



# Amazon Bedrock

## Workshop

김정곤

솔루션즈 아키텍트  
AWS Korea

# Agenda

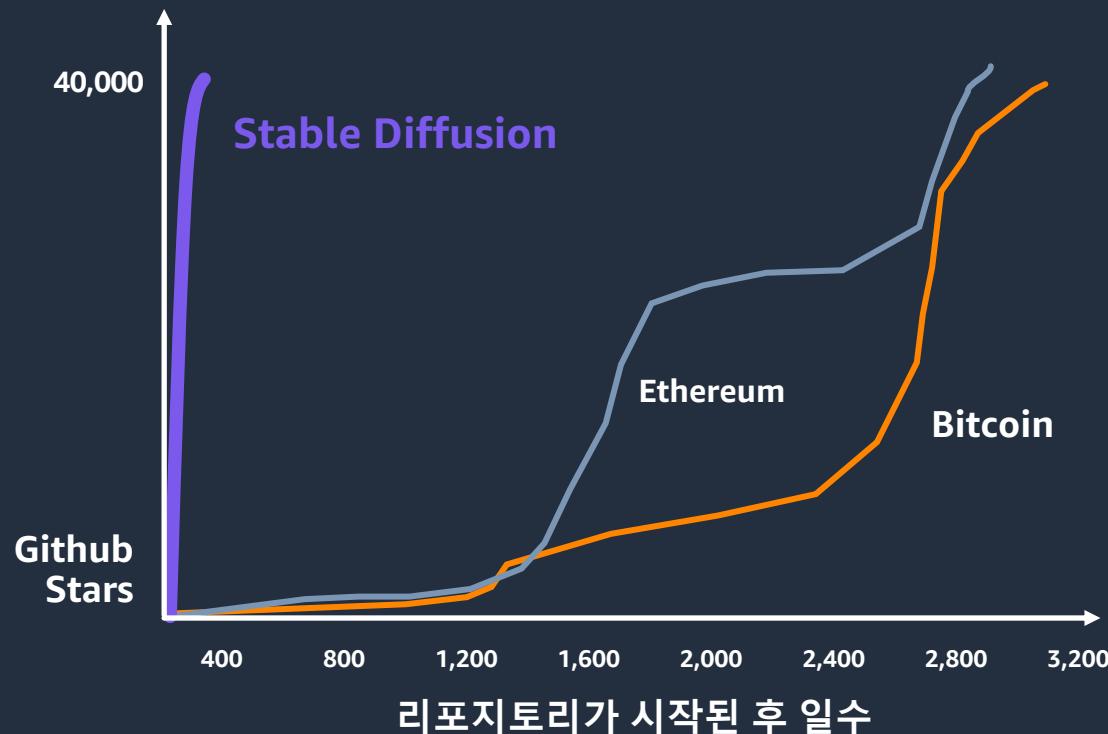
- What is Generative AI?
- Amazon Bedrock
- Architecture Patterns
- Security
- Extras/next step

# What is Generative AI?

# Generative AI is the fastest growing trend in AI

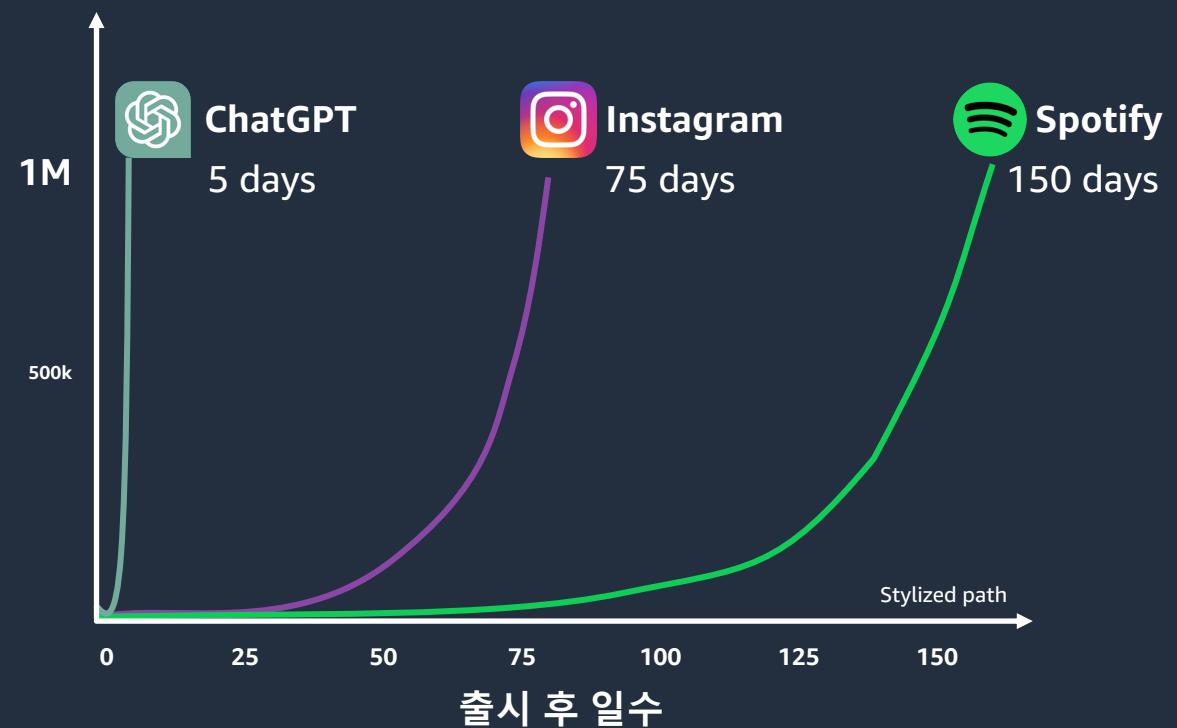
Developer adoption

Stable Diffusion은 첫 90일 동안  
GitHub에서 4만 개의 별을 누적했습니다.



Consumer adoption

단 5일 만에 ChatGPT 사용자 수  
100만 명 돌파



# What is generative AI?



실제 업무에서 사람이 만든 콘텐츠에 가까운 오리지널 콘텐츠를 제작할 수 있는 AI



수천억 개의 매개 변수가 포함된 대규모 데이터 세트에 대해 사전 학습된 기초 모델을 기반으로 합니다.



최소한의 파인 투닝으로 특정 도메인에 맞게 작업을 사용자 지정할 수 있습니다.



텍스트 요약, 질문 답변, 디지털 아트 제작, 코드 생성 등과 같은 다양한 사용 사례에 적용할 수 있습니다.



ML 모델 개발에 소요되는 시간과 비용 절감 및 혁신 속도 향상

# 다양한 영역에서 생성형 AI가 등장하고 있습니다 ...



## 고객 경험 향상

CHATBOTS

VIRTUAL ASSISTANTS

AI-POWERED CONTACT CENTER

PERSONALIZATION



## 직원 생산성 향상

CONVERSATIONAL SEARCH

SUMMARIZATION

CODE GENERATION

DATA TO INSIGHTS



## 창의성 및 콘텐츠 제작

WRITING

MEDIA

DESIGN

MODELING



## 비즈니스 운영 개선

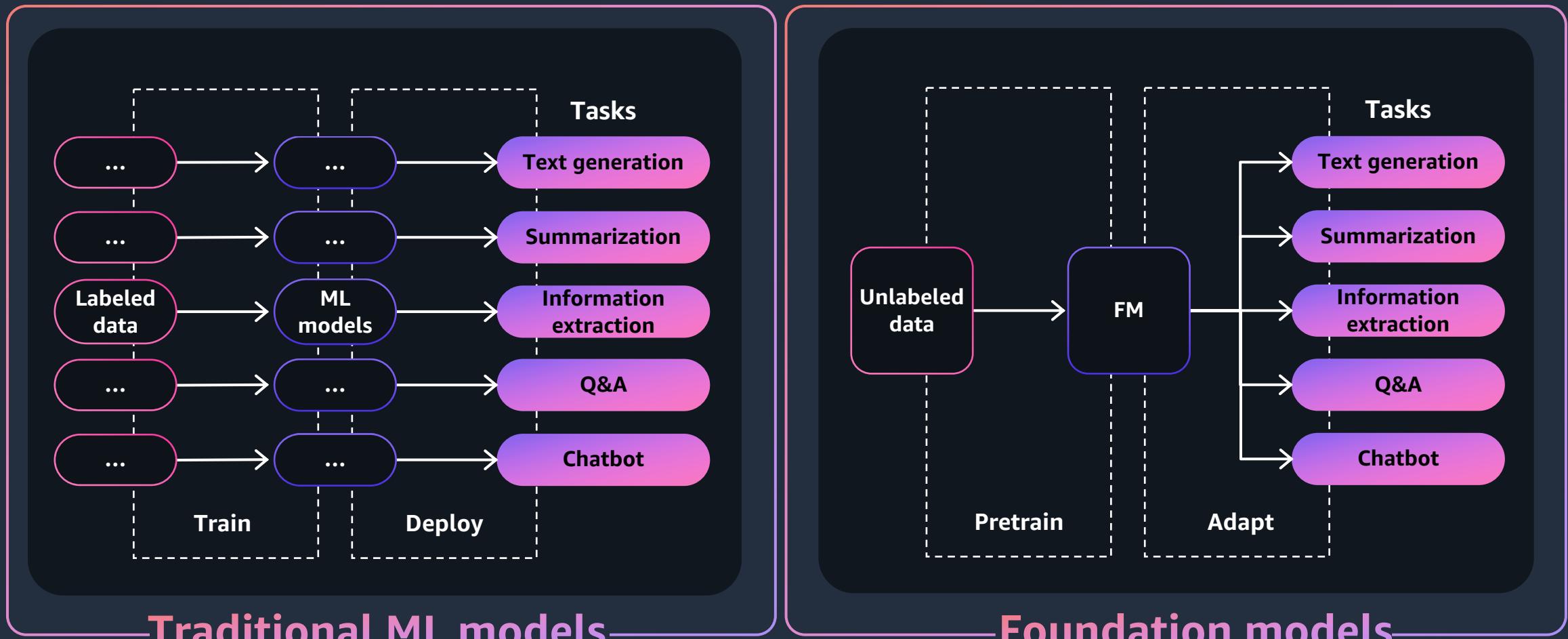
DOCUMENT PROCESSING

PROCESS OPTIMIZATION

CYBERSECURITY

DATA AUGMENTATION

# Why foundation models?



# 생성형 AI 애플리케이션 구축은 어려운 과제입니다.



여러 FM 및 최신  
버전에 액세스하기

FM  
커스터마이징은  
쉽지 않습니다.



데이터 개인정보  
보호 및 보안



FM이 작업을  
실행하도록 하기



데이터 원본에  
연결



인프라 관리의  
어려움



# Amazon Bedrock

FMS로 생성형 AI 애플리케이션을 구축하고 확장하는 가장 쉬운 방법



단일 API를 통해 다양한 주요 FM에 액세스하세요.



나만의 데이터로 FM을 비공개로 맞춤 설정하기



데이터 보안 및 규정 준수 지원



API를 동적으로 호출하여 복잡한 비즈니스 작업을 실행하는 에이전트를 구축하세요.

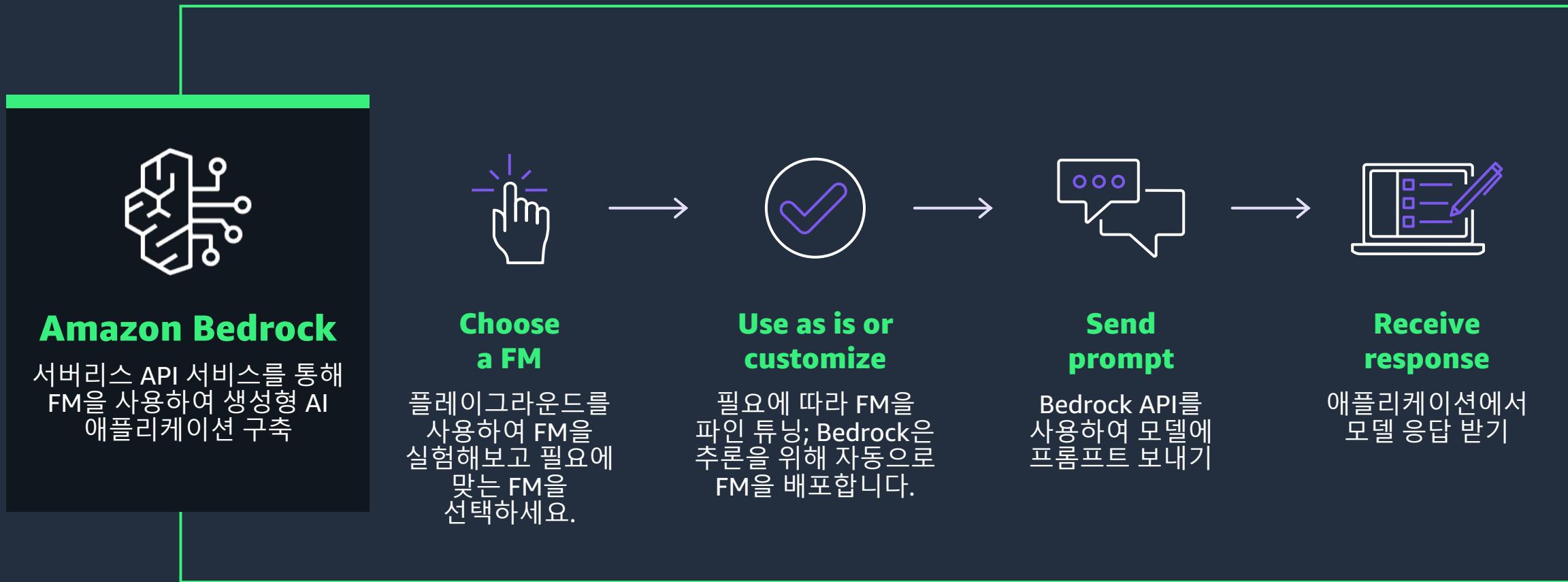


검색 증강 생성(RAG)을 사용하여 데이터로 FM의 성능을 확장하세요.



인프라 관리 없이 최고의 가격 대비 성능 확보

# How Amazon Bedrock works



# Benefits

-  주요 파운데이션 모델 선택
-  간편한 모델 커스터마이징
-  작업을 실행하는 완전 관리형 상담원
-  RAG에 대한 기본 지원
-  보안 및 규정 준수



# 다양한 주요 FM과 협력

단일 API 액세스를 사용하여 Amazon 및 주요 AI 기업의 FM에 액세스하세요.

대량의 데이터에 주석을 달지 않고도 특정 작업에 맞게 FM을 비공개로 사용자 지정할 수 있습니다.

관리형 인프라로 확장

The screenshot shows the 'Overview' section of the Amazon Bedrock console. It features a header with 'Amazon Bedrock > Overview'. Below the header, there's a section titled 'Foundation models' which includes logos for AI21 labs, Amazon, Anthropic, Cohere, and Stability AI. To the right, there's a 'Examples' section with a link to 'Browse all examples'. The main content area is divided into two main sections: 'Playgrounds' (Text and Image) and 'Customizations' (Fine-tuned models). Each section has a brief description and a 'View...' button.

Amazon Bedrock > Overview

## Overview

### Foundation models

Amazon Bedrock supports foundation models from industry-leading providers. Choose the model that is best suited to achieving your unique goals. [Explore models](#)

AI21 labs a A C S.

### Playgrounds

#### Text

Experiment using fast iterations on a vast range of language processing tasks, trying out various pre-trained models. In the playground, enter a text prompt to get started.

[Open text playground](#)

#### Image

Easily generate compelling images by providing text prompts to pre-trained models. In the playground, enter a text prompt to get started.

[Open image playground](#)

### Customizations

#### Fine-tuned models

Fine-tune a large language model with custom data and less training time to improve model performance in specific domains.

[View fine-tuned models](#)

# Foundation models





No one model  
**will rule them all**

# Amazon Bedrock

Broad choice of models

AI21labs

amazon

ANTHROPIC

cohere

Meta

MISTRAL  
AI

stability.ai

문맥에 맞는 답변,  
요약, 의역

텍스트 요약, 생성, Q&A,  
검색, 이미지 생성

요약, 복잡한 추론, 글쓰기,  
코딩

텍스트 생성, 검색,  
분류

Q&A 및 독해력

텍스트 요약, Q&A,  
텍스트 분류, 텍스트  
완성, 코드 생성

고품질 이미지와  
아트

Jurassic-2 Ultra

Amazon Titan Text Premier

Claude 3 Opus

Command

Llama 3 8B

Mistral Large

Stable Diffusion XL1.0

Jurassic-2 Mid

Amazon Titan Text Lite

Claude 3 Sonnet

Command Light

Llama 3 70B

Mistral 7B

Stable Diffusion XL 0.8

Amazon Titan Text Express

Claude 3 Haiku

Embed English

Llama 2 13B

Mixtral 8x7B

Amazon Titan Text  
Embeddings

Claude 2.1

Embed Multilingual

Llama 2 70B

Amazon Titan Text  
Embeddings V2

Claude 2

Command R+

Amazon Titan Multimodal  
Embeddings

Claude Instant

Command R

Amazon Titan Image  
Generator



**NEW****Titan Text Premier**

광범위한  
엔터프라이즈  
애플리케이션에서  
뛰어난 성능을  
제공하도록  
설계되었습니다.

Max Tokens: 32K

Languages: English

Fine Tuning: No

Use-case:  
Instruction  
following, Text  
Generation, Code  
generation.....

**Titan Text Express**

경량 다국어 소형  
비용 효율적인  
모델입니다.

Max Tokens: 4K

Languages: English

Fine Tuning: Yes

Use-case: RAG,  
Conversational  
chat, Text  
Generation, Code  
generation.....

**Titan Text Lite**

가격과 성능의  
균형을 제공하는  
LLM. 비용 효율적인  
파인 티닝에 이상적

Max Tokens: 8K

Languages: 100+  
languages

Fine Tuning: Yes

Use-case: RAG,  
Summarization,  
COT, Fine tuning,  
copywriting.....

**NEW****Titan Text Embeddings V2**

텍스트를 벡터  
표현으로  
변환합니다.

Max Tokens: 8K

Languages: 100+  
languages

Output: 256, 512,  
1024

Use-case: Semantic  
similarity ,  
Contextual  
responses, RAG,  
Q&A, Clustering

**Titan Image Generator**

텍스트에서  
이미지로 및  
이미지에서  
이미지로 생성.

Max Tokens: 77  
tokens and 25MB

Languages: English

Fine Tuning: Yes

Use-case: Generate  
image from text, In-  
painting, Image to  
Image generation.

**Titan MM Embeddings**

텍스트, 이미지 또는  
텍스트-이미지의  
조합을 허용하여  
임베딩을  
생성합니다.

Max Tokens: 128  
tokens and/or  
25MB image

Languages: English

Use-case: Semantic  
similarity ,  
Contextual  
responses, RAG,  
Q&A



# Titan Text Text Embeddings V2

RAG에 최적화된 유연한 임베딩 모델

NEW



유연한 차원 크기



RAG에 최적화



100개 이상의 사전 학습 언어

# Titan text embeddings model V2

NEW



## Amazon Titan Text Embeddings

v2.0

텍스트 입력을 벡터 표현(임베딩)으로 변환합니다.

Input Tokens: 512 (expandable to 8k)

Output Vectors: 1,024

Language: Multilingual (100 languages)



## Highlights

- Titan 텍스트 임베딩은 100개 언어로 빠르고, 비용 효율적이며, 고성능의 정확한 임베딩을 제공합니다.
- RAG, 텍스트 검색 작업, 의미적 유사성, 분류 및 클러스터링에 최적화되어 있습니다.
- 1024에서 512까지 검색 시 99% 정확도 유지, 1024에서 256까지 검색 시 97% 정확도 유지의 유연한 임베딩(1024, 512, 256)을 제공합니다.

# Amazon Titan Multi Modal Embedding



## Titan Multi Modal Embeddings

Max Tokens: **128 or 25 MB image size**

Output Vectors: **1,024**

Language: **English**

Model ID: *amazon.titan-embed-image-v1:0*



## Highlights



텍스트, 이미지 또는 텍스트-이미지의 조합을 허용하여 임베딩을 생성합니다.



고유하고 독점적인 데이터에 맞게 조정합니다.



편향된 검색 결과를 줄이는 데 도움이 되는 기본 제공 완화 기능

# Amazon Titan **Image Generator**

Generate realistic,  
studio-quality images



자연어 프롬프트를 사용하여 스튜디오 품질의  
이미지 생성



브랜드 스타일에 맞게 독점 데이터로 이미지  
사용자 지정



여러 주요 모델에 비해 텍스트-이미지 관련  
점수가 더 높습니다.

# Anthropic Claude

NEW

## ANTHROPIC Claude 3 Sonnet

대규모 배포를 위한  
인텔리전스와 속도의  
최상의 조합. 비전 기능.

Max Tokens: 200K

Languages: Multi-Lingual

Use-case: Search &  
Retrieval, RAG, Image  
analysis, Coding, Content  
Generation,  
Summarization, Advanced  
translation, .....

NEW

## ANTHROPIC Claude 3 Haiku

동급 모델보다 더 빠르고  
지능적이며 비용 효율적이며  
초고속 비전 기능을  
제공합니다.

Max Tokens: 200K

Languages: Multi-Lingual

Use-cases: Dialogue, Data  
extraction & classification,  
Search & Retrieval,  
Summarization, Image  
analysis, Translation, .....

NEW

## ANTHROPIC Claude 3 Opus

추론, 지식, 수학, 다국어 성능 및  
멀티모달 앱에 대한 선도적인  
벤치마크인 Anthropic의 가장  
지능적인 모델입니다.

Max Tokens: 200K

Languages: Multi-Lingual

Use-case: Advanced logical  
reasoning, Accelerate R&D tasks  
like literature review, hypothesis  
generation, end-to-end  
automation of complex tasks with  
agents & tool use, Graduate-level  
math, Data analysis, Interactive &  
complex coding, .....

## ANTHROPIC Claude 2.1

매우 큰 컨텍스트 창과  
신뢰성, 환각이 크게 개선된  
LLM

Max Tokens: 200K

Languages: Multi-Lingual

Use-case: Content  
generation, RAG, Q&A with  
citation, Summarization,  
COT, PII removal, Code  
Generation, Role play, Text  
analysis, .....

## ANTHROPIC Claude Instant 1.2

다양한 작업을 처리할 수  
있는 매우 빠르면서도  
저렴한 LLM

Max Tokens: 100K

Languages: Multi-lingual

Use-case: Content  
generation, RAG, Q&A with  
citation, Summarization,  
COT, PII removal, Code  
Generation, Role play, Text  
analysis, .....



# Claude3 Foundation Model



# Anthropic is founded by the world's forefront AI experts



**Dario Amodei**  
CEO



**Tom Brown**  
CTO



**Jared Kaplan**  
Chief Scientist



**Daniela Amodei**  
President



**Benjamin Mann**  
Co-Founder



**Chris Olah**  
Interpretability



**Jack Clark**  
Head of Policy

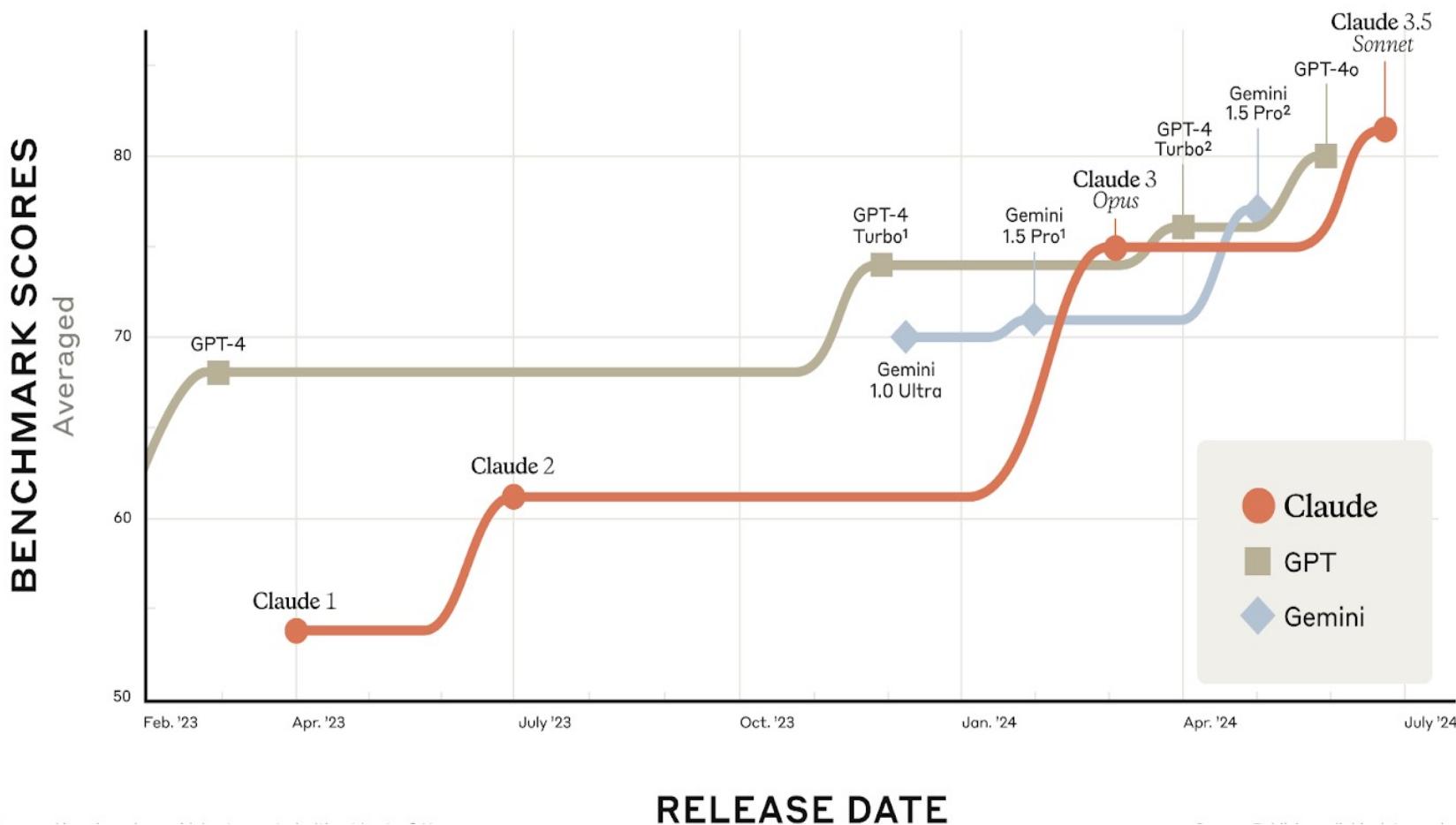


**Sam McCandlish**  
Scaling

## Frontier research includes:

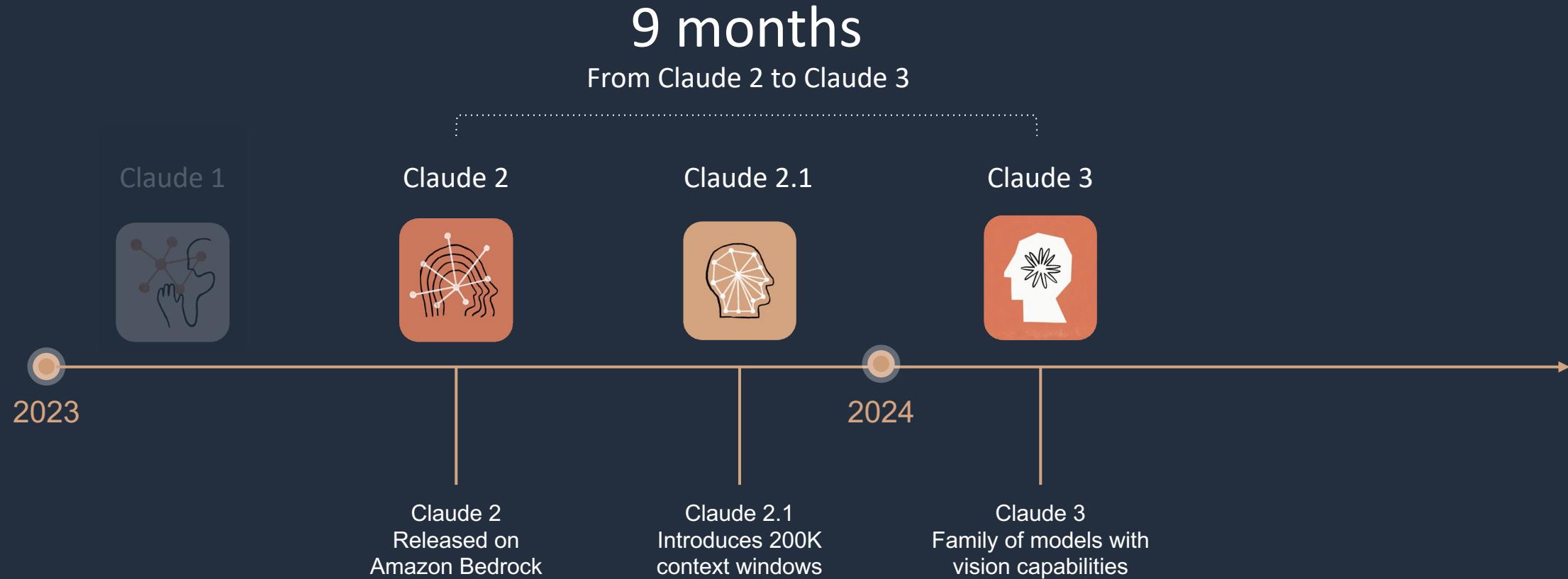
- Scaling Laws for Neural Language Models
- Language Models are Few-Shot Learners
- Constitutional AI
- Mapping The Mind of an LLM
- ...and more

# Anthropic is advancing the frontier at a rapid pace

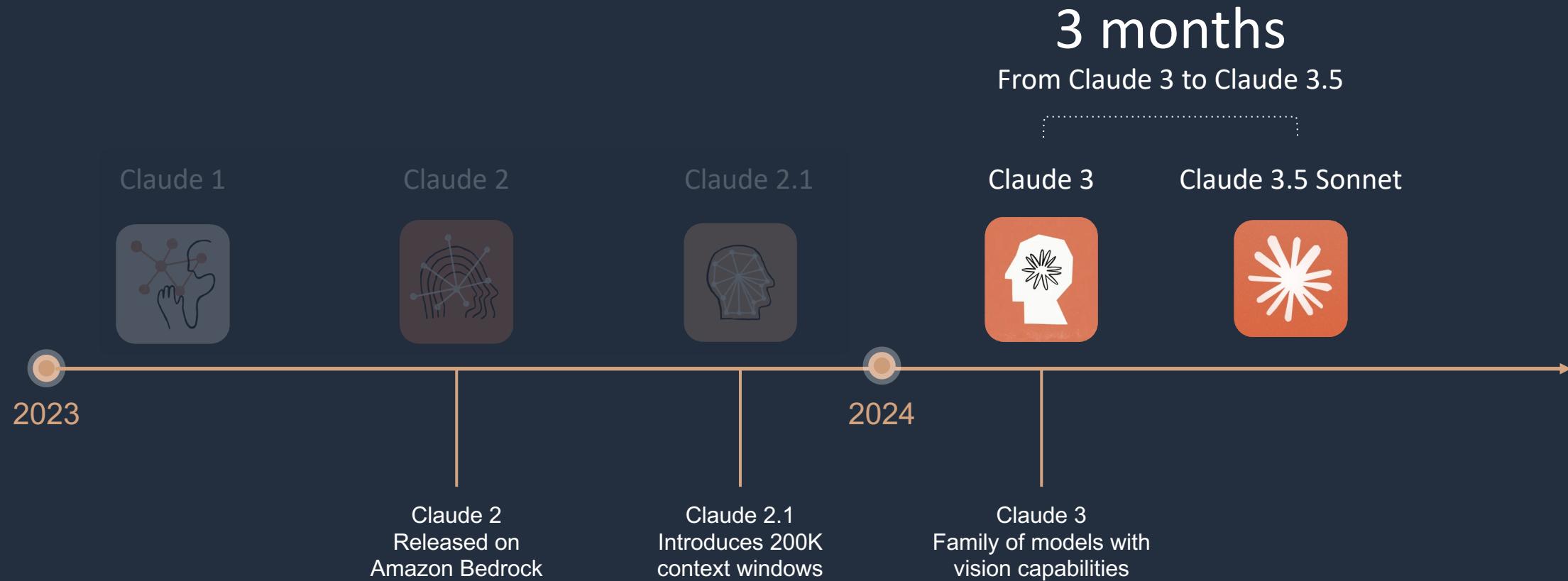


<sup>1.</sup> Measured via Anthropic's internal benchmarks using publicly available information

# Anthropic Claude on Amazon Bedrock

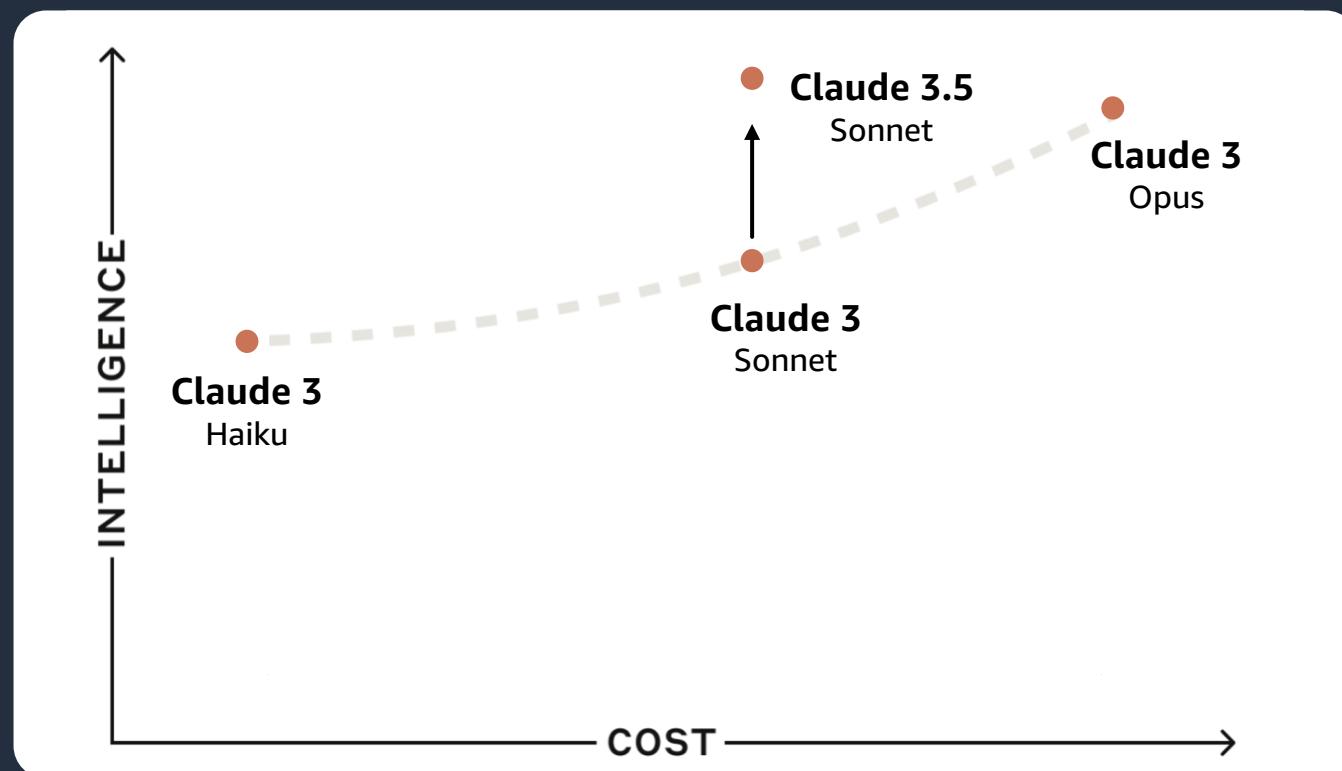


# Anthropic Claude on Amazon Bedrock



# Anthropic Claude Models

FOUNDATION MODELS CRAFTED FOR THE ENTERPRISE



## Claude model family

Claude 3.5  
Sonnet

Anthropic's most intelligent model to date, priced for high-volume use cases

Claude 3  
Opus

Powerful model for complex tasks that excels in sight-unseen scenarios

Claude 3  
Sonnet

Balance of speed and performance and intelligence

Claude 3  
Haiku

Fastest, most cost-effective model preferred for scaled tasks

# Claude 3.5 Sonnet





# Claude 3.5 Sonnet

MOST INTELLIGENT MODEL, PRICED FOR HIGH-VOLUME USER-FACING USE CASES

## Core features

- 다른 업계 선도 모델보다 더 지능적이고 비용 효율적입니다.
- 코딩, 추론, 글쓰기에 가장 적합한 모델
- 특히 차트 및 데이터 시각화를 통한 강력한 시각 기능

## Top skills

- Advanced reasoning
- Code generation
- Code translation
- Agentic tool use
- Visual reasoning
- Creative writing

## Top use cases

- マイグ레이션, 코드 수정 및 코드 번역을 포함한 소프트웨어 개발
- 데이터 과학, 특히 비정형 데이터 탐색 및 상담원 도구 사용 등
- 티켓 분류, 분류 및 해결에 대한 강력한 추론을 통한 고객 지원

# A new standard for model intelligence

Claude 3.5 Sonnet은 AI 시스템에 대한 대부분의 일반적인 평가 벤치마크에서 동급 제품보다 우수한 성능을 발휘합니다.



Undergraduate level expert  
knowledge (MMLU)



Graduate level expert reasoning  
(GPQA)



Multilingual reasoning abilities  
(MGSM, MMLU)



Interactive & complex coding  
(HumanEval)

Source: <https://www.anthropic.com/news/clause-3-5-sonnet>



© 2023, Amazon Web Services, Inc. or its affiliates.

# Frontier intelligence

## Claude 3.5 Sonnet

Claude 3.5 Sonnet은 다양한 평가에서 경쟁사 모델과 Claude 3 Opus를 능가하는 성능으로 인텔리전스에 대한 업계 기준을 높였으며, 중간급 모델인 Claude 3 Sonnet의 속도와 비용도 뛰어납니다.

	Claude 3.5 Sonnet	Claude 3 Opus	GPT-4o	Gemini 1.5 Pro
Graduate level reasoning <i>GPQA, Diamond</i>	<b>59.4%*</b> 0-shot CoT	<b>50.4%</b> 0-shot CoT	<b>53.6%</b> 0-shot CoT	—
Undergraduate level knowledge <i>MMLU</i>	<b>88.7%**</b> 5-shot  <b>88.3%</b> 0-shot CoT	<b>86.8%</b> 5-shot  <b>85.7%</b> 0-shot CoT	—  <b>88.7%</b> 0-shot CoT	<b>85.9%</b> 5-shot  —
Code <i>HumanEval</i>	<b>92.0%</b> 0-shot	<b>84.9%</b> 0-shot	<b>90.2%</b> 0-shot	<b>84.1%</b> 0-shot
Multilingual math <i>MGSM</i>	<b>91.6%</b> 0-shot CoT	<b>90.7%</b> 0-shot CoT	<b>90.5%</b> 0-shot CoT	<b>87.5%</b> 8-shot
Reasoning over text <i>DROP, F1 score</i>	<b>87.1</b> 3-shot	<b>83.1</b> 3-shot	<b>83.4</b> 3-shot	<b>74.9</b> Variable shots
Mixed evaluations <i>BIG-Bench-Hard</i>	<b>93.1%</b> 3-shot CoT	<b>86.8%</b> 3-shot CoT	—	<b>89.2%</b> 3-shot CoT
Math problem-solving <i>MATH</i>	<b>71.1%</b> 0-shot CoT	<b>60.1%</b> 0-shot CoT	<b>76.6%</b> 0-shot CoT	<b>67.7%</b> 4-shot
Grade school math <i>GSM8K</i>	<b>96.4%</b> 0-shot CoT	<b>95.0%</b> 0-shot CoT	—	<b>90.8%</b> 11-shot

\* Claude 3.5 Sonnet scores 67.2% on 5-shot CoT GPQA with maj@32

\*\* Claude 3.5 Sonnet scores 90.4% on MMLU with 5-shot CoT prompting

Source: <https://www.anthropic.com/news/clause-3-5-sonnet>



# Frontier visual reasoning

## Claude 3.5 Sonnet

이러한 개선 사항은 차트 및 그래프 해석과 같이 시각적 추론이 필요한 작업에서 가장 두드러지게 나타납니다. 클로드 3.5 소네트는 불완전한 이미지의 텍스트도 정확하게 전사할 수 있습니다.

	Claude 3.5 Sonnet	Claude 3 Opus	GPT-4o	Gemini 1.5 Pro
Visual math reasoning <i>MathVista (testmini)</i>	<b>67.7%</b> 0-shot CoT	<b>50.5%</b> 0-shot CoT	<b>63.8%</b> 0-shot CoT	<b>63.9%</b> 0-shot CoT
Science diagrams <i>AI2D, test</i>	<b>94.7%</b> 0-shot	<b>88.1%</b> 0-shot	<b>94.2%</b> 0-shot	<b>94.4%</b> 0-shot
Visual question answering <i>MMMU (val)</i>	<b>68.3%</b> 0-shot CoT	<b>59.4%</b> 0-shot CoT	<b>69.1%</b> 0-shot CoT	<b>62.2%</b> 0-shot CoT
Chart Q&A <i>Relaxed accuracy (test)</i>	<b>90.8%</b> 0-shot CoT	<b>80.8%</b> 0-shot CoT	<b>85.7%</b> 0-shot CoT	<b>87.2%</b> 0-shot CoT
Document visual Q&A <i>ANLS score, test</i>	<b>95.2%</b> 0-shot	<b>89.3%</b> 0-shot	<b>92.8%</b> 0-shot	<b>93.1%</b> 0-shot

Source: <https://www.anthropic.com/news/clause-3-5-sonnet>



# Anthropic Claude Models

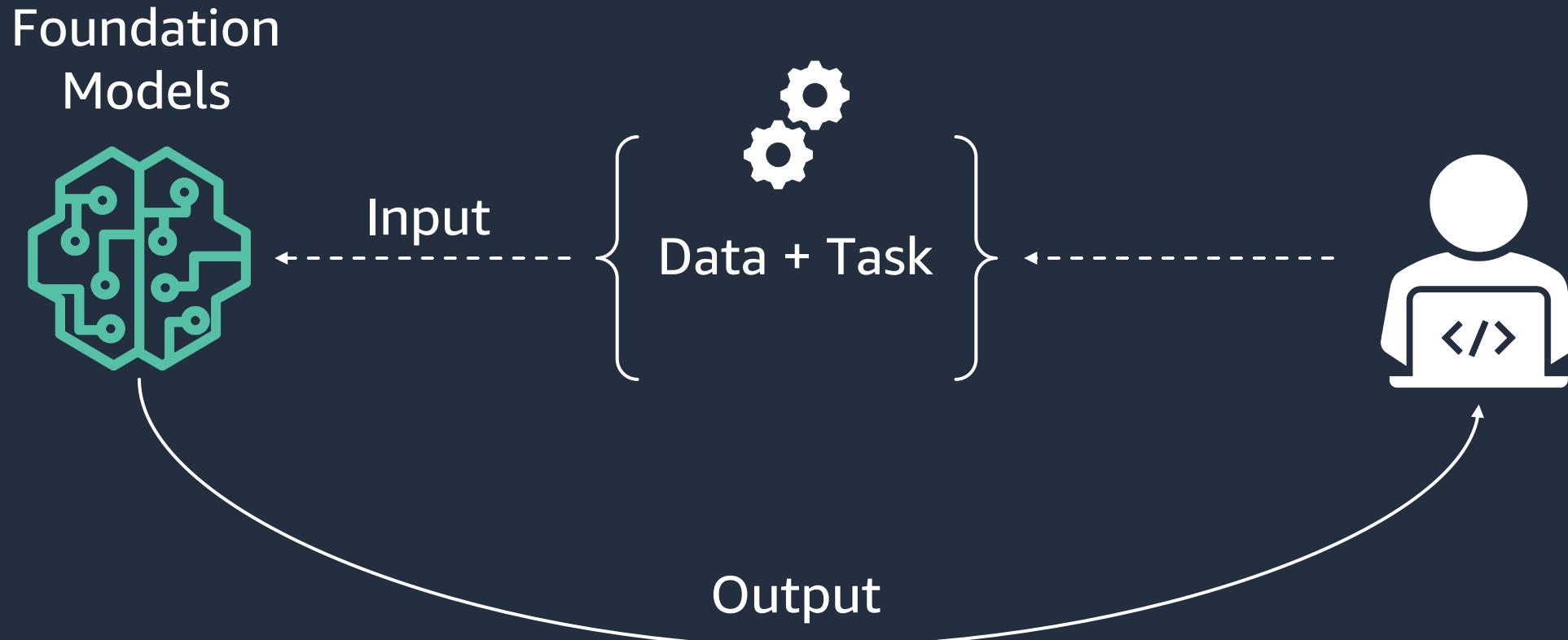
CHOOSE THE EXACT COMBINATION OF INTELLIGENCE, SPEED, AND COST TO SUIT YOUR NEEDS

	Claude 3.5 Sonnet	Claude 3 Haiku	Claude 3 Sonnet	Claude 3 Opus	
Use case	Most intelligent, built for high-volume use cases	Fastest performance at the lowest cost	Balance between intelligence, speed, and cost	Second-most intelligent overall; most intelligent in Claude 3 family	
Context	200K	200K	200K	200K	
Vision	✓	✓	✓	✓	
Cost*	Input: Output:	\$0.003 \$0.015	\$0.00025 \$0.00125	\$0.003 \$0.015	\$0.015 \$0.075

\*Per 1K tokens

# Prompt engineering

# 프롬프트 엔지니어링, ML을 사용하는 새로운 방법!



# 프롬프트의 요소

## Instruction

모델 수행 방법에 대한  
작업 설명 또는 지침

## Context

모델 성능을 조정하기  
위한 추가/외부 정보

## Input Data

모델이 출력을 제공해야  
하는 입력/질문입니다.

## Output Indicator

모델에 출력을 제공해야  
하는 지표/형식

### Instructions

Act as an IT technical expert providing customer service. Consider the Context below to answer the user's questions with a friendly tone. Answer in English in 2 sentences or less providing instructions.

### Context

Context: You work in the Support line of a technology company that commercializes Android smartphones.  
The user is calling because the phone is not charging

### User Input

User: Hi, how can I fix my phone?

### Output Indicator

Assistant:  
**I will try my best to assist you with this. Can you please tell me the model of your phone and what version of the Android operating system it is running?**



# 모델로 실험하기



## Playground experience

- 여러 모델과 제공업체 중에서 선택할 수 있습니다.
- 사용이 간편합니다. 프롬프트 필드에 텍스트를 입력한 다음 실행을 선택하면 됩니다. 응답 패널에서 콘솔에 모델의 응답이 표시됩니다.
- 추론 구성 매개변수를 조정한 다음 프롬프트를 다시 실행할 수 있습니다.

# Examples



## Try out real-life use cases

- 아마존 베드락 콘솔에는 각 모델 공급업체에 대해 최대 20개의 예시가 표시됩니다.
- 양식, 공급자 이름, 모델 이름 또는 예제 카테고리 중 하나 이상의 속성을 기준으로 예제 목록을 필터링할 수 있습니다.
- 예제를 플레이그라운드에서 열어 쉽게 실행할 수 있습니다.

# Experiment with models to try out real-life use cases

Examples (15) Info

Search in examples

**Titan Text Large v1.01**

**Action items from a meeting transcript**

This prompt creates a list of action items from a meeting transcript.

Summarization

**Claude Instant v1.3**

**Content Generation**

An example prompt to generate a paragraph from instructions

Text generation

**Stable Diffusion XL v2.2.2**

**Create an image**

This prompt creates an image based on a short description

Image gen

**Titan Text Large v1.01**

**Creating a table of product descriptions**

This is a prompt for generating product descriptions that incorporate keywords

Open ended text generation

**Creating a table of product descriptions**

This is a prompt for generating product descriptions that incorporate keywords

**Open in Playground**

 Titan Text Large v1.01  
By Amazon

**Prompt**

Product: Sunglasses.  
Keywords: polarized, designer, comfortable, UV protection, aviators.

Create a table that contains five variations of a detailed product description for the product listed above, each variation of the product description must use all the keywords listed.

**Inference configuration**

Temperature	0
Top P	1
Response length	4096
Stop sequences	N/A

**API request**

**Copy**

```
1 {
2   "modelId": "amazon.titan-tg1-large",
3   "contentType": "application/json",
4   "accept": "*/*",
5   "body": {
6     "inputText": "Product: Sunglasses. Keywords: polarized, de",
7     "textGenerationConfig": {
8       "maxTokenCount": 4096,
9       "stopSequences": [],
10      "temperature": 0,
11      "topP": 1
12    }
13 }
```

# Access Bedrock via the AWS console

Amazon ▾ Titan Large v1.01 ▾

Powerful, general-purpose models pretrained on large datasets, Titan FMs are powerful, general-purpose models that can be used as-is or customized to perform specific tasks.  
[View details](#) ⓘ

**What is generative AI?**

Generative AI is a form of artificial intelligence that can create text, images, or videos based on user prompts. It uses machine learning algorithms to analyze data and generate new content that is similar to the input it has been given.

One of the main applications of generative AI is in natural language processing (NLP). Generative AI can be used to create text that is similar to the way a human writes, and can also be used to generate translations between different languages. It can also be used to create social media posts, reviews, and other types of content that require human-like writing skills.

Another application of generative AI is in image and video creation. Generative AI can be used to create new images or videos that are similar to those that have been seen before. It can also be used to create animations or to modify existing images or videos in creative ways.

There are also generative AI applications in music and sound. Generative AI can be used to create new music or to modify existing music files in creative ways. It can also be used to create sound effects or to modify existing sound files.

Overall, generative AI is a powerful tool that has the potential to revolutionize the way we create and consume content. It can help to automate many of the tasks that were previously done by humans, and can also help to create new forms of content that were not possible before.

**Inference configuration** [Reset](#)

**Randomness and diversity** [Info](#)

Temperature

Top P

**Length** [Info](#)

Response length

Stop sequences  
  
Only | allowed (maximum 20 characters)

**Run** ⏪ ⏴ [View API request](#) [Save](#)

Stability AI ▾

Stable Diffusion XL v2.2.2 ▾

Deep learning, text-to-image model used to generate detailed images conditioned on text descriptions, inpainting, outpainting, and generating image-to-image translations.

[View details](#)

**Prompt**

photo taken from above of an italian landscape. cloud is clear with few clouds. Green hills and few villages, a lake

**Inference configuration** [Info](#)

Reset

Prompt strength

Generation step

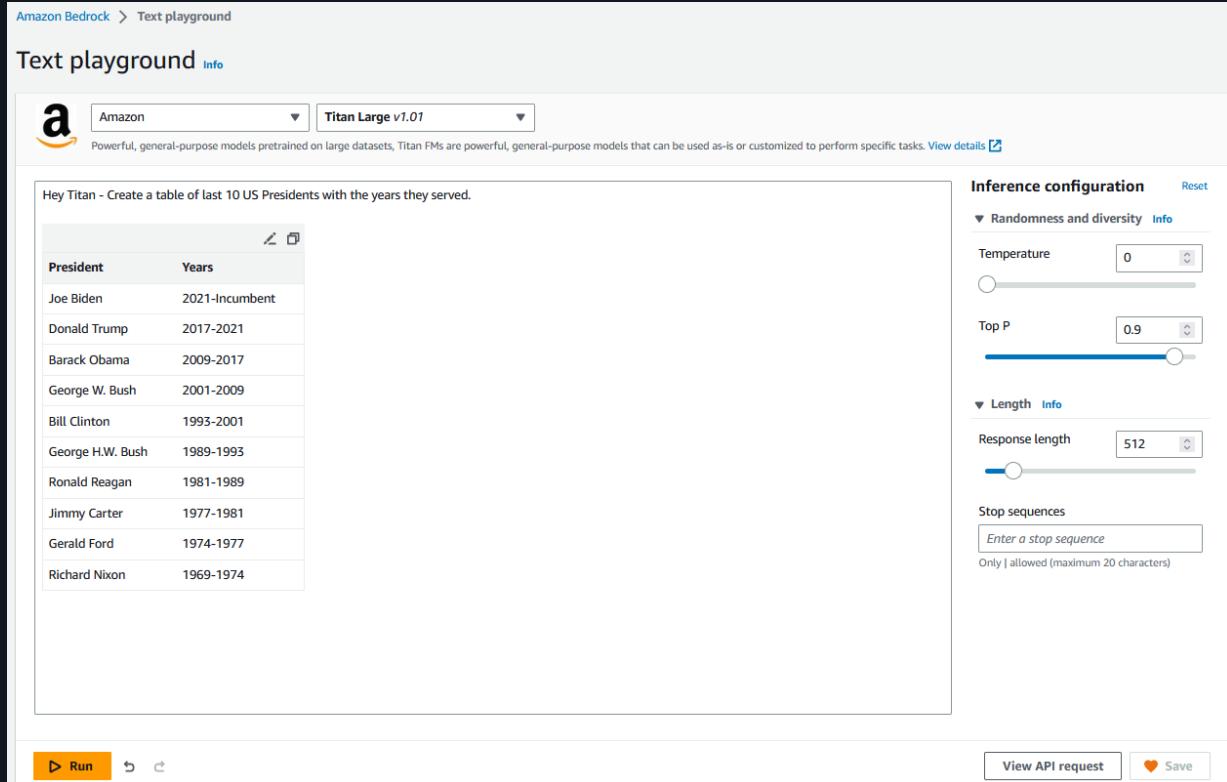
Seed

**Response**



# Playground – New features

## Markdown formatting



The screenshot shows the Amazon Bedrock Text playground interface. At the top, there are dropdown menus for 'Amazon' and 'Titan Large v1.01'. Below them is a message: 'Hey Titan - Create a table of last 10 US Presidents with the years they served.' To the right is an 'Inference configuration' panel with sections for 'Randomness and diversity' (Temperature: 0, Top P: 0.9), 'Length' (Response length: 512), and 'Stop sequences' (an input field). On the left, a table is displayed:

President	Years
Joe Biden	2021-Incumbent
Donald Trump	2017-2021
Barack Obama	2009-2017
George W. Bush	2001-2009
Bill Clinton	1993-2001
George H.W. Bush	1989-1993
Ronald Reagan	1981-1989
Jimmy Carter	1977-1981
Gerald Ford	1974-1977
Richard Nixon	1969-1974

At the bottom are buttons for 'Run' (orange), 'View API request', and 'Save'.

# Playground – New features

## Markdown formatting

The screenshot shows the Amazon Bedrock Text playground interface. On the left, there's a list of US Presidents from 1969 to 2009. A yellow circle highlights a portion of the interface where a user has typed "Hey Titan - Can I see a list of the last 10 US Presidents?" and received a response starting with "Copy this data over to Excel". Below this, a large gray box contains the text "Copy this data over to Excel". On the right, there's an "Inference configuration" panel with settings for temperature (0), top P (0.9), and response length (512). At the bottom, there are "Run" and "Save" buttons.

Hey Titan - Can I see a list of the last 10 US Presidents?

Copy this data over to Excel

President	Years
Joe Biden	2021-Present
Donald Trump	2017-2021
Barack Obama	2009-2017
George W. Bush	2001-2009
Bill Clinton	1993-2001
George H.W. Bush	1989-1993
Ronald Reagan	1981-1989
Jimmy Carter	1977-1981
Gerald Ford	1974-1977
Richard Nixon	1969-1974

Inference configuration

Randomness and diversity

Temperature: 0

Top P: 0.9

Length

Response length: 512

Stop sequences

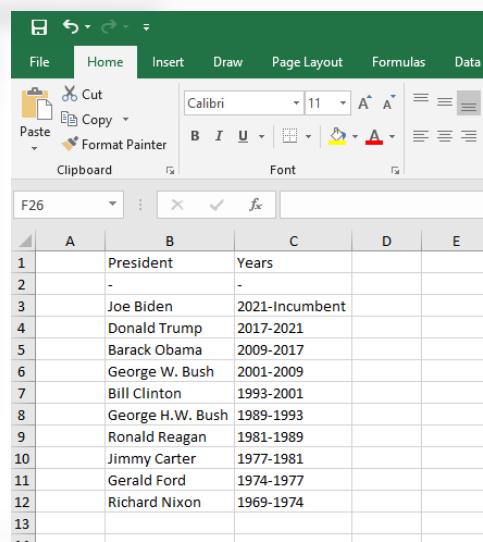
View API request

Save

# Playground – New features

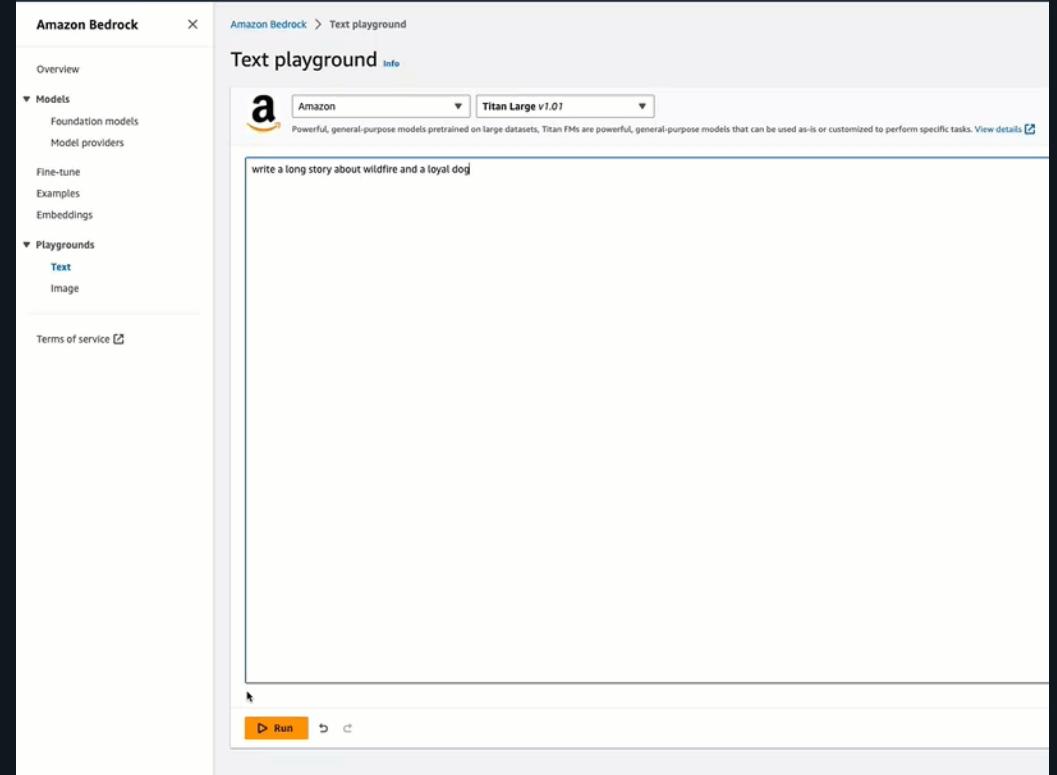
## Markdown formatting

Markdown support helps with carrying over  
formatting of the model output



	A	B	C	D	E
1	President	Years			
2	-	-			
3	Joe Biden	2021-Incumbent			
4	Donald Trump	2017-2021			
5	Barack Obama	2009-2017			
6	George W. Bush	2001-2009			
7	Bill Clinton	1993-2001			
8	George H.W. Bush	1989-1993			
9	Ronald Reagan	1981-1989			
10	Jimmy Carter	1977-1981			
11	Gerald Ford	1974-1977			
12	Richard Nixon	1969-1974			
13					

## Streaming responses



The screenshot shows the Amazon Bedrock Text playground interface. On the left, there's a sidebar with 'Amazon Bedrock' branding and sections for 'Overview', 'Models' (Foundation models, Model providers), 'Fine-tune', 'Examples', 'Embeddings', and 'Playgrounds' (Text, Image). The 'Text' tab is selected. On the right, the main area has a title 'Text playground' with a 'Info' link. It features a logo for 'Amazon' and 'Titan Large v1.01'. Below that is a text input field containing the prompt 'write a long story about wildfire and a loyal dog'. At the bottom, there are 'Run' and 'Stop' buttons.

# Chat Playground

채팅 플레이그라운드를 사용하여 파운데이션 모델을 빠르게 평가하기

대화 전체에 걸쳐 역할을 맡을 수 있는 지침을 제공하세요.

브라우저 세션이 완료되거나 '새 채팅 시작하기'를 클릭하면 메시지와 응답이 지워집니다.

Chat playground [Info](#)  Streaming

AI Anthropic Claude V2 v2 Start new chat

Next-gen AI assistant trained on helpful, honest, and harmless AI systems. Claude can help with summarization, search, creative writing, Q&A, coding, as well as take direction. Note: The terms "Human:" and "Assistant:" will automatically be added to all prompts using Anthropic models to improve your results. [View details](#)

Instructions update: Use a tone that is technical and scientific.

Human: Human: Hello, who are you? Can you tell me about the creation of blackholes?

Assistant: Assistant: Here is some information about the creation of black holes:

Black holes are formed when massive stars collapse at the end of their life cycles. When a star with a mass greater than about 3 times the mass of the sun runs out of fuel to burn, it can no longer produce enough heat and radiation pressure to counteract the gravitational force inward. This causes the star to collapse rapidly under its own weight. As it collapses, the gravitational force increases further, speeding up the collapse. If the remaining core of the star after collapse has a mass greater than about 3 solar masses, it will form a black hole. The collapse continues until the core reaches an extremely dense state called a singularity, where the laws of physics as we know them break down. At the singularity, the gravitational field becomes infinite. Surrounding the singularity is the event horizon, the boundary past which nothing, not even light, can escape the black hole's gravitational pull.

Some ways black holes can form:

Instructions: Use a tone that is technical and scientific. [Update](#) Default [Update inference configurations](#)

Human:

▶

# Amazon Bedrock API

## 생성형 AI를 구축할 수 있는 단일 API



# Single API to build with generative AI

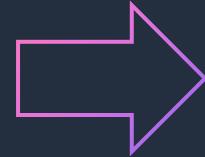


## Bedrock core API: InvokeModel

- 요청의 모델 ID, 콘텐츠 유형 및 본문을 전달합니다.
  - 본문에는 프롬프트 및 실행 매개변수가 포함됩니다.
  - 모델 응답 및 메타데이터를 반환합니다.
- 텍스트 대 텍스트, 텍스트 대 이미지, 이미지 대 이미지 등을 처리합니다.
- 현재 및 향후 Amazon Titan 모델, 타사 모델 및 파인 튜닝된 모델까지 지원

# Bedrock core API: InvokeModel

```
bedrock.invoke_model(  
    modelId = model_id,  
    contentType = "...",  
    accept = "...",  
    body = body)
```

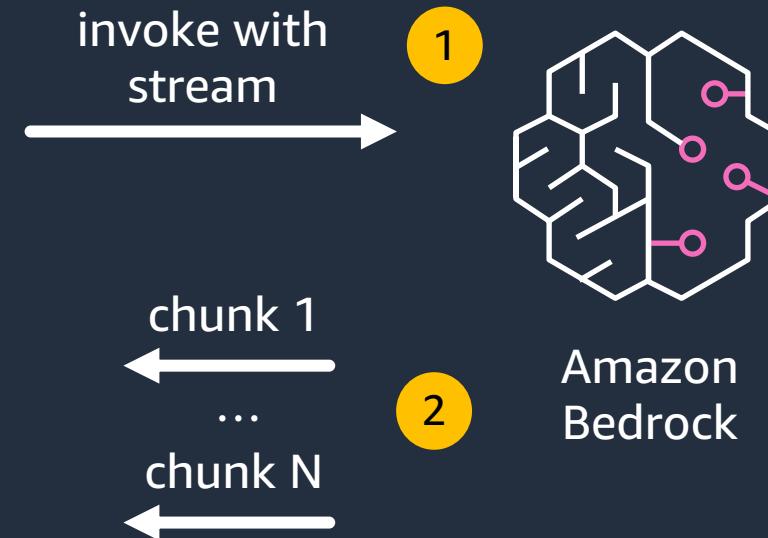


Access  
foundation  
models

- Amazon Titan models
- Third-party models
- Fine-tuned models NEW

# Bedrock core API: Streaming responses

```
response = bedrock.invoke_model_with_response_stream(  
    modelId = model_id, body = body)  
stream = response.get('body')  
if stream:  
    for event in stream:  
        chunk = event.get('chunk')  
        if chunk:  
            print(json.loads(chunk.get('bytes')).decode())
```



- 사용자는 첫 번째 청크가 제공되는 즉시 응답 읽기를 시작할 수 있습니다.

# Access Bedrock via Boto3: API operations

- list foundation models()

파운데이션 모델에 대한 정보를 검색하려면 ListFoundationModels 작업을 사용합니다.

- invoke model()

이 호출을 사용하여 원하는 모델을 호출합니다. API 매개변수와 결과는 호출하는 모델에 따라 달라집니다.

\* AWS CLI 및 Python용 AWS SDK(Boto3)를 사용하여 Amazon Bedrock API에 액세스할 수 있습니다.

# invoke\_model() - Anthropic Claude

```
prompt_data = """Command: Write me a blog about making strong business decisions as a leader."""

body = json.dumps({"prompt": prompt_data,
                   "max_tokens_to_sample": 500,
                   "temperature": 0.5,
                   "top_k": 250,
                   "top_p": 1,
                   "stop_sequences": ["\n\nHuman:"]
})
modelId = "anthropic.claude-instant-v1"
accept = "*/*"
contentType = "application/json"
response = bedrock_runtime.invoke_model(
    body=body, modelId=modelId, accept=accept, contentType=contentType
)
response_body = json.loads(response.get("body").read())
print(response_body.get("completion"))
```

# invoke\_model() – Titan Embeddings

```
prompt_data = """This is just some input I want to get embeddings for"""

body = json.dumps({"inputText": prompt_data})
modelId = "amazon.titan-embed-text-v1"
accept = "*/*"
contentType = "application/json"
response = bedrock_runtime.invoke_model(
    body=body, modelId=modelId, accept=accept, contentType=contentType
)
response_body = json.loads(response.get("body").read())
embedding = response_body.get("embedding")
print(embedding)
```

# invoke\_model() – Stability.ai Stable Diffusion

```
body = json.dumps({
    "text_prompts": [
        {"text": "a pizza", "weight": 1},
        {"text": "an unrealistic pizza shape", "weight": -1}
    ],
    "cfg_scale": 10,
    "seed": 20,
    "steps": 50,
    "style_preset": "cinematic",
    # For in-painting or image-to-image task, add following
    "init_image": "<base64 encoded string of image>",
    "start_schedule": 0.6 # 0-1 scale of relevancy
})
modelId = "stability.stable-diffusion-xl"
accept = "application/json"
contentType = "application/json"
response = bedrock_runtime.invoke_model(
    body=body, modelId=modelId, accept=accept, contentType=contentType
)
response_body = json.loads(response.get("body").read())
base_64_img_str = response_body["artifacts"][0]["base64"]
image = Image.open(io.BytesIO(base64.decodebytes(bytes(base_64_img_str, "utf-8"))))
print(image)
```

# Integrated with LangChain

```
pip install langchain
```

```
from langchain import Bedrock
from langchain.embeddings import BedrockEmbeddings

llm = Bedrock()
print(llm("what is generative
AI?"))
```

**Popular Python framework for developing applications powered by language models**

- 아마존 베드락을 위한 새로운 LLM 및 임베딩 클래스
- 대화 체인에서 LLM 클래스를 사용하기 위한 코드가 포함되어 있습니다.
- 텍스트에서 임베딩을 생성하는 코드 포함

# CloudWatch metrics



Amazon  
CloudWatch

## CloudWatch metrics now supported:

- 모델 호출 횟수
- 호출 지연 시간
- 오류 메트릭에는 호출 횟수가 포함됩니다:
  - 클라이언트 측 오류
  - 서버 측 오류
  - 스로틀링

"AWS/Bedrock" 네임스페이스 및 각 메트릭은 모델별("ModelId" 차원)입니다.

# Model invocation logging

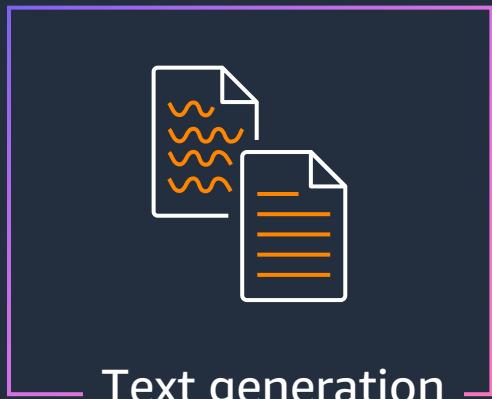
모델 호출 로깅을 사용하여 계정의 모든 모델 호출에 대한 메타데이터, 요청 및 응답을 수집하세요.

S3 및 CloudWatch 로그 모두 지원

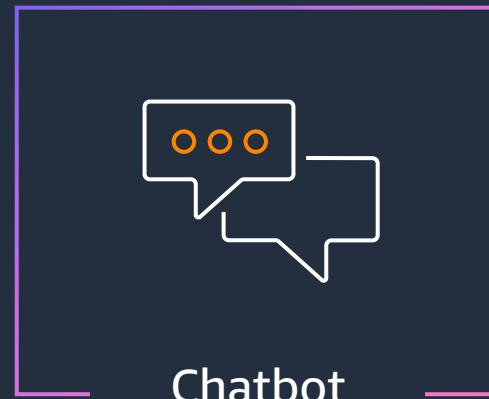
The screenshot shows the 'Settings' page in the Amazon Bedrock console. At the top, it says 'Amazon Bedrock > Settings'. Below that, it says 'Settings' and 'These are account level settings'. Under the heading 'Model invocation logging' (with a 'Info' link), it says 'Use model invocation logging to collect metadata, requests, and responses for all model invocations in your account.' A blue toggle switch is turned on next to the text 'Model invocation logging', with the subtext 'Enabling model invocation logging will start publishing invocation logs.' Below this, there's a section titled 'Select the data types to include with logs - Optional' with three checked checkboxes: 'Text', 'Image', and 'Embedding'. Then, there's a section titled 'Select the logging destinations' with three radio buttons: 'S3 only' (selected), 'Cloudwatch Logs only', and 'Both S3 and Cloudwatch Logs'. At the bottom, there are 'Cancel' and 'Save settings' buttons, with 'Save settings' being orange.

# Architecture patterns

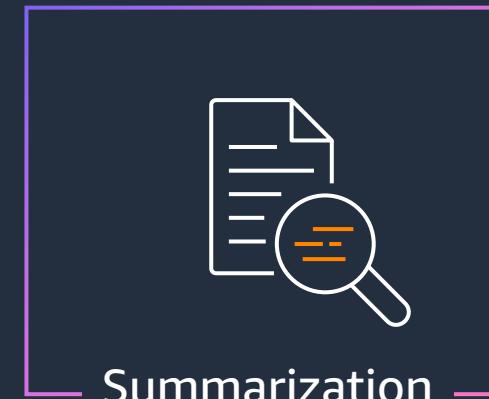
# Architecture patterns in:



Text generation



Chatbot



Summarization

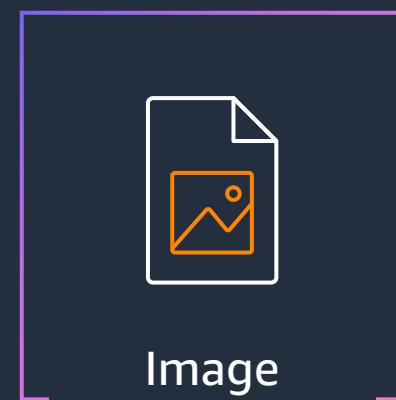


Image  
generation



Question  
answering

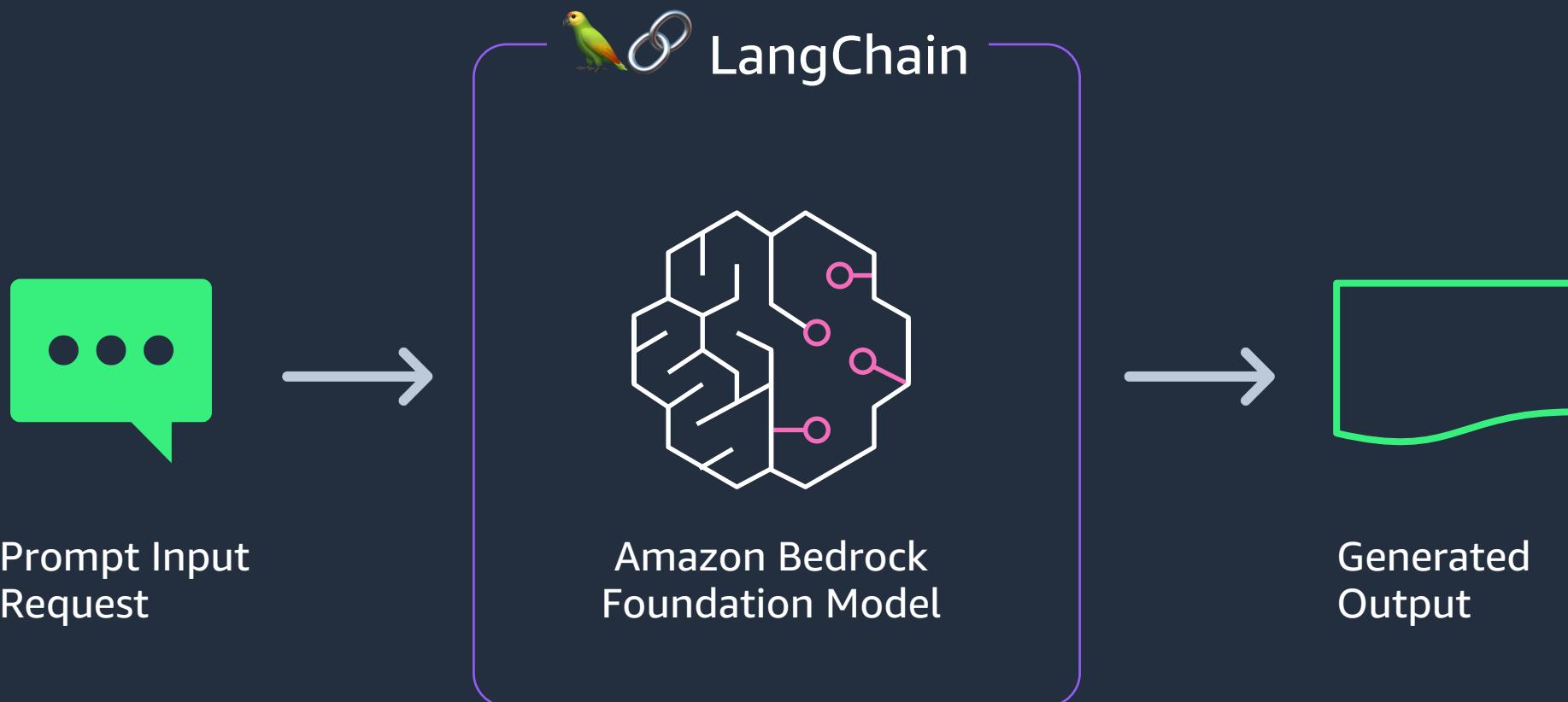
# Text Generation

WITH SIMPLE PROMPT



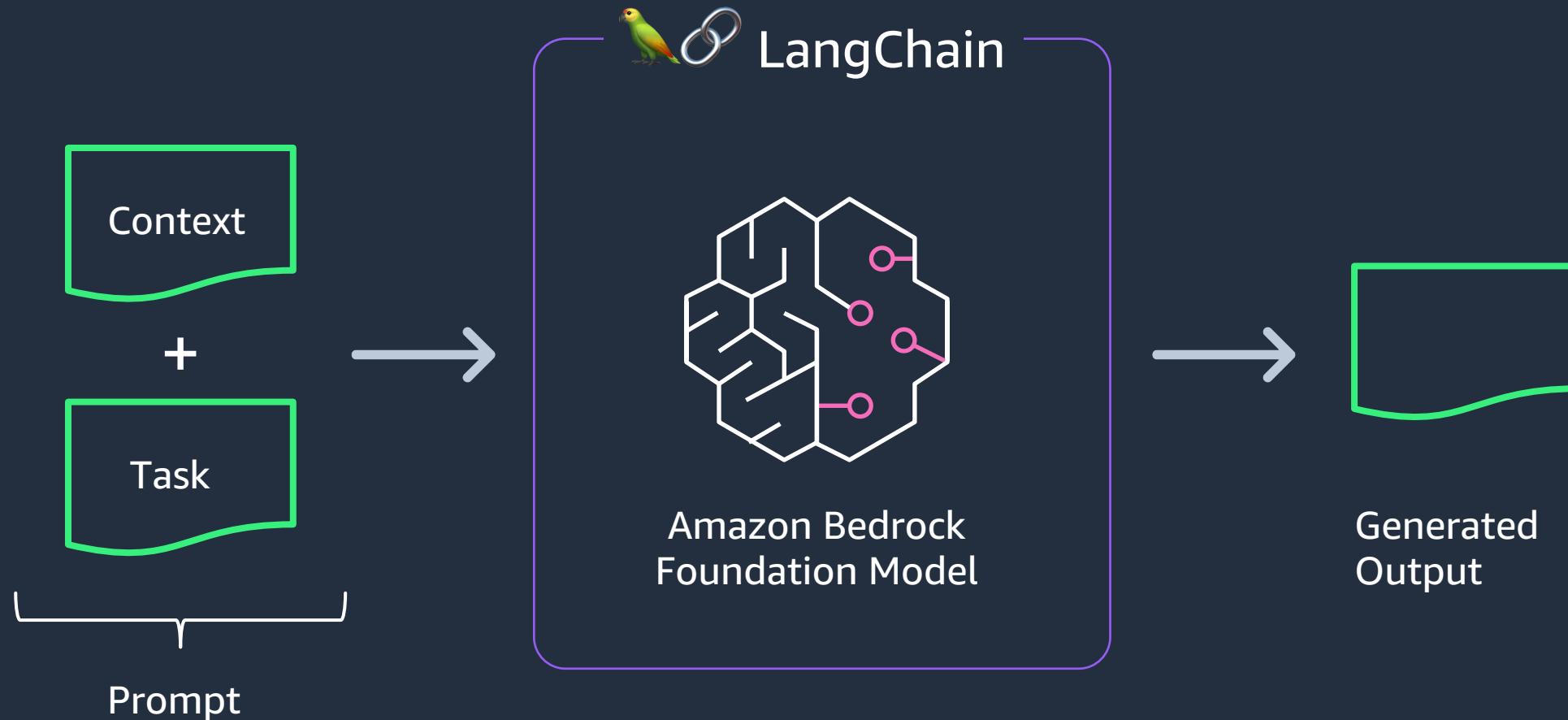
# Text Generation

WITH LANGCHAIN



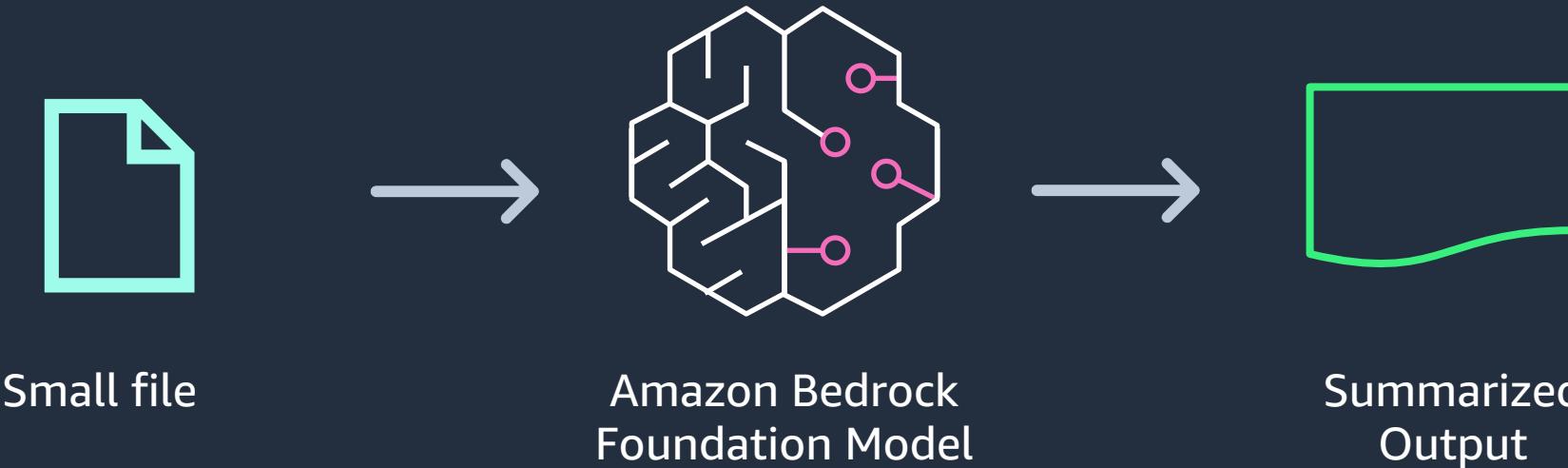
# Text Generation

WITH CONTEXT AND LANGCHAIN



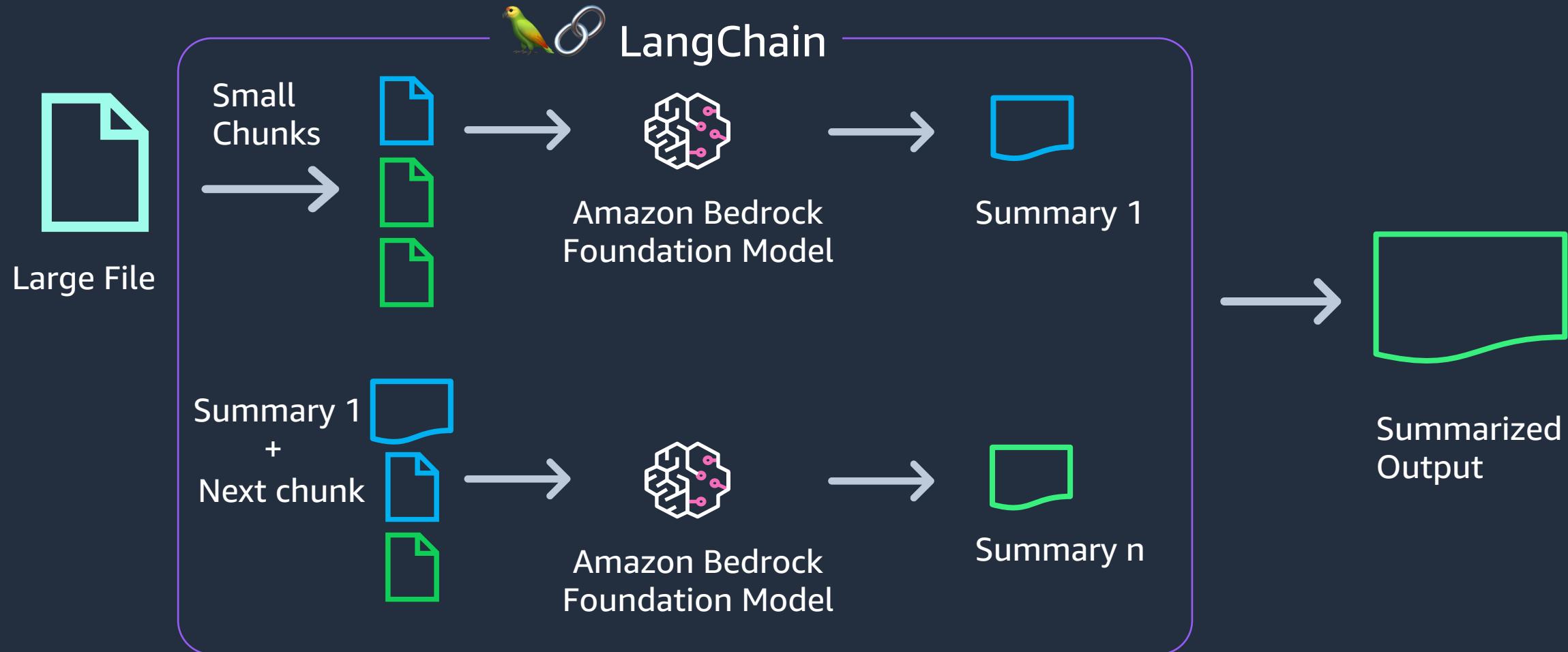
# Text Summarization

WITH SMALL FILES



# Text Summarization

WITH LARGE FILES AND LANGCHAIN



# Question Answering

WITH SIMPLE PROMPT



User  
Question

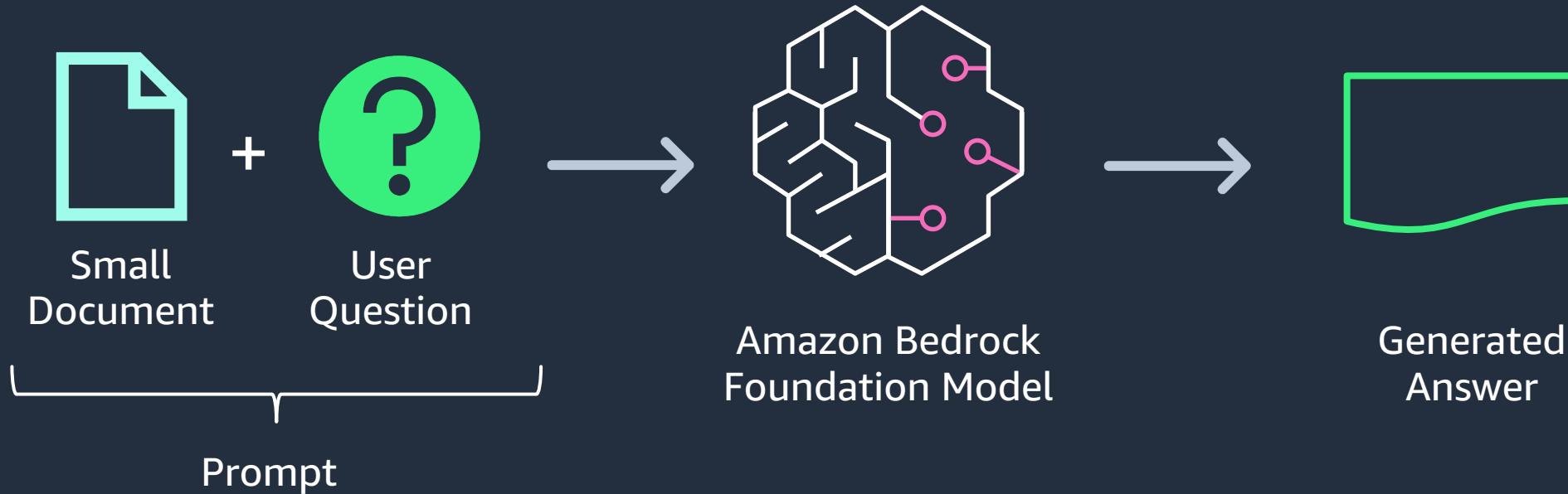


Amazon Bedrock  
Foundation Model

Generated  
Answer

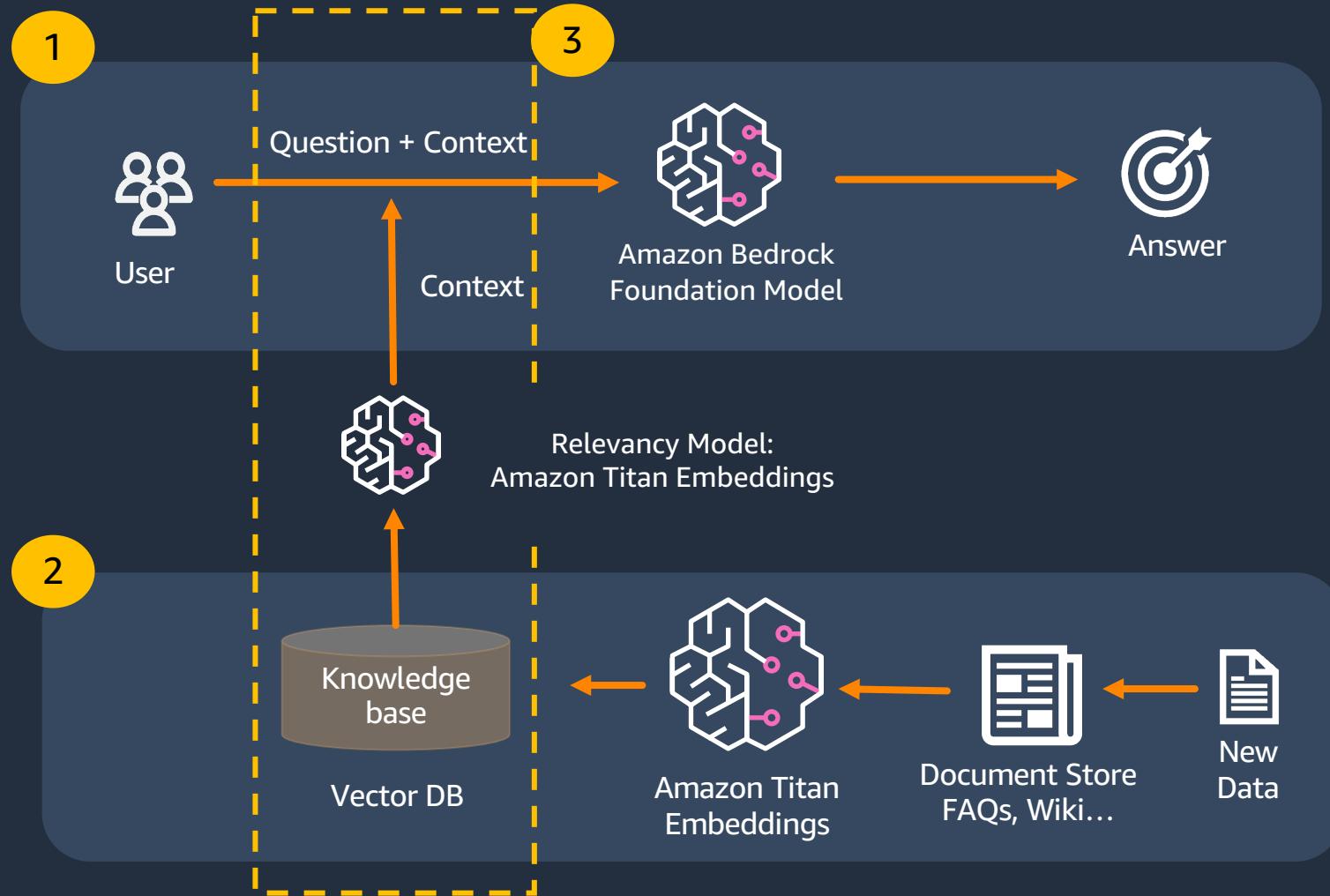
# Question Answering

WITH CONTEXT



# Question Answering

WITH RETRIEVAL-AUGMENTED GENERATION



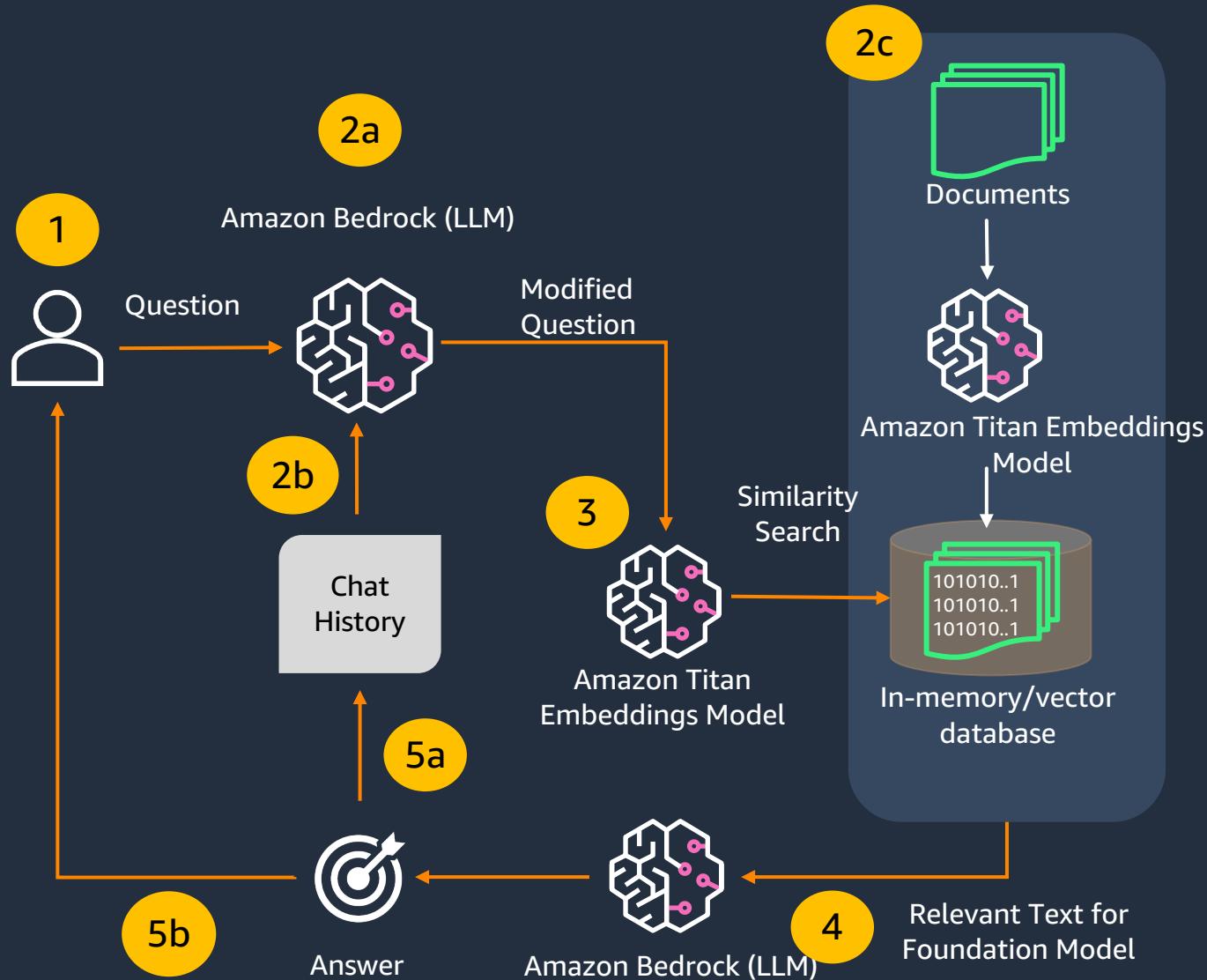
# Chatbot

## BASIC

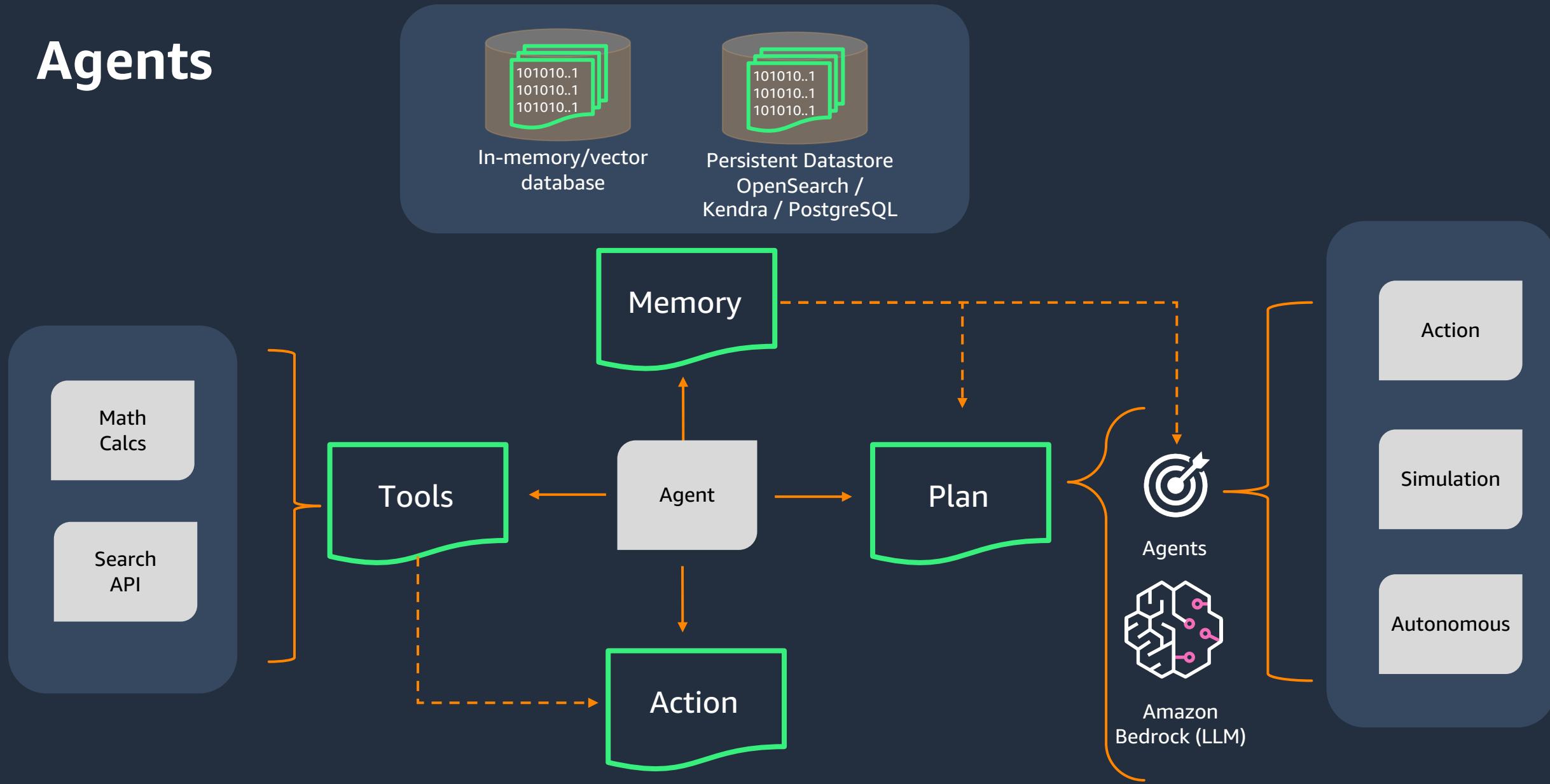


# Chatbot

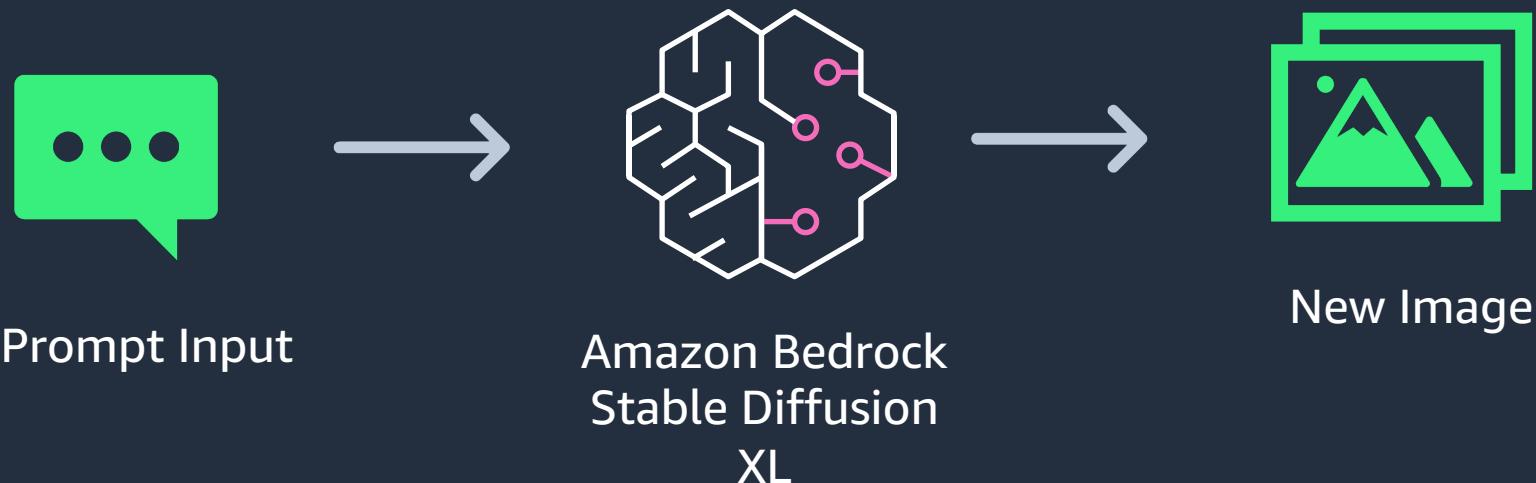
WITH CONTEXT



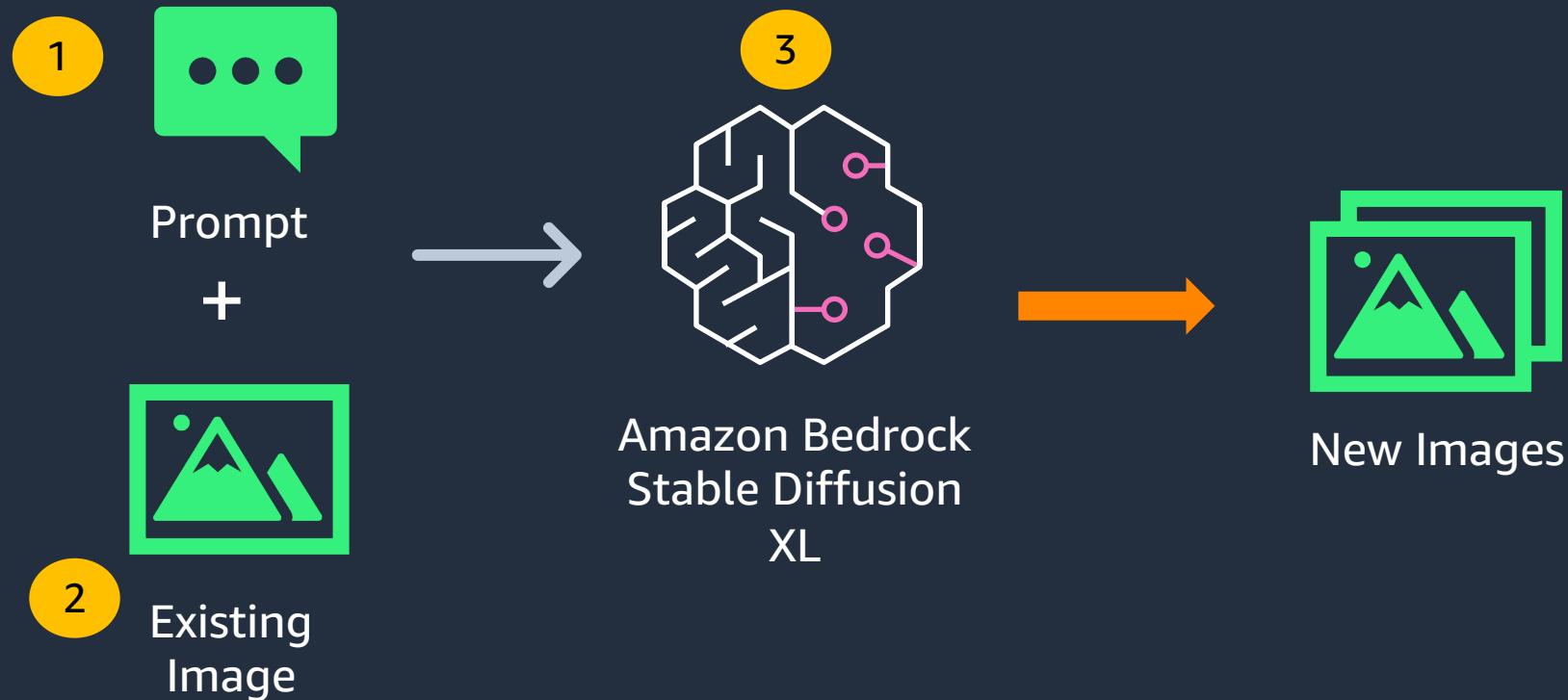
# Agents



# Text to Image



# Image to Image (In-painting)



# Security



# Data privacy



## You are always in control of your data

- 고객 데이터는 다른 고객을 위해 Amazon Titan 모델을 개선하는 데 사용되지 않으며 다른 기초 모델 제공업체와 공유되지 않습니다.
- 고객 데이터(프롬프트, 응답, 파인 투닝된 모델)는 생성된 지역에 유지됩니다.

# Data security

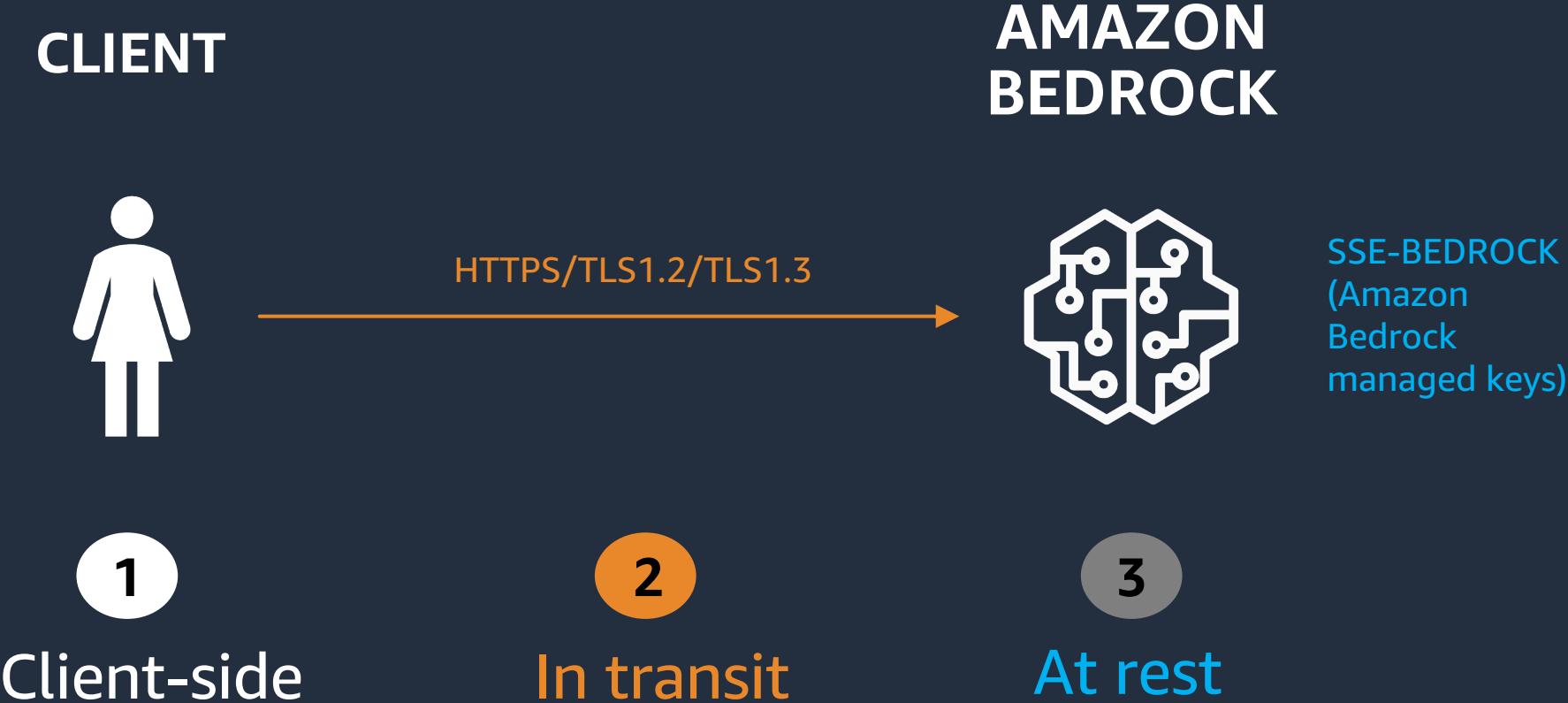


## You are always in control of your data

- 고객이 **VPC 엔드포인트**를 사용하여 가상 프라이빗 클라우드(VPC)와 베드락 서비스 간에 프라이빗 연결을 설정할 수 있도록 AWS PrivateLink를 지원합니다.
- 추론 액세스를 관리하고, 특정 모델에 대한 액세스를 거부하고, 콘솔 액세스를 활성화하기 위해 AWS IAM(신원 및 액세스 관리 서비스)과의 통합
- CloudTrail을 사용하여 솔루션을 구축할 때 API 활동을 모니터링하고 문제를 해결할 수 있습니다.
- 파인 투닝된(사용자 정의) 모델은 서비스 관리 키를 사용하여 암호화되고 저장되며, 사용자만 엔드포인트를 통해 사용자 정의 모델에 액세스할 수 있습니다.
- 고객이 키를 생성하고 제어하여 파인 투닝된 모델을 암호화할 수 있도록 **고객 관리 키(CMK)** 지원
- 파인 투닝 작업의 VPC 구성 지원

# Data protection

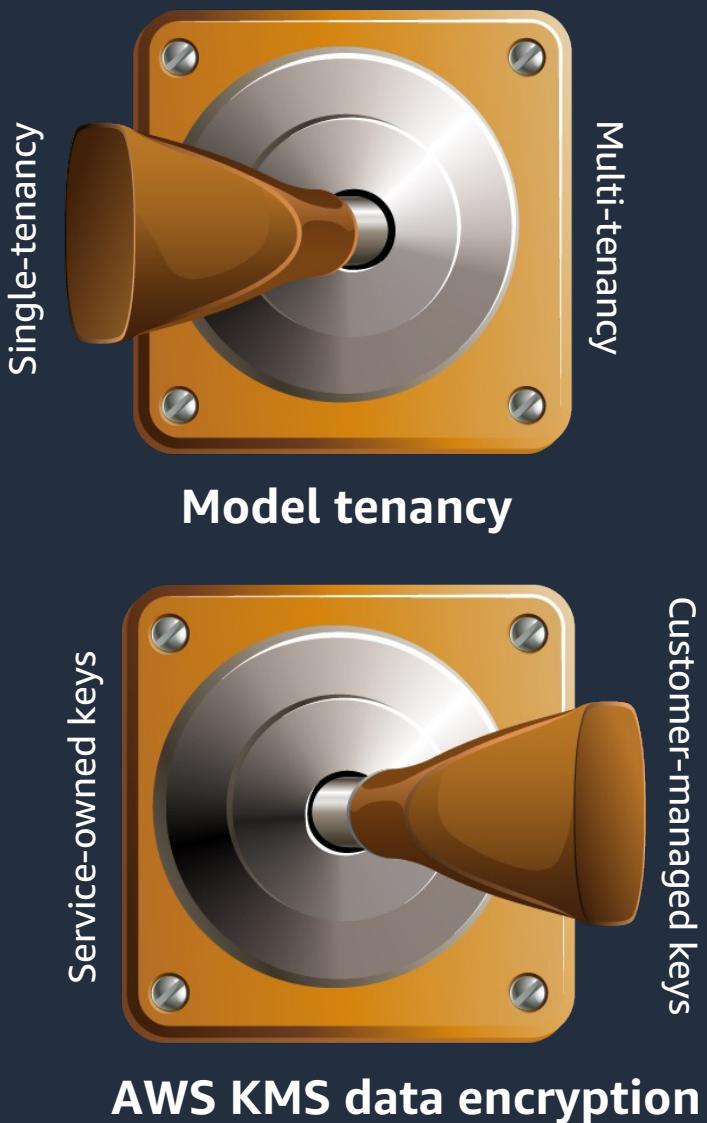
## ENCRYPTION



# Configurable security controls



**Data privacy**



**Model fine tuning**



**Access management**

# On-demand vs. provisioned compute capacity



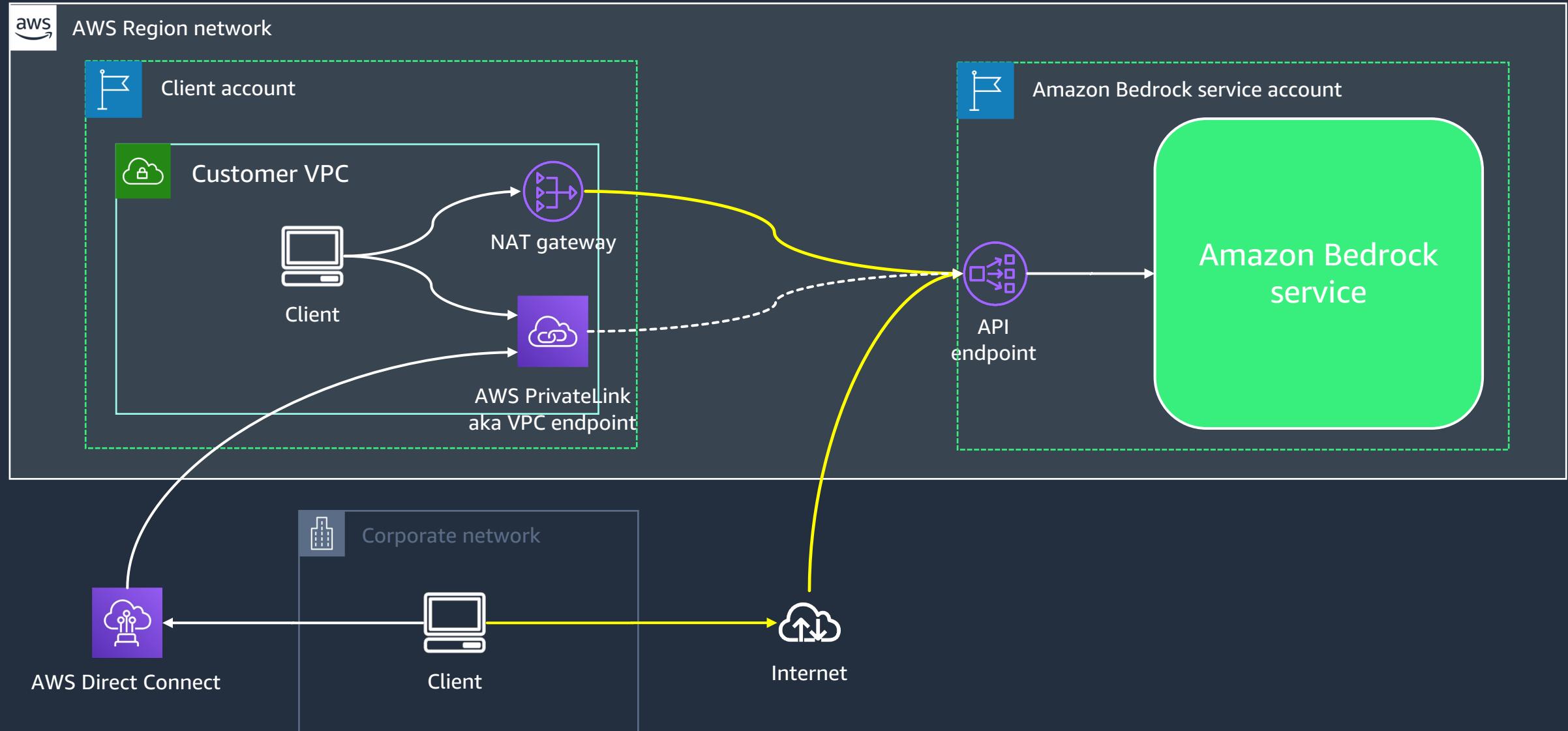
Provisioned  
capacity compute



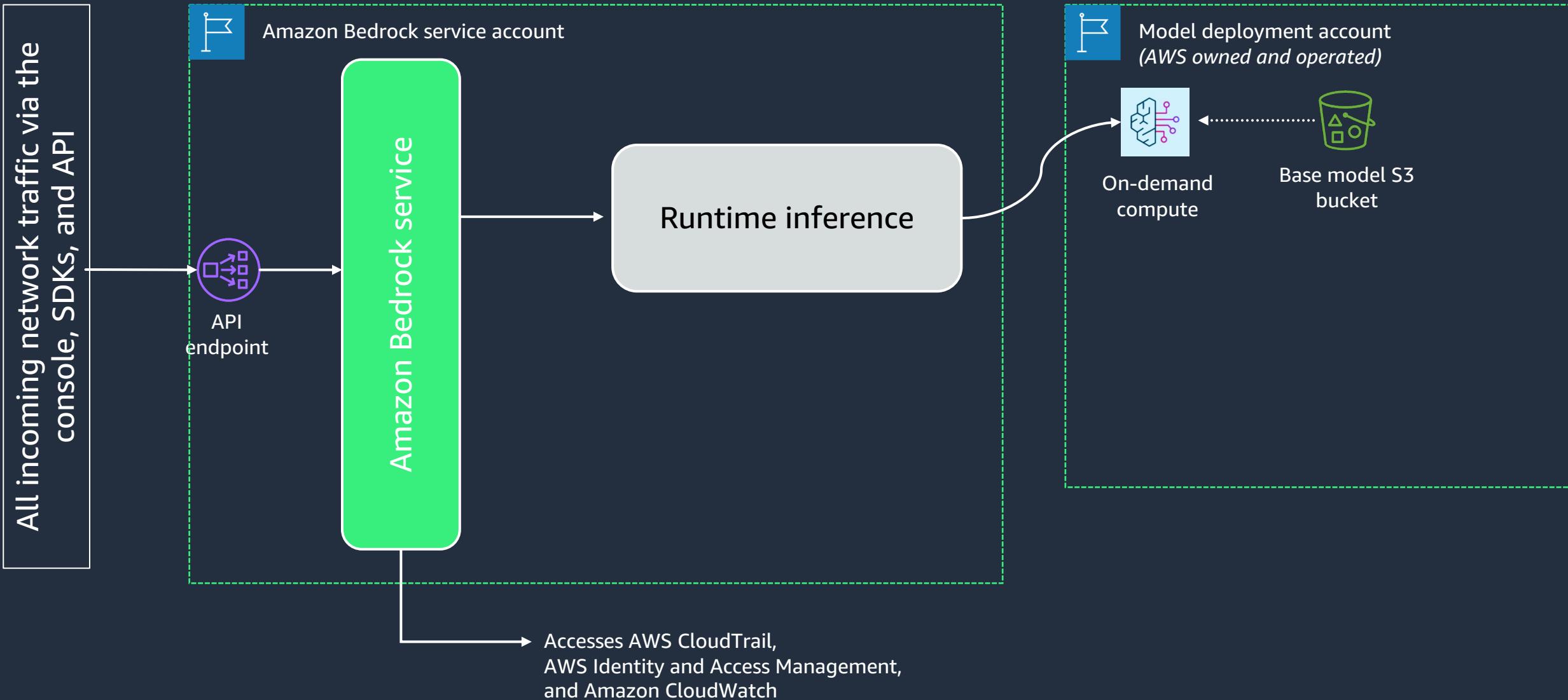
On-demand  
compute

- |  |   |
|--|---|
| <ul style="list-style-type: none"><li>1. 단일 고객에게 배포 가능</li><li>2. 고객이 파인 튜닝했을 수 있는 기본 모델의 비공개 사본을 보유합니다.</li></ul>   | <ul style="list-style-type: none"><li>1. 모든 고객이 배포 가능</li><li>2. 지원되는 모델의 기준 버전을 보유합니다.</li></ul> |
| <ul style="list-style-type: none"><li>3. 추론 요청의 입력 또는 출력 텍스트는 배포에서 모델을 학습하는 데 사용되지 않습니다.</li><li>4. 모델 배포는 Bedrock 서비스 팀이 소유하고 운영하는 AWS 계정 내에서 이루어집니다.</li></ul> | <ul style="list-style-type: none"><li>5. 모델 공급업체는 고객 데이터에 액세스할 수 없습니다.</li></ul>                  |

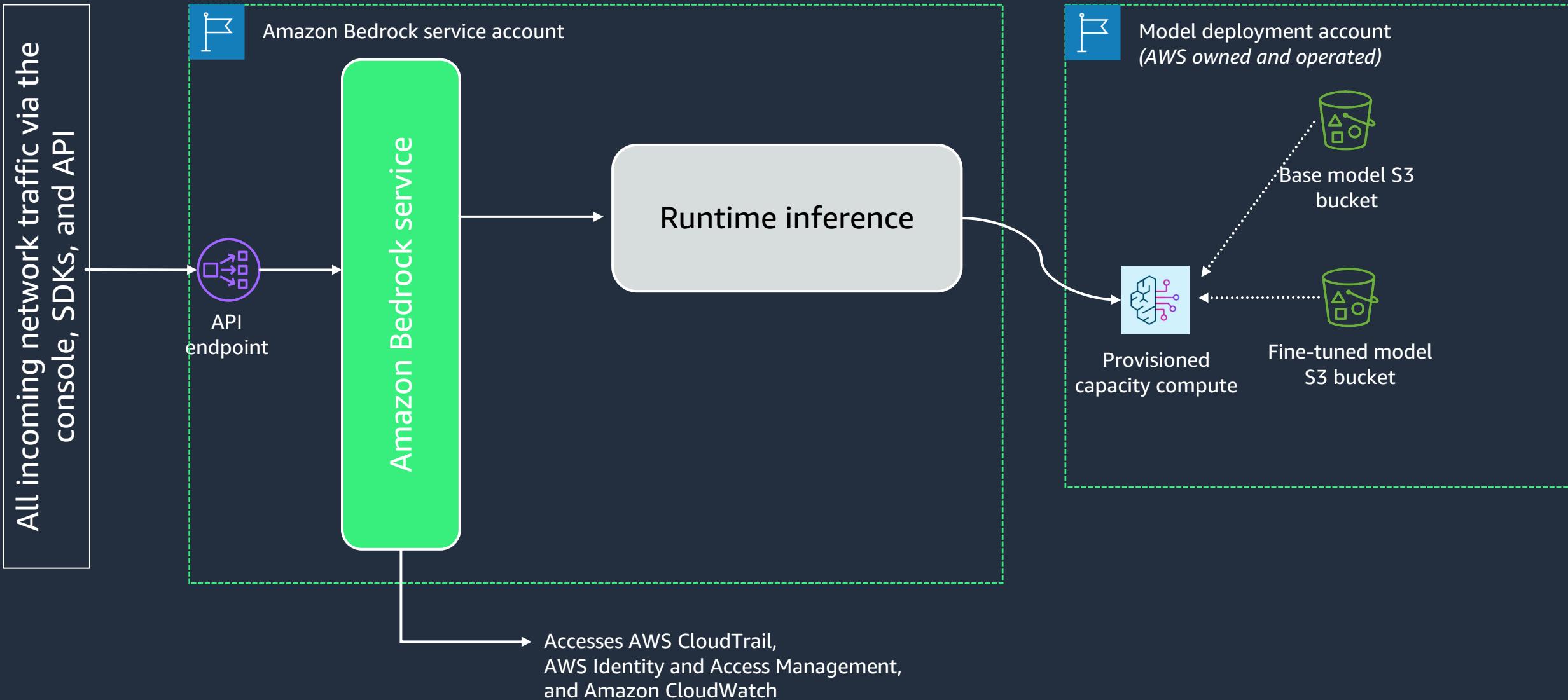
# Client connectivity



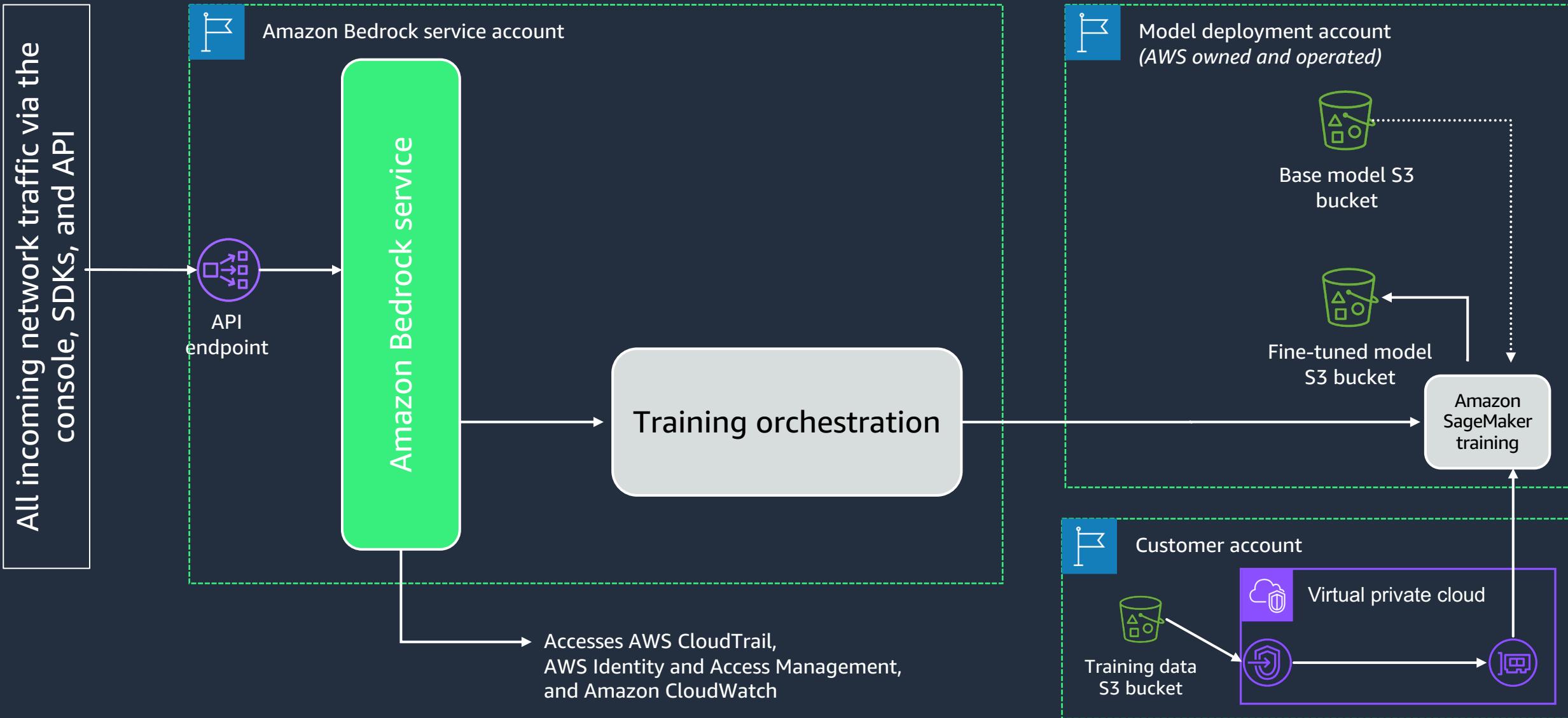
# On-demand compute architecture overview



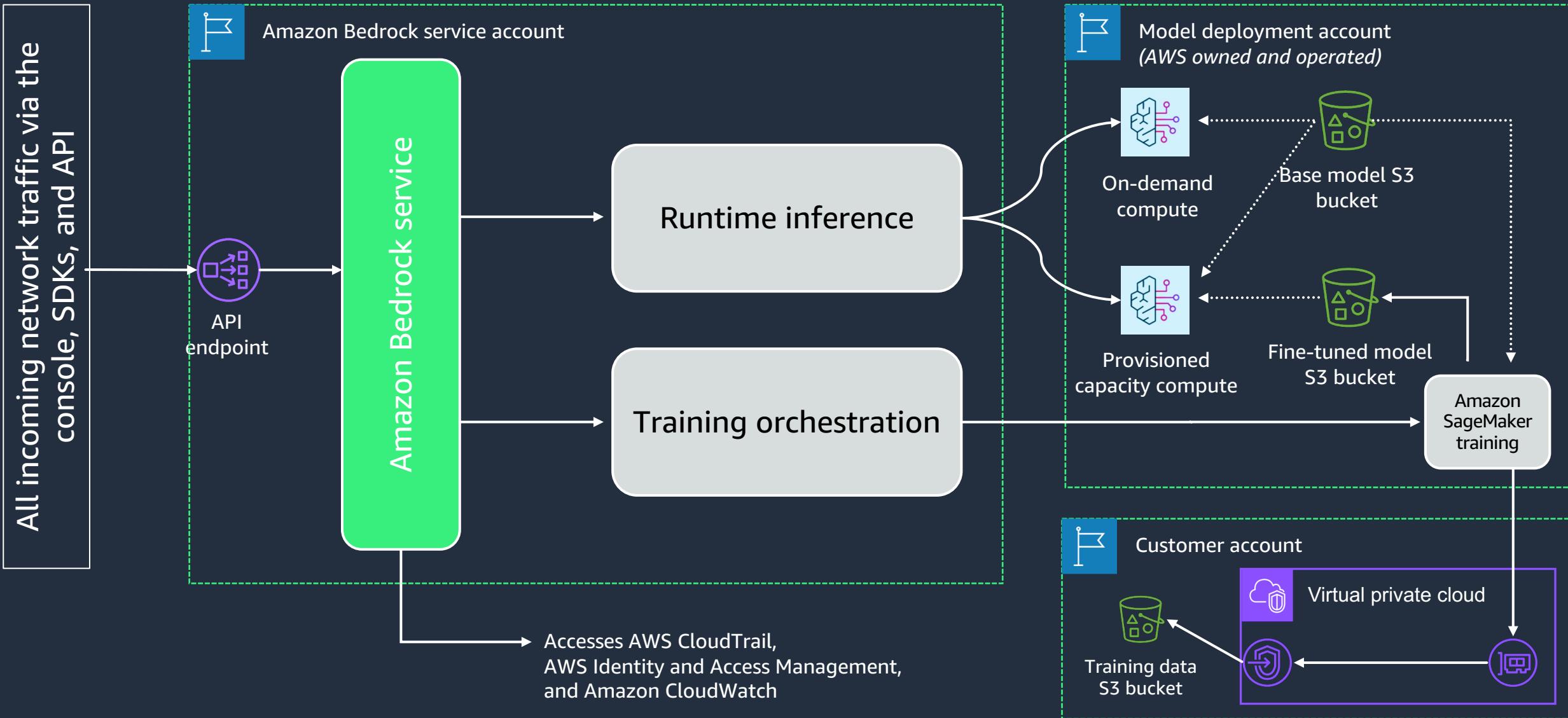
# Provisioned capacity architecture overview



# Model fine-tuning architecture overview



# Complete architecture overview



# AWS Identity and Access Management



IAM



- Identity-based policies
- Actions
- Resources
- Tags (ABAC)

# IAM/SCP – Example deny policy

```
{  
    "version": "2012-10-17",  
    "Statement":  
    {  
        "Sid": "DenyInferenceForModelX",  
        "Effect": "Deny",  
        "Action": "bedrock:InvokeModel",  
        "Resource": "arn:aws:bedrock::::foundation-model/<name-of-model>"  
    }  
}
```



# Understanding Amazon Bedrock inference rates, limits, and throughput

# Summary of Amazon Bedrock inference options



## On-demand

종량제, 약정 없이 사용량 기반  
결제

- › LLM의 입력 및 출력 토큰 수에 따른 가격 책정
- › 처리량과 자연 시간에 대한 요구 사항이 보다 완화된 프로토타이핑, POC, 소규모 워크로드에 적합합니다.
- › 분당 요청 수(RPM) 및 분당 토큰 수(TPM) 제한 적용



## Provisioned throughput

애플리케이션의 성능 요구 사항을 충족할 수 있는  
충분한 처리량 제공

- › 고정 비용으로 처리량 보장
- › 더 높은 처리량 제공
- › 1개월 또는 6개월의 유연한 약정 기간
- › 시간당 요금 지불, 약정 기간 연장 시 할인 적용
- › 프로덕션 워크로드 또는 사용자 정의 모델에 대한 추론에 적합

# Bedrock Workshop

HANDS-ON

Github



Workshop Studio





# Thank you!