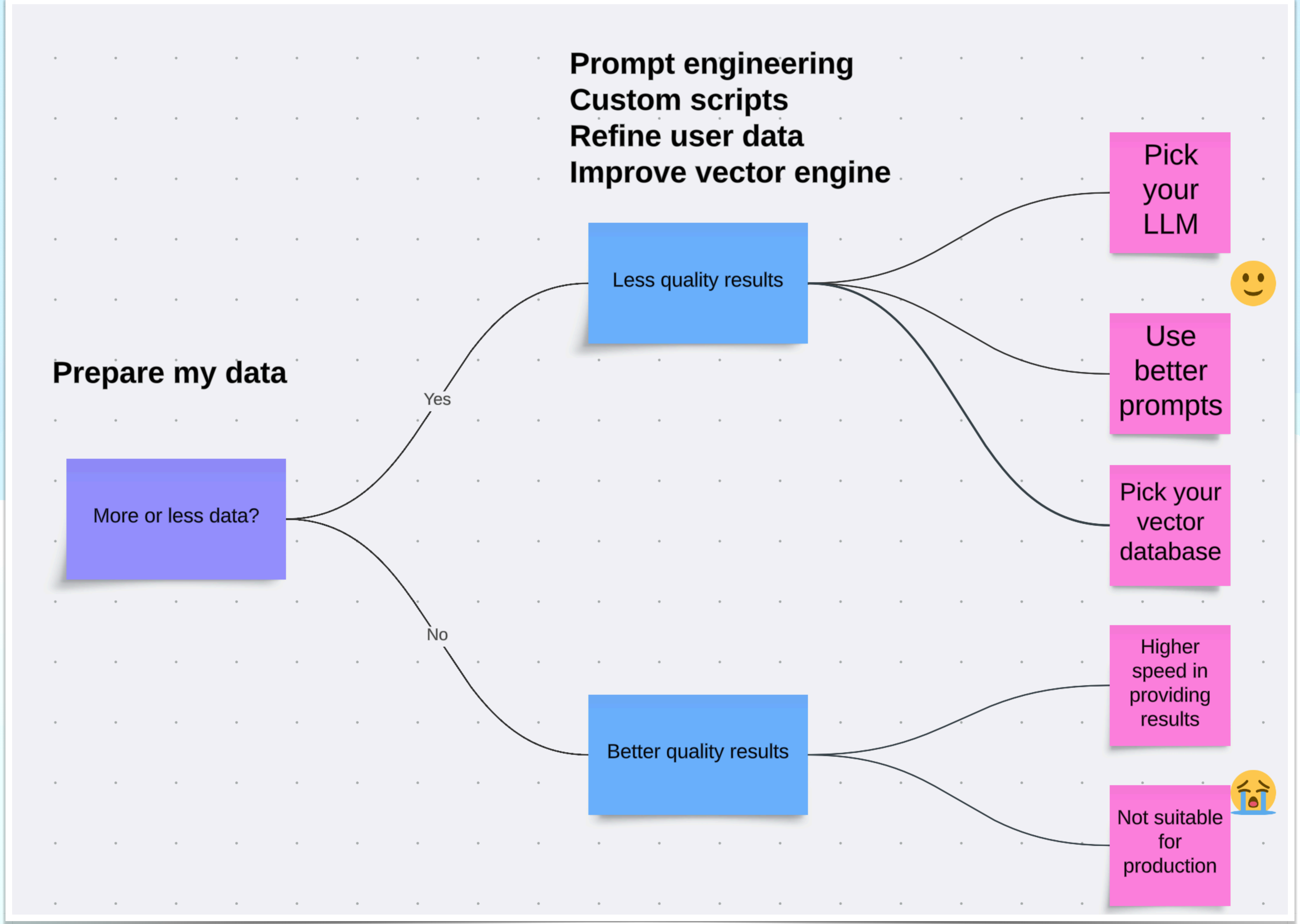
GenerativeAI

Coherent & contextually appropriate answer

Understand dialogue history from multi-turn conversations

User receives response

Is your infrastructure ready?



Prepare my data

Prompt engineering
Custom scripts
Refine user data
Improve vector engine

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

- Assess your data (incoming and outgoing)
- Fine tuning or RAG
- Choosing LLMs/SLMs
- Which vector database
- Upskill, train, review infrastructure
- Tweak your prompts
- Production ready?
- Do your users like it? Get feedback!

- 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

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

EQS CREATING TRUSTED COMPANIES »

www.eqs.com

WE ARE HIRING!