



GENERATIVE AI: APPLICATIONS IN THE SERVERLESS WORLD

DIANA TODEA - SITE RELIABILITY ENGINEER

DEVBCN JUNE 14th 2024

<https://www.linkedin.com/in/diana-todea-b2a79968>

[https://github.com/didiViking/
Conferences_Talks](https://github.com/didiViking/Conferences_Talks)

Generative AI in a Nutshell

Prompt Engineering/Design

- Bad prompt**: You Give me an agenda for a workshop.
- Good prompt**: You Give me an agenda for a workshop. **task context**: I'm meeting a leadership team at an aerospace consulting firm. The goal of the workshop is figure out how they can use AI. They are new to this. We have 8 people for 4 hours.
- Good prompt**: You Give me an agenda for a workshop. **questions**: Feel free to ask me any clarifying questions first.

Autonomous Agents with tools

Computers have gotten smarter

Execute Instructions

Side effect: Better communication skills overall

Better prompt engineering skills = Better results

Study, Practice, Learn

The role of Humans

Is human role X still needed? (doctor, developer, lawyer, ceo, teacher, etc...) Decide what to ask and how Evaluate results (beware of hallucinations!) Compensate for AI weaknesses Provide context Legal compliance Data security etc,etc...

Using vs Building AI-powered products

Product: chatGPT, copilot, etc... **Model**: GPT4, Llama, etc...

AI: Artificial Intelligence
Machine Learning
Computer vision, etc...

Generative AI: LLM Large Language Model Chat GPT

Terminology

GPT: Generative Transformer Pretrained

Training

- ① Unsupervised Generative Pretraining (lots + lots of text...)
- ② Reinforcement Learning with Human Feedback (RLHF)

Emergent Capabilities

Small model: "John fell" → "down"

Large model: etc etc teach coach legal/medical advice Roleplay Poetry etc etc etc etc

The Age of AI

Human Intelligence vs Artificial Intelligence

Fast Revolution vs Slow Revolutions

Timeline: ≈ 80 years

Mindset

Denial: Nah AI can't do my job

Positive: I will be insanely productive!

Panic: AI is going to steal my job! We will go bankrupt!

AI might not take your job, but people/companies using AI will

Autonomous Agents with tools

Here Be Dragons

How to survive and thrive in the Age of AI

Einstein in your basement

Biggest limitation is YOU

Imagination: what can I do? How do I do it?

Neural Network

How it works

Dogs are animals → animals

Dogs are animals → that

Dogs are animals that are known for... (bla bla)...

Models, models everywhere

Speed: Easy to use vs Specialized

Capability: Specialized vs Integrated

Cost: Integrated vs Open Source

Downloadable: etc etc

Open Source: I'm back!

Model Types

Text → Text: "opposite of yellow" → "Purple" (or code, JSON, HTML, etc)

Text → Image: "Einstein sitting in the basement" → Einstein sitting in the basement

Image → Image: "Fusion of a human and a cat, seated in an armchair" → Fusion of a human and a cat, seated in an armchair

Text → Image: "Ugly cat" → Ugly cat

Text → Audio: "People talking in a busy restaurant" → People talking in a busy restaurant

Text → Video: "Darth Vader surfing" → Darth Vader surfing

Multimodal Models

Whatever → Whatever

Where could I hide something valuable in this room?

ChatGPT: If you're looking to hide something valuable in this room, you have several options. Here are a few suggestions:

- Behind the map: You could tape something flat, like documents or cash, behind the large map on the wall.
- Inside the guitar: If the guitar has a case, you could hide something inside it. If not, small items could potentially be placed through the sound hole, although this is not advisable if it could damage the instrument.
- Under the couch cushions: A common hiding spot, but effective for small and flat items.
- Behind or under the wood stove: If there is any removable panel or space, although this is risky due to the heat when in use.
- In the lamp: If the bottom of the lamp opens or if there is space within the lampshade, it could conceal small items.

Taking AI for a walk

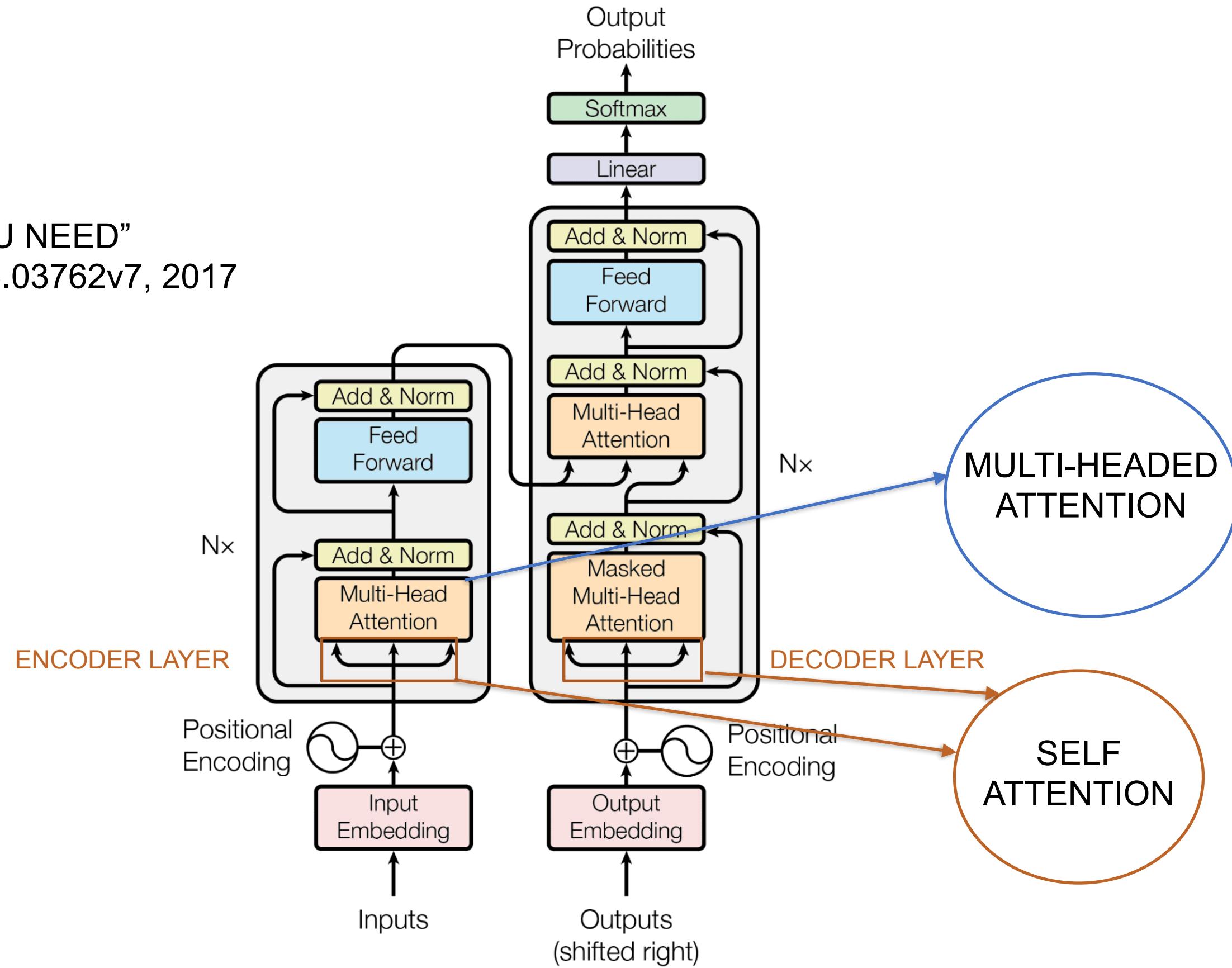
SOURCE: <https://blog.crisp.se/wp-content/uploads/2024/01/generative-AI-in-a-nutshell.png>

WHAT IS A TRANSFORMER?

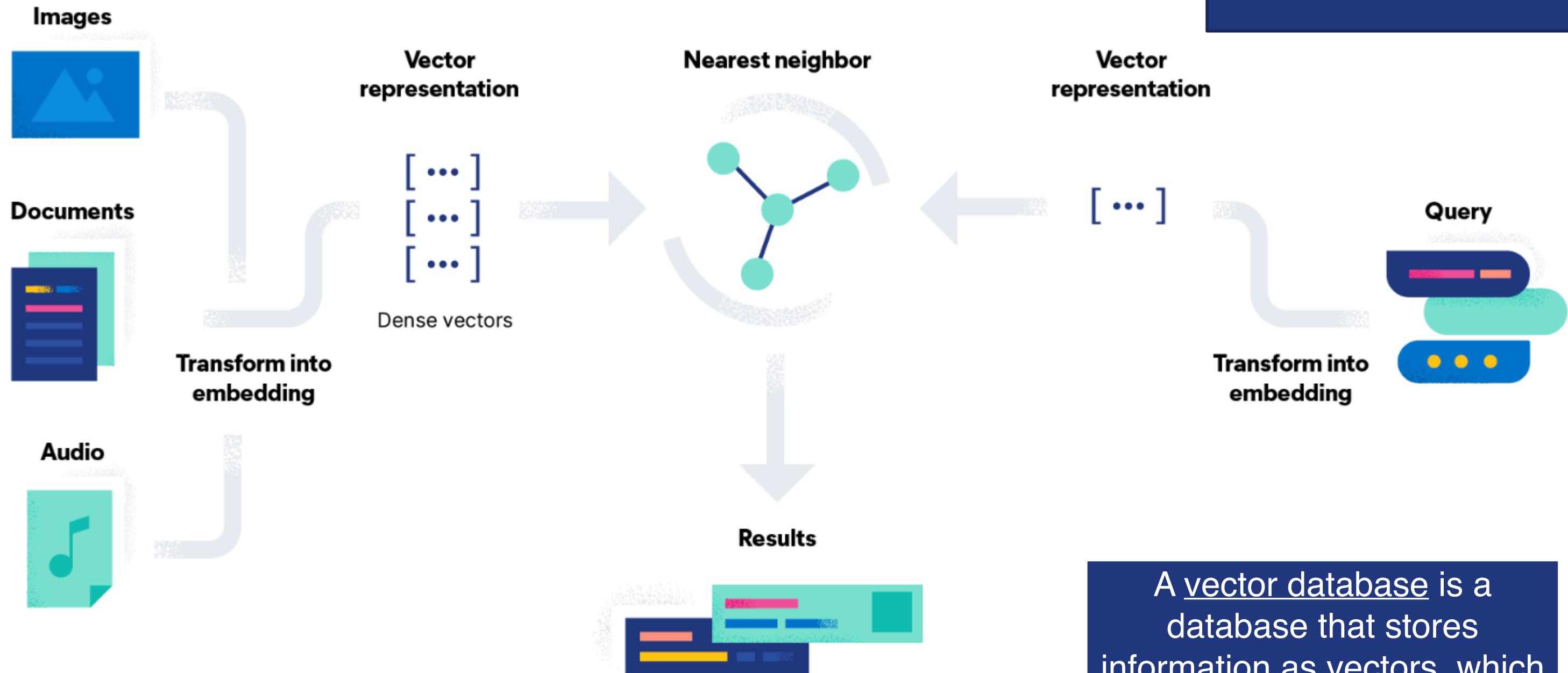
In machine learning, a transformer is a neural network that learns context and meaning by tracking relationships in sequential data like the words in the sentence.

TRANSFORMER ARCHITECTURE

“ATTENTION IS ALL YOU NEED”
<https://arxiv.org/abs/1706.03762v7>, 2017



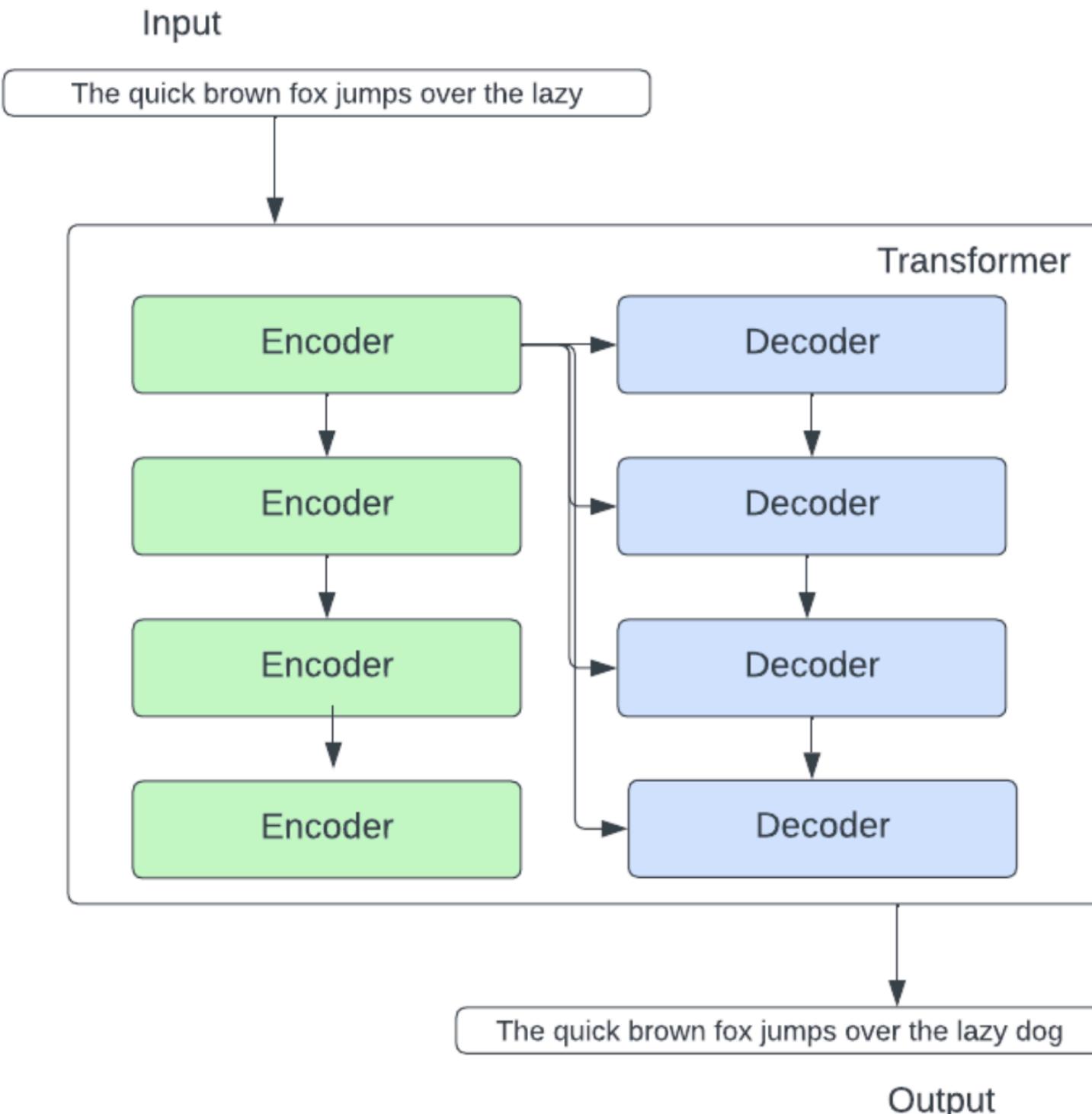
GEN AI ARCHITECTURE



Vector databases use algorithms to index and query vector embeddings. The algorithms enable approximate nearest neighbor (ANN) search through hashing, quantization, or graph-based search. To retrieve information, an ANN search finds a query's nearest vector neighbor.

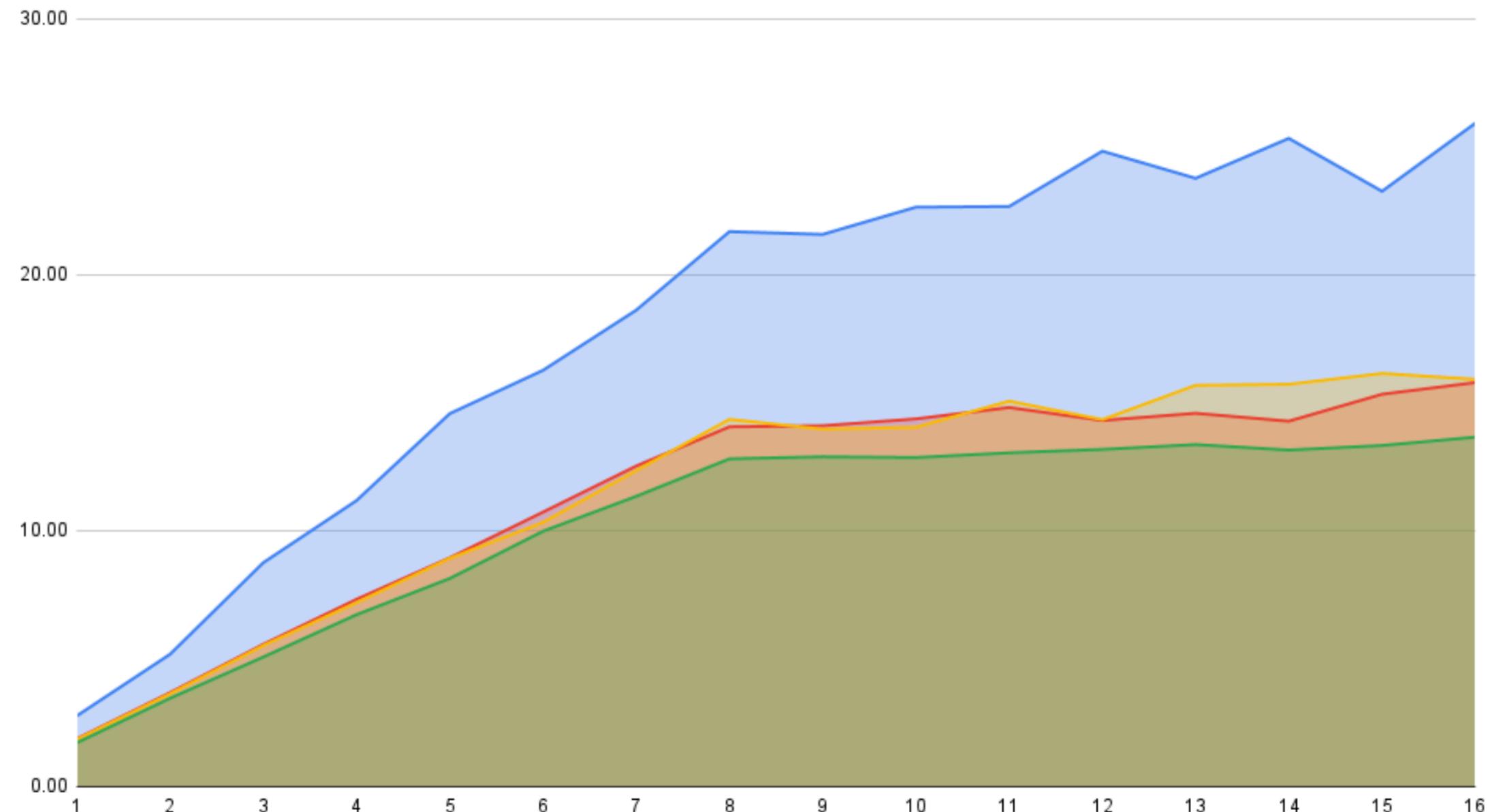
A vector database is a database that stores information as vectors, which are numerical representations of data objects, also known as vector embeddings.

RETRIEVAL AUGMENTED GENERATION (RAG)



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 + ELSER

Add a trained model

[Click to Download](#) [Manual Download](#)

ELSER (Elastic Learned Sparse EncodeR)

ELSER is Elastic's NLP model for English semantic search, utilizing sparse vectors. It prioritizes contextual meaning over literal term matching, optimized specifically for English document queries on the Elastic platform.

[View documentation](#)

Choose a model



Cross platform
.elser_model_2

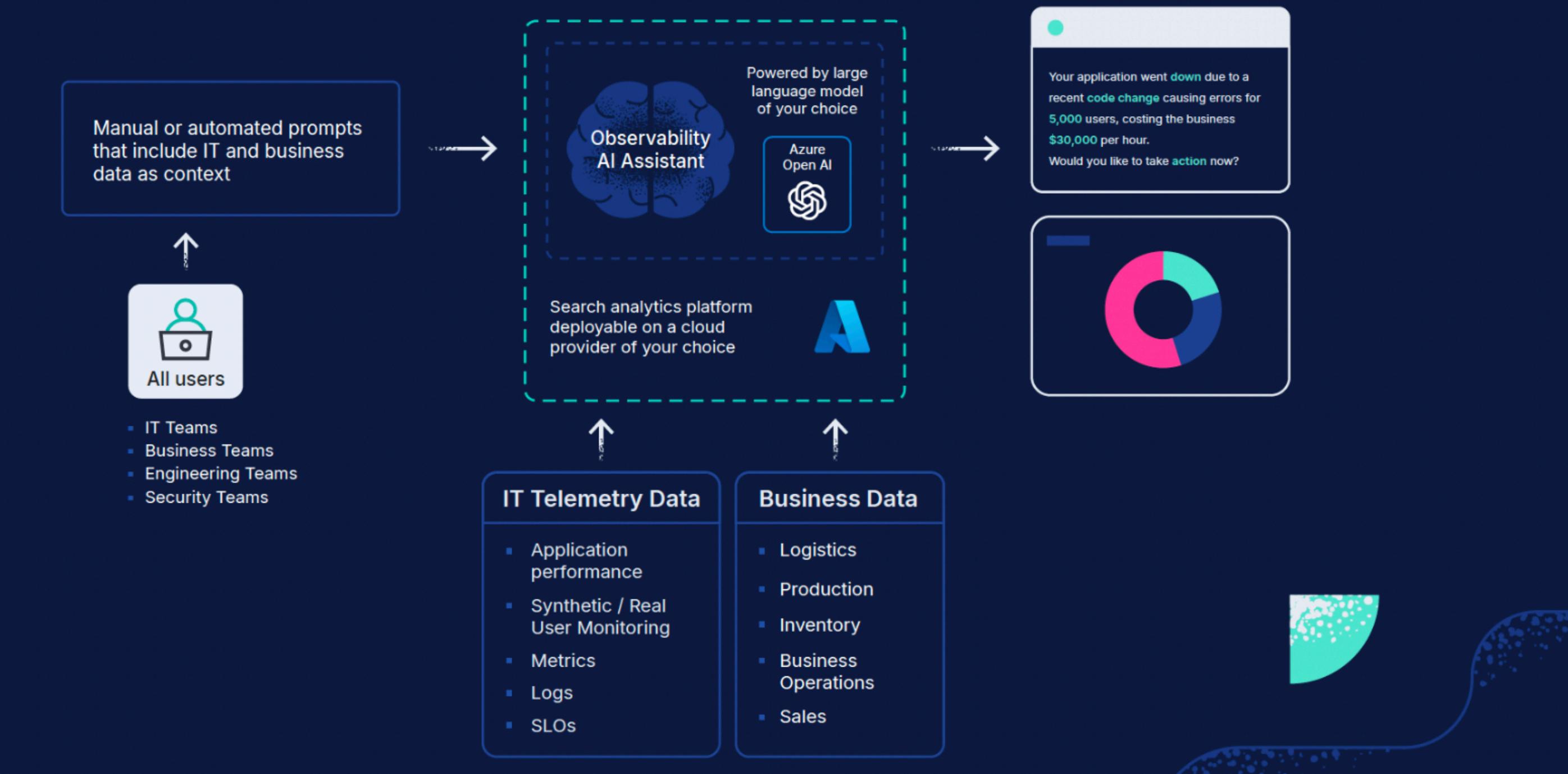


Intel and Linux optimized
.elser_model_2_linux-x86_64

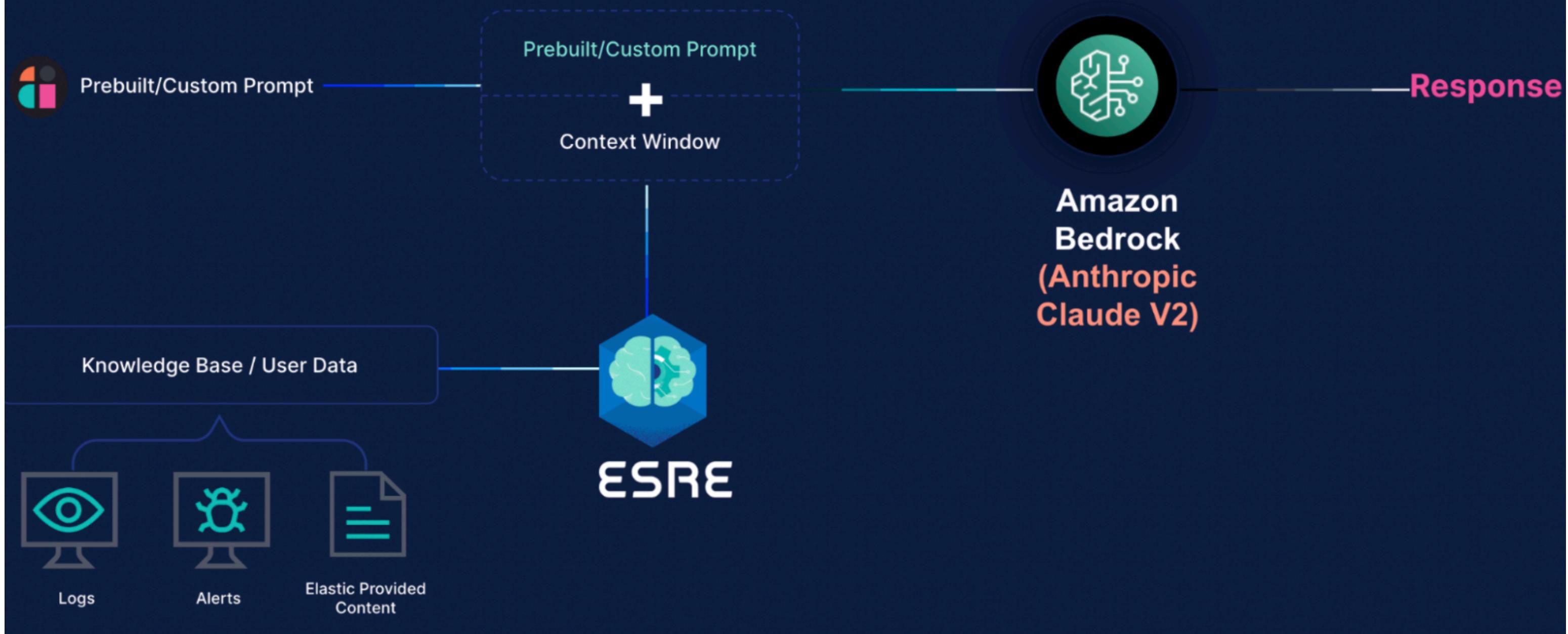
Recommended



Blueprint for Business Observability



Elastic AI Assistant - Behind the Scenes



elastic

Find apps, content, and more.

Add integrations AI Assistant

SLOs

Search your SLOs... Sort by slo status ▾

cartservice-otel APM availability cartservice-... 7 days Target 99.9% 99.519%

Checkout SLO APM availability checkoutSe... 7 days Target 95% 0%

cartservice-otel-latency APM latency cartservice-... 7 days Target 99% 99.498%

Rows per page: 25 < 1 >

Tell us what you think! Stop refreshing Create new SLO

Compact view

Observability Overview Alerts SLOs Cases

Logs Explorer BETA Stream Anomalies Categories

Infrastructure Inventory Metrics Explorer Hosts BETA

APM Services Traces Dependencies

Synthetics Monitors TLS Certificates

Uptime Uptime Monitors TLS Certificates

User Experience Dashboard

Universal Profiling

The screenshot shows the Elastic Observability SLOs dashboard. On the left, there's a sidebar with various monitoring categories like Logs, Infrastructure, APM, and User Experience. The main area displays three SLO cards. The first card, 'cartservice-otel', has a red background and shows a 99.519% success rate against a 99.9% target. It includes APM availability and cartservice-related metrics. The second card, 'Checkout SLO', also has a red background but shows 0% success against a 95% target, with APM availability and checkout-related metrics. The third card, 'cartservice-otel-latency', has a green background and shows a 99.498% success rate against a 99% target, with APM latency and cartservice-related metrics. Each card includes a 'Target' percentage, the current success percentage in large bold digits, and a status icon indicating performance relative to the target.



AI ASSISTANTS ON SERVERLESS

1. FUNCTION AS A SERVICE (FaaS)
2. SCALABILITY
3. COST-EFFECTIVENESS
4. INTEGRATION WITH OTHER SERVICES
5. EVENT-DRIVEN ARCHITECTURE

Where do you expect that the MOST time and resources will be spent when building Generative AI use cases for your applications?

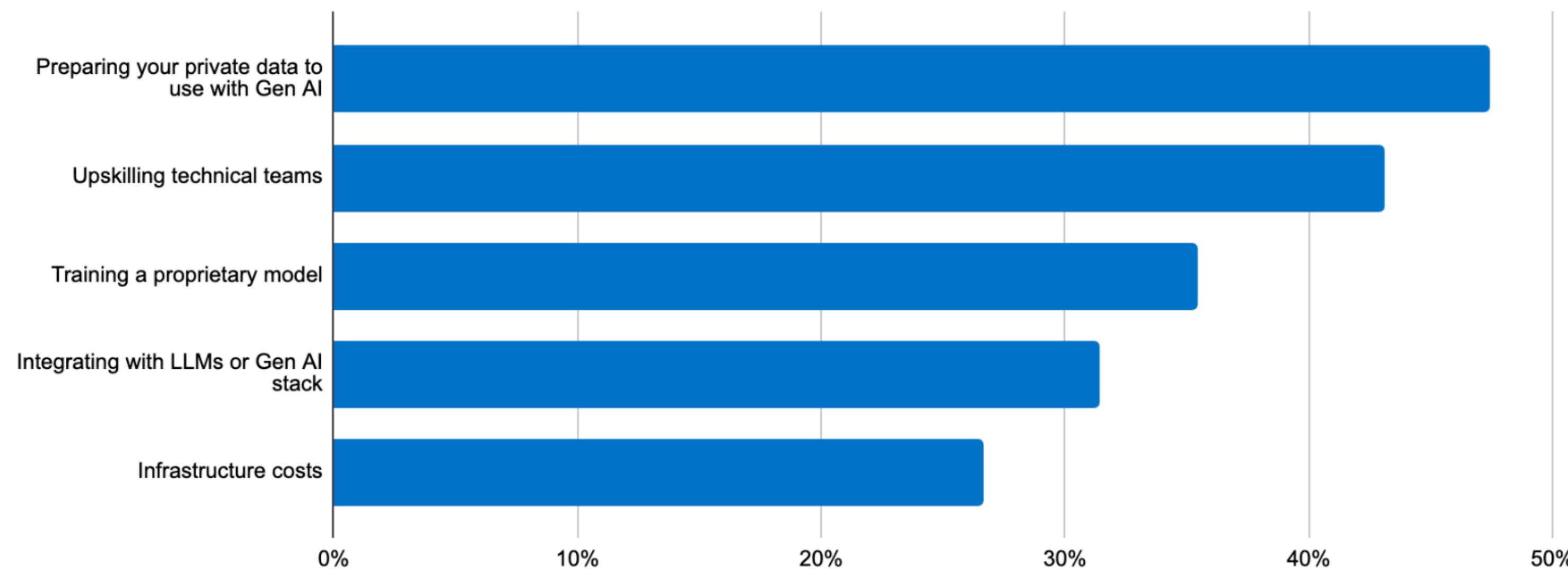


Chart: Where does your organization expect to spend the most time and resources when building generative AI use cases?

What are your organization's top considerations when selecting a vector search engine?

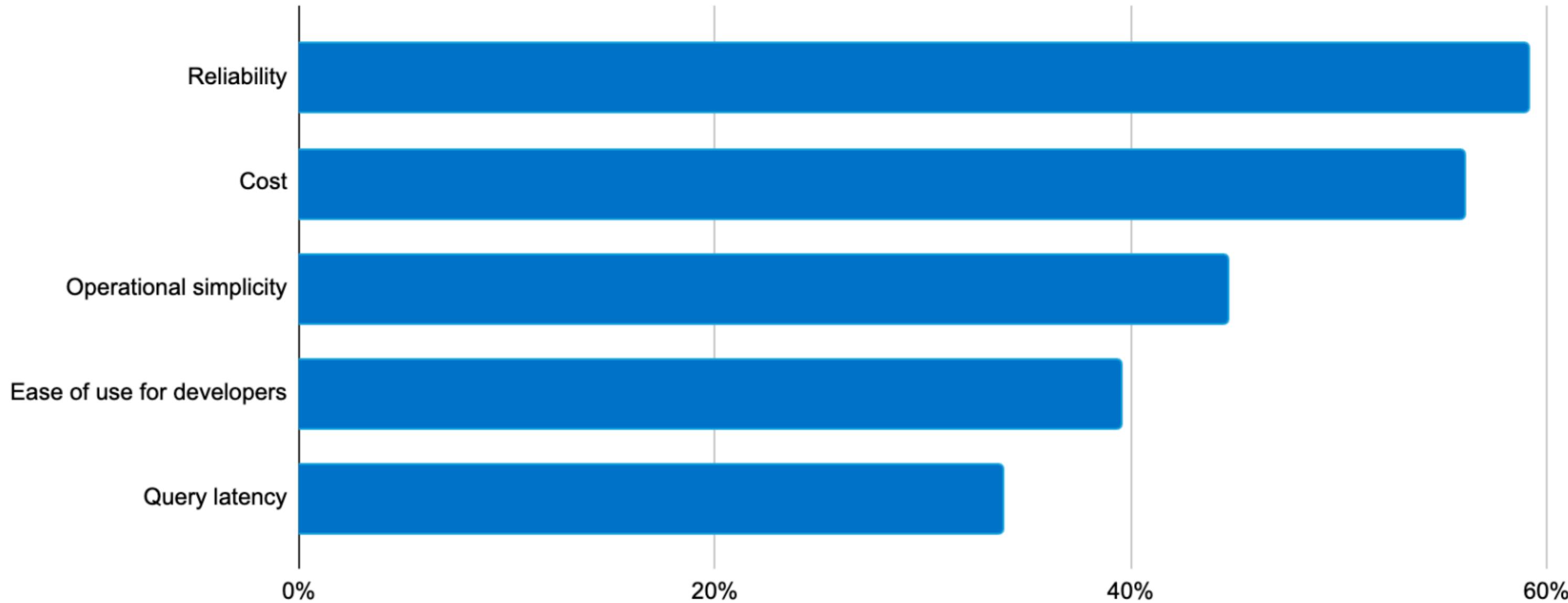


Chart: What are your organization's top considerations when selecting a vector search engine?

RESOURCES FOR DEVELOPERS | BY DEVELOPERS LIKE YOU!

Elasticsearch Labs

New!

BLOG / ML RESEARCH

Evaluating RAG: A journey through metrics



In 2020, Meta published a paper titled "[Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks](#)". This paper introduced a method for expanding the knowledge of Language

 **elasticsearch-labs** Public

About
Elasticsearch Guides, Notebooks & Example Apps for Search Applications
search-labs.elastic.co/search-labs

python search elasticsearch ai
vector applications openai elastic
chatlog chatgpt langchain
openai-chatgpt genai genaistack
vectordatabase

Readme Apache-2.0 license Security policy Activity
109 stars 178 watching 40 forks
Report repository

Languages

Jupyter Notebook	93.7%
Python	2.9%
TypeScript	1.3%
Handlebars	0.1%
JavaScript	1.6%
CSS	0.2%
Other	0.2%



Generative AI
ML Research
Vector Search
How-Tos
Integrations
Lucene

elastic.co/search-labs

github.com/elastic/elasticsearch-labs



Star our repo!



Starred 276



RESOURCES

<https://www.elastic.co/what-is/generative-ai>

<https://www.elastic.co/search-labs/blog/generative-ai-transformers-explained>

<https://www.elastic.co/what-is/vector-database>

<https://www.elastic.co/guide/en/observability/current/obs-ai-assistant.html>

<https://www.elastic.co/guide/en/observability/current/open-telemetry-direct.html>

<https://opentelemetry.io/blog/2023/ecs-otel-semconv-convergence/>

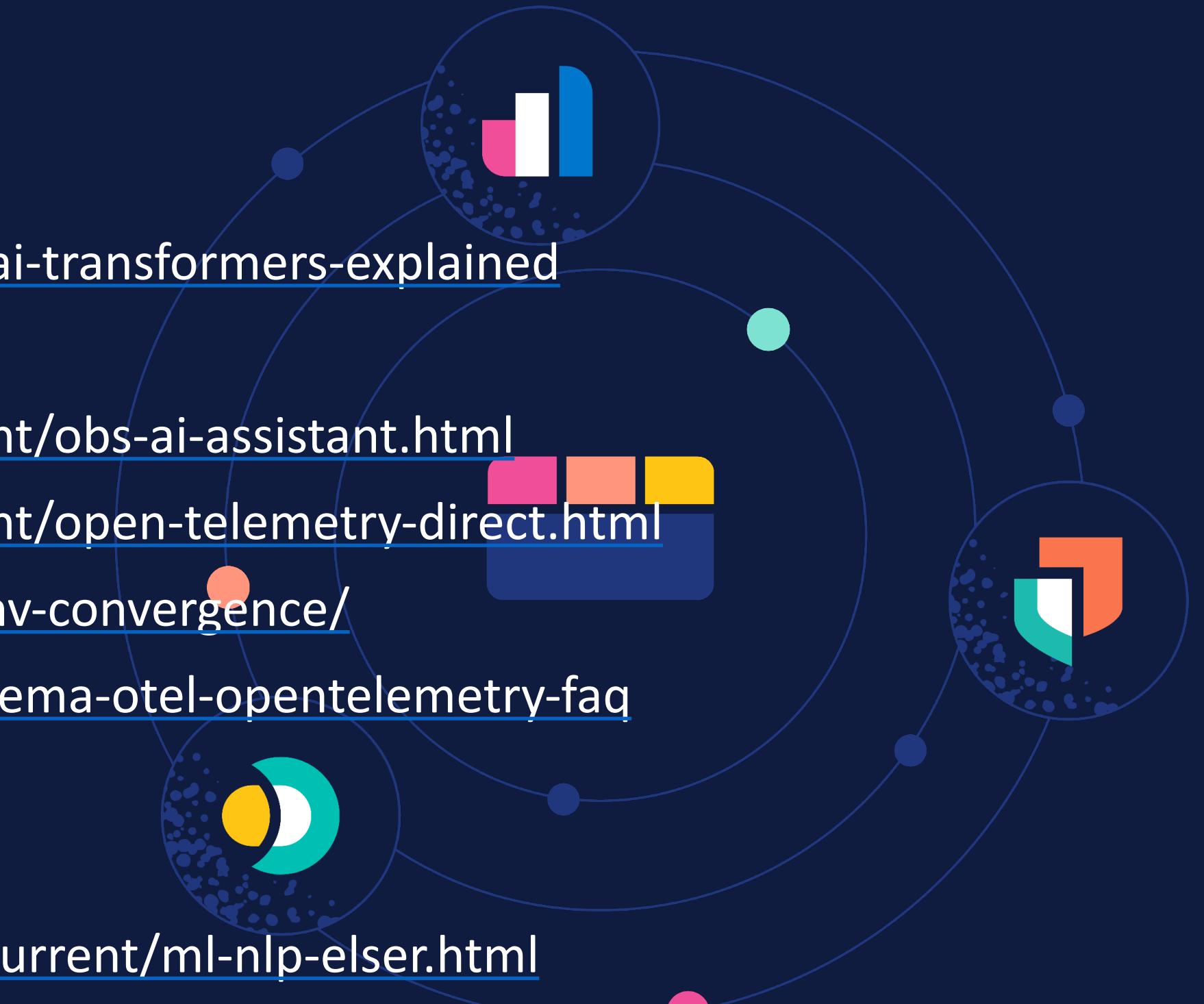
<https://www.elastic.co/blog/ecs-elastic-common-schema-otel-opentelemetry-faq>

<https://github.com/elastic/opentelemetry-demo>

<https://github.com/elastic/elasticsearch-labs>

<https://www.elastic.co/guide/en/machine-learning/current/ml-nlp-elser.html>

<https://www.elastic.co/guide/en/elasticsearch/reference/current/semantic-search-elser.html>



QUESTIONS?

THANK YOU!