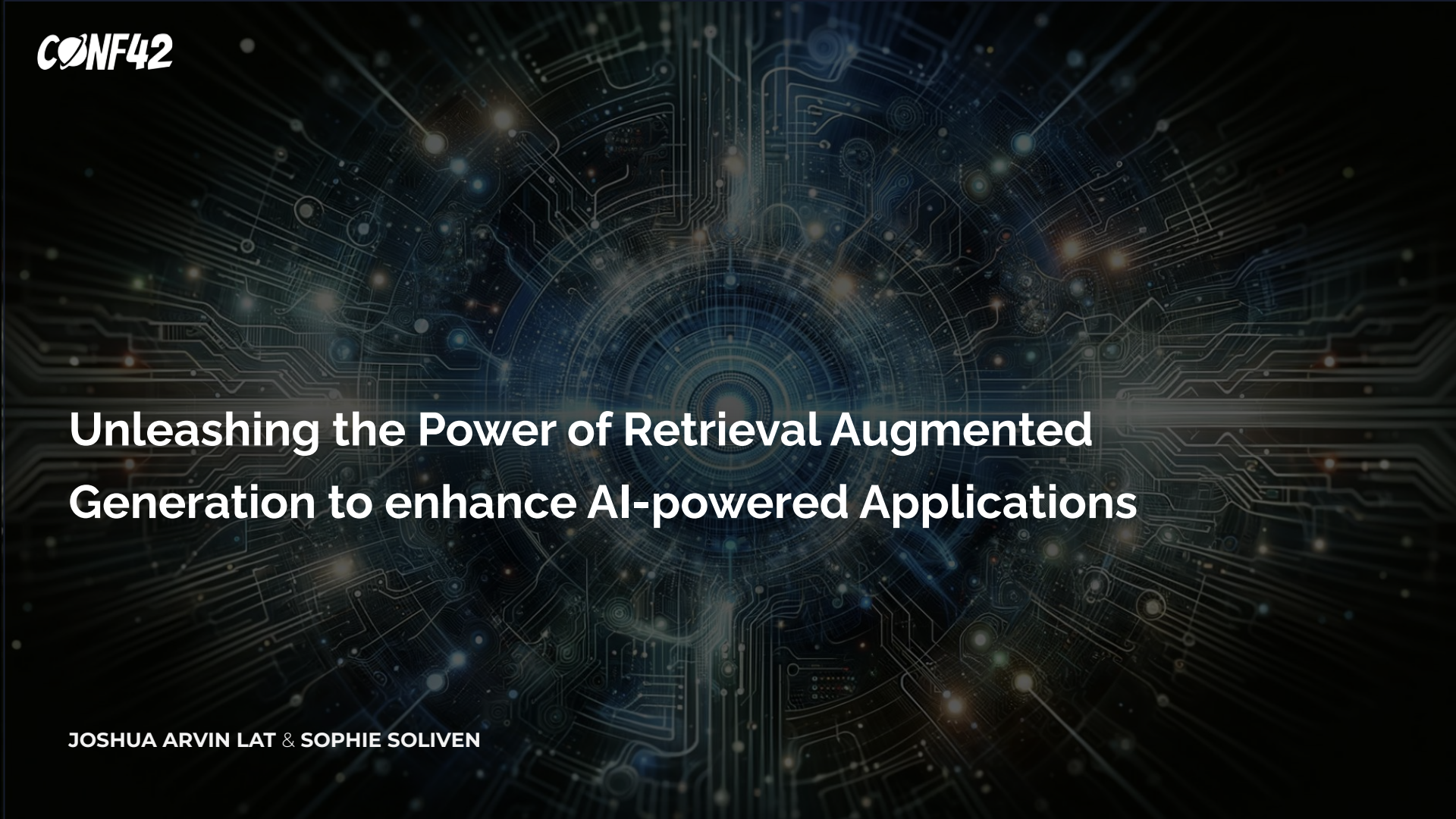
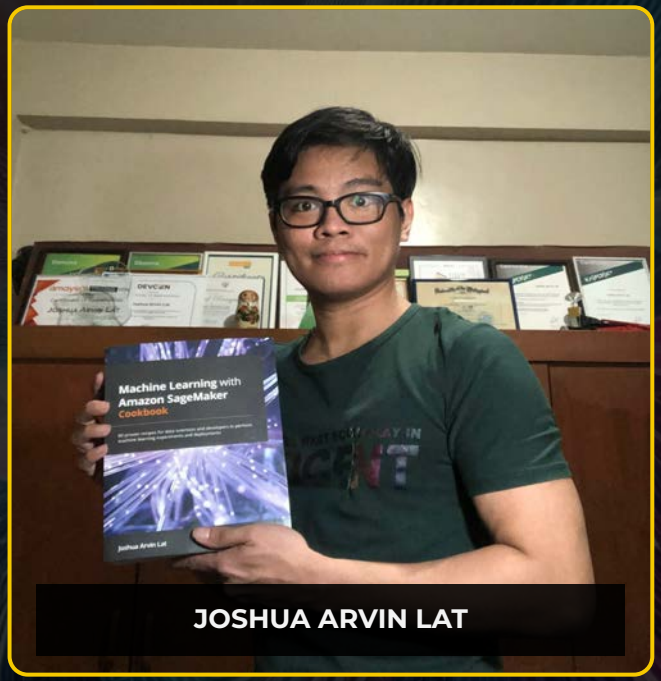


Unleashing the Power of Retrieval Augmented Generation to enhance AI-powered Applications

JOSHUA ARVIN LAT & SOPHIE SOLIVEN

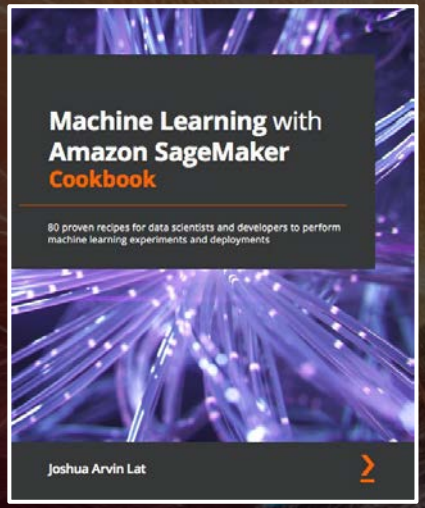
SOPHIE SOLIVEN

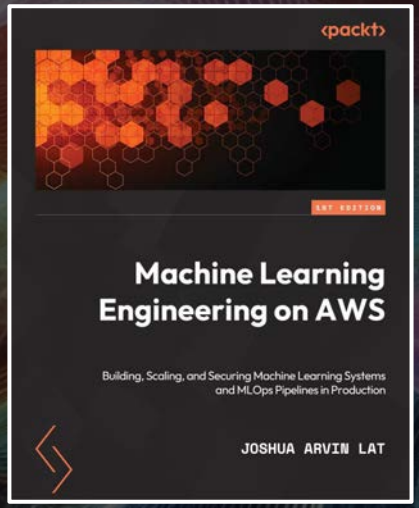
Operations Director of
Edamama



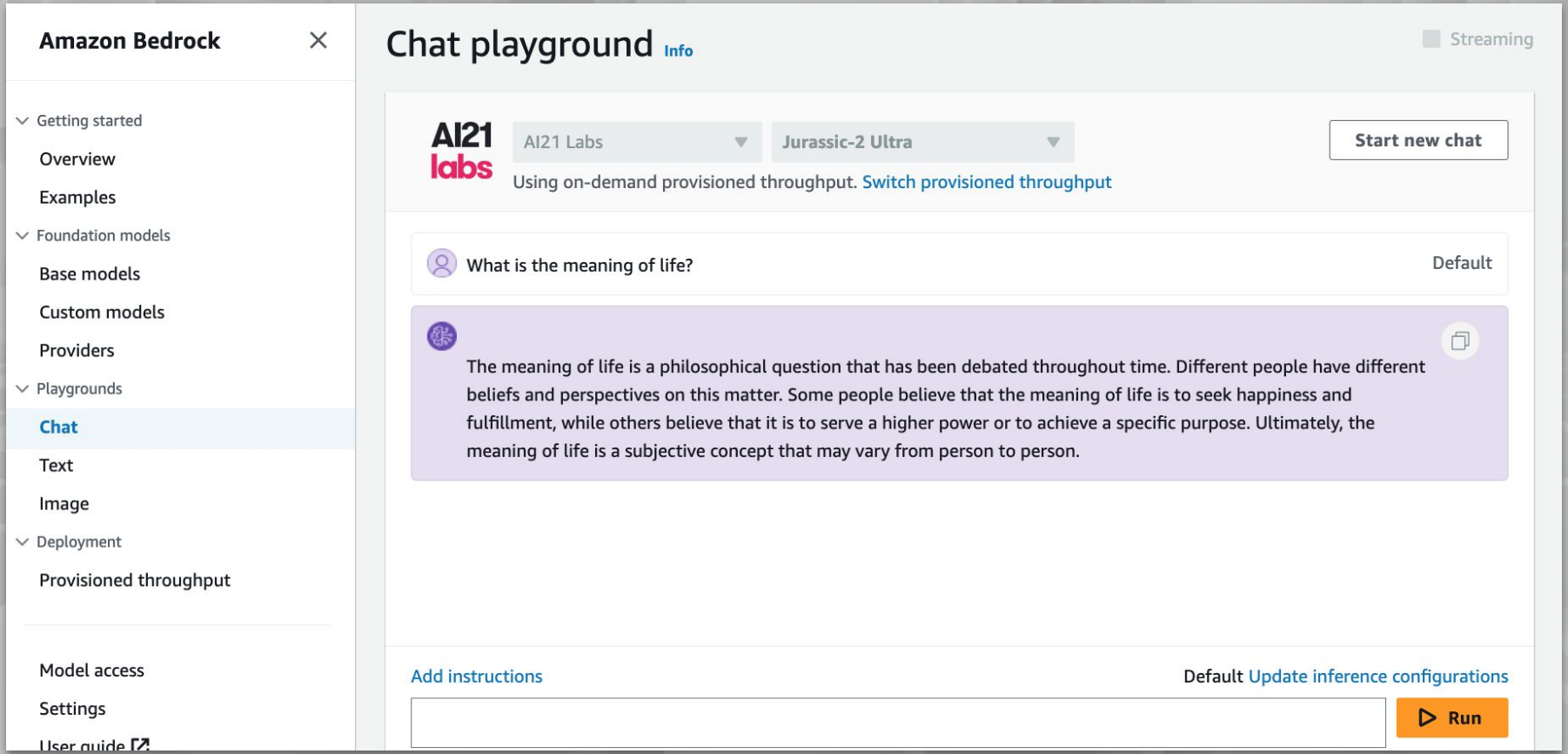
JOSHUA ARVIN LAT

Chief Technology Officer of
NuWorks Interactive Labs


Machine Learning with
Amazon SageMaker
Cookbook
80 proven recipes for data scientists and developers to perform machine learning experiments and deployments
Joshua Arvin Lat


packt
Machine Learning Engineering on AWS
Building, Scaling, and Securing Machine Learning Systems and MLOps Pipelines in Production
JOSHUA ARVIN LAT


packt
Building and Automating Penetration Testing Labs in the Cloud
Set up cost-effective hacking environments for learning cloud security on AWS, Azure, and GCP
JOSHUA ARVIN LAT


Amazon Bedrock
Getting started
Overview
Examples
Foundation models
Base models
Custom models
Providers
Playgrounds
Chat
Text
Image
Deployment
Provisioned throughput
Model access
Settings
User guide
Chat playground Info
Streaming
AI21 labs
AI21 Labs
Jurassic-2 Ultra
Start new chat
Using on-demand provisioned throughput. Switch provisioned throughput
What is the meaning of life?
Default
The meaning of life is a philosophical question that has been debated throughout time. Different people have different beliefs and perspectives on this matter. Some people believe that the meaning of life is to seek happiness and fulfillment, while others believe that it is to serve a higher power or to achieve a specific purpose. Ultimately, the meaning of life is a subjective concept that may vary from person to person.
Add instructions
Default Update inference configurations
Run

Amazon Bedrock

- Getting started
  - Overview
  - Examples
- Foundation models
  - Base models
  - Custom models
  - Providers
- Playgrounds
  - Chat
  - Text
  - Image
- Deployment
  - Provisioned throughput

- Model access
- Settings
- User guide

Image playground Info

Stability AI

Stable Diffusion XL
v0.8 - preview

Using on-demand provisioned throughput. Switch provisioned throughput

Prompt

meaning of life

Response

Inference configuration Info

Reset

Prompt strength 15.9

Generation step 50

Seed 0

Amazon Bedrock

- Getting started
  - Overview
  - Examples
- Foundation models
  - Base models
  - Custom models
  - Providers
- Playgrounds
  - Chat
  - Text
  - Image
- Deployment
  - Provisioned throughput

- Model access
- Settings
- User guide

Stability AI

Stable Diffusion XL
v0.8 - preview

Using on-demand provisioned throughput. Switch provisioned throughput

Prompt

cat flying with wings

Response

Inference configuration Info

Reset

Prompt strength 21.4

Generation step 70

Seed 71681485

Run | Download image | View API request | Save

Amazon Bedrock

- Getting started
  - Overview
  - Examples
- Foundation models
  - Base models
  - Custom models
  - Providers
- Playgrounds
  - Chat
  - Text
  - Image
- Deployment
  - Provisioned throughput

- Model access
- Settings
- User guide

Stability AI

Stable Diffusion XL
v0.8 - preview

Using on-demand provisioned throughput. Switch provisioned throughput

Prompt

dog flying with wings

Response

Inference configuration Info

Reset

Prompt strength 21.4

Generation step 70

Seed 158765648

Run | Download image | View API request | Save

CAN WE BUILD A

GENERATIVE AI-POWERED APPLICATION

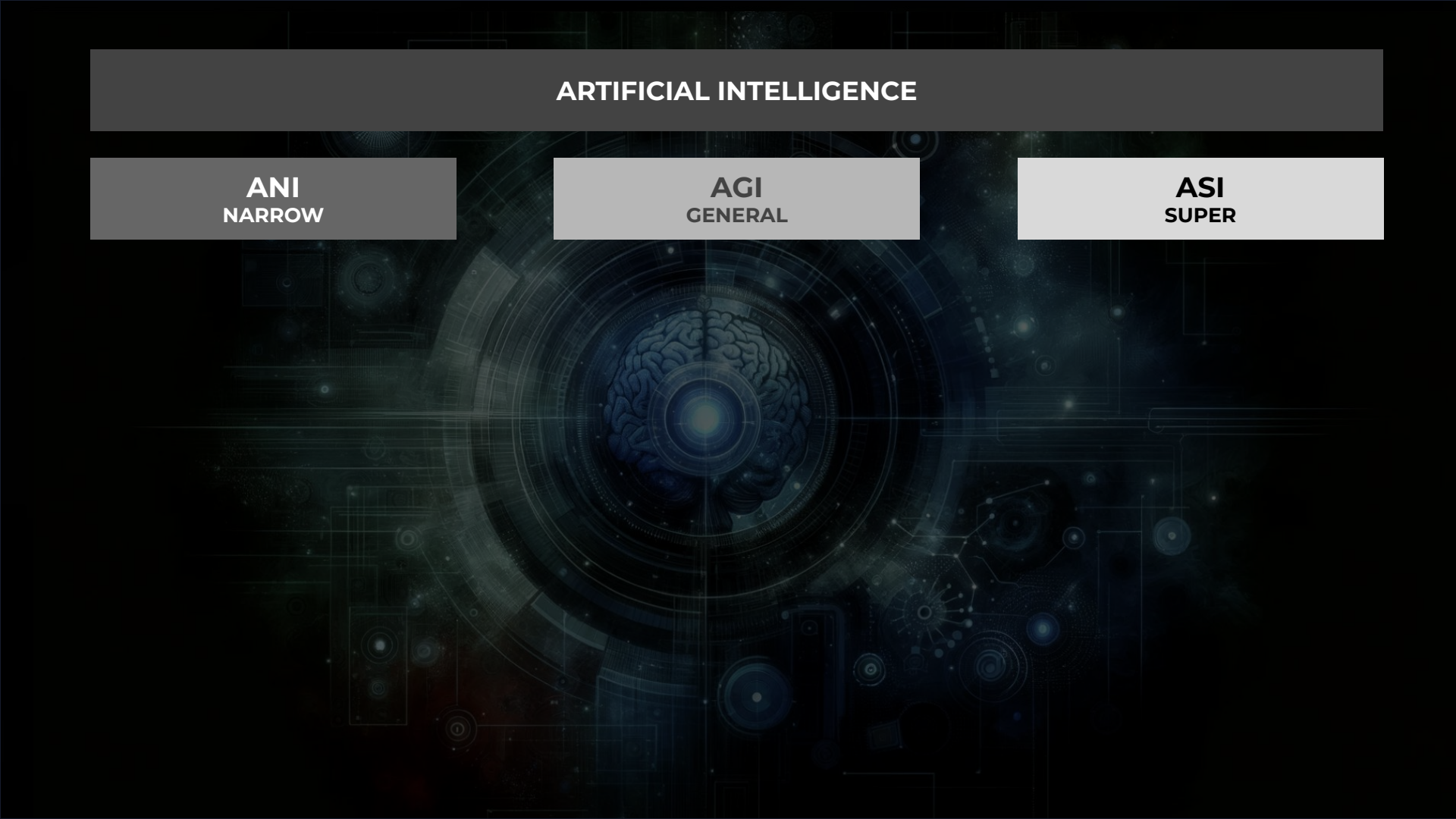
IN 24 HOURS?

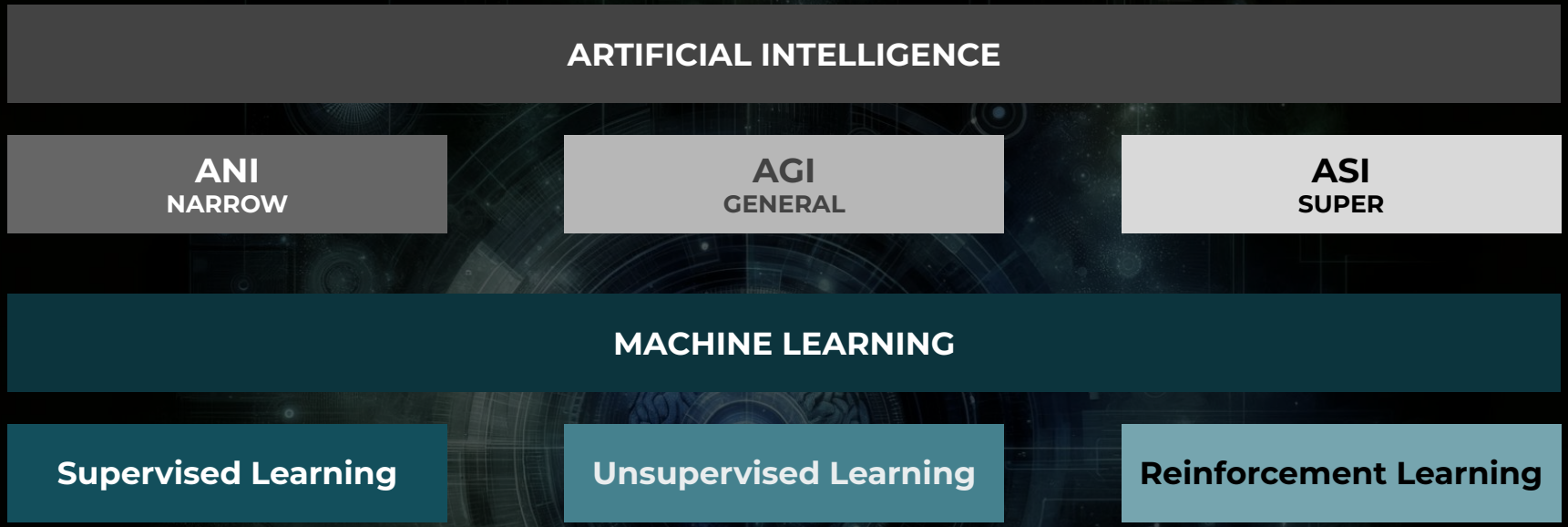
ARTIFICIAL INTELLIGENCE

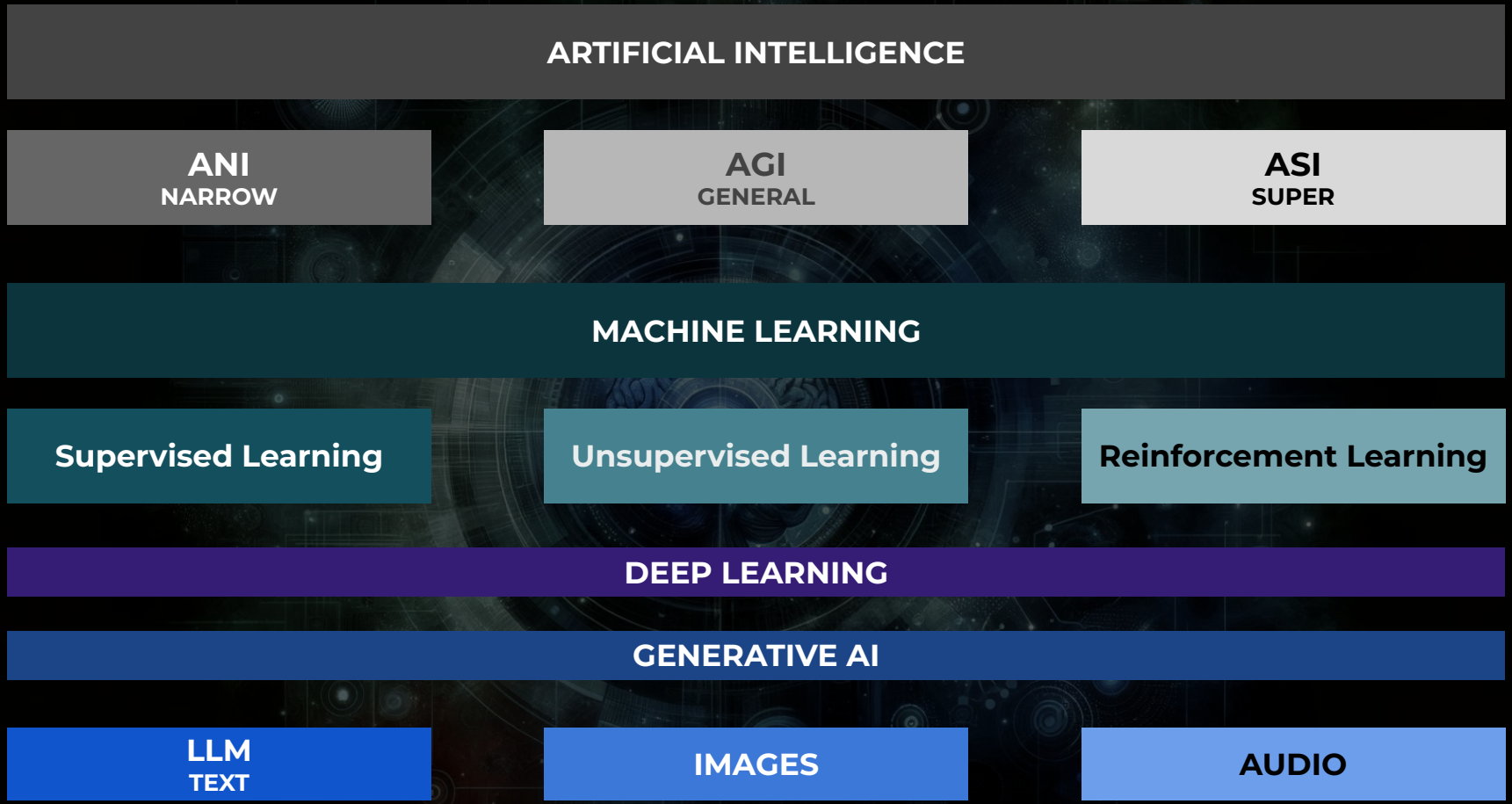
ANI
NARROW

AGI
GENERAL

ASI
SUPER




ARTIFICIAL INTELLIGENCE
ANI
NARROW
AGI
GENERAL
ASI
SUPER
MACHINE LEARNING
Supervised Learning
Unsupervised Learning
Reinforcement Learning


ARTIFICIAL INTELLIGENCE
ANI
NARROW
AGI
GENERAL
ASI
SUPER
MACHINE LEARNING
Supervised Learning
Unsupervised Learning
Reinforcement Learning
DEEP LEARNING
GENERATIVE AI
LLM
TEXT
IMAGES
AUDIO

Limitations of LLMs

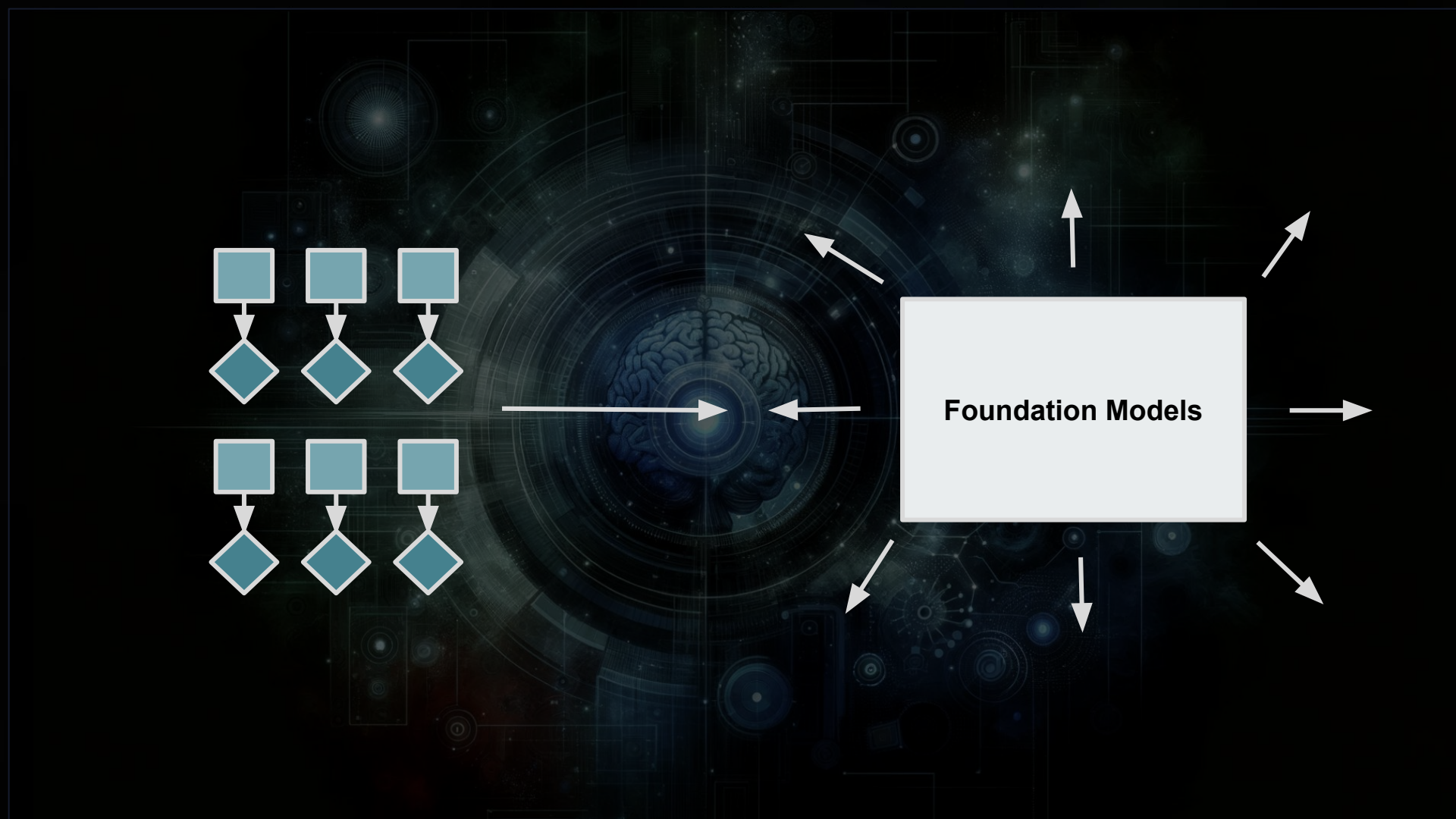
Fairness & Bias

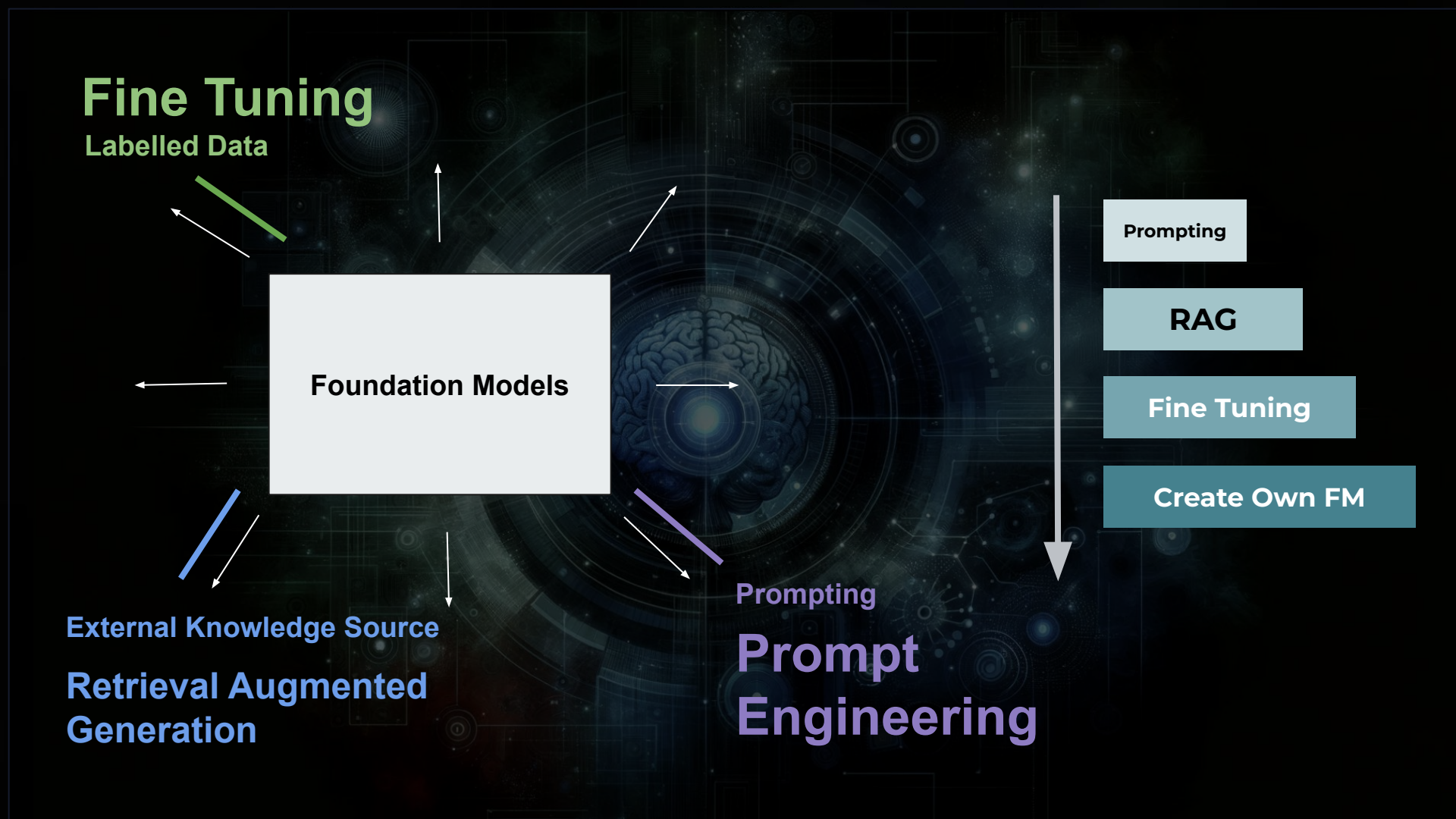
Security and Misuse

Hallucination

Speed and Cost

Interpretability and Explainability


Foundation Models


Fine Tuning
Labelled Data
Foundation Models
External Knowledge Source
Retrieval Augmented Generation
Prompting
Prompt Engineering
Prompting
RAG
Fine Tuning
Create Own FM

RETRIEVAL

AUGMENTATION

GENERATION

IMPROVE CONTENT QUALITY

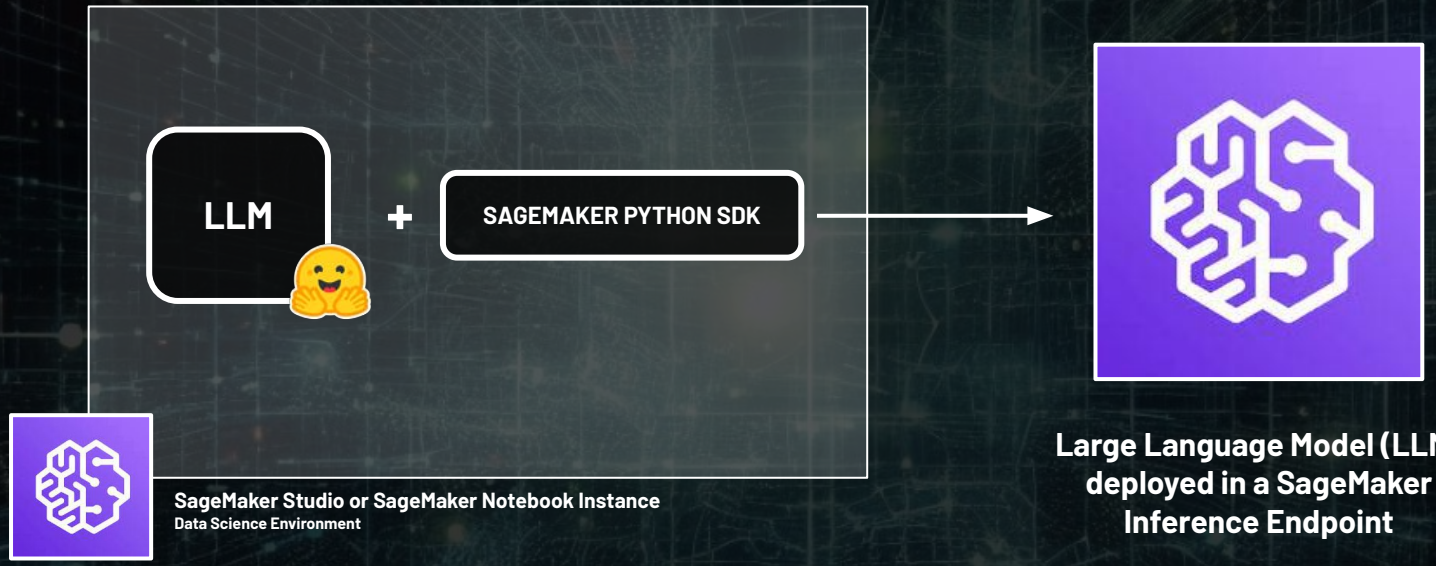
CONTEXT-BASED CHATBOT

PERSONALIZED SEARCH

NAIVE

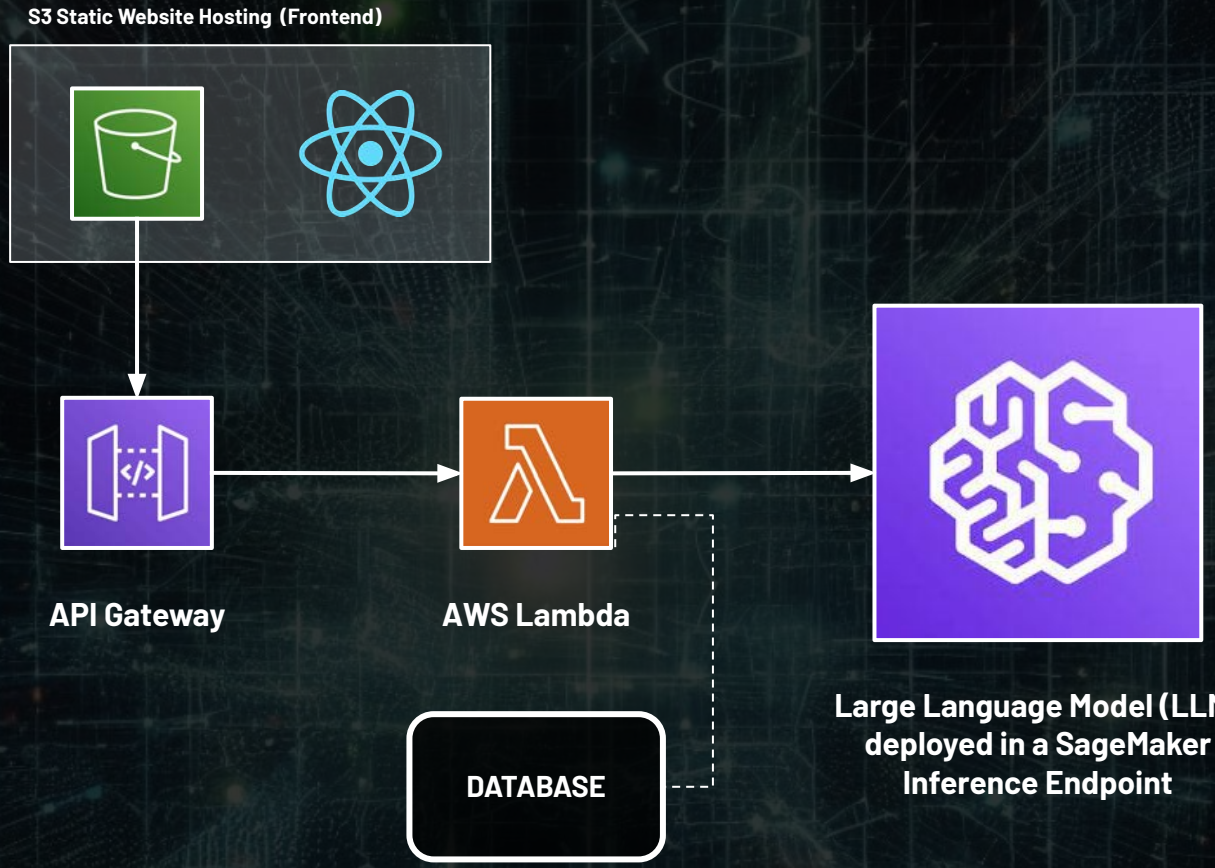
ADVANCED

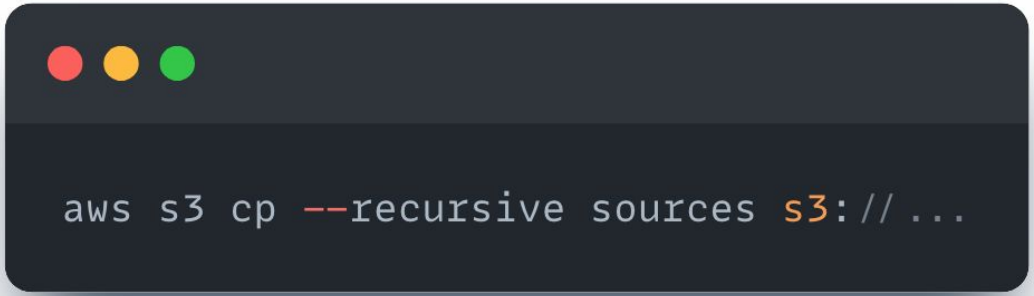
MODULAR


LLM
+
SAGEMAKER PYTHON SDK
SageMaker Studio or SageMaker Notebook Instance
Data Science Environment
Large Language Model (LLM) deployed in a SageMaker Inference Endpoint

SAGEMAKER PYTHON SDK

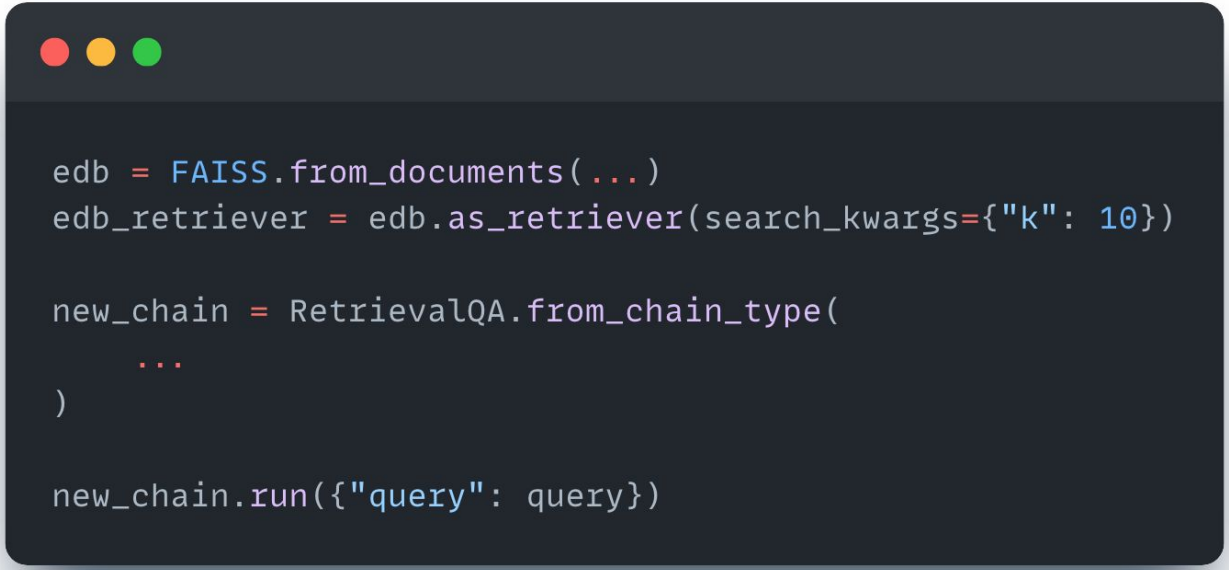
```
predictor = model.deploy(
  initial_instance_count=1,
  instance_type="ml.g5.2xlarge",
  container_startup_health_check_timeout=300,
)

...

answer = chain.run({query})
print(answer)
```


S3 Static Website Hosting (Frontend)
API Gateway
AWS Lambda
DATABASE
Large Language Model (LLM) deployed in a SageMaker Inference Endpoint

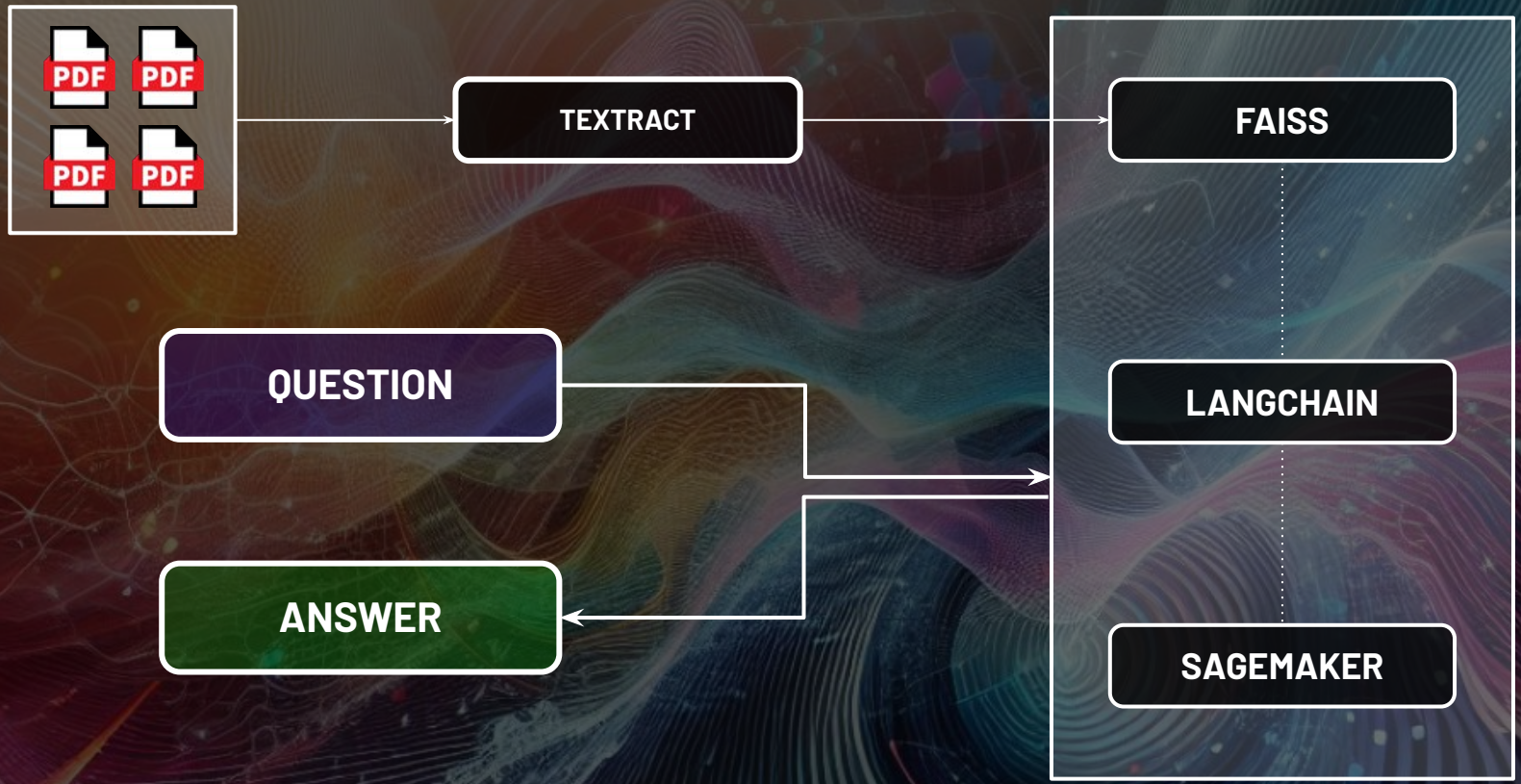

aws s3 cp --recursive sources s3:// ...


PDF
PDF
PDF
PDF


PDF
PDF
PDF
PDF
TEXTRACT
FAISS
LANGCHAIN
SAGEMAKER



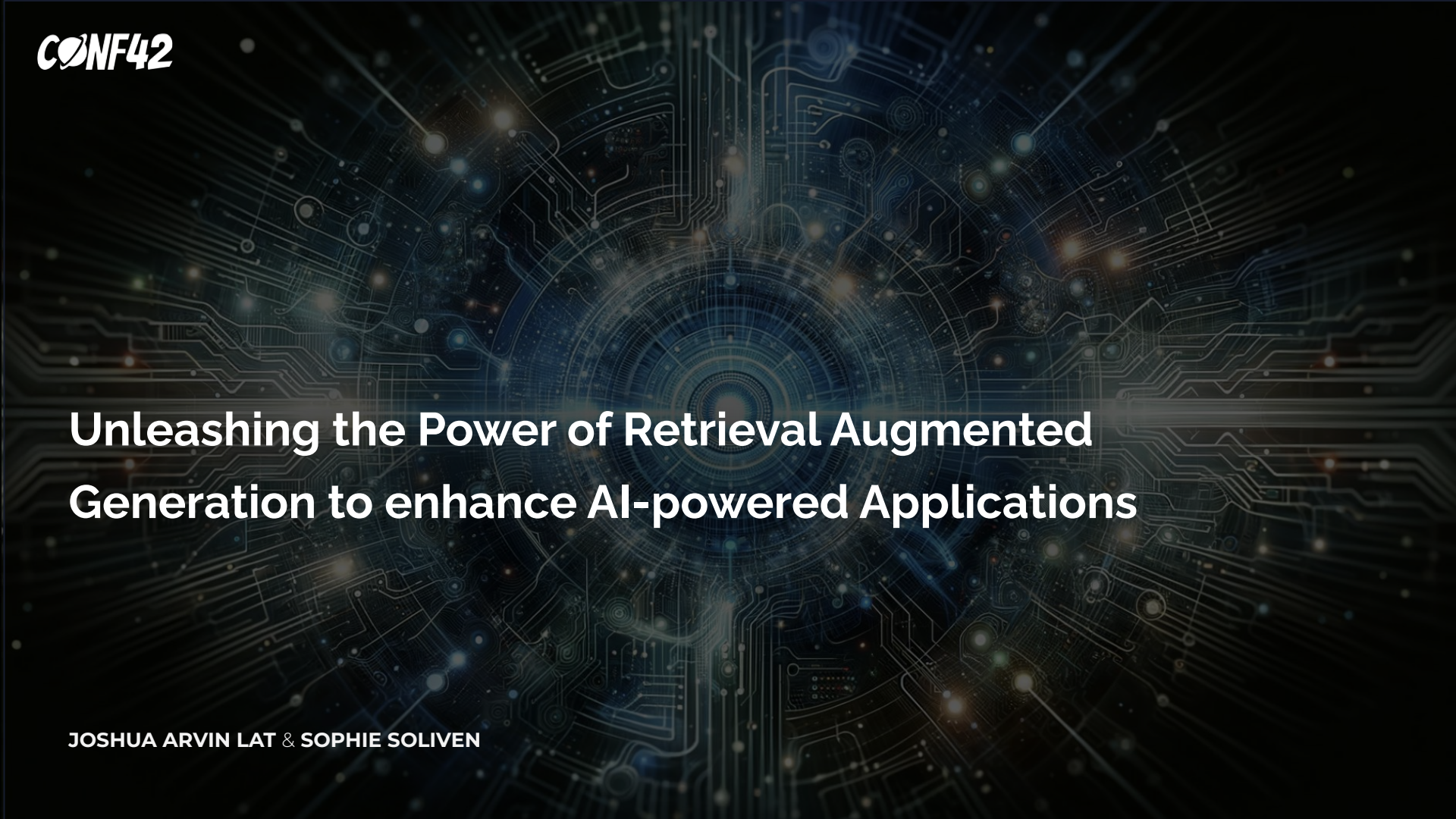
```
edb = FAISS.from_documents(...)
edb_retriever = edb.as_retriever(search_kwargs={"k": 10})

new_chain = RetrievalQA.from_chain_type(
    ...
)

new_chain.run({"query": query})
```


PDF
PDF
PDF
PDF
TEXTRACT
FAISS
QUESTION
LANGCHAIN
ANSWER
SAGEMAKER

CAN WE BUILD A

GENERATIVE AI-POWERED APPLICATION

IN 24 HOURS?



Unleashing the Power of Retrieval Augmented Generation to enhance AI-powered Applications

JOSHUA ARVIN LAT & SOPHIE SOLIVEN