Mastering Generative AI: Harnessing AWS GenAI for Your Solutions

SAMUEL BARUFFI

SENIOR SOLUTIONS ARCHITECT

Agenda

Generative AI Fundamentals

Tranium and Inferentia

Amazon SageMaker

Amazon Bedrock

Vector Databases

Amazon CodeWhisperer

Demo

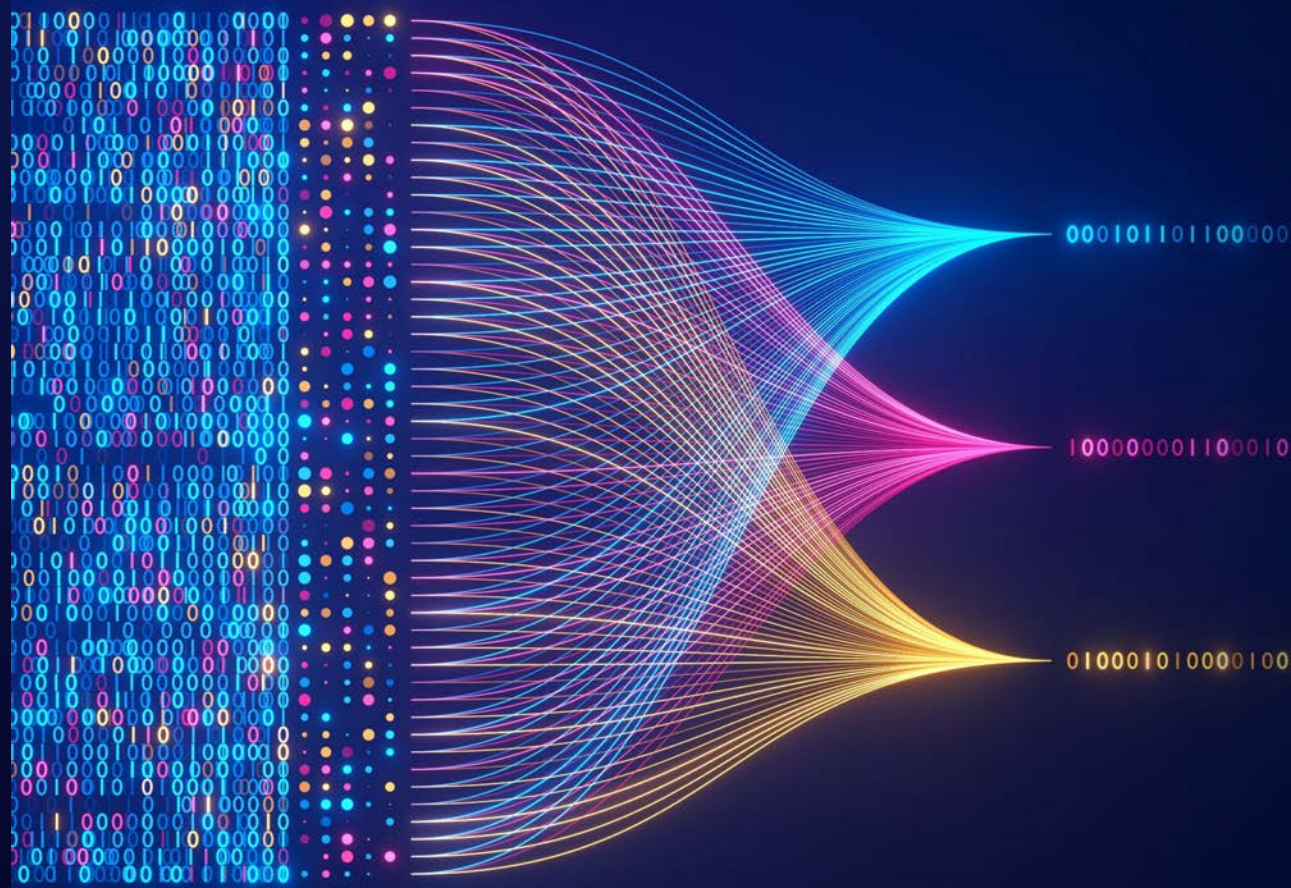
Generative AI is powered by foundation models

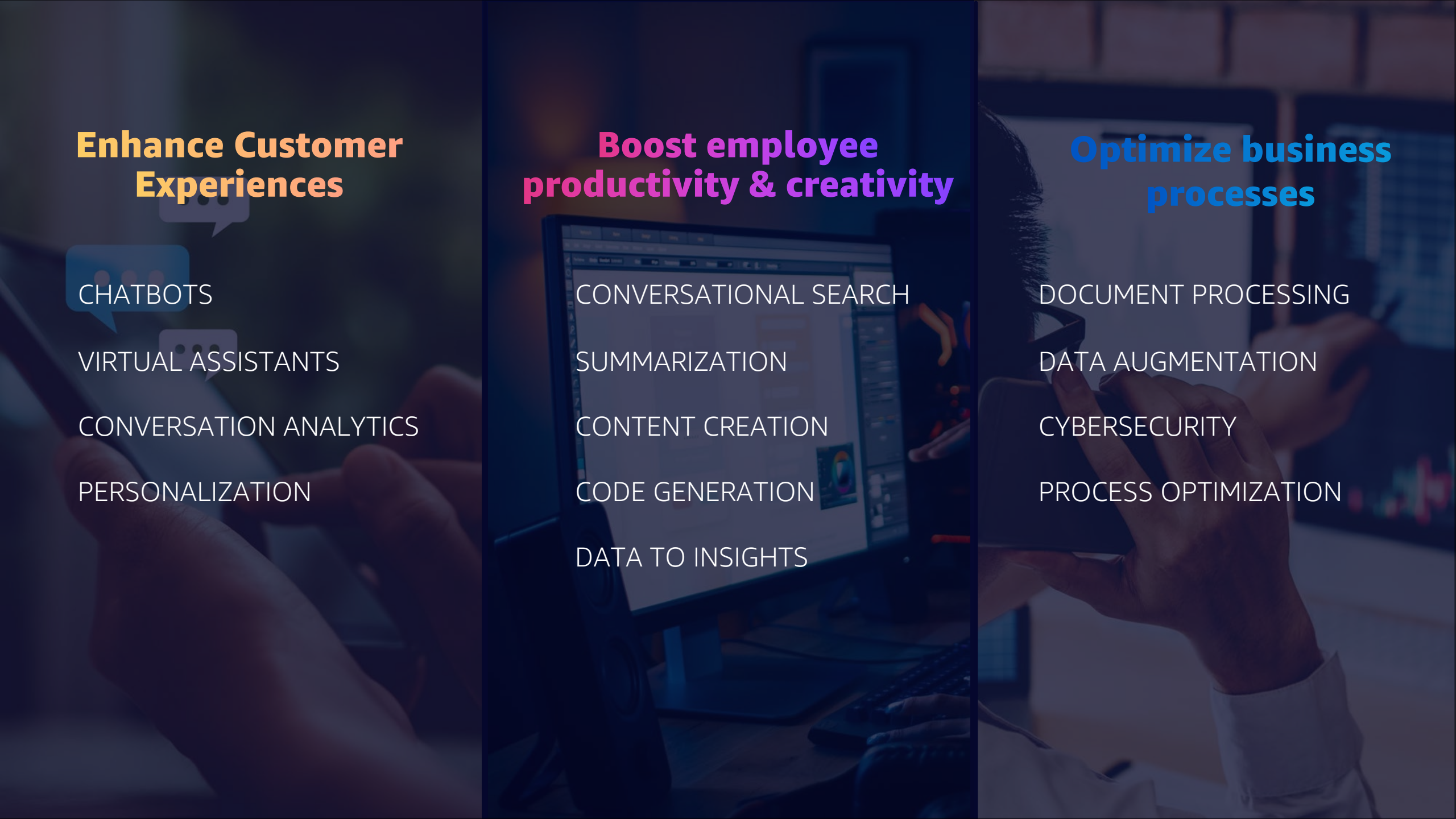
Pretrained on vast amounts of unstructured data

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

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

Innovating at the silicon level

AWS Trainium

AWS Inferentia

Amazon SageMaker

Build, train, and deploy ML models at scale

Automatic model fine-tuning & distributed training

Flexible model deployment options

Tools for ML operations

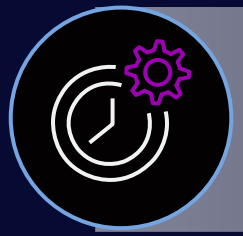
Built-in features for responsible AI

Amazon SageMaker JumpStart

ML hub with foundation models, built-in algorithms, and prebuilt ML solutions that you can deploy with just a few clicks

Machine learning hub

Browse through 400+ built-in algorithms with pretrained foundation models, solutions, and example notebooks



Pre-built training and inference scripts

Compatible with SageMaker and configurable with custom dataset

UI as well as API-based

Use the user interface for single click model deployment or API for the Python SDK-based workflow

Notebooks with examples

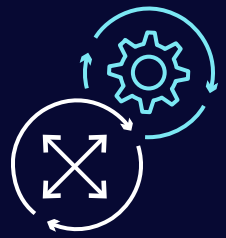
Jump into notebooks to use selected model with examples to guide you through the entire ML workflow

Share and collaborate within your organization

Share models and notebooks with others within your organization, and allow them to train with their own data or deploy as-is for inferencing

What generative AI customers are asking for

Which model should I use?



How can I move quickly?

How can I keep my data secure and private?

Amazon Bedrock

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

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

Amazon Bedrock supports leading foundation models

AI21labs

Jurassic-2

Contextual answers, summarization, paraphrasing

ANTHROP\C

Claude 3, Claude 2.1 & Claude Instant

Multimodal reasoning, vision capabilities, translation, summarization, writing, coding

cohere

Command & Embed

Text generation, search, classification

Mistral AI

Mistral 7B & Mixtral 8x7B

Text summarization, Q&A, Text classification, Text completion, code generation

Meta

Llama 2

Dialogue use cases and language tasks

stability.ai

Stable Diffusion XL 1.0

High-quality images and art

amazon

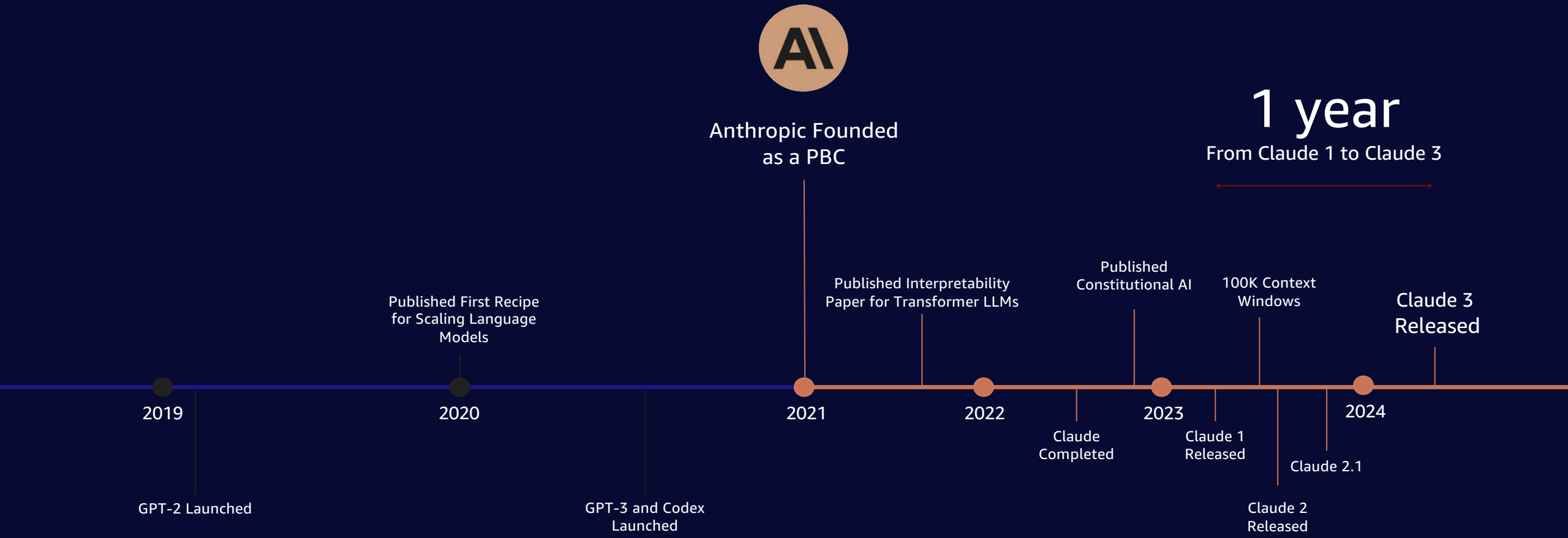
Amazon Titan

Summarization, image and text generation and search, Q&A

aws

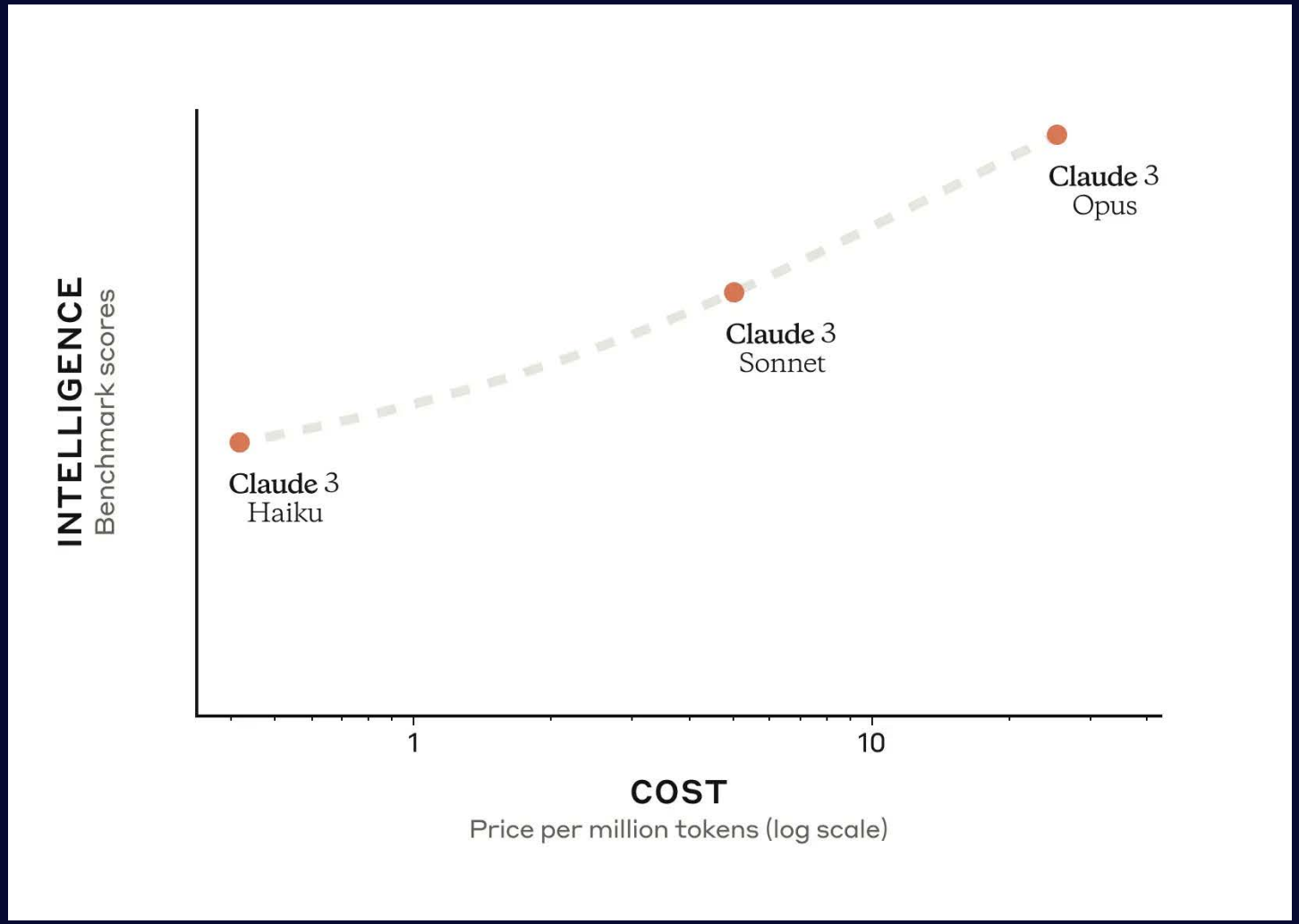

amazon
ANTHROP\C

Anthropic’s breakthroughs are setting the pace for the AI industry



Introducing Claude 3

Leading the frontier of **speed**, **intelligence,** and **cost-efficiency** for generative AI



Claude 3 family on Amazon Bedrock

CHOOSE THE EXACT COMBINATION OF INTELLIGENCE, SPEED, AND COST TO SUIT YOUR NEEDS.

	Coming soon	Now available on Amazon Bedrock	Coming soon
	Claude 3 Opus	**Claude 3 Sonnet**	**Claude 3 Haiku**
Use case	Most intelligence and highest performance	Balance between intelligence, speed and cost	Fastest performance at the lowest cost
Context	200K	200K	200K
Vison	✓	✓	✓

Privately customize models with your data

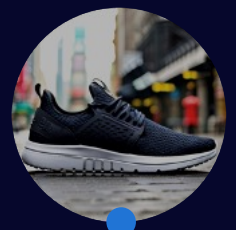
FINE-TUNING AND CONTINUED PRE-TRAINING

Deliver tailored, differentiated tail user experiences with customized FMs

Fine-tune Llama 2, Command, and Titan FMs for specific tasks with labeled data

Use continued pre-training to adapt Titan Text FMs to your domain with unlabeled data

None of your inputs to or outputs from Amazon Bedrock will be used to train the original base models



Knowledge bases for Amazon Bedrock

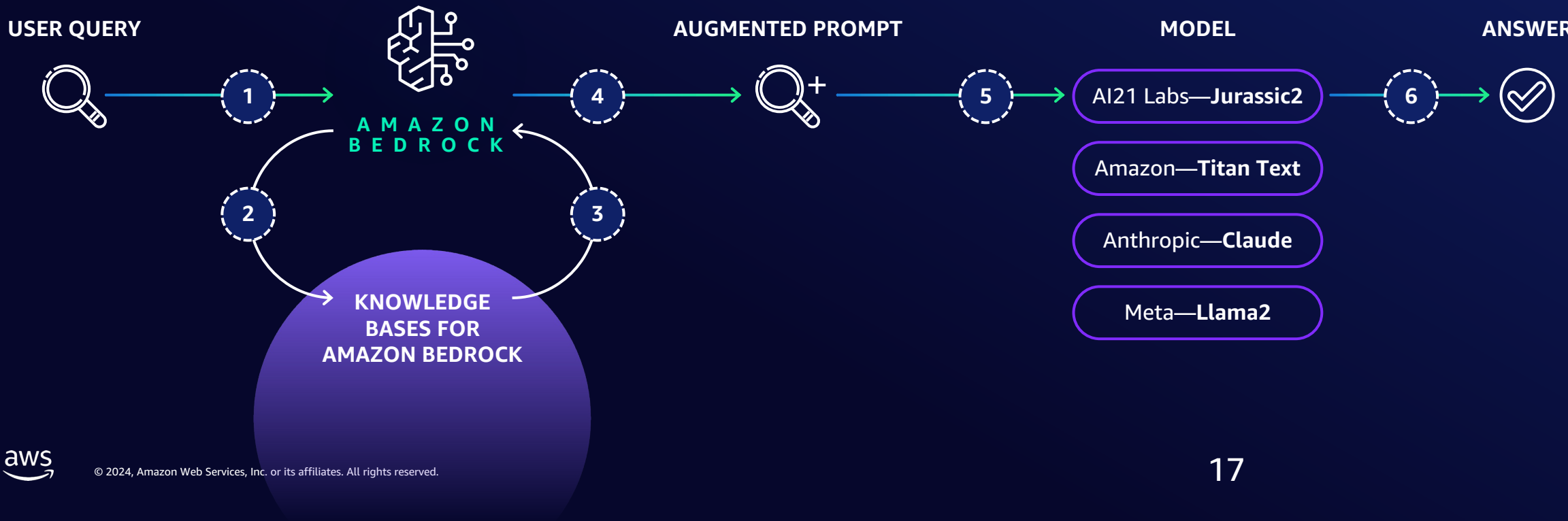
NATIVE SUPPORT FOR RETRIEVAL AUGMENTED GENERATION (RAG)

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 multi-turn 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

1

SELECT YOUR FOUNDATION MODEL

2

PROVIDE BASIC INSTRUCTIONS

3

SELECT RELEVANT DATA SOURCES

4

SPECIFY AVAILABLE ACTIONS

- Breaks down and orchestrates tasks
- Securely accesses and retrieves company data for RAG
- Takes action by invoking API calls on your behalf
- Chain-of-thought trace and ability to modify agent prompts

Guardrails for Amazon Bedrock

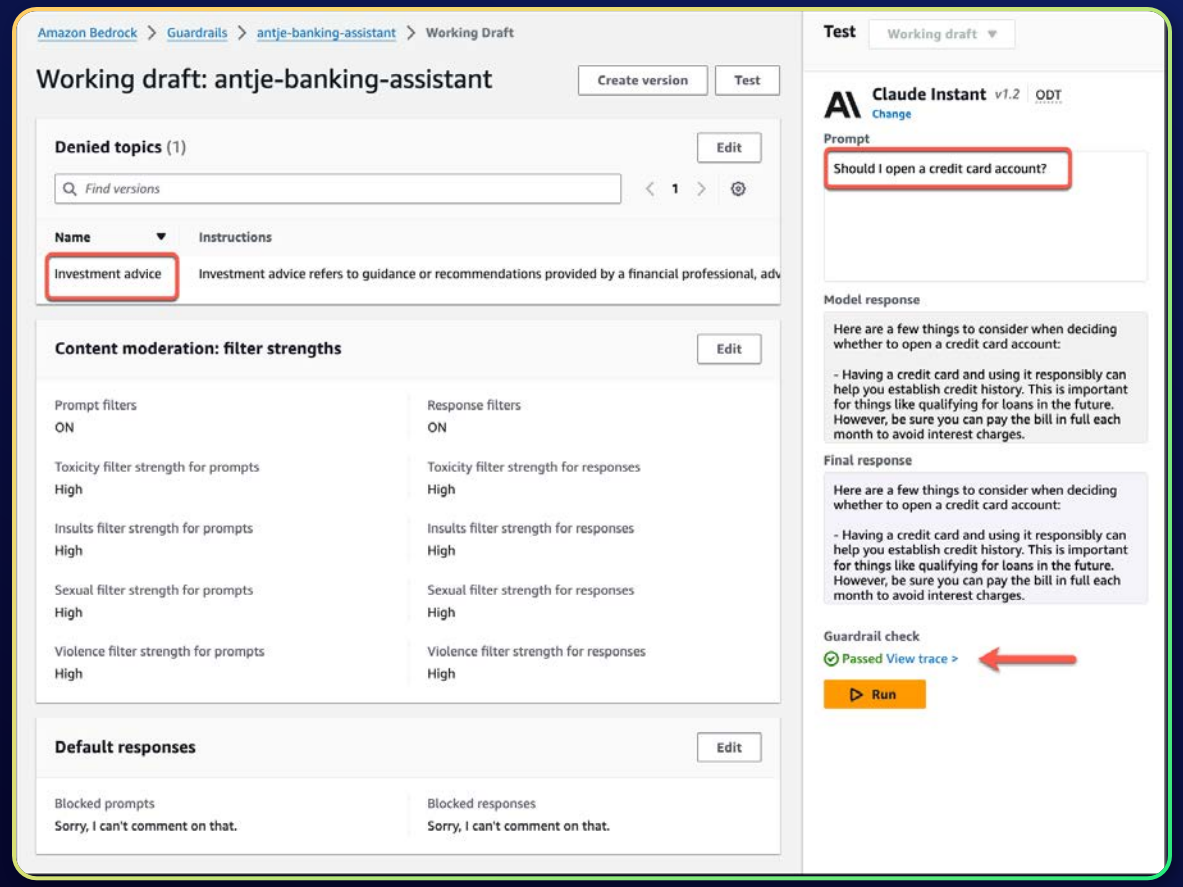
IMPLEMENT SAFEGUARDS TAILORED TO YOUR APPLICATION REQUIREMENTS AND RESPONSIBLE AI POLICIES

Preview

Apply guardrails consistently across FMs including fine-tuned models and agents

Configure filtering of harmful content and topics to avoid based on your responsible AI policies

Redact personally identifiable information (coming soon)



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

Support for GDPR, SOC, ISO, CSA compliance and HIPAA eligibility

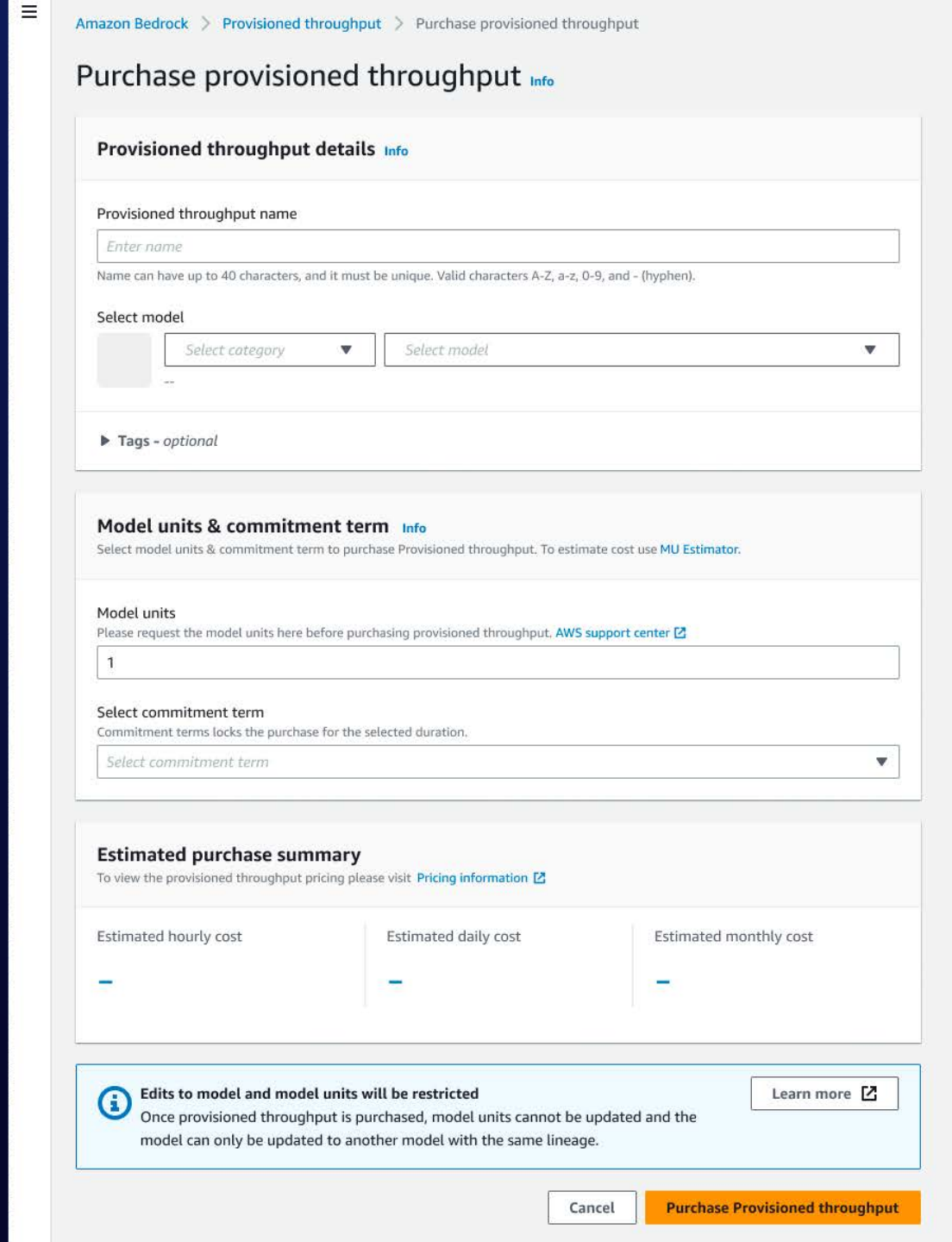
Provisioned throughput

Reserve throughput (input/output tokens per minute)

Ensure consistent user experience during traffic spikes

Purchase with commitment term of one month or six months

Pay hourly rate, discounted for extended commitment



Batch mode (preview)

Efficiently run model inference on large volumes of data

Avoids throttling when running large jobs

Fully-managed model invocation jobs

No need to write code to handle failures and restarts

Works with base and custom models

Batch mode (preview)

Efficiently run model inference on large volumes of data

- Avoids throttling when running large jobs
- Fully-managed model invocation jobs
- No need to write code to handle failures and restarts
- Works with base and custom models

Model Evaluation on Amazon Bedrock

EVALUATE FMS TO SELECT THE BEST ONE FOR YOUR USE CASE

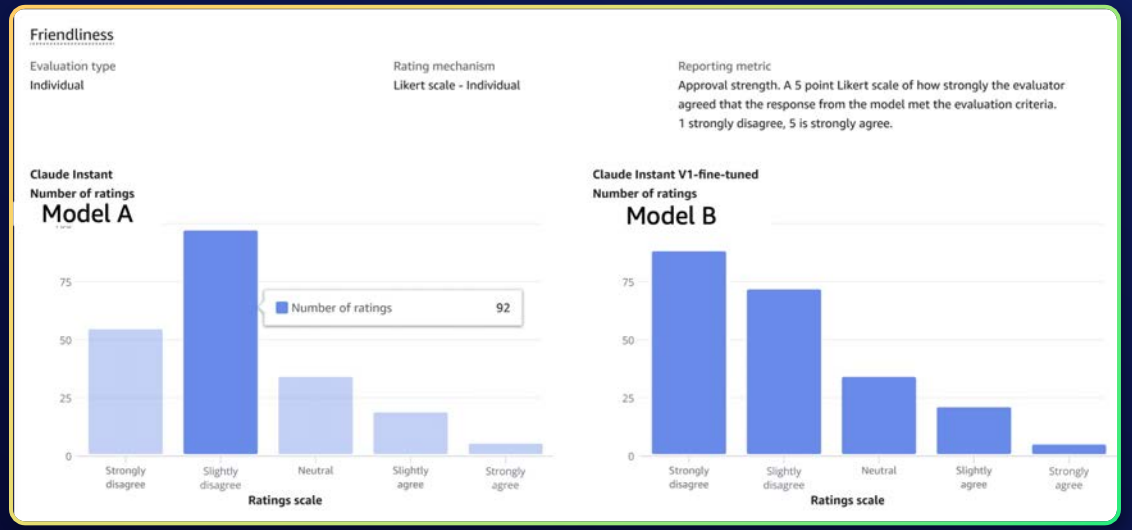
Preview

Choose automatic or human evaluation method

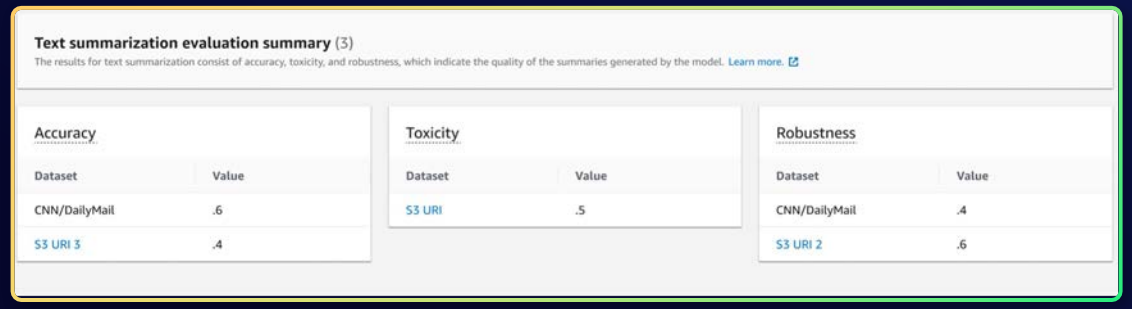
Curated datasets or bring your own

Pre-defined and custom metrics

Human evaluation report



Automatic evaluation report




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
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!

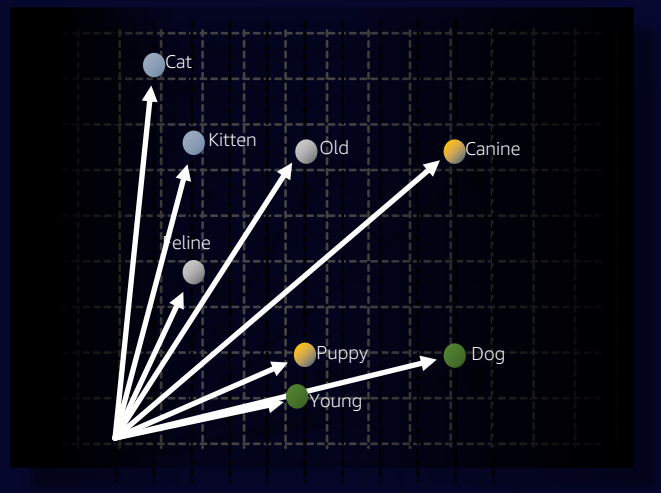
Vector Database Support on AWS

BUILDING GENAI APPLICATIONS OFTEN REQUIRES VECTOR SEARCH AND AWS IS EXTENDING SUPPORT

Embeddings encode all data types into vectors that capture meaning and context of an asset.

Many Generative AI models depend on reading embeddings data from a vector database

Vectors are critical for customizing generative AI applications



Vector Support for Amazon Databases

- DocumentDB NEW
- MemoryDB for Redis Preview
- Amazon RDS PostgreSQL
- Aurora PostgreSQL
- Amazon Neptune
- OpenSearch

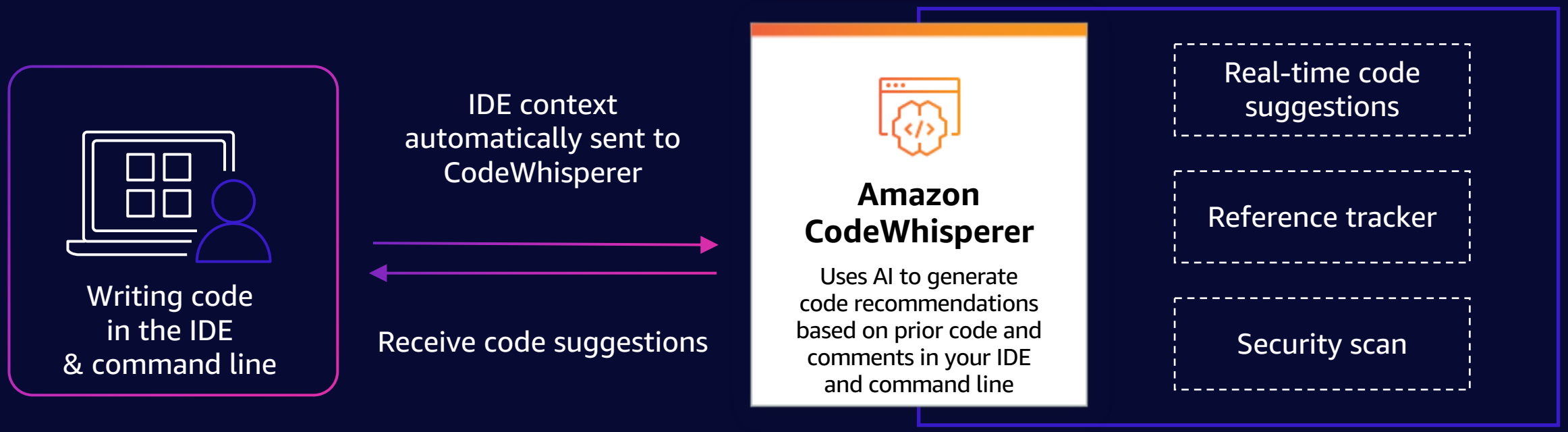
Vector Database Direct Integration for Amazon Bedrock

- Vector Engine For OpenSearch Serverless
- Redis Enterprise Cloud
- Pinecone
- COMING SOON — Amazon Aurora
- COMING SOON — MongoDB

CodeWhisperer

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

How it works



Content processed by CodeWhisperer Professional is not stored or used for service improvement

AWS Generative AI Demo