



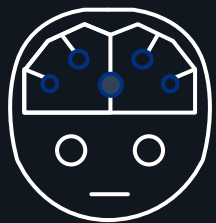
Generative AI with AWS

생성형 AI를 통해 대규모로 혁신하고
디지털 전환을 가속화

Kim, SeongHan

Sr. Industry Service SA
AWS

AI가 어떻게 진화해 왔을까요?



인공 지능 (AI)

컴퓨터가 인간의 지능을 모방할 수 있도록 하는 모든 기술
(예: 로직, If-then 조건, 기계 학습 등)



기계 학습 (ML)

인공 지능(AI)의 하위 집합으로, 기계를 사용하여 데이터에서 패턴을 찾아내 논리 모델을 자동으로 구축



딥 러닝 (DL)

기계 학습(ML)의 하위 집합으로, 음성이나 이미지 인식과 같은 작업을 수행하는 심층 다층 신경망으로 구성



생성형 AI

일반적으로 파운데이션 모델(FM)이라고 하는, 방대한 데이터에 대해 사전 학습된 대규모 모델을 기반으로 함

2023

개념 증명의 해



생성형 AI가 무엇인가요?

안전한가요?

프롬프트 엔지니어가 되려면 어떻게 하나요?

어떻게 모델을 선택 하나요?

어디서부터 시작하죠?



이것이 내 사업에 어떤 의미가 있나요?

파운데이션 모델이 무엇인가요?



어떤 모델을 사용하면 좋을까요?

FM이 뭐죠?

대형 언어 모델(LLM)이 무엇인가요?



2024 제품화의 해

(일부 고객)



프로젝트에서 우선순위를 어찌 정하죠?

어떻게 비용을 낮출 수 있죠?

어떻게 이것을 현실로 만들죠?

어떤 사용자 지정 방법을 사용해야 할까요?



어떻게 확장할 수 있나요?

어떤 모델을 사용하는 게 좋죠?

직접 모델을 훈련시켜야 할까요?

어떻게 리스크를 관리하죠?

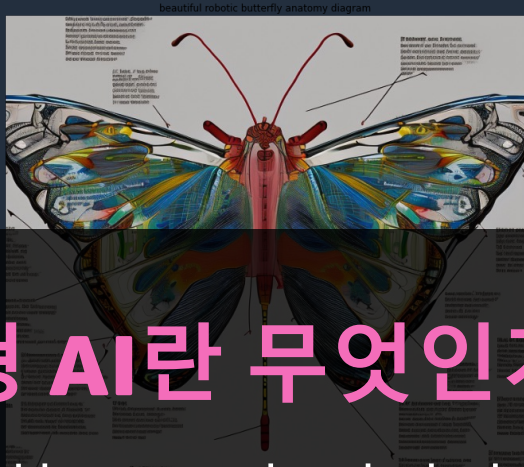


어떻게 빨리 전환할 수 있죠?

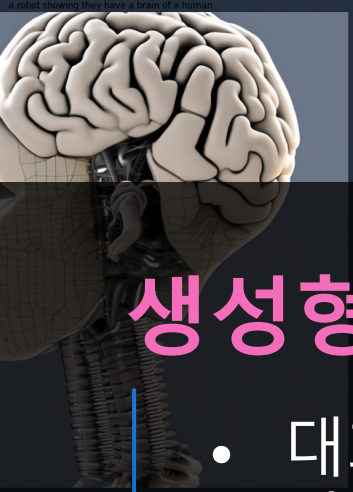
A golden retriever wearing glasses and a hat in a portrait painting



photo of a statue of a robot in university courtyard



beautiful robotic butterfly anatomy diagram

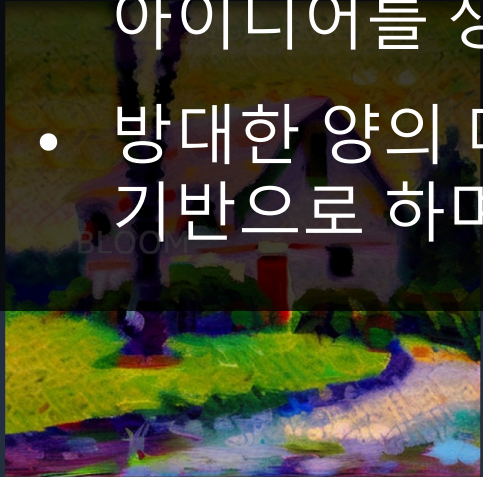


a robot thinking about a brain

생성형 AI란 무엇인가요?

• 대화, 스토리, 이미지, 동영상, 음악 등 새로운 콘텐츠와 아이디어를 생성할 수 있습니다

• 방대한 양의 데이터를 기반으로 하는 사전 훈련된 초대형 모델을 기반으로 하며, 일반적으로 파운데이션 모델(FM)이라고 합니다



생성형 AI는 상당한 사업적 가치를 창출할 잠재력을 가지고 있습니다



새로운 경험

고객 및 직원과 상호 작용할 수 있는 새롭고 혁신적이며 매력적인 방법을 만들 수 있습니다



생산성

모든 사업 부문의 생산성을 획기적으로 개선합니다



인사이트

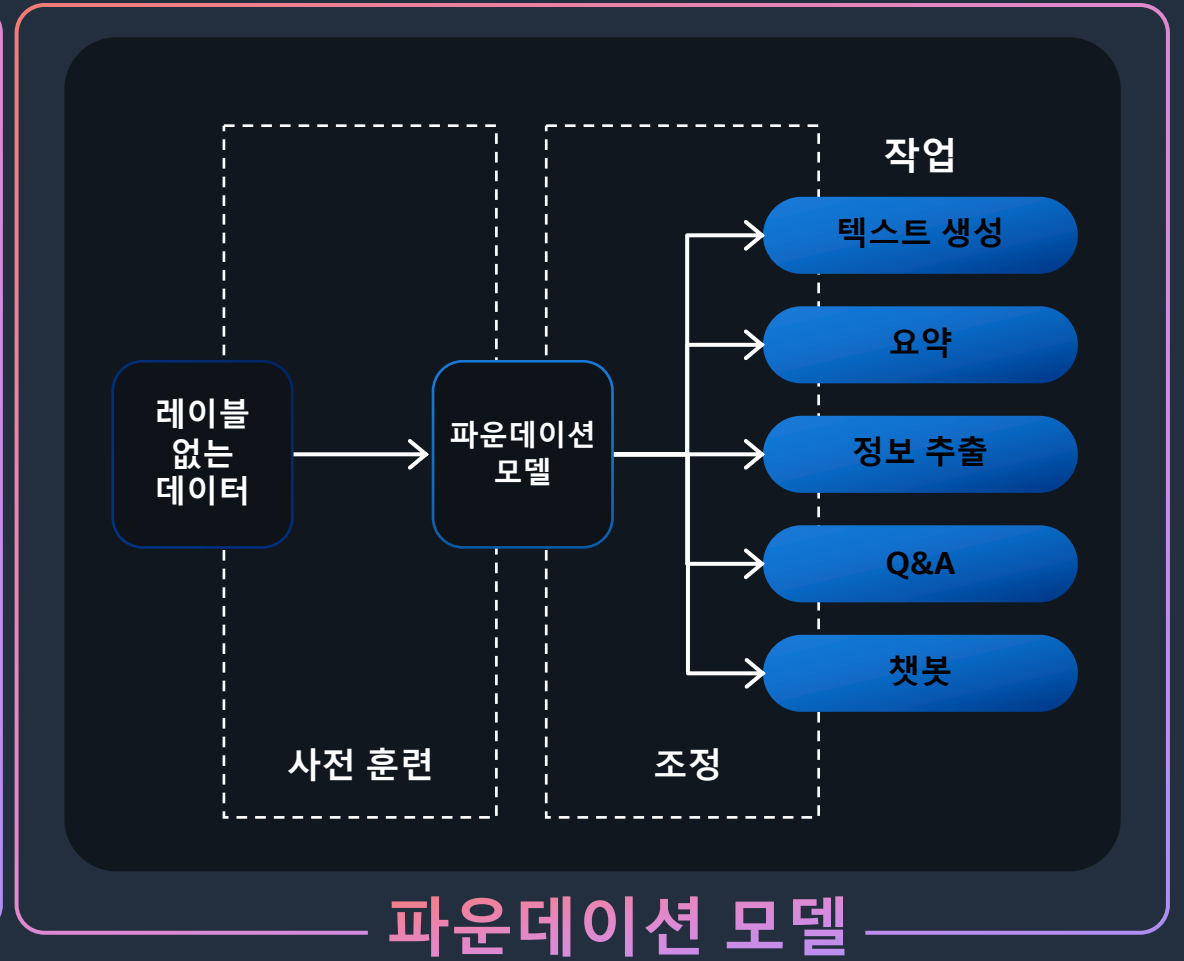
모든 기업 정보에서 인사이트와 명확한 답변을 추출하여 더 빠르고 더 나은 의사 결정을 내릴 수 있게 도와줍니다



창의성

대화, 스토리, 이미지, 동영상, 음악 등 새로운 콘텐츠와 아이디어를 만들어 낼 수 있습니다

파운데이션 모델과 다른 기계 학습(ML) 모델의 차이점



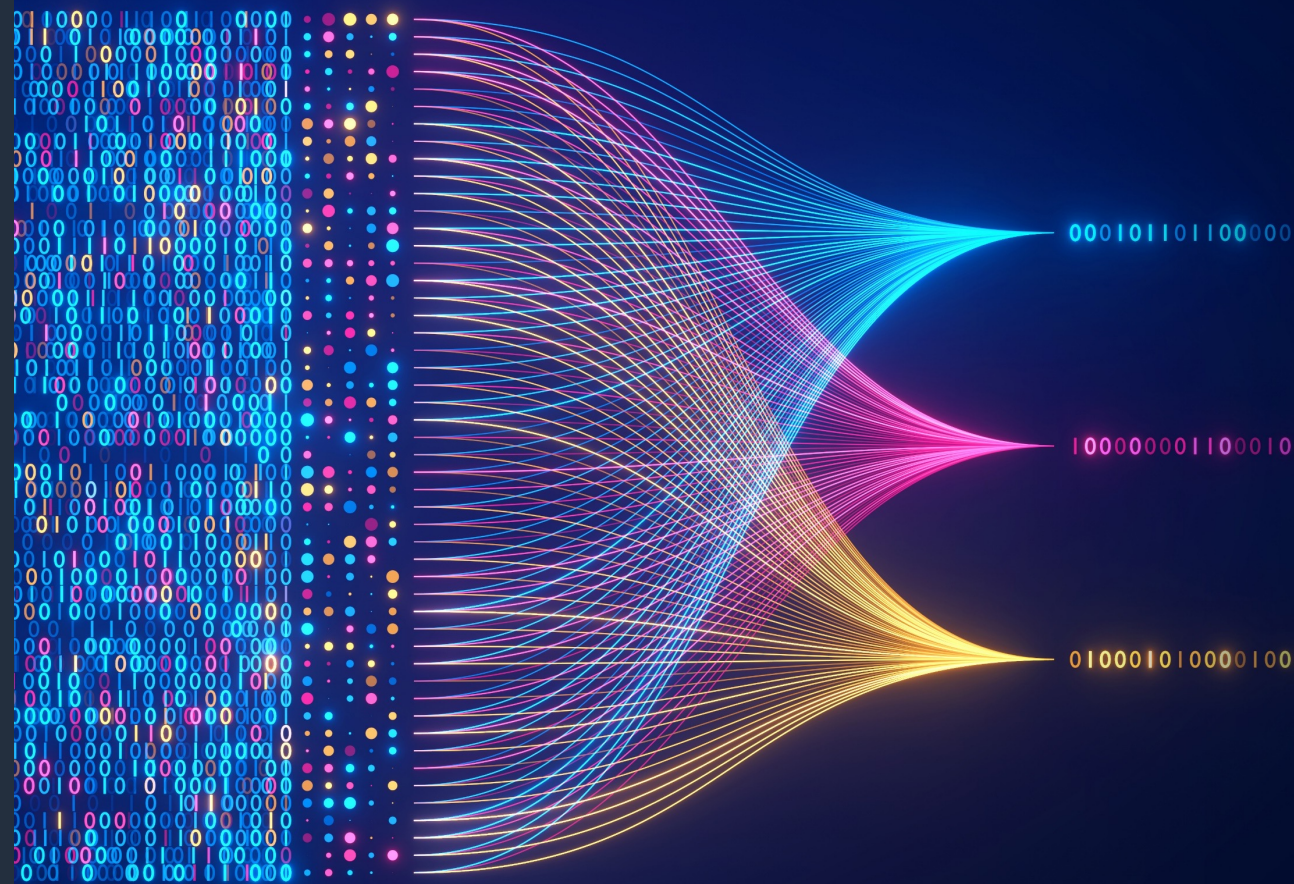
파운데이션 모델을 기반으로 구동되는 생성형 AI

방대한 양의 비정형 데이터에 대한 사전 훈련

복잡한 개념을 훈련할 수 있는 많은 파라미터 포함

다양한 상황에 적용 가능

각 분야별 작업에 데이터를 사용하여 FM 사용자 지정



업스테이지, Amazon SageMaker를 활용해 LLM 제품을 개발



CHALLENGE

LLM 학습과 서빙 시 많은 GPU 장비가 필요한데 자체 IDC 서버로 투자하기엔 비용적 부담 해결이 필요했습니다.

또한 기계 학습을 위한 다양한 목적별 기능 세트를 함께 활용해 고품질의 기계 학습 모델을 빠르게 훈련하고자 하는 니즈가 있었습니다.

SOLUTION

완전 관리형 인프라 및 워크플로우 뿐만 아니라 모든 사용 사례에 적합한 기계 학습 모델을 구축, 훈련, 배포할 수 있는 Amazon SageMaker 를 업스테이지 LLM 제품인 Solar의 pre-training 에 활용했습니다.

OUTCOME

- ✓ 수요가 많아 구하기 어려운 GPU 장비들을 확보할 수 있도록 적극적으로 도움 받아 pre-training 에 집중
- ✓ Amazon SageMaker로 100% pre-training 한 생성형 AI 모델 Solar가 허깅 페이스에서 운영하는 '오픈 LLM 리더보드' 평가 점수에서 ChatGPT 기반인 'GPT-3.5' 성능을 뛰어넘어 1위를 차지

글로벌 EdTech 기업들은 이미 생성형 AI로 혁신 중입니다

학생 경험 향상

 saasyan

BriBooks

 CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

 netex

창의성 및
생산성 향상

Blackboard

Dystch

 eTrivium

비즈니스 성과
최적화

 Pearson

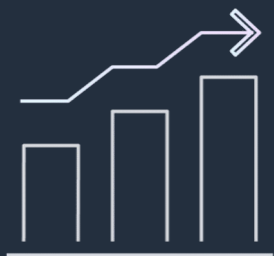
upstage 

LLM 사용 사례



고객 경험 개선

- 챗봇
- 통화 분석
- 에이전트 지원



직원 생산성 향상

- 대화형 지원
- 코드 생성
- 자동화된 보고서 생성



창의성 및 콘텐츠 제작 향상

- 마케팅
- 영업
- 제품 개발
- 미디어 및 엔터테인먼트
- 뉴스 생성



프로세스 최적화 가속

- 문서 프로세싱
- 사기 탐지
- 공급망 최적화

생성형 AI가 모든 것을 해결해 줄 수 있을까요?



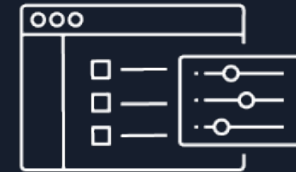
아직 빠르게 발전 중

생성형 AI는 아직 초기 기술 단계이고, 계속 발전 중인 기술입니다.
100% 완벽하지 않습니다.



파운데이션 모델의 한계

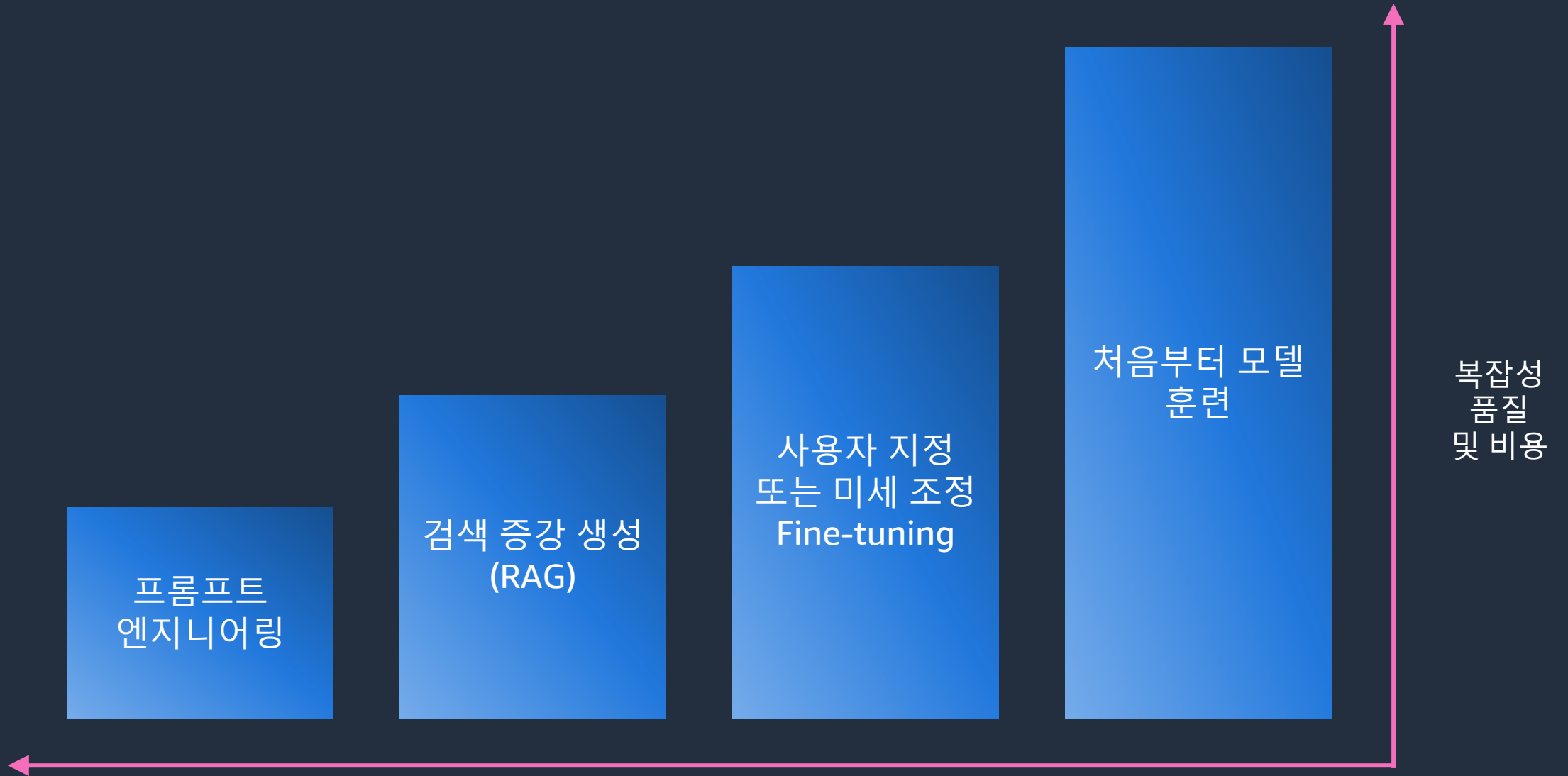
일반적인 지식으로 학습된 확률 모델이고, **최신 지식 또는 기업 고유 지식으로 학습되지 않아** 예기치 않은 응답을 할 수도 있습니다.



제품화 고려사항

제품은 **기업의 데이터(지식)으로 특화**되어야 하고, 이를 유기적으로 결합하기 위해 **보안, 거버넌스, 프로비저닝 등 여러 요구사항을 고려**해야 합니다.

파운데이션 모델을 적용하는 일반적인 접근 방식



프롬프트 엔지니어링

프롬프트 엔지니어링 예제

What is $10 + 10$?

$10 + 10 = 20$

$1 + 1$ is an addition problem.
 $1 - 1$ is a subtraction problem.
 1×1 is a multiplication problem.
 $1 / 1$ is a division problem.

What is $10 + 10$?

$10 + 10$ is an addition problem

프롬프트 엔지니어링 - 간단한 정의



- **프롬프트**란 무엇인가요?
 - ✓ AI 시스템에 응답을 유도하는 텍스트 입력 값
- **프롬프트 엔지니어링**이란 무엇인가요?
 - ✓ NLP(자연어 처리) 기술을 이용하여 파운데이션 모델(FM)/거대 언어 모델(LLM)이 원하는 답변을 하도록 프롬프트를 만드는 것
- 왜 중요한가요?
 - ✓ 모델의 행동에 대한 정교하고 전략적인 통제 가능
 - ✓ 원하는 성능을 달성
 - ✓ 위험을 완화



프롬프트의 요소



1

지침



2

맥락



3

입력 데이터



4

출력 예시

프롬프트 구현

XML 태그, 예시 사용 ←

데이터와 지시 사항 분리 ←

역할 주기 ←

아래의 <example>안에는 <information>과 <user>, <review>와 <result>가 있습니다. <information>은 사용자에게 추천할 제품의 정보이며, <user>는 상품 구매자의 정보를 담고 있습니다. <result>는 <review>를 분석한 감정인 Emotion과, 사장님의 입장에서 작성한 답변인 "Generated"가 json 형식으로 나옵니다.

사용자의 Emotion은 NEGATIVE, POSITIVE, MIXED로 표현됩니다. 분석한 Emotion이 POSITIVE, MIXED일 경우에는 <Information>을 반영해 "Generated"를 작성하여 다른 제품을 추천합니다. Emotion이 NEGATIVE일 경우에는 제품 추천을 하지 않고 정중하게 사장님의 입장에서 사과문을 작성합니다.

고객 데이터를 활용한 소통

- 고객 정보를 응대에 이용

개인화된 추천 자연스럽게 제공

- 판매 데이터 이용,
고객에 맞는 제품 추천

사람 간의 의사소통 모방

- 감정 분석 및 그 결과에 따른 응대

프롬프트

Prompt Engineering Techniques

Zero-shot

Instruction
and/or Question



LLM



Output

Few-shot

Instruction
and/or
Question

Examples
(input + output)



LLM



Output

Chain of Thoughts (CoT)

Instruction
and/or
Question

Reasoning
example
+
CoT
indicator



LLM



Output

Reasoning & Acting (ReAct)

Instruction
and/or
Question

Tools
Description



LLM

Action



Agent

Response



LLM



Output

Self Consistency

Instruction
and/or
Question

Reasoning
example
+
CoT
indicator



LLM

Reasoning Outputs



Aggregation



Output

Among others like:

- Generate Knowledge Prompting
- Automatic Prompt Engineer
- Active-Prompt
- Directional Stimulus Prompting
- Augmented Language Models
- Multi-Modal CoT
- Graph Prompting

Plus all the possible combinations, e.g.:

- Zero-shot + CoT
- Few-shot + CoT
- Few-shot + ReAct
- etc.

Plus other factors like:

- Aggregations
- Chaining
- etc.

검색 증강 생성 (RAG)

LLM의 환각(hallucination) 현상

조선왕조실록에 기록된 세종대왕의 맥북프로 던짐 사건에 대해 알려줘



세종대왕의 맥북프로 던짐 사건은 조선왕조실록에 기록된 사건 중 하나입니다. 세종대왕은 조선시대의 왕으로서 과학, 문학, 예술 등 다양한 분야에서 업적을 남긴 왕이었습니다.

조선왕조실록은 조선왕조의 업적과 사건들을 기록한 역사 서적으로, 세종대왕의 통치 기간 동안의 사건들을 상세하게 기록하고 있습니다. 이 중에서 세종대왕의 맥북프로 던짐 사건은 1443년 5월 5일에 일어났습니다.

해당 사건은 세종대왕이 맥북프로를 사용하다가 문제가 생겨 화가 나서 맥북프로를 던져버린 것으로 전해지고 있습니다. 이 사건은 세종대왕의 고집스러움과 열정, 그리고 인간적인 면모를 보여주는 에피소드로 유명합니다.

세종대왕은 과학 기술에 관심이 많았고, 이를 바탕으로 한글을 창제하였으며, 천문학, 의학, 지리학 등 다양한 분야에서도 업적을 남겼습니다. 그러나 그의 과학 기술에 대한 열정은 때로는 과도한 행동으로 이어질 때도 있었던 것으로 보입니다.

세종대왕의 맥북프로 던짐 사건은 세종대왕의 인간적인 모습을 엿볼 수 있는 사건 중 하나로, 그의 업적과 함께 전해지고 있습니다.

LLM의 지식단절 (knowledge cutoff) 문제

SU

이마트 경기광주점 영업시간은 언제까지야?



죄송하지만, 제가 가지고 있는 정보는 2021년 9월까지의 것이며, 현재의 영업 시간 정보를 제공할 수 없습니다. 이마트 경기광주점의 영업 시간은 시기와 요일에 따라 다를 수 있으며, 휴일에도 다를 수 있습니다. 따라서 가장 정확한 정보를 얻으려면 이마트 경기광주점의 공식 웹사이트를 확인하거나, 직접 전화로 해당 점포에 문의하시는 것이 좋습니다. 그렇게 하면 최신 정보를 확인하실 수 있을 것입니다.

SU

휴무일은 언제인데?



죄송하지만, 제가 가지고 있는 정보는 2021년 9월 이전의 정보이기 때문에 현재 이마트 경기광주점의 정확한 휴무일을 알 수 없습니다. 휴무일은 지역과 계점에 따라 다를 수 있으며, 또한 연휴나 특별한 일정에 따라 변경될 수 있습니다. 따라서 가장 정확한 휴무일 정보를 확인하려면 이마트 경기광주점의 공식 웹사이트나 고객 서비스에 문의하시는 것이 좋습니다. 그렇게 하면 최신 정보를 얻을 수 있을 것입니다.



생성형 AI 어플리케이션

생성형 AI 어플리케이션

스토리지

구조화 및 비구조화 데이터

운영 데이터베이스

SQL, NoSQL, 도큐먼트, 그래프, 벡터

분석 및 데이터 레이크

검색, 스트리밍, 배치, 대화형

데이터 통합

캡처, 변환, 스트리밍

데이터 거버넌스

카탈로그, 품질, 프라이버시, 접근 제어

데이터는 생성형 AI에서 차별화 요소가 됩니다



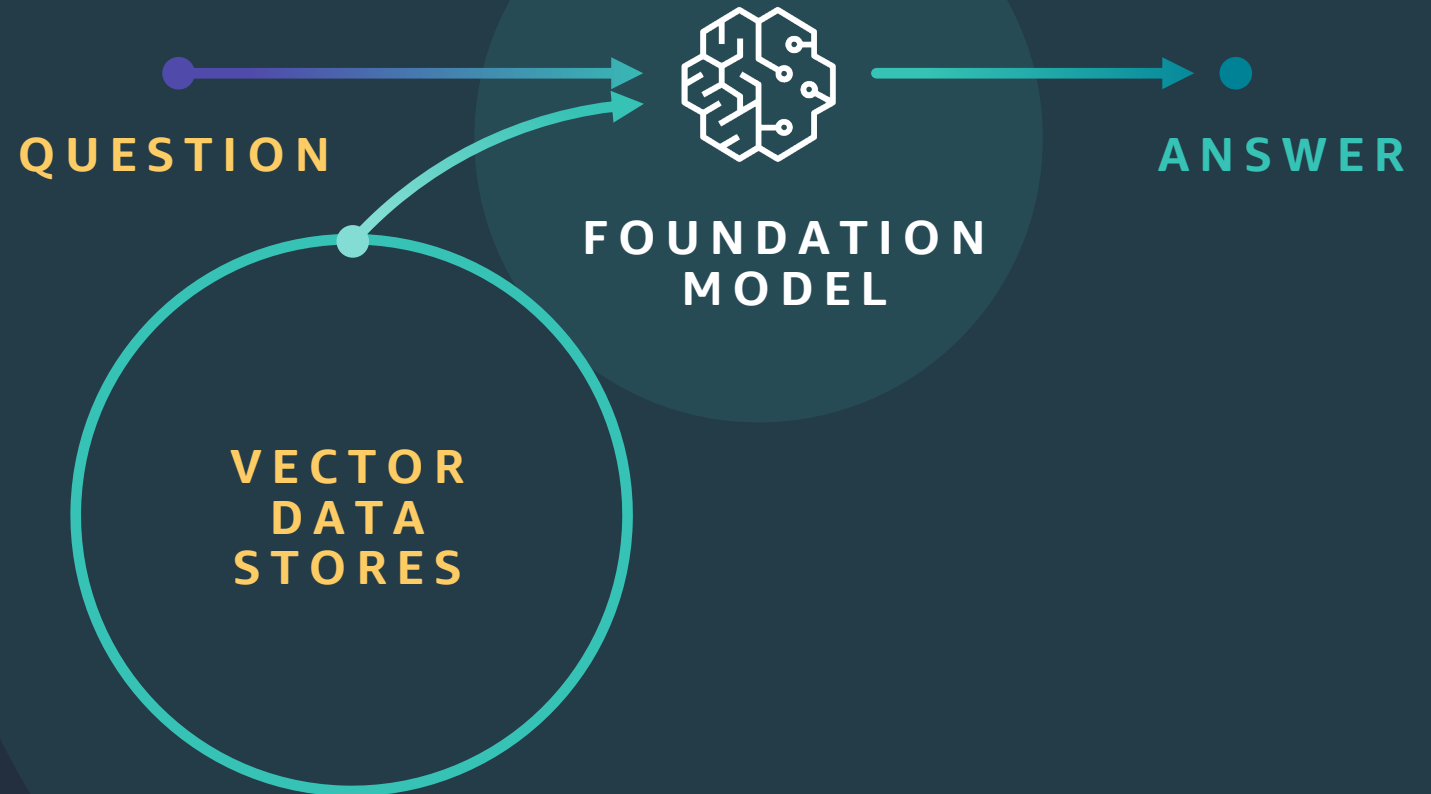
일반적인
생성형 AI



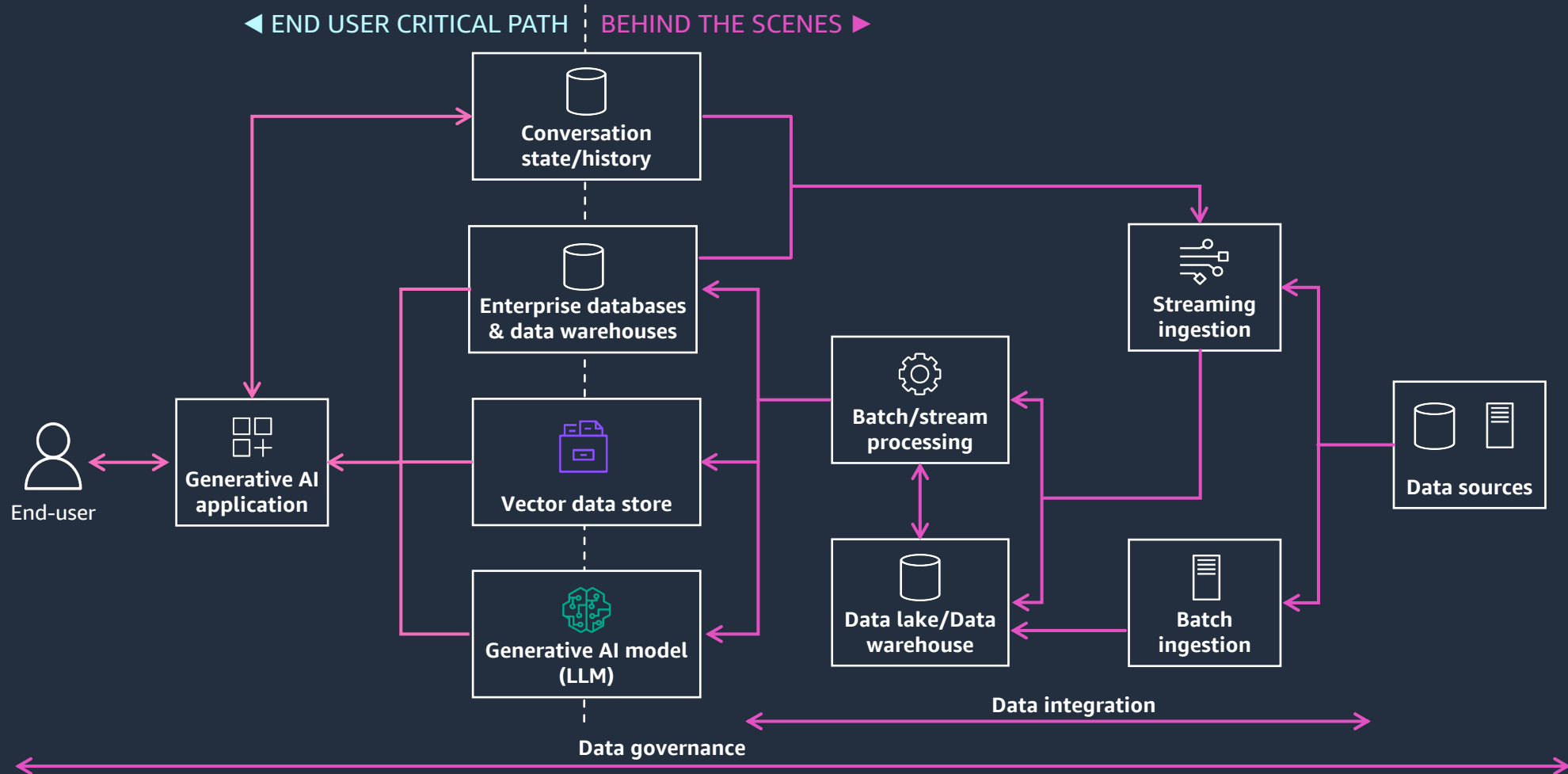
비즈니스와 고객을
이해하는 생성형 AI

Question Answering

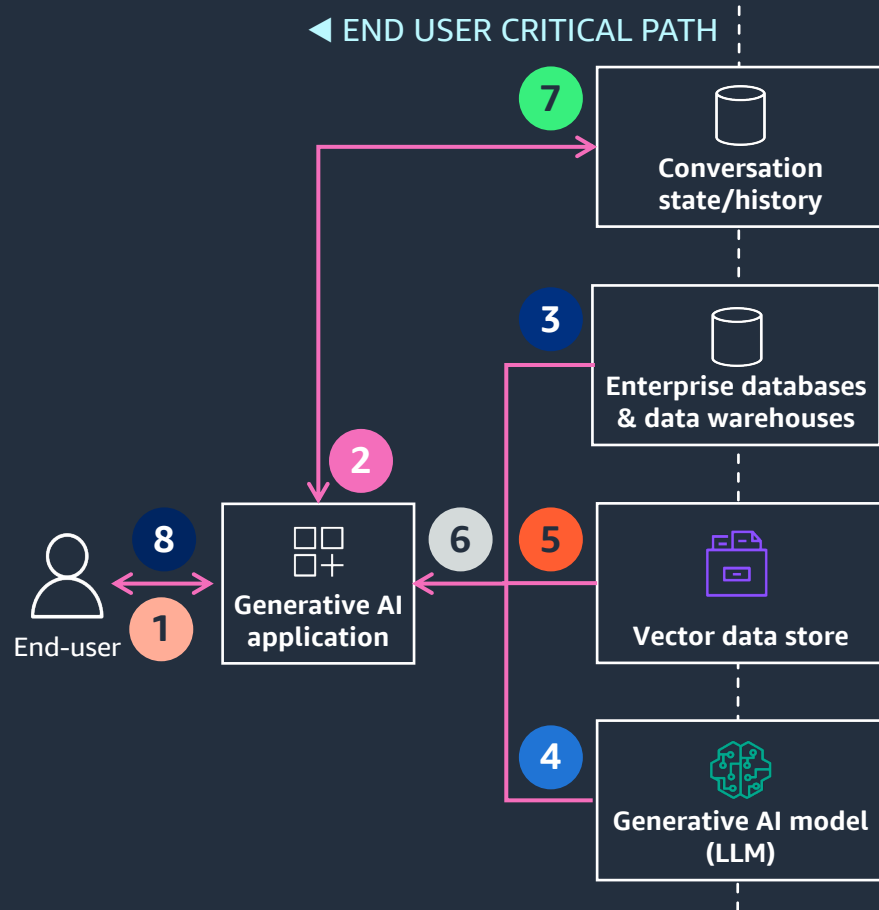
기업 지식 (FAQ, PDF, Doc)



RAG 레퍼런스 아키텍처



최종 사용자 중요 경로에서의 데이터 흐름



- 1 사용자 질문
- 2 대화 상태/이력 가져오기
- 3 상황 별 맥락 파악
- 4 토큰화, LLM으로부터 임베딩 가져오기
- 5 질문 임베딩에 대한 유사도 검색, 가장 많이 일치하는 텍스트를 가져오기
- 6 엔지니어링된 프롬프트로 LLM 호출
- 7 대화 상태/이력 업데이트
- 8 응답 반환

뒤에서 일어나는 데이터 이동

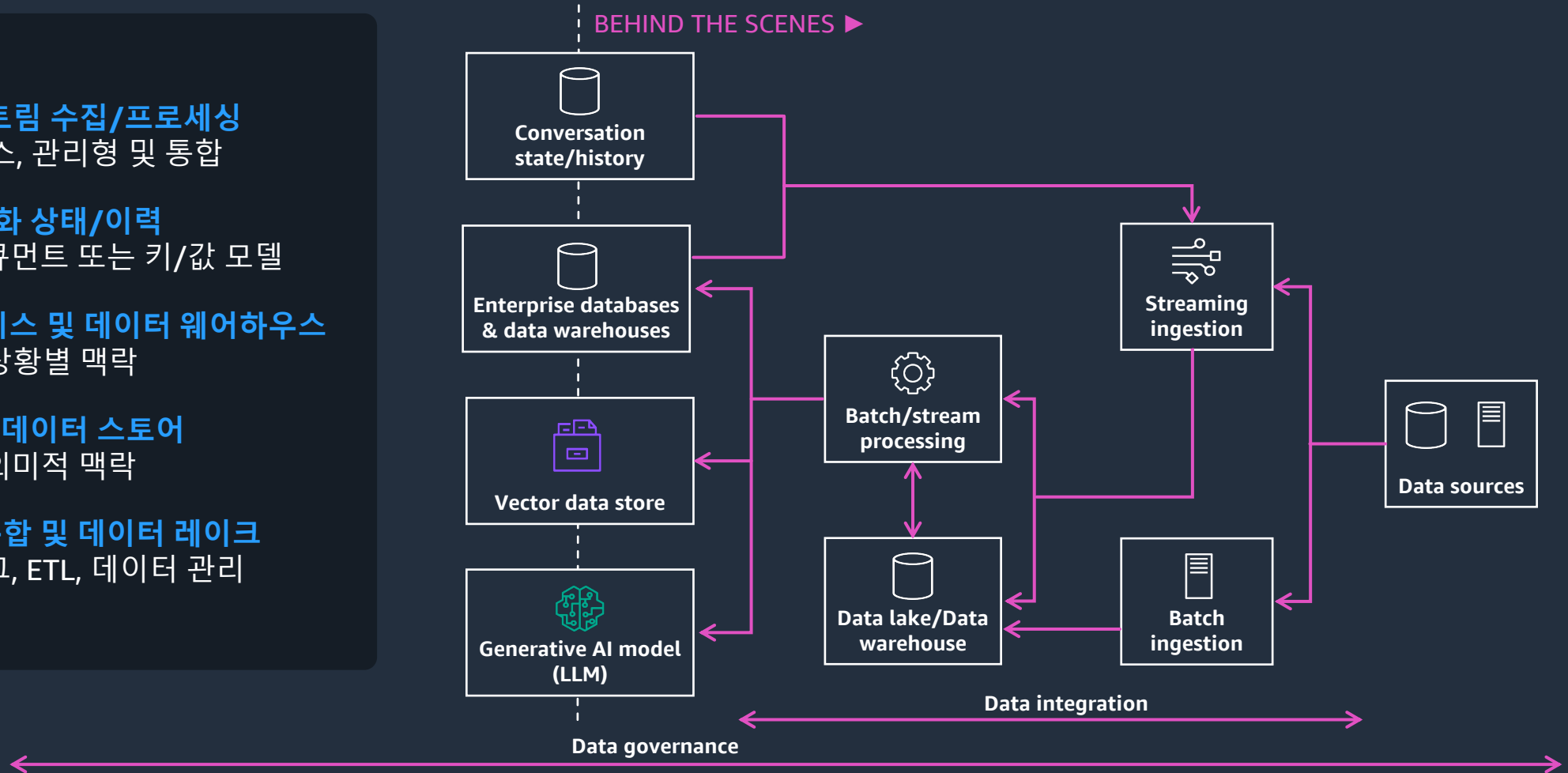
배치/스트림 수집/프로세싱
서버리스, 관리형 및 통합

대화 상태/이력
NoSQL, 문서 또는 키/값 모델

기업 데이터베이스 및 데이터 웨어하우스
상황별 맥락

벡터 데이터 스토어
의미적 맥락

데이터 통합 및 데이터 레이크
카탈로그, ETL, 데이터 관리



Global Alumni, 학생 경험을 개선하기 위해 챗봇 도우미를 구축



CHALLENGE

학생들은 특정 설명이 제공되는 비디오에서 정확한 순간을 찾아내거나, 강의에 문서, 비디오, 오디오 파일과 같은 방대한 자료가 포함된 경우 특정 답을 찾아야 하는 상황에서 어려움을 겪었습니다

SOLUTION

고객은 **Amazon Bedrock** 을 통한 대규모 언어 모델(LLM) 과 검색 증강 생성(RAG) 기술을 이용하여, 학생들이 자연어를 사용하여 빠르게 답을 찾아낼 수 있도록 챗봇을 개발하였습니다

OUTCOME

- ✓ 학생 참여도가 25% 향상
- ✓ 학생 학습 경험이 개선됨



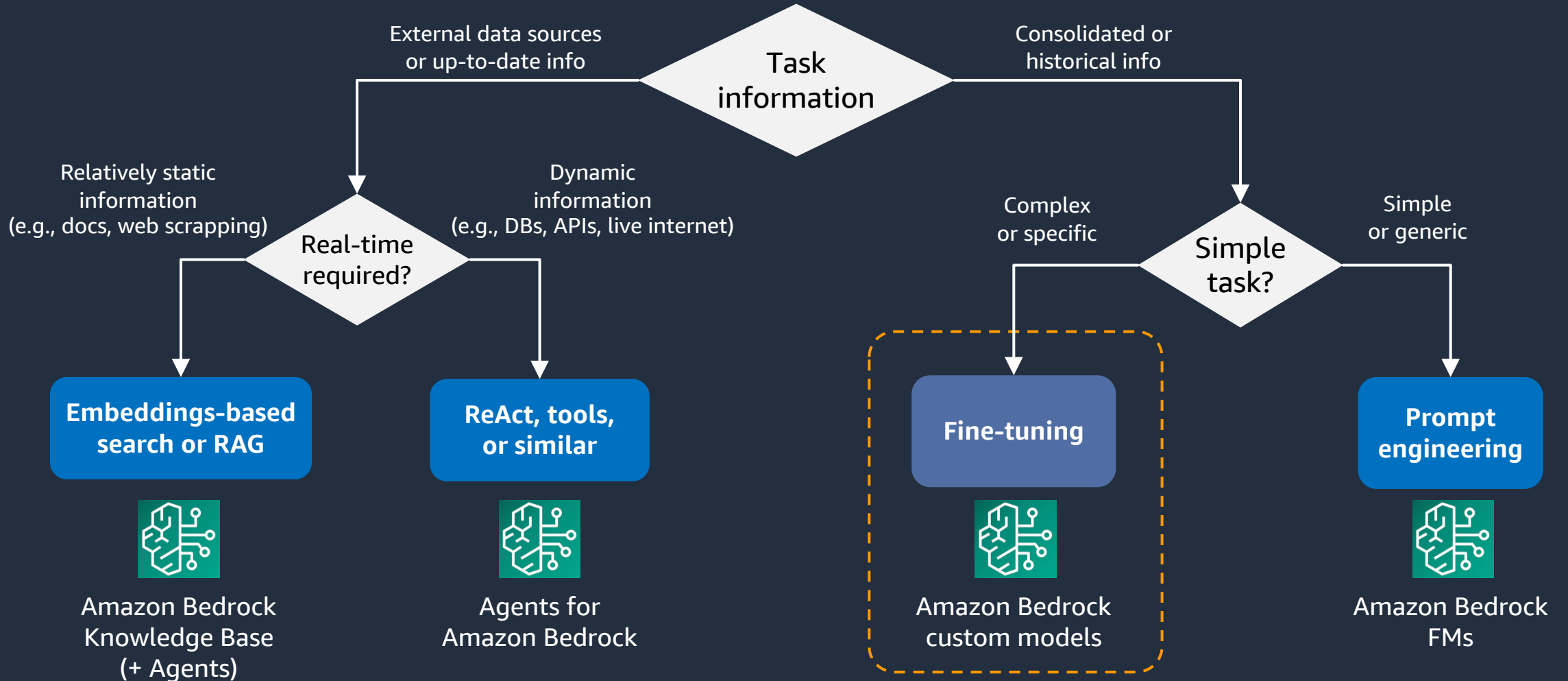
미세 조정 Fine-tuning

언제 Fine-Tuning 이 필요할까요?

- 특정 작업이나 데이터 세트에 대해 기본 모델의 성능이 만족스럽지 않은 경우 (일반적으로 Small Models)
- 특정 스타일이나 어조(예: 격식, 유머, 저널리즘)로 텍스트를 생성
- 유해하거나, 편파적이거나, 오해의 소지가 있는 생성을 피할 경우
- 기본 모델에서 제공하지 않은 의학, 법률 같은 새로운 도메인 및 새로운 언어가 필요한 경우(예: 한국어)
- 과거 정보 보다는 새로운 데이터로 업데이트가 필요한 경우
- 모델이 특정 영역에서 일관된 피드백이나 수정을 받는 경우

When to use fine-tuning

CHOOSING THE RIGHT STRATEGY FOR YOUR TASK



중요! : Fine-tuning은 꼭 필요하지는 않다! 특히나 최근엔..

질문

Summarize the chat dialogue:
Avery: You went to Ethan's house?
David: yeah I had to babysit
Avery: Aww, how do you babysit, just curious
David: I had to go through a lot :/
Avery: Was his sister naughty
David: Tooo much
Avery: Lol
David: I will just refuse net time :/
Avery: As you wish
David: :/
Avery: I just got his text
David: What is he saying
Avery: He is asking me to say thanks to you
David: yeah whatever<3
Avery: He was saying that your phone was switched off
David: Yeah i have just turned it on
Avery: I have told him about that
David: k
Avery: Gotta go now.
Please summarize it.

답변 (llama2-13b-chat)

The chat dialogue is about Avery going to Ethan's house to babysit his sister. David is just there to help him with his homework. They are talking about how they go through a lot of babysitting tasks, and Avery is asking David to say thank you to him.

- LLM 성능이 빠르게 발전하고 있음.
- Fine-tuning 보다는 기본적인 프롬프트 튜닝, 그리고 RAG 와 같은 아키텍처 개선 방법을 먼저 고려할 필요가 있음.

The Generative AI Journey

	사전 훈련 <small>Pretraining</small>	파인 튜닝 <small>Fine tuning</small>	프롬프트 엔지니어링 <small>Prompt engineering</small>
훈련 시간 (및 비용)	모델 및 인프라에 따라 며칠에서 몇 달까지 소요	대개 몇 분에서 몇 시간	없음
커스터마이징	모델 전체 <ul style="list-style-type: none"> 신경망 아키텍처 어휘 크기 컨텍스트 길이 훈련 데이터 	모델 일부 <ul style="list-style-type: none"> 특정 작업 튜닝 도메인별 훈련 데이터 추가 	프롬프트 <ul style="list-style-type: none"> 모델 커스터마이징 없음 프롬프트 커스터마이징 및 최적화
요구되는 ML 전문성	높음	중간	낮음

왜 생성형 AI를 AWS에서 사용해야 할까요?

Generative AI Stack

APPLICATIONS THAT LEVERAGE LLMs AND OTHER FMs



Amazon Q
Business



Amazon Q
Developer



Amazon Q in
QuickSight



Amazon Q in
Connect

TOOLS TO BUILD WITH LLMs AND OTHER FMs



Amazon Bedrock

Guardrails | Agents | Studio | Customization Capabilities | Custom Model Import

INFRASTRUCTURE FOR FM TRAINING AND INFERENCE



GPUs



Trainium



Inferentia



SageMaker



UltraClusters



EFA



EC2 Capacity Blocks



Nitro



Neuron

Generative AI Stack

APPLICATIONS THAT LEVERAGE LLMs AND OTHER FMs

TOOLS TO BUILD WITH LLMs AND OTHER FMs

INFRASTRUCTURE FOR FM TRAINING AND INFERENCE



GPUs



Trainium



Inferentia



SageMaker



UltraClusters



EFA



EC2 Capacity Blocks



Nitro



Neuron



Amazon SageMaker

파운데이션 모델을 포함하여
대규모 머신 러닝 모델 구축, 학습
및 배포

공개적으로 사용 가능한 최신 파운데이션
모델에 액세스

처음부터 파운데이션 모델 구축

파운데이션 모델 사용자 지정

추론 수행

파운데이션 모델 운영(FMOps) 및
거버넌스 구현

Generative AI Stack

APPLICATIONS THAT LEVERAGE LLMs AND OTHER FMs

TOOLS TO BUILD WITH LLMs AND OTHER FMs

 **Amazon Bedrock**

Guardrails | Agents | Studio | Customization Capabilities | Custom Model Import

INFRASTRUCTURE FOR FM TRAINING AND INFERENCE

 GPUs  Trainium  Inferentia  SageMaker

 UltraClusters  EFA  EC2 Capacity Blocks  Nitro  Neuron

Amazon Bedrock

파운데이션 모델(FM)을 사용하여
생성형 AI 어플리케이션을
구축하고 확장하는 가장 쉬운 방법

단일 API를 통한 주요 파운데이션 모델
선택

모델 사용자화

검색 증강 생성(RAG) 기능

다단계 작업을 수행할 수 있는 에이전트

보안, 프라이버시 및 안전함

Amazon Bedrock

데이터를 안전하게
보호하고 비공개로
유지



고객 데이터는 기본 모델 학습에
사용되지 않습니다

모든 데이터는 전송 및 저장 시 암호화
되며, 사용자 지정에 사용되는
데이터는 고객의 VPC 안에서 안전하게
전송됩니다

데이터는 API가 처리되는 리전에
유지됩니다

GDPR, SOC, ISO, CSA 규정 준수 및
HIPAA 적격 지원

Generative AI Stack

APPLICATIONS THAT LEVERAGE LLMs AND OTHER FMs



Amazon Q
Business



Amazon Q
Developer



Amazon Q in
QuickSight



Amazon Q in
Connect

TOOLS TO BUILD WITH LLMs AND OTHER FMs



Amazon Bedrock

Guardrails | Agents | Studio | Customization Capabilities | Custom Model Import

INFRASTRUCTURE FOR FM TRAINING AND INFERENCE



GPUs



Trainium



Inferentia



SageMaker



UltraClusters



EFA



EC2 Capacity Blocks



Nitro



Neuron

Amazon Q

AWS의 생성형 AI 기반
어시스턴트로 업무 혁신

데이터 소스, 저장소 및 기업 시스템 통합

업무가 진행되는 모든 곳에서 사용 가능

업계 최고의 지원 제공



Thank you!