오픈소스와 생성형 AI:
엔터프라이즈 환경에서의 적용 가능성과 전망

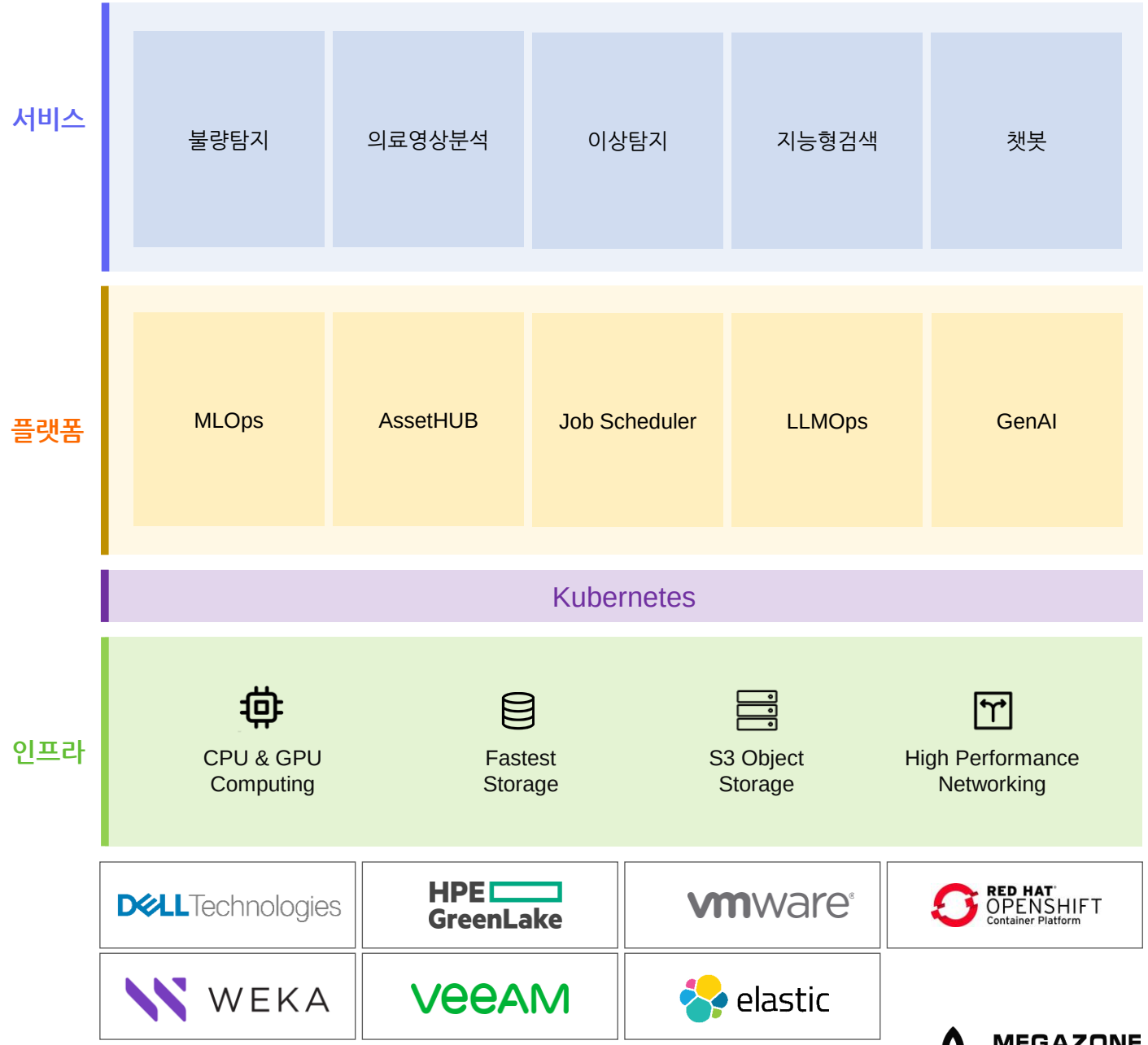
메가존클라우드 Hybrid & AI Platform Center | 강민주 센터장

2024. 09

MEGAZONE CLOUD

메가존클라우드

Hybrid & AI Platform Center



메가존클라우드의 AI 전문 연구소 Matilda Lab

Matilda Lab 소개

Digital Transformation 의 가속화를 위한 머신 인텔리전스, 인공지능, 데이터 분석 솔루션, 데이터 모델링, 소프트웨어 정의 데이터 센터 기반의 하이브리드 클라우드, 슈퍼 컴퓨팅 등의 융 복합 기술의 서비스화를 위해 연구하는 AI 전문가 그룹입니다.

AI 엔지니어 & 컨설턴트

머신러닝 컨설팅 &
사용자 모델 최적화

다양한 경험을 바탕으로
현업에서 바로 적용할 수 있는
AI 가이드를 만들고 문제를 해결

K8S 클러스터 전문가

대용량 AI컴퓨팅 쿠버네티스
클러스터 구축 및 운영

AI워크로드를 최고의 성능으로
처리할 수 있는 대용량 컴퓨팅
인프라와 운영기술을 제공

E2E AI 통합 플랫폼

AI 워크로드 통합
마틸다 플랫폼 개발

고도화된 산업에서 AI 프로세스를
자동화하고 자원 활용률을
최대화한 통합 플랫폼

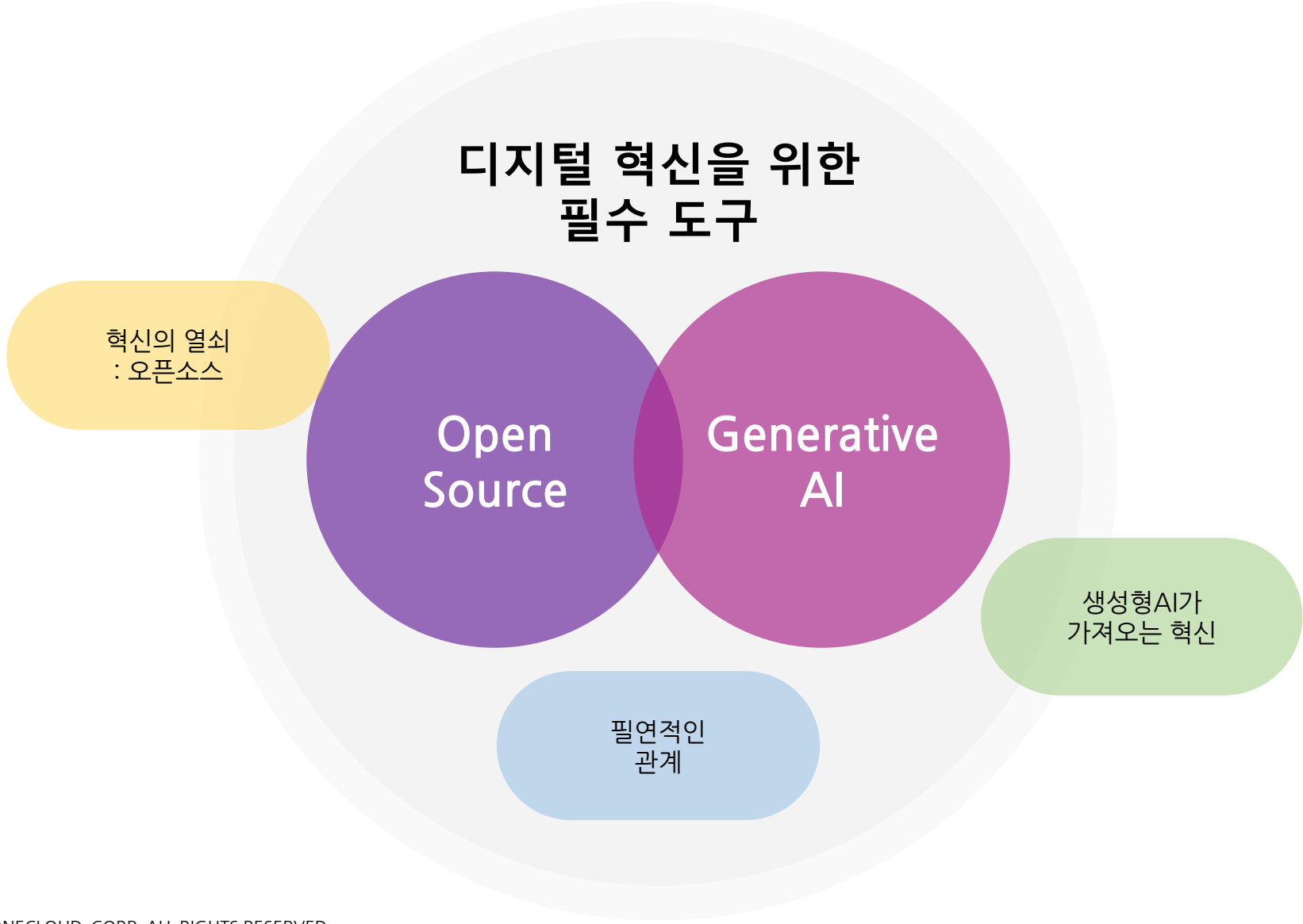
비즈니스 파트너 프로그램

특수 목적 AI 모델 서비스(SaaS)
공동 개발 및 전환

Trained Model을 활용한Inferencing
전용 머신 개발
(Eco Partner 공동개발)

MEGAZONE CLOUD

왜 오픈소스와 생성형 AI인가?



MEGAZONE CLOUD

오픈소스 이야기

1983년	1991년	1995년	2002년	2008년	2024년
					
GNU Project	Linux	Apache	Firefox	Github	Llama 3.1

OS

WEB / WAS

RDMBS / NoSQL

Search

Map

Bigdata

IDE

LLM

MEGAZONE CLOUD

History of Large Language Model: Generative AI

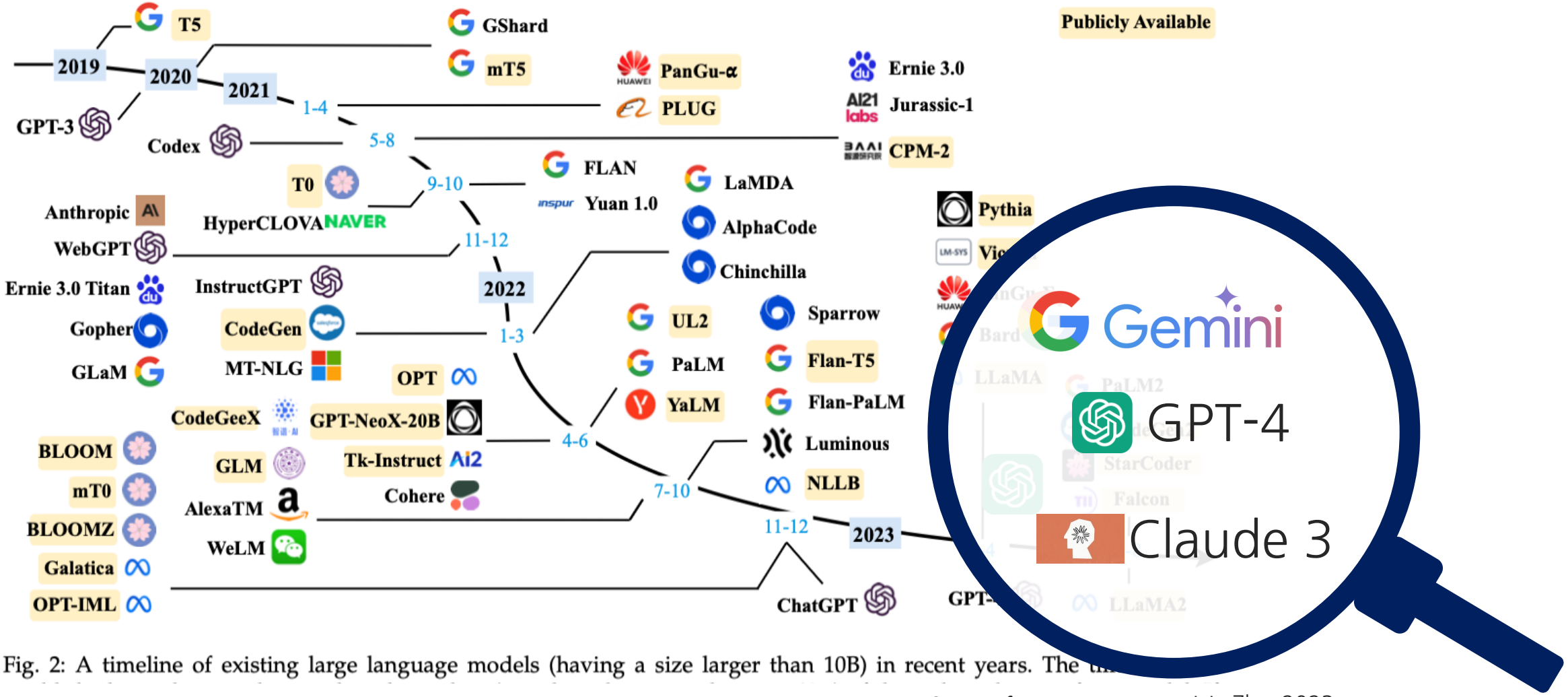
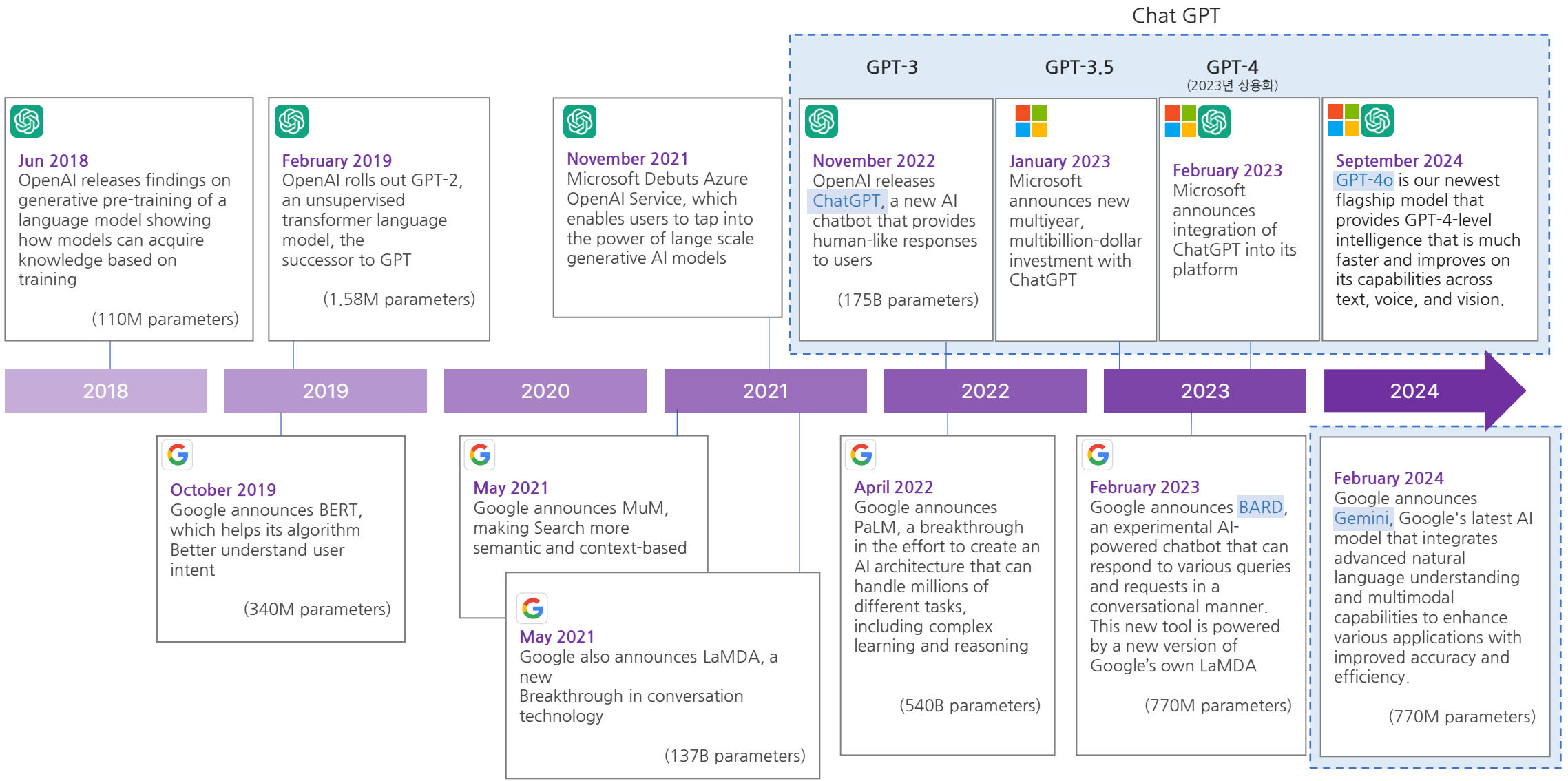


Fig. 2: A timeline of existing large language models (having a size larger than 10B) in recent years. The

A Survey of Large Language Models. Zha. 2023.

MEGAZONE CLOUD

Generative AI - History of Large Language Model

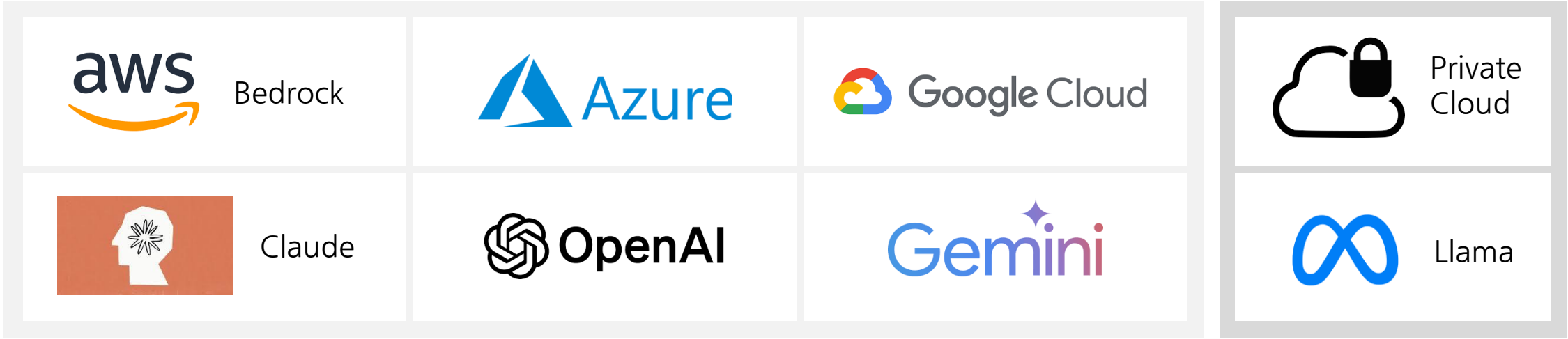


MEGAZONE CLOUD

일반 사용자들의 생성형AI 활용

MEGAZONE CLOUD

기업에서의 생성형AI 활용



1. 대고객 챗봇
2. 법률 검토
3. 코드 생성
4. 보안
5. 업무매뉴얼 챗봇
6. 계약서 검토
7. 문서 도우미

MEGAZONE CLOUD

Large Language Model: Toward the Private and Closed AI

AI 지속적 발전의 원동력이었던 “전체 공개”가 아닌, Private and closed AI로 변하는 추세

· 수익 창출이 가능한 뚜렷한 비즈니스 모델의 부재
· 치열한 시장 경쟁
· LLM 분야의 지속적인 주도권 확보

- 데이터셋, 학습과정, 모델의 비공개
- Private and closed AI로 변하는 추세

서울신문 구독

‘오픈소스’ 문이 닫힌다… 챗GPT가 촉발한 데이터 전쟁

입력 2023.04.20. 오후 5:27 기사원문

김민석 기자

GPT-4에 와서는 “‘경쟁 환경’과 ‘안정성’을 위해”
소스코드는 물론 모델 크기와 학습한 데이터,
사용한 하드웨어 등 어떤 정보도 공개하지 않고 있다.
유료 계약을 통해 GPT를 사용할 수 있는 API를
부여하고 있을 뿐이다.

MEGAZONE CLOUD

Large Language Model: Toward the Private and Closed AI

AI 지속적 발전의 원동력이었던 “전체 공개”가 아닌, Private and closed AI로 변하는 추세

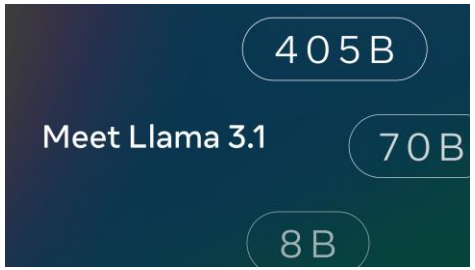
· 수익 창출이 가능한 뚜렷한 비즈니스 모델의 부재
· 치열한 시장 경쟁
· LLM 분야의 지속적인 주도권 확보

- 데이터셋, 학습과정, 모델의 비공개
- Private and closed AI로 변하는 추세

Meta

메타의 현존 가장 강력한 AI 모델, Llama 3.1을 소개합니다.

2024년 7월 23일



지금까지 오픈 소스 대규모 언어 모델은 기능이나 성능 면에서 대부분 폐쇄형 언어 모델에 비해 뒤쳐져 있었습니다. 이제 Meta는 오픈 소스 모델이 이끄는 새로운 시대를 열고자 합니다. Meta는 세계 최대 규모이자 가장 뛰어난 성능의 오픈 소스 파운데이션 모델인 Llama 3.1 405B를 공개 출시합니다. 현재까지 Llama의 다양한 버전 모델들은 총 3억 건 이상 다운로드됐으며, 이는 이제 시작에 불과합니다.

https://about.fb.com/ko/news/2024/07/introducing-llama-3-1-our-most-capable-models-to-date/

MEGAZONE CLOUD

Public Cloud는 Generative AI를 위한 서비스를 확대하고 있습니다.

AWS - 생성형 AI 구축 키트

Amazon Bedrock
FM을 사용하여 생성형 AI 애플리케이션을 구축하고 확장하는 가장 쉬운 방법입니다.

AWS Trainium
동급 Amazon EC2 인스턴스에 비해 교육 비용을 최대 50% 절감하면서 더 빠르게 교육할 수 있습니다.

AWS Inferentia
동급 Amazone EC2 인스턴스보다 최대 40% 낮은 추론당 비용으로 고성능 FM추론을 실행할 수 있습니다.

Amazon CodeWhisperer
애플리케이션을 더 빠르고 안전하게 구축할 수 있도록 도와주는 AI 코딩 도우미로 개인용은 무료로 이용할 수 있습니다.

AWS 기반 Hugging Face
AWS를 기반으로 하는 Hugging Face 모델을 훈련하고 미세하게 조정하며 배포합니다.

Amazon SageMaker
대규모로 FM 을 구축, 교육, 배포하세요.

Amazon Bedrock - API를 통한 FM 완전관리형 서비스

AI21labs

Jurassic-2
질문 답변, 요약, 텍스트 생성 등의 모든 언어 작업에 대한 지침을 따르는 LLM.

cohere

Command and Embed
비즈니스 애플리케이션 및 임베딩 모델을 위한 텍스트 생성 모델로, 100개 이상의 언어로 검색, 클러스터링 또는 분류 지원

stability.ai

Stable Diffusion XL 1.0
헌법적 AI 및 무해 훈련에 기반한 사려 깊은 대화, 콘텐츠 제작, 복잡한 추론, 창의성 및 코딩을 위한 LLM

ANTHROP\C

Claude 3
헌법적 AI 및 무해 훈련에 기반한 사려 깊은 대화, 콘텐츠 제작, 복잡한 추론, 창의성 및 코딩을 위한 LLM

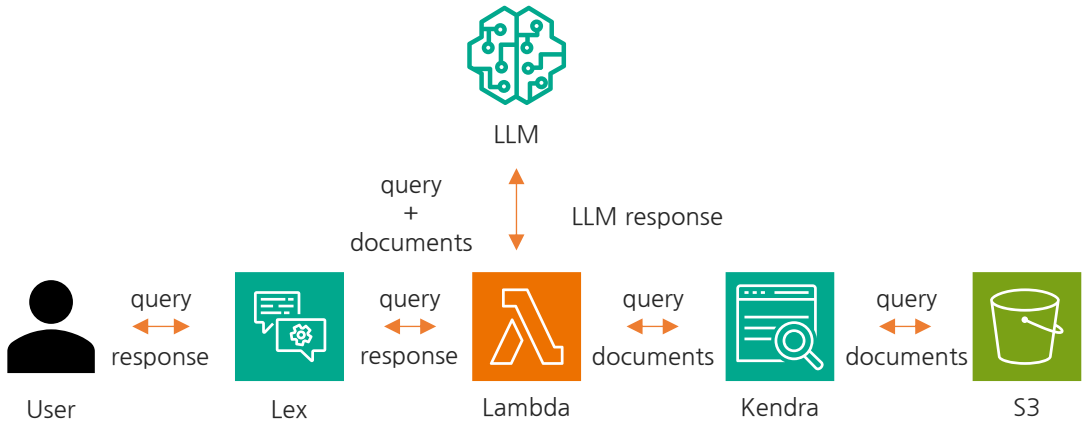
amazon

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

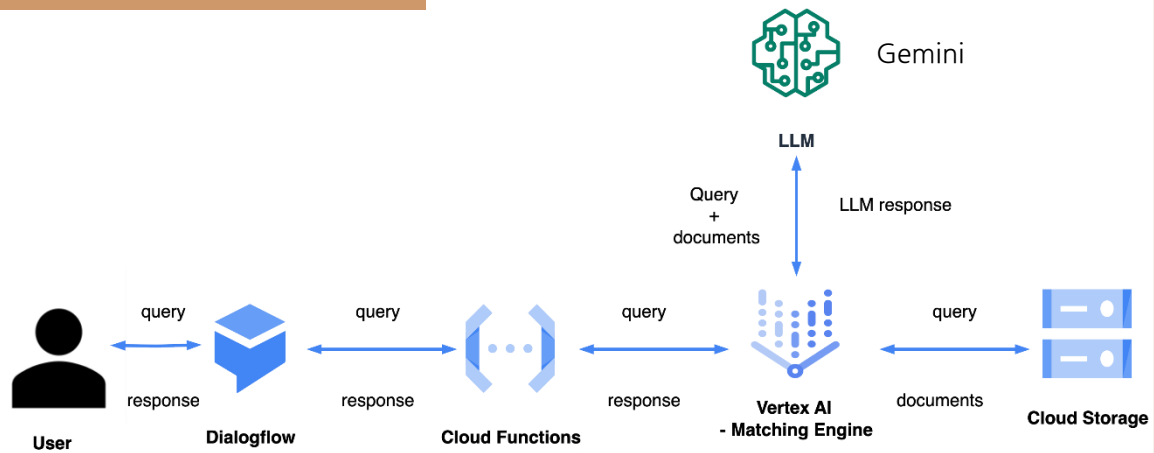
MEGAZONE CLOUD

LLM기반 챗봇 서비스 아키텍처와 Public Cloud Native Service

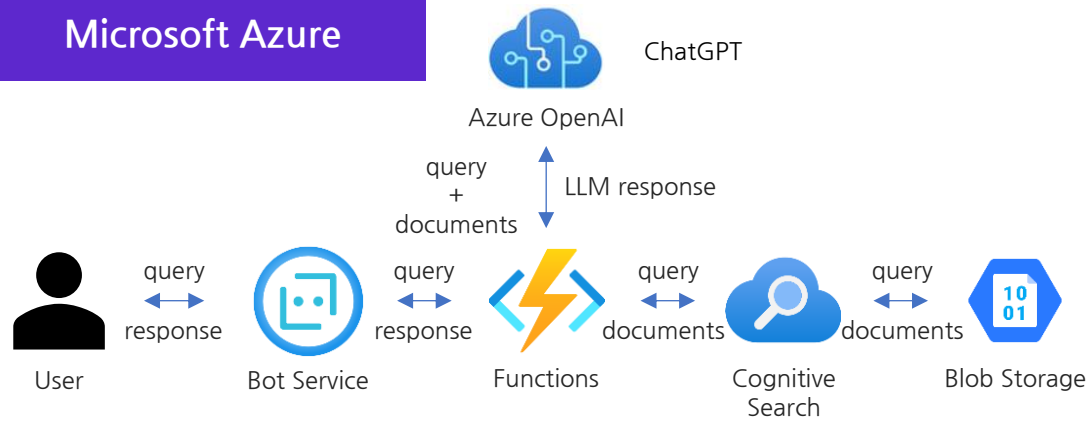
Amazon Web Service



Google Cloud Platform



Microsoft Azure



Private Cloud에서는?

MEGAZONE CLOUD

Generative AI (LLM) Service 이슈 1: 보안 및 신뢰

Public LLM을 사용하려면?

MEGAZONE CLOUD

Matilda: A Hybrid MLOps Platfrom

기업의 상황에 맞는 유연한 LLM 서비스 전략 지원

보안

비용

자원 관리

Private AI

Traditional Model

· 소형 언어 모델 (sLLM)

Public AI

Traditional Model

· 소형 언어 모델 (sLLM)
· 거대언어 모델 (LLM)

Hybrid AI Platform

MEGAZONE CLOUD

엔터프라이즈 생성형AI 도입전략 1: 보안 및 신뢰

- **주어진 참고자료 위주로만 답변을 생성**

  Retrieval-Augmented Generation (RAG)

- **기업 자체 데이터로 fine-tuning**

  보안 및 신뢰 확보한 상태로 정확도 향상

+

LLM

aws Bedrock

Azure OpenAI

Claude

Google Cloud Gemini

MEGAZONE CLOUD

엔터프라이즈 생성형AI 도입전략 2: 비용

오픈소스 LLM을 활용한 기업 자체 Generative AI 모델 개발

Public CSP LLM

OpenAI
aws
Google
LG AI Research
SK telecom
kt
NAVER
kakao
SAMSUNG SDS

오픈소스 sLLM (약 ~70B)

Meta
Llama2
DeepMind
Jurassic-2
Stanford University
Alpaca
Google AI
Megatron-Turing NLG

MEGAZONE CLOUD

엔터프라이즈 생성형AI 도입전략 3: 운영 및 관리

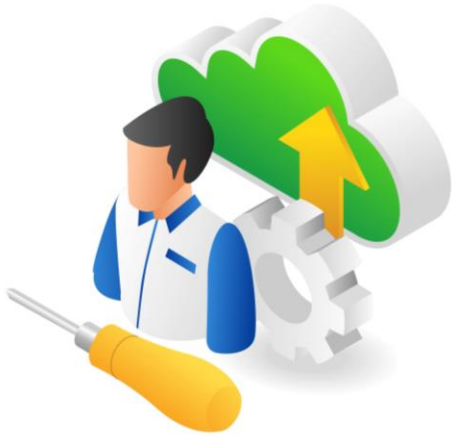
Generative AI

인프라

플랫폼

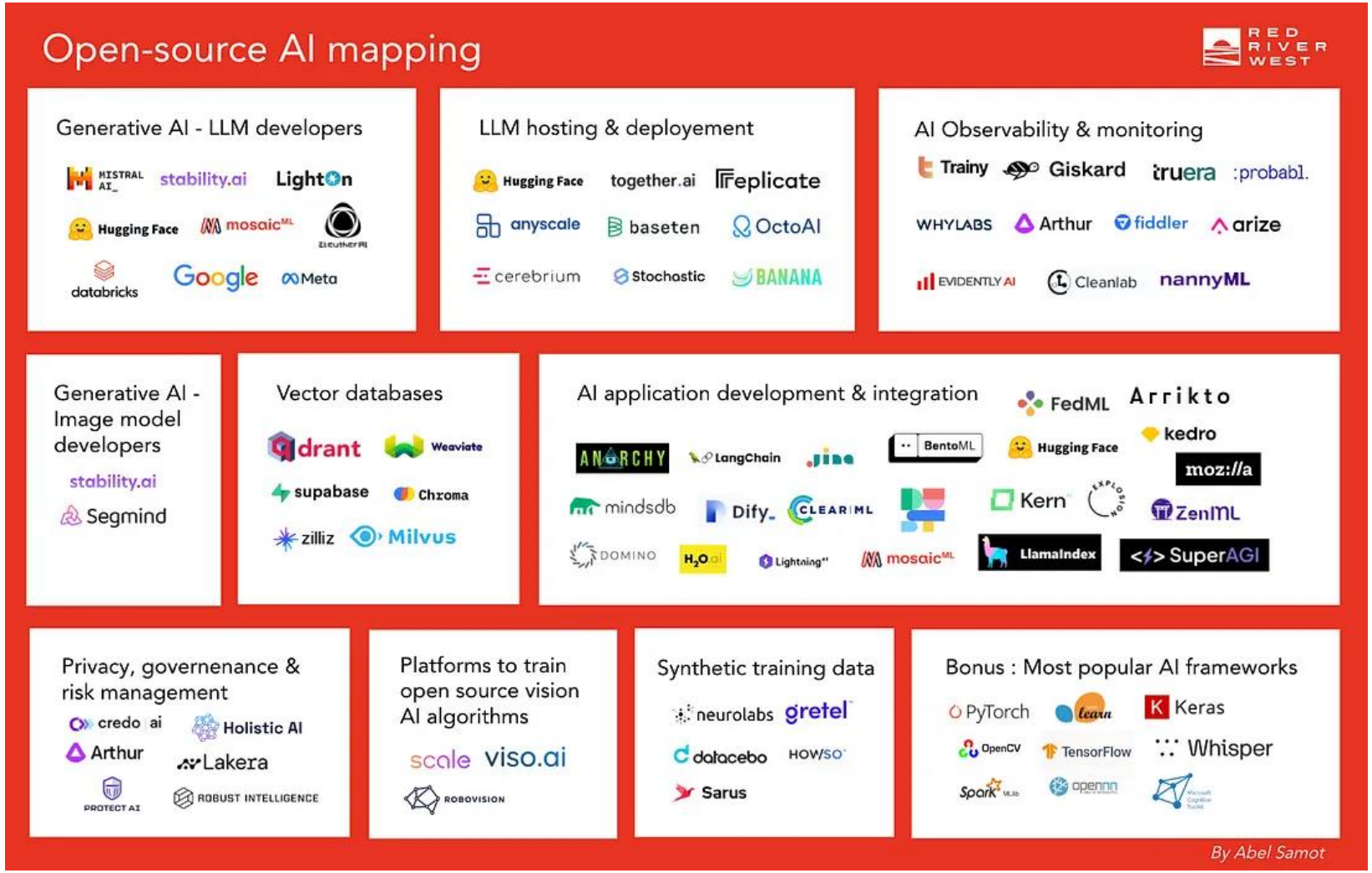
운영 및 서비스

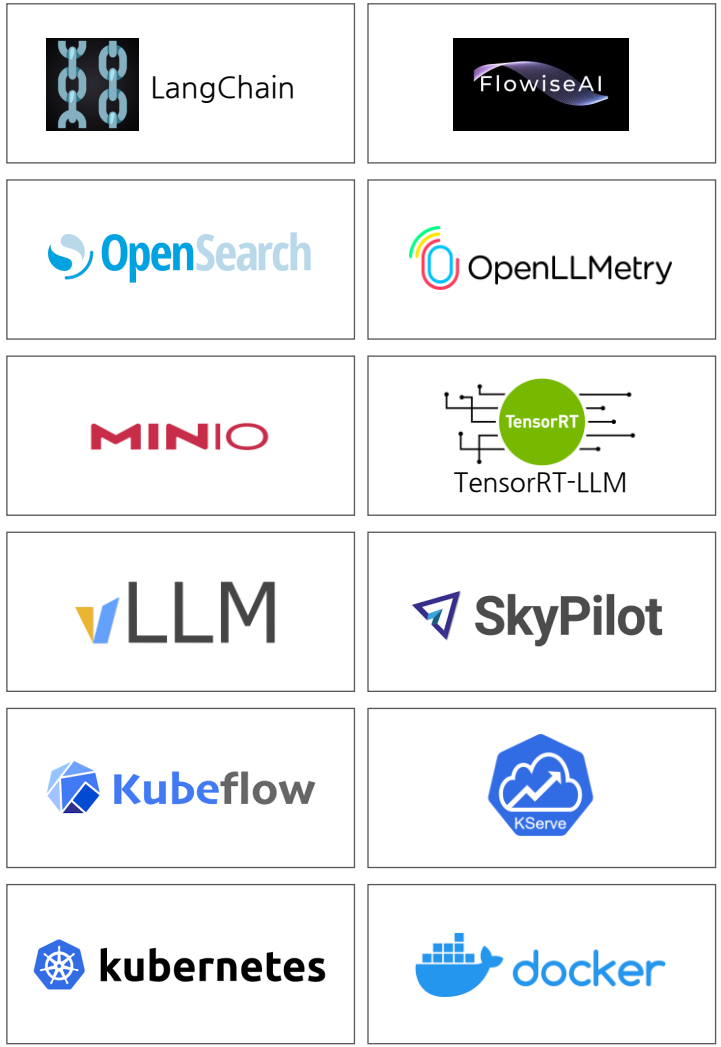
전문 인력



MEGAZONE CLOUD

오픈소스 생성형AI 에코시스템

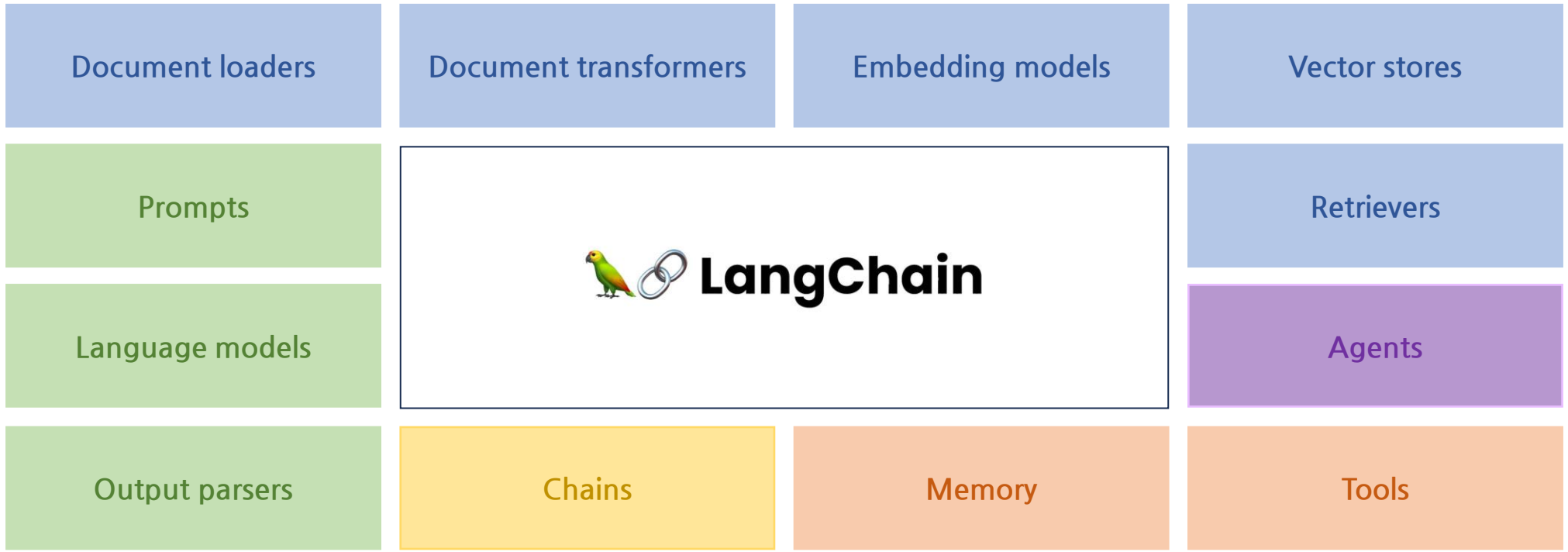




출처: Open-Source AI — Challenges, Opportunities & Ecosystem by Abel Samot https://medium.com/red-river-west/open-source-ai-mapping-advantages-debate-dd6be433eff6

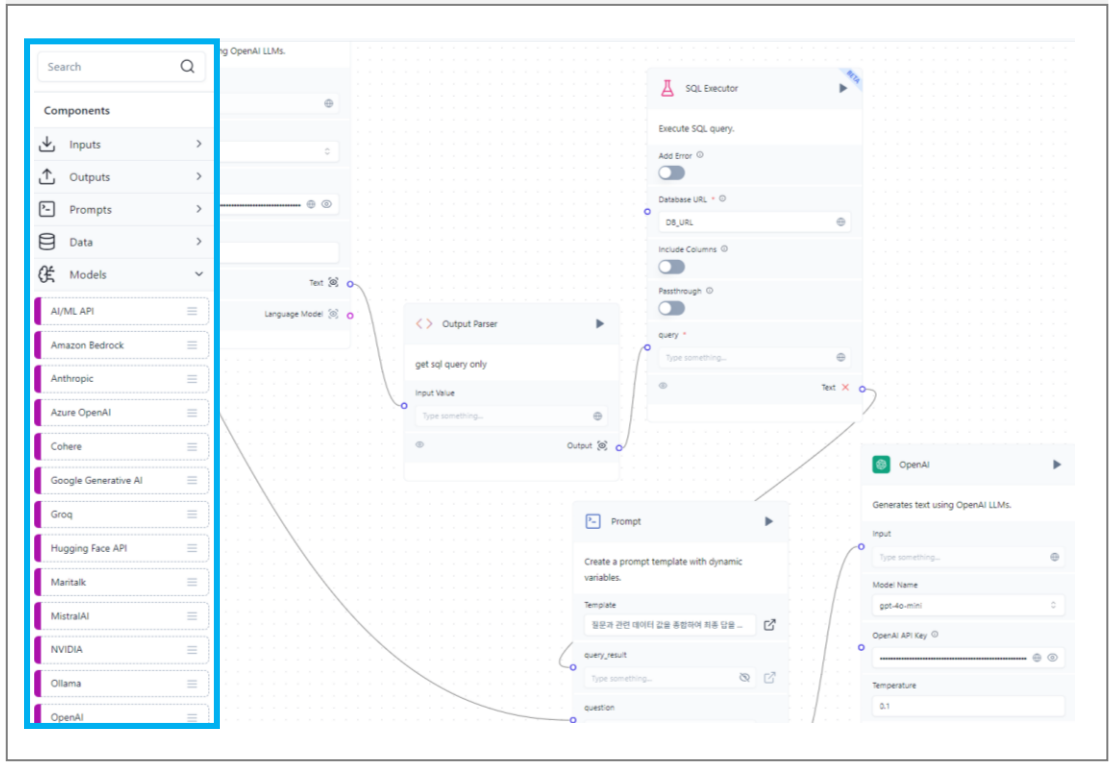
MEGAZONE CLOUD

LangChain 주요 구성요소

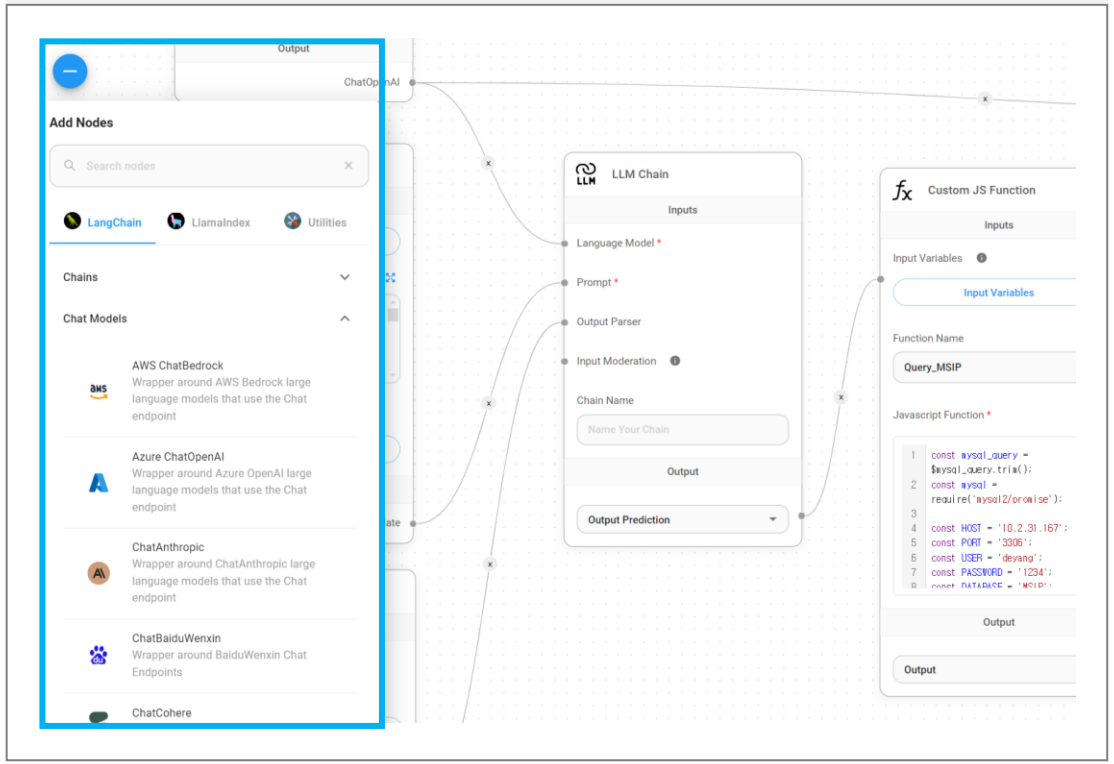


No-code & Low-code 저작도구

Langflow



FlowiseAI



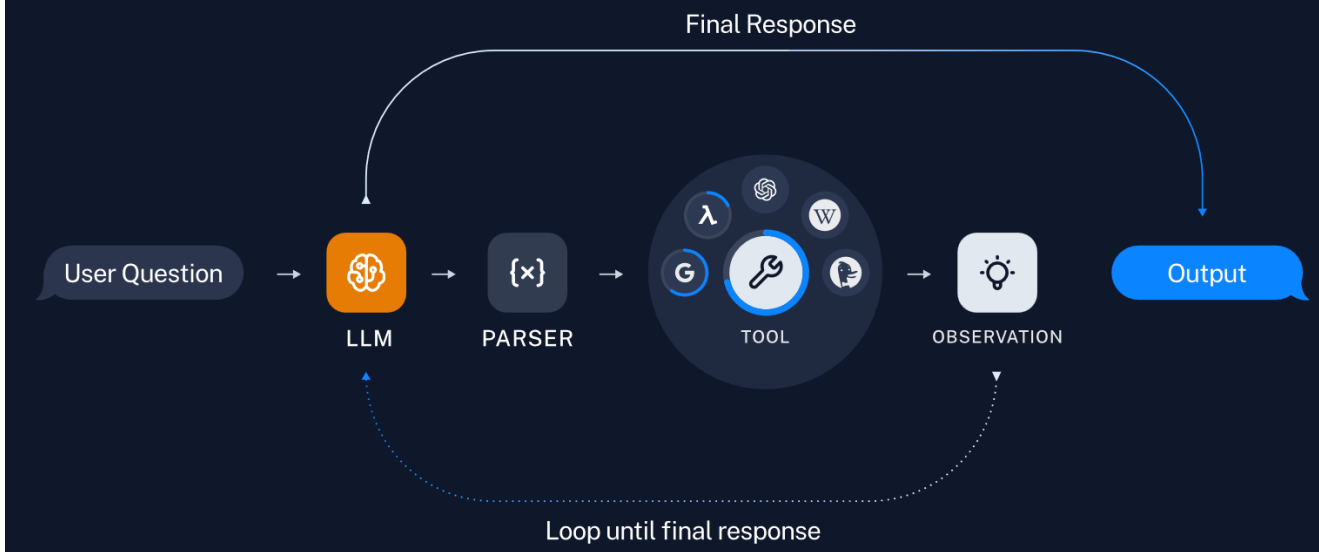
MEGAZONE CLOUD

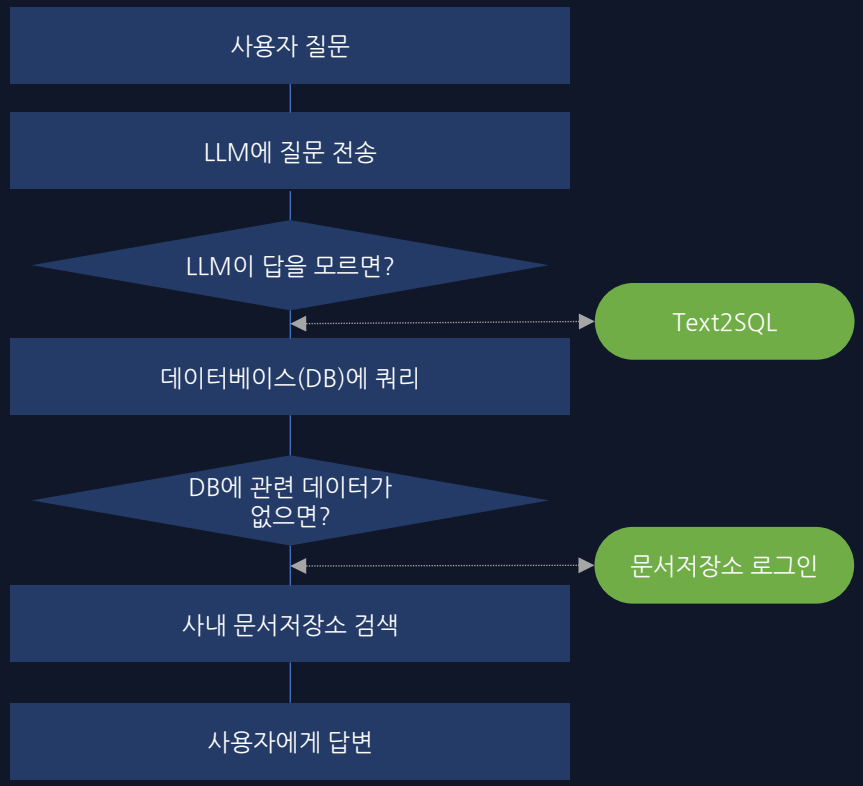
GenAI 에코 생태계의 빠른 확장

GenAI 에코 확장

다양한 목적의 특화 LLM	편리한 AI Agent 개발도구	보안 & 윤리 강화 요소

LangChain Agent

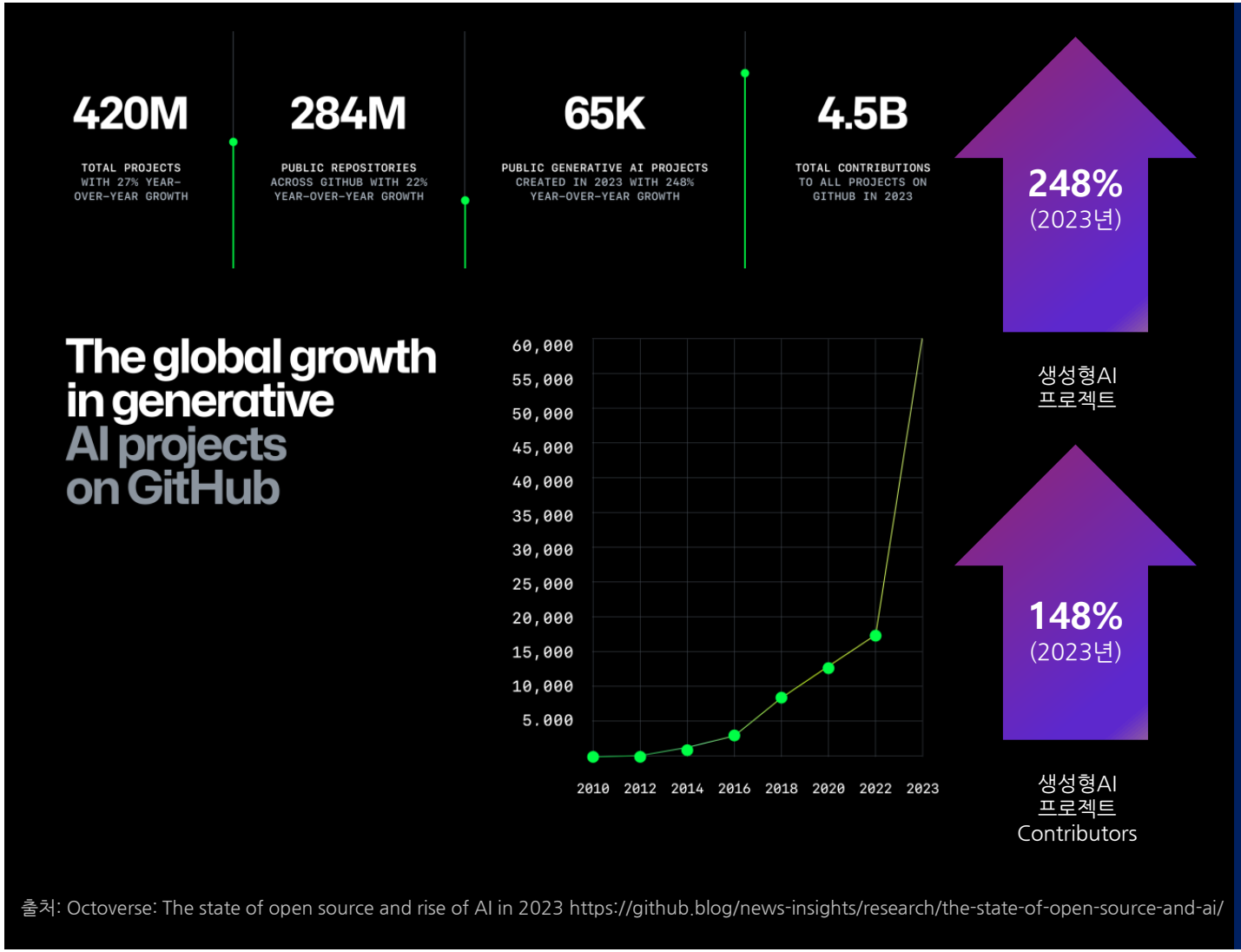




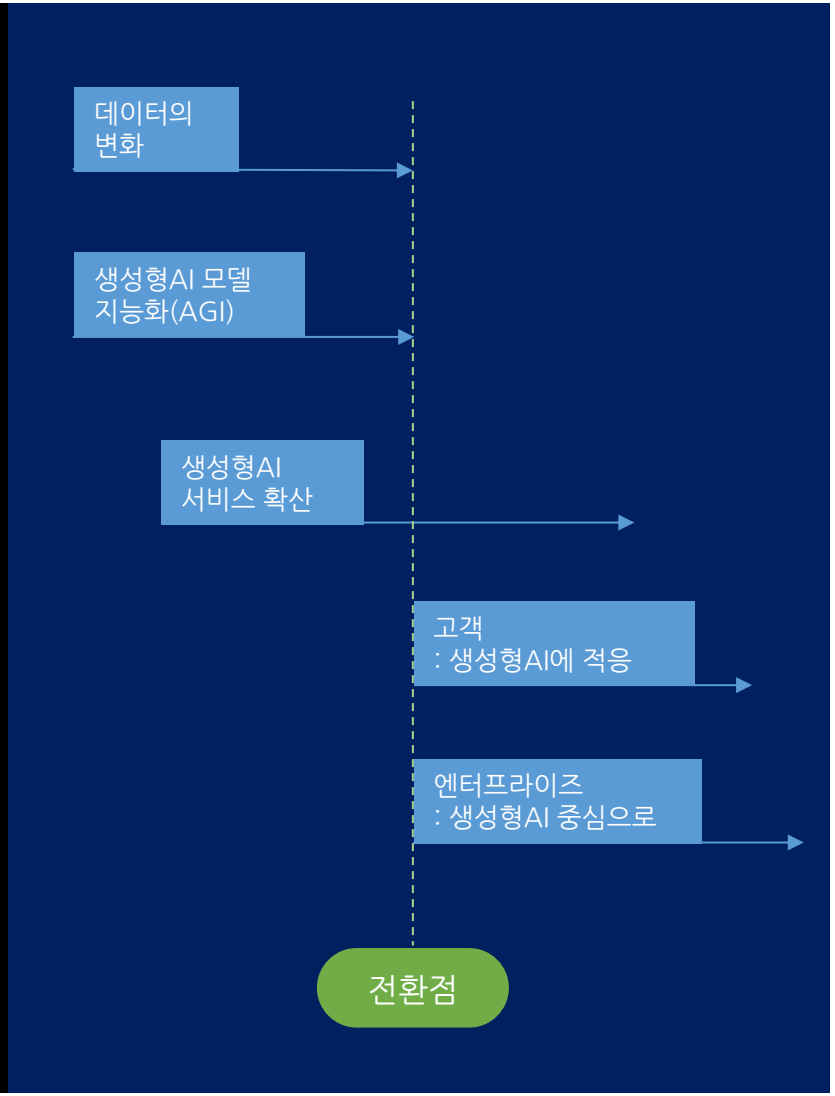
<GenAI Agent 예시>

MEGAZONE CLOUD

오픈소스, 그리고 엔터프라이즈 환경에서의 생성형AI 전망



출처: Octoverse: The state of open source and rise of AI in 2023 https://github.blog/news-insights/research/the-state-of-open-source-and-ai/



MEGAZONE CLOUD

맺는 말

“생성형AI와 함께하는 디지털 전환의 여정을 최적의 솔루션과 함께 메가존클라우드가 함께 하겠습니다”

MEGAZONE CLOUD