



AWS BUILDERS KOREA PROGRAM SPECIAL

Generative AI on AWS

하루만에 끝내는 생성형 AI

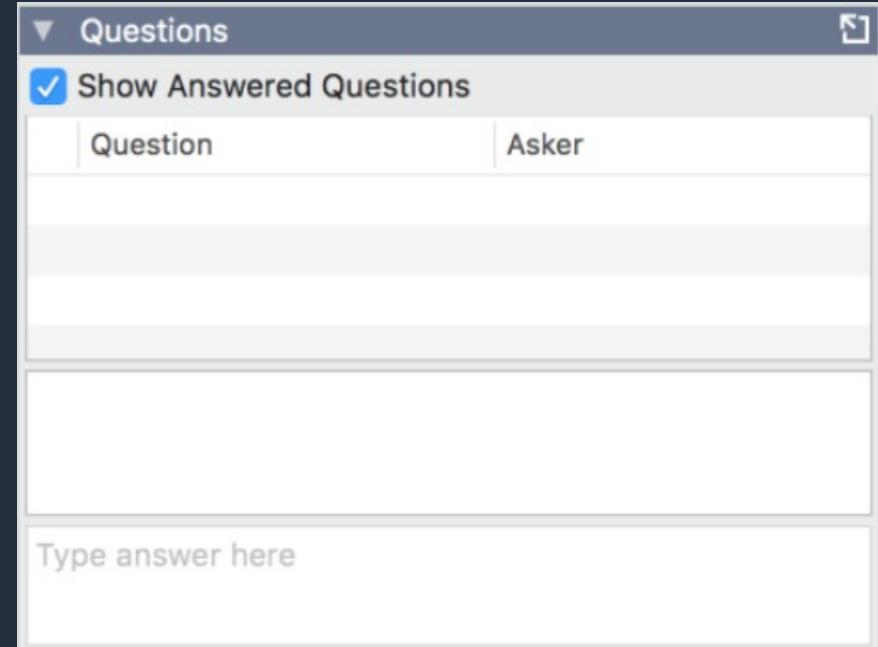
Suji Lee

Solutions Architect
Amazon Web Services

강연 중 질문하는 방법

AWS Builders Korea Go to Webinar “**Questions (질문)**” 창에
자신이 질문한 내역이 표시됩니다. 본인만 답변을 받고 싶으실
경우 (비공개)라고 하고 질문해 주시면 됩니다.

질문 주신 사항에 대해서는 질문창을 통해 답변을 드립니다.



고지 사항 (Disclaimer)

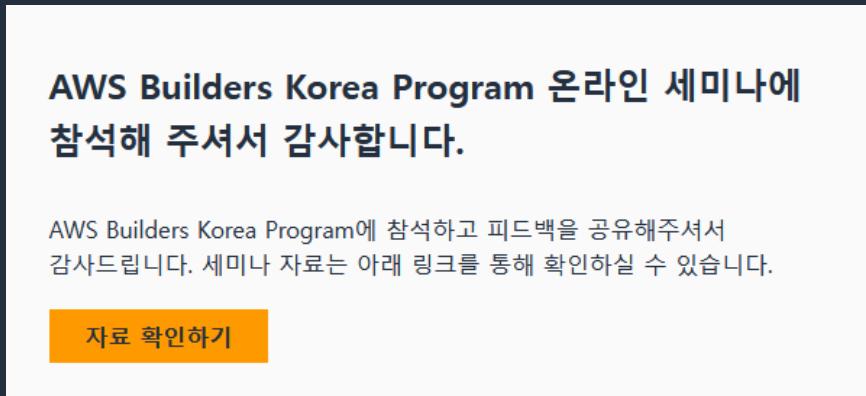
본 컨텐츠는 고객의 편의를 위해 AWS 서비스 설명을 위해 온라인 세미나용으로 별도로 제작, 제공된 것입니다. 만약 AWS 사이트와 컨텐츠 상에서 차이나 불일치가 있을 경우, AWS 사이트(aws.amazon.com)가 우선합니다. 또한 AWS 사이트 상에서 한글 번역문과 영어 원문에 차이나 불일치가 있을 경우(번역의 지체로 인한 경우 등 포함), 영어 원문이 우선합니다.

AWS는 본 컨텐츠에 포함되거나 컨텐츠를 통하여 고객에게 제공된 일체의 정보, 콘텐츠, 자료, 제품(소프트웨어 포함) 또는 서비스를 이용함으로 인하여 발생하는 여하한 종류의 손해에 대하여 어떠한 책임도 지지 아니하며, 이는 직접 손해, 간접 손해, 부수적 손해, 징벌적 손해 및 결과적 손해를 포함하되 이에 한정되지 아니합니다.

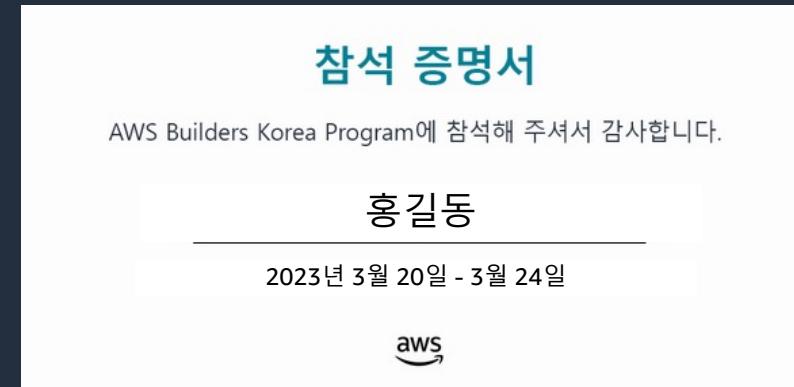
감사 메일 & 참석 증명서

- AWS Builders Korea 세션에 참석해 주신 분들께 행사 종료 후 1개월 내 감사메일과 참석 증명서가 순차 발송됩니다.
- 등록 진행 후 참석하지 않으실 경우 별도 메일 및 증명서는 발급되지 않습니다.

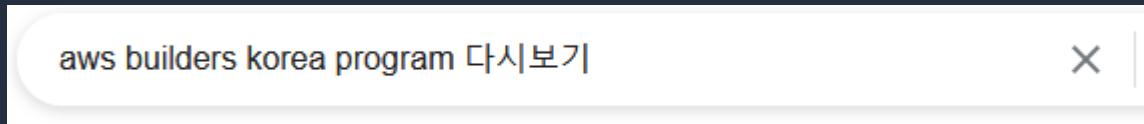
감사 메일 예시



참석 증명서 예시

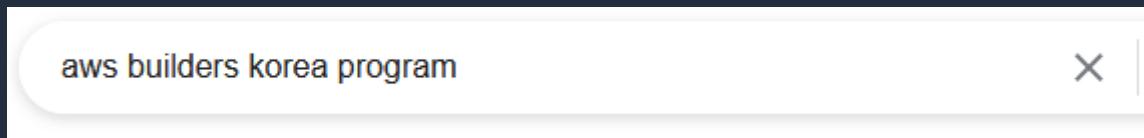


강연 다시보기



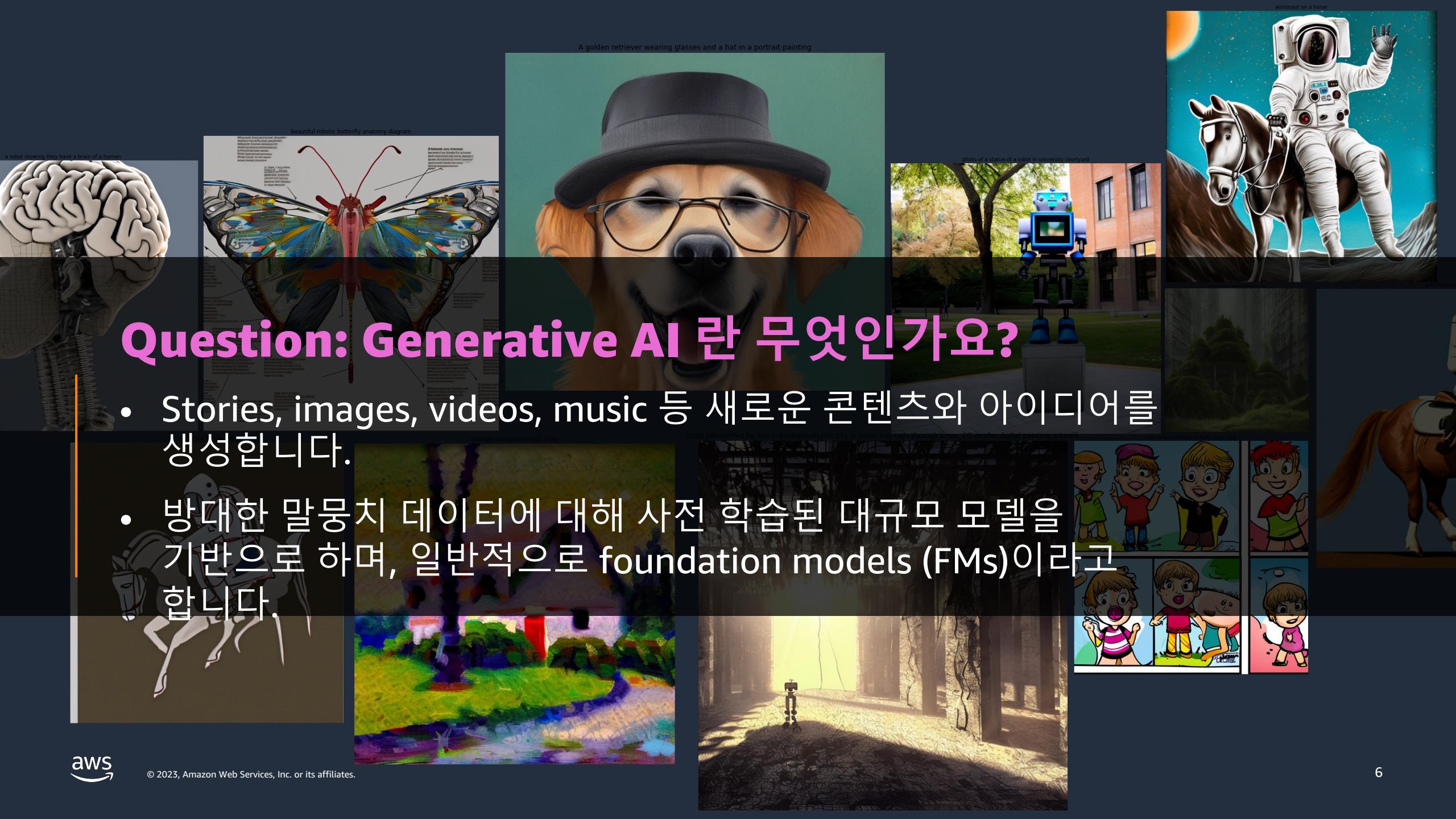
<https://kr-resources.awscloud.com/aws-builders-korea-program>

AWS Builders Korea 프로그램 정보



<https://aws.amazon.com/ko/events/seminars/aws-builders/>

Generative AI 소개



Question: Generative AI 란 무엇인가요?

- Stories, images, videos, music 등 새로운 콘텐츠와 아이디어를 생성합니다.
- 방대한 말뭉치 데이터에 대해 사전 학습된 대규모 모델을 기반으로 하며, 일반적으로 foundation models (FMs)이라고 합니다.

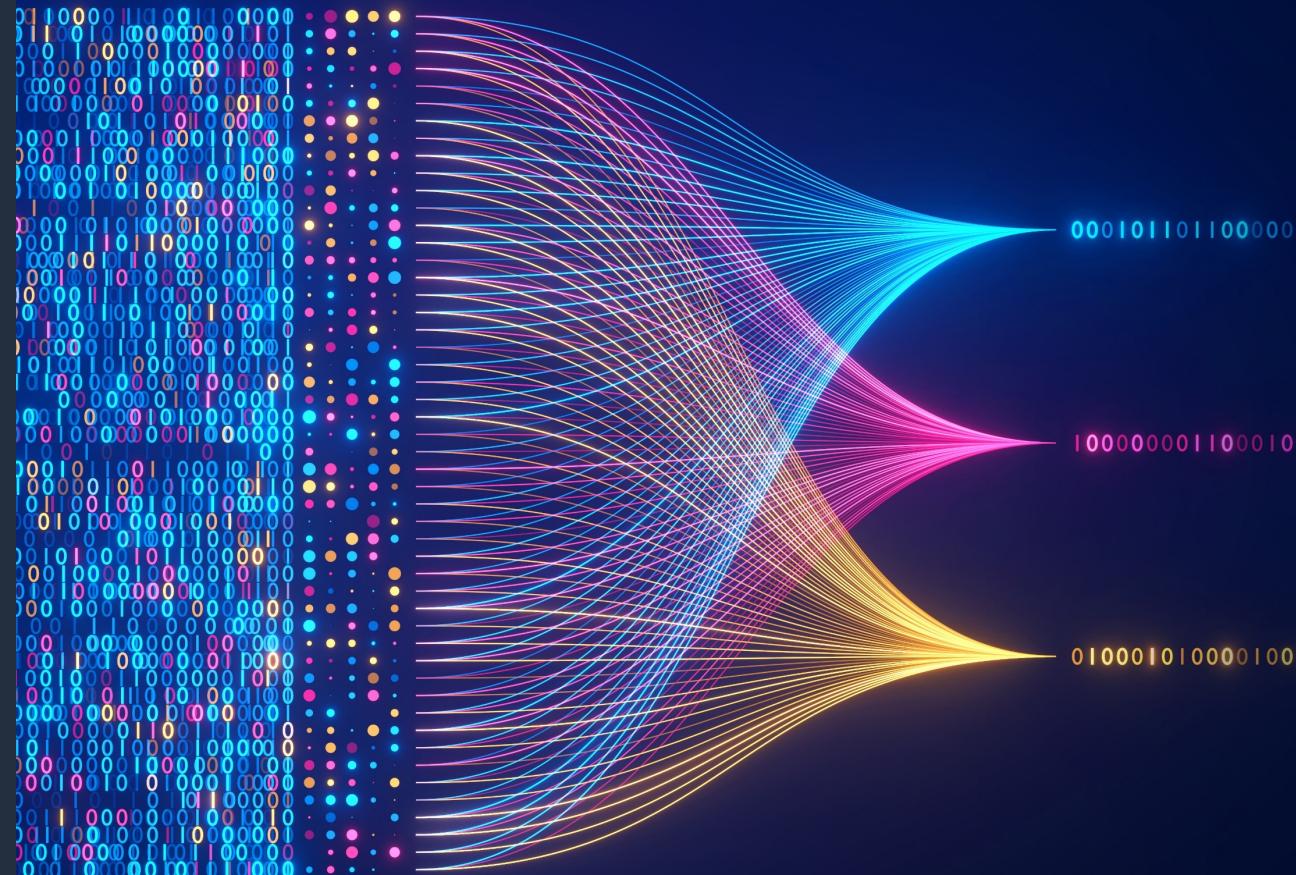
Foundation model을 기반으로 하는 Generative AI

방대한 양의 비정형 데이터에 대해
사전 학습된 기능

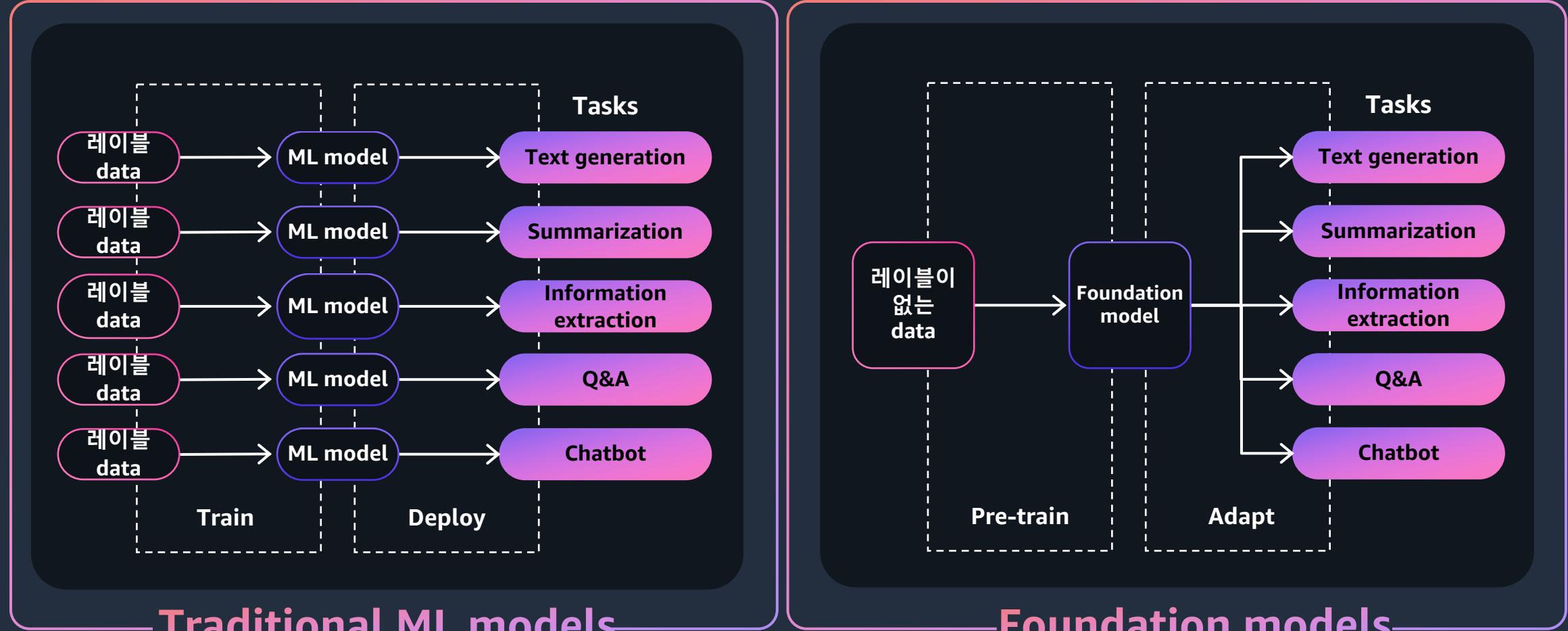
복잡한 개념을 학습할 수 있는
많은 수의 매개변수 포함

다양한 컨텍스트에 적용 가능

도메인별 작업에 맞게 데이터를 사용하여 FM 을
사용자 정의



Foundation model 이 다른 ML 모델과 어떻게 다른가



Foundation models 유형

Input

“걷기가 심장 건강에 미치는 영향에 대한 기사 요약”



FM

Text-to-text
자연어 프롬프트에서 텍스트 생성



Output

“건강한 심장을 유지하려면 하루 만보 걷기가 최적입니다.”

“손 비누”

Text-to-embeddings
텍스트의 숫자 표현 생성

다음을 숫자로 표현
“손 비누 리필
손 비누 디스펜서
항균 손 비누”

“화성에서 말을 타고 있는 우주 비행사의 사진”

Multimodal
자연어 프롬프트를 사용하여 이미지 생성 및 편집



중요한 비즈니스 가치를 창출하는 Generative AI



새로운 경험

고객 및 직원과 소통하는
새롭고 혁신적이며
매력적인 방법 창출.



생산성

모든 비즈니스 라인의
생산성을 획기적으로
개선하세요, 예를
들어 [Amazon](#)
[CodeWhisperer](#) 는 작업을
57% 더 빠르게 완료하도록
지원합니다.



인사이트

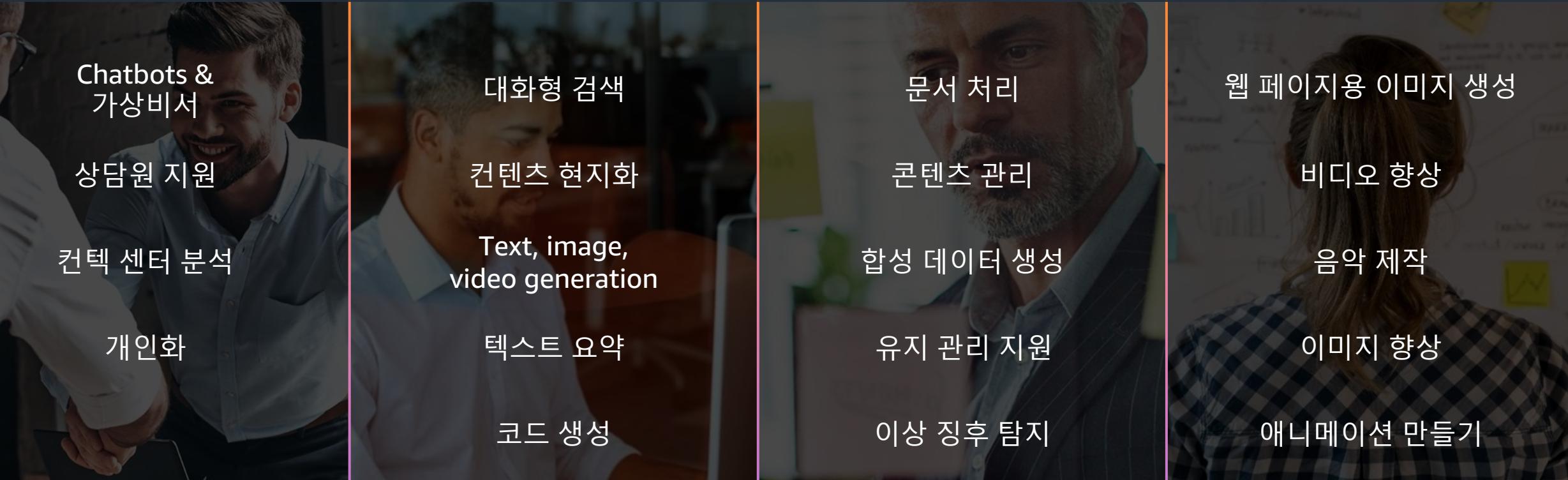
모든 기업 정보에서
인사이트와 명확한 답을
추출하여 더 빠르고 더 나은
의사 결정을 내릴 수 있습니다.



창의성

대화, 스토리, 이미지,
동영상, 음악 등 새로운
콘텐츠와 아이디어를
생성하세요.

Generative AI 는 다양한 사용 사례에 사용할 수 있습니다



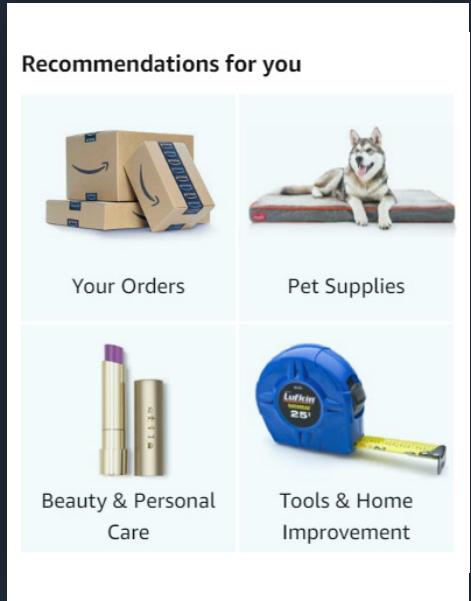
고객 경험 향상

직원 생산성 향상

비즈니스 운영 개선

창의성 향상

ML 혁신은 Amazon 의 DNA 에 있습니다



Amazon.com에서 **분당 4,000 개의 상품이 판매됩니다**



매일 **160만 건의 패키지 발송**



매주 **수십억 건의** of Alexa 상호작용



2016년 12월 7일 첫 프라임 에어 배송



100,000 명 이상의 고객이 ML을 위해 AWS를 사용하고 있습니다



모든 산업을 변화시키는 Generative AI



금융 서비스



의료 및 생명과학



자동차



제조



미디어 & 엔터테이먼트



리테일



통신



에너지



여행 & 숙박



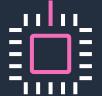
소비재

FMs 와 Generative AI 의 잠재력 활용하기

Generative AI 의 잠재력 활용



FMs 을 사용하여 구축하는 가장 쉬운 방법



가격 대비 성능이 가장 뛰어난 인프라



Generative AI 기반 애플리케이션



유연성

Generative AI 의 잠재력 활용



FMs 을 사용하여 구축하는 가장 쉬운 방법



가격 대비 성능이 가장 뛰어난 인프라



Generative AI 기반 애플리케이션



유연성

Amazon Bedrock

FMS로 GENERATIVE AI 애플리케이션을 구축하고 확장하는
가장 쉬운 방법



이점

- API를 통해 FM을 사용하여 Generative AI 애플리케이션 개발 가속화
- 인프라 관리 불필요
- AI21 Labs, Anthropic, Cohere, Stability AI, Amazon의 FM을 함께 사용
- 조직의 데이터를 사용하여 비공개적으로 FM을 사용자 정의 가능
- 포괄적인 AWS 보안 기능
- NEW – Bedrock용 에이전트를 사용하여 몇 번의 클릭만으로 작업을 완료할 수 있는 Generative AI 앱 지원

Amazon Bedrock supports leading foundation models



Amazon Titan

텍스트 요약, 생성, 분류,
개방형 Q&A, 정보 추출,
임베딩 및 검색



Jurassic-2

스페인어, 프랑스어,
독일어, 포르투갈어,
이탈리아어,
네덜란드어로 텍스트를
생성할 수 있는 다국어
LLM



new Claude 2

사려 깊은 대화, 콘텐츠
제작, 복잡한 추론, 창의성,
코딩을 위한 현법 AI 및
무해성 교육을 기반으로
하는 LLM



new Command + Embed

비즈니스 애플리케이션을
위한 텍스트 생성 모델 및
100개 이상의 언어로 검색,
클러스터링 또는 분류를
위한 임베딩 모델



new Stable Diffusion XL 1.0

독특하고 사실적인 고품질
이미지, 아트, 로고 및
디자인 생성



Bedrock 용 에이전트를 통해 FMs 작업을 지원



작업을 세분화하고
조율



회사 데이터에
안전하게
액세스하고 검색



사용자를 대신하여
API 호출을 실행하여
작업 수행



완전 관리형
인프라 제공

몇 가지 간단한 단계로 Bedrock용 에이전트 설정하기



1

Foundation Model
선택



2

기본 지침 제공



3

관련 데이터 소스 선택



4

개발자가
람다 함수를 지정

애플리케이션에 FMs 을 빠르게 통합



Amazon SageMaker 및 Amazon S3와 같은
깊이 있고 광범위한 AWS 기능 및
서비스와의 친숙한 제어 및 통합을 사용하여
**AWS에서 실행되는 애플리케이션 및
워크로드에 FMs를 빠르게 통합 및
배포하세요**

안전한 사용자 지정으로 차별화 촉진



Fine-tune

목적

특정 작업에 대한 정확도 극대화

데이터
요구사항

작은 수의 레이블이 지정된 예제

Amazon Titan

AMAZON의 고성능 FMS로 책임감 있게 혁신하기



NLP 작업에 중점을 둔
Titan Text



검색 및 개인화와 같은
엔터프라이즈 작업용
Titan Embeddings

이점

- 20년 이상의 Amazon ML 경험을 바탕으로 구축
- Amazon Titan Text FM으로 요약 및 텍스트 생성과 같은 언어 작업 자동화
- Amazon Titan Embeddings FM으로 검색 정확도 향상 및 개인화된 추천 개선
- 부적절하거나 유해한 콘텐츠를 줄여 AI의 책임감 있는 사용 지원

Deloitte 는 Amazon Bedrock 을 통해 연간 수만 시간을 절약할 수 있을 것으로 기대합니다.

“Deloitte는 고객이 AI의 힘을 활용할 수 있도록 제너레이티브 AI 역량을 발전시키고 있습니다. 이러한 노력의 일환으로 딜로이트는 제휴 관계를 통해 아마존 베드락과 같은 선도적인 서비스를 활용하여 이러한 기능을 발전시키고 있습니다. 베드락을 통해 AWS 고객이 제너레이티브 AI 애플리케이션을 구축할 수 있는 비용 효율적인 서비스 API를 고객에게 제공할 수 있습니다.”

Nishita Henry

Amazon/AWS Alliance Global Chief Commercial Officer,
Deloitte Consulting LLP

Deloitte.

Salesforce, Amazon Bedrock으로 제너레이티브 AI 앱 가속화

“ 생성형 AI를 위해 'Bring Your Own AI' 통합을 Amazon Bedrock으로 확장합니다! 제로-ETL을 통해 Salesforce Data Cloud 데이터에 안전하고 쉽게 액세스할 수 있으며, 해당 회사 데이터를 사용하여 Bedrock을 사용하여 원하는 FM을 빠르고 안전하게 사용자 지정할 수 있습니다. 이렇게 귀사에 맞게 맞춤화된 기초 모델을 Data Cloud에서 쉽게 호출하여 Salesforce 전체에서 사용할 수 있습니다.”

Gabrielle Tao
Vice President, Data Cloud, Salesforce



Philips 아마존 베드락을 사용하여 제너레이티브 AI 애플리케이션 개발

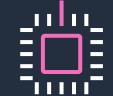
아마존 베드락은 필립스가 효율적인 임상 워크플로우를 지원하고 진단 역량을 강화하는 제너레이티브 AI 애플리케이션을 개발할 수 있도록 지원하는 포트폴리오의 일부로, **의료진이 인력 부족 속에서 증가하는 워크로드를 관리하고 진단 및 치료 시간을 단축할 수 있도록 지원합니다.**

PHILIPS

Generative AI 의 잠재력 활용



FMs 을 사용하여 구축하는 가장 쉬운 방법



가격대비 성능이 가장 뛰어난 인프라



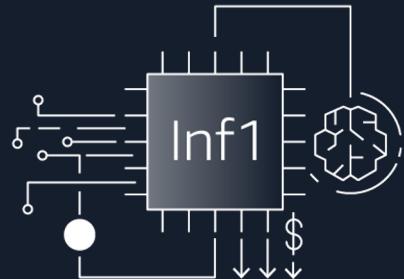
Generative AI 기반 애플리케이션



유연성

Generative AI 를 위해 특별히 설계된 가속기

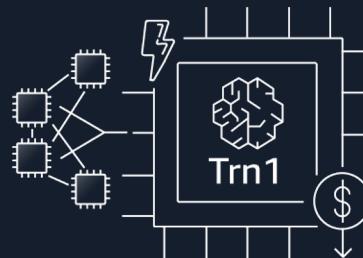
AWS Inferentia



딥 러닝(DL) 모델 실행을 위한
클라우드에서 추론 당 최저
비용

동급 아마존 EC2 인스턴스 대비
추론당 최대
70% 저렴한 비용

AWS Trainium



가장 비용 효율적인 고성능
LLM 및 확산 모델 학습

동급 아마존 EC2 인스턴스 대비
교육 비용 50% 절감

AWS Inferentia2



LLM 및 확산 모델을 위한 최저
추론 당 비용으로 고성능 제공

동급 아마존 EC2 인스턴스 대비
최대 40% 더 나은 가격 대비
성능

GPU 기반 컴퓨팅을 제공하는 탁월한 경험

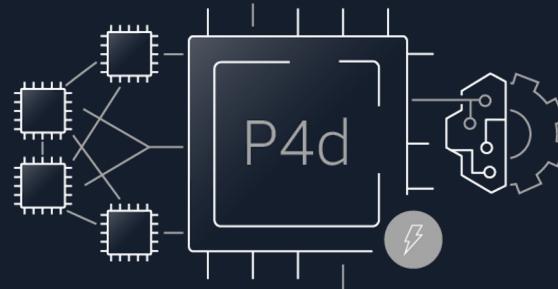
Amazon EC2 P5 instances

Coming soon!

NVIDIA H100 텐서 코어 GPU 기반

이전 세대 GPU 기반 인스턴스보다
최대 6배 빠르고 훈련 비용
최대 40% 절감

Amazon EC2 P4d/P4de instances



NVIDIA A100 텐서 코어 GPU 기반

이전 세대 P3/P3dn 인스턴스보다
최대 2.5배 빨라진 속도와
최대 60% 절감된 트레이닝
비용

Amazon EC2 G5 instances



NVIDIA A10G 텐서 코어 GPU 기반

이전 세대 G4dn 인스턴스보다
최대 3.3배 더 높은 성능

Finch Inferentia로 80% 비용 절감



CHALLENGE

FM 추론 작업을 지원하기 위해 GPU를 사용하는 비용 효율적인 방법이 필요했습니다.

SOLUTION

Finch Computing은 언어 번역, 텍스트 요약, 헤드라인 생성 등 다양한 사용 사례에서 FM 추론을 위해 AWS Inferentia를 사용합니다.

OUTCOME

- ✓ 비용 80% 절감
- ✓ 3개 언어 추가 지원 확대
- ✓ 시장 출시 시간 단축



Runway: Inf2로 모델 처리량 2배 증가

AI Magic Tools

- All
- Generative
- Image
- Video
- 3D
- Audio



Photo of a tree
Text to Image
Generate images from text descriptions



Erase and Replace
Erase and replace parts of an image with generated content



A summer, duotone filter
Text to Color Grade
Color grade your video with only text



Image to Image
Modify an existing image with text



Infinite Image
Expand an image by generating outside the original canvas



Frame Interpolation
Create an animated sequence video from uploaded images

“런웨이에서는 AI 매직 툴 제품군을 통해 사용자가 이전과는 전혀 다른 방식으로 콘텐츠를 생성하고 편집할 수 있습니다. Amazon EC2 Inf2 인스턴스를 통해 일부 모델을 최대 2배 더 높은 처리량으로 실행할 수 있습니다. 이를 통해 더 많은 기능을 도입하고 더 복잡한 모델을 배포할 수 있으며, 궁극적으로 런웨이를 사용하는 수백만 명의 크리에이터에게 더 나은 경험을 제공할 수 있습니다.”

Cristobal Valenzuela, CEO



Generative AI 의 잠재력 활용



FMs 을 사용하여 구축하는 가장 쉬운 방법



가격대비 성능이 가장 뛰어난 인프라



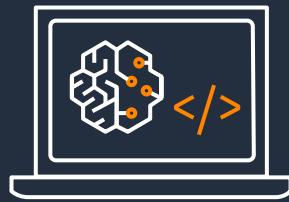
Generative AI 기반 애플리케이션



유연성

Amazon CodeWhisperer

개인 개발자가 무료로 사용할 수 있는 AI 코딩 컴패니언으로
애플리케이션을 더 빠르고 안전하게 빌드하세요.



실시간으로
코드 제안 생성



찾기 어려운 취약점이
있는지 코드 스캔하기



오픈 소스 학습 데이터와
유사한 플래그 코드 또는
기본 필터링

미리 보기 기간 동안 Amazon은 생산성 챌린지를 진행했는데, Amazon CodeWhisperer를 사용한 참가자는 사용하지 않은 참가자에 비해 작업을 성공적으로 완료할 확률이 27% 더 높았으며 평균 57% 더 빨리 완료했습니다.

Amazon CodeWhisperer로 개발자 생산성을 개선한 Accenture



“Accenture는 Velocity 플랫폼에서 소프트웨어 엔지니어링 모범 사례 이니셔티브의 일환으로 코딩을 가속화하기 위해 Amazon CodeWhisperer를 사용하고 있습니다. Velocity 팀은 개발자의 생산성을 향상할 방법을 찾고 있었습니다. 여러 가지 옵션을 검색한 끝에 개발 노력을 30% 절감할 수 있는 Amazon CodeWhisperer를 발견했습니다.

이제 보안, 품질, 성능 개선에 더욱 집중하고 있습니다.”

Balakrishnan Viswanathan

Tech Architecture Sr. Manager, Accenture

Generative AI 의 잠재력 활용



FMs 을 사용하여 구축하는 가장 쉬운 방법



가격대비 성능이 가장 뛰어난 인프라



Generative AI 기반 애플리케이션



유연성

Amazon SageMaker를 사용하여 대규모로 나만의 FM 구축하기



관리형 인프라스트럭쳐

가성비가 뛰어난 관리형 인프라로 모델 트레이닝을 완벽하게 제어하세요.



효율적인 분산 학습

최대 40% 더 빠르게 분산 교육 완료



디버깅 및 실험 도구

실시간으로 메트릭을 캡처하고 트레이닝 작업을 프로파일링하여 성능 문제를 빠르게 해결하세요. ML 모델 반복을 쉽게 추적할 수 있습니다.



가격 대비 성능 추론

최고의 가성비로 모든 사용 사례에 맞는 모델을 프로덕션 환경에 배포하세요.



반복 및 재현 가능한 MLOps

ML 라이프사이클 전반의 프로세스 자동화 및 표준화



거버넌스

책임감 있게 ML을 구현하는 데 도움이 되는 목적에 맞게 설계된 거버넌스 도구



Human-in-loop 지원

고품질 데이터 세트를 생성하고 사람의 선호도에 맞게 모델 결과물을 조정합니다.

SageMaker JumpStart에서 더 많은 모델 살펴보기

Publicly available			Proprietary models		
stability.ai			co:here		
Models Text2Image Upscaling	Models AlexaTM 20B	Models Flan T-5 (8 variants) DistilGPT2, GPT2 Bloom (3 variants)	Models Cohere Command	Models Lyra-Fr 10B	Models Jurassic-2 Jumbo
Tasks Generate photo-realistic images from text input Improve quality of generated images	Tasks Machine translation Question answering Summarization Annotation Data generation	Tasks Machine translation Question answering Summarization Annotation Data generation	Tasks Text generation Information extraction Question answering Summarization	Tasks Text generation Keyword extraction Information extraction Question answering Summarization	Tasks Text generation Long-form generation Summarization Paraphrasing Chat Information extraction Question answering Classification

FMs 를 사용하여 구축하는 데 JumpStart를 사용하는 방법

1

모델 공급자가 제공하는 더 많은
FMs 중에서 선택

AI21labs

Light^{on}
We bring Light to AI

stability.ai

co:here



alexa

2

모델링 및/또는
배포 체험



AWS Console을 통해
모델 사용해보기



단일 노드가 포함된
SageMaker 호스팅 옵션을
사용하여 추론을 위한 모델
배포

3

모델 Fine Tune 및
ML 워크플로 자동화



선택한 모델만
Fine tune 가능



ML 워크플로
자동화

모델, 인스턴스, 로그,
모델입력, 모델
출력을 포함한
**데이터는 계정에
유지 됩니다.**

Amazon
SageMaker와
완벽하게 통합 가능

Amazon SageMaker를 활용한 LG AI Research FM 개발

“분산 학습을 최적화하여 모델을 59% 더 빠르게 학습할 수 있었습니다(Amazon SageMaker를 사용하지 않았을 때보다).”

Seung Hwan Kim
Vice President, Vision Lab Leader, LG AI Research



LG AI Research's Tilda, the AI artist powered by EXAONE



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

Amazon SageMaker로 FM 개발을 가속화한 AI21 Labs

AI21 labs

CHALLENGE

AI21 Labs는 대규모 언어 모델을 구축합니다. 이러한 모델은 수백 억에서 수천 억 개의 매개변수가 포함된 거대한 신경망으로, 복잡성과 리소스 요구 사항으로 인해 이를 학습시키는 것은 큰 과제입니다.

SOLUTION

AI21 Labs는 Amazon SageMaker를 사용하여 170억 개의 파라미터로 쥬라기-그란데 모델을 훈련했습니다. Amazon SageMaker는 모델 학습 프로세스를 더 쉽고 효율적으로 만들었으며, DeepSpeed 라이브러리와 완벽하게 작동했습니다.

OUTCOME

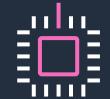
- ✓ 더 쉽고 효율적인 모델 트레이닝
- ✓ 분산된 트레이닝 작업을 수백 개의 NVIDIA A100 GPU로 쉽게 확장
- ✓ 추론 비용 절감



FM과 Generative AI의 잠재력 활용하기



FMs 을 사용하여 구축하는 가장 쉬운 방법



가격대비 성능이 가장 뛰어난 인프라



Generative AI 기반 애플리케이션



유연성



지금 바로 Generative AI 여정을 시작하세요



지금 바로 Generative AI 여정을 시작하세요

1

지금 Amazon
CodeWhisperer로
생산성 향상
시작하기

2

Amazon Bedrock
및 JumpStart의
다른 FMs을 통해
FMs 탐색하기

3

주요 사용 사례에
대한
PoC 시작하기



AI 사용 사례 탐색기

관련 콘텐츠 및 지침을 통해 가장 관련성이 높은 AI 사용 사례를 쉽게 찾아 실제 적용하세요.

The screenshot shows the AWS AI Use Case Explorer landing page. The header features the AWS logo and a search bar with the URL https://aiexplorer.aws.amazon.com. Below the header, the title "AI Use Case Explorer" is displayed. Three main navigation links are shown: "Explore Use Cases" (selected), "Discover Success Stories", and "Mobilize Your Team". A central call-to-action button says "Explore The Art Of The Possible In AI". Below it is a search bar with the placeholder "Search by industry, business function, or desired business outcome". Underneath the search bar are three dropdown menus: "Industry", "Business Function", and "Business Outcome", followed by an "Explore" button.

aiexplorer.aws.amazon.com





Thank you!

Suji Lee

awsjlee@amazon.com