



# Amazon Q

Your generative AI assistant designed for work that can be tailored to your business, data, code, and operations

**PREVIEW**

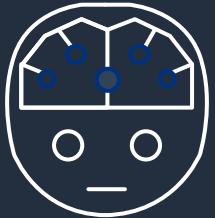
Birender Pal

Senior Solutions Architect  
AWS

# Agenda

- What is Generative AI and Foundational Models
- Use cases and challenges
- Using your data - RAG
- Amazon Q
- Amazon Q Business - Features
- Demo

# Where does Generative AI fit?



## Artificial intelligence (AI)

Any technique that allows computers to mimic human intelligence using logic, if-then statements, and machine learning



## Machine learning (ML)

A subset of AI that uses machines to search for patterns in data to build logic models automatically



## Deep learning (DL)

A subset of ML composed of deeply multi-layered neural networks that perform tasks like speech and image recognition



## Generative AI

Powered by large models that are pretrained on vast corpora of data and commonly referred to as foundation models (FMs)

# Generative AI is powered by foundation models

Pretrained on vast amounts of unstructured data

---

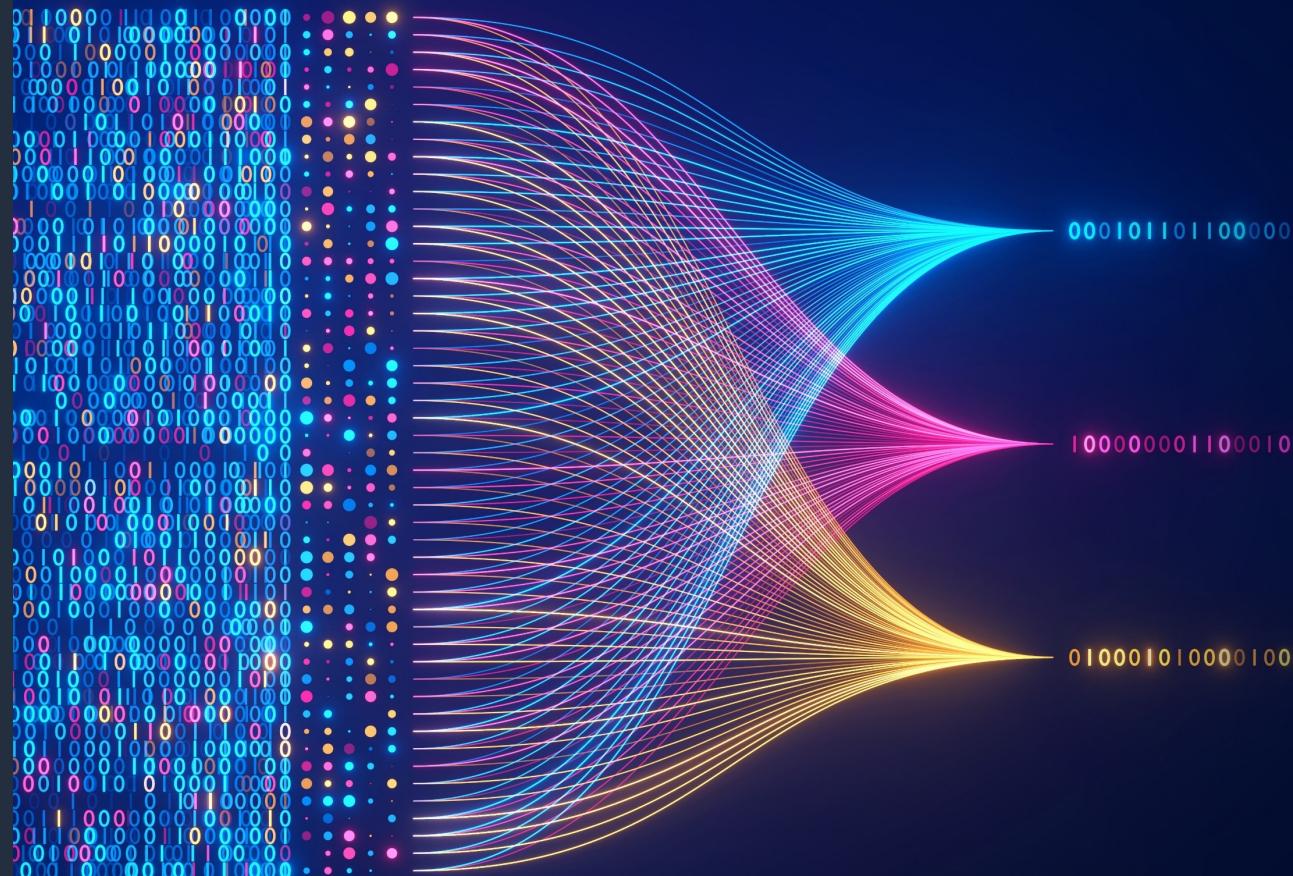
Contain large number of parameters that make them capable of learning complex concepts

---

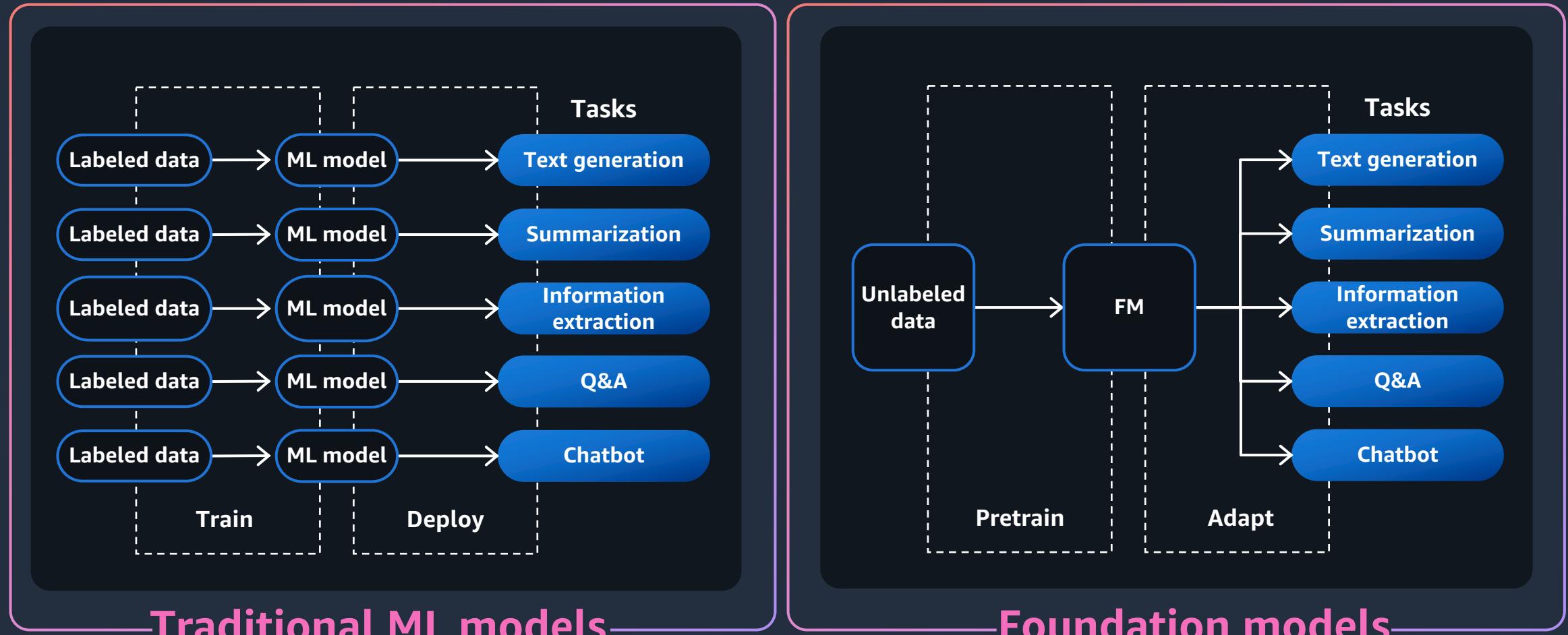
Can be applied in a wide range of contexts

---

Customize FMs using your data for domain specific tasks

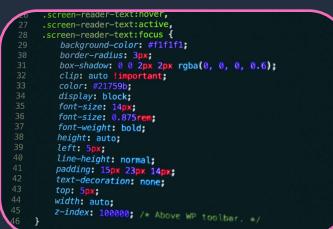
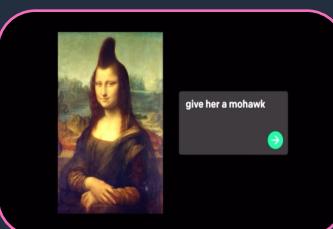
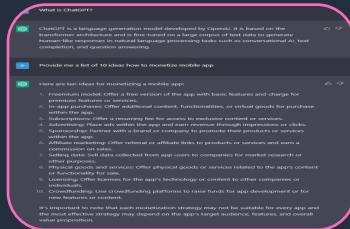


# How FMs differ from other machine learning (ML) models



# General Use Cases for Generative AI

## Content Creation



Text

Images

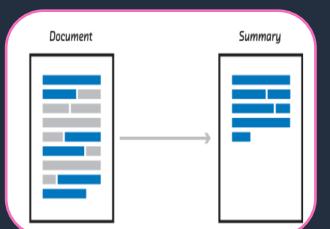
Code

Create copy for things like product descriptions, blogs, and marketing content.

Create product images or even show apparel on models. This is much cheaper than photography.

Generate software to accomplish specific tasks. Can really enhance programmer productivity.

## Natural Language Interactions



Translations

Search

Create copy for things like product descriptions, blogs, and marketing content.

Create product images or even show apparel on models. This is much cheaper than photography.

Generate software to accomplish specific tasks. Can really enhance programmer productivity.

Have more natural conversations with customers online, via voice, or even as an avatar or hologram. Provide answers to their questions.

Provide a summaries of bulk data such as weekly sales, competitive analysis, supply chain risks, or industry trends.

Translate copy to different languages across the globe. Done on the fly, this can be more cost-effective.

Better understand a user's intent, and assist them in finding products.

A photograph of a man with dark hair and a beard, wearing glasses and a blue denim shirt. He is sitting at a wooden desk, looking down at his laptop with a weary expression, his hand near his face. The background is blurred, showing an office environment with a lamp and other furniture.

**Information overload impedes  
employee productivity**

# Opportunity with generative AI



Generative AI's ability to understand natural language enables automation for work activities that account for

**25%** of total  
work time

---



**Imagine a generative AI powered assistant that saves  
you**

every  
day

**2 hours**

McKinsey Digital June 2023



© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# Risks and challenges with generative AI



**Wrong answers  
(hallucinations)**



**Sharing confidential  
information**



**Insufficient or obsolete  
data**



**Don't know about your  
business and your customers**

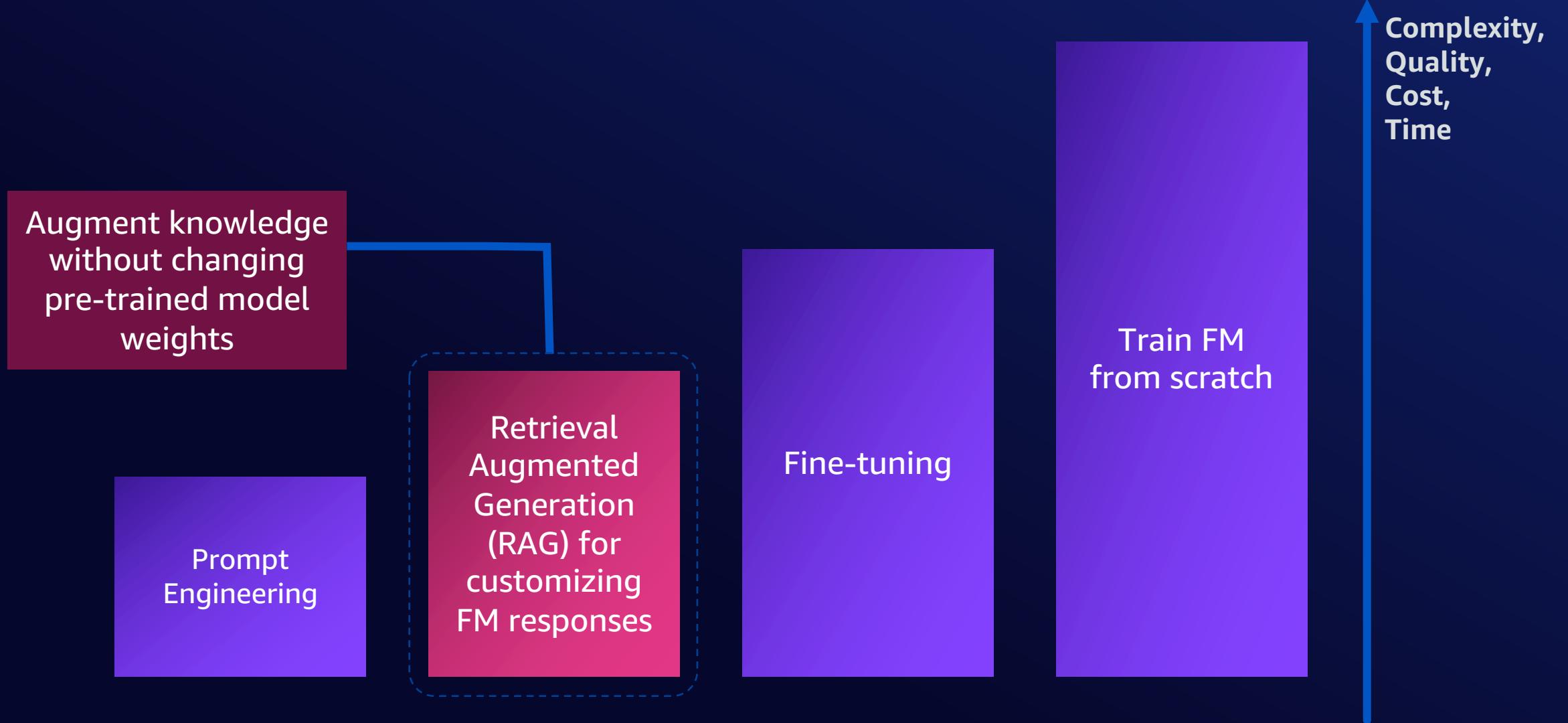


**Don't know about  
your role**



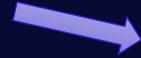
**Can't connect to your  
data**

# Common approaches for customizing foundation models (FMs)



# Elements of the prompt example

Instructions  
and  
output indicator



Context



Input data



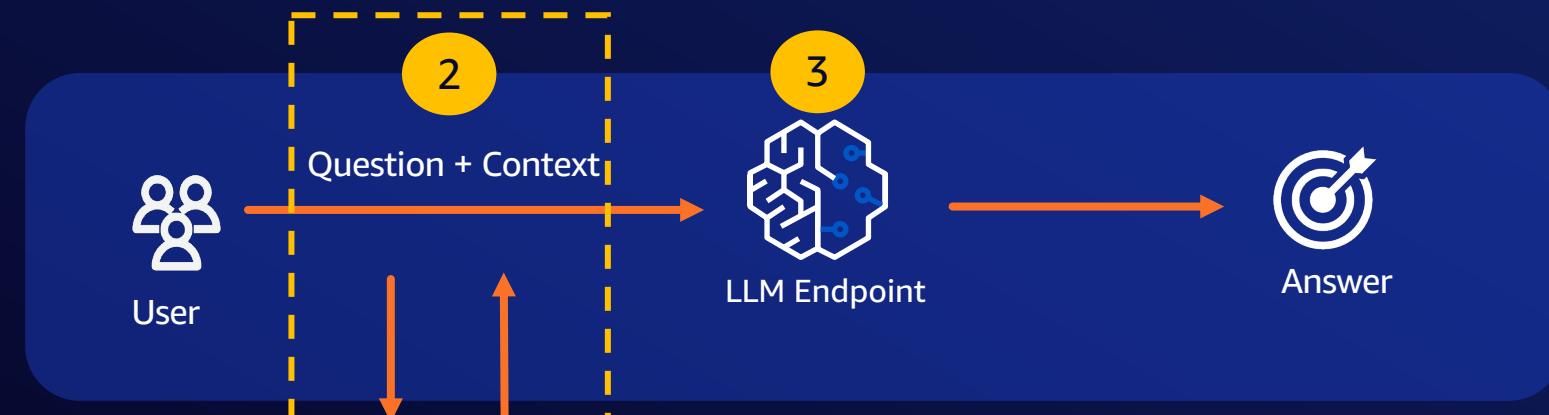
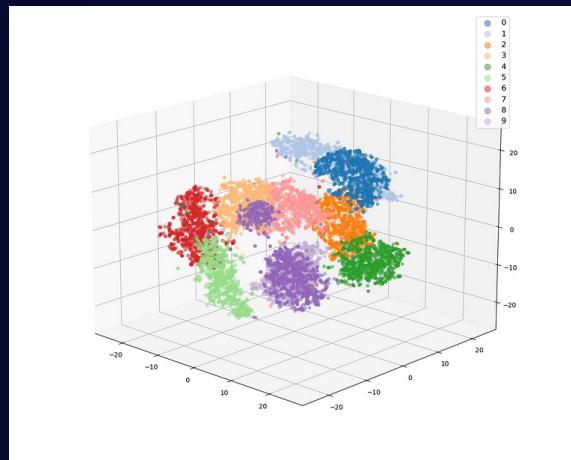
Prompt	Output
<p>Write a summary of a service review using two sentences.</p> <p>Store: Online Service: Shipping</p> <p>Review: Amazon Prime Student is a great option for students looking to save money. Not paying for shipping is the biggest save in my opinion. As a working mom of three who is also a student, it saves me tons of time with free 2-day shipping, and I get things I need quickly and sometimes as early as the next day, while enjoying all the free streaming services and books that a regular Prime membership has to offer for half the price. Amazon Prime Student is only available for college students, and it offers so many things to help make college life easier. This is why Amazon Prime is the no-brainer that I use to order my school supplies, my clothes, and even to watch movies in between classes. I think Amazon Prime Student is a great investment for all college students.</p> <p>Summary:</p>	Amazon Prime Student is a fantastic option for college students, offering free 2-day shipping, streaming services, books, and other benefits for half the price of a regular Prime membership. It saves time and money, making college life easier.

1

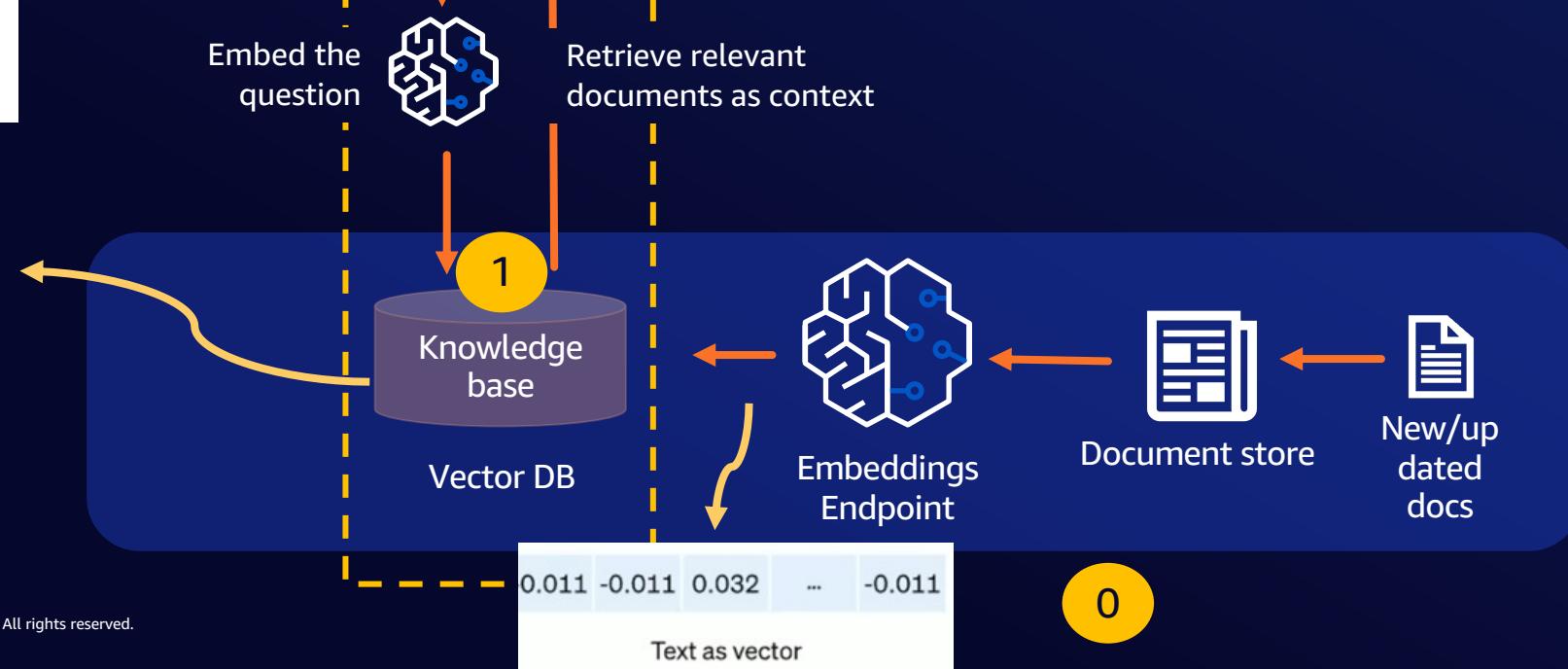
2

3

# RAG – Retrieval Augmented Generation



Embedding Chunk 1	
-0.011	-0.011 0.032 ... -0.011
Embedding Chunk 2	
-0.011	-0.011 0.032 ... -0.011
...	
Embedding Chunk n	
-0.011	-0.011 0.032 ... -0.011



# Implementing RAG can be resource intensive and time consuming



# Introducing Amazon Q



**Your generative AI assistant designed for work that can be tailored to your business, data, code, and operations**

# Amazon Q areas of expertise

Your  
business



Building  
on AWS



Amazon  
QuickSight



Amazon  
Connect



AWS  
Supply Chain



# Supercharge productivity



## Streamline search experience

Connect internal wikis, SharePoint sites, Confluence, Salesforce, and 40+ other applications for a unified, conversational search experience.



## Accelerate content creation

- Draft an email
- Provide 5 conference session titles on the topic of "Sustainable workplace"
- Generate 3 social posts for the launch of Jasper



## Generate summaries

Quickly understand the essence of documents by creating summaries of uploaded files or existing enterprise documents.



## Extract key insights

- What is the 5- year CAGR?
- How has revenues and margins changed over the last 3 quarters? Why?

# Supercharge productivity

FIND ACCURATE AND REFERENCEABLE ANSWERS



Trusted answers generated from enterprise data



In-context conversations



Source references for fact-checking



Conversation history

The screenshot displays the Amazon Q interface with four distinct conversational threads:

- Thread 1:** A user asks "What is the reliability of S3?" The AI responds by explaining S3's durability and redundancy across multiple facilities and devices, mentioning 10,000,000 data assets for 10,000 years. It also discusses the infrastructure of Availability Zones. A "Sources" button is present at the bottom.
- Thread 2:** A user asks "Tell me more about availability zones." The AI provides a detailed explanation of how availability zones are physically separated within a region to protect against single location failures. It mentions independent infrastructure, redundant power, and networking. A "Sources" button is present at the bottom.
- Thread 3:** A user asks "How many availability zones should my data be stored across?" The AI recommends storing data across multiple availability zones for higher availability and fault tolerance. It lists several key reasons: isolation of resources, maximization of resource availability, reduced impact of single zone failure, and the placement of read replicas. A "Sources" button is present at the bottom.
- Thread 4:** This thread is partially visible at the bottom of the interface.

Each thread includes standard interaction buttons like upvote, downvote, share, and copy. The interface also features an "Info" button and a "New Conversation" search bar at the top left.

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

# Supercharge productivity

## UPLOAD FILES AND ANALYZE CONTENT

The screenshot shows the Amazon Q interface. On the left, there's a sidebar with a 'New Conversation' button and two recent queries: 'Summarize this document' (November 24, 2023) and 'What is the reliability of S3?' (November 24, 2023). The main area displays two AI-generated responses. The first response is for summarizing the US Constitution, which is described as outlining the structure of the federal government and separation of powers between the legislative, executive, and judicial branches. It also mentions the establishment of citizens' rights and limitations on government power. The second response is about the First Amendment, stating it prohibits Congress from making laws respecting an establishment of religion or prohibiting its free exercise, and guarantees freedom of speech, freedom of the press, the right to peaceably assemble, and the right to petition the government. Both responses include a 'No sources' button and sharing icons. At the bottom, there's a note about using generative AI and a link to the AWS Responsible AI Policy.

- Summarize documents
- Ask questions about content in the documents
- Analyze data
- Multiple file types supported – pdf, docx, csv, ppt, txt and more

# Supercharge productivity

## EXECUTE ACTIONS USING PLUGINS

 Create a Jira ticket

Project

Project B

Priority - optional

Select a priority

Highest

High

Normal

Low

Lowest

Bug

Enhancement

To set up the VPN connection, the customer needs to first create the VPN components including a customer gateway and VPN gateway or transit gateway. The customer gateway represents their on-premises device and network and needs to be created by providing information like the external IP address or certificate. Then a VPN connection can be established between the customer gateway and VPN

Cancel Create

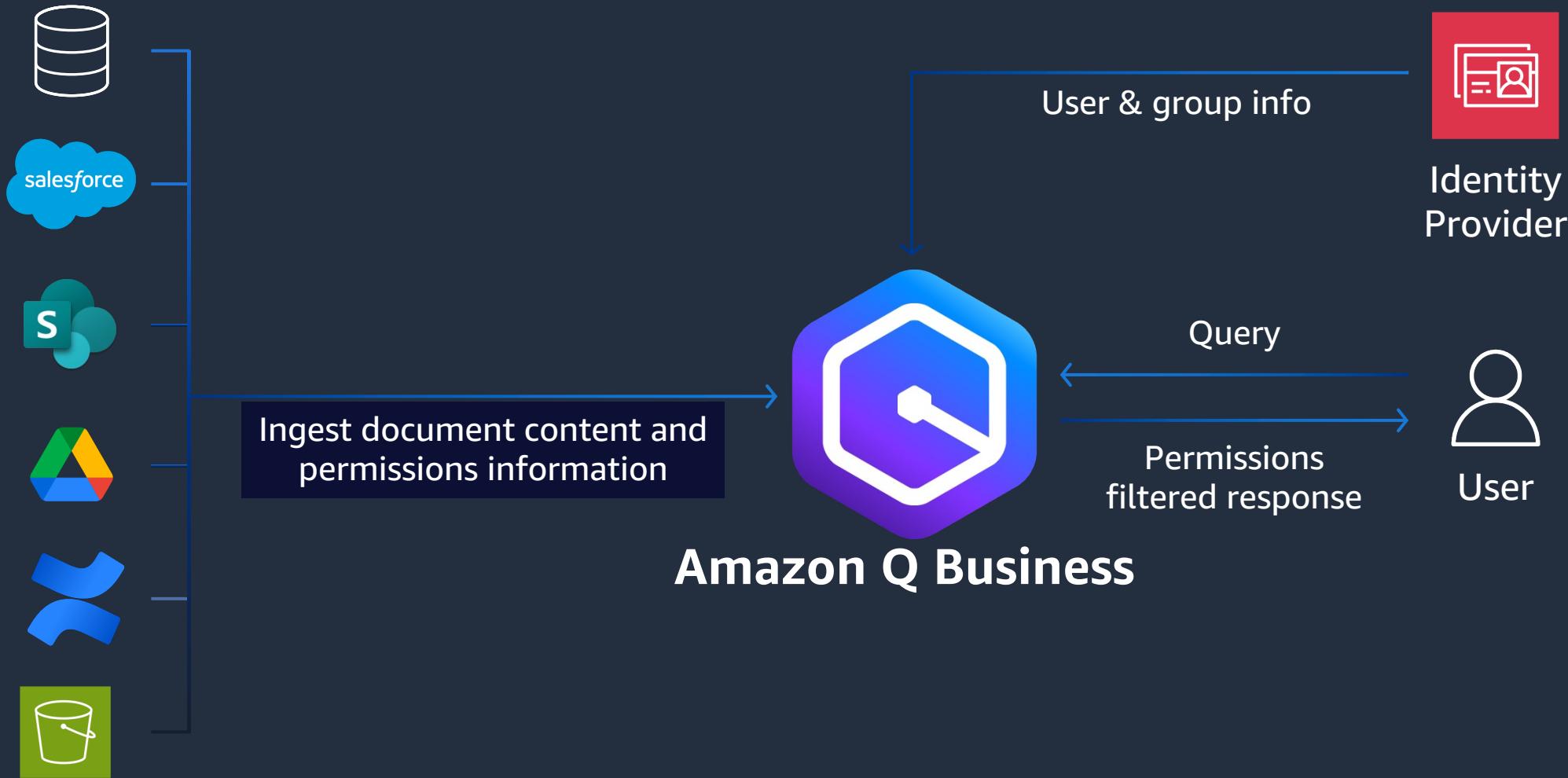
Enable end-users to perform actions on SaaS applications

*"summarize conversation and create ticket in Jira"*



# Key Features – Safety and Security

BUSINESS Q IS AWARE OF ENTERPRISE USER PERMISSIONS



# 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  
 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.

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**

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.

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

**Application tags** **Web experience tags**

**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.

**Create**

# Adhere to data privacy and security needs

## PROTECT AGAINST TOXIC TOPICS WITH PRE-BUILT GUARDRAILS

Update global controls Info

**Global controls** 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.

Enter blocked words

You can block 18 more words.

Account vulnerabilities  Project X

**Messaging shown for blocked words**  
I cannot complete this request as the response contains content that is blocked by your Admin. Please contact your Admin for help.

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

**Feature settings** Info  
Configure features 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.

Use pre-built guardrails for toxicity

Restrict responses to enterprise content only

Specify blocked words or phrases that never appear in responses

# Adhere to data privacy and security needs

## ESTABLISH GUARDRAILS AND CONFIGURE CUSTOM TOPICS

Create topic specific control [Info](#)

Name and description [Info](#)

Name  
Gaps in our security architecture

Description  
Outline how the model should use this guardrail.  
Do not discuss gaps in our company's security architecture

Example chat messages - optional (2) [Info](#)

Add representative phrases that you expect a user to type to invoke this topic.

Example chat message

List vulnerabilities in our security architecture [Remove](#)

Assess the effectiveness of our security controls [Remove](#)

Add new example chat message

You can add 3 more example chat messages.

▼ Rule 1

Behavior in response to topic control [Info](#)

Define how Enterprise Q should handle the topic.

Behavior  
Block completely

Messaging shown

I cannot complete this request as the response contains content that is blocked by your Admin. Please contact your Admin for help.

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

User handling [Info](#)

Specify this rule to user groups  
Define included or excluded user groups.

Include Rule only applies to the list of user groups

Exclude Rule applies to all except the list of user groups

User groups

Specify user groups that this topic control applies to.

Search

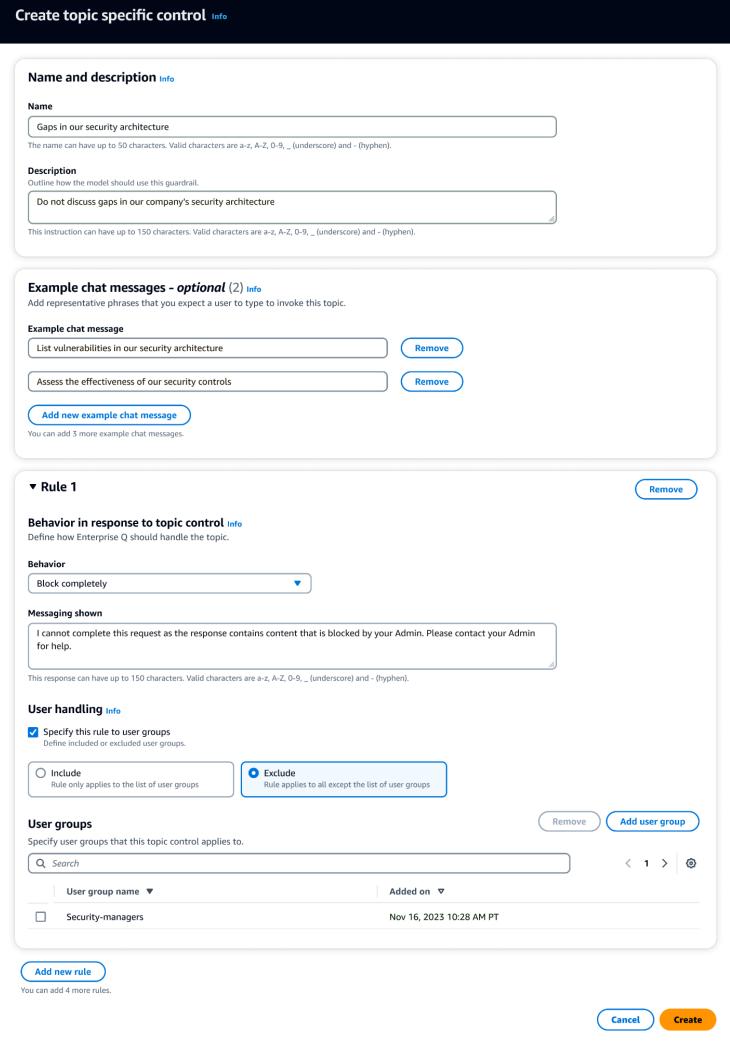
User group name [Remove](#) Added on [Nov 16, 2023 10:28 AM PT](#)

Security-managers

Add new rule

You can add 4 more rules.

[Cancel](#) [Create](#)



Define special topics and configure 4 levers of guardrails for such topics –

- Respond with an pre-defined message
- Restrict responses to enterprise content
- Restrict responses to enterprise content with metadata filters
- Apply guardrails to specific users and groups in the enterprise



# Boost Time to Value

**Unify content from all your enterprise sources together in a few clicks!**

Adobe Experience Manager	Jira
Alfresco	Microsoft Exchange
Amazon Simple Storage Service (Amazon S3)	Microsoft OneDrive
Atlassian Confluence	Microsoft SharePoint
Aurora (MySQL, PostgreSQL)	Microsoft Teams
Box	Microsoft Yammer
DB2	Microsoft SQL Server
Dropbox	Quip
Drupal	Salesforce
Custom Connector	ServiceNow
FSX for Windows	Slack
Github	Web Crawler
Gmail	Workdocs
Google Drive	Zendesk



# Demo



© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.



# Thank you!

Birender Pal

[palbiren@amazon.com](mailto:palbiren@amazon.com)

@biren\_pal