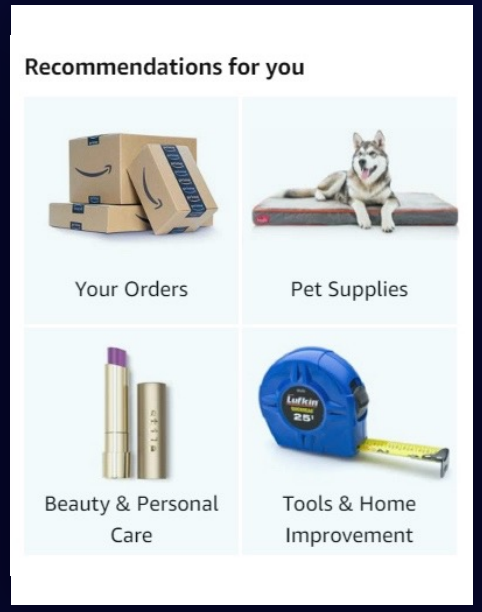
Generative AI on AWS using Amazon Bedrock

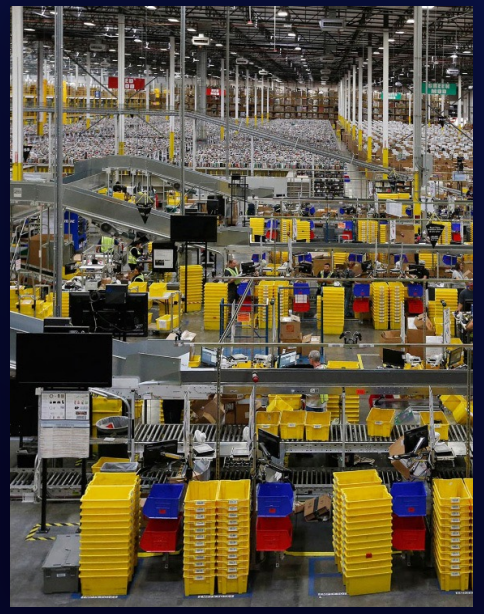
ERENAKBABA
AMAZON WEB SERVICES


ERENAKBABA

ML innovation is in Amazon's DNA



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

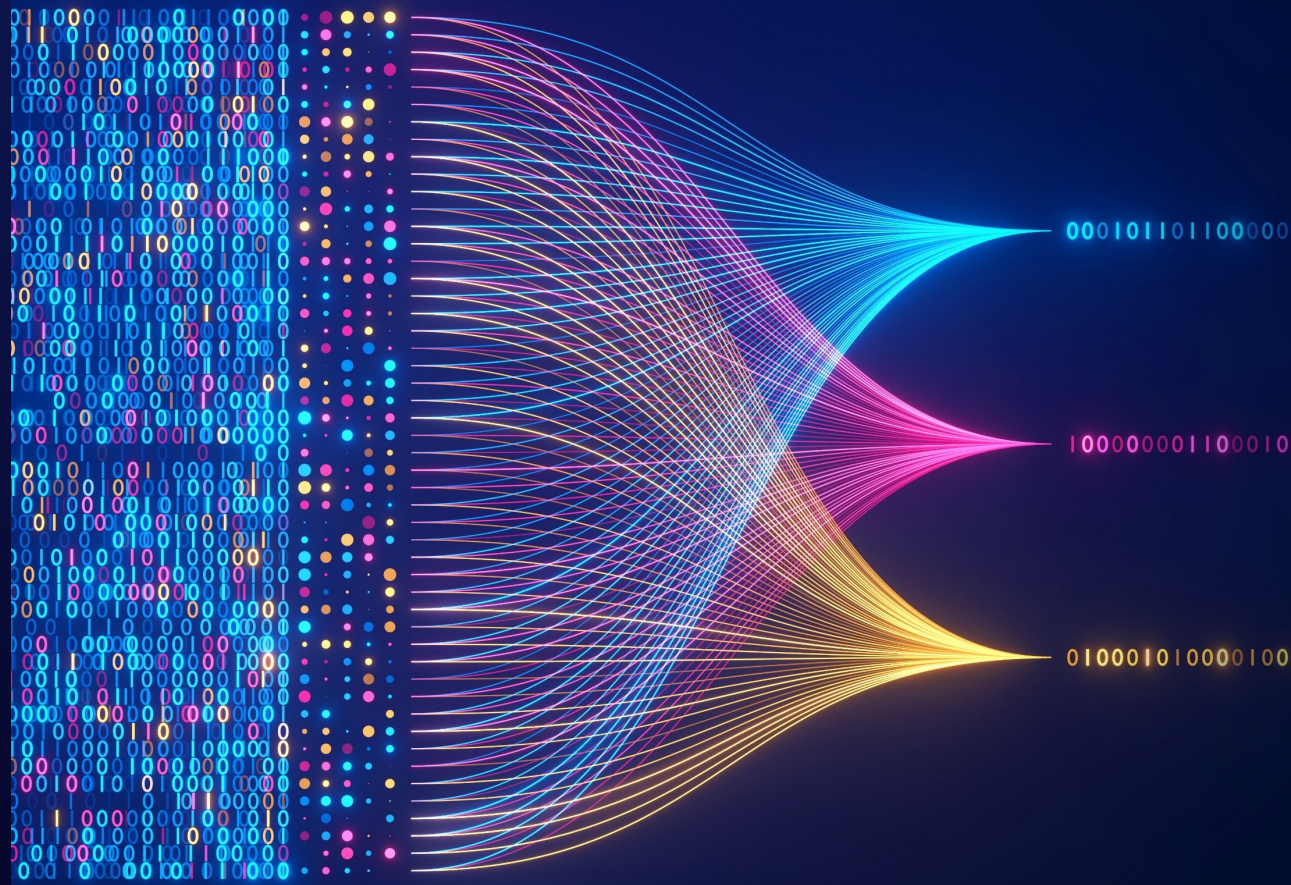
Generative AI is powered by foundation models

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



Enhance Customer Experiences

- CHATBOTS
- VIRTUAL ASSISTANTS
- CONVERSATION ANALYTICS
- PERSONALIZATION

Boost employee productivity & creativity

- CONVERSATIONAL SEARCH
- SUMMARIZATION
- CONTENT CREATION
- CODE GENERATION
- DATA TO INSIGHTS

Optimize business processes

- DOCUMENT PROCESSING
- DATA AUGMENTATION
- CYBERSECURITY
- PROCESS OPTIMIZATION

Healthcare & Life Sciences

- Ambient digital scribe
- Medical imaging
- Drug discovery
- Enhance clinical trials
- Research reporting

Industrial & Manufacturing

- Product design
- Operational efficiency
- Maintenance Assistants
- Supply chain optimization
- Equipment diagnostics

Financial Services

- Portfolio management
- Financial documentation
- Intelligent advisory
- Fraud detection
- Compliance assistant

Retail

- Pricing optimization
- Virtual try-ons review
- Marketing Optimization
- Product descriptions
- Pers. Recommendations

Media & Entertainment

- HQ content at scale
- Enrich broadcast content
- Automated content tagging
- Optimize subscriber exper.
- Automated highlights gen.

Generative AI Stack

APPLICATIONS THAT LEVERAGE LLMs AND OTHER FMs

TOOLS TO BUILD WITH LLMs AND OTHER FMs

INFRASTRUCTURE FOR FM TRAINING AND INFERENCE

Generative AI Stack

APPLICATIONS THAT LEVERAGE LLMs AND OTHER FMs

TOOLS TO BUILD WITH LLMs AND OTHER FMs

INFRASTRUCTURE FOR FM TRAINING AND INFERENCE

CG1	G2	P2	G3	P3	G4	P4	G5	G5g	P5
NVIDIA Tesla M2050 “Fermi” GPUs	NVIDIA GRID GK104 “Kepler” GPUs	NVIDIA K80 GPUs	NVIDIA Tesla M60 GPUs	NVIDIA V100 Tensor Core GPUs	NVIDIA T4 Tensor Core GPUs	NVIDIA A100 Tensor Core GPUs	NVIDIA A10G Tensor Core GPUs	NVIDIA T4G Tensor Core GPUs	NVIDIA H100 Tensor Core GPUs

4x

HIGHER THROUGHPUT

10x

LOWER LATENCY

Amazon SageMaker

Build, train, and deploy ML models at scale

Automatic model fine-tuning & distributed training

Flexible model deployment options

Tools for ML operations

Built-in features for responsible AI

Tens of thousands of customers

Used to train Falcon 180 billion model

Support for Hugging Face models and Hugging Face AWS Deep Learning Containers

Generative AI Stack

APPLICATIONS THAT LEVERAGE LLMs AND OTHER FMs

TOOLS TO BUILD WITH LLMs AND OTHER FMs

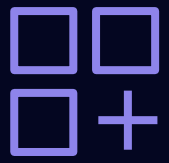
INFRASTRUCTURE FOR FM TRAINING AND INFERENCE

 GPUs  Trainium  Inferentia  SageMaker

 UltraClusters  EFA  EC2 Capacity Blocks  Nitro  Neuron

aws

Customers have questions…



Which model should I use?

How can I move quickly?

How can I keep my data secure & private?

Generative AI Stack

APPLICATIONS THAT LEVERAGE LLMs AND OTHER FMs

TOOLS TO BUILD WITH LLMs AND OTHER FMs

INFRASTRUCTURE FOR FM TRAINING AND INFERENCE

GPUs Trainium Inferentia SageMaker

UltraClusters EFA EC2 Capacity Blocks Nitro Neuron

aws

Amazon Bedrock

The easiest way to build and scale generative AI applications with LLMs and other FMs

Choice of industry-leading FMs from AI21 Labs, Amazon, Anthropic, Cohere, Meta, Mistral AI, and Stability AI

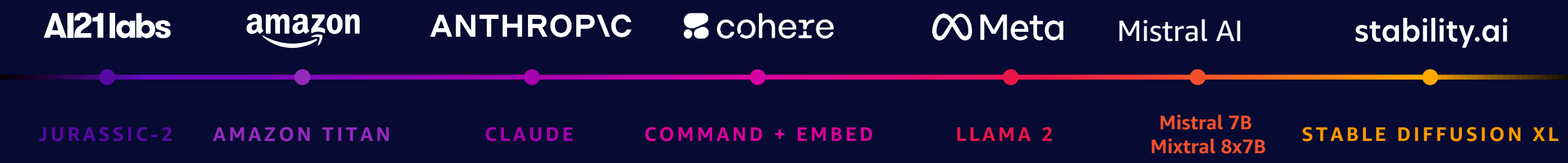
Customize FMs using your organization's data

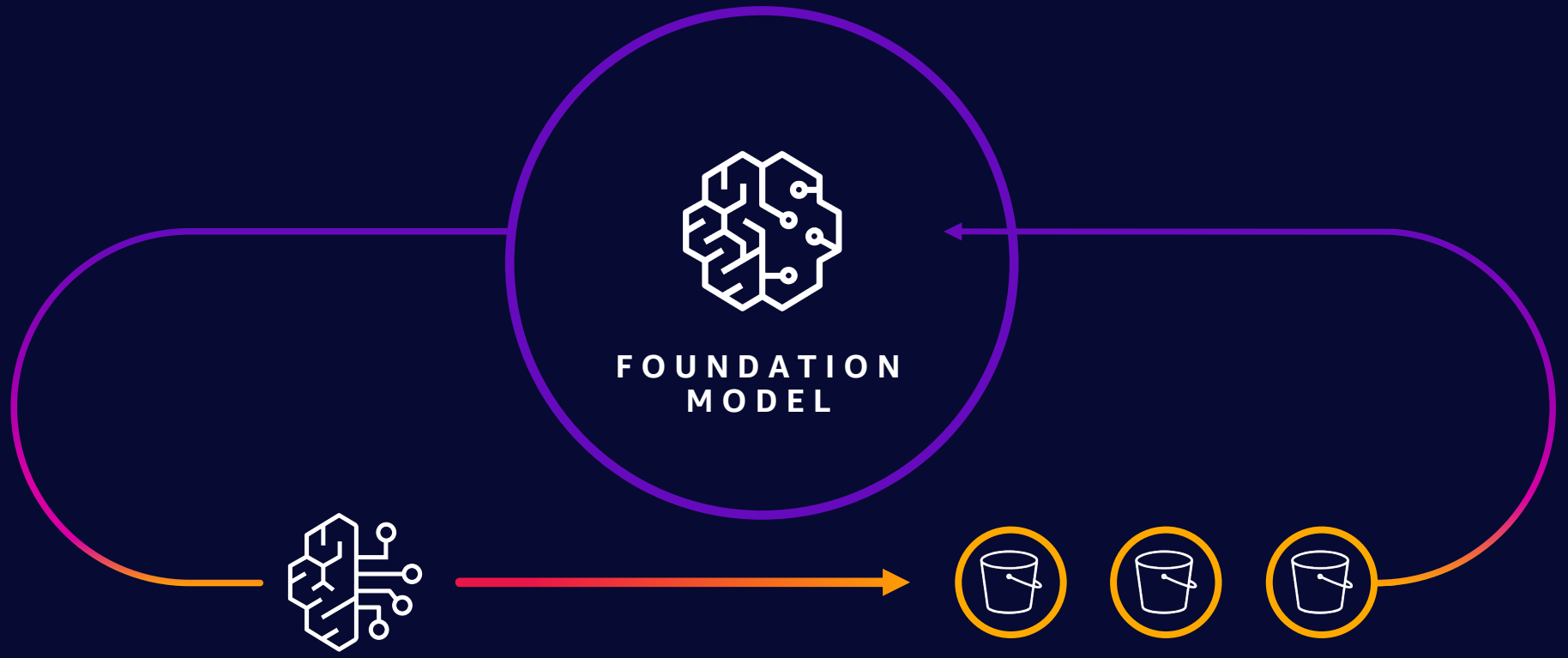
Enterprise-grade security and privacy


adidas
Adobe
ADP
Asahi
ATLASSIAN
AstraZeneca
BBVA
BMW GROUP
Booking.com
BRIDGEWATER
Broadridge
Canva
Carrier
Chegg
cimpress
CLARIANT
coda
Cox AUTOMOTIVE
coinbase
druva
DUKE ENERGY
ERICSSON
Expedia
Fidelity INVESTMENTS
FOX
Genesys
Georgia-Pacific
GoDaddy
Grab
government technology
happyfox
hellmann
WORLDWIDE LOGISTICS
ifood
KONE
LaunchDarkly
LexisNexis
lonely planet
MarathonOil
MERCK
Nasdaq
NatWest
nexxiot
NOMURA
OmnicomGroup
PHILIPS
proofpoint
RareJob
RICOH
imagine. change.
salesforce
serasa experian
SIEMENS
Schneider Electric
Showpad
smartsheet
STELLANTIS
Sun Life
TAKENAKA
Vanguard
THOMSON REUTERS
twilio
UNITED
VERITONE
WPS
yahoo!

Amazon Bedrock

Broad choice of models





Fine tuning | Retrieval Augmented Generation (RAG) | Continued Pre-training

NEW

Agents for Amazon Bedrock

Execute multi-step tasks across company systems and data sources

GENERALLY AVAILABLE

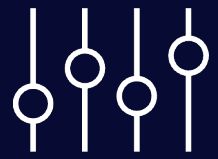
Enables generative AI applications to take action in just a few clicks

Breaks down and orchestrates tasks and executes API calls on your behalf

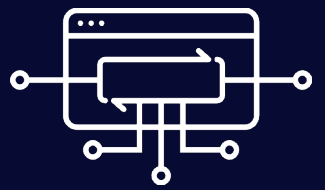
Securely accesses and retrieves company data

Amazon Bedrock simplifies

Choice



Customization



Integration

Amazon Bedrock keeps data secure & private

None of the customer's data is used to train the underlying model

All data is encrypted in transit and at rest

Data used to customize models remains within your VPC

Support for standards, including GDPR & HIPAA

Amazon Bedrock

Recently added security capabilities

CloudWatch integration

Track usage metrics and build customized dashboards

CloudTrail integration

Monitor API activity and troubleshoot issues

SOC compliance

SOC 1, 2 & 3

NEW

Guardrails for Amazon Bedrock

Safeguard your generative AI applications with your responsible AI policies

AVAILABLE IN PREVIEW

Easily configure harmful content filtering based on your responsible AI policies

Apply Guardrails to any FM or agent

Redact PII information in FM responses (coming soon)

Generative AI Stack

APPLICATIONS THAT LEVERAGE LLMs AND OTHER FMs

TOOLS TO BUILD WITH LLMs AND OTHER FMs

Amazon Bedrock

Guardrails | Agents | Customization Capabilities

INFRASTRUCTURE FOR FM TRAINING AND INFERENCE

GPUs | Trainium | Inferentia | SageMaker

UltraClusters | EFA | EC2 Capacity Blocks | Nitro | Neuron

Generative AI Stack

APPLICATIONS THAT LEVERAGE LLMs AND OTHER FMs

TOOLS TO BUILD WITH LLMs AND OTHER FMs

Amazon Bedrock

Guardrails | Agents | Customization Capabilities

INFRASTRUCTURE FOR FM TRAINING AND INFERENCE

GPUs | Trainium | Inferentia | SageMaker

UltraClusters | EFA | EC2 Capacity Blocks | Nitro | Neuron

Amazon CodeWhisperer

AI-powered code suggestions in the IDE and the command line

Amazon CodeWhisperer customization capability

Better and more relevant code suggestions

Your data is never used to train the underlying model

Onboard developers faster

Amazon CodeWhisperer customization capability

Amazon CodeWhisperer

AMAZON CODE

PUBLICLY AVAILABLE CODE

YOUR CODE

YOUR APIs

```
const apiUrl = 'https://shoppingCartAPI.exec
shoppingCartAPI(apiUrl)

  .then((data) => {
    console.log('API response:', data);

  })
  .catch((error) => {
    console.error('Error:', error);

  });
```

More relevant suggestions; better code

28%

faster

CODEWHISPERER
CUSTOMIZATIONS

NEW

Amazon Q

A generative AI-powered assistant for work that is tailored to your business

Provides interactive answers, solves problems, generates content, and takes action

Understands your company information, code, and systems

Personalizes interactions based on your role and permissions

Built to be secure and private

Amazon Q

Your expert assistant for building on AWS

Trained on 17 years of AWS knowledge

Assists you everywhere you work with AWS – in the console, IDE, and documentation

Converses with you to explore new AWS capabilities, learn unfamiliar technologies, and architect solutions

Works with you to troubleshoot, build new features, and upgrade languages on AWS

Amazon Q knocks down obstacles for developers

Amazon Q

Your business expert

Delivers quick, accurate, and relevant answers to your business questions, securely and privately

Connects to over 40 popular data sources including S3, Salesforce, Google Drive, Microsoft 365, ServiceNow, Gmail, Slack, Atlassian, and Zendesk

Respects existing access controls - only returns info you're authorized to see based on your role

Amazon Q

in Amazon QuickSight

Generative dashboard authoring

Visually compelling data stories

Reimagined Q&A experience

Amazon Q is

AMAZON Q
YOUR AWS EXPERT

AMAZON Q IN AMAZON QUICKSIGHT
YOUR BI EXPERT

AMAZON Q
YOUR BUSINESS EXPERT

AMAZON Q IN AMAZON CONNECT
YOUR CONTACT CENTER EXPERT

Generative AI Stack

APPLICATIONS THAT LEVERAGE LLMs AND OTHER FMs

Amazon Q | Amazon Q in Amazon QuickSight | Amazon Q in Amazon Connect | Amazon CodeWhisperer

TOOLS TO BUILD WITH LLMs AND OTHER FMs

Amazon Bedrock

Guardrails | Agents | Customization Capabilities

INFRASTRUCTURE FOR FM TRAINING AND INFERENCE

GPUs | Trainium | Inferentia | SageMaker

UltraClusters | EFA | EC2 Capacity Blocks | Nitro | Neuron

aws


aws