



AWS 클라우드를 기반으로 AI 엔지니어링 시작하기

윤석찬

수석 테크 에반젤리스트, AWS

AI Engineering이란?

AI Engineering의 정의



AI 엔지니어링은 시스템 엔지니어링 원리, 소프트웨어 엔지니어링, 컴퓨터 과학, 인간 중심 설계를 결합하여 특정 작업을 완료하거나 특정 목표를 달성할 수 있는 지능형 시스템을 만드는 분야입니다. AI 엔지니어는 방대한 양의 데이터를 통해 만든 정교한 알고리즘으로 가상 비서, 자율 주행, 중증 진단 등의 기능을 수행합니다.

- AI 모델을 구축, 개발, 테스트 및 배포합니다.
- AI 모델을 API로 변환하여 다른 애플리케이션에서 활용할 수 있도록 합니다.
- 데이터 수집 및 데이터 변환 인프라를 구축하고, 데이터 분석가와 협력합니다.
- 통계 분석을 실행하고 결과를 조정하여 더 나은 통찰력을 추출합니다.
- 데이터 과학 팀에서 사용하는 인프라를 자동화합니다.
- AI 모델 개발 및 프로덕션을 위한 인프라를 만들고 관리합니다.



AI Engineering의 정의



AI 엔지니어링이란 AI 및 기계 학습 기술을 사용하여 기업 내 효율성을 높이고, 비용을 절감하고, 수익을 늘리고, 더 나은 비즈니스 결정을 내리는 데 도움이 되는 애플리케이션과 시스템을 개발하는 분야입니다. AI 엔지니어링은 인공 지능을 현실 세계에 적용할 수 있도록 하는 도구, 시스템 및 프로세스를 개발하는 데 중점을 둡니다. 즉, 인간의 지능을 모방하는 모든 애플리케이션이 대상이 됩니다.

- 생성형 AI 모델을 기반한 애플리케이션 개발 및 운영을 담당합니다.
- 데이터 사이언스팀을 위한 AI 인프라를 자동화합니다.
- AI 모델을 분석하고, 구축하고, 이를 다른 애플리케이션과 통합할 수 있는 API를 제공합니다.
- 타 팀 간의 협업을 통해 생성형 AI 채택 및 모범 사례를 지원합니다.

<https://www.coursera.org/articles/ai-engineer>



**AI Engineer**

World's Fair

June 25-27 2024 • San Francisco

The Biggest Technical AI conference in San Francisco

Where the leading AI companies, AI founders, VPs of AI & AI Engineers meet. Featuring an Expo floor showcasing 30+ companies pioneering the AI Engineering landscape, and over 100+ speakers delivering talks and workshops across 9 tracks.

Buy Now SOLD OUT!**Sign up for livestream**

Jun 25 Jun 26 Jun 27

Jun 25

8:00AM

Registration & Breakfast *Plenary*

9:00AM Lance Martin / [LangChain](#)

Architecting and Testing Controllable A

9:00AM Daniel Han / [Unsloth](#)

Low Level Technicals of LLMs *Workshops*

9:00AM Manuel Odendahl / [The Tree Center](#)

All Tracks

- Multimodality
- Open Models
- RAG & LLM Frameworks
- Agents
- AI Leadership
- Evals & LLM Ops
- Expo Sessions
- CodeGen & Dev Tools
- Keynote
- AI in the Fortune 500
- GPUs & Inference
- Workshops
- Uncategorized

Add Plenary Schedule

AWSKRUG AI Engineering 소모임

AI 엔지니어링 소모임은 생성형 AI 시대에 새롭게 바뀌는 실제 개발 현장의 문제를 다루는 모임입니다. 클라우드 및 소프트웨어 엔지니어들이 꼭 알아두어야 하는 **프롬프트 엔지니어링**, **검색증강생성(RAG)** 기법, **생성형 AI 개발 도구 및 활용법** 등을 주로 다룹니다.

단, 대규모언어모델(LLM) 훈련 및 추론, 파인튜닝, GPU나 CUDA 활용, LLM MLOps와 관련된 주제는 다루지 않으며, 가급적 AWS 클라우드 내 Amazon Bedrock, Amazon Q Business, Amazon Q Developer 및 관련 벡터데이터베이스 (Amazon Aurora, OpenSearch 등) 서비스를 주로 다룹니다.

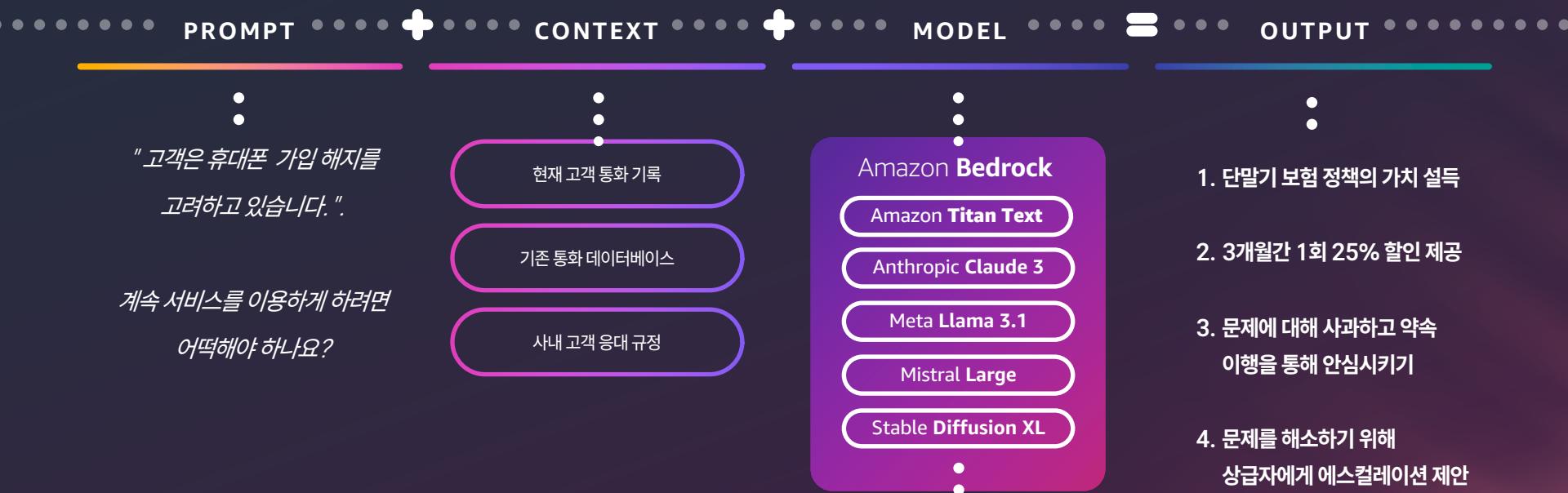


생성형 AI를 통한 기반 앱 구현 기술

생성형 AI 앱 만들기



생성형 AI 앱 만들기



Amazon Bedrock 

▼ Getting started

[Overview](#)[Examples](#)[Providers](#)

▼ Foundation models

[Base models](#)[Custom models](#)[Imported models](#) [Preview](#)

▼ Playgrounds

[Chat](#)[Text](#)[Image](#)

▼ Builder tools

[Prompt management](#) [Preview](#)[Knowledge bases](#)[Agents](#)[Prompt flows](#) [Preview](#)[Amazon Bedrock](#) > [Text playground](#)Text playground [Info](#)[Load examples](#) [Select model](#)*Try one of these examples [or view more examples](#)***Titan**
Text G1 -
Express**Action items
from a
meeting
transcript****Claude**
**Advanced
Q&A with
Citation****An example
prompt for****Llama 2**
Chat 13B**Chain of
thought****An example
prompt that****Jamba-
Instruct****Clinical
documentati
on Q&A****Gather
reliable****Command**
**Contract
Entity
Extraction****Use
generative****Mistral
Large
(24.02)****Finding the
Difference in
Payment
Dates**

Amazon Bedrock – 폭 넓은 파운데이션 모델 제공

30+ 이상의 성능 및 요금에 따른 모델 선택 옵션

AI21labs



ANTHROPIC



Meta



stability.ai

JURASSIC-2
JAMBA-INSTRUCT

TITAN:
TEXT,LITE,EXPRESS
TEXT PREMIERE
IMAGE GENERATOR V2
EMBEDDINGS V2
MULTIMODAL EMBEDDINGS

CLAUDE 3.5 SONNET
CLAUDE 3
HAIKU, SONNET,
OPUS

COMMAND R+
COMMAND R
COMMAND
EMBED

LLAMA 3.1
8B 70B 405B

MISTRAL 7B
MIXTRAL 8x7B
MISTRAL LARGE
MISTRAL SMALL

STABLE DIFFUSION
XL 1.0

CUSTOM MODEL IMPORT
파인튜닝 된 다양한 파운데이션 모델 가져오기 기능 제공



Amazon SageMaker JumpStart



LG AI Research



upstage



Hugging Face

50+ 이상의 상용 및 공개 모델 제공



생성형 AI 앱 만들기

프롬프트 엔지니어링

Amazon Bedrock 기반 신제품 출시 예제



Prompt Engineering

Introduction



LLM 설정

프롬프트의 기초

프롬프트의 구성 요소

프롬프트 설계에 관한 일반적인
팁

프롬프트 예시

Techniques



Zero-shot Prompting

Few-shot Prompting

Chain-of-Thought
Prompting

Self-Consistency

Generate Knowledge

Prompt Engineering

프롬프트 엔지니어링 가이드

프롬프트 엔지니어링은 다양한 애플리케이션과 연구 주제에 언어모델(LM)을 효율적으로 사용할 수 있도록 프롬프트를 개발하고 최적화하는 비교적 새로운 분야입니다. 프롬프트 엔지니어링 기술은 대규모언어모델(LLM)의 기능과 한계를 더 잘 이해하는 데 도움이 됩니다.

연구자는 프롬프트 엔지니어링을 사용하여 질의응답 및 산술 추론과 같은 일반적 작업부터 복잡한 대화형 인터페이스를 향상시킵니다. 개발자는 프롬프트 엔지니어링을 사용하여 LLM 및 기타 도구와 인터페이스를 설계합니다.

프롬프트 엔지니어링은 단순히 프롬프트를 설계하고 개발하는 것만이 아닙니다. 프롬프트 엔지니어링은 LLM과 인터페이스를 형성하는 중요한 기술입니다. 프롬프트 엔지니어링을 사용하여 LLM의 안전성을 개선하고 도메인 지식을 활용해 새로운 기능을 구축할 수 있습니다.

LLM을 기반으로 한 개발에 대한 높은 관심에 힘입어 이 새로운 프롬프트 엔지니어링 가이드를 만들었습니다. 프롬프트 엔지니어링에 관한 모든 최신 논문, 학습 가이드, 모델, 강의, 참고 문헌, 새로운 LLM 기능 및 도구가 포함되어 있습니다.

<https://www.promptingguide.ai/kr>

한국어



예시 - 신제품 출시를 위한 다양한 마케팅 작업



제품 홍보 설명 작성



제품 이미지 제작



소셜 미디어 홍보카피



검색 최적화 용어 선택

참고 – 본 예시는 이해를 돋기 위해 영어를 기반한 결과를 한국어로 번역한 것으로 각 파운데이션 모델이 한국어 생성을 직접 지원하는 것은 아닙니다.

Anthropic
Claude

Stability AI
Stable Diffusion

AI21 Labs
Jurassic-2

Amazon
Titan Text



PROMPT

편안함, 내구성, 메쉬 패브릭, 고무 밑창,
반사 안전 기능과 같은 디테일과 소재로
런던을 걸어 다니기에 좋은 신발에 대한
제품 설명을 만들어 주세요

OUTPUT

편안하고 튼튼한 런던 워커 신발

런던 워커 신발의 편안함과 내구성을 경험해 보세요.
도시 곳곳을 걷는 모든 모험에 적합합니다. 갑피는
통기성이 좋은 메쉬 원단으로 제작되어 더운 여름날
발을 시원하고 통풍이 잘되게 해줍니다. 고무 밑창은
고르지 않은 자갈길에서 뛰어난 견인력과 접지력을
제공합니다. 측면 패널과 뒤크치에 추가된 반사
디테일은 야간이나 어두운 곳에서 걸을 때 향상된
가시성과 안전성을 제공합니다.”

Anthropic
Claude

Stability AI
Stable Diffusion

AI21 Labs
Jurassic-2

Amazon
Titan Text



편안하고 튼튼한 런던 워커 신발

런던 워커 신발의 편안함과 내구성을 경험해 보세요. 도시
곳곳을 걷는 모든 모험에 적합합니다. 갑피는 통기성이 좋은
메쉬 원단으로 제작되어 더운 여름날 발을 시원하고 통풍이
잘되게 해줍니다. 고무 밀창은 고르지 않은 자갈길에서
뛰어난 견인력과 접지력을 제공합니다. 측면 패널과
뒤꿈치에 추가된 반사 디테일은 아간이나 어두운 곳에서
걸을 때 향상된 가시성과 안전성을 제공합니다.”

OUTPUT



PROMPT

런던을 배경으로 제품
사진을 만들어 주세요!



참고 – 본 예시는 이해를 돋기 위해 영어를 기반한 결과를 한국어로 번역한
것으로 각 파운데이션 모델이 한국어 생성을 직접 지원하는 것은 아닙니다.



© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Anthropic
Claude

Stability AI
Stable Diffusion

AI21 Labs
Jurassic-2

Amazon
Titan Text



편안하고 튼튼한 런던 워커 신발

런던 워커 신발의 편안함과 내구성을 경험해 보세요. 도시 곳곳을 걷는 모든 모험에 적합합니다. 갑피는 통기성이 좋은 메쉬 원단으로 제작되어 더운 여름날 발을 시원하고 통풍이 잘되게 해줍니다. 고무 밑창은 고르지 않은 자갈길에서 뛰어난 견인력과 접지력을 제공합니다. 측면 패널과 뒤풀침에 추가된 반사 디테일은 야간이나 어두운 곳에서 걸을 때 향상된 가시성과 안전성을 제공합니다."



PROMPT

제품 세부 정보를 기반으로 제품 발표를 위한 소셜 미디어 트윗을 작성해 주세요.

OUTPUT

"새 신발을 찾고 계신가요? London Walker만 찾으세요! 이 스타일리시하고 편안한 신발은 도시 곳곳을 걷는 모험에 적합합니다. 통기성 메쉬 패브릭, 고무 밑창, 반사 디테일이 있는 London Walker는 하루 종일 안전하고 편안하게 유지하십시오.

#LondonWalker #walkingadventures"

Anthropic Claude

Stability AI Stable Diffusion

AI21 Labs Jurassic-2

Amazon Titan Text

편안하고 튼튼한 런던 워커 신발

런던 워커 신발의 편안함과 내구성을 경험해 보세요. 도시 곳곳을 걷는 모든 모험에 적합합니다. 갑피는 통기성이 좋은 메쉬 원단으로 제작되어 더운 여름날 발을 시원하고 통풍이 잘되게 해줍니다. 고무 밀창은 고르지 않은 자갈길에서 뛰어난 견인력과 접지력을 제공합니다. 측면 패널과 뒤품지에 추가된 반사 디테일은 야간이나 어두운 곳에서 걸을 때 향상된 가시성과 안전성을 제공합니다."



"새 신발을 찾고 계신가요? London Walker만 찾으세요! 이 스타일리시하고 편안한 신발은 도시 곳곳을 걷는 모험에 적합합니다. 통기성 메쉬 패브릭, 고무 밀창, 반사 디테일이 있는 London Walker는 하루 종일 안전하고 편안하게 유지하십시오.

#LondonWalker #walkingadventures"



PROMPT

검색 가능성을
최대화하기 위해
사이트에서 이
제품과 관련된 주요
검색어 목록을
만들어 주세요!

OUTPUT

런던워커 슈즈
런던용 운동화
편안한 운동화
미끄럼 방지 운동화
안전을 위한 반사 운동화
남성용 운동화
여성용 운동화
메쉬 워킹 슈즈
방수 운동화
속건성 운동화
도시의 거리를 위한
운동화
관광객을 위한 운동화
걷기 운동화

Anthropic
Claude

Stability AI
Stable Diffusion

AI21 Labs
Jurassic-2

Amazon
Titan Text

Comfortable and Durable London Walking Shoe
This shoe is designed for walking around London. It features:

- A breathable and water-resistant upper for comfort and durability. The mesh will keep your feet ventilated while walking, but also protect from rain and puddles.
- A slip-on design with elastic inserts for easy on and off. No laces to tie, so you can quickly put them on when rushing out the door or take them off again when drying off.
- A durable, slip-resistant rubber sole. The sole provides traction for walking on wet, uneven London pavements. It is also long-lasting, so it can withstand the many miles of walking around London.
- Reflective heel and toe accents for visibility and safety. The reflective accents will make you more visible to traffic when walking at night, providing an important safety feature for London pedestrians.



제품 홍보 설명 작성



소셜 미디어 홍보 카피

"Not suitable for running or sports",
"Limited color options"

]

}



검색 최적화 용어

London walking shoe
Walking shoes for men
Comfortable walking shoes
Slip-resistant walking shoes
Reflective walking shoes for safety
Walking shoes for men
Walking shoes for women
Quick-drying walking shoes
Walking shoes for city streets
Walking shoes for tourists
Sneakers for walking



Amazon Bedrock



파운데이션 모델의 한계



Q: 조선왕조실록에 기록된 세종대왕 맥북프로 던짐 사건에 대해 알려줘.



세종대왕의 맥북프로 던짐 사건은 조선왕조실록에 기록된 일화로, 15세기 세종대왕이 새로 개발한 훈민정음(한글)의 초고를 작성하던 중 문서 작성 중단에 대해 담당자에게 분노해 맥북프로와 함께 그를 방으로 던진 사건입니다.

환각 현상



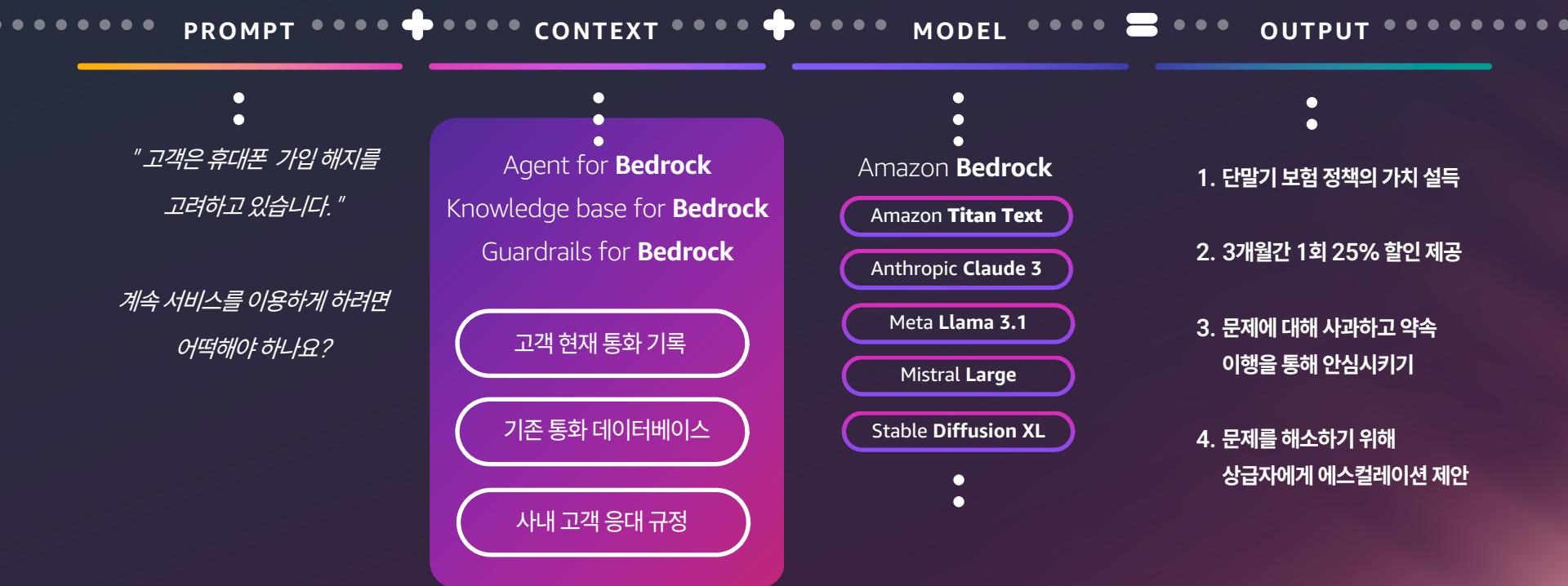
Q: 2024년의 한국 코스닥 주가 흐름에 대해 알려줘.



저는 인공 지능 모델이며, 2023년 9월 이후의 실시간 정보나 예측을 제공할 수 없습니다. 2024년 한국의 주가 동향에 대해 정확한 정보를 알고싶다면, 신뢰할 수 있는 금융 사이트나…

지식단절현상

생성형 AI 앱 만들기



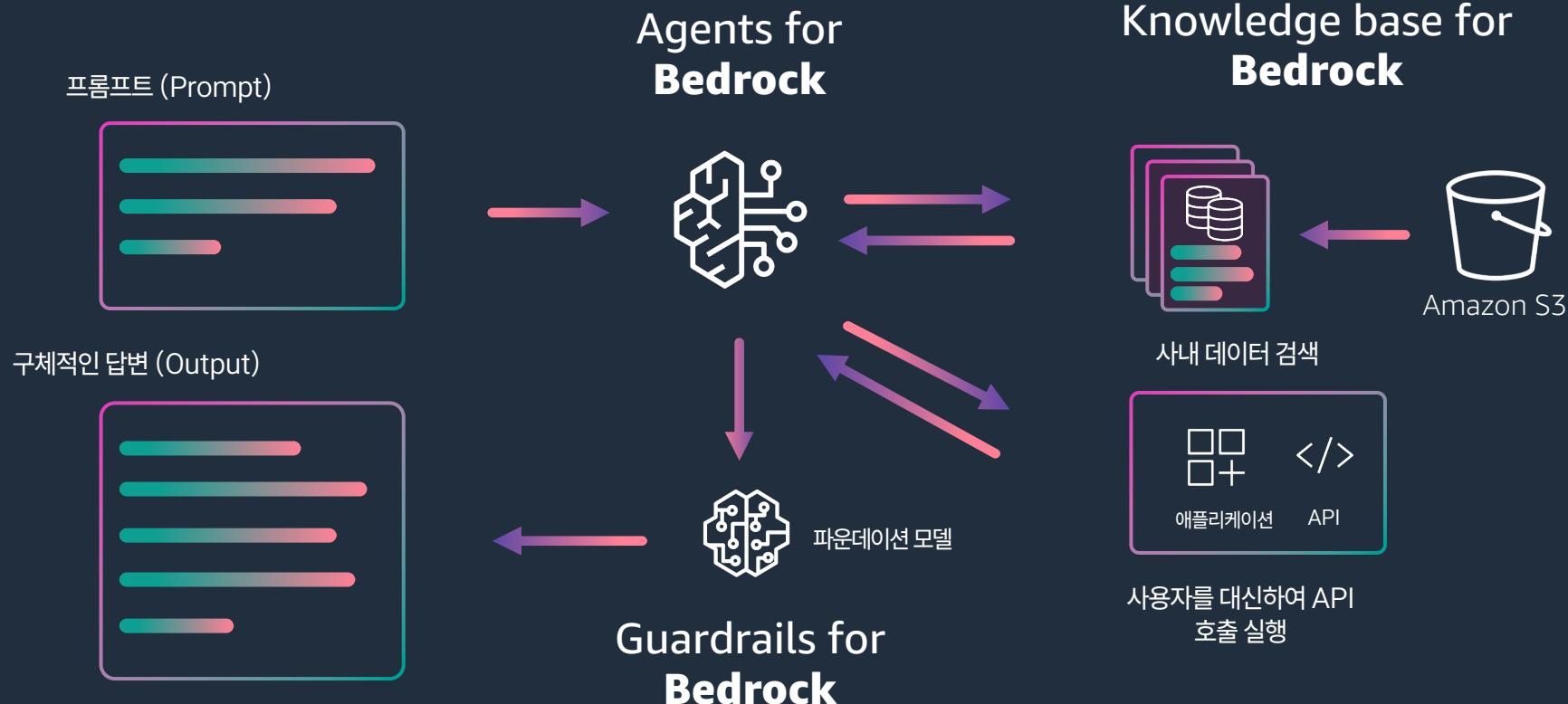
생성형 AI 앱 만들기

검색 증강 생성

Retrieval-Augmented Generation (RAG)



검색 증강 생성 (RAG)

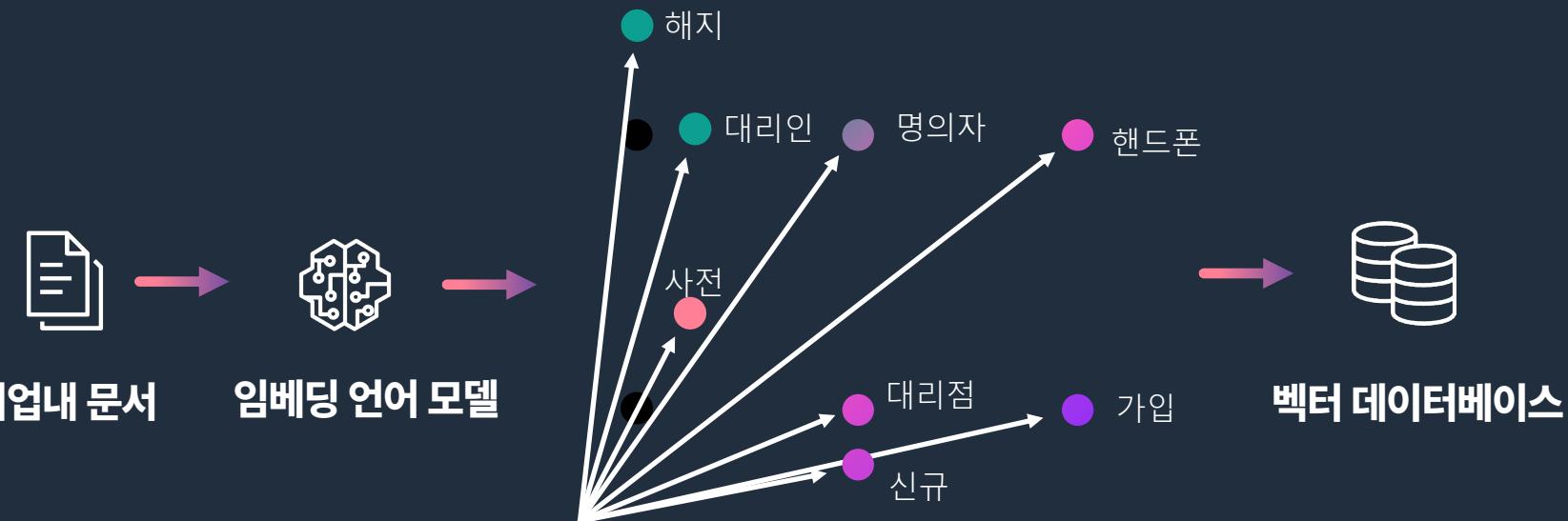


생성형 AI 지식 기반을 위한 벡터 임베딩

사람이 읽을 수 있는 텍스트

해지 대리인 사전 지정 제도는 휴대폰 명의자 본인이 직접 휴대폰을 해지할 수 없는 경우에 대비하기 위하여 미리 해지 대리인을 지정해 두는 제도입니다. 나중에 휴대폰을 해지할 때 명의자 본인 또는 해지 대리인이 지점/대리점에 방문하여 휴대폰을 해지하실 수 있습니다.

생성형 AI 지식 기반을 위한 벡터 임베딩



벡터 데이터베이스 및 벡터 검색 기능

- Vector databases for Amazon Bedrock



Amazon
OpenSearch
Serverless



Amazon
Aurora



Pinecone



Redis Enterprise
Cloud



MongoDB
Atlas

- Vector search in AWS Databases



Amazon
DocumentDB



Amazon
DynamoDB



Amazon
Memory DB



Amazon Neptune
Analytics

검색 증강 생성 (RAG)

프롬프트 (Prompt)

“엄마가 제 핸드폰 대신 해지할 수 있나요?”

Agents for Bedrock



Knowledge base for Bedrock



0.00526205 0.0057839 0.00856714 ...
0.00083171 -0.00273974 -0.01069805
[0.02154857 0.01325288 0.00202334
... -0.01588322 -0.00311089
0.01122954] [0.00033921 0.00205388

벡터 데이터베이스

휴대폰 (0.9) 해지 (0.8) 대리인 (0.7)

구체적인 답변 (Output)

“대리인이 핸드폰을 해지하려면, 대리인 해지 제도를 이용하면 됩니다. 미리 해지 대리인을 지정하시면 가능합니다. 혹시 더 많은 질문이 있으시면, 고객 상담 센터 챗봇에게 남겨 주시면 연락드리겠습니다.”

Guardrails for Bedrock

규정 및 잘못된 응답 검증

증강된 프롬프트

대리인 해지 제도

휴대폰 명의자 본인이 직접 휴대폰을 해지할 수 없는 경우에 대비하여 미리 해지 대리인을 지정해 두는 제도

“대리인이 핸드폰 해지 할 경우,
어떻게 해야 하나요?”

Playground

Model

Anthropic Claude 3 Sonnet



Reset

Quick start prompts

Rewrite the below email with a professional tone

Summarize the following text for me

Simulate a job interview with a candidate

Enter prompt

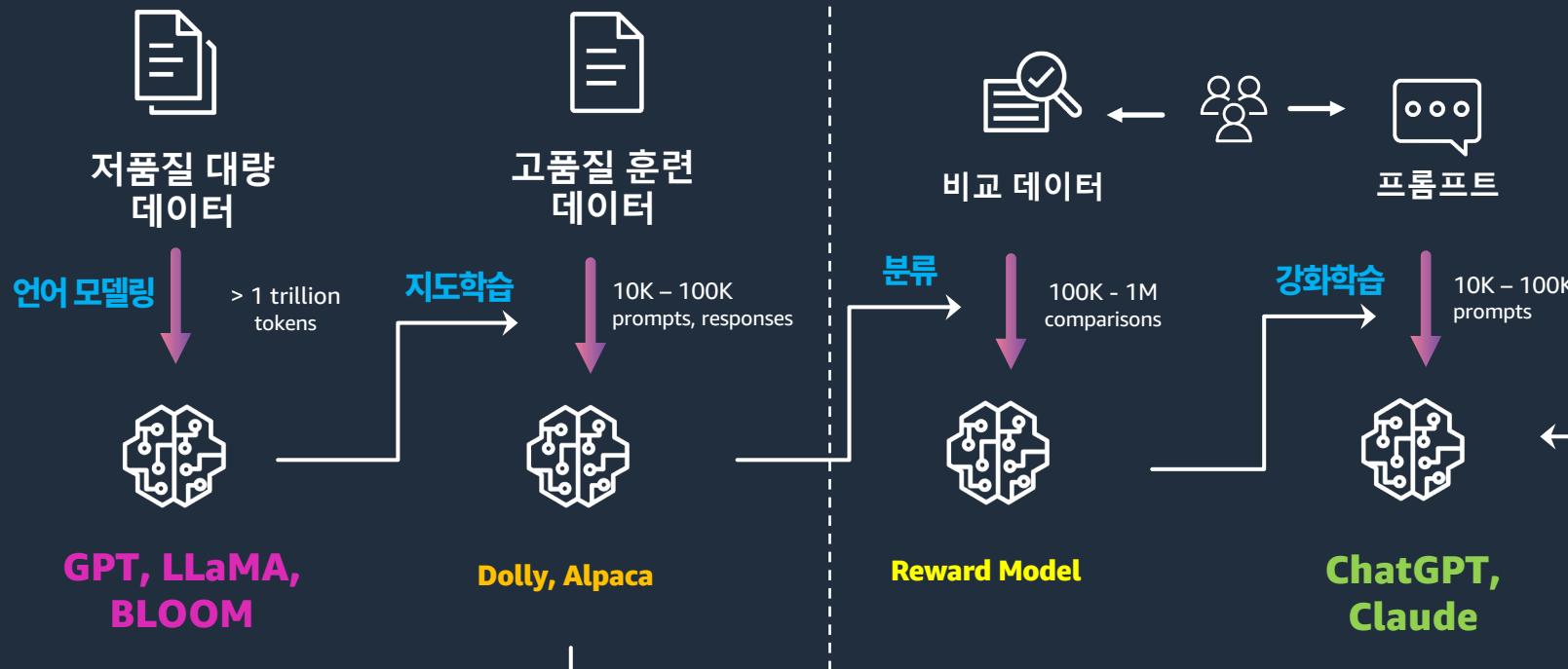
생성형 AI 앱 만들기

모델 미세 조정

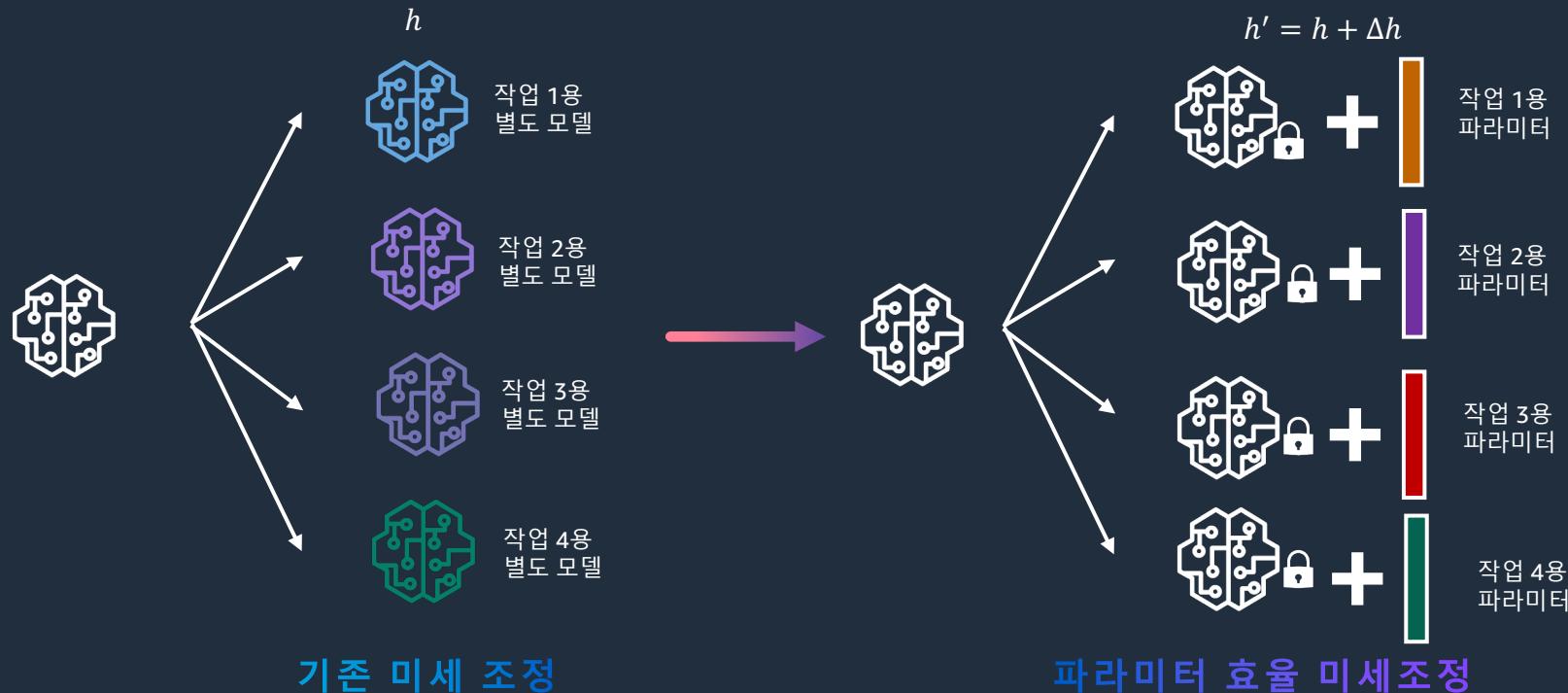
Model Fine-Tuning



파운데이션 모델의 미세 조정 과정

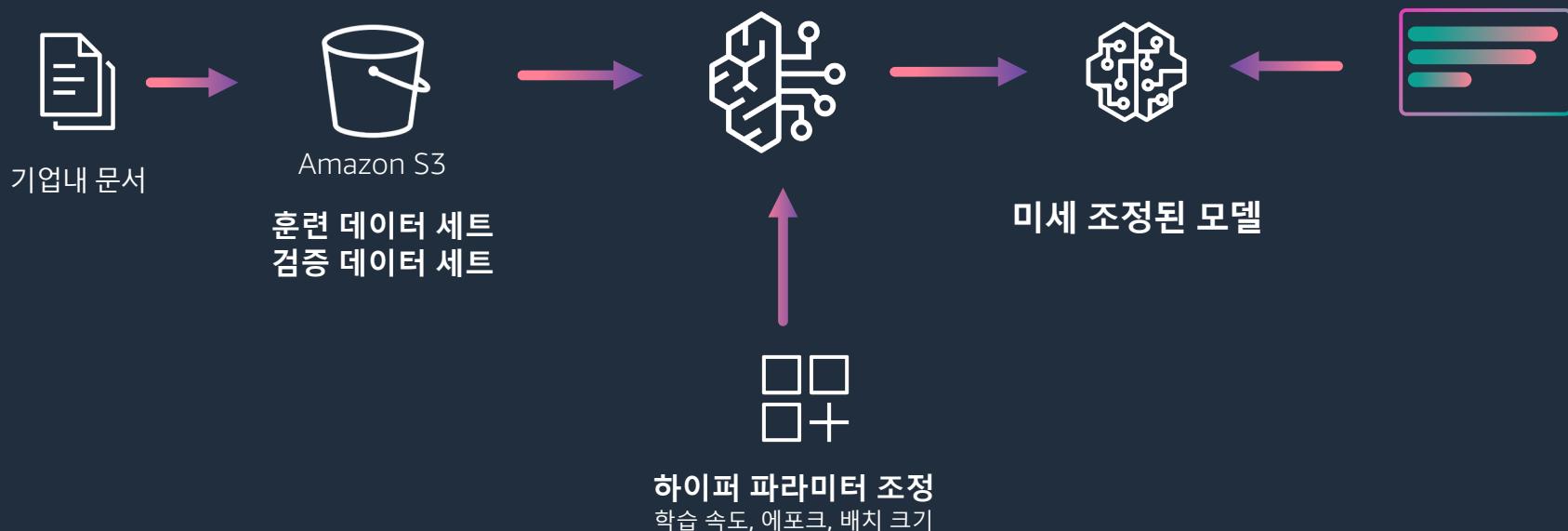


파운데이션 모델의 미세 조정 과정



Amazon Bedrock 미세 조정 방법

Bedrock Fine-tuning



Amazon Bedrock



Getting started

[Overview](#)[Examples](#)[Providers](#)

Foundation models

[Base models](#)[Custom models](#)[Imported models](#) [Preview](#)

Playgrounds

[Chat](#)[Text](#)[Image](#)

Builder tools

[Prompt management](#) [Preview](#)[Knowledge bases](#)[Agents](#)[Prompt flows](#) [Preview](#)[Amazon Bedrock](#) > [Custom models](#)

Custom models



Customize a model with Fine-tuning or Continued Pre-training. You can also share or copy custom models.

▼ How it works

Create a model



Select a model to customize and provide the Amazon S3 bucket that contains the training (and validation, if applicable) dataset. You can share your custom models with other accounts if you have an [AWS Organizations](#) account. You can copy a model that you own or that has been shared with you to the region that you want to use it in.

Purchase provisioned throughput to use the custom model



Purchase provisioned throughput to use a custom fine-tuned model, a continued pre-trained model or a copied model. You can then use your model in the Amazon Bedrock playgrounds or in your applications.

[Models](#)[Jobs](#)[Models \(0\)](#)[Purchase provisioned throughput](#)[Customize model ▾](#)



AWS GenAI Loft Seoul

TIPS Town S1, 747-2 역삼1동 Gangnam District,
Seoul, South Korea

[Register for upcoming events](#)



10월 30일 – 11월 6일 | 역삼동 팀스타운





AWS Community Day 2024 발표 및 스폰서 모집

안녕하세요! AWS Community Day 2024 행사가 11월 2일(토)에 개최됩니다. AWS 사용자(AWSKRUG)이 주최하는 이번 행사는 AWS 클라우드에 대한 최신 기술에 대한 소개와 아울러 용자들이 어떻게 서비스 개발에 클라우드를 활용하고 있는지 사례를 공유하는 행사이입니다. 또한 AWS 클라우드 기반 개발 경험 및 서비스 구현 사례에 대한 발표 신청을 받습니다.

행사 개요

- 행사 일자 및 장소: 2024년 11월 2일(토) 13:30-18:00 AWS Korea Office 및 AWS Loft Seoul (총 300명)
- 발표자 선정 일정: 신청 마감: 9월 28일(토) / 발표자 통보: 9월 30일(월)

