ADC-B2

The architectural patterns of generative AI and your data

Connor Gervin

Lead Architect
KX Partnerships

Davi Baccan

Solutions Architect
Amazon Web Services

Agenda

- Foundation models (FMs)
- Approaches and where to start
- Customise a foundation model
- KX (AWS Partner track)
- Survey

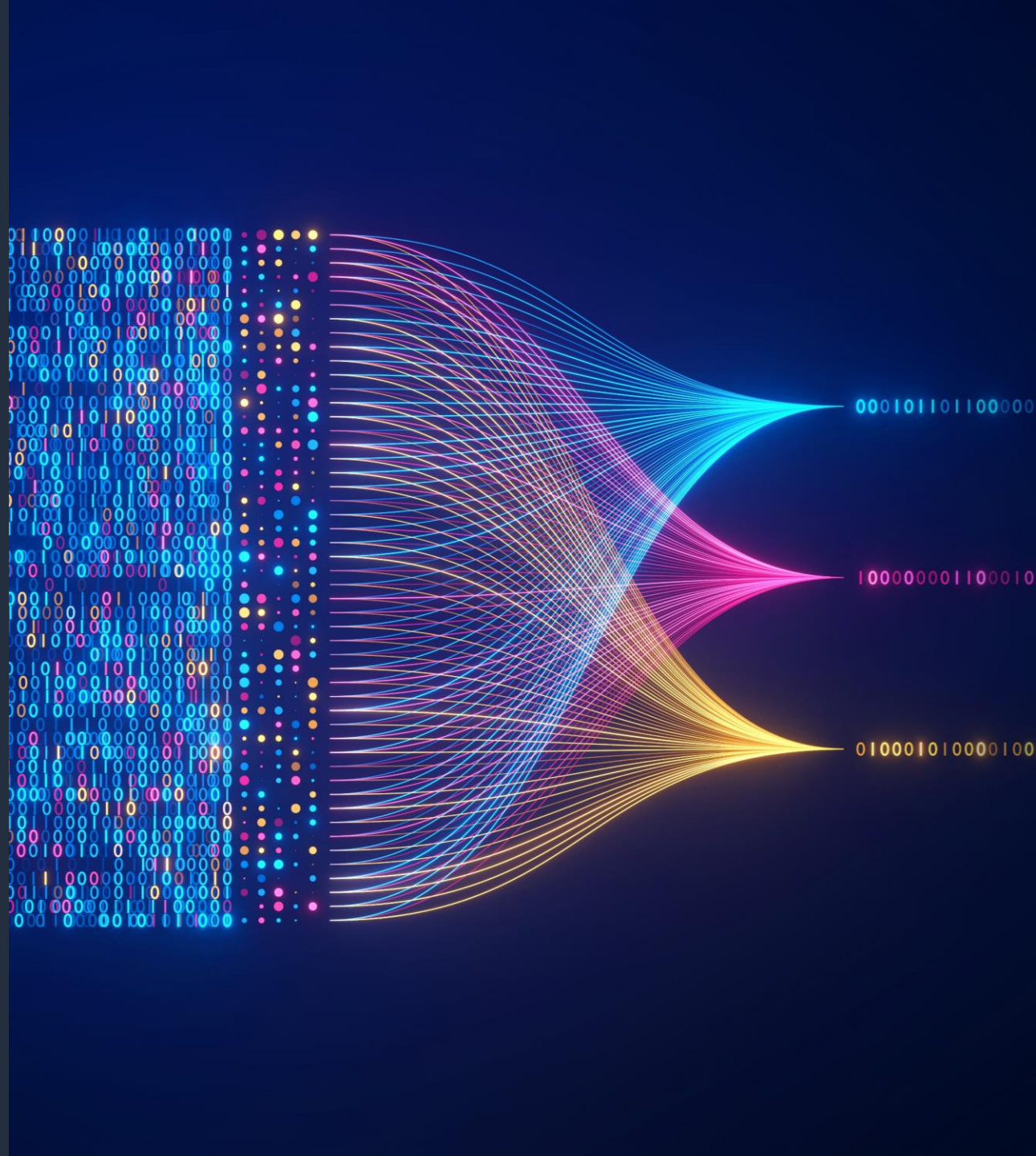
Generative AI is powered by FMs

Pretrained on vast amounts of unstructured data

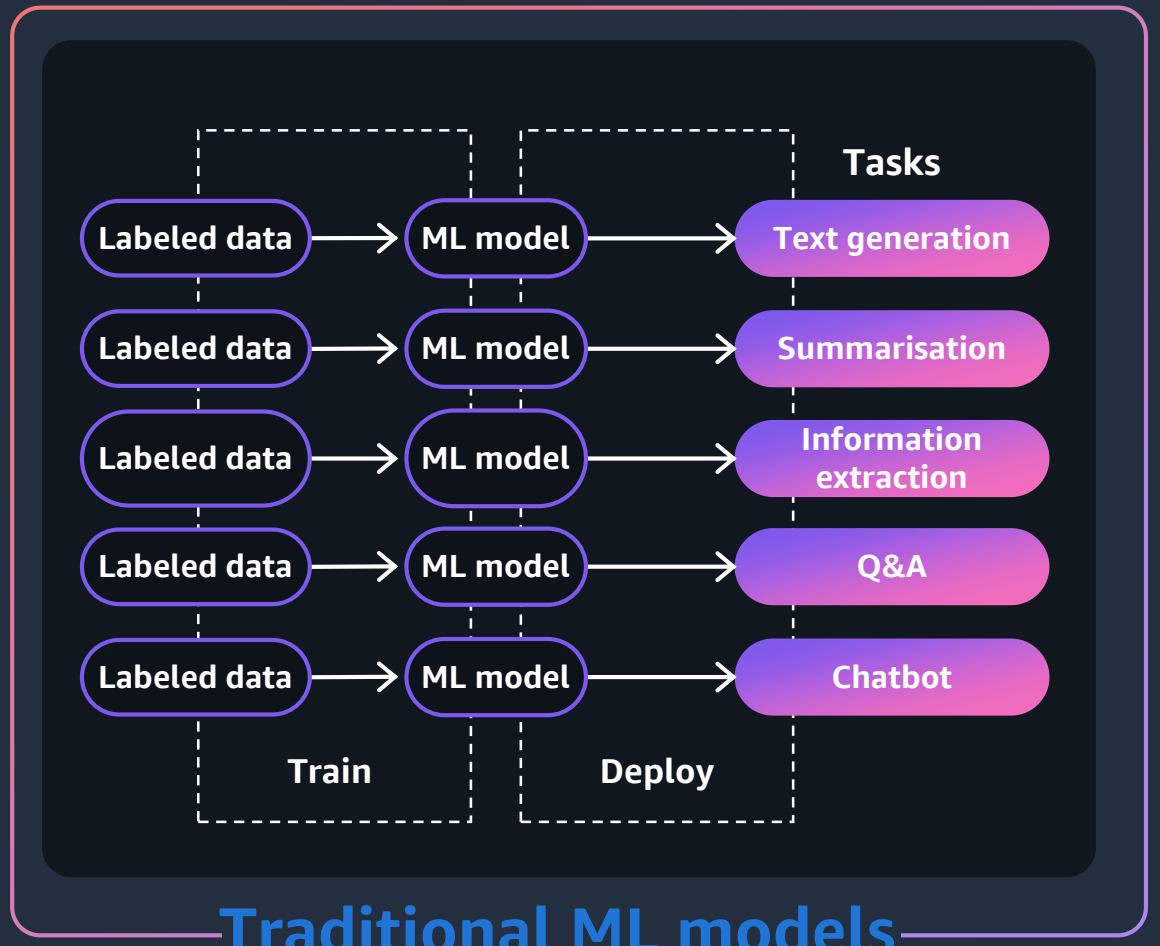
Contain large number of parameters that make them capable of learning complex concepts

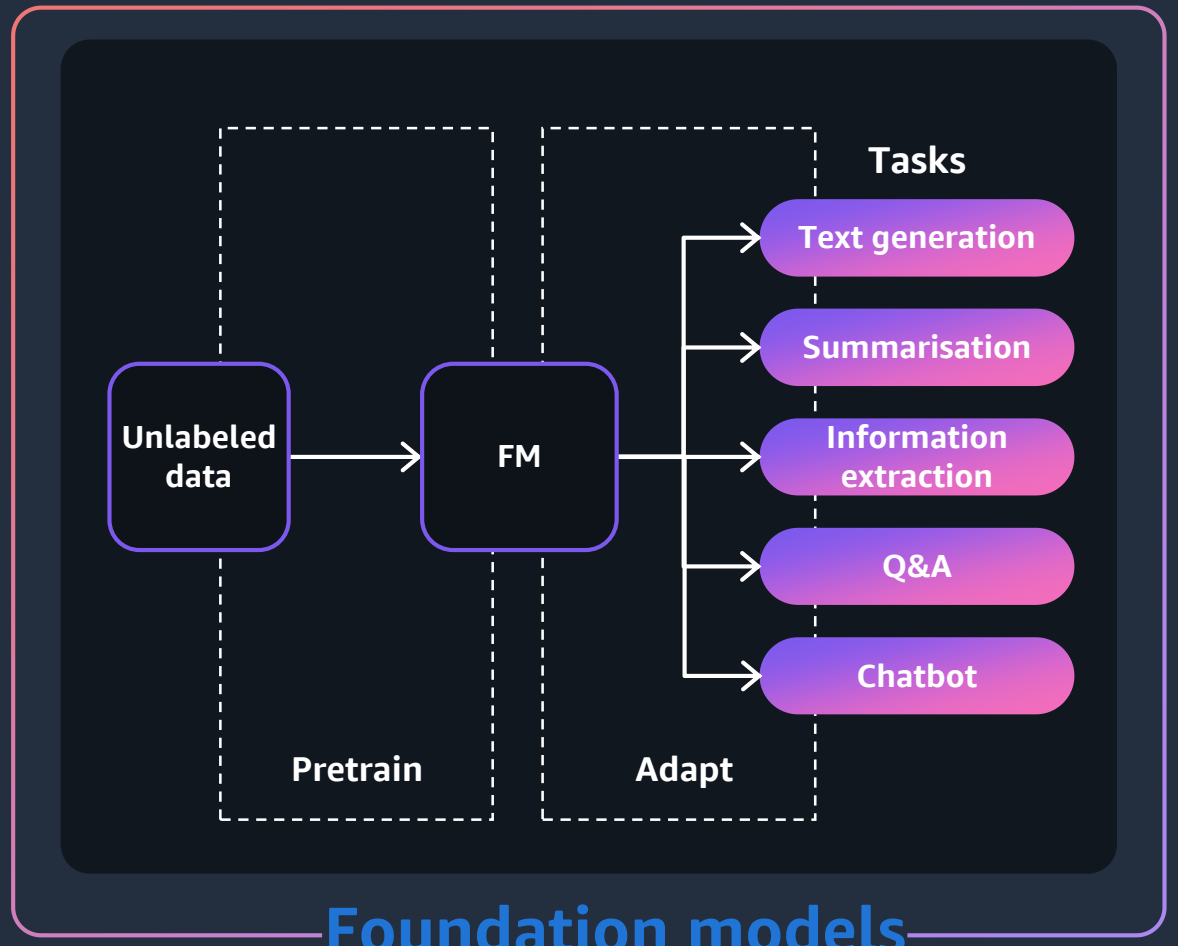
Can be applied in a wide range of contexts

Customise FMs using your data for domain-specific tasks



How FMs differ from other machine learning (ML) models

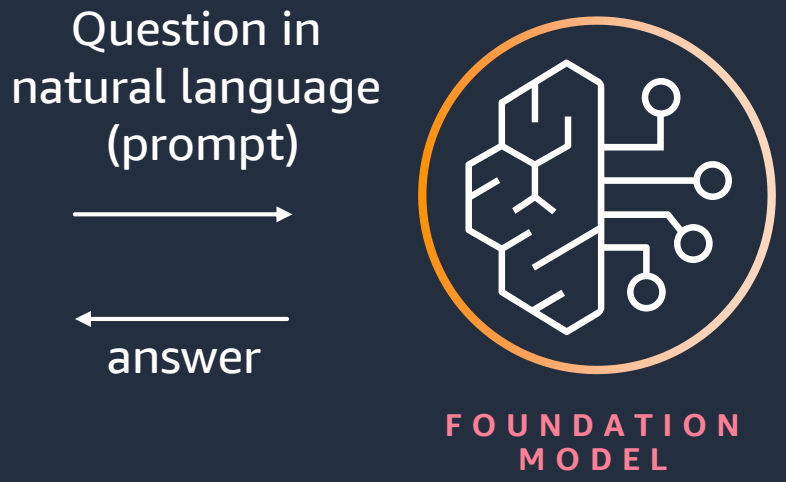




Types of FMs

Input	FM	Output
“summarise the articles on impact of walking on heart health”	**Text-to-text** Generate text from natural language prompts	“Ten thousand steps per day is optimum for maintaining a healthy heart”
“hand soap”	**Text-to-embeddings** Generate numerical representation of text	Numerical representation of “Hand soap refills Hand soap dispenser Hand soap antibacterial”
“a photo of an astronaut riding a horse on mars”	**Multimodal** Create and edit images using natural language prompts	

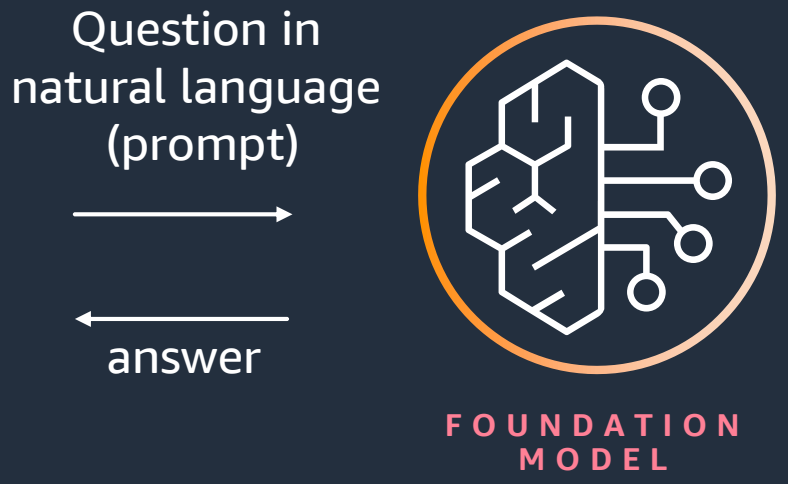
Prompt and answer



Prompt and answer, but …

Prompt	Answer
Who is the best men's soccer player of all time?	There are many great soccer players in the world, and it is difficult to say who is the best of all time. Some people might say that Lionel Messi or Cristiano Ronaldo are the best, while others might argue that players like Diego Maradona or Pele were better. Ultimately, it is up to personal opinion to decide who is the best soccer player of all time.
Where is Lionel Messi playing today?	Lionel Messi is currently playing for Paris Saint-Germain (PSG). He signed with the club in August 2021 after leaving FC Barcelona.
Where is Lionel Messi playing as of September 2023?	As of September 2023, Lionel Messi is playing for Paris Saint-Germain Football Club (PSG). He signed with PSG in August 2021 and his current contract with the club runs until June 2023.

Prompt and answer, but ...



Examples of special challenges posed by generative AI
Toxicity
Hallucinations
Intellectual property
Traceability
Recency of information

Approaches and where to start

- Create your own foundation model (FM)
- Use an existing foundation model

Approaches and where to start

- Create your own foundation model (FM)

Challenges
Complex
Time-consuming process
Potentially (probably) expensive

Key considerations
Do you have the necessary resources (e.g., data and compute)?
How are you going to train your model in light of the special challenges mentioned before (e.g., toxicity)?
Does an existing model fit your needs?
Do you have ML expertise?
How are you going to evaluate if your model is better than the existing models?

- Then… self-manage!

Approaches and where to start

- Use an existing foundation model
  - Self-managed
  - Fully-managed

Self-managed: Amazon SageMaker JumpStart

Choose from more FMs offered by model providers

1

AI21labs

Lighton
We bring Light to AI

stability.ai

co:here

alexa

Try out model or deploy

2

Try out models through the AWS Management Console

Deploy the model for inference using SageMaker hosting options includes single node

Fine-tune model and automate ML workflow

3

Only selected models can be fine-tuned

Automate ML workflow

Data stays in your account including model, instances, logs, model inputs, model outputs

Fully integrated with Amazon features

Fully-managed: Amazon Bedrock

Easily build with FMs

- Choose from a wide range of FMs
- Access FMs through a single API
- Fully managed experience
- Enable generative AI apps to complete tasks with agents

Securely build generative AI apps with your data

- Comprehensive data protection and privacy
- Secure your generative AI applications
- Support for governance and auditability

Deliver customised experiences using your data

- Fine-tune FMs privately
- Supplement organisation-specific information to the FM
- Deliver customised search capabilities for your organisation

Customise a foundation model

- Prompt Engineering
- Fine-tune a Model
- Retrieval Augmented Generation (RAG)

Customise a foundation model

- Prompt Engineering

Techniques	Example
Zero-shot learning	"Classify the following text as either sports, politics, or entertainment: [input text]."
Few-shot learning	"[image 1], [image 2], and [image 3] are examples of [target class]. Classify the following image as [target class]"

Key considerations
Cost effective in comparison with other options
0 to low ML expertise is needed
Fast time to market

Customise a foundation model

- Fine-tune

Might be useful if you need:
To customise your model to specific business needs
Your model to successfully work with domain-specific language, such as industry jargon, technical terms, or other specialised vocabulary
Enhanced performance for specific tasks
Accurate, relative, and context-aware responses in applications
Responses that are more factual, less toxic, and better-aligned to specific requirements

Customise a foundation model

- Fine-tune

Approach	Purpose
Domain adaptation fine-tuning	Adapt FMs to specific tasks using limited domain-specific data. You can use domain adaption fine-tuning to get your model working with domain-specific language, such as industry jargon, technical terms, or other specialised data.
Instruction-based fine-tuning	Uses labelled examples to improve the performance of FM on a specific task. The labelled examples are formatted as prompt, response pairs and phrased as instructions.

Customise a foundation model

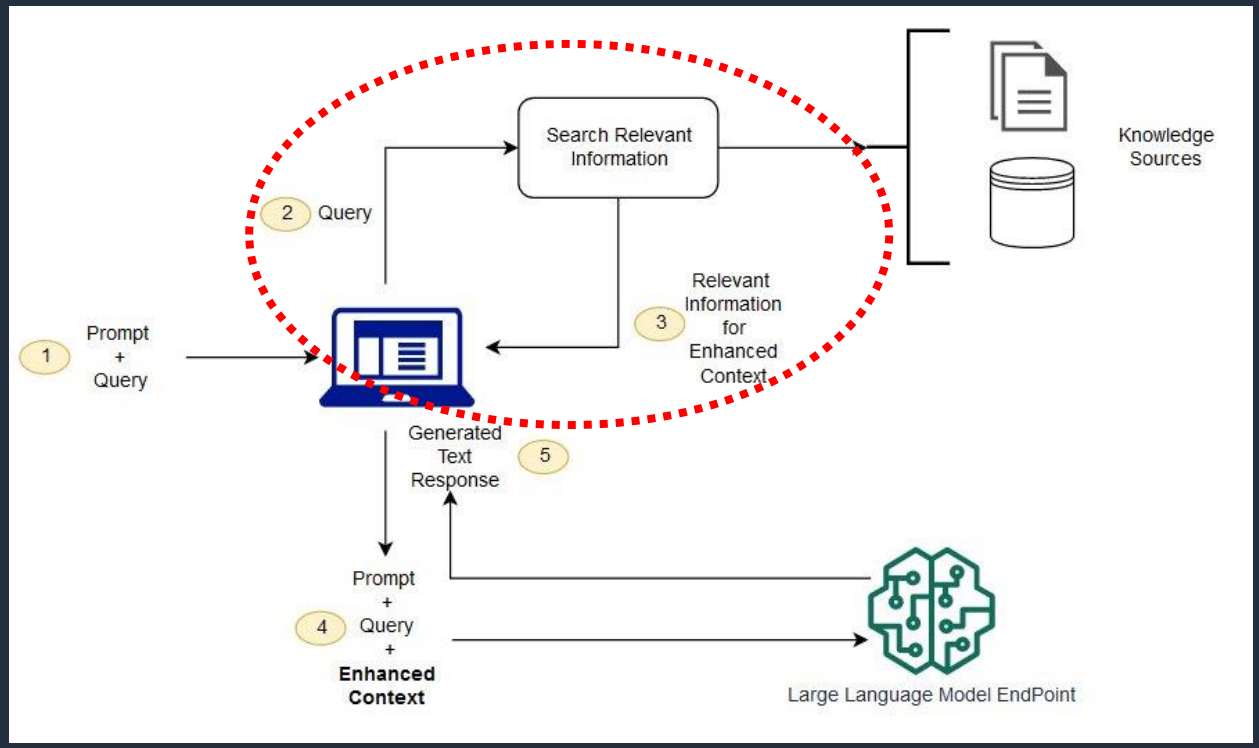
- Retrieval Augmented Generation (RAG)
  - Retrieve data from outside a foundation model and augment your prompts by adding the relevant retrieved data in context

Examples of special challenges posed by generative AI
Toxicity
Hallucinations
Intellectual property
Traceability
Recency of information

Customise a foundation model

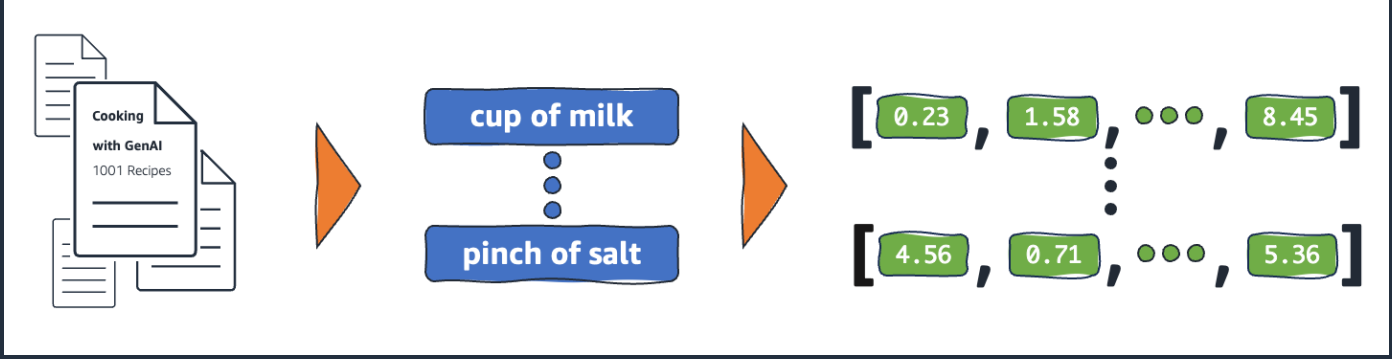
- Retrieval Augmented Generation (RAG)
  - Retrieve data from outside a foundation model and augment your prompts by adding the relevant retrieved data in context

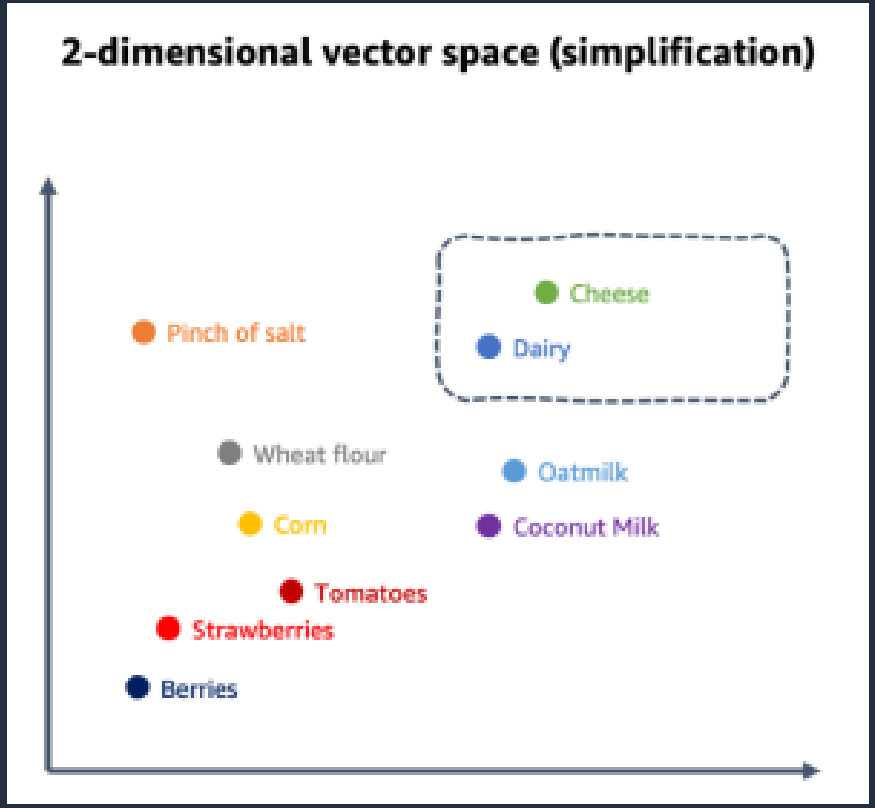
Semantic Search	
Technique	**Components**
Deep learning	Amazon Kendra
Embeddings	Embedding model + Vector datastores



Customise a foundation model

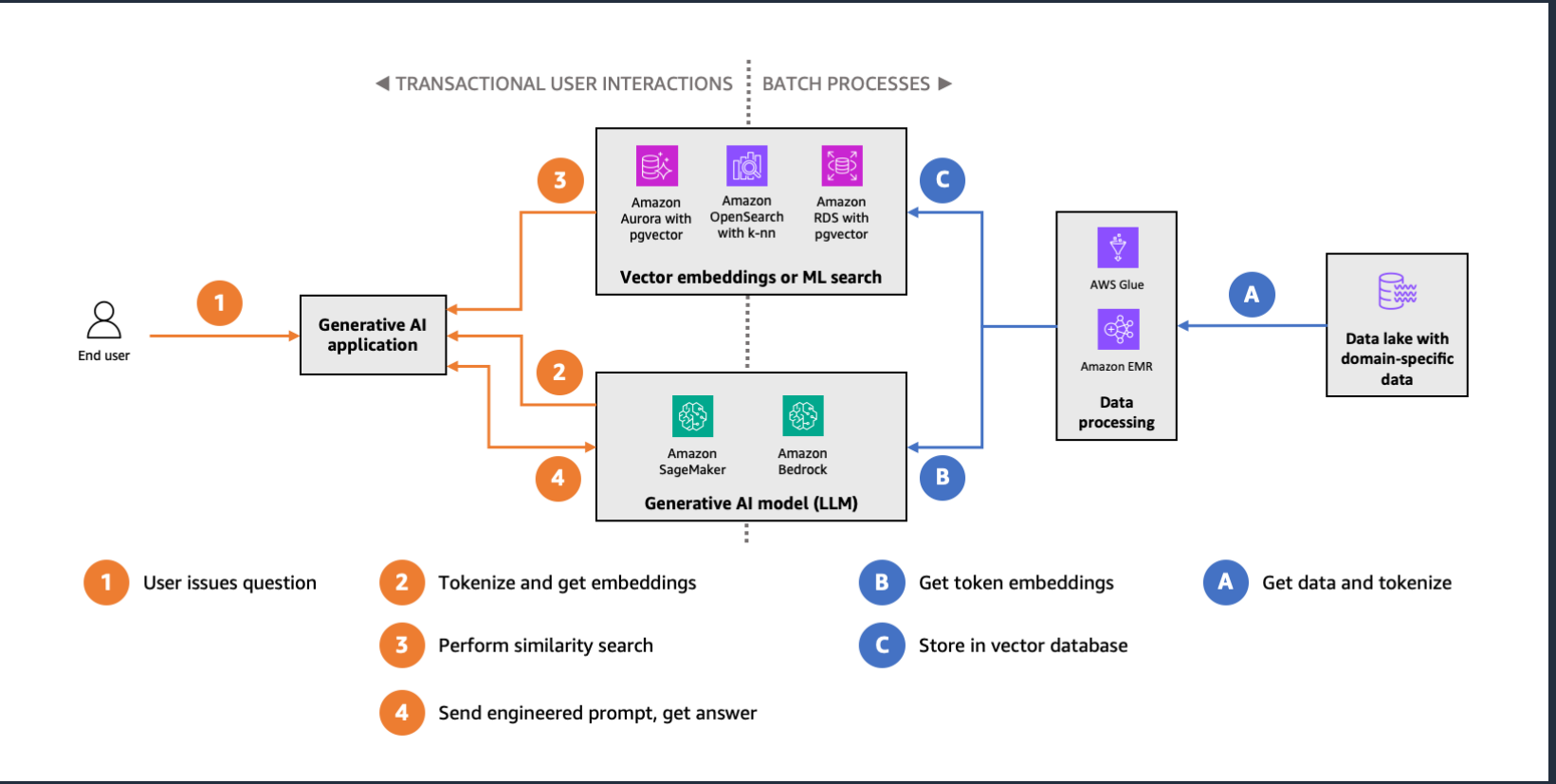
- Embeddings





Customise a foundation model

- Using vector datastores for RAG



Introduction

Each week state-of-the-art LLMs become available, yet keeping LLMs up-to-date is hard.

“I'm unable to provide real-time information or access the internet as I was only trained up to September 2021.”

Frequently retraining or fine-tuning models with new data is slow, expensive and resource intensive.

How can we **bring down costs**, add **real-time context**, introduce **guard rails** for LLMs, while also **improving explainability**?

… Vector Databases!

KX powers actionable decision-making for insights-driven businesses by uniquely enriching real-time data analytics with historical insight.

Who Are KX?

- Proven in capital markets since 1993 – **our client list includes the Top 40 Tier 1 Banks**
- Headquarters in Newry (County Down) with 650 employees globally across 16 Locations

What are KX known for?

- Real-time Streaming and Historical Analysis
- World's Fastest Vector and Time-series Engine
- Powering Wall Street Trading for 30 Years

AWS Partnership

- kdb Insights on Amazon FinSpace
  - Time-series Managed Service (PaaS)
  - Launched June 2023
- KDB.AI Cloud on AWS
  - Vector DB offering by KX (SaaS)
  - Launched Sept 2023

So why the big fuss about Vector Databases?

By leveraging a process called Retrieval Augmented Generation (RAG), vector databases can help keep LLMs accurate while allowing all reasoning to happen in the model.

This has the following benefits:

- We can ground LLM responses in truth and facts

- We can supplement LLMs with real-time, contextual information

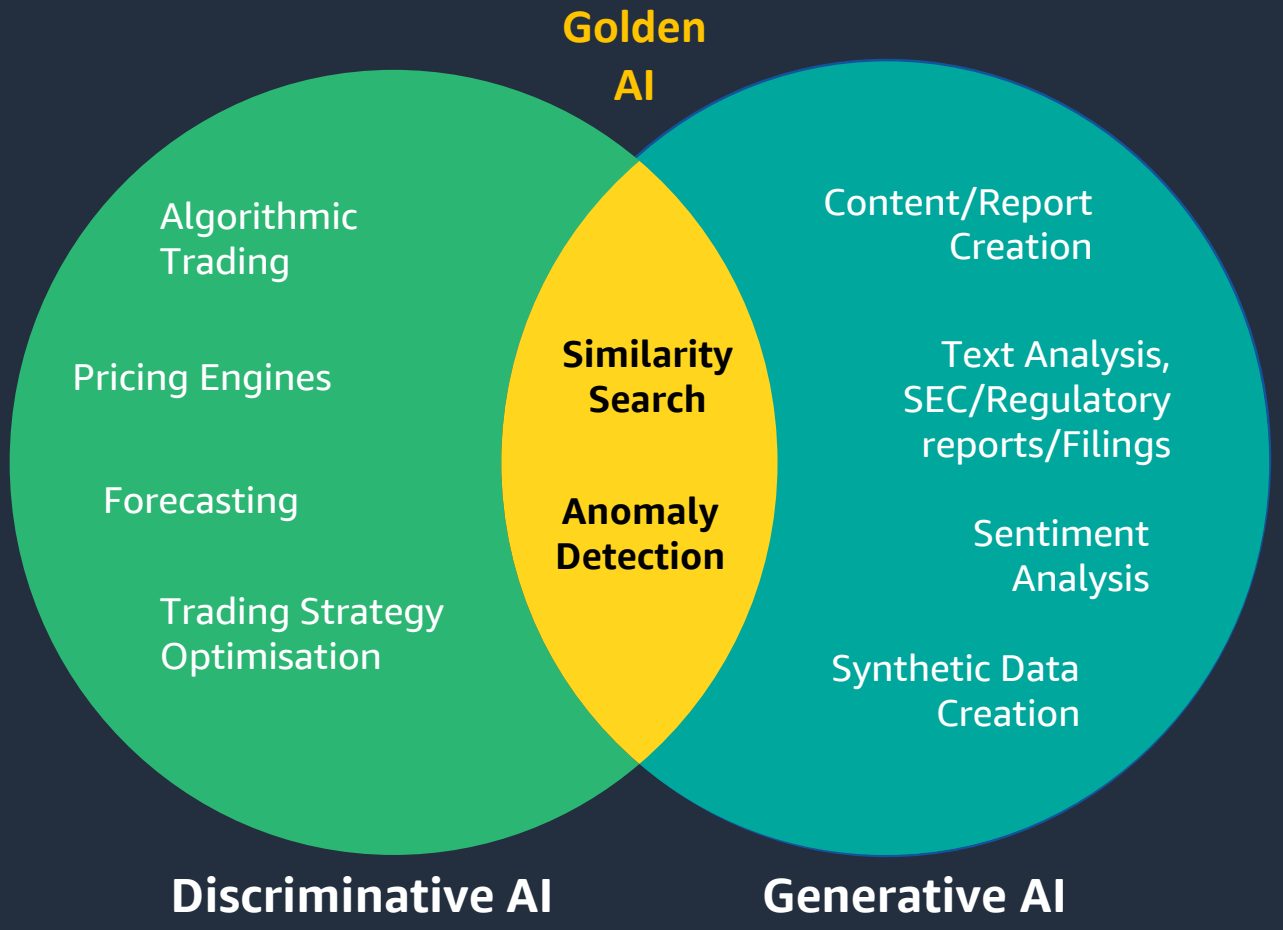
- Facilitate long-term memory and statefulness with LLMs through a native knowledge base

- Improve time-to-market by reducing upfront costs for building responsible AI Apps.

English is the hottest new programming language

Golden AI Fits Between

Discriminative & Generative AI



Golden Use Cases

Pre-Trade

- Accelerate research on industry trends.
- Market predictions on news and social based on sentiment
- Prompt recommendations and support for traders and investors

Trading

- Market-making and matching are heavily mathematical scenarios that many not have affinity for LLMs
- Back-testing and automation of event-driven trading strategies using SEC filings and market data

Clearing and Settlement

- Settlement is a complex multi-party process that is heavily dependent on mathematical scenarios that many not have affinity for LLMs

Custody and Funds

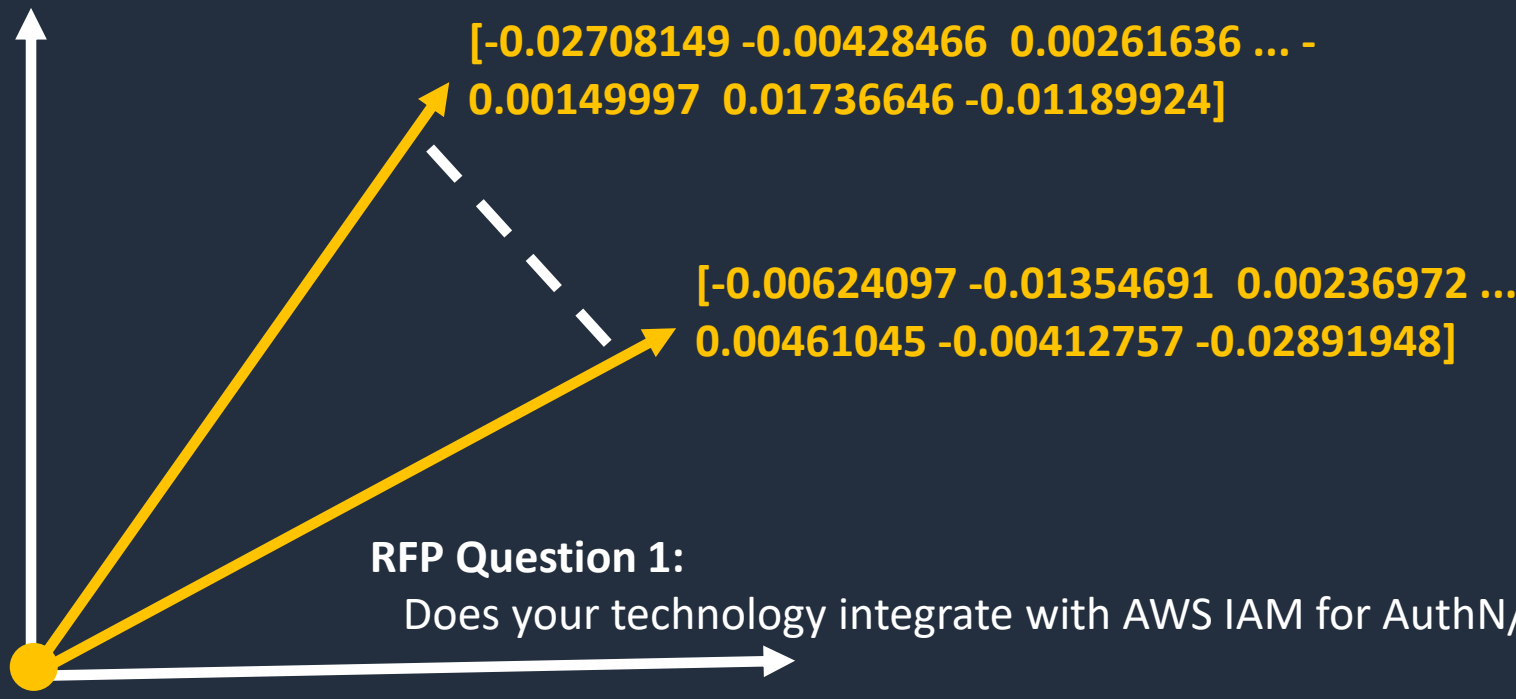
- Monitor transactions to ensure compliance with regulatory requirements.
- Analyse and monitor risk

KDB.AI

Technical Use Cases

Example RFP Q&A Bank

RFP Number	Requirement	Vendor Responses
198	Technology and Infra- > Cloud Management and Integration- >Integration and capability; Integration to external sources (e.g. SaaS application) that are to connect to the end client must be available and secured at point of access and integration	KX will use native AWS security groups and NACL to control point of access and integration. Please see below for the definitions: 1. Security Group – Security Group is a stateful firewall to the instances where the security group keeps a track of the State and operates at the instance level. 2. NACL – NACLs are stateless firewalls which work at Subnet Level, meaning NACLs act like a Firewall to an entire subnet or subnets. A default NACL allows everything both Inbound and Outbound Traffic.
199	Technology and Infra- > Cloud Management and Integration- >Capacity & Scalability; The solution must be able to scale up and horizontal based on required workload and workspace required	KX supports both vertical and horizontal scaling, so as additional compute or storage is required this can be accommodated for. Based on the information provided to date, KX have calculated the expected data volumes for CUSTOMERNAME allowing for annual growth. If the calculations as laid out are incorrect KX would welcome a discussion to revise these to ensure that the system is sized correctly based on any updated information.
200	Technology and Infra- > Cloud Management and Integration- >Security - Data Privacy and Assurance; Production data shall not be replicated or used in non-production environments. Any use of customer data in non-production environments requires explicit, documented approval from all customers whose data is affected, and must comply with all legal and regulatory requirements for scrubbing of sensitive data elements.	Confirmed. Production data will not be replicated or used in non-production environments.
201	Technology and Infra- > Cloud Management and Integration- >Security - Data Privacy and Assurance; Supplier is require to request permission from CUSTOMERNAME if they intend to collect or create metadata about tenant data usage through the use of inspection technologies (search engines, etc.)	Confirmed. KX will require permission from CUSTOMERNAME if there is intention to collect or create metadata about tenant data usage through the use of inspection technologies.
202	Technology and Infra- > Cloud Management and Integration- >Security - Data Privacy and Assurance; Supplier is required to provide a high level of assurance for protection, retention, and lifecycle management of audit logs, adhering to applicable legal, statutory or regulatory compliance obligations (PDPA, Bank Negara) and providing unique user access accountability to detect potentially suspicious network behaviors and/or file integrity anomalies, and to support forensic investigative capabilities in the event of a security breach.	Confirmed. KX will provide a high level of assurance for protection, retention, and lifecycle management of audit logs, adhering to applicable legal, statutory or regulatory compliance obligations (PDPA, Bank Negara) and providing unique user access accountability to detect potentially suspicious network behaviors and/or file integrity anomalies, and to support forensic investigative capabilities in the event of a security breach.



LangChain, KDB.AI and Amazon Bedrock

RAG Response

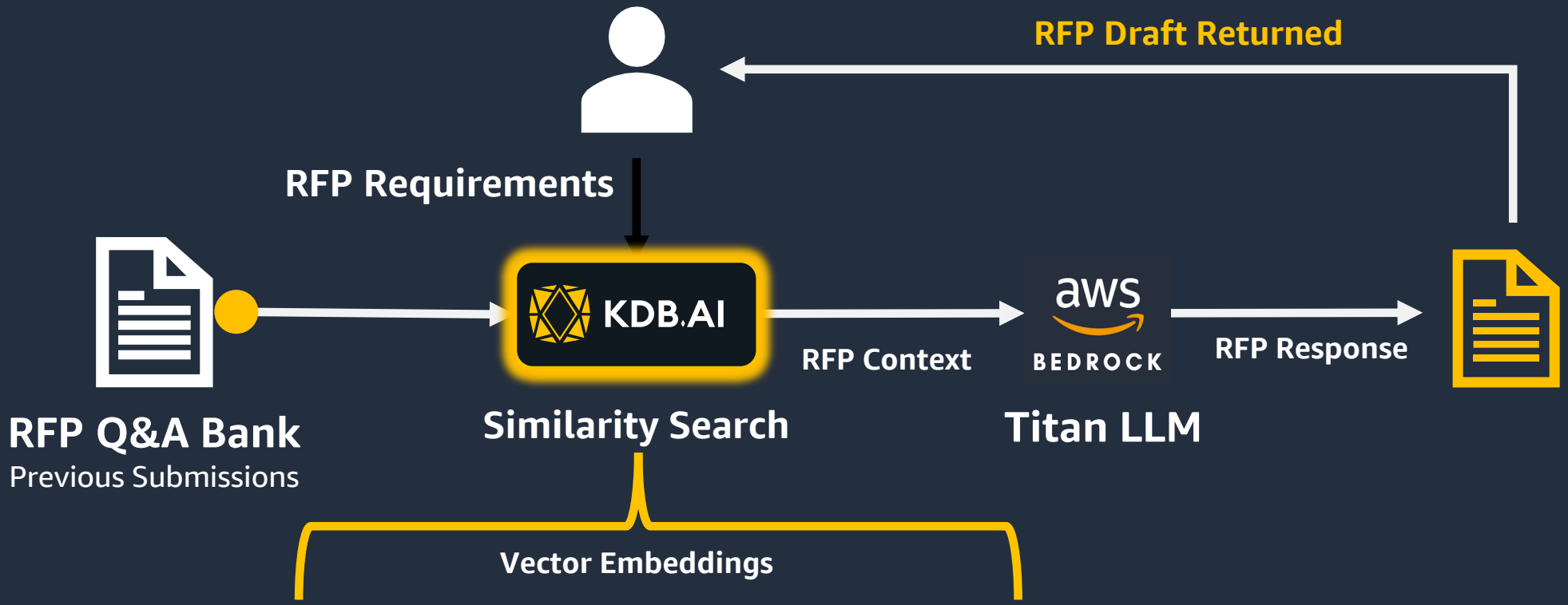
```
# Requirements pulled from an existing RFPs to compare generated responses with real responses
requirements = """
    1. Market Abuse Detection: The solution must process each days trading data, apply alert rules and generate summary reports by market open.
    2. Does your system integrate with computer vision techniques for monitoring suspicious behaviour on the trading floor?
"""

# Giving LLM constructed context and original question to be answered
context = """
    You write responses to RFP requirements provided.
    Use only the below information to write a response to the requirement.
    If you are unsure whether the requirement can be fulfilled, respond with 'Not enough information to answer at this time'.
"""

num_neighbours = 10

generate_RFP_responses_from_Requirements_Amazon_Bedrock(requirements, context, num_neighbours)
```

Using LLMs for RFP Q&A with RAG



The Next Wave

KX have had 150 Customer Conversations over 3 months

Climb the AI value ladder by combining real-time datasets

Customer Support and Document Q&A chatbots were just the beginning

Customers will build their own intellectual property around stateless LLMs. (golden vs non-core use-cases)

Customers will now build AI enabled apps with **Contextual Reasoning**

- Tap into all their existing multi-modal datastores on AWS
- Agents for Bedrock (automation / long running tasks).

Why did my portfolio decline in value?

Generative AI Only

There could be several reasons why your portfolio declined in value: market volatility, company-specific news, currency fluctuations, and interest rate changes.

Generative AI + Enterprise Data

There could be several reasons why your portfolio declined in value: market volatility, company-specific news, currency fluctuations, and interest rate changes. **In your case, you're heavily weighted toward the S&P 500, which declined 6.3% in the last six months; your portfolio declined by just 1.2%.**

That's good news: your portfolio outperformed its comparable index.

Call to action

You have the data, now decide are you a taker, shaper or a maker?

Check out our Free Developer Sandbox on AWS https://kdb.ai/

The #1 Vector Database for 30 Years:

- As independently STAC benchmarked for Tick Analytics in FSI
- As ranked by DB-Engines in the Vector DBMS category