Leveraging RAG architecture for generative AI applications

Jatin Pal Singh

Principal Solutions Architect
Amazon Web Services

Ashutosh Dubey

Sr Solutions Architect
Amazon Web Services

Agenda

Customizing foundation model for your use case

Introduction to retrieval augmented generation (RAG)

Use cases RAG

Demo of RAG pattern using Vector DB and LLM(Amazon Bedrock)

Overview and demo Knowledge bases for Amazon Bedrock

Overview and demo Amazon Q for business

Why customize?

Adapt to domain-specific language

E.g., Healthcare – Understand medical terminology and provide accurate responses related to patient's health

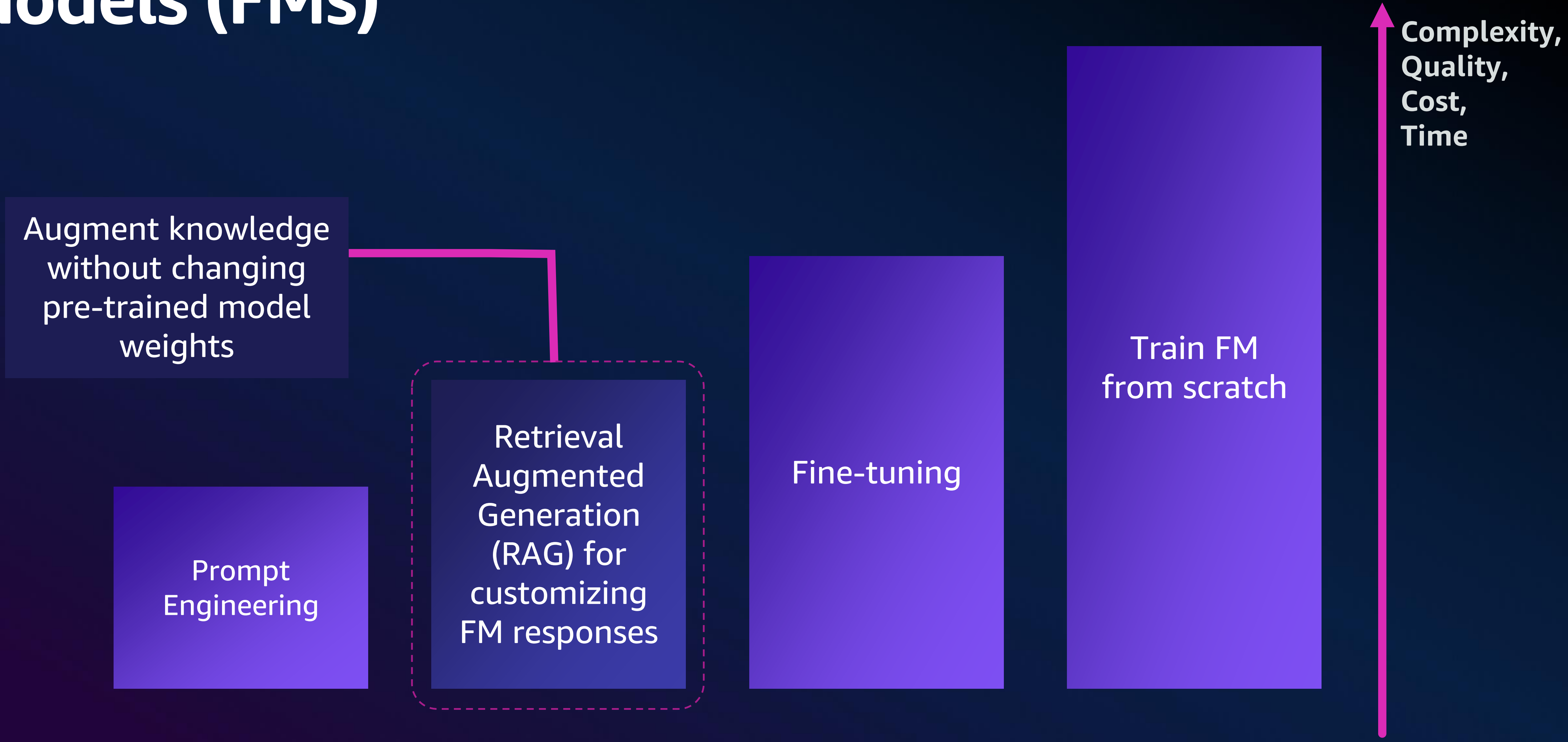
Enhance performance for specific tasks

E.g., Finance – Teach financial & accounting terms to provide good analysis for earnings reports

Improve context-awareness in responses

E.g., Customer Service – Improve ability to understand and respond to customer's inquires and complaints

Common approaches for customizing foundation models (FMs)



Augment knowledge without changing pre-trained model weights

Prompt Engineering

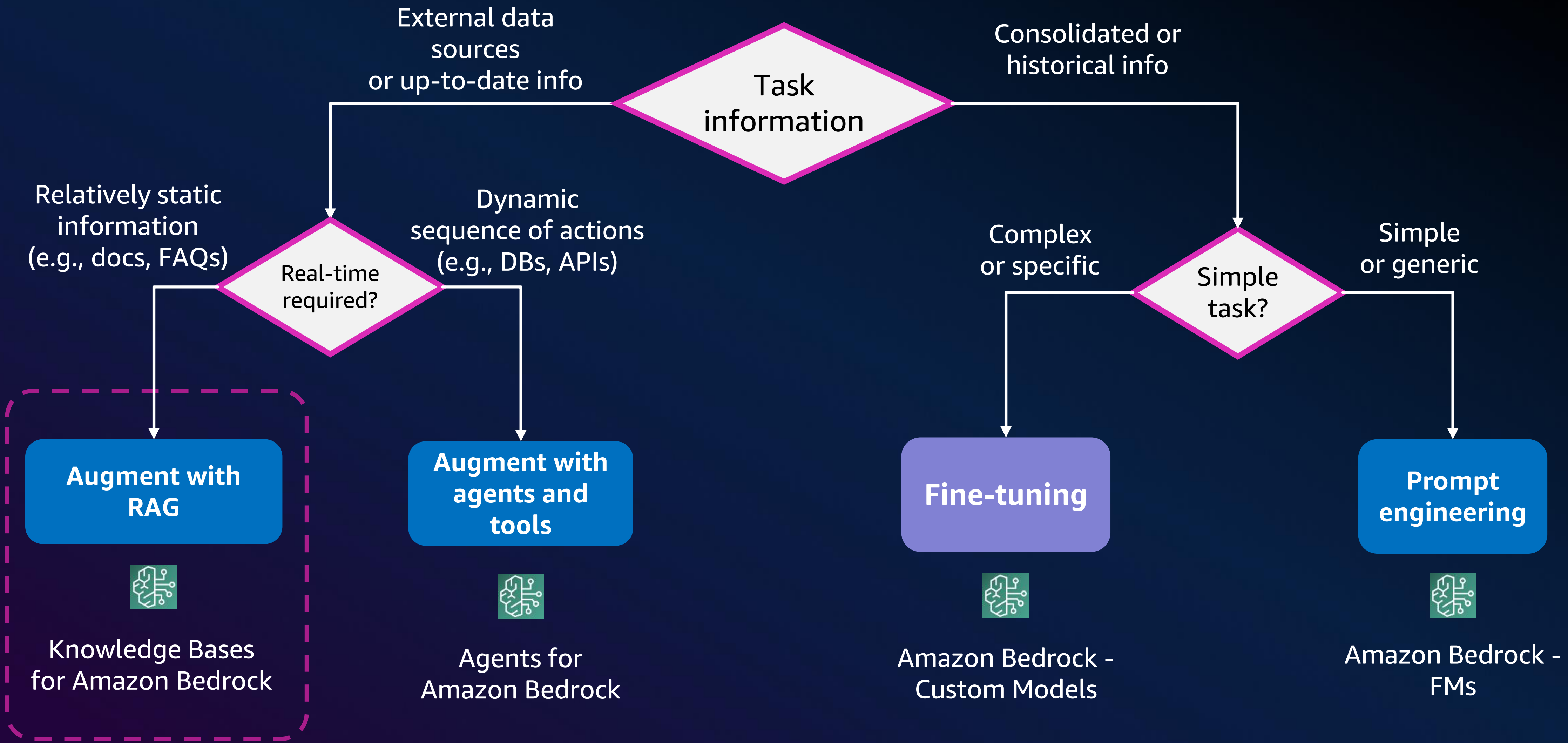
Retrieval Augmented Generation (RAG) for customizing FM responses

Fine-tuning

Train FM from scratch

Complexity, Quality, Cost, Time

Customize vs. augment



External data sources or up-to-date info
Task information
Consolidated or historical info
Relatively static information (e.g., docs, FAQs)
Real-time required?
Dynamic sequence of actions (e.g., DBs, APIs)
Complex or specific
Simple task?
Simple or generic
Augment with RAG
Knowledge Bases for Amazon Bedrock
Augment with agents and tools
Agents for Amazon Bedrock
Fine-tuning
Amazon Bedrock - Custom Models
Prompt engineering
Amazon Bedrock - FMs

What is Retrieval Augmented Generation?

Retrieval

Fetches the relevant content from the external knowledge base or data sources based on a user query

Augmentation

Adding the retrieved relevant context to the user prompt, which goes as an input to the foundation model

Generation

Response from the foundation model based on the augmented prompt

RAG use cases

Improved content quality

E.g., helps in reducing hallucinations and connecting with recent knowledge including enterprise data

Contextual chatbots and question answering

E.g., enhance chatbot capabilities by integrating with real-time data

Personalized search

E.g., searching based on user previous search history and persona

Real-time data summarization

E.g., retrieving and summarizing transactional data from databases, or API calls

Types of retrieval

Rule Based

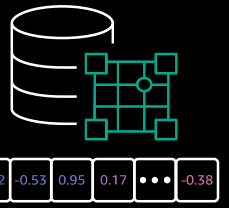
Fetches unstructured data like documents

e.g., Key word searches

Structured data

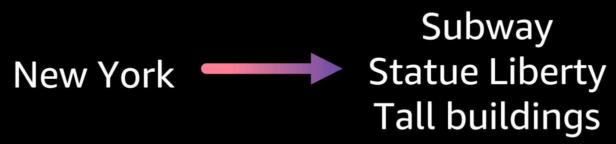
Transactional retrieval from database or API

e.g., Select customers from All_orders where order == ‘XYZ’



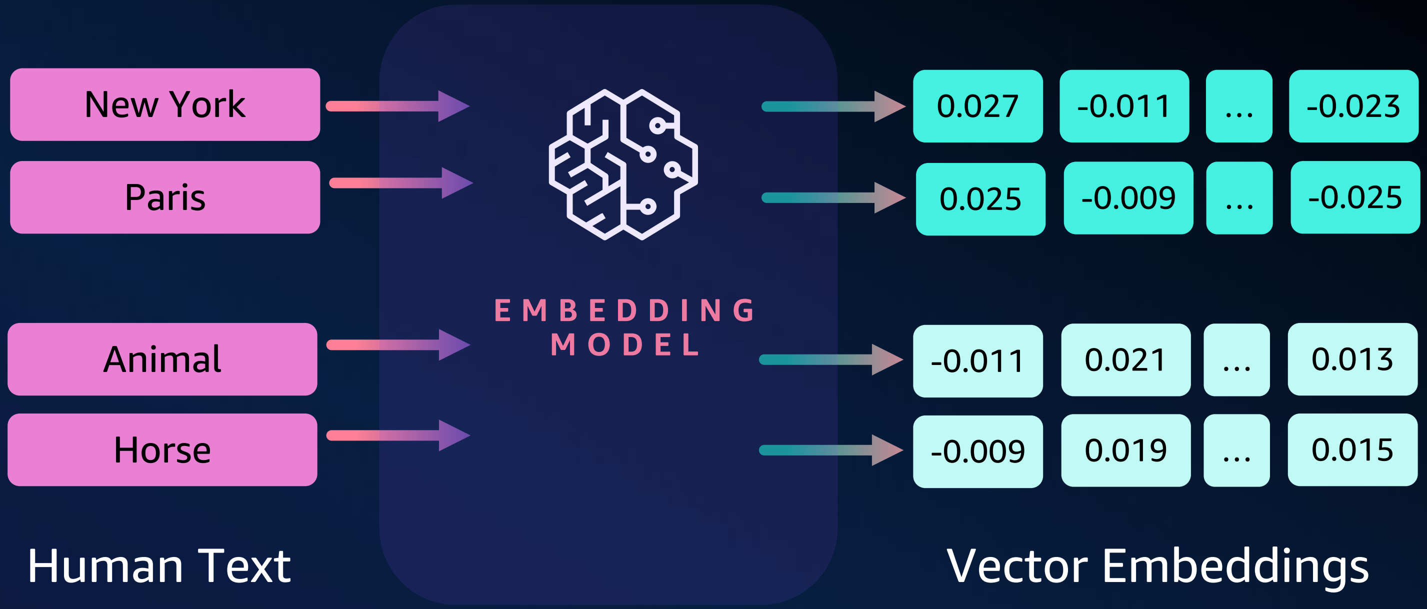
Semantic Search

Get relevant documents based on text embeddings



What are embeddings?

- Numerical representation of text (vectors) that captures semantics and relationships between words.
- Embedding models capture features and nuances of the text.
- Rich embeddings can be used to compare text similarity.
- Multilingual Text Embeddings can identify meaning in different languages.



Why are embeddings important for RAG?

Powers **text retrieval based on semantic meaning.**

Used to augment prompts with **more accurate context** from vector stores using the Retrieval Augmented Generation (RAG).

High-accuracy embeddings leads to improved context **and higher quality LLM-generated** responses to a user query.

Titan text embeddings model

Amazon Titan Text Embeddings

V2.0

Translates text inputs (words, phrases) into numerical representations (embeddings). Comparing embeddings produces more relevant and contextual responses than word matching.

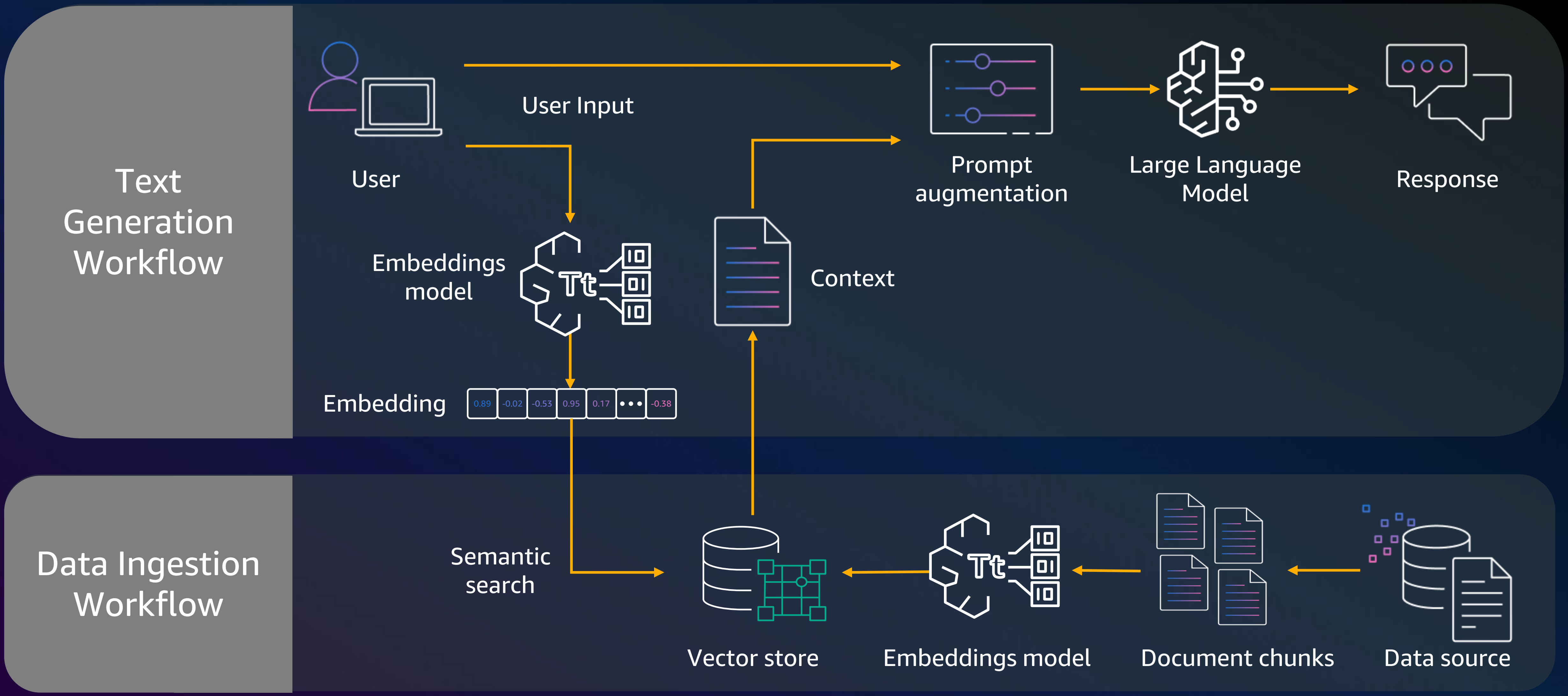
Max Tokens: **8,000**
Output Vectors: **1,536**
Language: **Multilingual** (25 languages)

Model ID: *amazon.titan-embed-g1-text-02*

Highlights

- Titan Text Embeddings offers fast, cost effective, high-performance, accurate embeddings in 25 languages.
- Optimized for text retrieval tasks, semantic similarity and clustering.
- Applications of this model includes semantic search and personalization.

RAG in Action



However, when it comes to implementing RAG, there are challenges…

Managing multiple data sources

Creating vector embeddings for large volumes of data

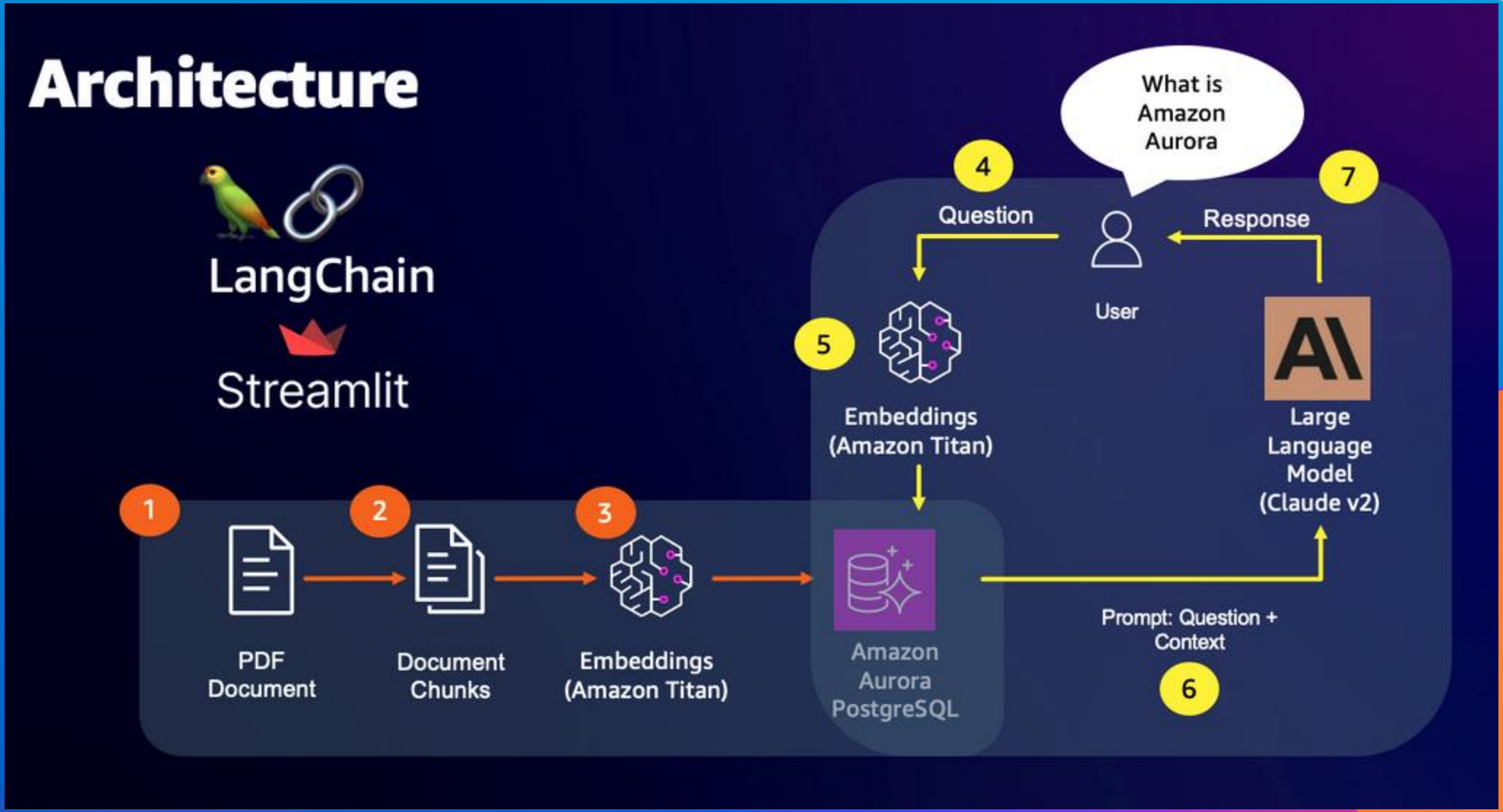
Incremental updates to vector store

Coding effort

Scaling retrieval mechanism

Orchestration

Demo – Vector DB and Bedrock



Knowledge Bases for Amazon Bedrock

Gives FMs and agents contextual information from your private data sources for Retrieval Augmented Generation (RAG) to deliver more relevant, accurate, and customized responses.

Fully managed support for end-to-end RAG workflow

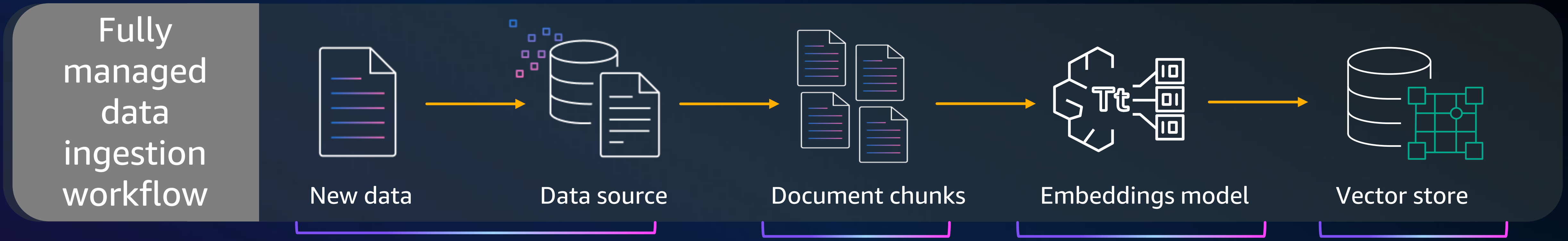
Securely connect FMs and agents to data sources

Easily retrieve relevant data and augment prompts

Provide source attribution

Data Ingestion Workflow

KNOWLEDGE BASES FOR AMAZON BEDROCK



Fully managed data ingestion workflow

New data → Data source → Document chunks → Embeddings model → Vector store

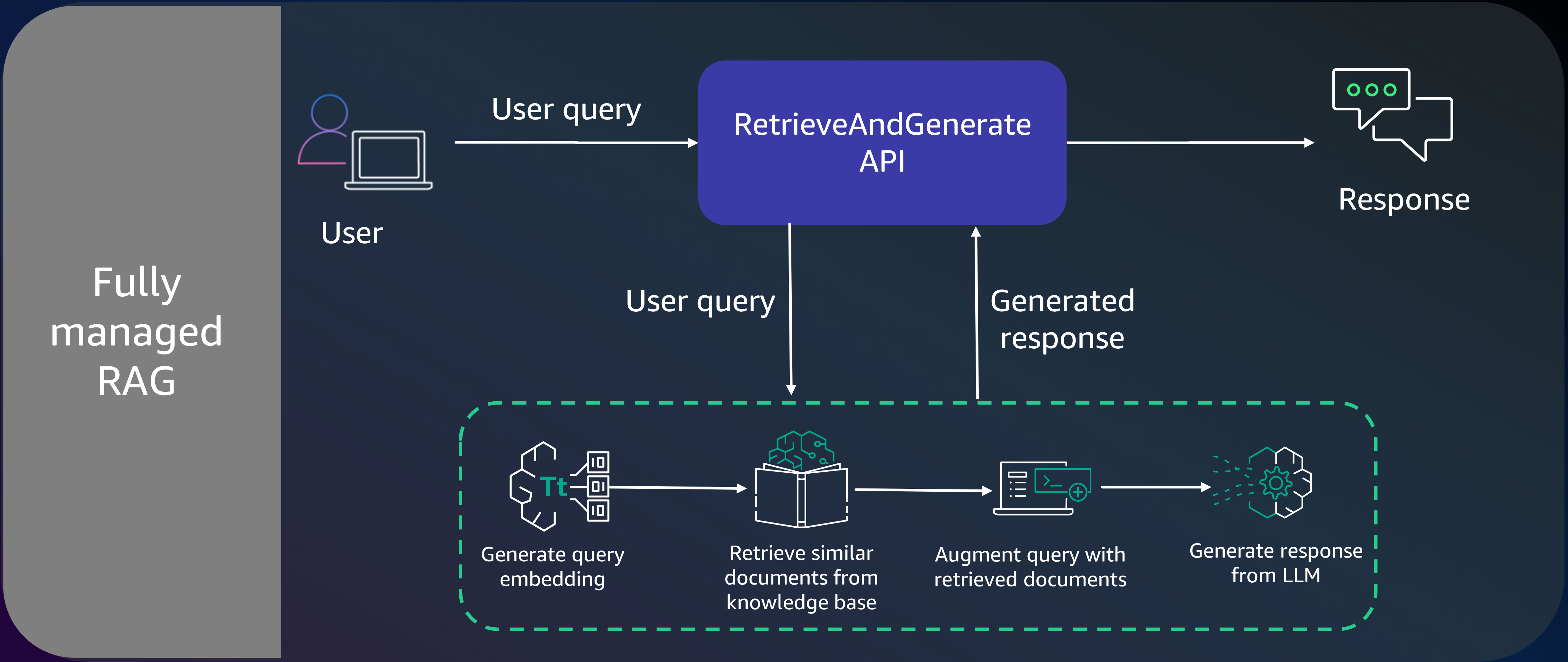
- Choose your data source (Amazon S3)
- Support for incremental updates
- Multiple data file formats supported

- Choose your chunking strategy
  - Fixed chunks
  - No chunking
  - Default (300 tokens)

- Choose your embedding model
  - Amazon Titan
  - Cohere

- Choose your vector store
  - Open search serverless
  - Pinecone
  - Redis
  - Aurora PostgreSQL pgvector

RetrieveAndGenerate API



Fully managed RAG

User

User query

RetrieveAndGenerate API

Response

User query

Generated response

Generate query embedding

Retrieve similar documents from knowledge base

Augment query with retrieved documents

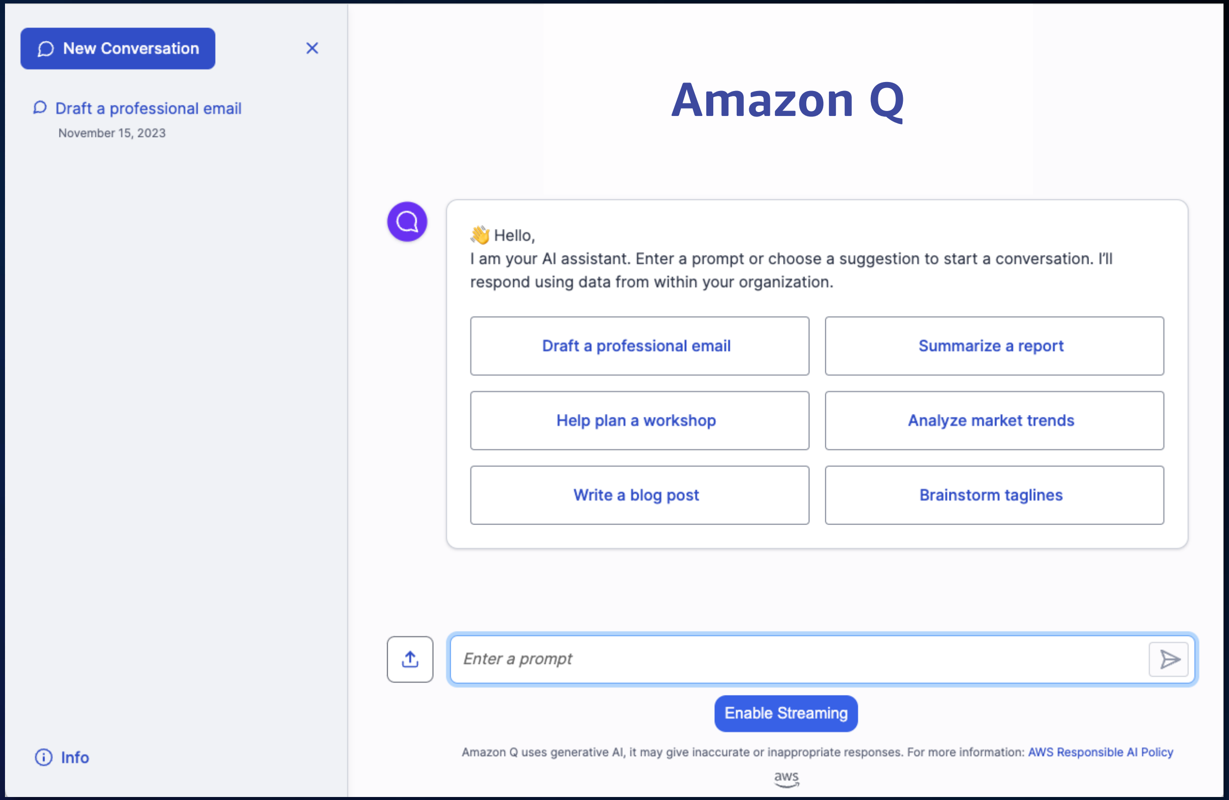
Generate response from LLM

Demo - Knowledge Bases for Amazon Bedrock

Amazon Q is your business expert (Preview)

BOOST YOUR WORKFORCE PRODUCTIVITY WITH GENERATIVE AI

- Delivers quick, accurate, and relevant answers to your business questions, securely and privately and document repositories.
- Provides responses with references and citations for easy fact-checking
- Respects existing access control based on user permissions
- Connects to over 40 popular enterprise applications and document repositories
- Enables administrators to easily apply guardrails to customize and control responses



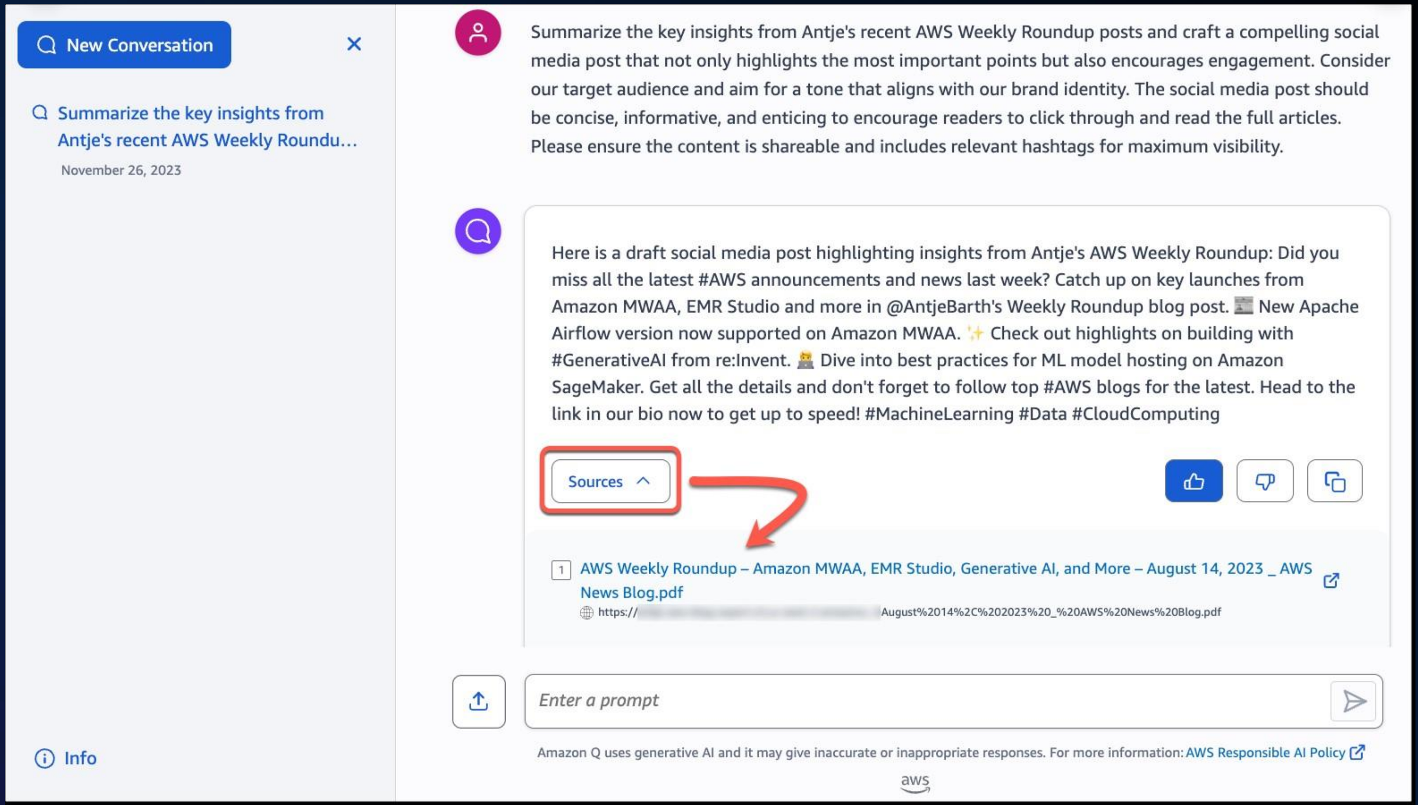
👋 Hello,
I am your AI assistant. Enter a prompt or choose a suggestion to start a conversation. I'll respond using data from within your organization.

Amazon Q uses generative AI, it may give inaccurate or inappropriate responses. For more information: AWS Responsible AI Policy

Key Features

BOOSTING EMPLOYEE PRODUCTIVITY

- Trusted answers generated from enterprise data
- Citations and source attribution
- Conversation history and context
- Upload files and analyze content
- Execute Actions across multiple Enterprise Apps



Summarize the key insights from Antje's recent AWS Weekly Roundu…

November 26, 2023

Summarize the key insights from Antje's recent AWS Weekly Roundup posts and craft a compelling social media post that not only highlights the most important points but also encourages engagement. Consider our target audience and aim for a tone that aligns with our brand identity. The social media post should be concise, informative, and enticing to encourage readers to click through and read the full articles. Please ensure the content is shareable and includes relevant hashtags for maximum visibility.

Here is a draft social media post highlighting insights from Antje's AWS Weekly Roundup: Did you miss all the latest #AWS announcements and news last week? Catch up on key launches from Amazon MWAA, EMR Studio and more in @AntjeBarth's Weekly Roundup blog post. 📰 New Apache Airflow version now supported on Amazon MWAA. ✨ Check out highlights on building with #GenerativeAI from re:Invent. 🧑‍💻 Dive into best practices for ML model hosting on Amazon SageMaker. Get all the details and don't forget to follow top #AWS blogs for the latest. Head to the link in our bio now to get up to speed! #MachineLearning #Data #CloudComputing

1 AWS Weekly Roundup – Amazon MWAA, EMR Studio, Generative AI, and More – August 14, 2023 _ AWS News Blog.pdf

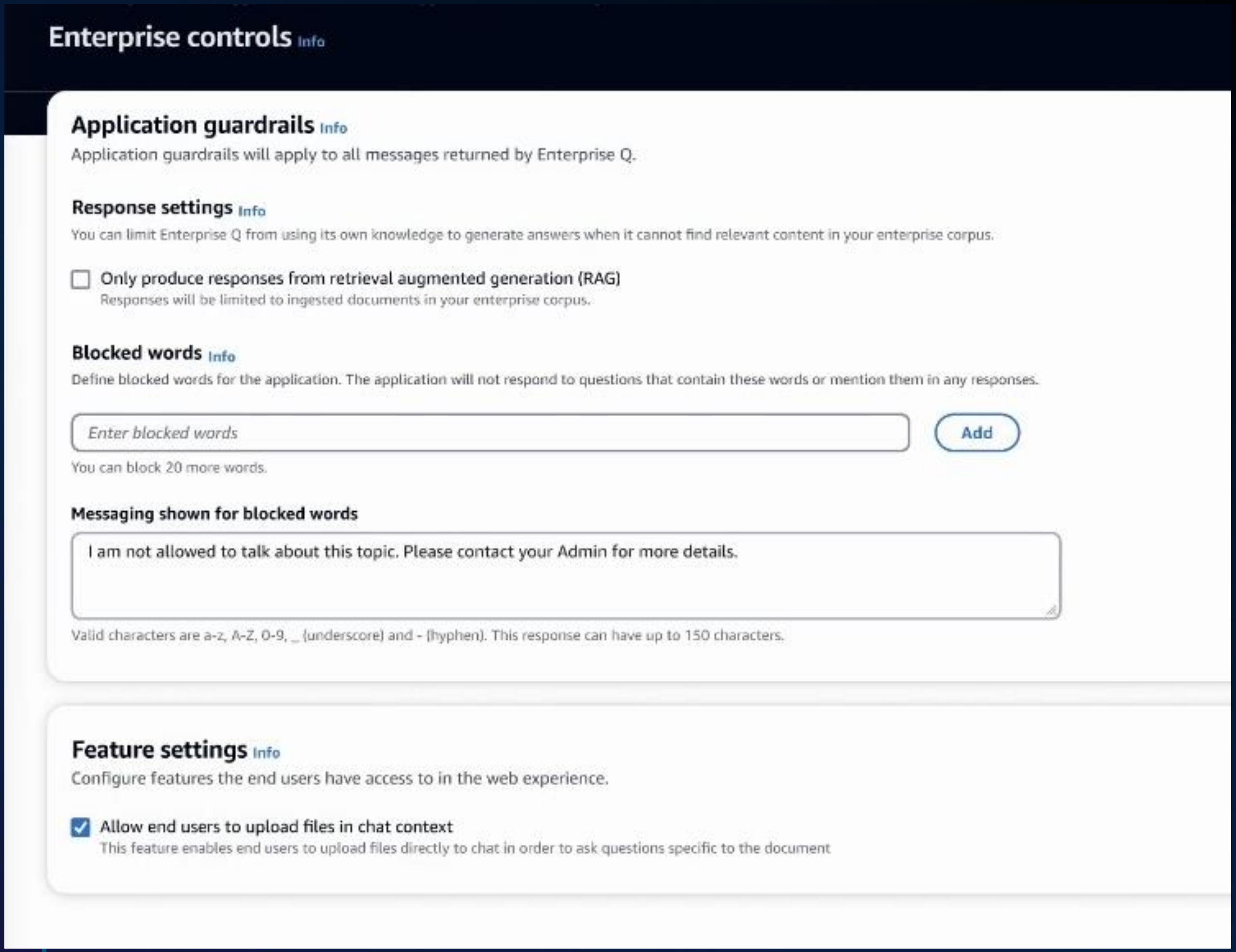
https://[redacted]August%2014%2C%202023%20_%20AWS%20News%20Blog.pdf

Amazon Q uses generative AI and it may give inaccurate or inappropriate responses. For more information: AWS Responsible AI Policy

Key Features

SAFETY AND SECURITY

- Use pre-built guardrails for toxicity
- Restrict responses to enterprise content only
- Specify blocked words or phrases that never appear in responses
- Define special topics and configure guardrails for such topics



Enterprise controls Info

Application guardrails Info

Application guardrails will apply to all messages returned by Enterprise Q.

Response settings Info

You can limit Enterprise Q from using its own knowledge to generate answers when it cannot find relevant content in your enterprise corpus.

☐ Only produce responses from retrieval augmented generation (RAG)
Responses will be limited to ingested documents in your enterprise corpus.

Blocked words Info

Define blocked words for the application. The application will not respond to questions that contain these words or mention them in any responses.

You can block 20 more words.

Messaging shown for blocked words

I am not allowed to talk about this topic. Please contact your Admin for more details.

Valid characters are a-z, A-Z, 0-9, _ (underscore) and - (hyphen). This response can have up to 150 characters.

Feature settings Info

Configure features the end users have access to in the web experience.

☑ Allow end users to upload files in chat context
This feature enables end users to upload files directly to chat in order to ask questions specific to the document

Key Features

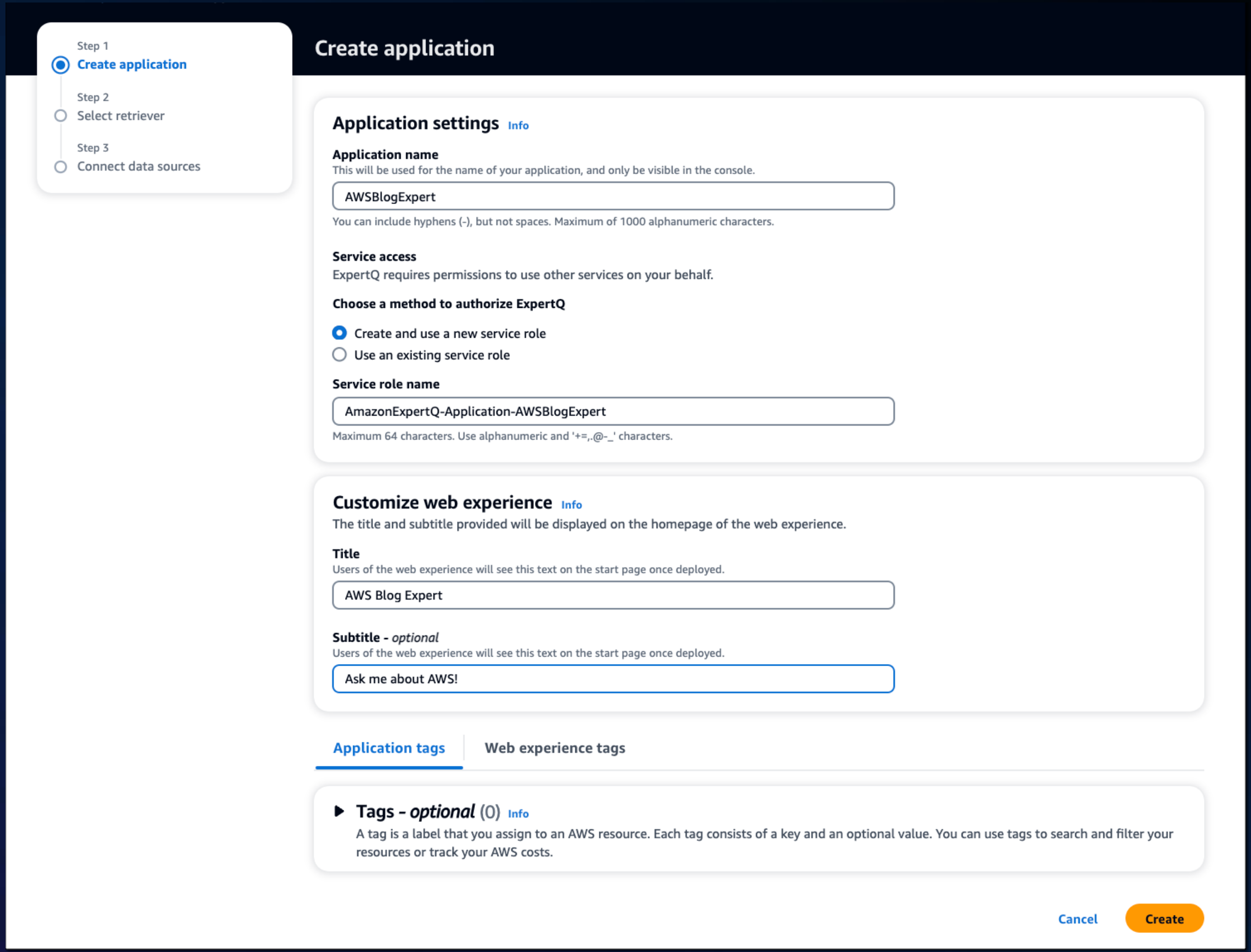
FASTER TIME TO MARKET

In-built vector index with managed ingestion

In-built application with SSO

3 click setup:
Settings, retriever, and data sources

Accuracy of retriever-augmented generation (RAG)



Step 1
● Create application

Step 2
○ Select retriever

Step 3
○ Connect data sources

Create application

Application settings Info

Application name
This will be used for the name of your application, and only be visible in the console.

AWSBlogExpert

You can include hyphens (-), but not spaces. Maximum of 1000 alphanumeric characters.

Service access
ExpertQ requires permissions to use other services on your behalf.

Choose a method to authorize ExpertQ

● Create and use a new service role
○ Use an existing service role

Service role name

AmazonExpertQ-Application-AWSBlogExpert

Maximum 64 characters. Use alphanumeric and '+=,.@-_' characters.

Customize web experience Info
The title and subtitle provided will be displayed on the homepage of the web experience.

Title
Users of the web experience will see this text on the start page once deployed.

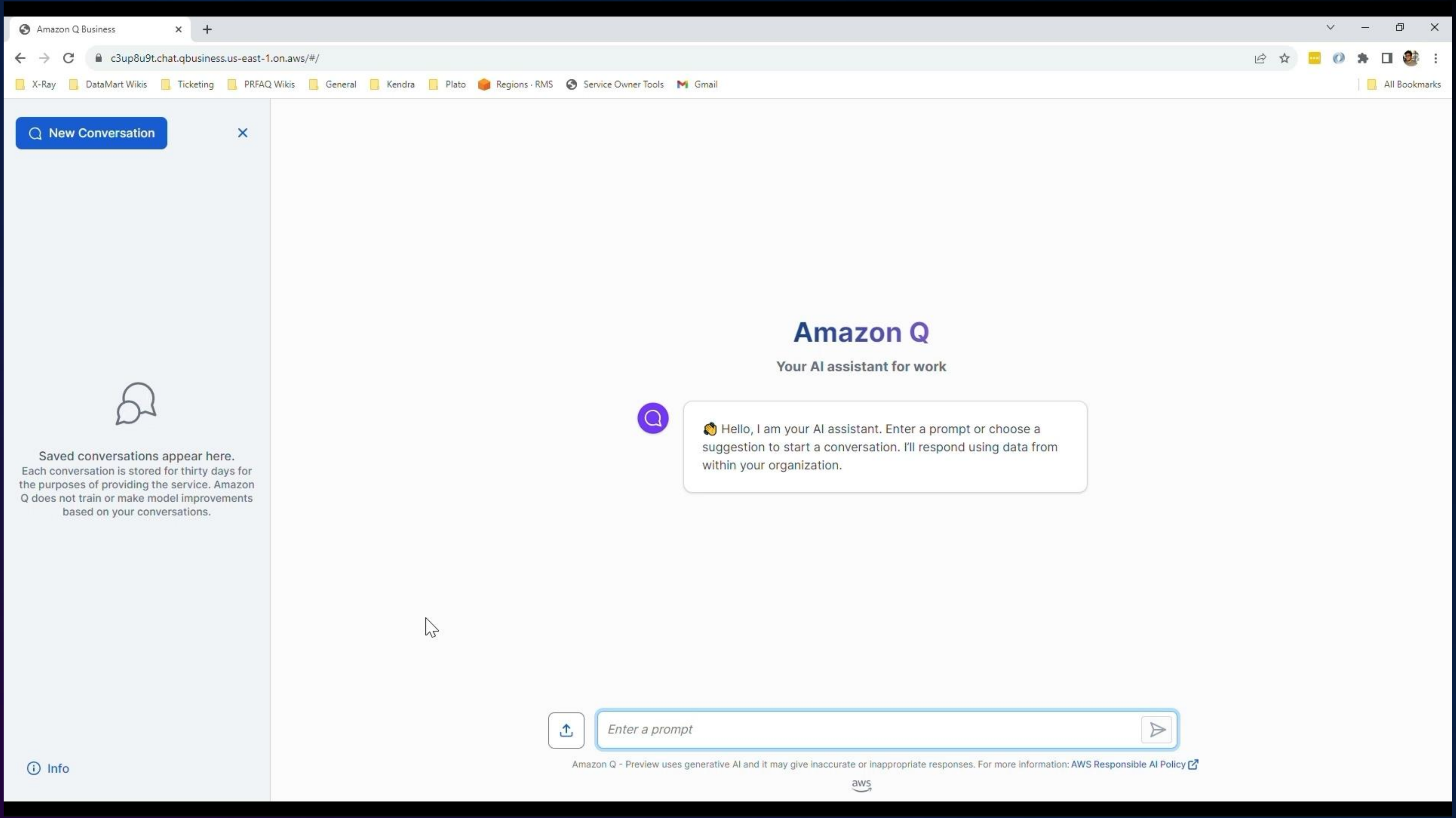
AWS Blog Expert

Subtitle - *optional*
Users of the web experience will see this text on the start page once deployed.

Ask me about AWS!

▶ **Tags** - *optional* (0) Info
A tag is a label that you assign to an AWS resource. Each tag consists of a key and an optional value. You can use tags to search and filter your resources or track your AWS costs.

Amazon Q for business Demo



Saved conversations appear here.
Each conversation is stored for thirty days for the purposes of providing the service. Amazon Q does not train or make model improvements based on your conversations.

Amazon Q

Your AI assistant for work

👋 Hello, I am your AI assistant. Enter a prompt or choose a suggestion to start a conversation. I'll respond using data from within your organization.

Enter a prompt

Amazon Q - Preview uses generative AI and it may give inaccurate or inappropriate responses. For more information: AWS Responsible AI Policy

Thank you!

Jatin Pal Singh

sinjatin@amazon.com

Ashutosh Dubey

ashdbey@amazon.com