Introduction to GenAI

Adi Jabkowsky

Sr. GenAI Specialist

Liat Tzur

Sr. Technical Account Manager

Mobileye’s AWS account team

NEW ADDITIONS TO THE TEAM

Shira Mayer
Account Manager

Olga Netzer Nevzorov
Customer Solution Manager

Daniel Goldenstein Bar
Solution Architect

Boyd McGeachie
Executive Sponsor

Liat Tzur
Lead Technical Account Manager

Irmi Tzuri
Technical Account Manager

Udi Dahan
Technical Account Manager

Dani Yakubov
Concierge

Agenda

- Intro to GenAI
- AWS offering
  - Amazon Q + demos
  - Amazon Bedrock + demos
- AV – Showcasing their Amazon Bedrock use case
- Next steps for your GenAI journey
- Quiz
- Survey

A glimpse into history

Stone Age tools

-2.6M years

One invention leads to another

Language was invented

-2M years

From fire to language

Bronze Age tools

-3K years

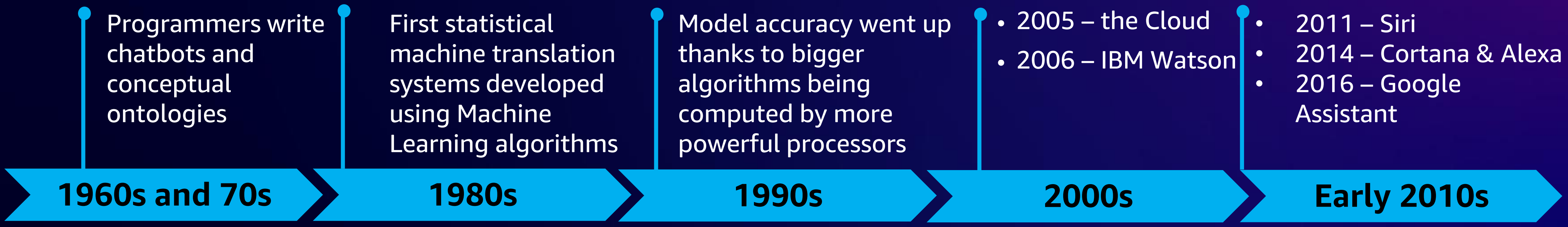
From writing to machines & technology

The industrial revolution

18th century

From horses to cars

Generative AI builds on 50 years of progress



1960s and 70s

Programmers write chatbots and conceptual ontologies

1980s

First statistical machine translation systems developed using Machine Learning algorithms

1990s

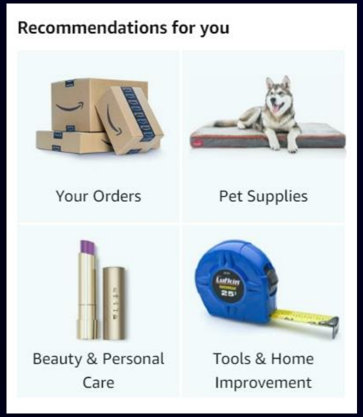
Model accuracy went up thanks to bigger algorithms being computed by more powerful processors

2000s

- 2005 – the Cloud
- 2006 – IBM Watson

Early 2010s

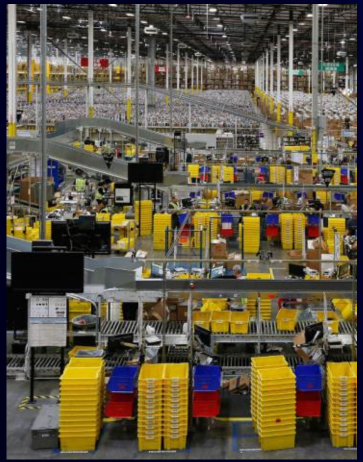
- 2011 – Siri
- 2014 – Cortana & Alexa
- 2016 – Google Assistant



Recommendations for you

Your Orders	Pet Supplies
Beauty & Personal Care	Tools & Home Improvement

4,000 products per minute sold on Amazon.com



1.6M packages every day

Billions of Alexa interactions each week

Just Walk Out technology in airports, stadiums and more

Generative AI builds on 50 years of progress

- 2011 – Siri
- 2014 – Cortana & Alexa
- 2016 – Google Assistant

Early 2010s

- Transformer architecture developed
- Data scientists realize bigger models are better
- GPT 2, T5, and BERT models released

2017 - 2022

November 2022 – Chat GPT launched

2023 – the year of experimentation & democratization

New Models every week, trend towards big and small models

- Private models such as Anthropic - Claude, Amazon – Titan, Meta - Llama, OpenAI – GPT, etc.
- Open Source models have evolved very fast and are catching up with private models (Hugging Face)

2024 – the year of production and scale

How smart are models these days?

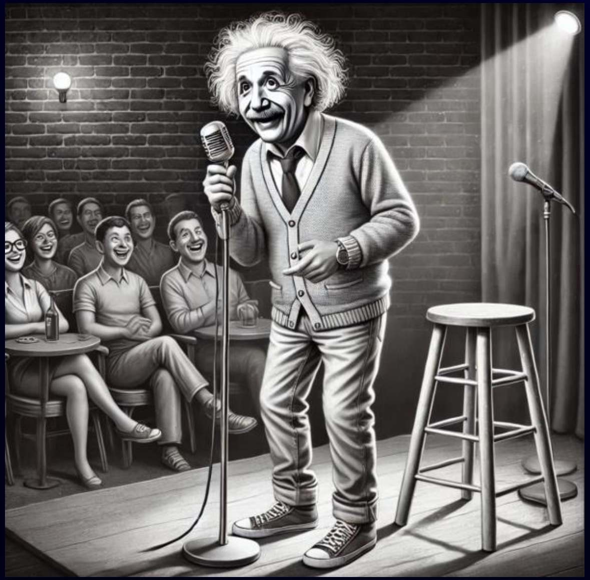
If someone did nothing but read 24 hours a day for their entire life they would consume 8 Billion words

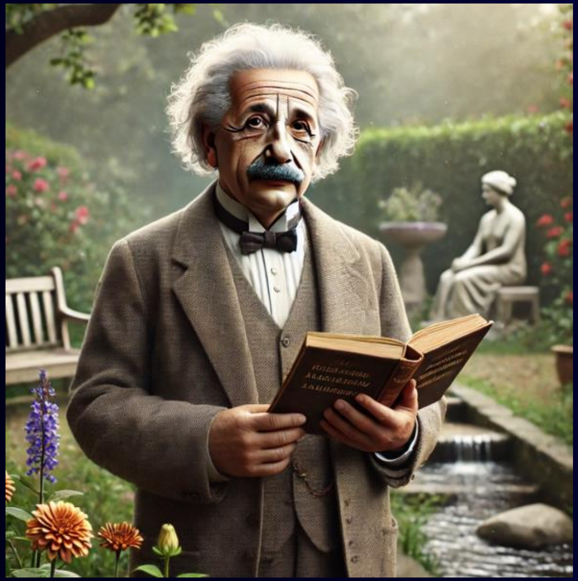
How smart are models these days?

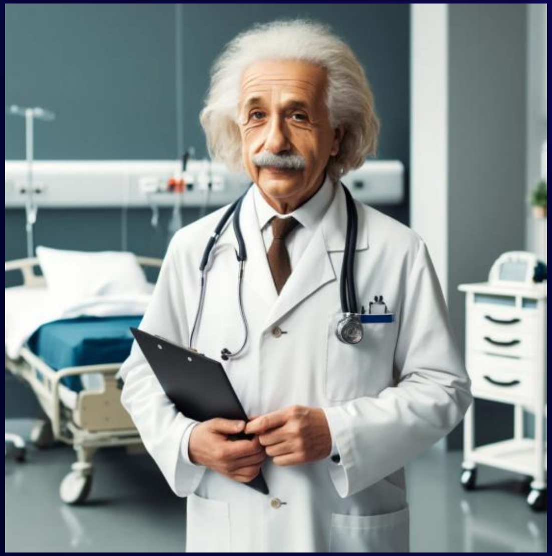
The most advanced models today consumed more than 8 Trillion words

So what is Generative AI?

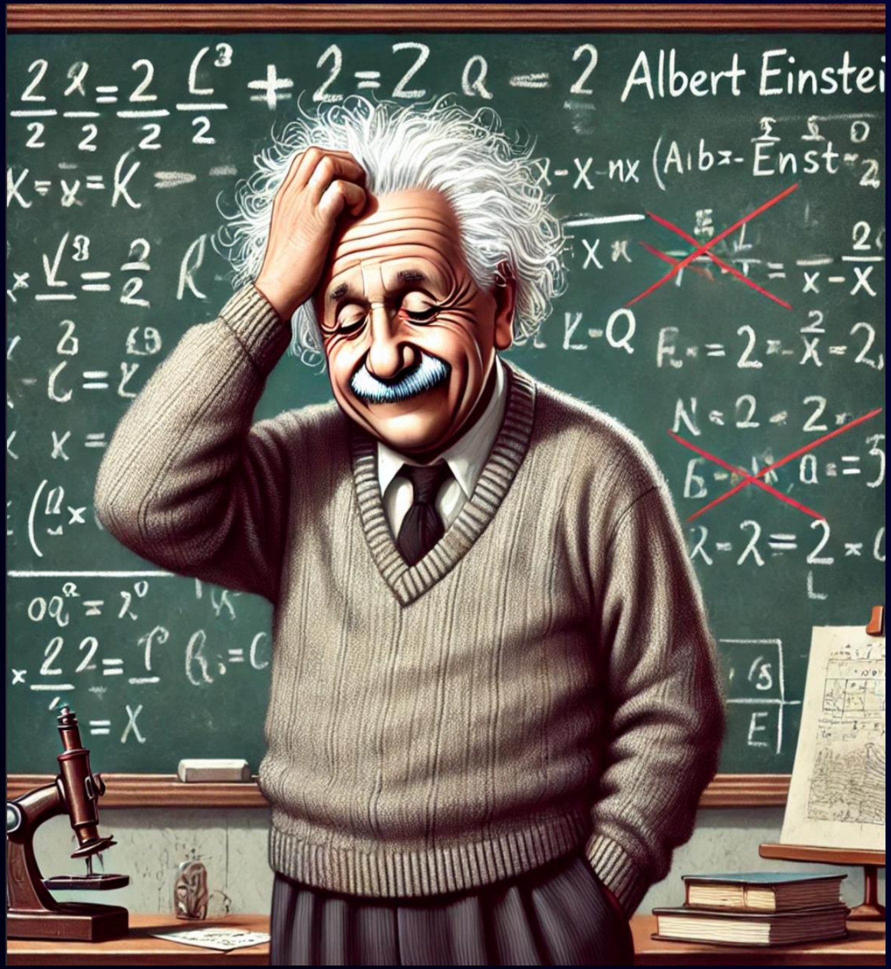
Think of it as getting access to the world's smartest brain







Is Generative AI mistakes free?



Although models are very smart they can make human like mistakes, such as:

- Jump into conclusions (the bias effect)
- Make mistakes / hallucinate
- Misunderstand you

The biggest model’s limitation is your imagination & prompt

Let’s start with the basics…

AI – Artificial Intelligence
Netflix suggestions, Waze, parking barrier arm.

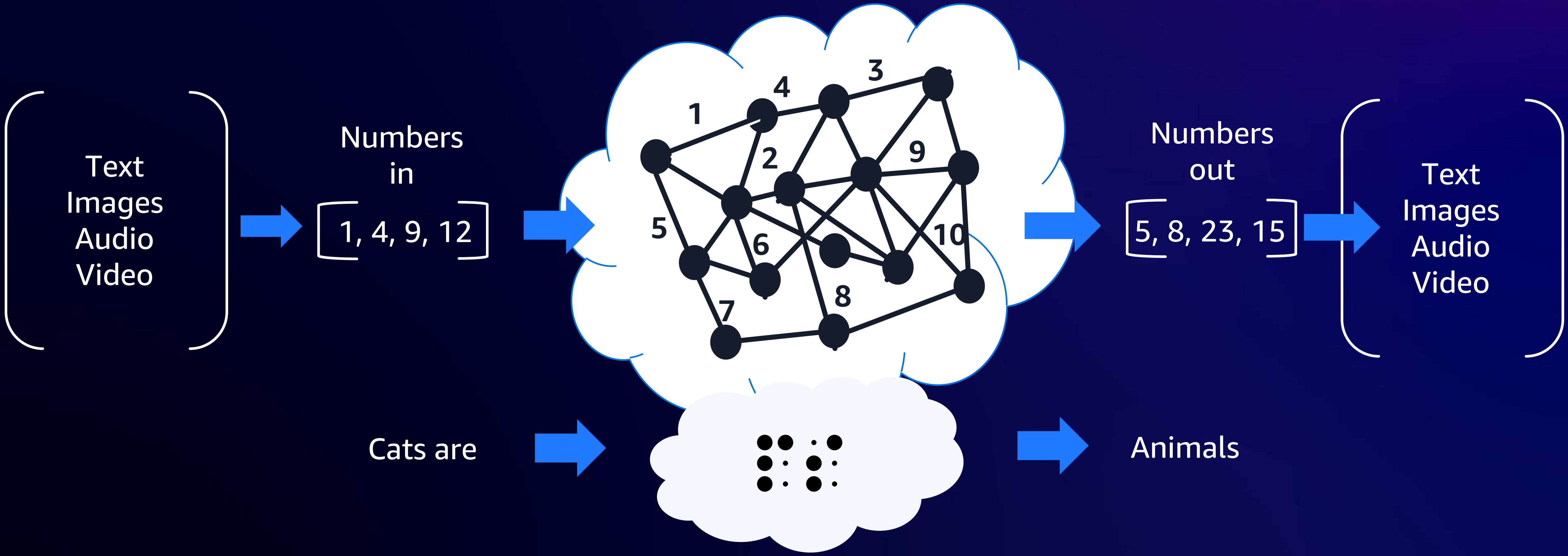
GenAI
Generating new content from scratch (in contradiction to finding existing data). The “G” in GPT stands for generative.

LLM - Large Language Model
A type of Gen AI that can communicate in human language.

Prompt
The text we send to a model. Instructions & Questions.

ChatGPT
A famous chatbot, developed by OpenAI, based on Transformer Architecture that was invented by Google in an article called “Attention is all you need”. The “T” in ChatGPT stands for Transformer.

How does LLM work?



Text
Images
Audio
Video

Numbers
in

1, 4, 9, 12

Numbers
out

5, 8, 23, 15

Text
Images
Audio
Video

Cats are

Animals

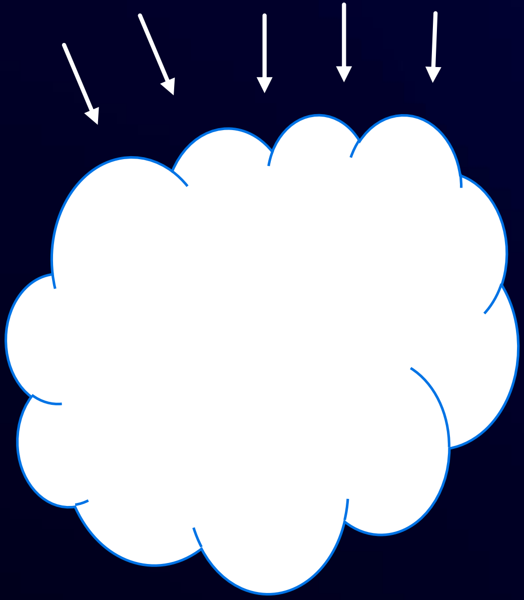
How do these models know what they know?

Training!

Training breakdown

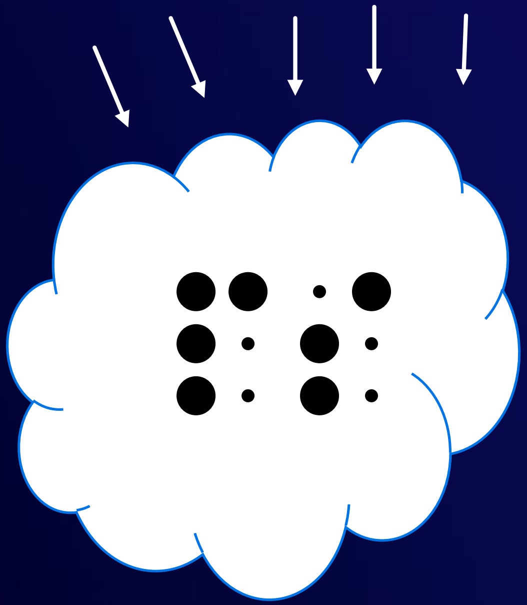
1

Unsupervised training (stupendously large amount of text)



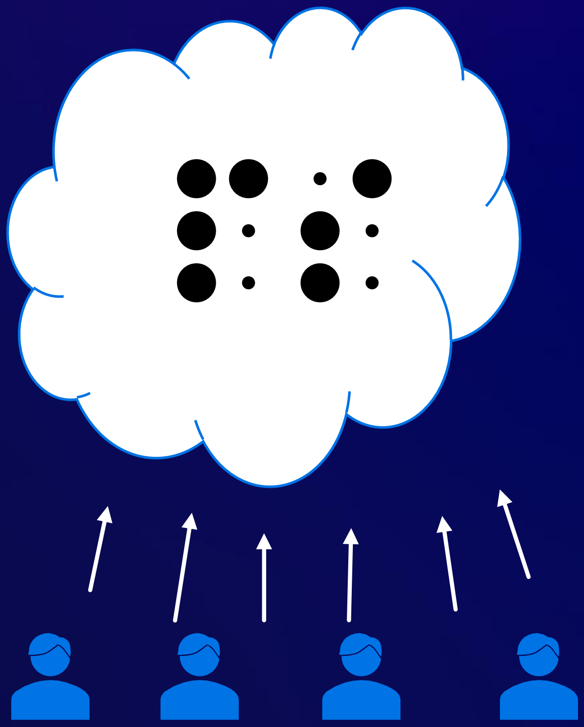
2

Guess the next word! (Neural network is beginning to form)

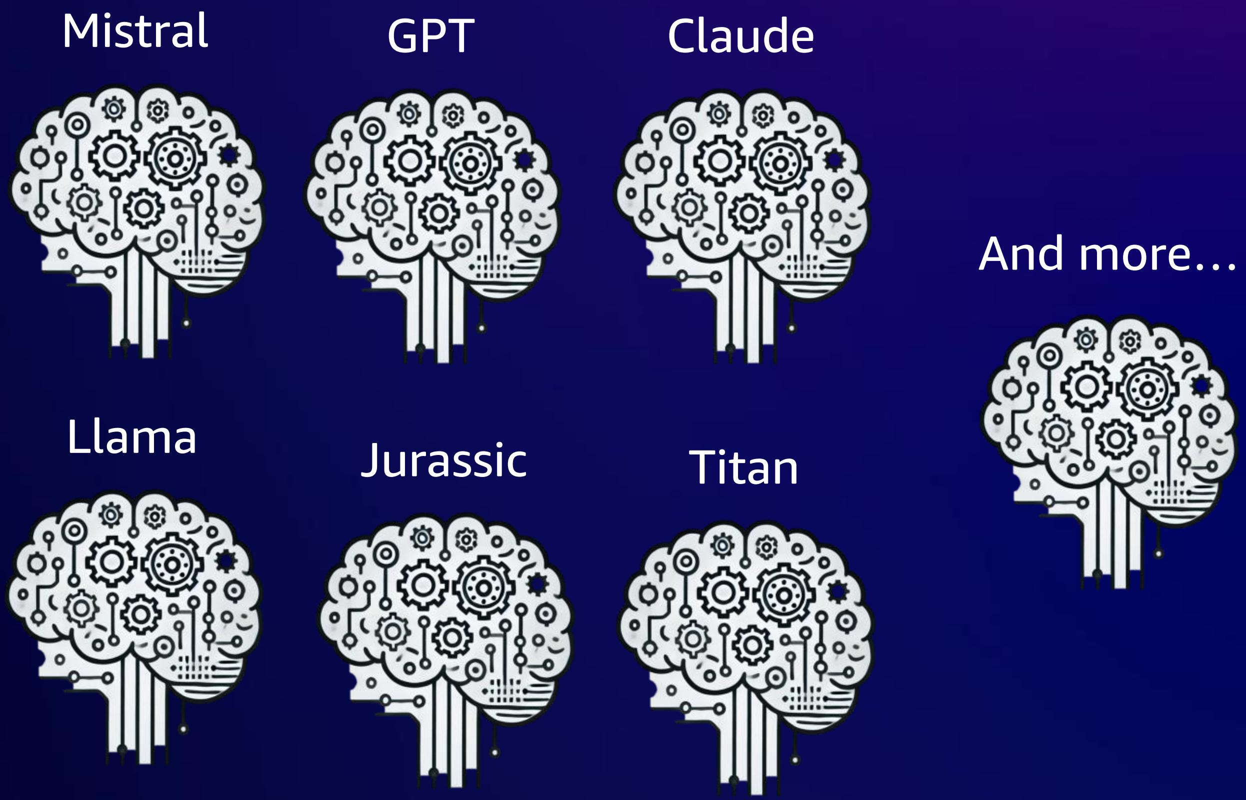


3

Reinforcement Learning with Human Feedback



Models, models and some more models!



Mistral

GPT

Claude

Llama

Jurassic

Titan

And more…

Type of models

Text to Text:
Code
Summarization
Content creation

Speech to text:
Voice to text
Translation

Text to Audio:
Create sounds & music

Text to Video:
Create videos using text
Sora

Text to Image:
"Create image of Albert Einstein and an image of a cat"

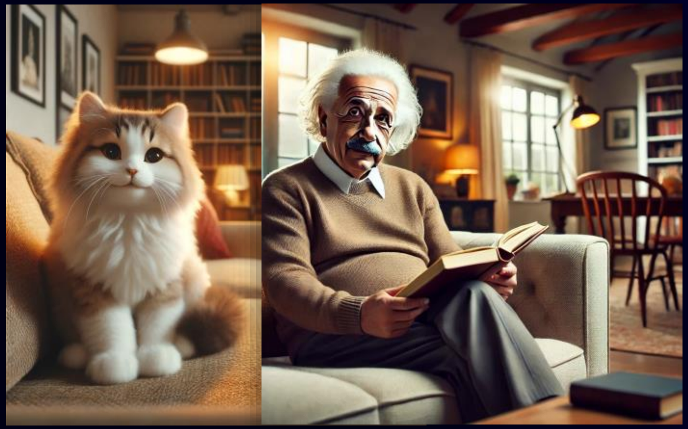


Image to image:
Combine the two images

Image to Text:
Describe the image

"The image depicts Albert Einstein sitting on a chair in a cozy study room. His face has been creatively fused with that of a cat"

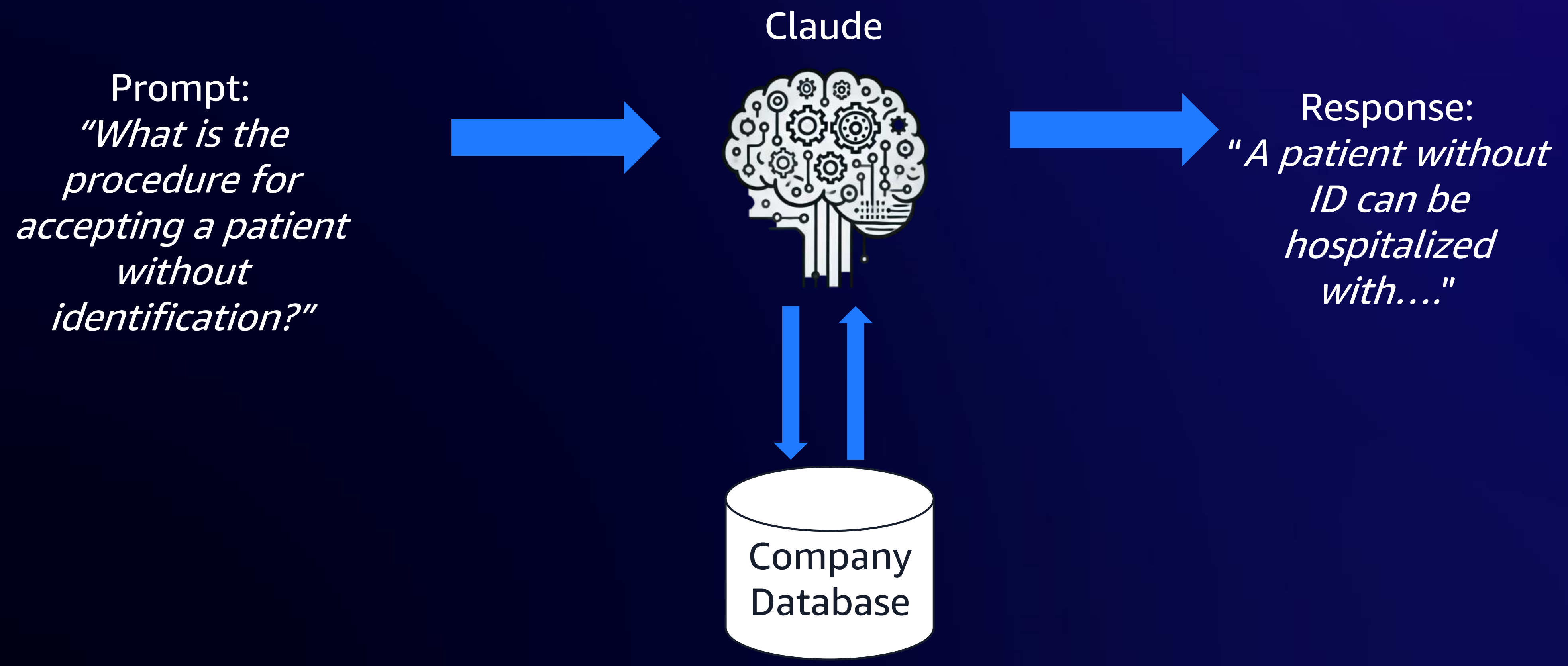
Type of models

Multi Modal
models

Generative AI terms

RAG – Retrieval Augmented Generation

A technique to give a model access to data that is not part of it's training.


Prompt:
"What is the procedure for accepting a patient without identification?"
Claude
Response:
" A patient without ID can be hospitalized with…."
Company Database

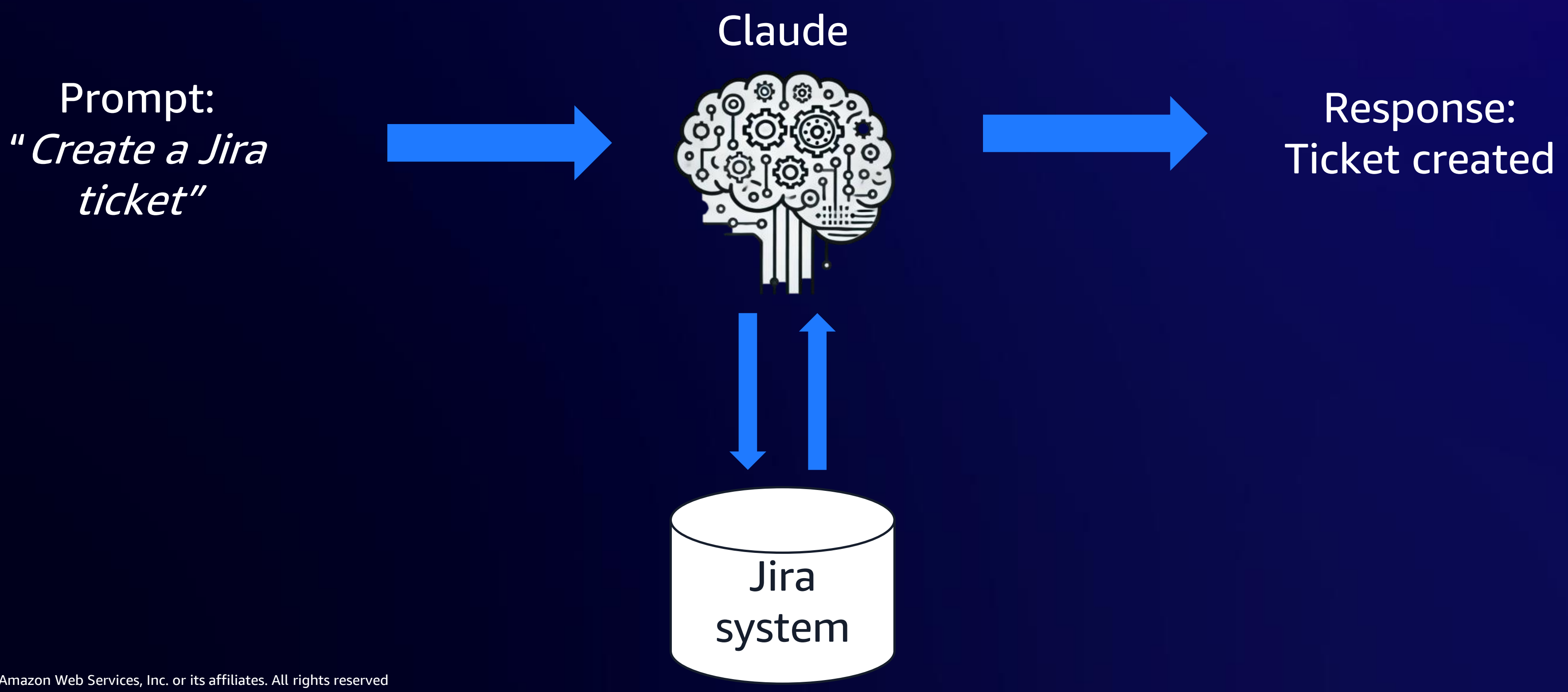
Generative AI terms

Fine Tuning

Training a pre-trained model with additional data

Agent

a component that models can use to perform actions. Can be used to automate processes.



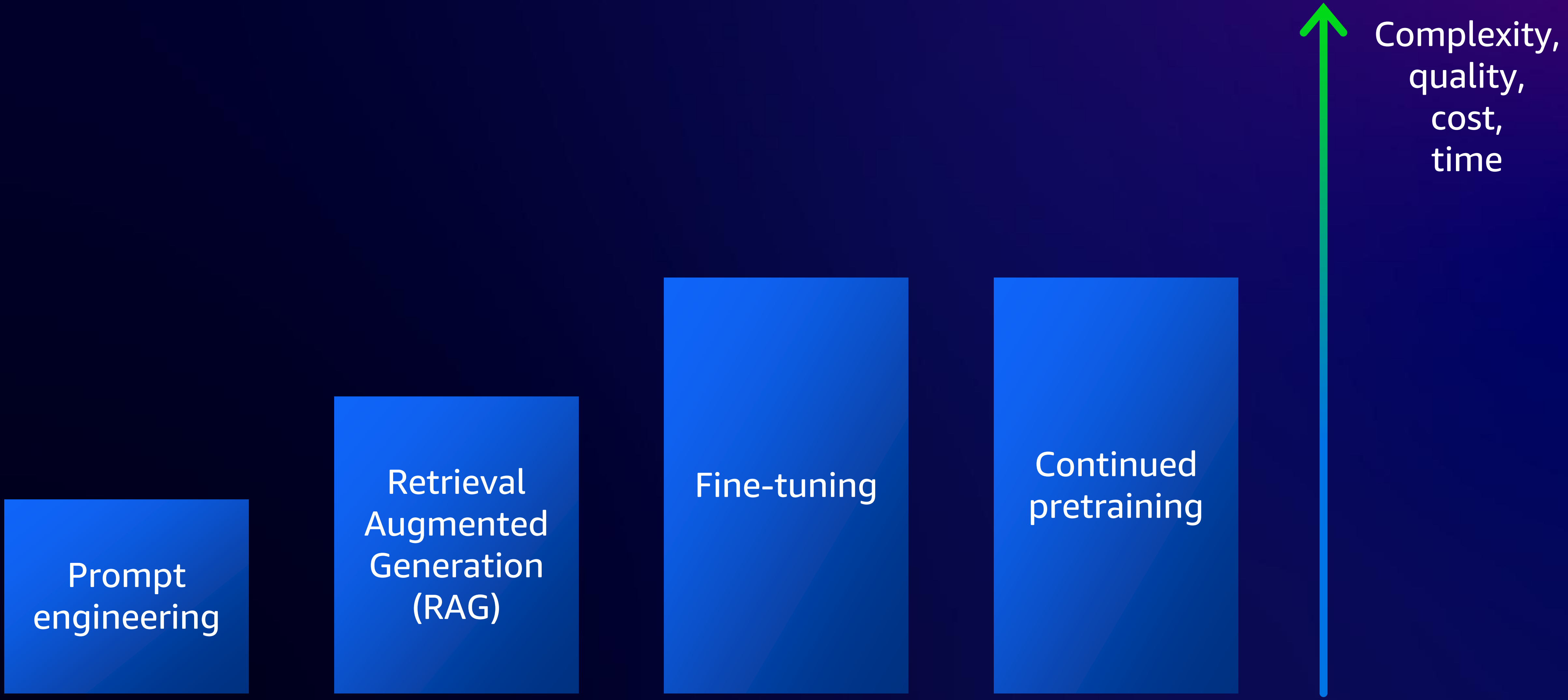
Claude

Prompt:
“ Create a Jira ticket”

Response:
Ticket created

Jira system

Common approaches for customizing FMs



Prompt engineering

Retrieval Augmented Generation (RAG)

Fine-tuning

Continued pretraining

Complexity, quality, cost, time

What generative AI does well (for now)

- ✓ Summarization
- ✓ Content generation (Text, image, audio, video)
- ✓ Language Translation
- ✓ Correction/paraphrasing
- ✓ Classification

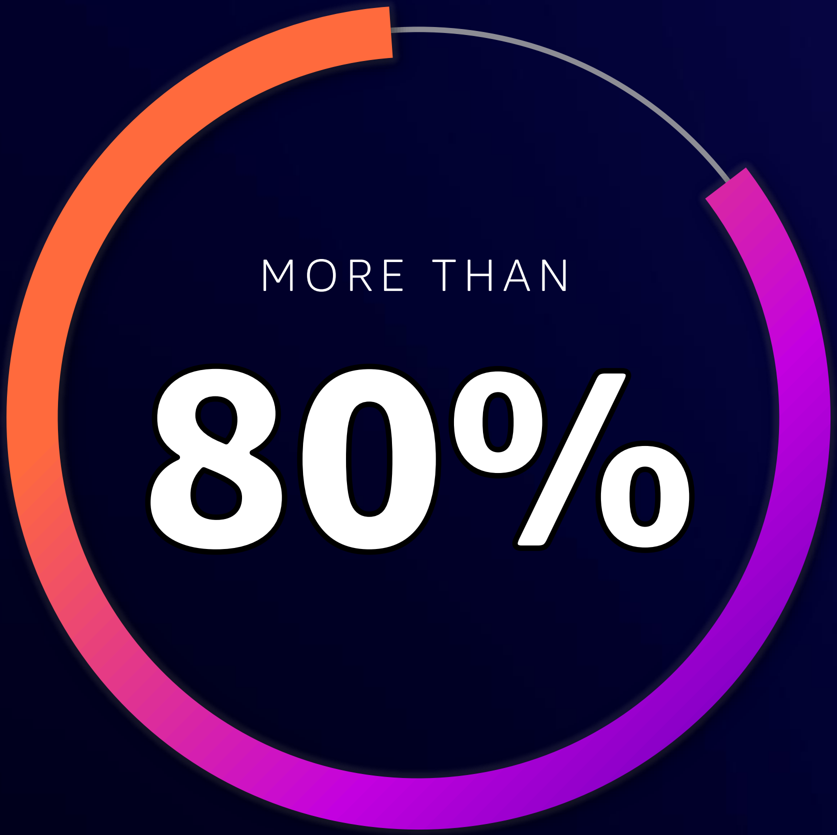
Use case examples

- Extract insights from your documents
- Create a product description
- Generate draft marketing copy
- Create a job posting
- Summarize your meetings
- Create an ad from a product description
- Explain complex data in plain English
- Easily draft documents, emails, or design

Current limitations

- Explainability of the model and results
- Hallucinations and Biases
- Data Staleness/update
- Not good at complex math and reasoning (yet)
- Not good at large scale code translation (yet)

Generative AI apps have captured public's imagination



ACCORDING TO GARTNER, INC.®

of enterprises will have used generative AI APIs or deployed generative AI-enabled apps by 2026

Gartner, "More than 80% of Enterprises," October 11, 2023.

Organizations want to enable employees with generative AI

ACCORDING TO GARTNER, INC.®

Generative AI–supported work tends to be more efficient and of higher quality than work produced by unsupported human workers

Gartner, Four GenAI Use Cases for the Digital Workplace, October 10, 2023.

PRODUCTIVITY IMPROVES MORE THAN

30%

ON AVERAGE

AWS offers a full generative AI stack of tools and services

APPLICATIONS THAT USE LLMs AND OTHER FMs

- Amazon Q Business
- Amazon Q Developer
- Amazon Q in QuickSight
- Amazon Q in Connect

TOOLS TO BUILD WITH LLMs AND OTHER FMs

Amazon Bedrock

Guardrails | Agents | Customization capabilities

INFRASTRUCTURE FOR FM TRAINING AND INFERENCE

- GPUs
- AWS Trainium
- AWS Inferentia
- Amazon SageMaker
- Amazon EC2 UltraClusters
- Elastic Fabric Adapter (EFA)
- Amazon EC2 Capacity Blocks
- AWS Nitro
- AWS Neuron

Amazon Q

Reinvent work with AWS' generative AI–powered assistant

Generally Available

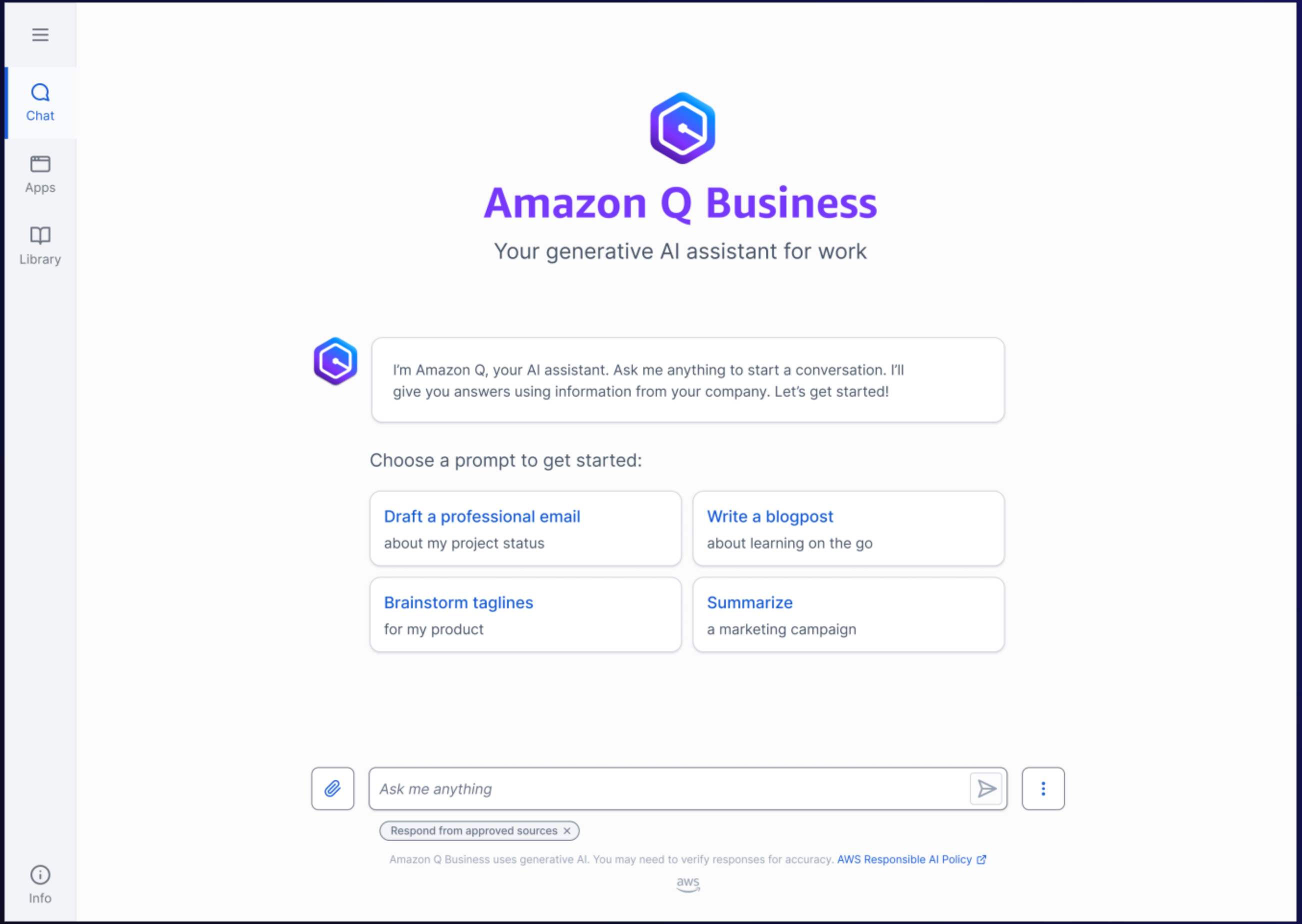
Knowledgeable of your company, code and systems

Available wherever your work

Attains superior generative AI performance on tasks

Amazon Q Business Overview

BOOST YOUR WORKFORCE PRODUCTIVITY WITH GENERATIVE AI



Amazon Q Business

Your generative AI assistant for work

I'm Amazon Q, your AI assistant. Ask me anything to start a conversation. I'll give you answers using information from your company. Let's get started!

Choose a prompt to get started:

- **Draft a professional email** about my project status
- **Write a blogpost** about learning on the go
- **Brainstorm taglines** for my product
- **Summarize** a marketing campaign

Amazon Q Business uses generative AI. You may need to verify responses for accuracy. AWS Responsible AI Policy

- Delivers quick, accurate, and relevant answers to your business questions, securely, and privately
- Execute actions using out-of-the-box or custom plugins **NEW**
- 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
- Streamlines daily tasks with user-created lightweight applications **NEW**

Top capabilities

Unified conversational search experience

Is Amazon Transcribe available in Africa?

How do I create a PO?

What is my co-pay for Lipitor?

Generate summaries and extract key insights

Provide a summary of that last interactions with Customer X

What is the 5- year CAGR?

How has revenues and margins changed over the last 3 quarters? Why?

Accelerate content creation

Provide 5 conference session titles on the topic of "Sustainable workplace"

Generate 3 social posts for the launch of Jasper

Streamline tasks

Create an application that generates customer stories based on the customer name, industry, challenges, solution and impact

Create an application that checks if a document is compliant with the guidelines available in this URL

Amazon Q use cases across the organization

Sales

Personalized outreach | Sales scripts | Sales forecasting | Customer sentiment
Intelligent search| Personalized responses and offers | Competitive analysis

Marketing

Content generation | Personalized content delivery | Insights extraction
Automated compliance check | SEO optimization | Sentiment analysis

Product management

Knowledge retrieval | Product design | Feature prioritization
Dynamic FAQ generation | Demos & tutorials | User persona creation

Engineering

Technical documentation search | Project precedents & lessons learned
Design proposals | Project performance analytics | Feedback cycles

Human resources

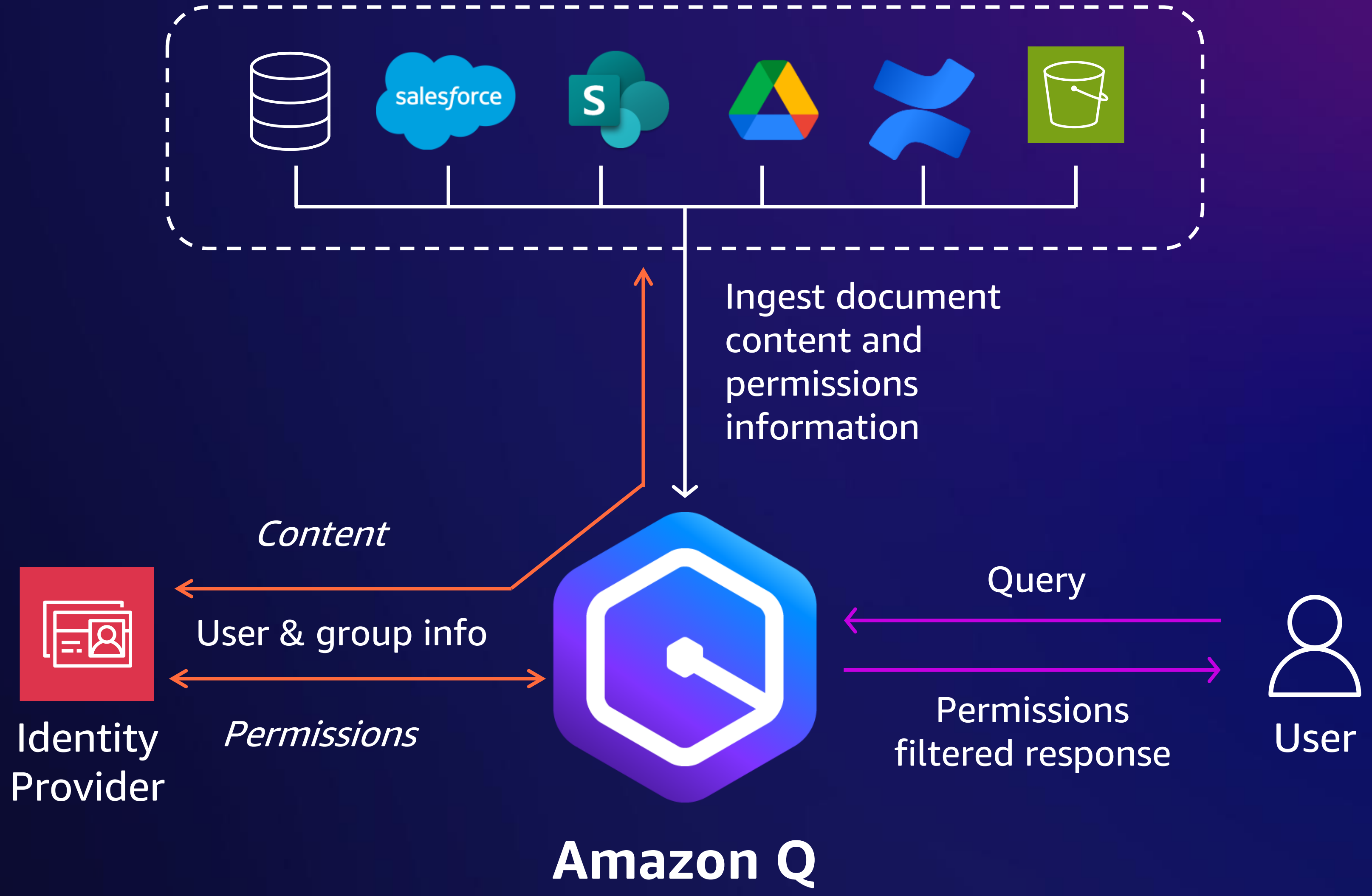
Job descriptions and postings | Personalized learning paths | Surveys
Generate surveys | Onboarding plans | Training material generation

Operations

Design dashboards and graphs | Extract and summarize insights
Trend analysis & forecasting | Policy & document drafting | Process optimization

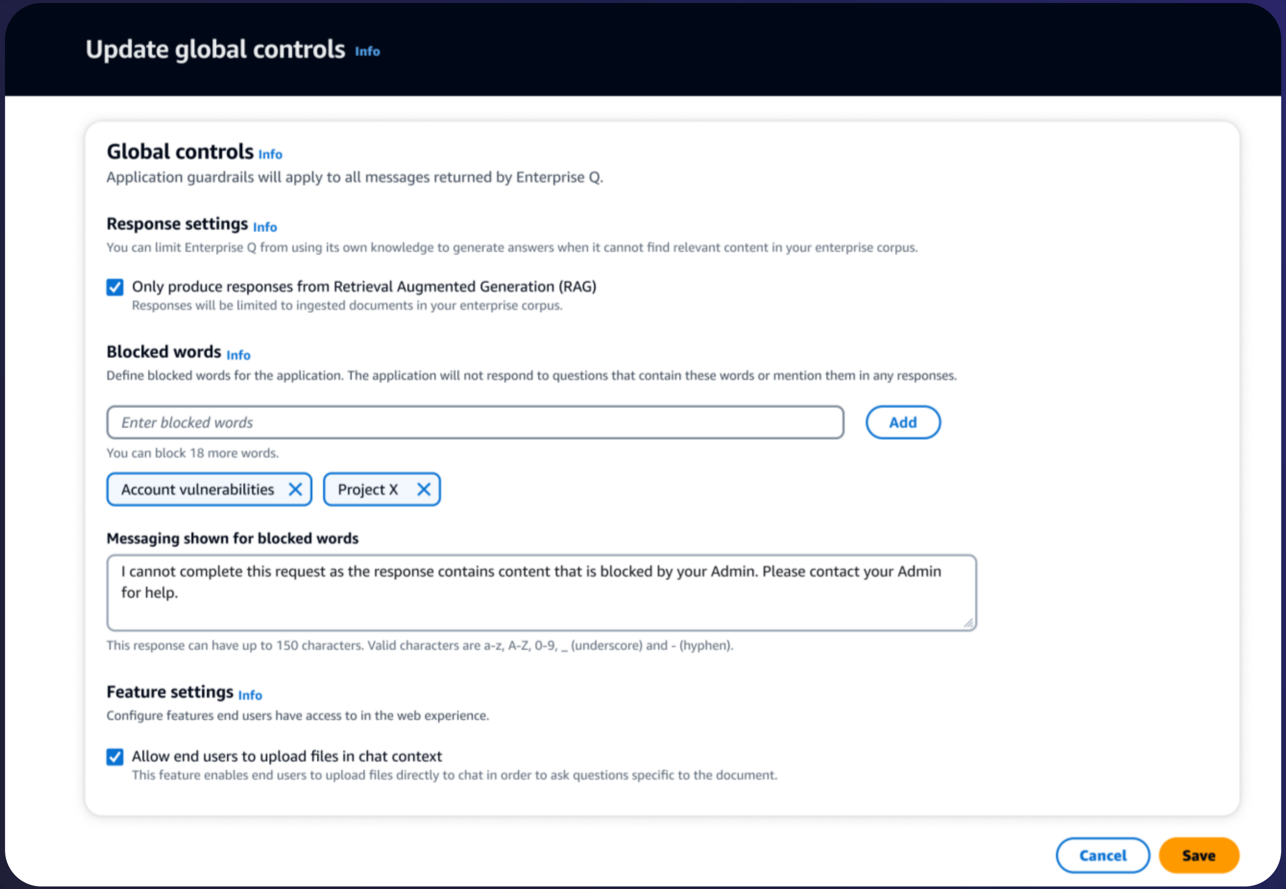
Safety and security

AMAZON Q BUSINESS IS AWARE OF ENTERPRISE USER PERMISSIONS


salesforce
S
Ingest document content and permissions information
Content
User & group info
Permissions
Identity Provider
Amazon Q
Query
Permissions filtered response
User

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 [Add]

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

Amazon Q Demos

AWS offers a full generative AI stack of tools and services

APPLICATIONS THAT USE LLMs AND OTHER FMs

- Amazon Q Business
- Amazon Q Developer
- Amazon Q in QuickSight
- Amazon Q in Connect

TOOLS TO BUILD WITH LLMs AND OTHER FMs

Amazon Bedrock

Guardrails | Agents | Customization capabilities

INFRASTRUCTURE FOR FM TRAINING AND INFERENCE

- GPUs
- AWS Trainium
- AWS Inferentia
- Amazon SageMaker
- Amazon EC2 UltraClusters
- Elastic Fabric Adapter (EFA)
- Amazon EC2 Capacity Blocks
- AWS Nitro
- AWS Neuron

Amazon Bedrock

The easiest way to build and scale generative AI applications with foundation models (FMs)

- Choice of leading FMs through a single API
- Model customization
- Retrieval Augmented Generation (RAG)
- Agents that execute multistep tasks
- Security, privacy, and safety

Amazon Bedrock simplifies

Choice

Customization

Integration

Security and governance

Amazon Bedrock

BROAD CHOICE OF MODELS

AI21 labs	amazon	ANTHROP\C	cohere	Meta	MISTRAL AI_	stability.ai
Contextual answers, summarization, paraphrasing	**Text summarization, generation, Q&A, search, image generation**	**Summarization, complex reasoning, writing, coding**	**Text generation, search, classification**	**Q&A and reading comprehension**	**Text summarization, text classification, text completion, code generation, Q&A**	**High-quality images and art**
Jamba-Instruct Jurassic-2 Ultra Jurassic-2 Mid	Amazon Titan Text Premier Amazon Titan Text Lite Amazon Titan Text Express Amazon Titan Text Embeddings Amazon Titan Text Embeddings V2 Amazon Titan Multimodal Embeddings Amazon Titan Image Generator	Claude 3.5 Sonnet Claude 3 Opus Claude 3 Sonnet Claude 3 Haiku Claude 2.1 Claude 2 Claude Instant	Command Command Light Embed English Embed Multilingual Command R+ Command R	Llama 3 8B Llama 3 70B Llama 2 13B Llama 2 70B	Mistral Small Mistral Large Mistral 7B Mixtral 8x7B	Stable Diffusion XL1.0 Stable Diffusion XL 0.8

Knowledge bases now simplifies asking questions on a single document

Ask questions and summarize data from a document, without setting up a vector database

1. Ask questions, summarize content, and more without needing to ingest data into a vector database.
2. Documents are retained only for the session. Low-cost method to use your single document for content retrieval and generation related tasks.
3. No data preparation required.

Knowledge Bases for Amazon Bedrock

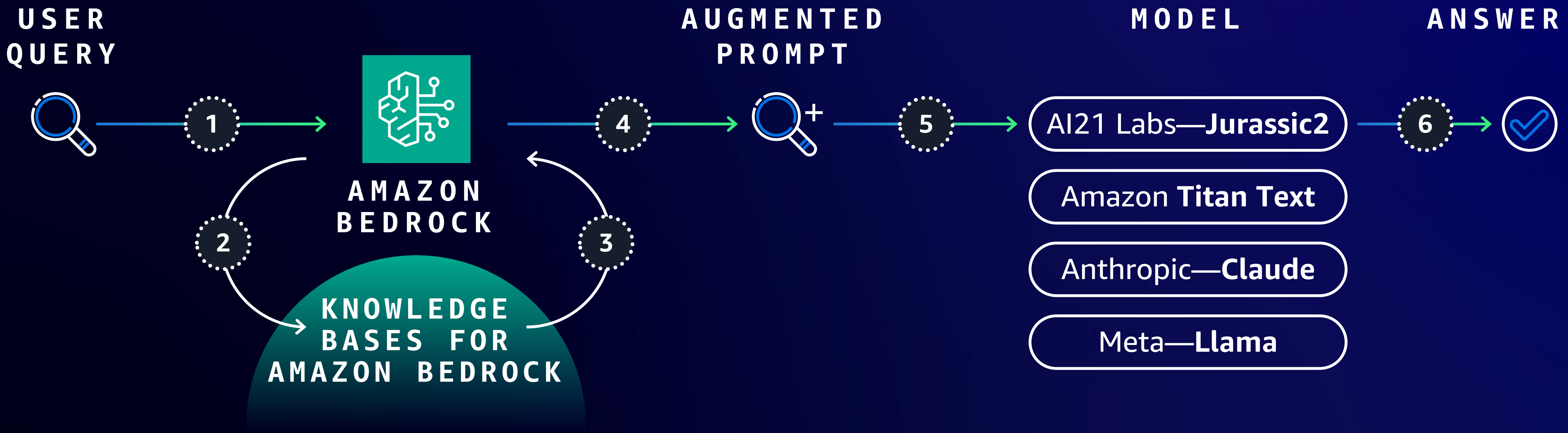
NATIVE SUPPORT FOR RAG

Securely connect FMs to data sources for RAG to deliver more relevant responses

Fully managed RAG workflow including ingestion, retrieval, and augmentation

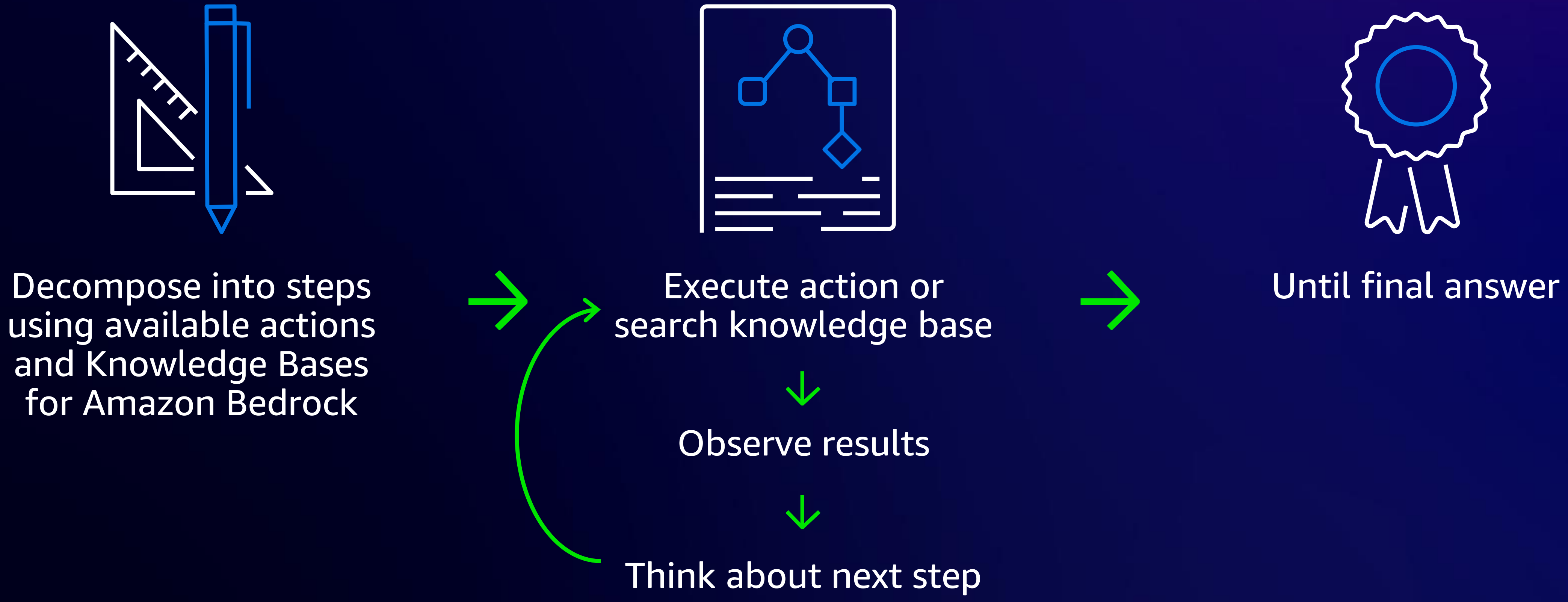
Built-in session context management for multiturn conversations

Automatic citations with retrievals to improve transparency



Agents for Amazon Bedrock

ENABLE GENERATIVE AI APPLICATIONS TO EXECUTE MULTISTEP TASKS USING COMPANY SYSTEMS AND DATA SOURCES



Decompose into steps using available actions and Knowledge Bases for Amazon Bedrock

Execute action or search knowledge base

Observe results

Think about next step

Until final answer

Guardrails for Amazon Bedrock

IMPLEMENT SAFEGUARDS CUSTOMIZED TO YOUR APPLICATION REQUIREMENTS AND RESPONSIBLE AI POLICIES

Apply guardrails to multiple foundation models and Agents for Amazon Bedrock

Configure harmful content filtering based on your responsible AI policies

Define and disallow denied topics with short natural language descriptions

Redact or block sensitive information such as PIIs, and custom Regex

Amazon Bedrock

HELPS KEEP YOUR DATA SECURE AND PRIVATE

- None of the customer's data is used to train the underlying models
- All data is encrypted in transit and at rest; data used for customization is securely transferred through customer's VPC
- Data remains in the Region where the API is processed
- Support for GDPR, SOC, ISO, CSA compliance, and HIPAA eligibility

Amazon Bedrock Demos

GenAI path to production

You need help with:	Enablement	Ideation	Model Evaluation/ Performance	POC Development
Then consider this :	Immersion days/Workshops	GenAI Factory	GenAI Tiger Team	Prototyping Team
One-liner:	A full or half day led by an AIML specialist and the AWS account team	A journey to a GenAI use case in production focusing on aligning product and tech teams towards a shared strategy	Engagement with AWS Data Scientist and AI/ML SSA to help customers solve challenges	A 3-6 weeks hands-on engagement with AWS Data Scientists to help customers build a GenAI POC

Reach out for more info:

aws-mobileye-team@amazon.com

Survey – your feedback is important !

Thank you!