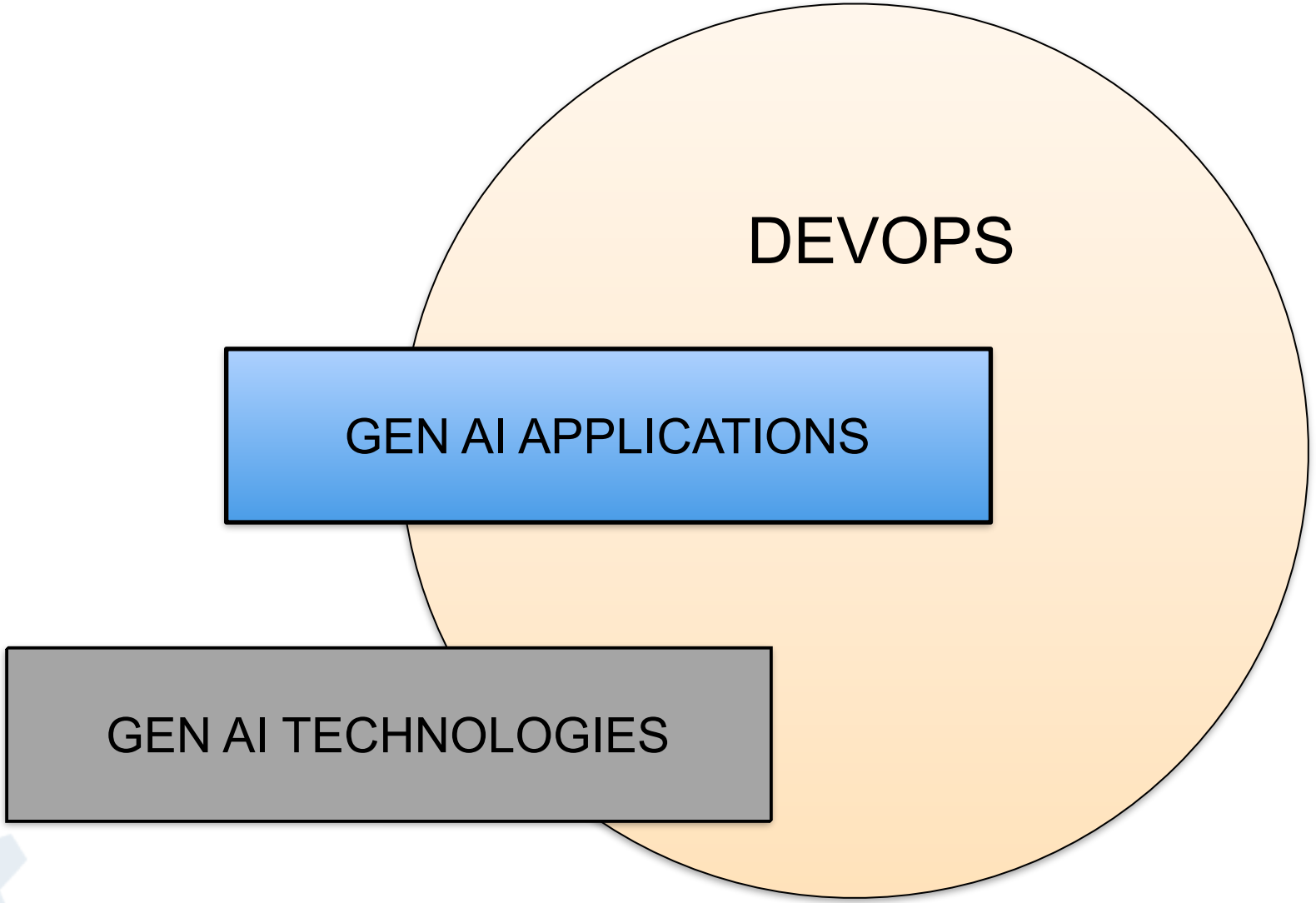
GENERATIVE AI: APPLICATIONS IN THE SERVERLESS WORLD

DIANA TODEA - SITE RELIABILITY ENGINEER

DEVOPS DAYS RALEIGH 2024



@DevOpsDaysRDU

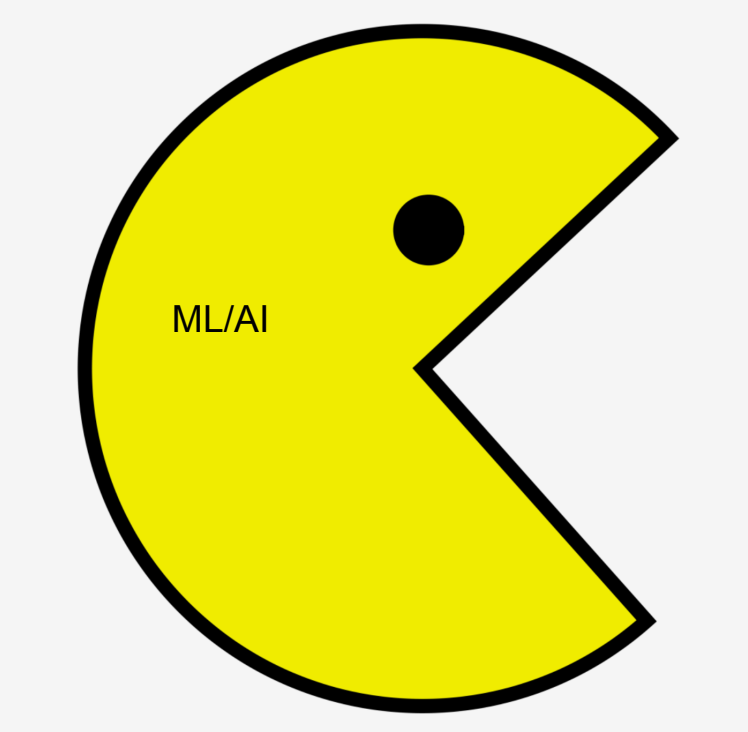
TOP 5 TRENDS FOR SEARCH AND AI IN 2024

1. ANSWERS NOT LINKS
2. DATA SAFETY AND PRIVACY
3. PROVIDE PROPER CONTEXT FOR GENERATIVE AI
4. THE NEXT WAVE OF LARGE LANGUAGE MODELS
5. AI SKILLS AND EXPERTISE FOR DEVELOPERS

@DevOpsDaysRDU

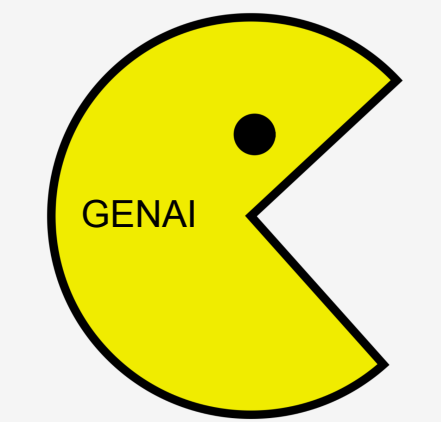
ALGORITHMS PROGRAMMED TO MAKE PREDICTIONS ON DATA

IMAGE RECOGNITION, NLP, SPEECH RECOGNITION



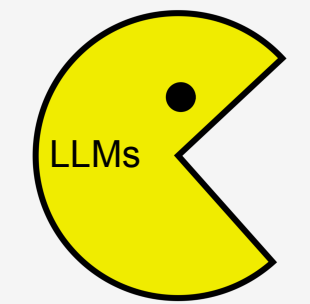
AI ALGORITHMS DESIGNED TO CREATE HIGH QUALITY CONTENT, SPECIFICALLY: TEXT, IMAGE, AUDIO

CHATBOTS, TEXT, IMAGE, MUSIC GENERATORS



DEEP LEARNING ALGORITHMS THAT CAN GENERATE TEXT

CHATBOTS, TEXT GENERATORS, TRANSLATION, WRITING, ANSWERING QUESTIONS

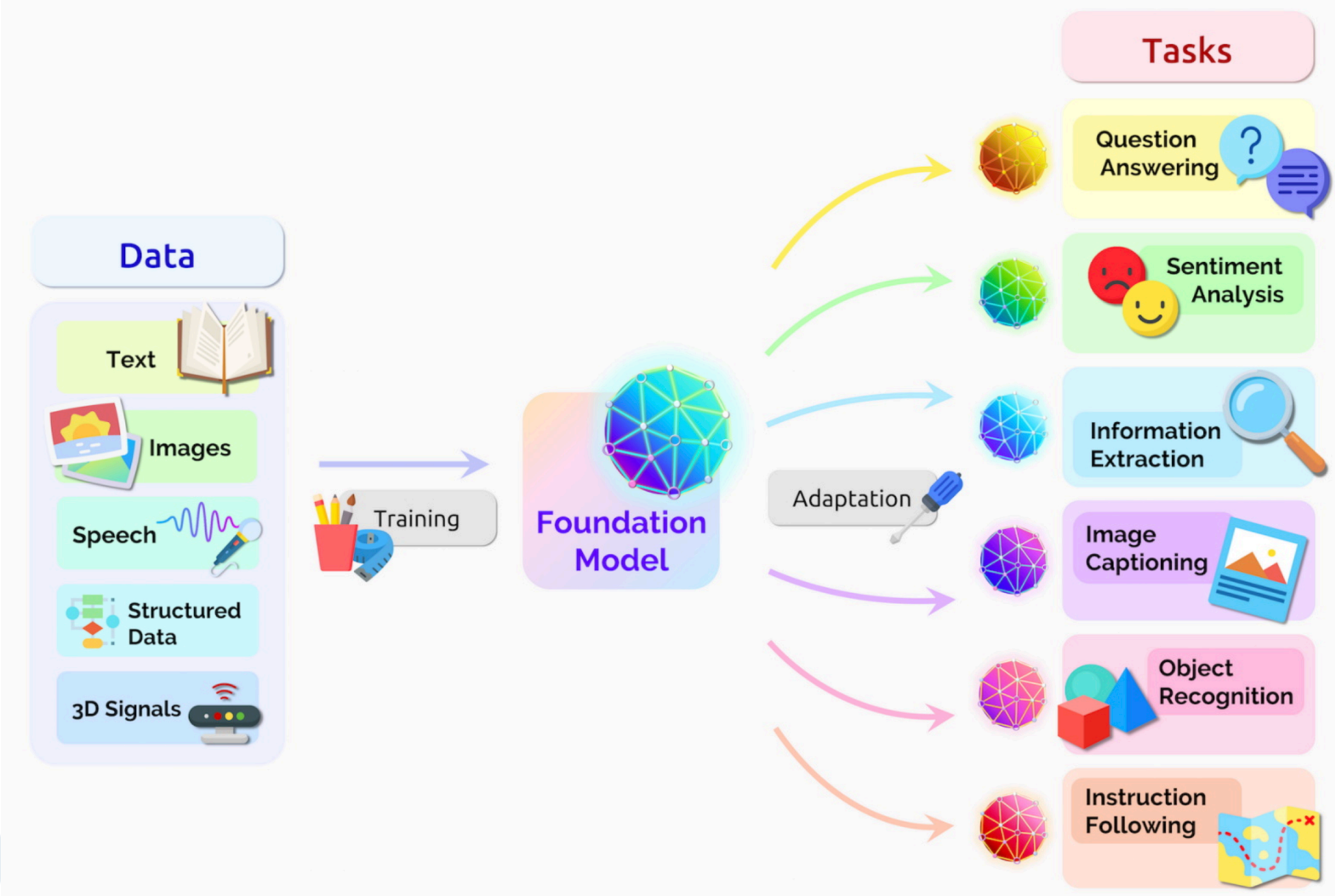


@DevOpsDaysRDU

TRANSFORMER DEFINITION

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.

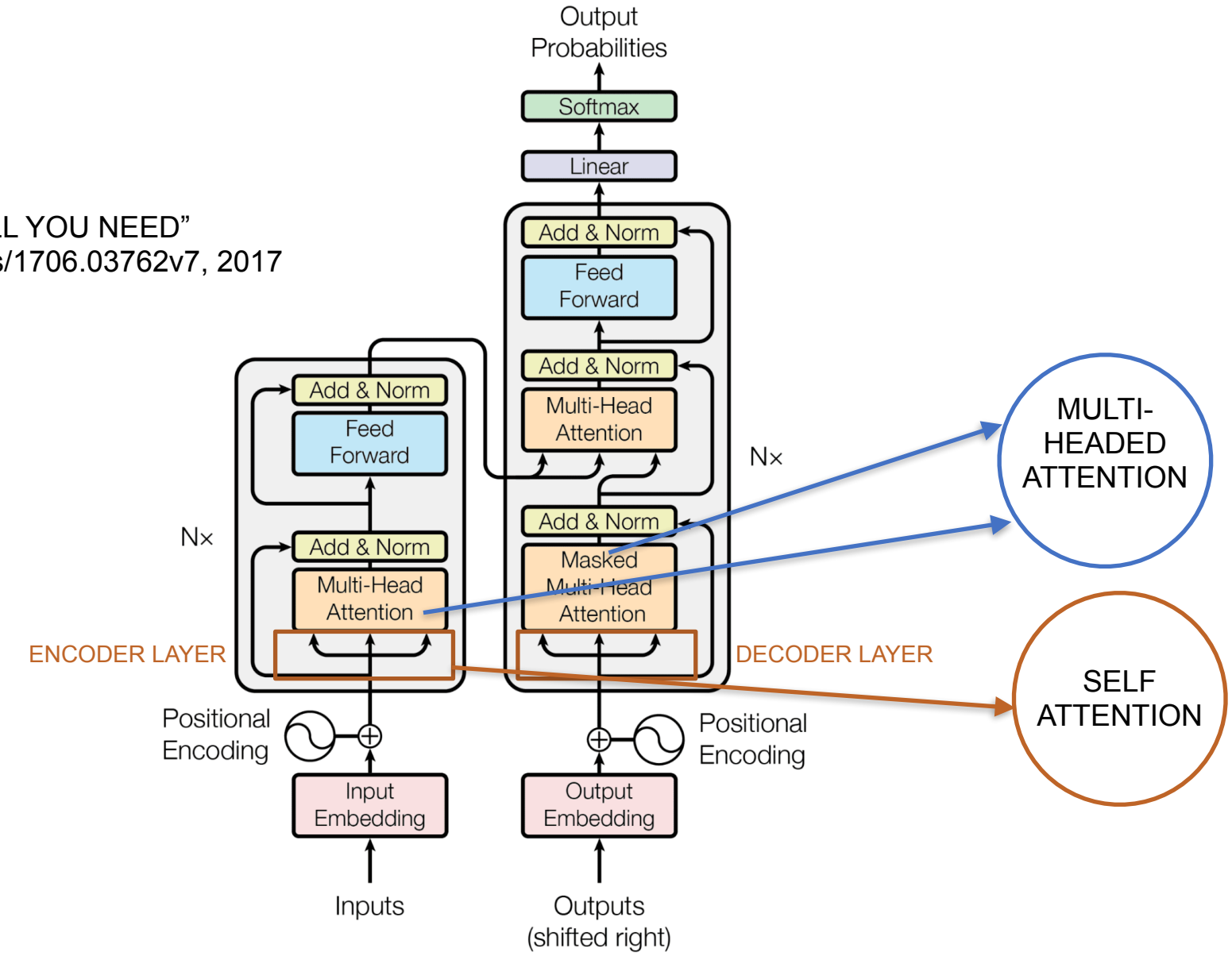
WHAT CAN TRANSFORMERS DO?



SOURCE: NVIDIA

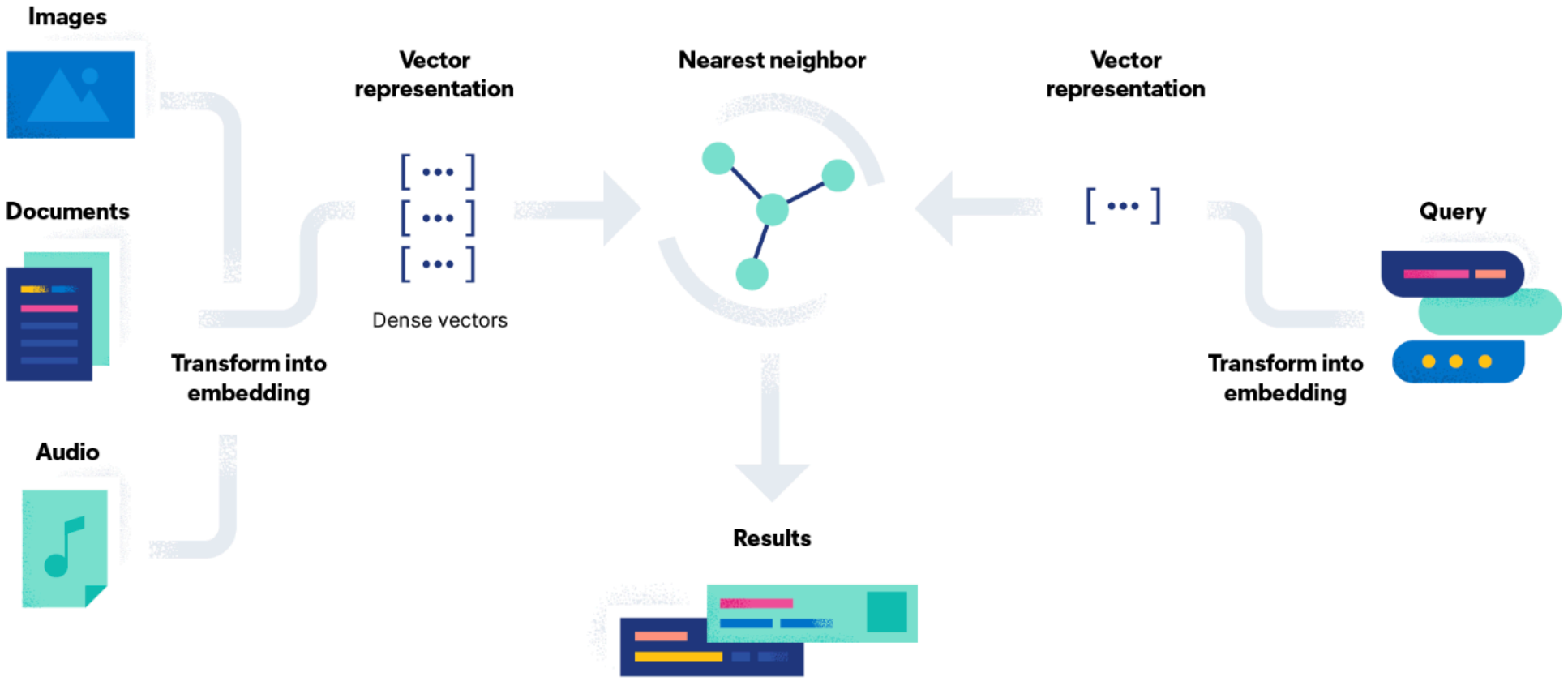
TRANSFORMER ARCHITECTURE

"ATTENTION IS ALL YOU NEED"
https://arxiv.org/abs/1706.03762v7, 2017

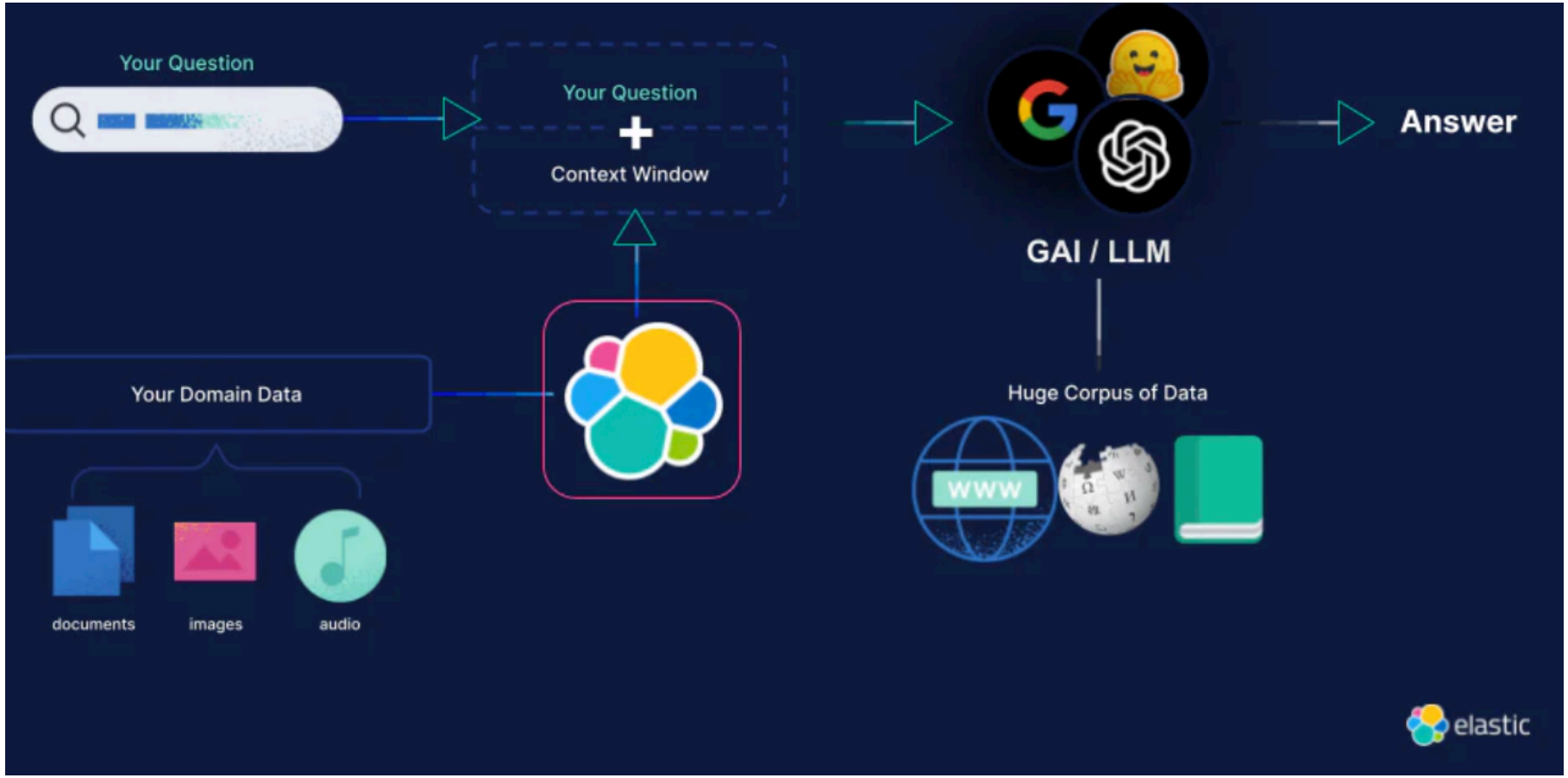


@DevOpsDaysRDU

GEN AI ARCHITECTURE

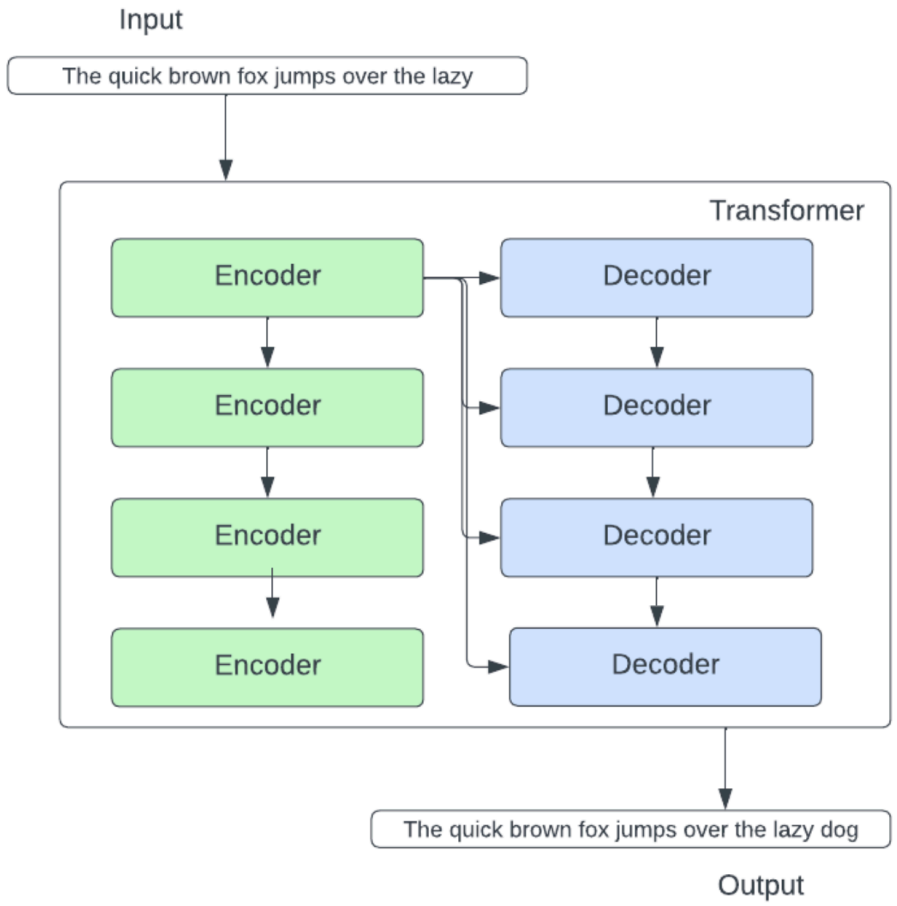


@DevOpsDaysRDU



@DevOpsDaysRDU

RETRIEVAL AUGMENTED GENERATION (RAG)



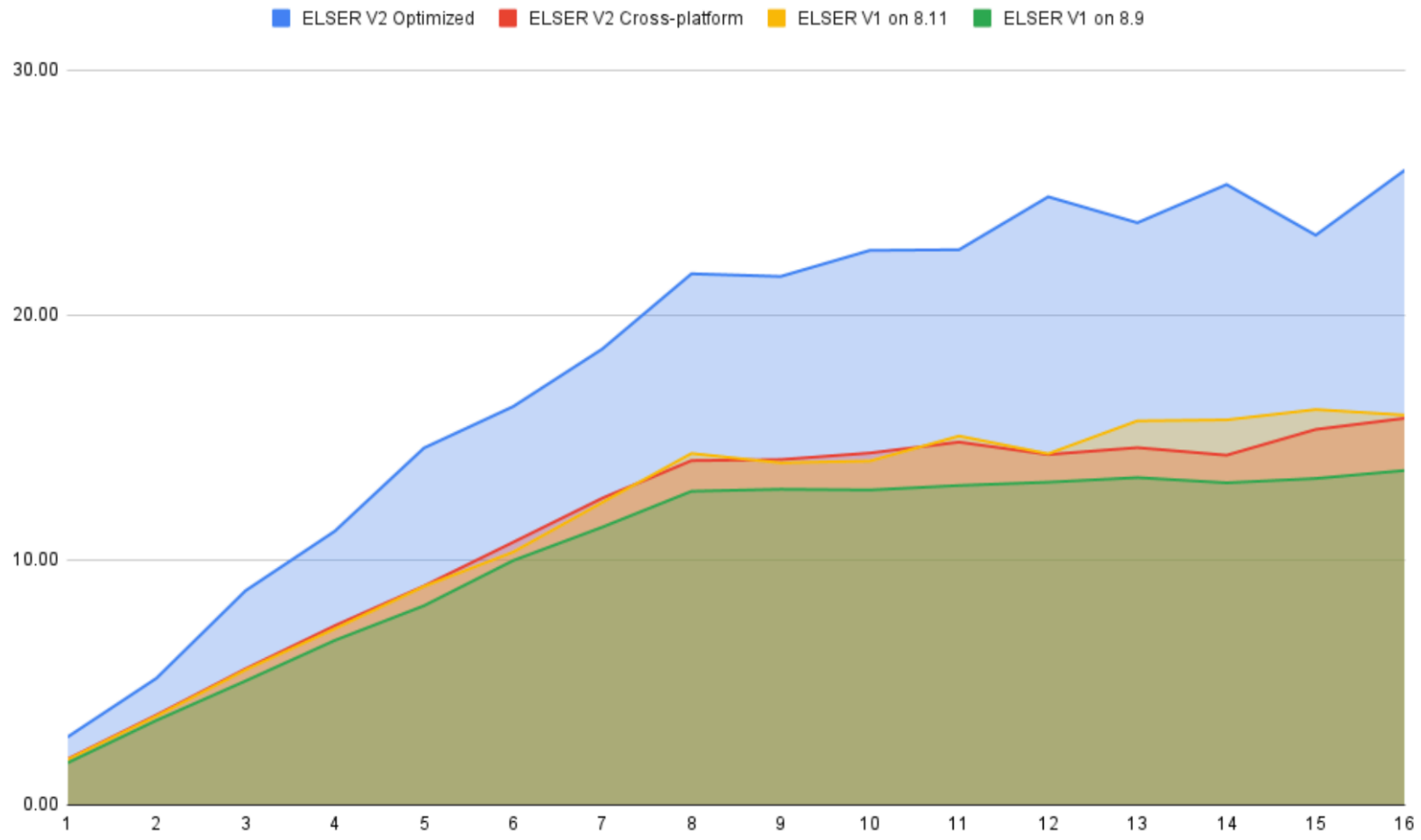
Retrieval augmented generation (RAG) is a technique that supplements text generation with information from private or proprietary data sources.

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.

DEVOPSDAYS RALEIGH

@DevOpsDaysRDU

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 sp
and contextual meaning over literal term matching, optimized specific
queries on the Elastic platform.

View documentation

Choose a model

- **Cross platform**
  .elser_model_2
- **Intel and Linux optimized** — Recommended
  .elser_model_2_linux-x86_64

DEVOPSDAYS RALEIGH

@DevOpsDaysRDU

Sherlock

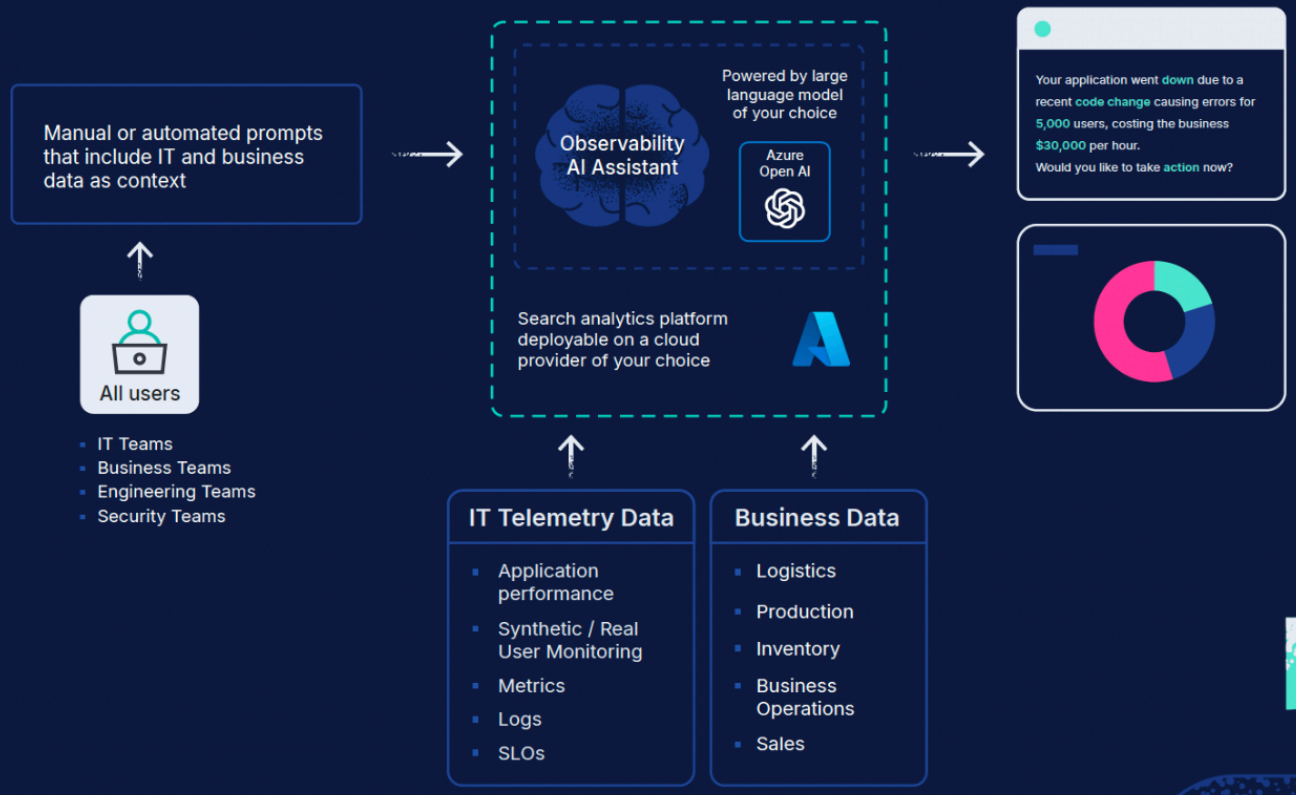
Research

Explain

@DevOpsDaysRDU

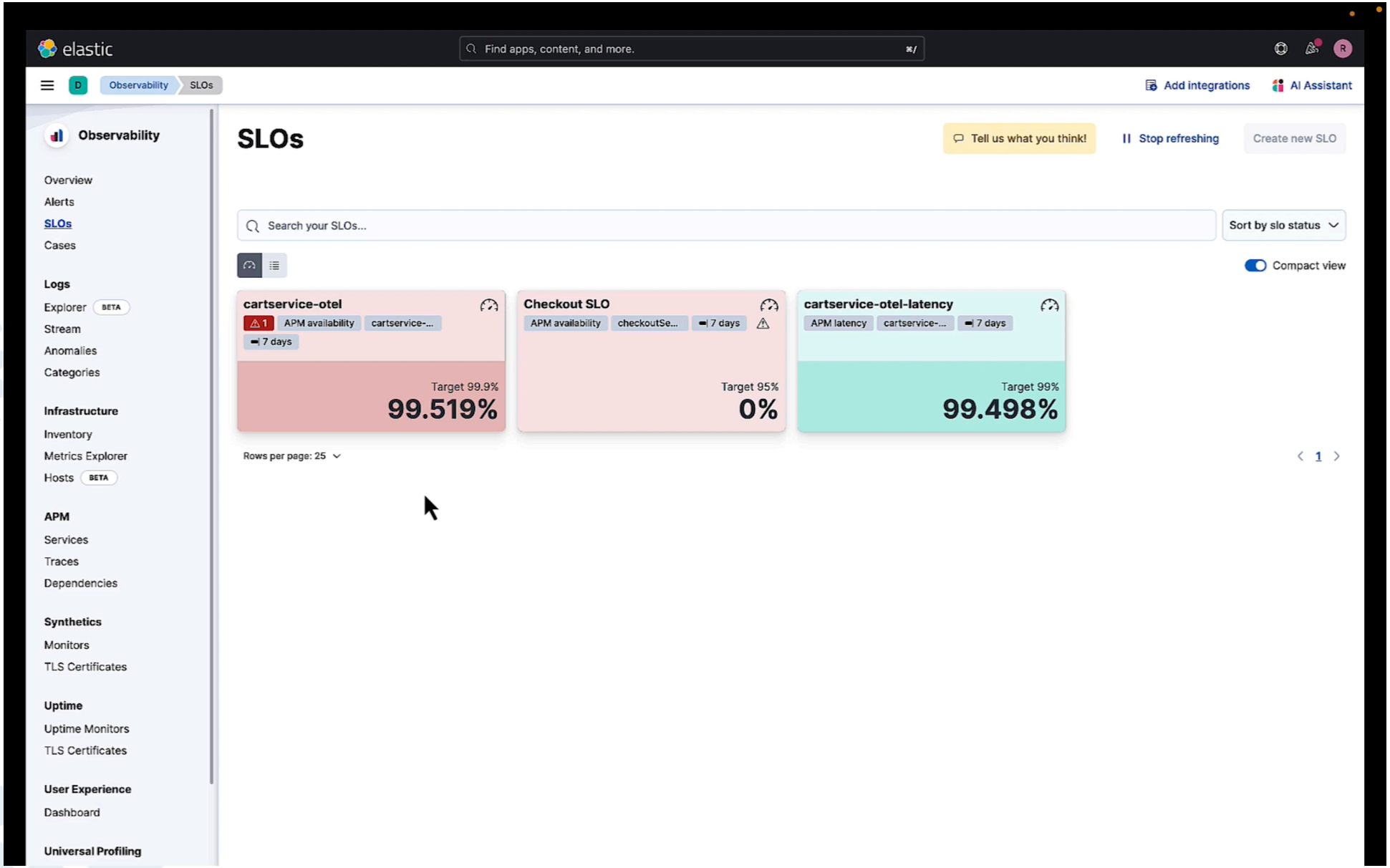
elastic Microsoft

Blueprint for Business Observability



DEVOPSDAYS RALEIGH

@DevOpsDaysRDU



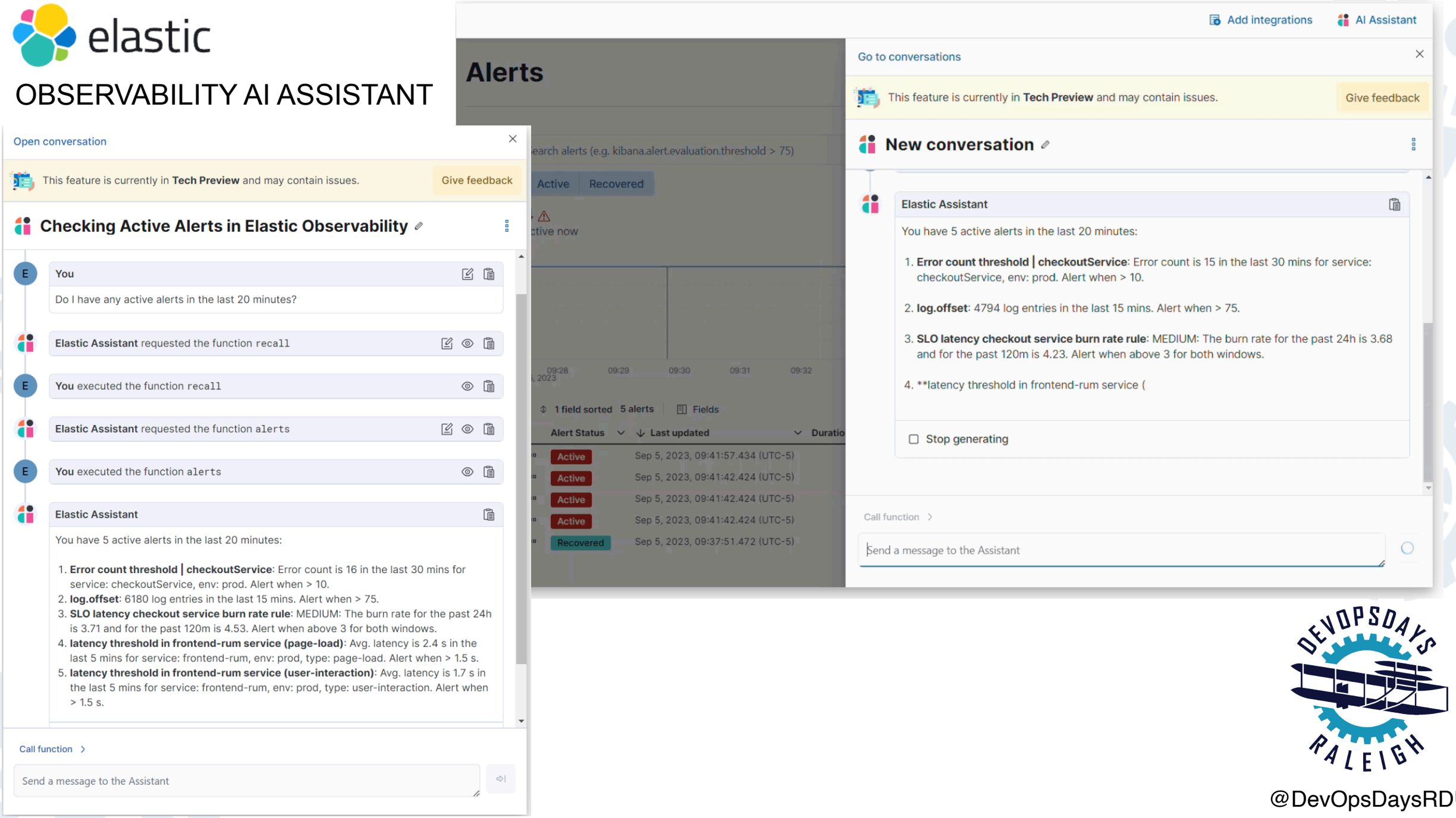
OpenTelemetry

DEVOPSDAYS RALEIGH

@DevOpsDaysRDU

elastic

OBSERVABILITY AI ASSISTANT



DEVOPSDAYS RALEIGH

@DevOpsDaysRDU

Discussing ArgoCD Workqueue Depth

Elastic Assistant

The alert '[ML] Stateful: FATAL Log Messages' indicates that there are fatal log messages in your Elasticsearch cluster. Here are some general troubleshooting steps you can take:

1. **Check the Logs:** The first step in troubleshooting this issue would be to check the logs for any fatal errors. You can do this by accessing the Kibana Discover app and filtering for fatal log messages. The alert indicates that these messages are coming from the 'logging-*:cluster-elasticsearch-*' index.

2. **Identify the Source:** Try to identify the source of these fatal messages. They could be coming from a specific service, node, or part of your application. This can help you narrow down where the issue might be.

3. **Check for Recent Changes:** Have there been any recent changes to your Elasticsearch cluster or the application that could have caused these errors? This could include things like configuration changes, updates, or new deployments.

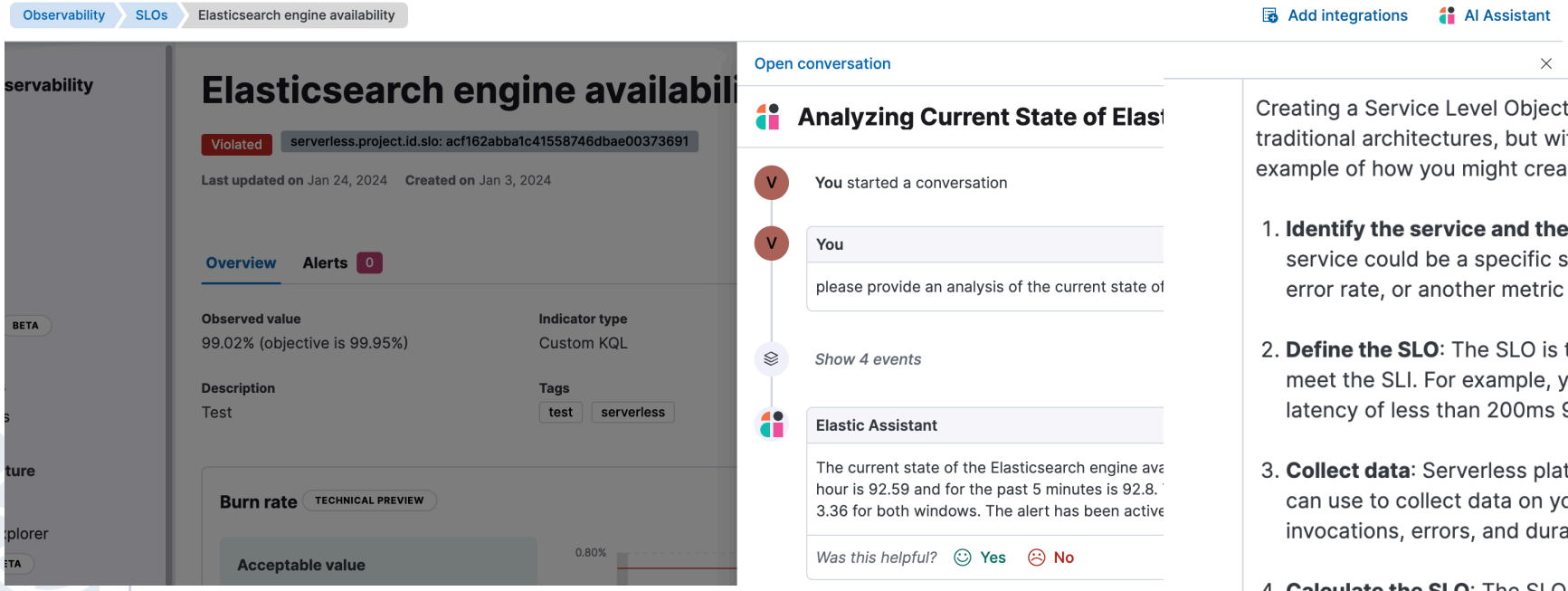
4. **Investigate the Error Messages:** The error messages themselves can often provide clues as to what's going wrong. They might indicate a problem with a specific part of your application, a resource issue, a network problem, etc.

5. **Monitor Resource Usage:** Check the resource usage of your Elasticsearch cluster. Fatal errors can sometimes be caused by resource issues, such as running out of memory or disk space.

6. **Contact Support:** If you're unable to resolve the issue yourself, you may need to contact Elastic support for further assistance.

DEVOPSDAYS RALEIGH

@DevOpsDaysRDU



Elastic Assistant

Creating a p99 SLO (Service Level Objective) involves defining the "good" events that contribute towards meeting the SLO, and the total events that are relevant for the SLO.

Here's a general process on how to create a p99 SLO:

1. **Identify the service and the SLI (Service Level Indicator)**: The service could be any part of your system that you want to measure the reliability of. The SLI is a specific metric that you want to measure. For a p99 SLO, this could be the latency of a service.

2. **Define the SLO**: The SLO is the target percentage of the time that you want your service to meet the SLI. For a p99 SLO, this would be 99%.

3. **Collect data**: You need to collect data on the SLI and the total events. This could involve instrumenting your service to collect this data, or using existing data if it's already available.

4. **Calculate the SLO**: The SLO is calculated as the percentage of good events out of the total events. For a p99 SLO, you would calculate the percentage of events where the latency is below a certain threshold.

5. **Monitor and alert on the SLO**: Once you've defined and calculated your SLO, you need to set up monitoring and alerting. This will let you know when you're at risk of not meeting your SLO, so you can take action to improve reliability.

Creating a Service Level Objective (SLO) for a serverless architecture involves similar steps as with traditional architectures, but with a focus on the specific characteristics of serverless. Here's an example of how you might create an SLO for a serverless function:

1. **Identify the service and the SLI (Service Level Indicator)**: In a serverless architecture, the service could be a specific serverless function. The SLI could be the latency of the function, the error rate, or another metric that is important for your application.

2. **Define the SLO**: The SLO is the target percentage of the time that you want your service to meet the SLI. For example, you might set an SLO that your serverless function should have a latency of less than 200ms 99% of the time.

3. **Collect data**: Serverless platforms often provide monitoring and logging capabilities that you can use to collect data on your SLIs. For example, AWS Lambda provides metrics on function invocations, errors, and duration.

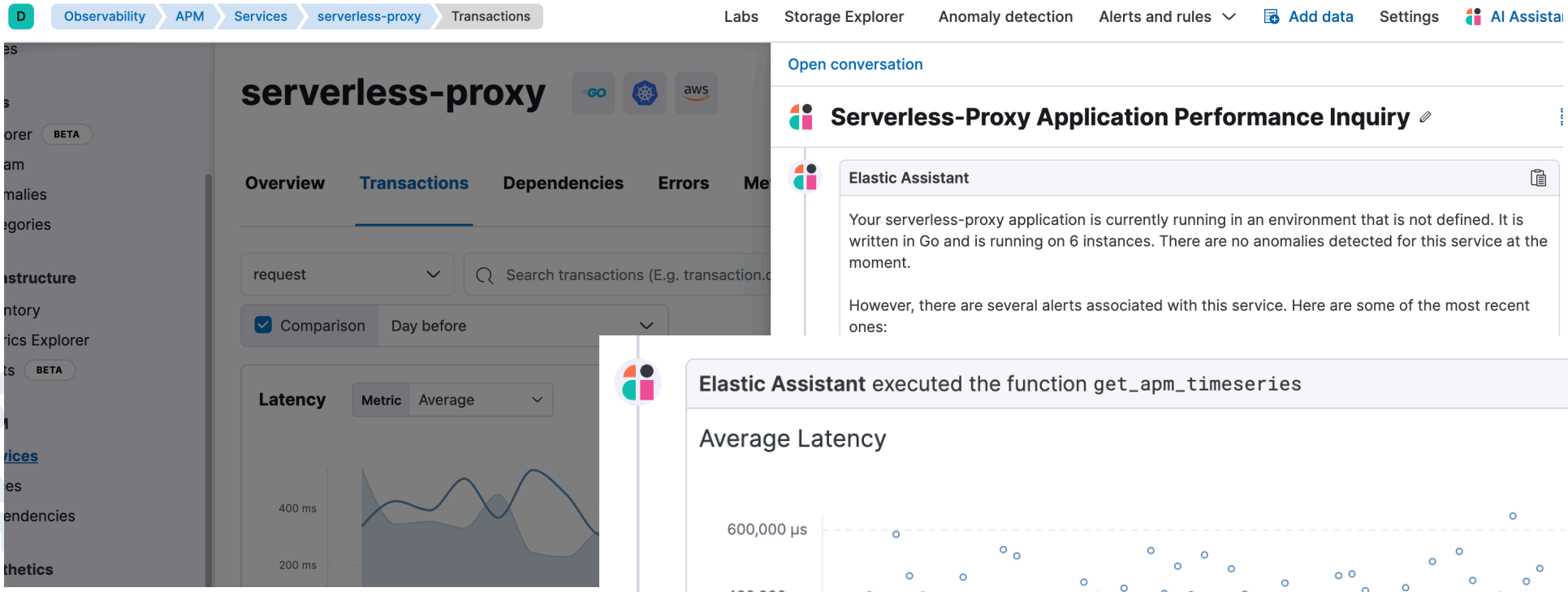
4. **Calculate the SLO**: The SLO is calculated as the percentage of good events out of the total events. For the example SLO above, you would calculate the percentage of function invocations where the latency is less than 200ms.

5. **Monitor and alert on the SLO**: Once you've defined and calculated your SLO, you need to set up monitoring and alerting. This will let you know when you're at risk of not meeting your SLO, so you can take action to improve reliability.

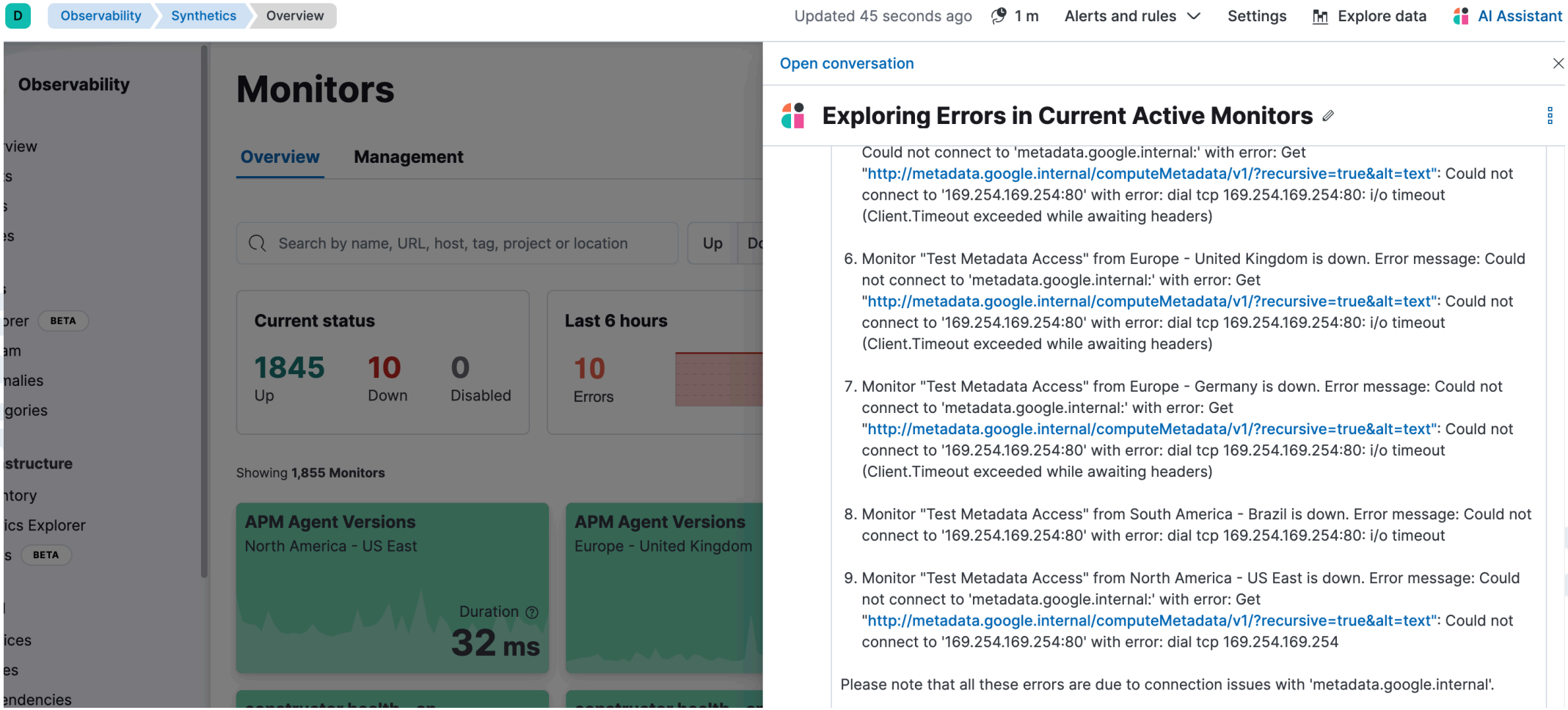
Remember, the goal of an SLO is to help you understand the reliability of your service and make informed decisions about where to invest in improving reliability. The exact SLIs and SLOs you choose will depend on what is most important for your application and users.

DEVOPSDAYS RALEIGH

@DevOpsDaysRDU



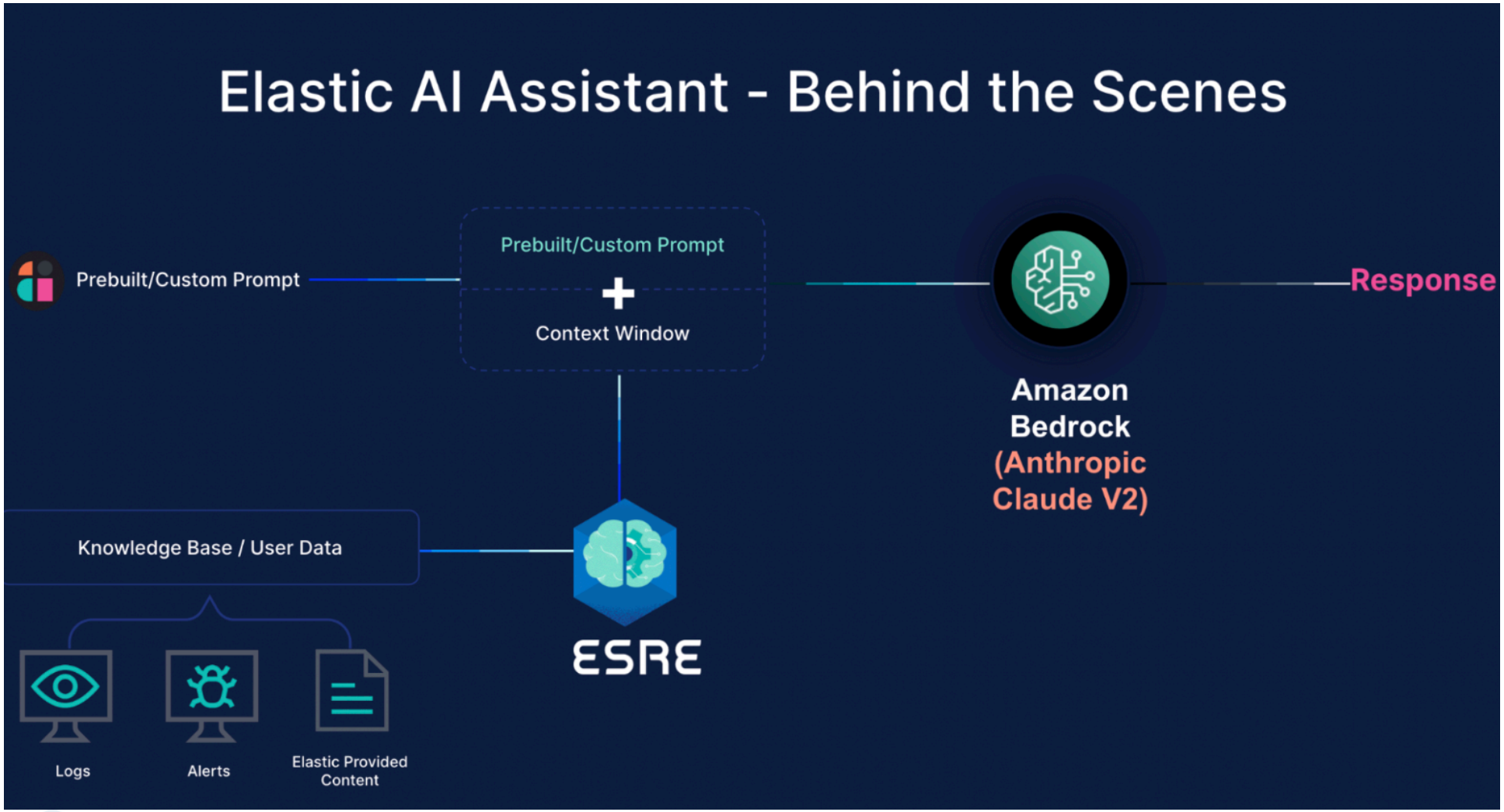
@DevOpsDaysRDU



DEVOPSDAYS RALEIGH

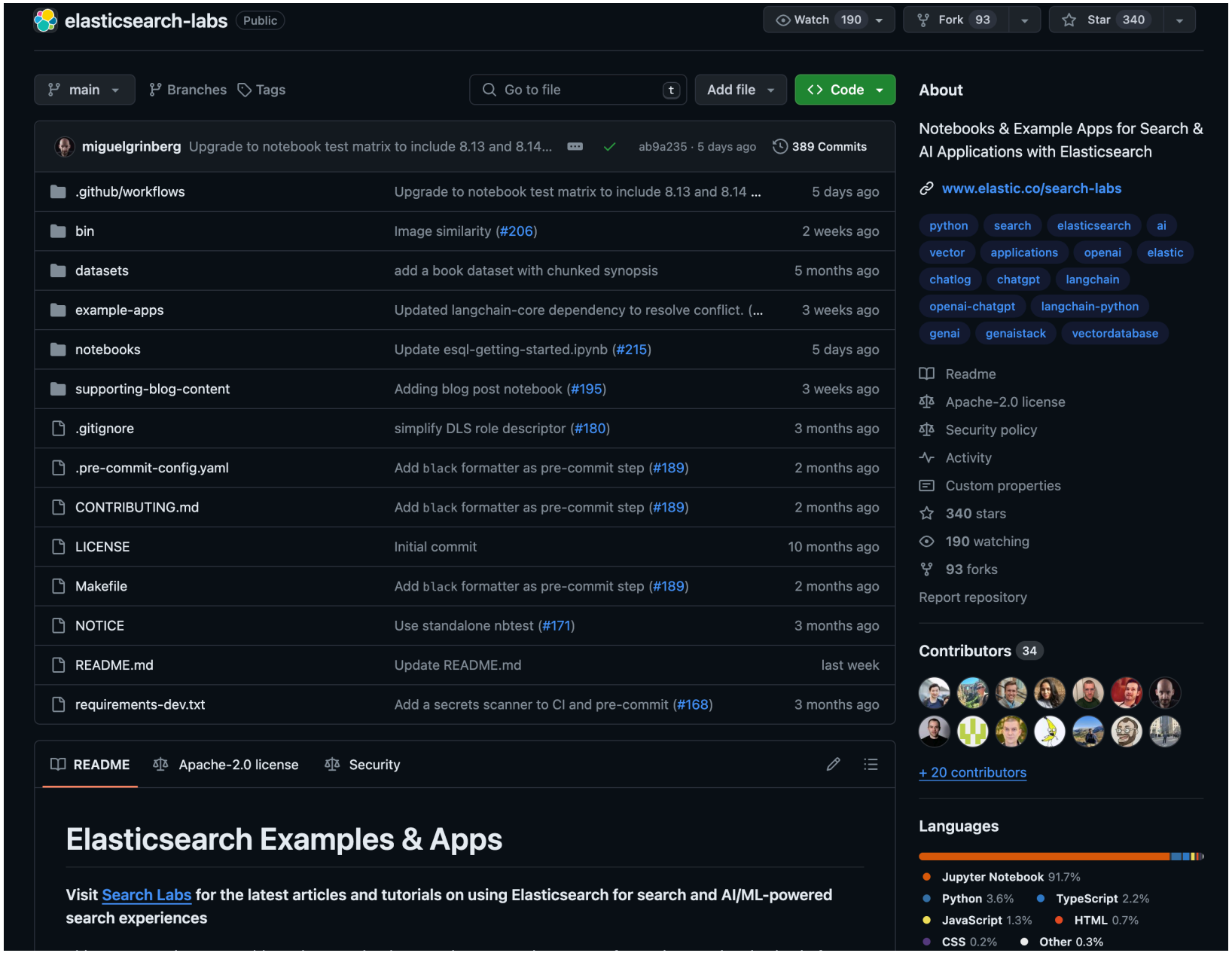
@DevOpsDaysRDU

Elastic AI Assistant - Behind the Scenes



@DevOpsDaysRDU

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



DEVOPSDAYS RALEIGH

@DevOpsDaysRDU

AI ASSISTANTS ON SERVERLESS

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

MOVING FORWARD

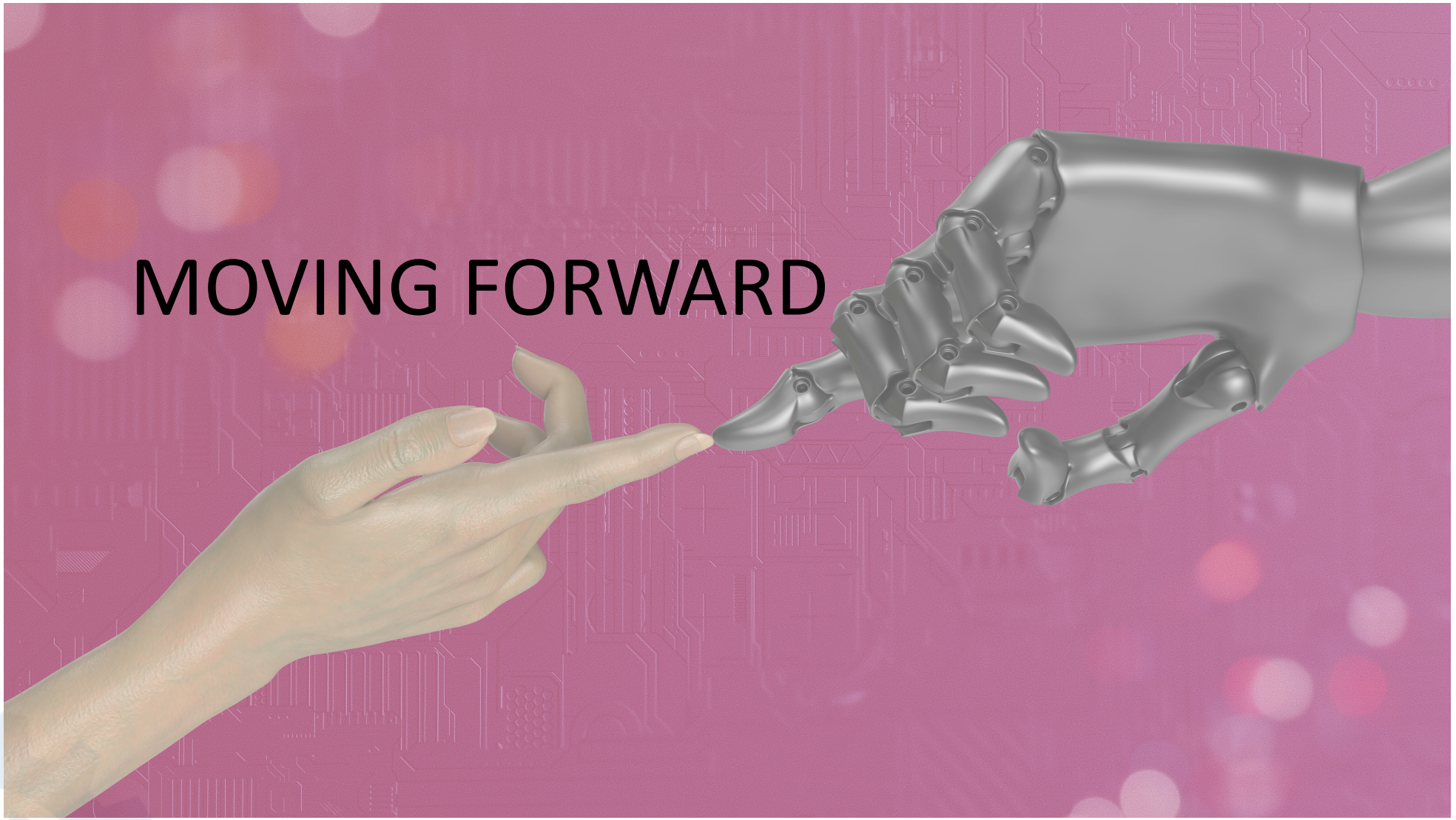


Photo by Igor Omilaev on Unsplash

@DevOpsDaysRDU

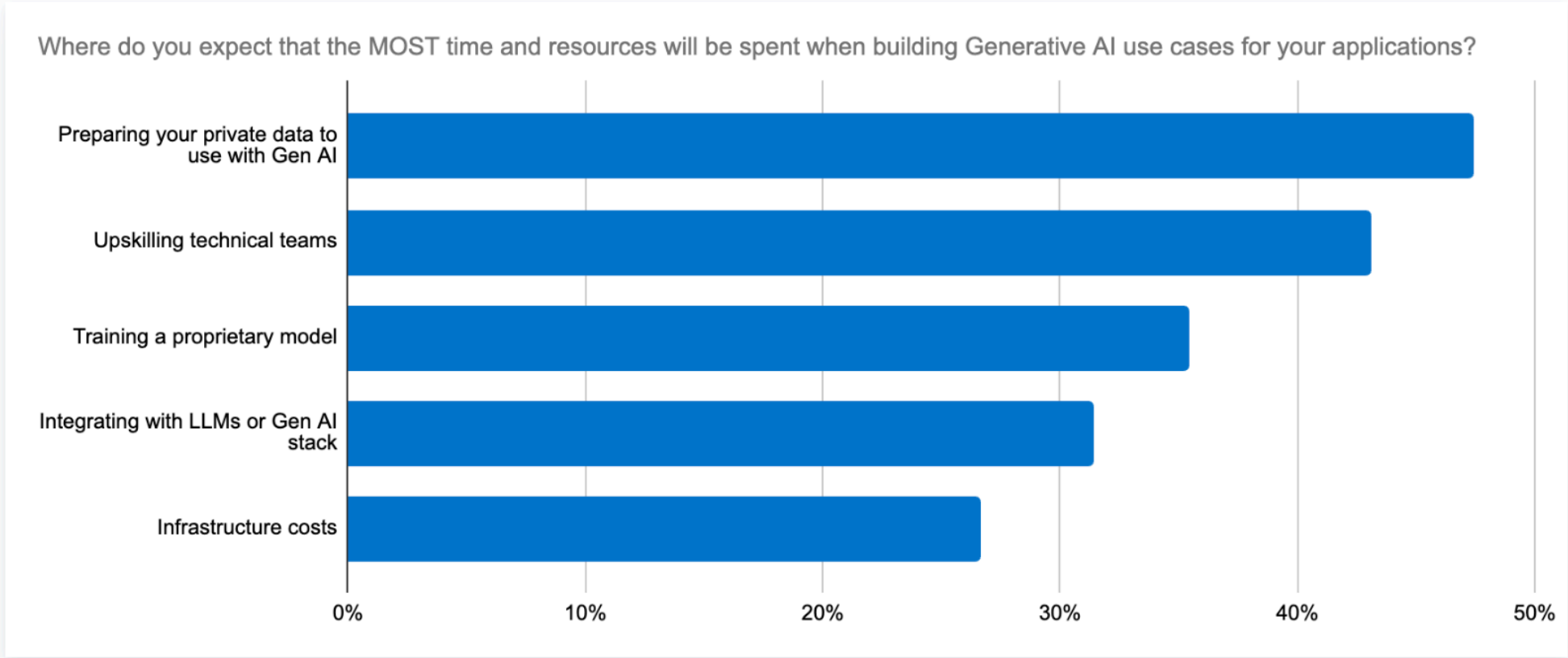


Chart: Where does your organization expect to spend the most time and resources when building generative AI use cases?

@DevOpsDaysRDU

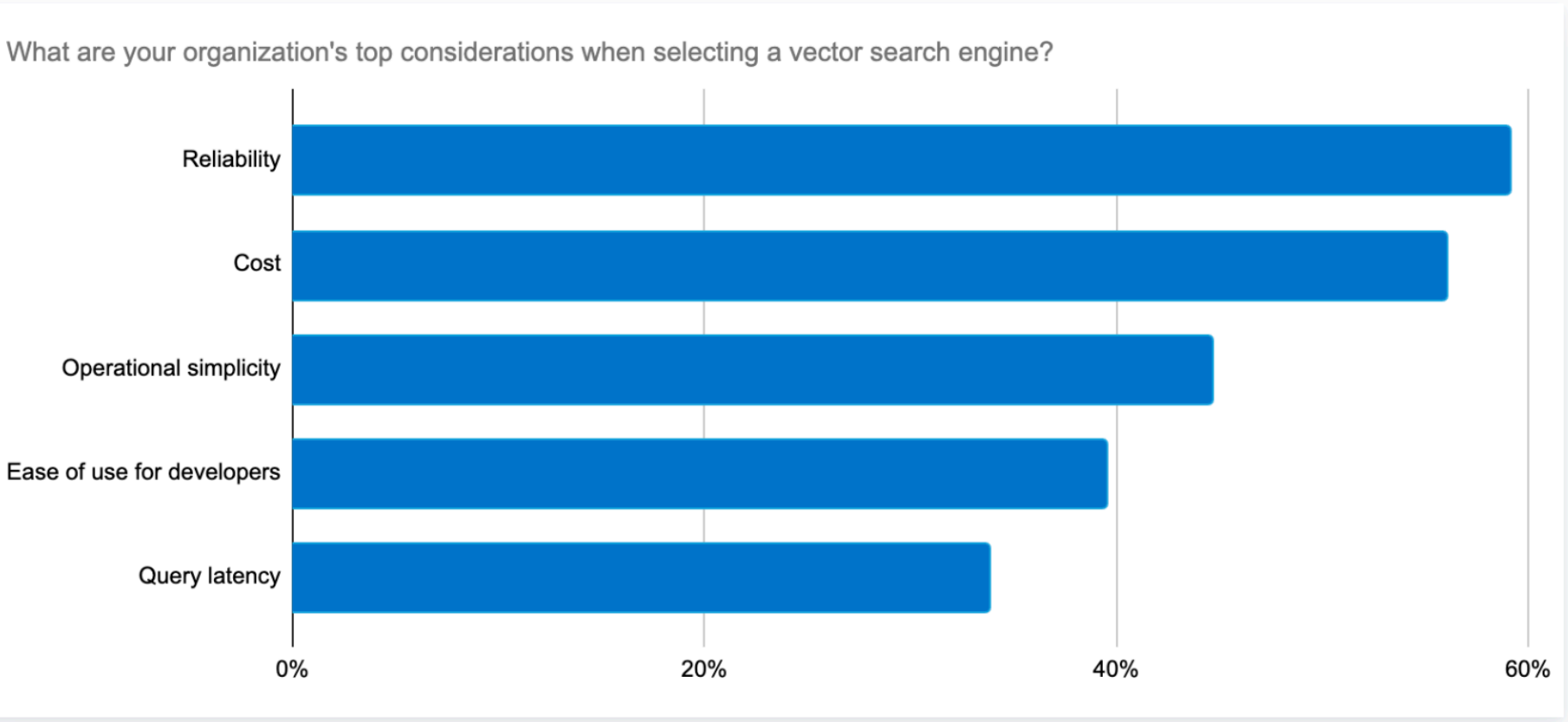


Chart: What are your organization's top considerations when selecting a vector search engine?

@DevOpsDaysRDU

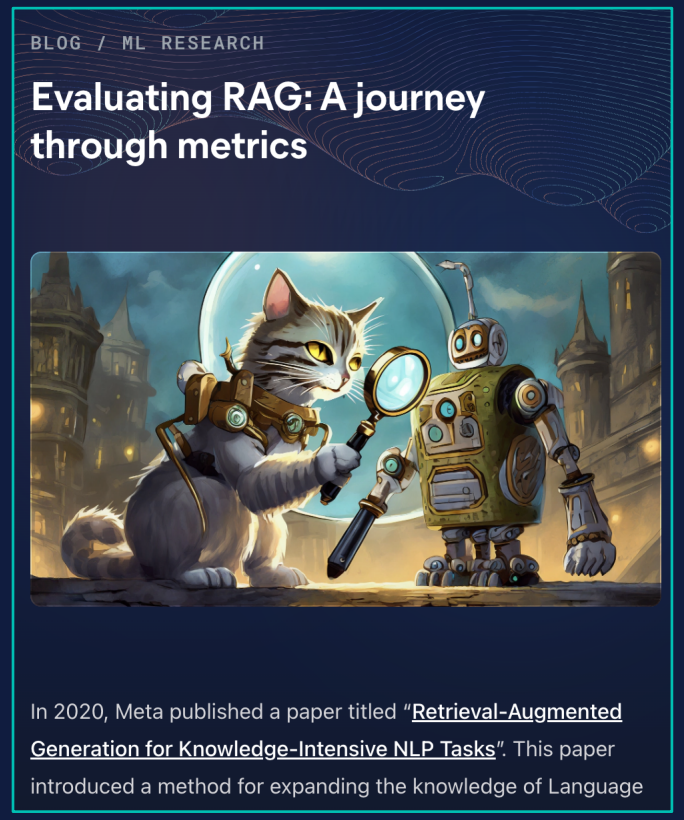
RESOURCES

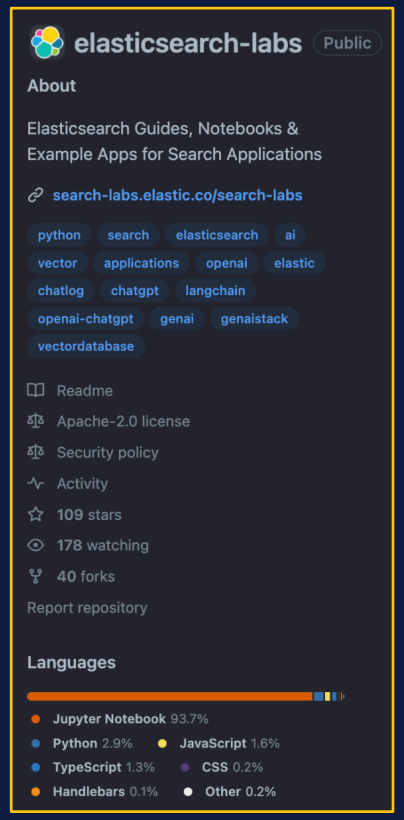
https://www.elastic.co/what-is/generative-ai
https://www.elastic.co/search-labs/blog/generative-ai-transformers-explained
https://www.elastic.co/guide/en/observability/current/obs-ai-assistant.html
https://www.elastic.co/guide/en/observability/current/open-telemetry-direct.html
https://opentelemetry.io/blog/2023/ecs-otel-semconv-convergence/
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

RESOURCES FOR DEVELOPERS | BY DEVELOPERS LIKE YOU!

Elasticsearch Labs

New!





Generative AI
ML Research
Vector Search
How-Tos
Integrations
Lucene

elastic.co/**search-labs**

github.com/elastic/**elasticsearch-labs**

Star our repo!

Starred 276

elastic

QUESTIONS?

THANK YOU!

YOUR FEEDBACK HERE!

elastic