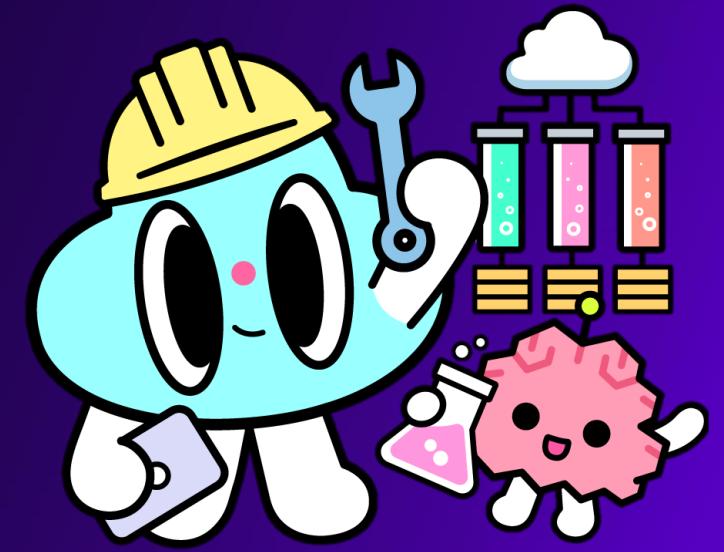


AWS RE:INVENT

re:Cap



2024-12-17

AWS re:Invent 2024 re:Cap > Amazon Bedrock

장문기

aws COMMUNITY



Agenda

- AWS re:Invent 2024 소개
- Keynote 소개
- Keynote에서 Bedrock
- Bedrock Blogs & Breakout Sessions

About AWS re:Invent 2024

AWS re:Invent는 글로벌 클라우드 컴퓨팅 커뮤니티를 위해 AWS가 주최하는 학습 컨퍼런스입니다.

60 000+

attendees

400 000+

watching keynotes (*online*)

900+

sessions

3500

speakers

546

updates (Oct 30th – Dec 6th)

422

pre-re:Invent (Oct 30th – Nov 30th)

124

re:Invent (Dec 1st – Dec 6th)

1

deck with selected 221 updates

관련 링크

[Homepage] – 영상 , 세션등

- <https://reinvent.awsevents.com/on-demand/>

[영상 - Youtube]

- <https://www.youtube.com/@AWSEventsChannel/videos>

[영상 검색]

- <https://lilys.ai/collections/21534>

[블로그 : Top announcements of AWS re:Invent 2024]

- 한국어
<https://aws.amazon.com/ko/blogs/korea/top-announcements-of-aws-reinvent-2024/>
- 영어
<https://aws.amazon.com/blogs/aws/top-announcements-of-aws-reinvent-2024/>

[Bedrock 관련 영상 요약]

- <https://lilys.ai/collections/21363> by 김도율님

기조 연설

Monday Night Live with Peter DeSantis

SVP, UTILITY COMPUTING, AWS

CEO Keynote with Matt Garman

CEO, AWS

Dr. Swami Sivasubramanian Keynote

VP, AI AND DATA, AWS

AWS Partner Keynote with Dr. Ruba Borno

VP, GLOBAL SPECIALISTS AND PARTNERS, AWS

Dr. Werner Vogels Keynote

VP AND CTO, AMAZON.COM

요약 - <https://lilys.ai/collections/21534>

영상 - https://www.youtube.com/playlist?list=PL2yQDdvIhXf_aPLMfxECsw-UIbEg6uy42



ADVANCED PACKAGING

Trainium2

Two compute chips and four HBM modules
in a single package

키노트 측면에서, AWS re:Invent는 항상
유틸리티 컴퓨팅 수석 부사장 Peter DeSantis의
먼데이 나이트 라이브로 시작합니다.



Most performant
AWS-built chip

+30%

Compute performance
for databases, web, and
Java apps vs Graviton3



DAVID BROWN
VP, AWS COMPUTE & NETWORKING

AWS의 Infra의 핵심 중 하나는
ARM Architecture 기반의 Graviton



AWS Nitro System

PERFORMANCE

Better networking and storage performance

AGILITY

Ability to turn any server into a full EC2 instance

SECURITY

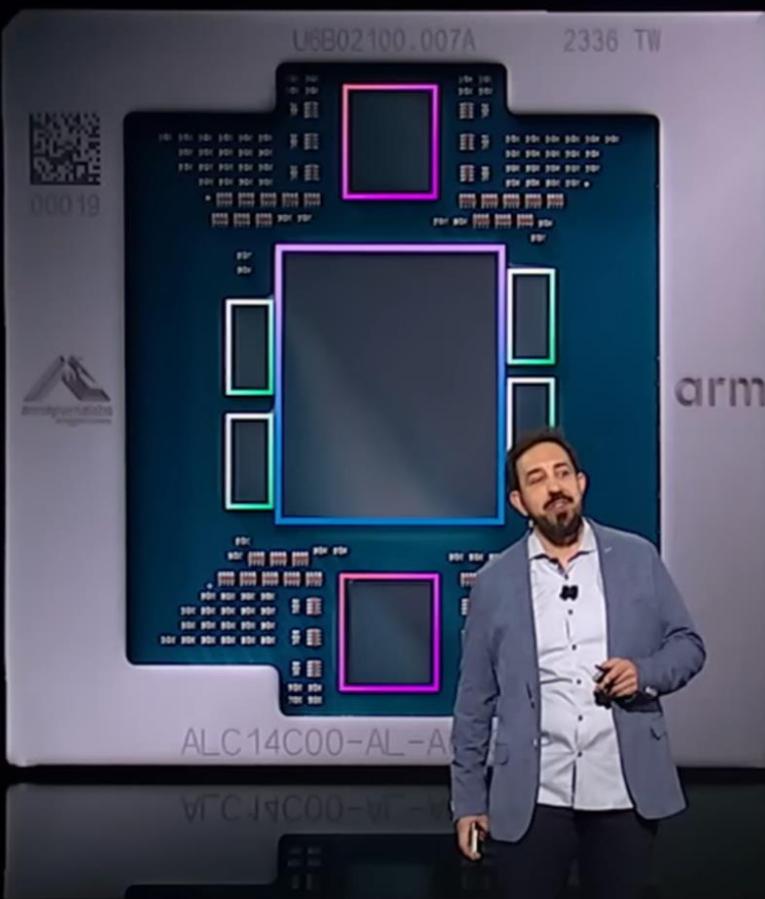
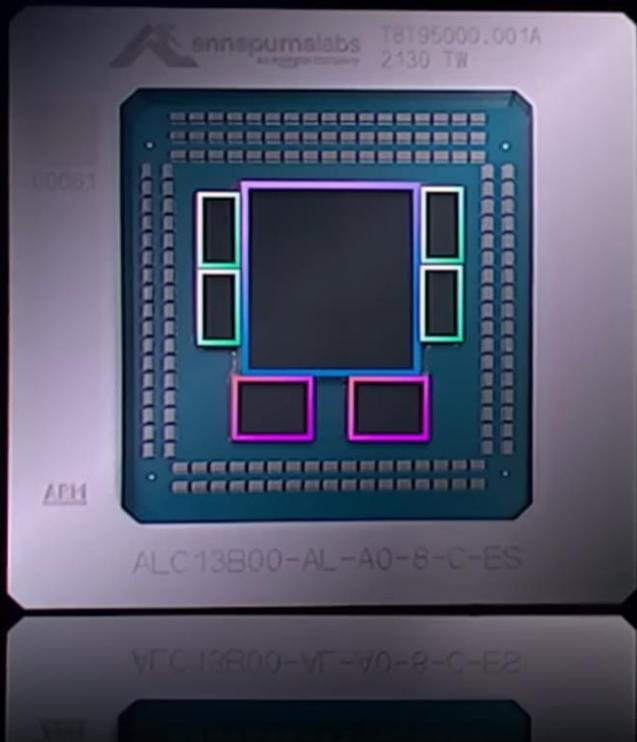
Hardware-enforced separation



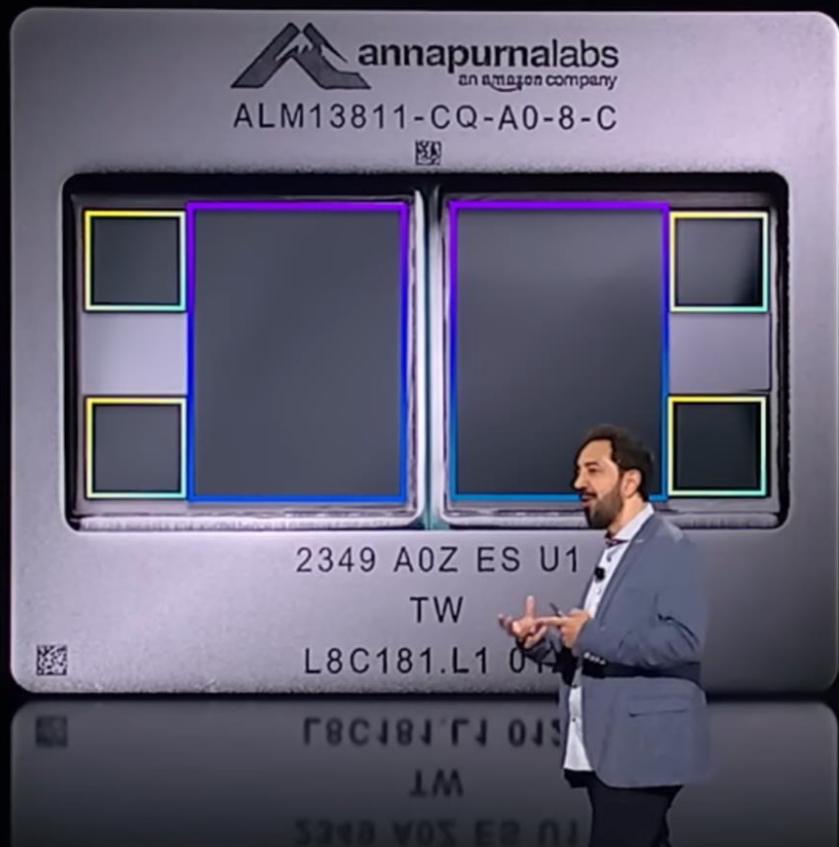
Advanced
packaging

GRAVITON4

GRAVITON3



- COMPUTE CHIP
- HBM



ADVANCED PACKAGING

Trainium2

Two compute chips and four HBM modules
in a single package



AWS Trainium2 UltraServer

2 racks, 4 servers, and 64 Trainium2 accelerators connected together via ultra-fast NeuronLink to provide you with the single greatest performance instance

83.2

PFLOPS
COMPUTE PERF
(DENSE)

332.8

PFLOPS
COMPUTE PERF
(SPARSE)

6

TB
HBM CAPACITY

185

TB/s
HBM BANDWIDTH

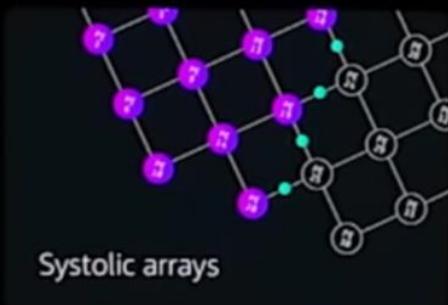
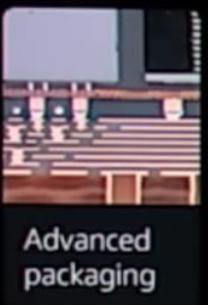
CHIP-CHIP INTERCONNECT
NeuronLink

2

TB/s
BANDWIDTH

1

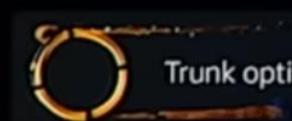
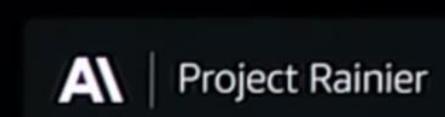
μs
LATENCY



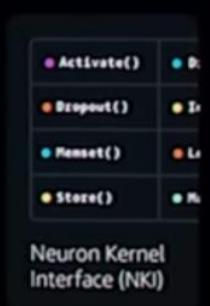
AWS Trainium2
Server



AWS Trainium2
UltraServer



SIDR





NEW

Latency-optimized inference option for Amazon Bedrock

Lower latency, performance-optimized
option for on demand inference

AVAILABLE IN PREVIEW TODAY

LEADING INFERENCE PERFORMANCE
Llama 3.1 405B
on Trn2

3.9s

TIME TO OUTPUT 100 TOKENS
(10K INPUT TOKENS)

PUBLIC BENCHMARK DATA FROM ARTIFICIAL ANALYSIS



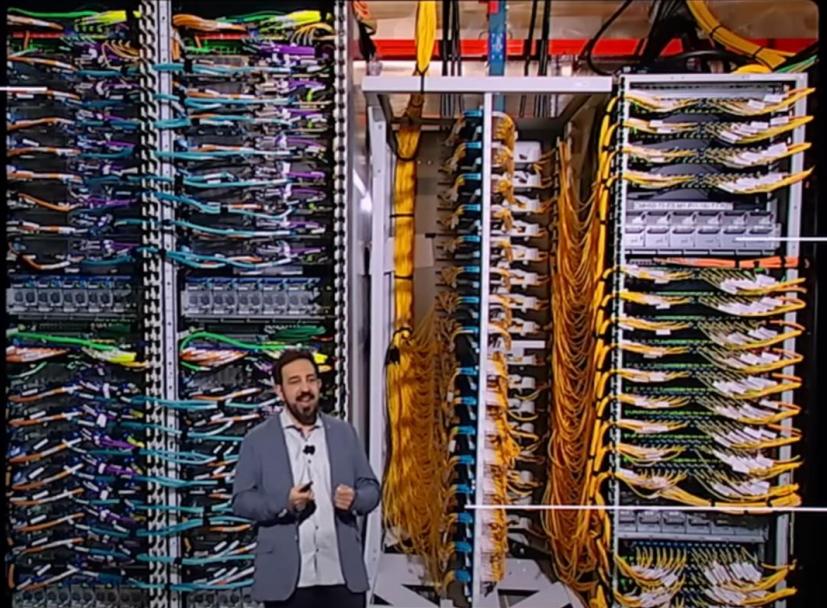


수십만개의 Trainium2 Cluster에서 Anthropic 학습

10p10u

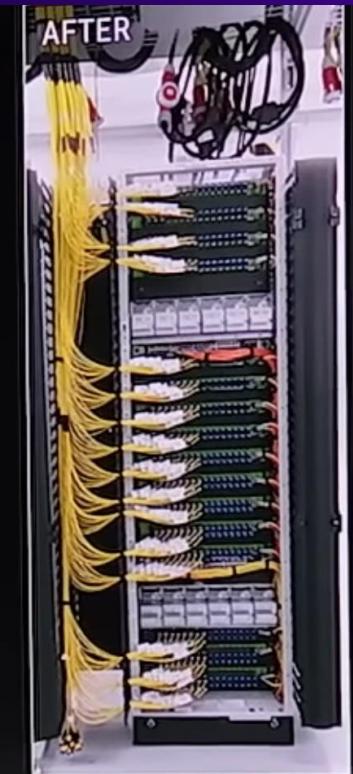
AWS network fabric designed to deliver 10s of petabits of bandwidth with under 10 microseconds of latency

COMPUTE RACK



NETWORK BRICK

PATCH RACK



16:1

REDUCTION IN NUMBER
OF CONNECTORS,
IN A SINGLE CABLE

Firefly Optic Plug

CUSTOM AWS-DESIGNED
Enabling loopback testing
and dust mitigation

CUSTOM NETWORK SWITCHES



CUSTOM AWS-DESIGNED
Fiber optical trunk cable



AWS re:Invent

다음 날에는 Matt Garman(AWS CEO)이 진행하는
메인 키노트가 있었습니다.

Amazon Bedrock

The easiest and fastest way to build and scale generative AI applications

Broadest choice of leading models

Optimize for cost, latency, and accuracy

Customize with your data

Apply safety and responsible AI checks

Build and orchestrate agents

Custom Model Import

Latency-optimized inference

Fine-tuning

Guardrails

Agents

Model Evaluation

NEW Model Distillation

Knowledge Bases managed RAG

NEW Automated Reasoning checks

NEW Multi-agent collaboration



AWS SEC , PRIVACY, AND RELIABILITY BUILT-IN

NEW

Amazon Bedrock Model Distillation

Create faster, more cost-effective models

AVAILABLE IN PREVIEW TODAY

Easily transfer knowledge from a large, complex model to a smaller one

Distilled models up to 500% faster and 75% less expensive

Anthropic, Meta, and Amazon models



**TEACHER
MODEL**
Highly intelligent model



Add your specific prompts

→ DISTILL INTO →

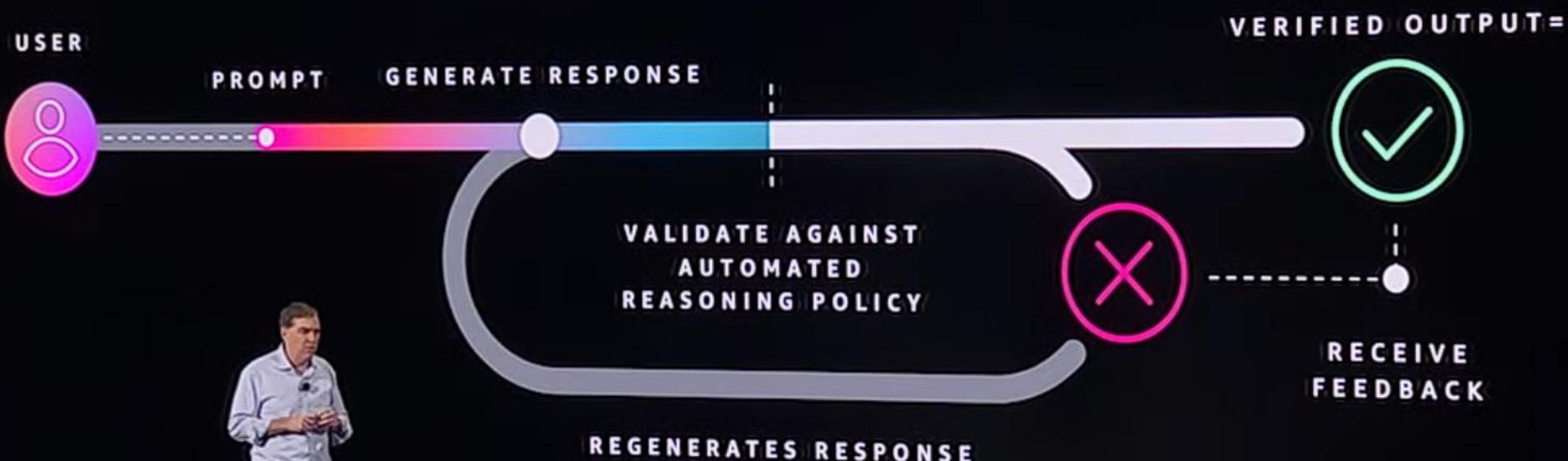


**STUDENT
MODEL**
Faster, cost-efficient, custom model

Proprietary data synthesis to enhance training data and fine-tune



Automated Reasoning checks



Amazon Bedrock allows you
to easily build agents for complex tasks



SEND PROMPT TO AGENT:
FILL FORM



AGENT USES
REASONING TO
BREAK DOWN TASK



ACCESES
RELEVANT DATA
SOURCES AND TOOLS



EXECUTE ACTION:
FILL FORM AND
CLICK SUBMIT



NEW

Amazon Bedrock multi-agent collaboration

Easily build, deploy, and orchestrate teams of agents that work together to handle complex, multi-step tasks

AVAILABLE IN PREVIEW TODAY



Orchestrate agents
without complex coding

Accelerate tasks with
agents working in parallel



Amazon Nova

State-of-the-art foundation models that deliver frontier intelligence and industry-leading price performance

Amazon Nova **Micro**

Text only model that
delivers the lowest latency
responses at very low cost

GENERALLY AVAILABLE

Amazon Nova **Lite**

Very low-cost multimodal
model that is lightning-fast
for processing image, video,
and text inputs

GENERALLY AVAILABLE

Amazon Nova **Pro**

Highly capable multimodal model
with the best combination of
accuracy, speed, and cost for a
wide range of tasks

GENERALLY AVAILABLE

Amazon Nova **Premier**

Most capable multimodal model
for complex reasoning tasks and
for use as the best teacher for
distilling custom models

COMING SOON



NEW

Amazon Nova

NOVA MICRO | NOVA LITE | NOVA PRO | NOVA PREMIER

State-of-the-art foundation models that deliver frontier intelligence and industry-leading price performance



75% more cost effective

Fastest models in their respective intelligence classes in Bedrock

Support for fine-tuning to boost accuracy

Distillation to train smaller, more efficient models that are highly accurate, fast, and cheaper to run

Integrated with Bedrock Knowledge Bases for RAG to ground responses in your own data

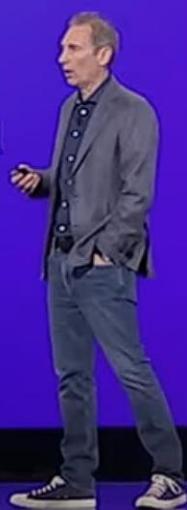
Optimized for agentic applications that require interacting with your systems and tools through APIs

NEW

Amazon Nova Canvas

State-of-the-art image generation model

GENERALLY AVAILABLE



Edit images using natural language text inputs

Provides controls for color scheme and layout

Built-in controls for safe and responsible use of AI, including watermarking for traceability, and content moderation to limit the generation of harmful content

NEW

Amazon Nova Reel

State-of-the-art video generation model

GENERALLY AVAILABLE

Provides camera motion control, including pan motion, 360-degree rotation, and zoom

Built-in controls for safe and responsible use of AI, including watermarking and content moderation

Support for 6-second videos today,
2-minute coming soon



What's next for Amazon Nova?

NEW

Amazon Nova **Speech-to-Speech**

COMING SOON

NEW

Amazon Nova **Any-to-Any**

COMING SOON



Amazon Bedrock

The easiest and fastest way to build and scale generative AI applications

Choose the best model

Broadest choice of industry leading models
Custom model import
Model evaluation

Optimize for cost, latency, and accuracy

Latency-optimized inference
Model distillation

Customize with your data

Fine-tuning
Knowledge Bases

Apply safety and responsible AI checks

Guardrails
Automated reasoning checks

Build and orchestrate agents

Agents
Multi-agent collaboration

AWS SECURITY, PRIVACY, AND RELIABILITY

BUILT-IN

Amazon Nova

State-of-the art foundation models that deliver frontier intelligence and industry leading price performance



Micro

Text only model that delivers the lowest latency responses at very low cost



Lite

Very low-cost multimodal model that is lightning-fast for processing image, video, and text inputs

Canvas

State-of-the-art image generation model

Pro

Highly capable multimodal model with the best combination of accuracy, speed, and cost for a wide range of tasks

Premier

Most capable multimodal model for complex reasoning tasks and for use as the best teacher for distilling custom models

Reel

State-of-the-art video generation model





Software engineering AI
for large enterprises



NEW

poolside Assistant

malibu

Tackles complex software
engineering challenges

COMING SOON

point

Low latency code completion
that predicts developers' needs

COMING SOON

NEW

Luma AI

High-quality video generation
from text and images

COMING SOON



Hyper-fast creation time for high-quality generations
Experiment with fluid, cinematic, and naturalistic camera motions
Create videos with great character consistency and accuracy

NEW

Amazon Bedrock supports prompt caching

Cache repetitive context in prompts across multiple API calls



- Securely cache entire prompts
- Enhance accuracy with longer prompts
- Reduce latency and save on costs

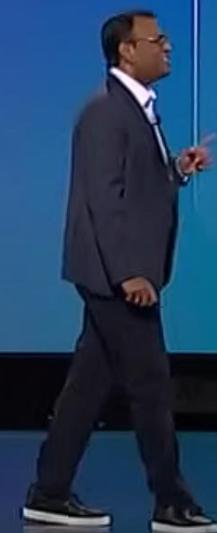
AVAILABLE IN PREVIEW TODAY

NEW

Amazon Bedrock Intelligent Prompt Routing

Automatically route prompts to different foundation models to optimize response quality and lower costs

AVAILABLE IN PREVIEW TODAY



Provides a single endpoint to efficiently route prompts

Meets cost and latency thresholds with advanced prompt matching techniques

Reduces application development costs by up to 30%

NEW

Amazon Bedrock Knowledge Bases supports structured data retrieval

Seamlessly integrate structured data for RAG

GENERALLY AVAILABLE

Use data stored in Amazon SageMaker Lakehouse, Redshift and S3 Tables

Reduce application development time from months to days

Improve the accuracy of your queries with customization context



NEW

Amazon Bedrock Knowledge Bases now supports GraphRAG

Generate more relevant responses for generative AI
applications using knowledge graphs

AVAILABLE IN PREVIEW TODAY



Generate knowledge graphs to link
relationships across data sources

Build more comprehensive, explainable
generative AI applications

Enhance transparency of source
information for better fact verification

NEW

Amazon Bedrock Data Automation

Transform unstructured multimodal data for generative AI applications and analytics

AVAILABLE IN PREVIEW TODAY



Extract, transform, and generate structured data from multimodal content

Generate customized outputs based on your requirements and business rules

Streamline application workflows with a fully managed, single API experience

NEW

Amazon Bedrock Guardrails Multimodal toxicity detection

Configurable safeguards for image content

AVAILABLE IN PREVIEW TODAY



Enhance security of multimodal
generative AI applications

Enable consistent policy control

Available for all foundation models in
Amazon Bedrock with image support

Amazon Bedrock

The easiest and fastest way to build and scale generative AI applications



Luma RAY2

The slide features a dark background with four generated video frames arranged in a row. Each frame is accompanied by descriptive text below it.

- 

High production quality video generation
- 

Intelligent text-to-video, image-to-video, video-to-video
- 

Consistent characters & scenes in minute-long videos
- 

Real-time generation, for the first time ever

NEW

Amazon Bedrock Marketplace

Discover, test, and use over 100 popular, emerging, and specialized models in Amazon Bedrock

GENERALLY AVAILABLE

Streamline development workflows with a unified console experience

Deploy models on managed endpoints with custom scaling policies

Leverage Amazon Bedrock APIs, tools, and security

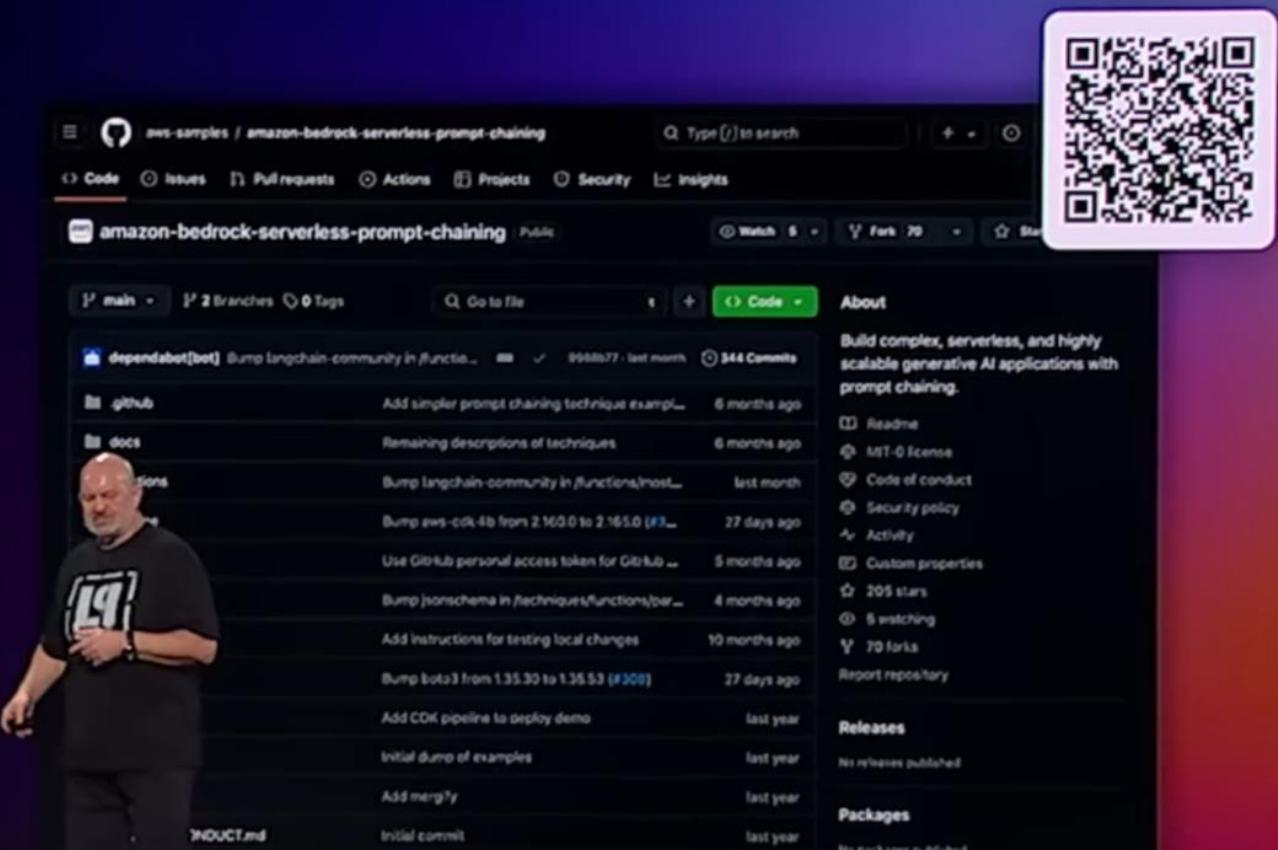




Dr. Werner Vogels
VP and CTO Amazon.com

CLARE LIGUORI'S

Amazon Bedrock Serverless Prompt Chaining



The GitHub repository page shows the following commit history:

- [dependabot[bot]] Bump langchain-community in /functions/main · 998fb77 · last month · 344 Commits
- [It's_github] Add simpler prompt chaining technique example · 6 months ago
- [It's_docs] Remaining descriptions of techniques · 6 months ago
- [It's_docs] Bump langchain-community in /functions/main · last month
- [It's_docs] Bump aws-cdk-lib from 2.160.0 to 2.165.0 (#308) · 27 days ago
- [It's_docs] Use GitHub personal access token for GitHub Actions · 5 months ago
- [It's_docs] Bump jsonschema in /techniques/functions/bar · 4 months ago
- [It's_docs] Add instructions for testing local changes · 10 months ago
- [It's_docs] Bump boto3 from 1.35.30 to 1.35.53 (#308) · 27 days ago
- [It's_docs] Add CDK pipeline to deploy demo · last year
- [It's_docs] Initial dump of examples · last year
- [It's_docs] Add mergeify · last year
- [INDUCT.md] Initial commit · last year

About
Build complex, serverless, and highly scalable generative AI applications with prompt chaining.

Readme
MIT-0 license

Code of conduct

Security policy

Activity

Custom properties

205 stars

5 watching

70 forks

Report repository

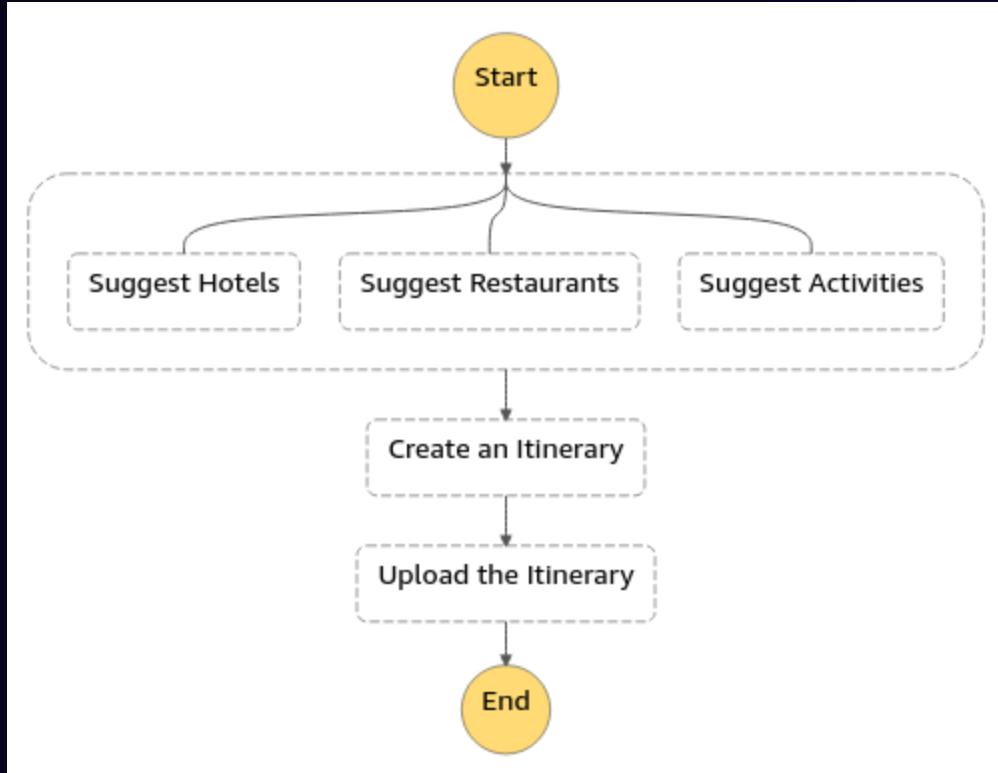
Releases
No releases published

Packages
No packages published

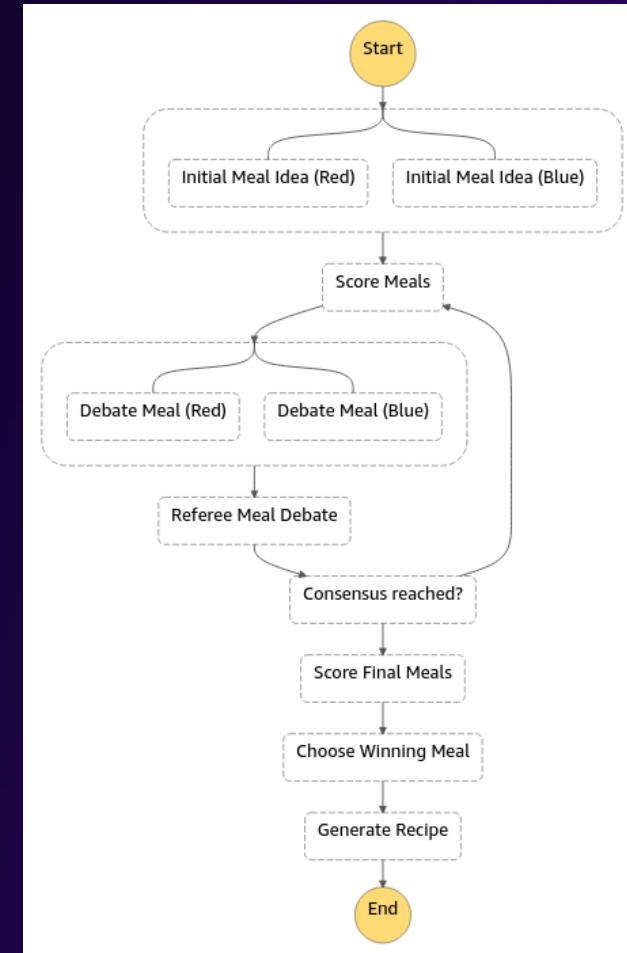
QR Code

<https://github.com/aws-samples/amazon-bedrock-serverless-prompt-chaining>

AWS Step Functions 와 Amazon Bedrock을 사용하여 프롬프트 체이닝으로
복잡하고, 서버리스하며, 높은 확장성을 가진 생성형 AI 애플리케이션을 구축하는 예제를 제공



여행 계획하기



식사 계획하기

GUIDED LEARNING

Serverless Agentic Workflows with Amazon Bedrock

The video player displays a course titled "jupyter Lesson_1" from DeepLearning.AI. The video content shows a man speaking, and the Jupyter notebook interface is overlaid on the video frame. The notebook code is as follows:

```
import boto3
bedrock_agent = boto3.client(service_name='bedrock-agent', region_name='us-east-1')

create_agent_response = bedrock_agent.create_agent(
    agentName="mugs-customer-support-agent",
    foundationModel="anthropic.claude-3-haiku-20240307-v1:0",
    instruction="""
        You are an advanced AI agent acting as a front line customer support agent.
        Your goal is to provide helpful and informative responses to customer inquiries.
        You should always be polite and friendly.
        You must not generate any harmful or inappropriate content.
    """,
    resourceRoleArn=roleArn
)

agent_response = create_agent_response['agent']
agent_id = agent_response['id']
agent_status = agent_response['status']

if agent_status != 'NOT_PREPARED':
    print(f"Agent {agent_id} is ready to serve!")

prepare_agent(response)
```

A QR code is located in the top right corner of the video player.

<https://www.deeplearning.ai/short-courses/serverless-agentic-workflows-with-amazon-bedrock/>

AWS re:Invent 2024: Launch Summary - Bedrock

AWS re:Invent 2024 Launch Summary (1/5)

Analytics

[Announcing scenarios analysis capability of Amazon Q in QuickSight \(preview\)](#)

[Announcing Amazon S3 Metadata \(Preview\) – Easiest and fastest way to manage your metadata](#)

[Amazon DynamoDB zero-ETL integration with Amazon SageMaker Lakehouse](#)

[AWS expands data connectivity for Amazon SageMaker Lakehouse and AWS Glue](#)

[Amazon S3 Access Grants now integrate with AWS Glue](#)

[Amazon SageMaker Lakehouse integrated access controls now available in Amazon Athena federated queries](#)

[Introducing AWS Glue 5.0](#)

[Introducing the next generation of Amazon SageMaker](#)

[AWS announces Amazon SageMaker Lakehouse](#)

[Amazon SageMaker Lakehouse and Amazon Redshift support for zero-ETL integrations from eight applications](#)

[AWS Glue Data catalog now automates generating statistics for new tables](#)

Analytics

[Introducing Amazon SageMaker Data and AI Governance](#)

[Announcing the preview of Amazon SageMaker Unified Studio](#)

[Data Lineage is now generally available in Amazon DataZone and next generation of Amazon SageMaker](#)

[Amazon Q in QuickSight unifies insights from structured and unstructured data](#)

[Amazon Q Business now provides insights from your databases and data warehouses \(preview\)](#)

[Amazon OpenSearch Service zero-ETL integration with Amazon Security Lake](#)

[AWS Clean Rooms now supports multiple clouds and data sources](#)

Business Applications

[Amazon Q Business now provides insights from your databases and data warehouses \(preview\)](#)

[Amazon Connect Contact Lens now supports external voice](#)

[Amazon Connect now supports external voice transfers](#)

[Amazon Connect Contact Lens now automatically categorizes your contacts using generative AI](#)

[Amazon Connect launches AI guardrails for Amazon Q in Connect](#)

[Amazon Connect launches new intraday forecast dashboards](#)

[Amazon Connect launches AI assistant for customer segments and trigger-based campaigns](#)

[Amazon Connect now supports WhatsApp Business messaging](#)

[Amazon Connect Contact Lens now automates agent performance evaluations using generative AI](#)

[Amazon Connect launches generative AI-powered self-service with Amazon Q in Connect](#)

[AWS announces Salesforce Contact Center with Amazon Connect \(Preview\)](#)

Application Integration

[Amazon SageMaker Lakehouse and Amazon Redshift support for zero-ETL integrations from eight applications](#)

[Amazon Redshift now supports multi-agent collaboration](#)

[Amazon EventBridge and AWS Step Functions announce integration with private APIs](#)

AWS re:Invent 2024 Launch Summary (2/5)

Compute

[Amazon EC2 Trn2 instances are generally available](#)

[Announcing Amazon Elastic VMware Service \(Preview\)](#)

[Deploy GROW with SAP on AWS from AWS Marketplace](#)

[Amazon EC2 P5en instances, optimized for generative AI and HPC, are generally available](#)

[Announcing Amazon EC2 I8g instances](#)

[Amazon Web Services announces declarative policies](#)

[Amazon CloudWatch Container Insights launches enhanced observability for Amazon ECS](#)

[Introducing Amazon EC2 next generation high density Storage Optimized I7ie instances](#)

[AWS Marketplace now offers EC2 Image Builder components from independent software vendors](#)

[Amazon EC2 introduces Allowed AMIs to enhance AMI governance](#)

[Announcing Amazon EKS Auto Mode](#)

[Announcing Amazon EKS Hybrid Nodes](#)

[AWS simplifies the use of third-party block storage arrays with AWS Outposts](#)

Containers

[VPC Lattice now includes TCP support with VPC Resources](#)

[Announcing Amazon EKS Auto Mode](#)

[Announcing Amazon EKS Hybrid Nodes](#)

[Amazon Connect launches simplified conversational AI bot creation](#)

[Amazon Connect now supports WhatsApp Business messaging](#)

[Amazon Connect launches generative AI-powered self-service with Amazon Q in Connect](#)

[Amazon Connect now makes it easier to collect sensitive customer data within chats](#)

Customer Engagement

[Introducing the AWS Digital Sovereignty Competency](#)

[Amazon Connect now provides the ability to record audio during IVR and other automated interactions](#)

[Amazon Connect Contact Lens now supports external voice](#)

[Amazon Connect now supports external voice transfers](#)

[Amazon Connect launches AI guardrails for Amazon Q in Connect](#)

[Amazon Connect Contact Lens launches built-in dashboards to analyze conversational AI bot performance](#)

Databases

[Announcing Amazon Aurora DSQL \(Preview\)](#)

[Amazon DynamoDB global tables previews multi-Region strong consistency](#)

[AWS DMS Schema Conversion now uses generative AI](#)

[Announcing the general availability of Amazon MemoryDB Multi-Region](#)

[Oracle Database@AWS is now in limited preview](#)

[Amazon Bedrock Knowledge Bases now supports GraphRAG \(preview\)](#)

[Amazon DynamoDB zero-ETL integration with Amazon SageMaker Lakehouse](#)

AWS re:Invent 2024 Launch Summary (3/5)

Databases (continued)

[Amazon Q Business now provides insights from your databases and data warehouses \(preview\)](#)

[Announcing Amazon EC2 I8g instances](#)

[AWS DMS Schema Conversion now uses generative AI](#)

Developer Tools

[Amazon Q Developer adds operational investigation capability \(Preview\)](#)

[Amazon Q Developer now provides transformation capabilities for .NET porting \(Preview\)](#)

[Amazon Q Developer can now automate code reviews](#)

[Announcing GitLab Duo with Amazon Q \(Preview\)](#)

[Amazon Q Developer can now generate documentation within your source code](#)

[Amazon Q Developer transformation capabilities for mainframe modernization are now available \(Preview\)](#)

Front End Web & Mobile

[Storage Browser for Amazon S3 is now generally available](#)

[Amazon Connect now supports WhatsApp Business messaging](#)

Machine Learning

[Amazon Bedrock Knowledge Bases now supports structured data retrieval](#)

[Task governance is now generally available for Amazon SageMaker HyperPod](#)

[AWS announces Amazon SageMaker Partner AI Apps](#)

[Amazon Bedrock Guardrails supports multimodal toxicity detection for image content \(Preview\)](#)

[Announcing new AWS AI Service Cards to advance responsible generative AI](#)

[Amazon SageMaker HyperPod now provides flexible training plans](#)

[Amazon Bedrock Marketplace brings over 100 models to Amazon Bedrock](#)

[Amazon Bedrock announces preview of prompt caching](#)

[Amazon Q Developer can now guide SageMaker Canvas users through ML development](#)

[Amazon Bedrock Data Automation now available in preview](#)

[Amazon Bedrock Knowledge Bases now supports GraphRAG \(preview\)](#)

[Announcing Amazon SageMaker HyperPod recipes](#)

[Amazon Bedrock Knowledge Bases now processes multimodal data](#)

[Announcing GenAI Index in Amazon Kendra](#)

[Amazon Q Developer adds operational investigation capability \(Preview\)](#)

[Amazon Q Developer can now automate code reviews](#)

[Amazon Bedrock Model Distillation is now available in preview](#)

[Amazon Q Business introduces over 50 actions for popular business applications and platforms](#)

[Amazon SageMaker Lakehouse and Amazon Redshift support for zero-ETL integrations from eight applications](#)

[Amazon Bedrock now supports multi-agent collaboration](#)

[Amazon Bedrock Guardrails now supports Automated Reasoning checks \(Preview\)](#)

AWS re:Invent 2024 Launch Summary (4/5)

Machine Learning (continued)

[Introducing Amazon SageMaker Data and AI Governance](#)

[Data Lineage is now generally available in Amazon DataZone and next generation of Amazon SageMaker](#)

[Amazon Q in QuickSight unifies insights from structured and unstructured data](#)

[Amazon Q Developer can now generate documentation within your source code](#)

[Announcing Amazon Bedrock IDE in preview as part of Amazon SageMaker Unified Studio](#)

[Announcing Amazon Nova foundation models available today in Amazon Bedrock](#)

[Amazon Q Business now provides insights from your databases and data warehouses \(preview\)](#)

[Amazon Q Developer transformation capabilities for mainframe modernization are now available \(Preview\)](#)

[Amazon EC2 P5en instances, optimized for generative AI and HPC, are generally available](#)

[Introducing latency-optimized inference for foundation models in Amazon Bedrock](#)

[Amazon Bedrock Knowledge Bases now supports RAG evaluation \(Preview\)](#)

[Amazon Bedrock Model Evaluation now includes LLM-as-a-judge \(Preview\)](#)

[Amazon Bedrock Knowledge Bases now provides auto-generated query filters for improved retrieval](#)

[Amazon Q Business adds support to extract insights from visual elements within documents](#)

[Amazon Bedrock Knowledge Bases now supports streaming responses](#)

[Amazon Bedrock now supports Rerank API to improve accuracy of RAG applications](#)

[Amazon Bedrock Knowledge Bases now supports custom connectors and ingestion of streaming data](#)

[PartyRock improves app discovery and announces upcoming free daily use](#)

[Announcing Amazon Q Developer transformation capabilities for VMware \(Preview\)](#)

[The Amazon Q index enhances software vendors' AI experiences](#)

Customer Engagement

[Amazon Web Services announces declarative policies](#)

[Amazon CloudWatch and Amazon OpenSearch Service launch an integrated analytics experience](#)

[Amazon CloudWatch adds network performance monitoring for AWS workloads using flow monitors](#)

[Amazon CloudWatch Container Insights launches enhanced observability for Amazon ECS](#)

[AWS Control Tower launches managed controls using declarative policies](#)

[AWS announces Amazon CloudWatch Database Insights](#)

Migration and Transfer

[Announcing AWS Transfer Family web apps](#)

[Announcing Amazon Q Developer transformation capabilities for VMware \(Preview\)](#)

Networking & Content Delivery

[AWS announces access to VPC resources over AWS PrivateLink](#)

AWS re:Invent 2024 Launch Summary (5/5)

Nova

(Generative AI)

[Announcing Amazon Nova foundation models available today in Amazon Bedrock](#)

Partner

[Buy with AWS accelerates solution discovery and procurement on AWS Partner websites](#)

[Respond and recovery more quickly with AWS Security Incident Response Partners](#)

[Start collaborating on multi-partner opportunities with Partner Connections \(Preview\)](#)

[Introducing the AWS Digital Sovereignty Competency](#)

[New streamlined deployment experience for Databricks on AWS](#)

[Introducing the AWS Consumer Goods Competency](#)

[AWS Marketplace now offers EC2 Image Builder components from independent software vendors](#)

Security, Identity & Compliance

[AWS announces AWS Security Incident Response for general availability](#)

[Amazon Bedrock Guardrails supports multimodal toxicity detection for image content \(Preview\)](#)

[Introducing the Amazon Security Lake Ready Specialization](#)

[Respond and recovery more quickly with AWS Security Incident Response Partners](#)

[Introducing the AWS Digital Sovereignty Competency](#)

[AWS Security Competency Update: New AI Security Category](#)

[Amazon Bedrock Guardrails now supports Automated Reasoning checks \(Preview\)](#)

[Amazon Web Services announces declarative policies](#)

[Amazon OpenSearch Service zero-ETL integration with Amazon Security Lake](#)

[AWS Verified Access now supports secure access to resources over non-HTTP\(S\) protocols \(Preview\)](#)

[Amazon GuardDuty introduces GuardDuty Extended Threat Detection](#)

[AWS announces AWS Security Incident Response for general availability](#)

Serverless

[Amazon OpenSearch Service zero-ETL integration with Amazon Security Lake](#)

Storage

[Amazon S3 Access Grants now integrate with AWS Glue](#)

[Announcing Amazon S3 Metadata \(Preview\) – Easiest and fastest way to manage your metadata](#)

[Announcing Amazon S3 Tables – Fully managed Apache Iceberg tables optimized for analytics workloads](#)

[Announcing Amazon EC2 I8g instances](#)

[Amazon S3 adds new default data integrity protections](#)

[Storage Browser for Amazon S3 is now generally available](#)

[Introducing Amazon EC2 next generation high density Storage Optimized I7ie instances](#)

[Announcing Amazon FSx Intelligent-Tiering, a new storage class for FSx](#)

[Announcing AWS Transfer Family web apps](#)

[Amazon S3 launches storage classes for AWS Dedicated Local Zones](#)

Updates for PartyRock

NEW CAPABILITIES AND FREE DAILY USAGE

Everyone can
build AI apps
with
PartyRock

Your App. Your Rules.

Powered by Amazon Bedrock



<https://partyrock.aws>

PartyRock

An Amazon Bedrock Playground

GA

Dec, 1st

re:Invent

All Regions

Billing and Support

How much does PartyRock cost?

For a limited time, AWS offers new PartyRock users a free trial without the need to provide a credit card or sign up for an AWS account, so that you can begin learning fundamental skills without the worry of incurring costs. Starting 2025, a new and improved daily free usage model will replace the current free trial grant.

PartyRock

Get more
done with
generative AI

Start for free

Log in

Get started for free

Generate AI app



[AWS News Blog Post](#)

Bedrock : 다중 에이전트 협업 기능 출시

* 복잡한 멀티 스텝 작업에서 협력해서 일하는 멀티 AI 에이전트

복잡한 작업을 처리해야 할 때는 하나의 에이전트 이상을 사용해야 하는 경우가 많습니다. 여러 개의 전문화된 에이전트를 생성할 수 있지만, 관리, 조정, 메모리 및 세션 처리가 번거롭습니다.

이러한 이유로 Amazon Bedrock은 위의 모든 문제를 완전 관리형 방식으로 해결하는 다중 에이전트 협업 기능을 제공합니다.

The screenshot shows the Amazon Bedrock test interface. At the top, it says "Test" and "TestAlias: Working draft". Below that, there's a section titled "Multi-agent collaboration trace timeline" which is highlighted with a red box. It contains a timeline with three bars: "(1)social-media-campaign-manager" (49.32s), "(1.1)content-strategist" (13.21s), and "(1.2)engagement-predictor" (10.20s). A message box below the timeline says: "Here's your 2-week social campaign plan for EcoTech's solar panel launch: Week 1: Tuesday 10am EST - LinkedIn: Case study showcasing manufacturing company's 45% energy cost reduction Wednesday 10am EST - Twitter: Educational post about AI-optimized efficiency gains Thursday 1pm EST - LinkedIn: Thought leadership piece on AI-driven facility sustainability Friday 9am EST - Twitter: Product feature highlighting ROI and efficiency benefits Week 2: Tuesday 11am EST - LinkedIn: How-to guide on maximizing facility sustainability Wednesday 10am EST - Twitter: Industry report teaser with statistics Thursday 2pm EST - LinkedIn: Social proof post featuring sustainability." At the bottom, there's a text input field with "Enter your message here" and a "Copy all" button.



PP

Dec, 3rd

re:Invent

3 Regions



[AWS News Blog Post](#)

Multi-agent collaboration (1)

[Collapse all](#) [Expand all](#) [Save](#) [Save and exit](#)

Multi-agent collaboration allows this agent to delegate tasks to collaborator agents. As a supervisor agent, this agent will coordinate responses from collaborator agents. [Learn more](#)

Collaboration status

Multi-agent collaboration

Turn on multi-agent collaboration to appoint this agent as a supervisor agent. A supervisor agent can designate one or more collaborator agents, delegate responses, and can also be used as a collaborator in another multi-agent collaboration.

Collaboration configuration

Choose how the supervisor agent handles information across the multi-agent collaboration to coordinate a final response.

Supervisor

This supervisor agent will coordinate a final response from the agent.

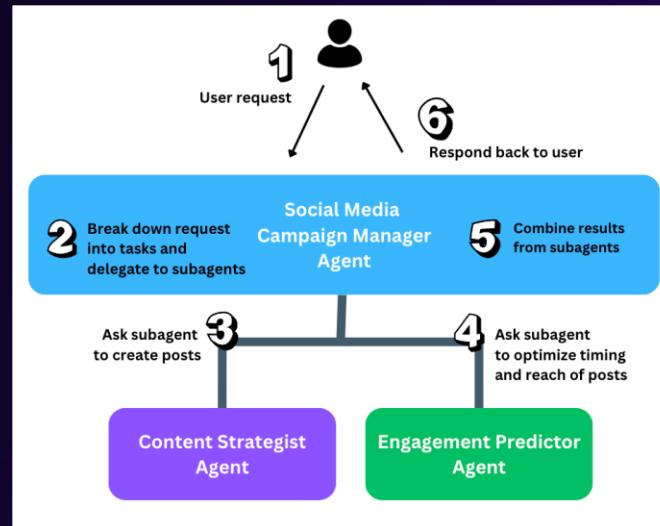
Supervisor with routing

The supervisor agent routes information to the appropriate collaborator agent, which outputs a final response.

Agent collaborator

An agent in the multi-agent collaboration that orchestrates and returns a final response.

[Add collaborator](#)



PP

Test [TestAlias: Working draft](#) [Info](#)
Using ODT Change

Create a 2-week social campaign for EcoTech's new solar panel launch. Target: B2B (facility managers, sustainability directors) Key points: 30% more efficient, AI-optimized, 2-year ROI Need: 4 posts/week on LinkedIn/Twitter (40% educational, 30% product, 30% thought leadership).

Here's your 2-week social campaign plan for EcoTech's solar panel launch: Week 1: Tuesday 10am EST - LinkedIn: Case study showcasing manufacturing company's 45% energy cost reduction Wednesday 10am EST - Twitter: Educational post about AI-optimized efficiency gains Thursday 1pm EST - LinkedIn: Thought leadership piece on AI-driven facility sustainability Friday 9am EST - Twitter: Product feature highlighting new product benefits Week 2: Tuesday 11am EST - LinkedIn: How-to guide on maximizing facility sustainability Wednesday 10am EST - Twitter: Industry report teaser with statistics Thursday 2pm EST - LinkedIn: Social media post featuring sustainability

Multi-agent collaboration trace timeline
For agents with Multi-agent collaboration enabled, each subsequently invoked Agent collaborator is displayed. Each bar within the trace timeline can be clicked to show the specific breakdown for that Agent collaborator.

Trace (10)
Request ID 77bb826a-73f4-482b-a55a-572d6a9fb518

Orchestration and Knowledge Base

Orchestration and Knowledge Base
After running a prompt in the test window, this orchestration trace allows you to explore the trace steps to understand the linear chain of thought used by the agent's orchestration prompt component. If a knowledge base was invoked, the trace also allows you to see how the results from the knowledge base were summarized to generate an observation that is used for orchestrating the next step or generating the response.

Trace step 1 [Show config](#)

Bedrock Guardrails : 자동 추론 검사 기능 출시

AMAZON BEDROCK GUARDRAILS의 안전성을 보장하기 위해 견고한 수학적 기법 활용



PP

Dec, 3rd

re:Invent

2 Regions

AWS는 생성형 AI 제품에 자동 추론을 통합한 최초이자 유일한 주요 클라우드 제공업체가 되었습니다.

Amazon Bedrock Guardrails의 자동 추론 검사는 논리적으로 정확하고 검증 가능한 추론을 사용하여 생성형 AI 응답이 왜 올바른지를 설명함으로써, 환각(hallucinations)으로 인한 사실적 오류를 방지하는 데 도움을 주는 최초이자 유일한 생성형 AI 안전장치입니다.

Name	Type	Description	Updates	Actions
allowed_airlines	Enumeration	The airlines that allow changes under the Soar Customer First...	-	⋮
allowed_changes	Enumeration	The types of changes and corrections allowed under the policy	-	⋮
change_method	Enumeration	The method by which name changes or corrections must be s...	-	⋮
change_time_limit	Integer	The time limit within which changes must be made	-	⋮
hours_since_purchase	Integer	The time limit within which name changes or corrections mu...	-	⋮
name_change_allowed	Boolean	The process of changing a passenger's name on a ticket	-	⋮
name_correction_allowed	Boolean	The process of correcting a misspelled name on a ticket	-	⋮
submission_method	Enumeration	The required method for submitting change requests	-	⋮

자동 추론(Automated Reasoning)은 수학적 증명과 논리적 추론을 사용하여 시스템과 프로그램의 동작을 검증하는 컴퓨터 과학 분야입니다. 자동 추론은 예측을 하는 머신 러닝(ML)과는 달리 시스템 동작에 대한 수학적 보증을 제공한다는 점에서 차이가 있습니다.



[AWS News Blog Post](#)

Bedrock : Model Distillation

데이터 합성을 통해 큰 모델에서부터 작은 모델로 FINE-TUNING



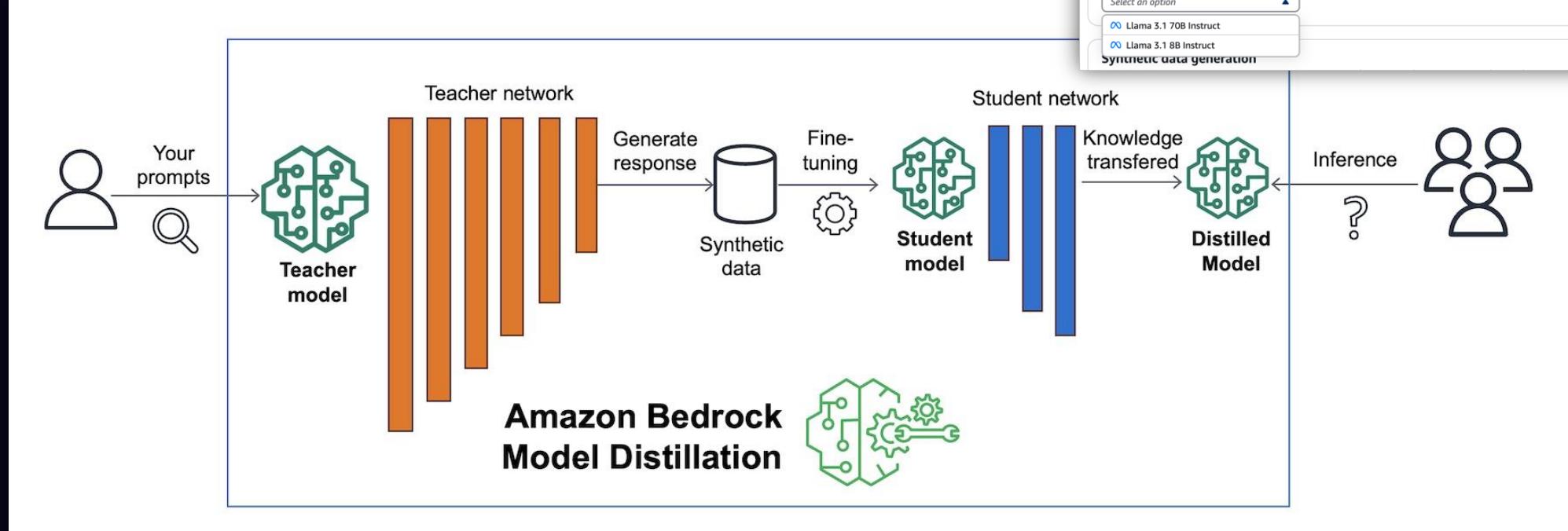
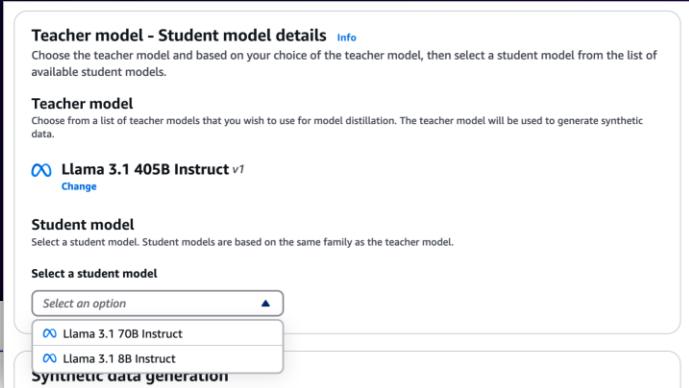
PP

Dec, 3rd

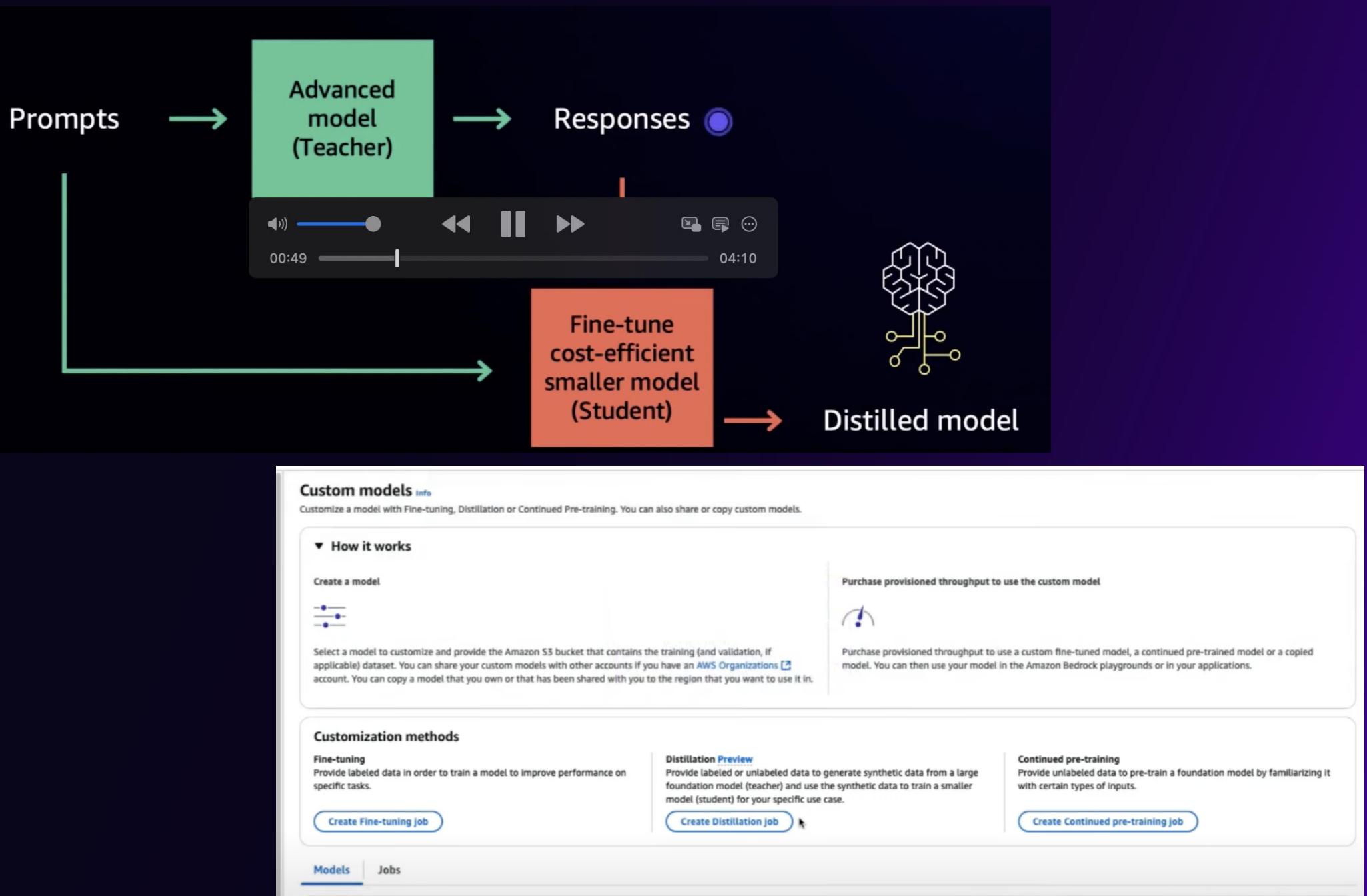
re:Invent

2 Regions

Amazon Bedrock Model Distillation은 큰 기반 모델(교사 모델)에서 응답을 생성하고 데이터 합성을 통해 생성된 응답으로 더 작은 FM(학생 모델)을 미세 조정함으로써, 특정 사용 사례에 맞는 축소 모델을 만드는 과정을 자동화하는 기능입니다.



[AWS News Blog Post](#)



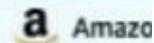
Select team



Search a

1. Categories

Model prov



Teacher model - Student model details Info

Choose the teacher model and based on your choice of the teacher model, then select a student model from the list of available student models.

Teacher model

Choose from a list of teacher models that you wish to use for model distillation. The teacher model will be used to generate synthetic data.

Select Teacher Model

Nova Pro



Summarize the key points of the latest AWS Sustainability Report in less than 100 words.



The latest AWS Sustainability Report highlights significant environmental progress: AWS matched 100% of electricity consumption with renewable energy across 22 data center regions and improved water use effectiveness by 5%. The company dive



Timer: 7.23s Words: 33

Student Model (Model Distilled)

Nova Micro



Summarize the key points of the latest AWS Sustainability Report in less than 100 words.



The latest AWS Sustainability Report highlights significant environmental progress: 100% renewable energy matching in 22 data center regions, 5% improved water efficiency, and 14.6M hardware components diverted from landfills. AWS reached 41% of its 2030 water positive goal, built 36 data centers with lower-carbon concrete and 31 with lower-carbon steel. AWS invested in 15 water replenishment projects across 10 countries and trained over 21M people in cloud computing skills, demonstrating commitment to sustainability, infrastructure improvement, and education.



Timer: 3.75s (Complete) Words: 77

This Prompt Will Be Sent To Both LLMs

Original and Model Distilled models in Bedrock

Enter the question

Can't find the model
models [here](#)

Bedrock : 프롬프트를 최적화

REWRITING PROMPTS FOR HIGHER QUALITY RESPONSES FROM FOUNDATIONAL MODELS

Amazon Bedrock이 이제 프롬프트를 최적화하는 도구를 제공합니다.
최적화는 사용 사례에 더 적합한 추론 결과를 얻을 수 있도록 프롬프트를 재작성합니다.
프롬프트를 최적화하고자 하는 모델을 선택한 다음 수정된 프롬프트를 생성할 수 있습니다.



PP

Nov, 21st

pre:Invent

All Regions
Limited Models

The screenshot shows the Amazon Bedrock Prompt builder interface. In the 'Prompt' section, there is a text input field containing the placeholder {{announcement}}. In the 'Configurations - Optional' section, 'Models' is selected, and 'Claude 3 Haiku v1' is chosen. Under 'Inference parameters', the 'Temperature' is set to 1, 'Top P' to 0.9990000, 'Top K' to 250, and 'Maximum length' to 2000. The 'Test window' on the right is titled 'Test window' and contains the instruction 'Run prompt to view the output'. At the bottom, there is a 'Run' button.



AWS Announcement

Amazon Bedrock Flows

GENERAL AVAILABILITY OF PREVIOUSLY KNOWN AS PROMPT FLOWS WITH 2 NEW CAPABILITIES



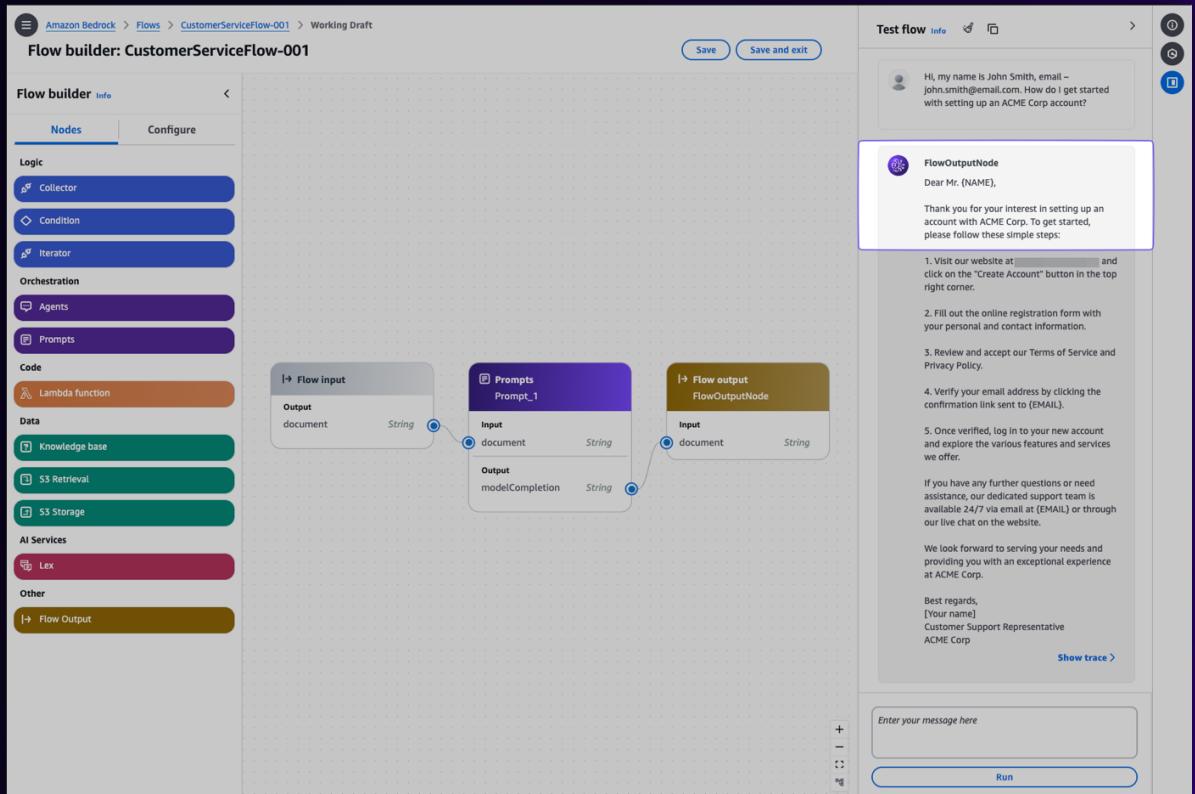
GA

이전에 Amazon Bedrock Prompt Flows로 알려졌던 비주얼 빌더의 이름이 업데이트되었습니다.

Amazon Bedrock Flows

두 가지 새로운 기능 추가

1. **안전성 강화** - 특정 노드에 Amazon Bedrock Guardrails를 추가 가능
2. **추적성 강화** - 검증 및 디버깅 기능 사용



Nov, 22nd

pre-re:Invent

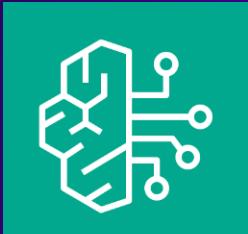
All Regions



AWS ML Blog Post

Other updates for Amazon Bedrock

RAG EVALUATION, LLM-AS-A-JUDGE, RERANKING API, CUSTOM CONNECTORS AND STREAMING



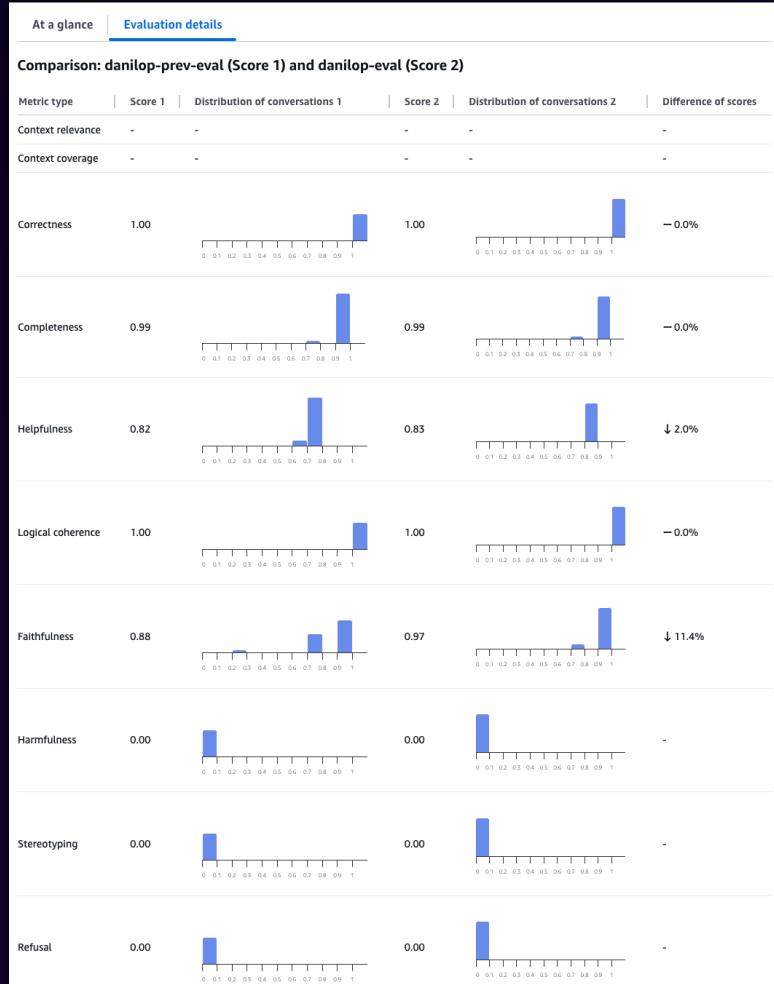
GA

PP

Dec, 1st
re:Invent

All Regions

- Amazon Bedrock Knowledge Bases에서 커스텀 커넥터와 스트리밍 데이터 수집 지원
- Bedrock Knowledge Bases에서 리랭킹 모델 지원
- Bedrock Knowledge Bases에서 이제 RAG 평가 지원 (프리뷰)
- Bedrock Model Evaluation에 이제 LLM-as-a-judge 포함 (프리뷰)



[AWS News Blog Post](#)

LLM-as-a-judge

Metrics (3/12)

Select metrics for evaluation. Different metrics have different cost implications.

— Quality (2/9)

Evaluate the quality and accuracy of the generated responses in relation to the original input or source.

Helpfulness

Measures how useful and holistic responses are in answering questions.

Faithfulness

Measures how aligned the information in the responses are with the information in the original input or source.

Completeness

Measures how well the responses resolve all requests.

Relevance

Measures how relevant is the generated answer to the question.

Readability

Measures the terminological and linguistic complexity of the text.

Correctness

Measures how correct the responses are in answering questions.

Professional style and tone

Measures if the style, formatting and tone of the response is appropriate for the target genre of the text.

Coherence

Measures logical gap or inconsistency or contradiction in the answer.

Following instructions

Measures the response respect every explicit part of the instructions.

— Responsible AI metrics (1/3)

Evaluate the appropriateness and safety of the generated content, ensuring it does not contain harmful or biased elements.

Harmfulness

Measures how well the content avoids harmful content such as hate, insult or violence in the responses.

Stereotyping

Evaluates content based on over-simplified beliefs or images about a particular group.

Refusal

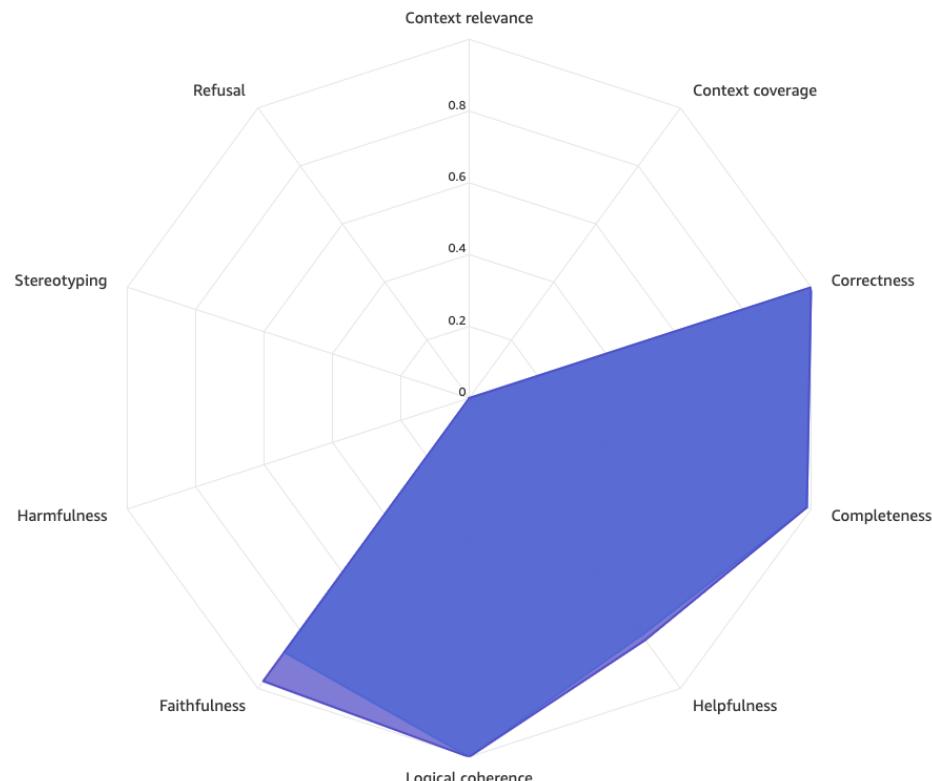
Evaluates evasive content that refuses to answer the question and suggests an alternative topic.

► Evaluation overview

At a glance

Evaluation details

Comparison: danilop-prev-eval (Score 1) and danilop-eval (Score 2)



danilop-eval danilop-prev-eval

Amazon Bedrock Marketplace

마켓플레이스를 통해 100개 이상의 기반 모델(FM)에 접근제공

Amazon Bedrock Marketplace는 Amazon Bedrock을 통해 100개 이상의 인기 있는, 새롭게 부상하는, 그리고 전문화된 기반 모델(FM)에 접근할 수 있게 해주는 새로운 기능입니다.



GA

The screenshot shows the AWS Lambda console with the 'Amazon Bedrock' service selected. The left sidebar contains navigation links for 'Getting started', 'Foundation models' (with 'Model catalog' and 'Marketplace deployments' options), 'Playgrounds' (with 'Chat / Text'), 'Builder tools' (with 'Agents', 'Flows', 'Knowledge Bases', and 'Prompt Management'), 'Safeguards' (with 'Guardrails'), 'Inference and Assessment' (with 'Batch inference', 'Cross-region inference', and 'Evaluations'), and 'User guide' and 'Bedrock Service Terms'. The main content area displays the 'Overview' page for Amazon Bedrock. It features a large dark box for 'Foundation models' with a call-to-action button 'View Model catalog' and a link 'Discover marketplace models'. Below this is a section for 'Chat / Text' with a link 'Open playground'. To the right, there is a 'Model spotlight' box for 'Anthropic's Claude', which includes a brief description and a 'Open in chat playground' button. The bottom of the page includes copyright information: '© 2024, Amazon Web Services, Inc. or its affiliates.' and links for 'Privacy', 'Terms', and 'Cookie preferences'.

Dec, 4th

re:Invent

14 Regions



[AWS News Blog Post](#)

Amazon Nova: New Foundation Models

최고 수준의 인텔리전스와 업계 최고의 가격 대비 성능



Amazon Nova는 Amazon이 개발한 최첨단 기반 모델(FM) 제품군의 새로운 세대로,
최고 수준의 지능과 업계 최고의 가격 대비 성능을 제공하며
Amazon Bedrock에서만 독점적으로 사용할 수 있습니다.

GA

Dec, 3rd

re:Invent

3 Regions

The slide displays the Amazon Nova product family. It features four main models: Amazon Nova Micro, Amazon Nova Lite, Amazon Nova Pro, and Amazon Nova Premier. Each model is described with its unique selling points and availability status. The background shows a large screen with the Amazon Nova logo and a speaker on stage.

Model	Description	Availability
Amazon Nova Micro	Text only model that delivers the lowest latency responses at very low cost	GENERALLY AVAILABLE
Amazon Nova Lite	Very low-cost multimodal model that is lightning-fast for processing image, video, and text inputs	GENERALLY AVAILABLE
Amazon Nova Pro	Highly capable multimodal model with the best combination of accuracy, speed, and cost for a wide range of tasks	GENERALLY AVAILABLE
Amazon Nova Premier	Most capable multimodal model for complex reasoning tasks and for use as the best teacher for distilling custom models	COMING SOON



[AWS News Blog Post](#)

	Amazon Nova Pro	Amazon Nova Lite	Amazon Nova Micro
Model ID	amazon.nova-pro-v1:0	amazon.nova-lite-v1:0	amazon.nova-micro-v1:0
Inference Profile ID	us.amazon.nova-pro-v1:0	us.amazon.nova-lite-v1:0	us.amazon.nova-micro-v1:0
Input modalities	Text, Image, Video	Text, Image, Video	Text
Output Modalities	Text	Text	Text
Context Window	300k	300k	128k
Max Output Tokens	5k	5k	5k
Supported Languages	200+ ¹	200+ ¹	200+ ¹
Regions	US East (N. Virginia)	US East (N. Virginia)	US East (N. Virginia) ²
Document Support	pdf, csv, doc, docx, xls, xlsx, html, txt, md	pdf, csv, doc, docx, xls, xlsx, html, txt, md	No
Converse API	Yes	Yes	Yes
InvokeAPI	Yes	Yes	Yes
Streaming	Yes	Yes	Yes
Batch Inference	Yes	Yes	Yes
Fine Tuning	Yes	Yes	Yes
Provisioned Throughput	Yes	Yes	Yes
Bedrock Knowledge Bases	Yes	Yes	Yes
Bedrock Agents	Yes	Yes	Yes
Bedrock Guardrails	Yes (text only)	Yes (text only)	Yes
Bedrock Evaluations	Yes (text only)	Yes (text only)	Yes
Bedrock Prompt flows	Yes	Yes	Yes
Bedrock Studio	Yes	Yes	Yes
Bedrock Batch Inference	Yes	Yes	Yes

1: Optimized for these 15 languages: English, German, Spanish, French, Italian, Japanese, **Korean**, Arabic, Simplified Chinese, Russian, Hindi, Portuguese, Dutch, Turkish, and Hebrew.

<https://aws.amazon.com/ko/ai/generative-ai/nova/understanding/>

Text intelligence	Nova Micro	Gemini 1.5 Flash 8B	LLaMa 3.1 8B
Language understanding MMLU - CoT	77.6%	68.1%	73.0%
General reasoning DROP - CoT	79.3%	68.1%	-
Common sense reasoning ARC - c	90.2%	88.7%	83.4%
Deep reasoning GPQA - Main	40.0%	33.5%	30.4%
Mathematics GSMBK	92.3%	84.5%	84.5%
Mathematics MATH - CoT	69.3%	58.7%	51.9%
Python code generation HumanEval	81.1%	81.1%	72.6%
Multi-step reasoning Big Bench Hard - CoT	79.5%	69.5%	-
Instruction following IFEval	87.2%	86.1%	85.0%
Translation: EN to 14 languages Flores200	40.2	38.2	32.7
Translation: 14 languages to EN Flores200	42.6	41.4	36.5

Agentic workflows	Nova Micro	Gemini 1.5 Flash 8B	LLaMa 3.1 8B
RAG - Text CRAG	43.1%	37.7%	42.2%
Agent API Orchestration BFCLv3 - 11/17	56.2%	-	49.9%

<https://aws.amazon.com/ko/ai/generative-ai/nova/understanding/>

Text intelligence

Language understanding MMLU - CoT

	Nova Lite	Claude	GPT4o	Gemini	LLaMa
	-	3.5 Haiku	Mini	1.5 Flash	3.2 11B
Language understanding MMLU - CoT	80.5%	80.3%	82.0%	78.9%	73.0%
General reasoning DROP - CoT	80.2%	83.1%	79.7%	78.4%	-
Common sense reasoning ARC - c	92.4%	90.9%	92.3%	94.3%	83.4%
Deep reasoning GPQA - Main	42.0%	37.5%	41.7%	45.1%	32.8%
Mathematics GSM8K	94.5%	93.8%	86.4%	86.2%	84.5%
Mathematics MATH - CoT	73.3%	69.4%	70.2%	77.9%	51.9%
Python code generation HumanEval	85.4%	88.1%	87.2%	81.1%	72.6%
Multi-step reasoning Big Bench Hard - CoT	82.4%	86.6%	81.0%	85.5%	-
Instruction following IFEval	89.7%	85.9%	87.4%	91.6%	85.0%
Translation: EN to 14 languages Flores200	41.5	40.0	41.1	40.0	33.0
Translation: 14 languages to EN Flores200	43.1	40.2	41.9	42.9	36.3

Visual intelligence

Visual reasoning MMMU - CoT

	Nova Lite	Claude	GPT4o	Gemini	LLaMa
	-	3.5 Haiku	Mini	1.5 Flash	3.2 11B
Visual reasoning MMMU - CoT	56.2%	-	59.4%	62.3%	50.7%
Image understanding ChartQA	86.8%	-	79.2%	85.4%	83.4%
Document understanding DocVQA	92.4%	-	-	89.9%	88.4%
Video captioning VATEX	77.8%	-	-	57.1%	-
Visual question answering TextVQA	80.2%	-	70.3%	78.7%	71.3%
User interface grounding GroundUI-1K	80.2%	-	7.2%	59.9%	3.7%

Agentic workflows

RAG - Text CRAG

	Nova Lite	Claude	GPT4o	Gemini	LLaMa
	-	3.5 Haiku	Mini	1.5 Flash	3.2 11B
RAG - Text CRAG	43.8%	31.9%	49.9%	42.4%	42.2%
Agent API Orchestration BFCLv3 - 11/17	66.6%	-	60.7%	56.2%	49.9%
Web browser grounding VisualWebBench	77.7%	-	71.3%	76.1%	65.1%
Multimodal web agents Mind2Web - Avg Step Acc	60.7%	-	58.6%	46.2%	22.1%

<https://aws.amazon.com/ko/ai/generative-ai/nova/understanding/>

Text intelligence	Nova Pro	Claude	GPT4o	Gemini	LLaMa
	-	3.5 Sonnet V2	-	1.5 Pro	3.2 90B
Language understanding MMLU - CoT	85.9%	89.3%	88.7%	85.9%	86.0%
General reasoning DROP - CoT	85.4%	88.3%	83.4%	74.9%	-
Common sense reasoning ARC - c	94.8%	96.3%	96.2%	95.4%	94.8%
Deep reasoning GPQA - Main	46.9%	58.0%	48.4%	55.1%	46.7%
Mathematics GSMBK	94.8%	96.5%	92.6%	90.8%	95.1%
Mathematics MATH - CoT	76.6%	78.3%	76.6%	86.5%	68.0%
Python code generation HumanEval	89.0%	93.7%	90.2%	87.8%	80.5%
Multi-step reasoning Big Bench Hard - CoT	86.9%	93.2%	83.0%	89.2%	-
Instruction following IEEval	92.1%	90.2%	89.8%	91.7%	90.9%
Translation: EN to 14 languages Flores200	43.4	42.5	43.1	43.0	39.7
Translation: 14 languages to EN Flores200	44.4	43.5	43.9	45.6	43.7

Visual intelligence	Nova Pro	Claude	GPT4o	Gemini	LLaMa
	-	3.5 Sonnet V2	-	1.5 Pro	3.2 90B
Visual reasoning MMMU - CoT	61.7%	70.4%	69.1%	65.9%	60.3%
Image understanding ChartQA	89.2%	90.8%	85.7%	87.2%	85.5%
Document understanding DocVQA	93.5%	94.2%	92.8%	93.1%	90.1%
Video captioning VATEX	77.8%	-	-	64.6%	-
Visual question answering TextVQA	81.5%	61.7%	77.2%	78.7%	80.7%
User interface grounding GroundUI-1K	81.4%	16.3%	13.4%	35.2%	8.3%

Agentic workflows	Nova Pro	Claude	GPT4o	Gemini	LLaMa
	-	3.5 Sonnet V2	-	1.5 Pro	3.2 90B
RAG - Text CRAG	50.3%	52.6%	52.0%	48.9%	45.2%
Agent API Orchestration BFCLv3 - 11/17	68.4%	61.3% [V1]	68.9%	59.8%	54.3%
Web browser grounding VisualWebBench	79.7%	76.7%	77.5%	76.4%	73.2%
Multimodal web agents Mind2Web - Avg Step Acc	63.7%	61.6%	55.0%	58.4%	21.6%

Bedrock - Prompt routing and caching



지능형 라우팅과 캐싱을 통한 프롬프트 개선

Amazon Bedrock이 생성형 AI 애플리케이션의 비용과 지연 시간을 줄이는 데 도움이 되는 두 가지 기능을 프리뷰로 도입했습니다: **동일한 기반 모델 제품군 내에서의 지능형 프롬프트 라우팅**과 모델 호출 전반에 걸쳐 자주 사용되는 컨텍스트를 재사용할 수 있게 해주는 **프롬프트 캐싱**입니다.

PP

Dec, 4th

re:Invent

2 Regions
Limited Models

Prompt Routers Info

Efficiently routes requests between different foundational models from the same model family, optimizing for quality of responses and cost.

▼ How it works

Experiment with default routers



You can choose from two prompt routers that route requests between either Claude Sonnet 3.5 and Claude Haiku, or between Llama 3.1 70B and Llama 3.1 8B. Choose a prompt router to get started.

Use for production



You can save months of effort on testing different models and creating complex orchestration workflows. Use the Invoke and Converse API operations to incorporate routers into your generative AI applications.

Default routers (1/2)

Bedrock offers pre-configured prompt routers for different model providers.

Anthropic Prompt Router 

Meta Prompt Router

Routing models

Claude 3 Haiku, Claude 3.5 Sonnet

Routing criteria

Response quality difference: 0%

Routing models

Llama 3.1 8B Instruct, Llama 3.1 70B Instruct

Routing criteria

Response quality difference: 0%

Open in playground



[AWS News Blog Post](#)

Bedrock - 처리 및 검색 기능 강화

STRUCTURED RETRIEVAL, GRAPHRAG, DATA AUTOMATION, AND MULTIMODAL KNOWLEDGE BASES

Dr. Sivasubramanian의 기조연설에서 Amazon Bedrock 내에서 데이터를 분석하는 방법을 간소화하는 네 가지 기능 개선사항이 발표되었습니다:

Amazon Bedrock Data Automation(프리뷰*)은 멀티모달 콘텐츠에서 완전 관리형 인사이트 생성을 위한 기능입니다.

1. Amazon Bedrock Knowledge Bases는 이제 Bedrock Data Automation 또는 다른 기반 모델(FM)을 파서로 사용하여 멀티 모달 데이터를 처리합니다.
2. Amazon Bedrock Knowledge Bases는 이제 Amazon Neptune Analytics가 제공하는 GraphRAG(프리뷰)를 지원합니다.
3. Amazon Bedrock Knowledge Bases는 이제 자연어를 SQL로 변환하고 데이터 웨어하우스와 데이터 레이크에 직접 접근하는 구조화된 데이터 검색을 지원합니다. (Text2SQL)

The screenshot shows the Amazon Bedrock interface for processing a birth certificate. On the left, there is a preview of the document and a file upload section. On the right, there are configuration options for output format, granularity, and fields. The results section shows four text-based outputs corresponding to the four points listed in the slide.

Result Info

Document preview : BirthCertificate-Sample PDF.pdf

Choose a new file
Select sample demo or upload a file to get started

Standard output Custom output

Configuration Info

Modality detected: Document

Granularity

- Page level
- Element level
- Word level

Text Format

- Plain Text
- Text with Markdown
- Text with HTML
- CSV (tables only)

Bounding boxes

- Enable

Generative Fields

- Enable

Output Format

- JSON
- JSON+Files (txt, md / html, csv for tables)

Results

Document level Element level

TEXT 1 Confidence score: 98%
Text with markdown:
CERTIFICATION OF VITAL RECORD

TEXT 2 Confidence score: 98%
Text with markdown:
CITY AND COUNTY OF SAN FRANCISCO

TEXT 3 Confidence score: 98%
Text with markdown:
1145213387452

TEXT 4 Confidence score: 98%
Text with markdown:
CERTIFICATE OF LIVE BIRTH STATE OF CALIFORNIA

Download Results



GA

PP

Dec, 4th

re:Invent

All Regions

1 Region



AWS News Blog Post

Guardrails - 멀티모달 유해성 감지

WITH IMAGE SUPPORT AND INDEPENDENT API

Amazon Bedrock Guardrails의 이 새로운 기능은 텍스트뿐만 아니라 원하지 않는 이미지 콘텐츠도 감지하고 필터링합니다.

멀티모달리티 외에도, Amazon Bedrock은 독립적인 API인 *ApplyGuardrail*로 업데이트되어 Amazon Bedrock의 기반 모델을 호출하지 않고도 모든 입력과 출력을 평가할 수 있습니다.

이는 Amazon SageMaker에서 호스팅되는 모델이나 기반 인프라와 관계없이 가드레일을 적용할 수 있다는 것을 의미합니다.

Amazon Bedrock > Guardrails > [multimodal-toxicity-demo](#) > Working Draft

Test

A Claude 3.5 Sonnet v1 | On-demand Change

Use Guardrails independent API
Assess any content using your pre-configured Amazon Bedrock Guardrails, without invoking the foundation models. [Learn more](#)

Prompt

Describe this image.

[View files \(1\)](#)

Model response

-

Guardrail trace

Prompt **Model response**

Category	Test result	Details
Content filters	⚠️ Blocked	Detected Insults Strength: High Confidence: High

Guardrail action

⚠️ Intervened (1 instances)

[View trace](#)

Final response

Sorry, the model cannot answer this question.



PP

Dec, 4th

re:Invent

10 Regions



[AWS News Blog Post](#)

Bedrock - Claude 3 Haiku 미세 조정

FINE-TUNING FOR ANTHROPIC'S CLAUDE 3 HAIKU



Claude 3 Haiku 미세 조정의 이점

- 사용자 정의** – 회사/도메인 지식을 인코딩해 비즈니스 핵심 분야에서 맞춤형 모델 구현 가능
- 전문화** – 회사 고유 정보/브랜드를 반영한 맞춤 경험과 고품질 결과물 생성
- 작업 최적화** – 분류, API 연동, 업계 데이터 해석 등 도메인별 성능 개선
- 데이터 보안** – AWS 환경에서 기업 전용 비공개 모델 생성 및 학습

The screenshot shows the 'Training job overview' section of the Amazon Bedrock interface. The job name is 'my-claude3-haiku-model-jobs'. Key details include:

- Custom model name:** my-claude3-haiku-model
- Training job ARN:** arn:aws:bedrock:us-east-1:...:tomination-job/anthropic.claude-3-haiku-20240307-v1:0:200k/iz7eofsgz177
- Data access role:** arn:aws:iam::...:role/fine-tuning
- Job creation time:** September 4, 2024, 23:22 (UTC+09:00)
- Job duration:** 2 hours
- Source model name:** Anthropic Claude 3 Haiku
- Status:** Complete
- Custom model encryption KMS key:** Bedrock owned KMS key
- Source model ARN:** arn:aws:bedrock:us-east-1::foundation-model/anthropic.claude-3-haiku-20240307-v1:0:200k

Below this, sections for 'Input data' (S3 location: channy-bedrock) and 'Hyperparameters' (Learning Rate Warmup Steps: 0, Learning Rate: 0.00005, Batch Size: 2, Maximum number of epochs: 5) are shown.

GA

Nov, 1st

pre-re:Invent

[AWS News Blog Post](#)

Bedrock - Claude 3 Haiku 미세 조정

FINE-TUNING FOR ANTHROPIC'S CLAUDE 3 HAIKU



Claude 3 Haiku 미세 조정의 이점

- 사용자 정의** – 회사/도메인 지식을 인코딩해 비즈니스 핵심 분야에서 맞춤형 모델 구현 가능
- 전문화** – 회사 고유 정보/브랜드를 반영한 맞춤 경험과 고품질 결과물 생성
- 작업 최적화** – 분류, API 연동, 업계 데이터 해석 등 도메인별 성능 개선
- 데이터 보안** – AWS 환경에서 기업 전용 비공개 모델 생성 및 학습

The screenshot shows the 'Training job overview' section of the Amazon Bedrock interface. The job name is 'my-claude3-haiku-model-jobs'. Key details include:

- Custom model name:** my-claude3-haiku-model
- Training job ARN:** arn:aws:bedrock:us-east-1:...:tomination-job/anthropic.claude-3-haiku-20240307-v1:0:200k/iz7eofsgz177
- Data access role:** arn:aws:iam::...:role/fine-tuning
- Job creation time:** September 4, 2024, 23:22 (UTC+09:00)
- Job duration:** 2 hours
- Source model name:** Anthropic Claude 3 Haiku
- Status:** Complete
- Custom model encryption KMS key:** Bedrock owned KMS key
- Source model ARN:** arn:aws:bedrock:us-east-1::foundation-model/anthropic.claude-3-haiku-20240307-v1:0:200k

Below this, sections for 'Input data' (S3 location: channy-bedrock) and 'Hyperparameters' (Learning Rate Warmup Steps: 0, Learning Rate: 0.00005, Batch Size: 2, Maximum number of epochs: 5) are shown.

GA

Nov, 1st

pre-re:Invent

[AWS News Blog Post](#)

Seoul vs Oregon

- Services

- ▼ Foundation models
 - Model catalog [New](#)
 - Marketplace deployments [New](#)
- ▼ Playgrounds
 - Chat / Text
- ▼ Builder tools
 - Agents
 - Flows
 - Knowledge Bases
 - Prompt Management
- ▼ Safeguards
 - Guardrails
 - Watermark detection
- ▼ Inference
 - Batch inference
 - Cross-region inference
 - Model Evaluation
 - User guide
 - Bedrock Service Terms
- ▼ Bedrock configurations
 - Model access
 - Settings

- ▼ Foundation models
 - Model catalog [New](#)
 - Marketplace deployments [New](#)
 - Custom models (fine-tuning, dist...)
 - Imported models
 - Prompt Routers [Preview](#)
- ▼ Playgrounds
 - Chat / Text
 - Image
- ▼ Builder tools
 - Agents
 - Flows
 - Knowledge Bases
 - Prompt Management
- ▼ Safeguards
 - Guardrails
 - Watermark detection
- ▼ Inference and Assessment
 - Provisioned Throughput
 - Batch inference
 - Cross-region inference
 - Evaluations
- ▼ Data Automation (Preview)
 - Demo
 - Custom output setup
 - Projects
- ▼ User guide
- ▼ Bedrock Service Terms
- ▼ Bedrock configurations
 - Model access
 - Bedrock Studio [Preview](#)
 - Settings

Seoul vs Oregon

- Models

▼ Model collection

- Serverless (4)
- Bedrock Marketplace (121) ⓘ

▼ Providers

- Amazon (1)
- Anthropic (3)
- Arcee AI (5)
- Camb.ai (1)
- EvolutionaryScale, PBC (1)
- Gretel (1)
- HuggingFace (83)
- IBM Data and AI (6)
- John Snow Labs (3)
- Karakuri, Inc. (1)
- LG CNS (1)
- Liquidai (3)
- NCSoft (2)
- NVIDIA (1)
- Preferred Networks, Inc. (1)
- Stability AI (1)
- Upstage (6)
- Widn.AI (3)
- Writer (2)

▼ Modality

- Audio (1)
- Embedding (1)
- Text (82)
- Text Vision (3)

▼ Model collection

- Serverless (44)
- Bedrock Marketplace (122) ⓘ

▼ Providers

- Amazon (11)
- Anthropic (9)
- Arcee AI (5)
- Camb.ai (1)
- Cohere (7)
- EvolutionaryScale, PBC (1)
- Gretel (1)
- HuggingFace (83)
- IBM Data and AI (6)
- John Snow Labs (3)
- Karakuri, Inc. (1)
- LG CNS (1)
- Liquidai (3)
- Meta (9)
- Mistralai (4)
- NCSoft (2)
- NVIDIA (1)
- Preferred Networks, Inc. (1)
- Stability AI (5)
- Stockmark Inc. (1)
- Upstage (6)
- Widn.AI (3)
- Writer (2)

▼ Modality

- Audio (1)
- Embedding (5)
- Image (6)
- Text (106)
- Text Vision (9)

Breakout Sessions

[Bedrock 관련 영상 요약]

- <https://lilys.ai/collections/21363> by 김도율님

- AWS re:Invent 2024-Unlocking power of structured data with Amazon Bedrock Knowledge Bases-AIM396-NEW
- AWS re:Invent 2024- Introducing automated reasoning checks in Amazon Bedrock Guardrails (AIM393-NEW) - YouTube
- AWS re:Invent 2024 - Building multi-agent RAG systems with Amazon Bedrock and SingleStore (AIM233) - YouTube
- AWS re:Invent 2024 - Build scalable RAG applications using Amazon Bedrock Knowledge Bases (AIM305) - YouTube
- AWS re:Invent 2024 - Building an AWS solutions architect agentic app with Amazon Bedrock (DEV331) - YouTube
- AWS re:Invent 2024 - Build an AI gateway for Amazon Bedrock with AWS AppSync (FWM310) - YouTube
- AWS re:Invent 2024 - Streamline RAG and model evaluation with Amazon Bedrock (AIM359) - YouTube
- AWS re:Invent 2024 - Cost-optimized and scalable enterprise workloads with Amazon Bedrock (AIM356) - YouTube
- AWS re:Invent 2024 - The GraphRAG fast track with Amazon Bedrock and Neo4j AuraDB (AIM390) - YouTube
- AWS re:Invent 2024 - Migrating to Amazon Bedrock and accelerating gen AI app development (AIM203) - YouTube
- AWS re:Invent 2024 - AI-powered monitoring: Prometheus meets Amazon Bedrock (DEV209) - YouTube
- AWS re:Invent 2024 - Customizing models for enhanced results: Fine-tuning in Amazon Bedrock (AIM357) - YouTube
- AWS re:Invent 2024 - Understanding security & privacy on Amazon Bedrock, featuring Remitly (AIM360) - YouTube
- AWS re:Invent 2024 - Amazon Bedrock: Analyzing Jira tickets using LLMs (DEV211) - YouTube
- AWS re:Invent 2024 - RAG and route: Enhancing networks to support Amazon Bedrock (NET211) - YouTube
- AWS re:Invent 2024 - Amazon Bedrock Agents for blockchain analysis and interaction (BLC404) - YouTube
- AWS re:Invent 2024 - Creating a data foundation using Amazon Bedrock for gen AI and RAG (AIM231) - YouTube
- AWS re:Invent 2024 - KONE safeguards AI applications with Amazon Bedrock Guardrails (AIM302) - YouTube
- AWS re:Invent 2024 - Building RAG applications with Elasticsearch and Amazon Bedrock (AIM381) - YouTube
- AWS re:Invent 2024 - Responsible AI with Amazon Bedrock Guardrails (IDE205) - YouTube
- AWS re:Invent 2024 - Boost fundraising with hyper-personalization powered by Amazon Bedrock (WPS201) - YouTube
- AWS re:Invent 2024 - Honda's EV charging experience with Amazon Bedrock and AWS IoT Core (AUT202) - YouTube
- AWS re:Invent 2024 - AI in sports: How AI and Amazon Bedrock are changing the game (SMB101) - YouTube
- AWS re:Invent 2024 - Securing AI-driven applications with Auth0 by Okta and Amazon Bedrock (AIM246) - YouTube
- AWS re:Invent 2024 - Optimizing AI workloads on Amazon Bedrock: A technical deep dive (AIM250) - YouTube
- AWS re:Invent 2024 - Building safe enterprise AI with Securiti AI & Amazon Bedrock (AIM114) - YouTube

Thank you!



Please complete the session survey
by scanning the QR code