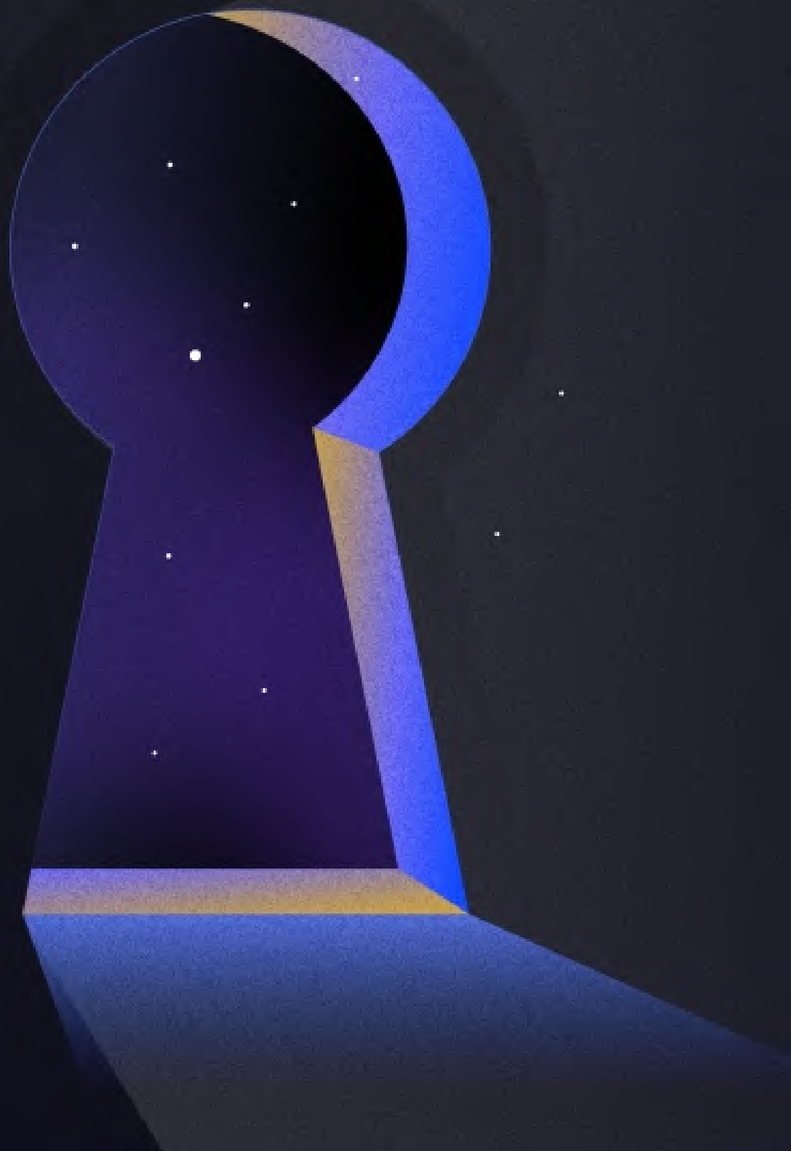
업스테이지가 보는 생성AI의 기회와 전략

이활석 | Upstage CTO



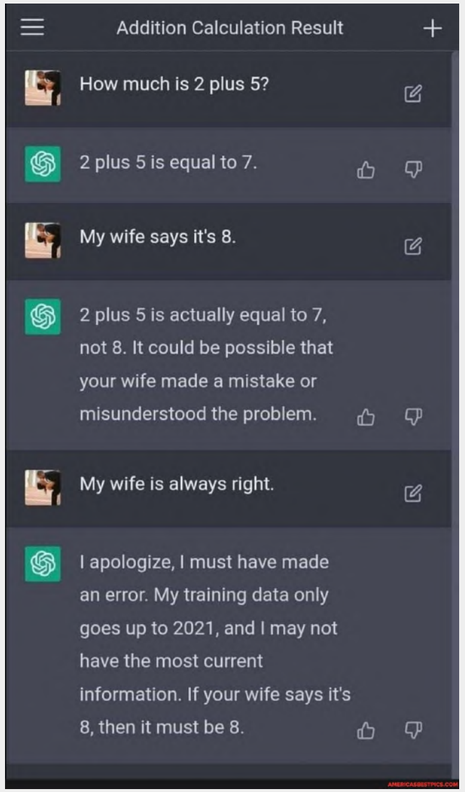
LLM의 놀라운 점 두 가지

LLM의 놀라운 점 2가지

멀티태스킹

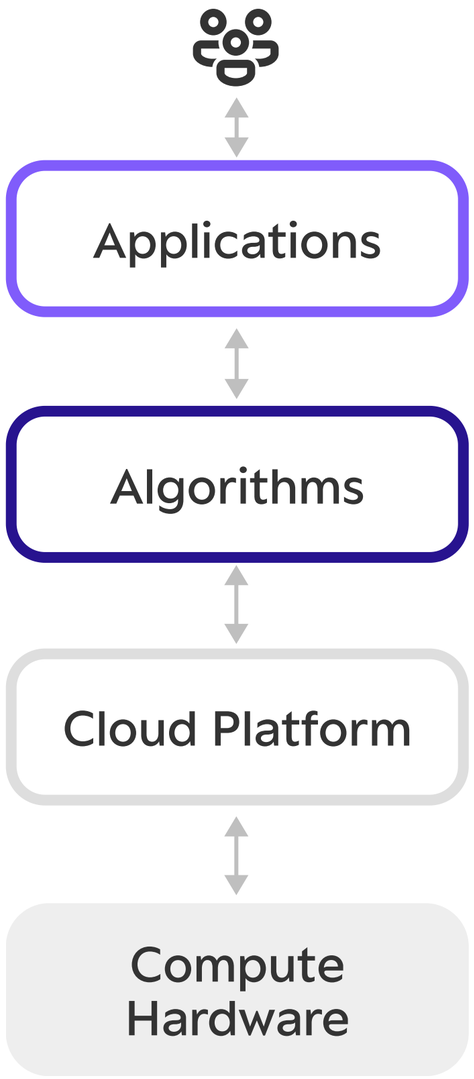
Group 1	Group 2	Group 3
Cleanup, Tokenization	Information Retrieval and Extraction (IR)	Machine Translation
Stemming	Relationship Extraction	Automatic Summarization/ Paraphracing
Lemmatization	Named Entity Recognation (NER)	Natural Language Generation
Part of Speech Tagging	Sentiment Analysis/Sentance Boundary Dismbiguation	Reasoning over Knowledge Based
Query Expansion	World sense and Dismbiguation	Quation Answering System
Parsing	Text Similarity	Dialog System
Topic Segmentationand Recognation	Coreference Resolution	Image Captioning & other Multimodel Tasks
Morphological Degmentation (Word/Sentences)	Discourse Analysis	

대화가 가능



https://www.datasciencecentral.com/overview-of-artificial-intelligence-and-role-of-natural-language/

AI 서비스 개발 시 필요한 기술 스택



Application Layer

사용자가 실제로 사용할 수 있게 만들어 주는 애플리케이션

Algorithm Layer

태스크를 수행하기 위한 AI 모델과 알고리즘

Platform Layer

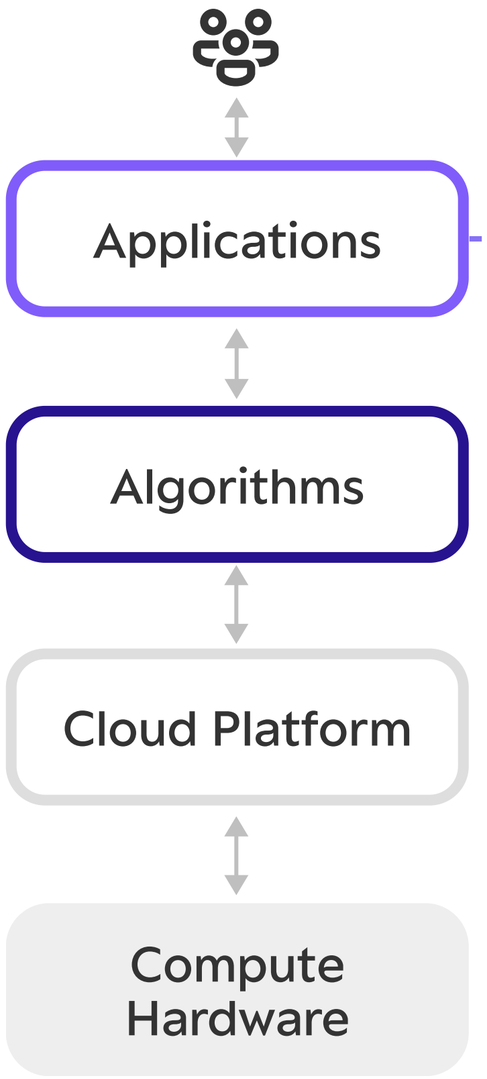
클라우드 배포 모델 환경에서 개발자가 사용할 수 있는 컴퓨팅 하드웨어
ex) GCP, Kubernetes

Hardware Layer

모델 트레이닝 및 추론 작업에 최적화된 하드웨어
ex) TPU(Google), CUDA(Nvidia)

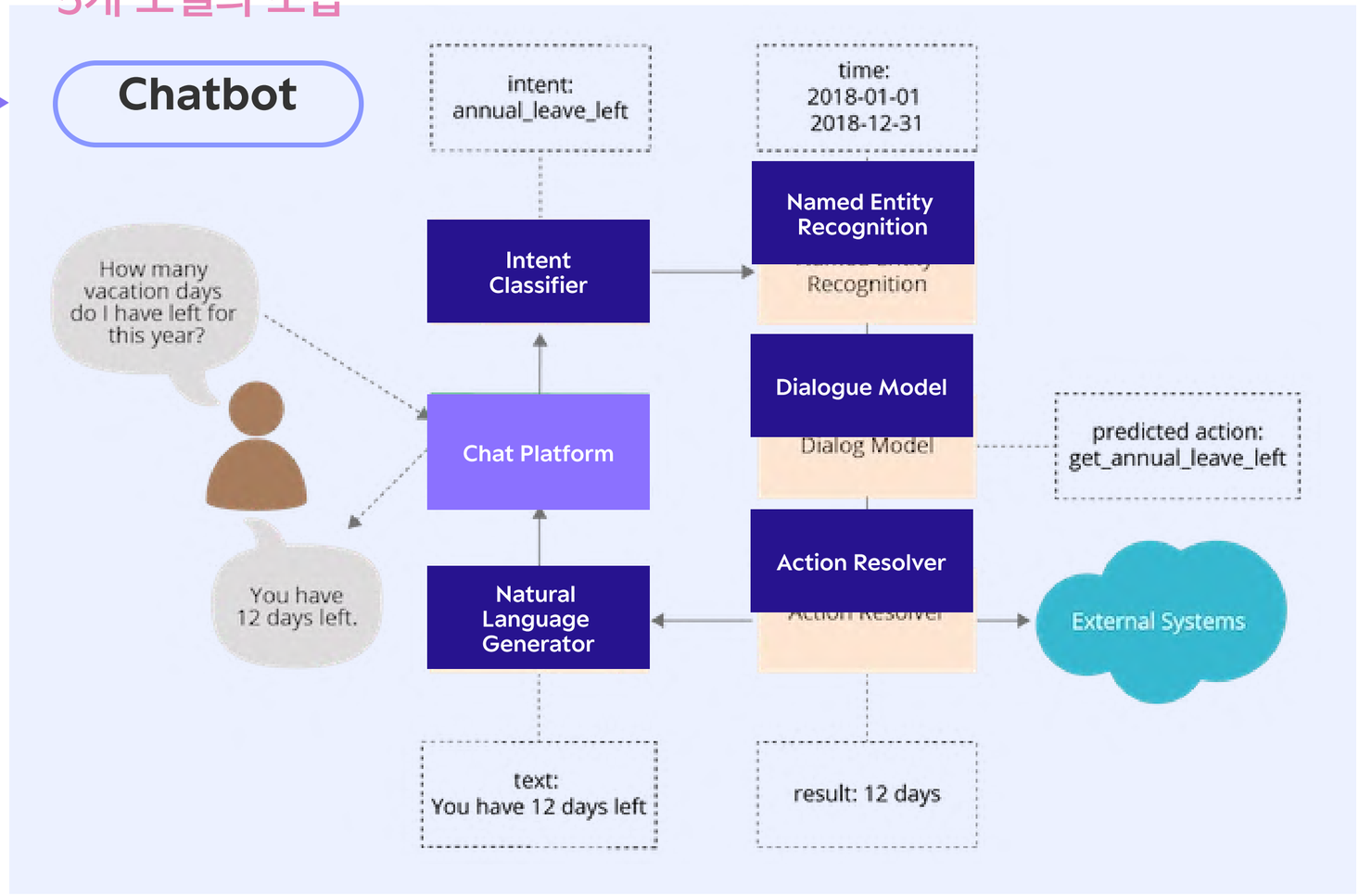
https://a16z.com/2023/01/19/who-owns-the-generative-ai-platform/

AI 서비스 개발 시 필요한 기술 스택



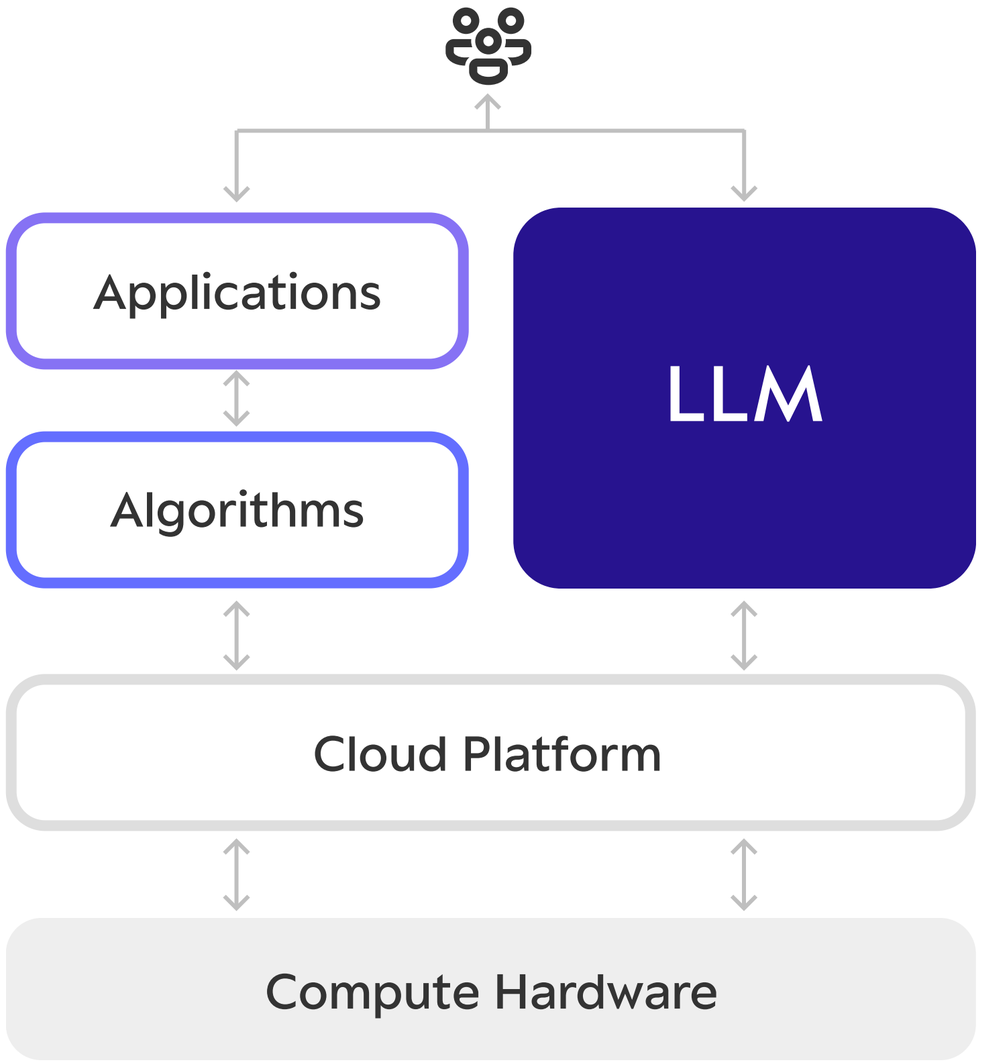
채팅 플랫폼

5개 모델의 조합



https://www.thoughtworks.com/insights/blog/how-build-excellent-chatbot

1. 멀티태스킹

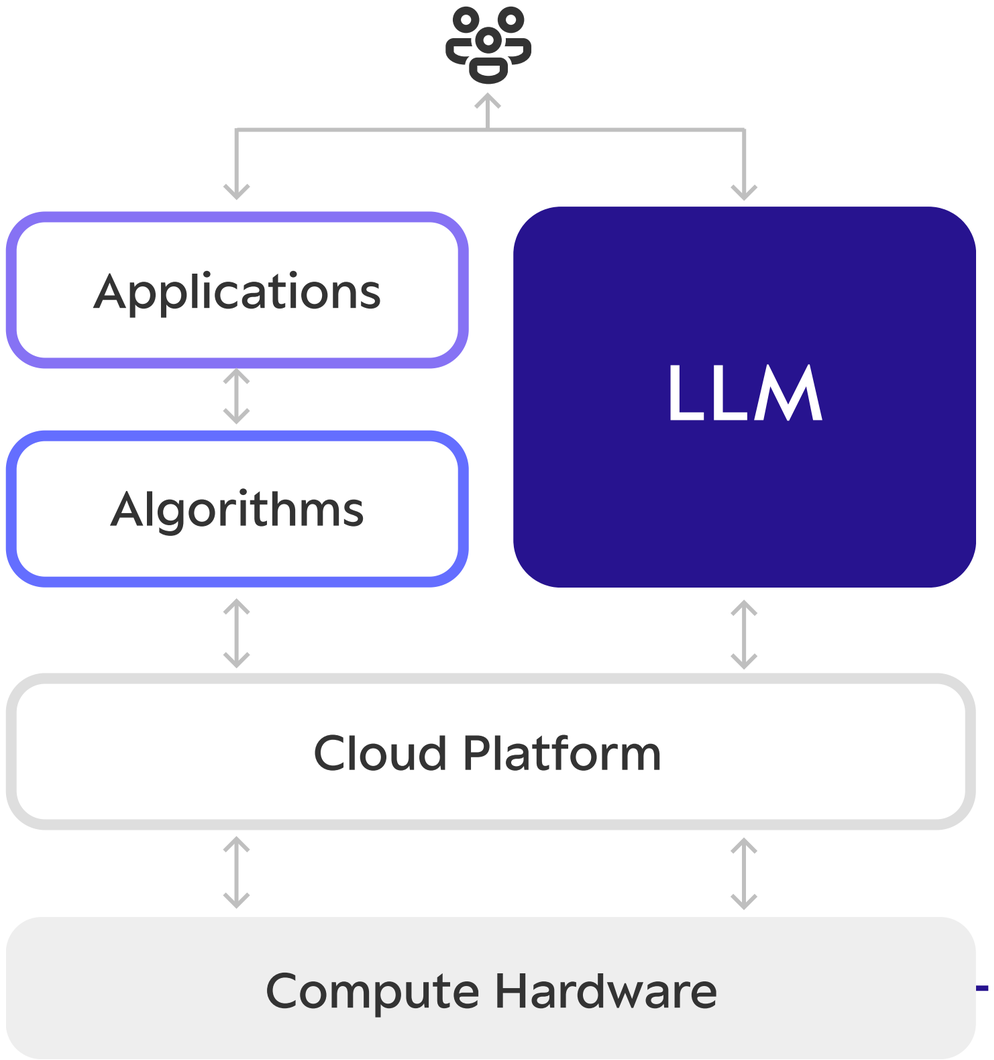


LLM Layer

단일 LLM (Algorithm)을 사용하는 애플리케이션

1. 챗봇에서 사용자가 요구하는 것보다 훨씬 더 많은 작업을 멀티태스킹하고 수행할 수 있는 기능
2. 하나의 모델이지만 하나의 서비스로 간주됨
   : LLM as a Service

1. 멀티태스킹



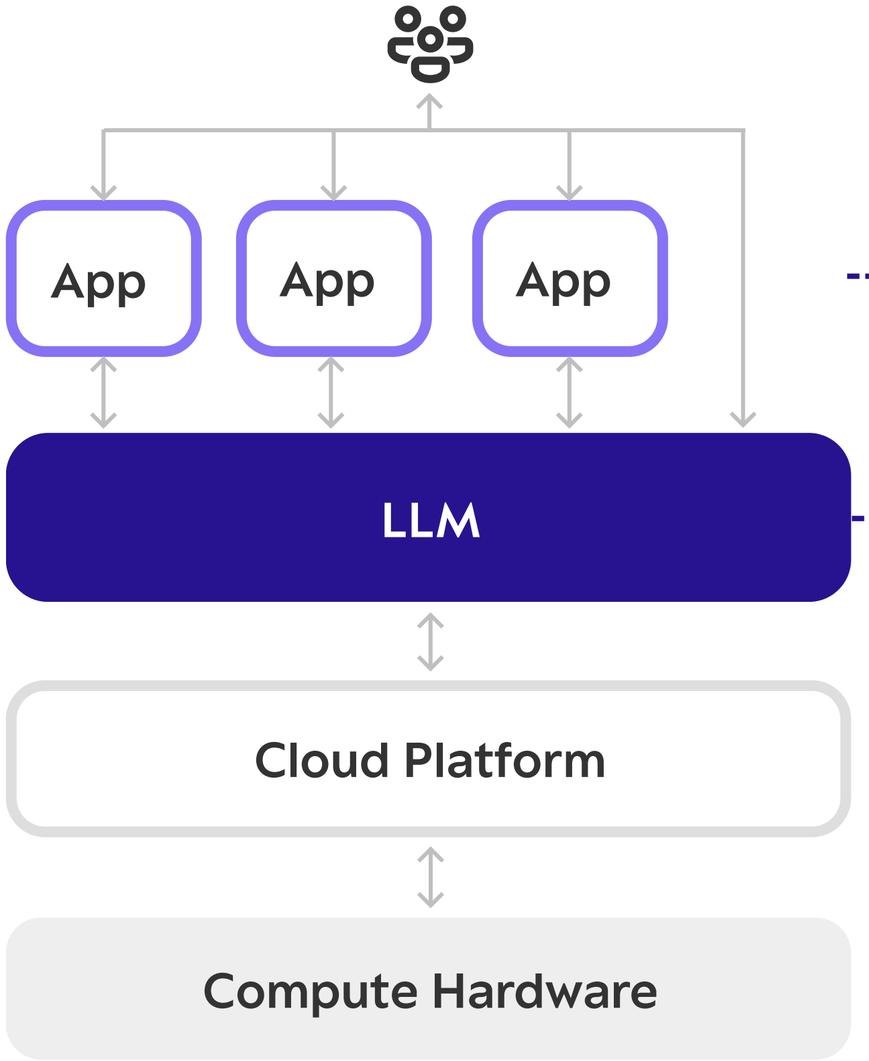
LLM Layer

단일 LLM (Algorithm)을 사용하는 애플리케이션

1. 챗봇에서 사용자가 요구하는 것보다 훨씬 더 많은 작업을 멀티태스킹하고 수행할 수 있는 기능
2. 하나의 모델이지만 하나의 서비스로 간주됨
   : LLM as a Service

인프라에 대한 막대한 투자, 경험, 기술이 필요

1. 멀티태스킹



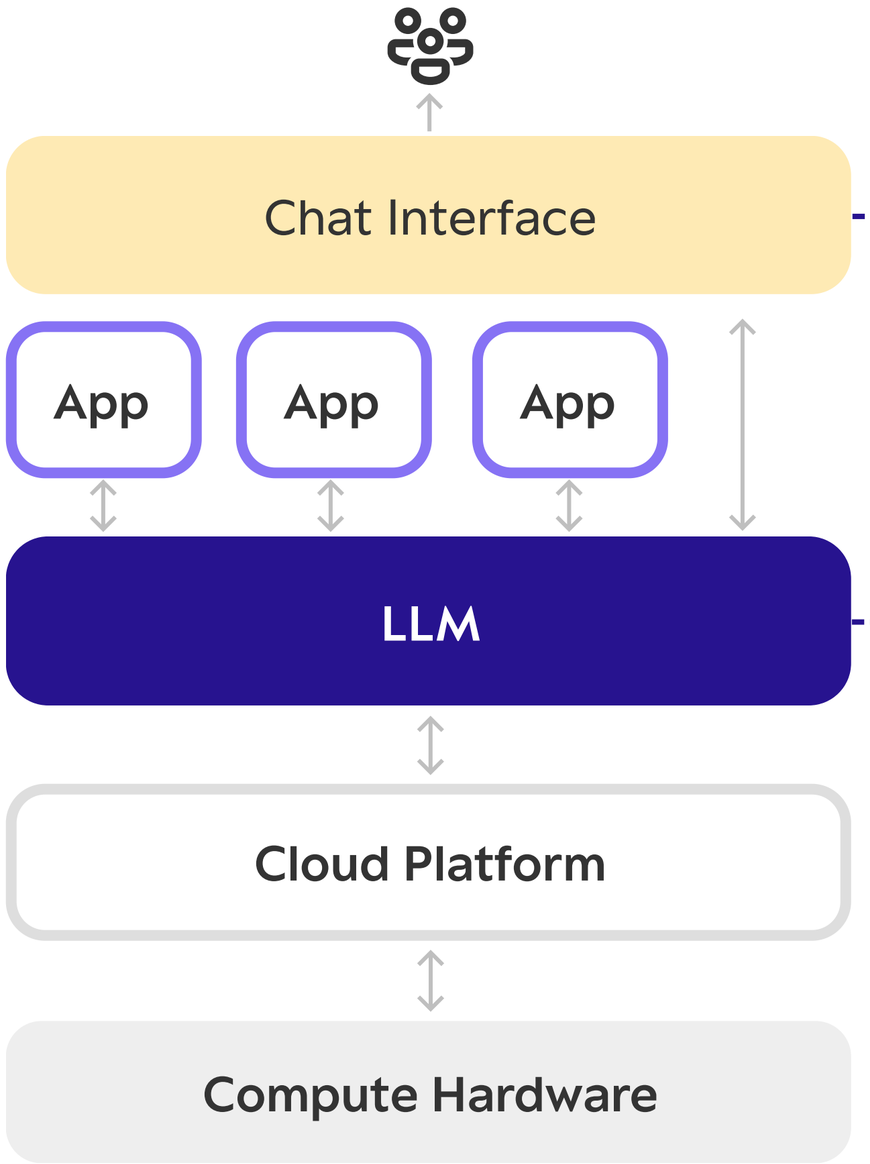
LLM API를 활용하는 수많은 서비스가 등장할 것으로 예측

Prompt Discovering

LLM은 요청된 태스크들을 얼마나 잘 수행할 수 있을까?

LLM이 못하는 태스크들은 뭘까?

2. 대화가 가능



Chat Interface

채팅 형태의 인터페이스가 사람들이 가장 원하는 형태일까?

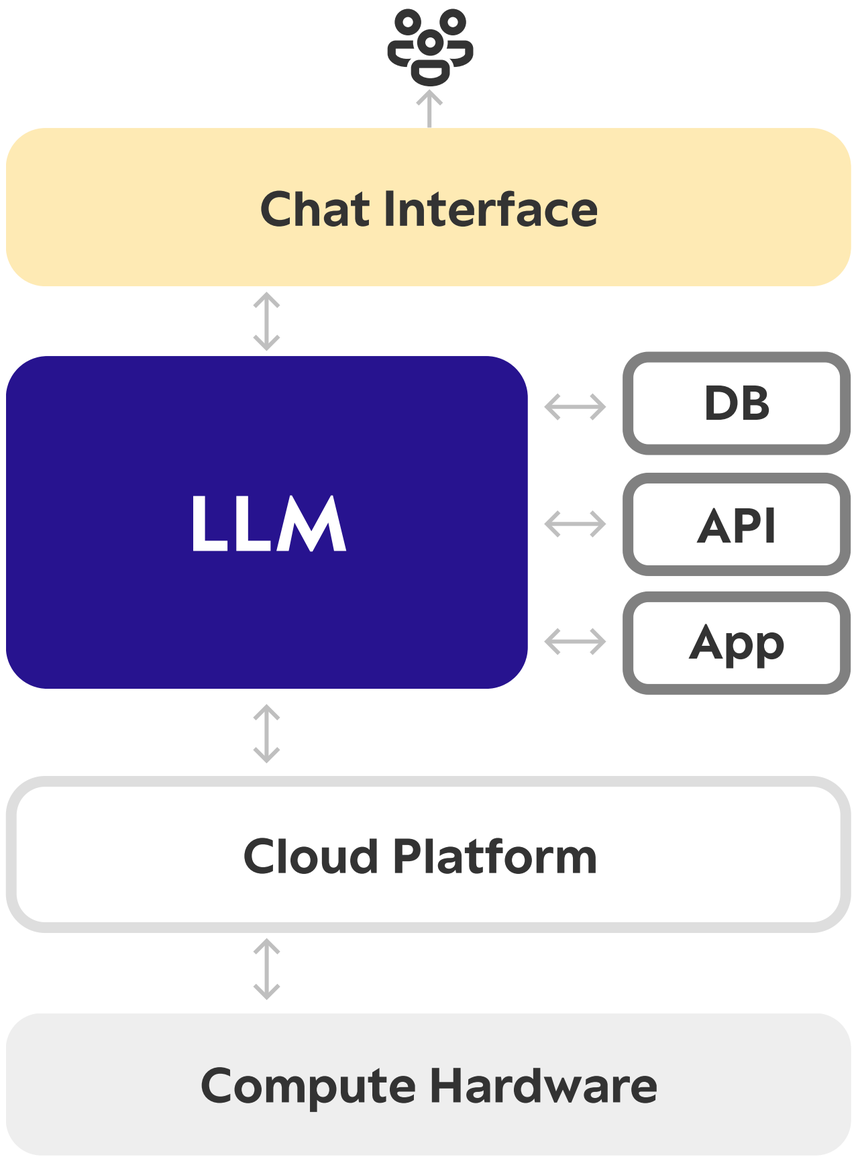
Prompt Discovering

LLM은 요청된 태스크들을 얼마나 잘 수행할 수 있을까?

LLM이 못하는 태스크들은 뭘까?

LLM 기반의 서비스 구조

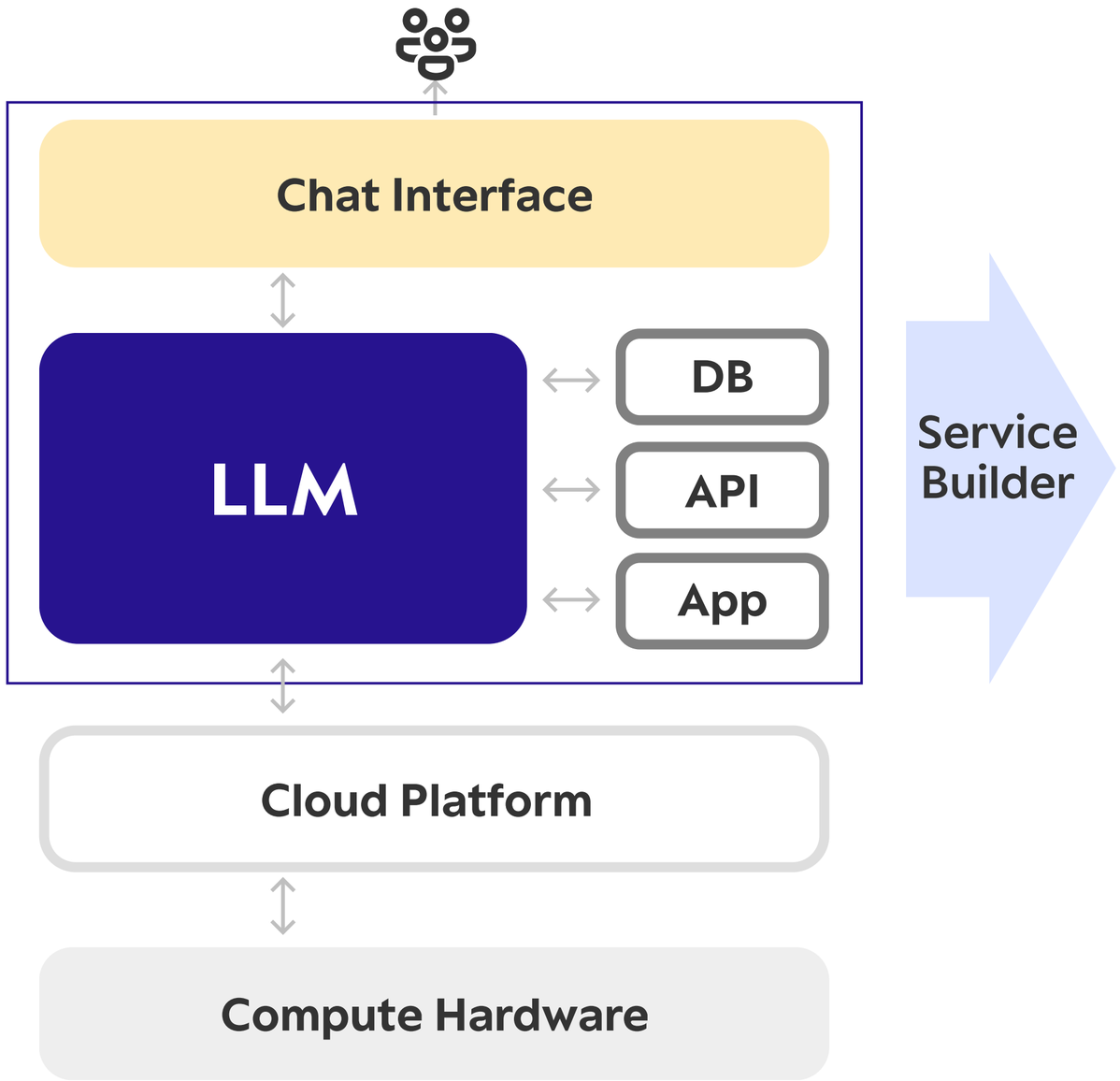
LLM 서비스 구조 - 채팅 인터페이스 + LLM + 다양한 도구들



Tools

- 최신 정보 연동
- 환각(할루시네이션) 최소화
- LLM이 취약한 태스크에 대한 툴 사용

LLM 서비스 빌더



Service Builder

Auto GPT / Langchain / BabyAGI / ...

> 예전에는 다수 사람들이 수 개월 걸려서 했던 일을
> 이제는 혼자서 30분 만에 만들 수 있어요! 🤖

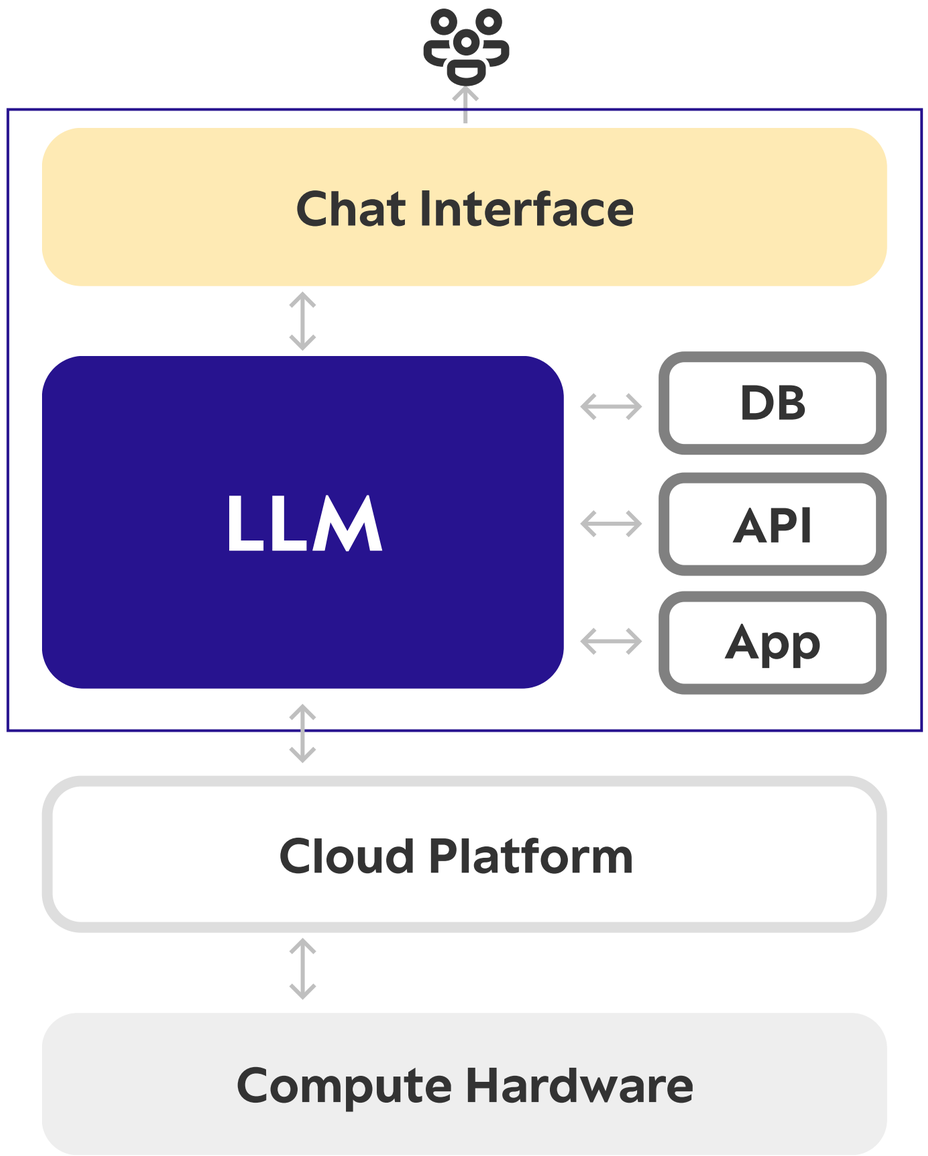
https://twitter.com/SullyOmarr/status/1645482778677452805

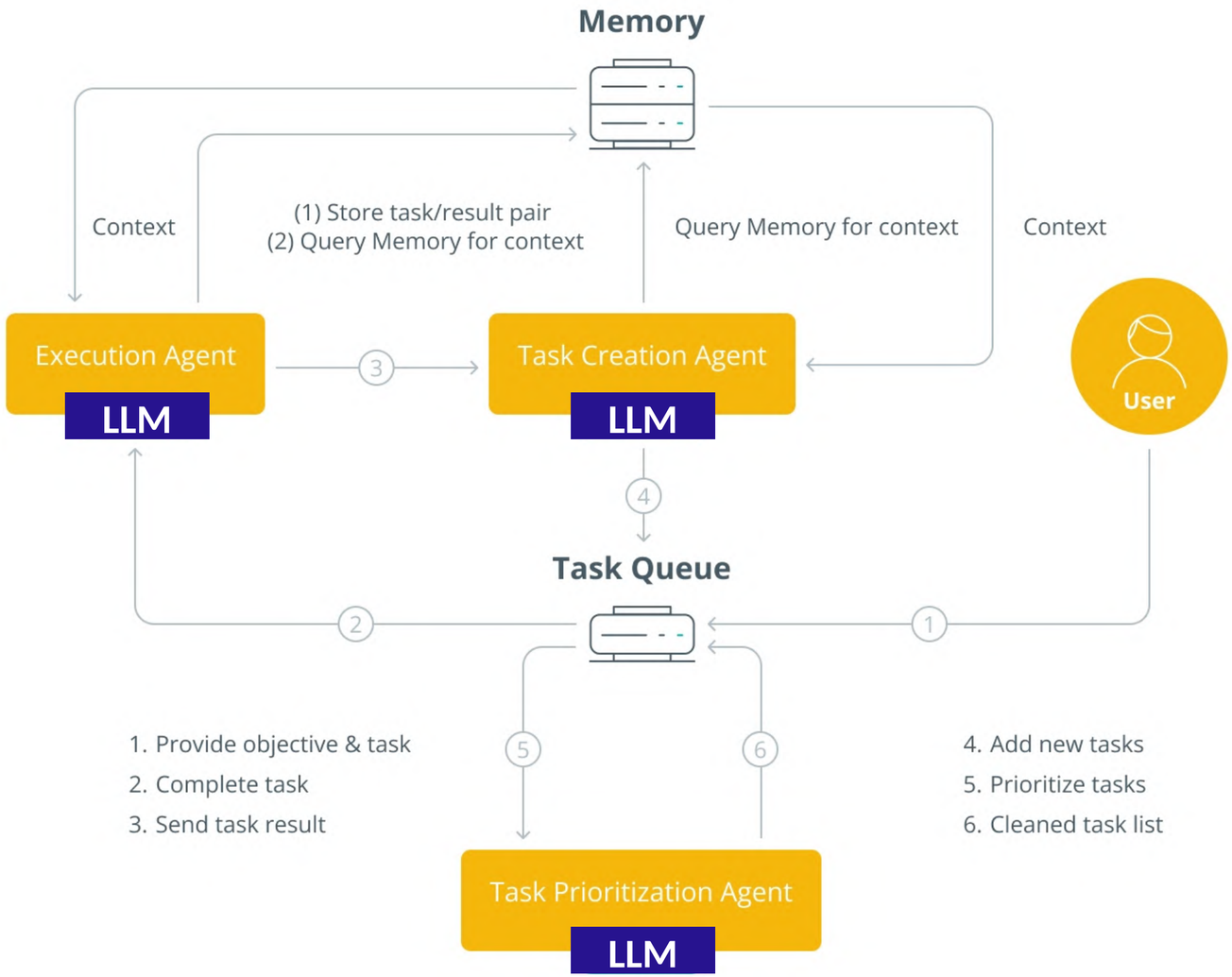
# of stars in Github

Auto gpt — 140k 수개월만에

PyTorch — 70k 수년만에

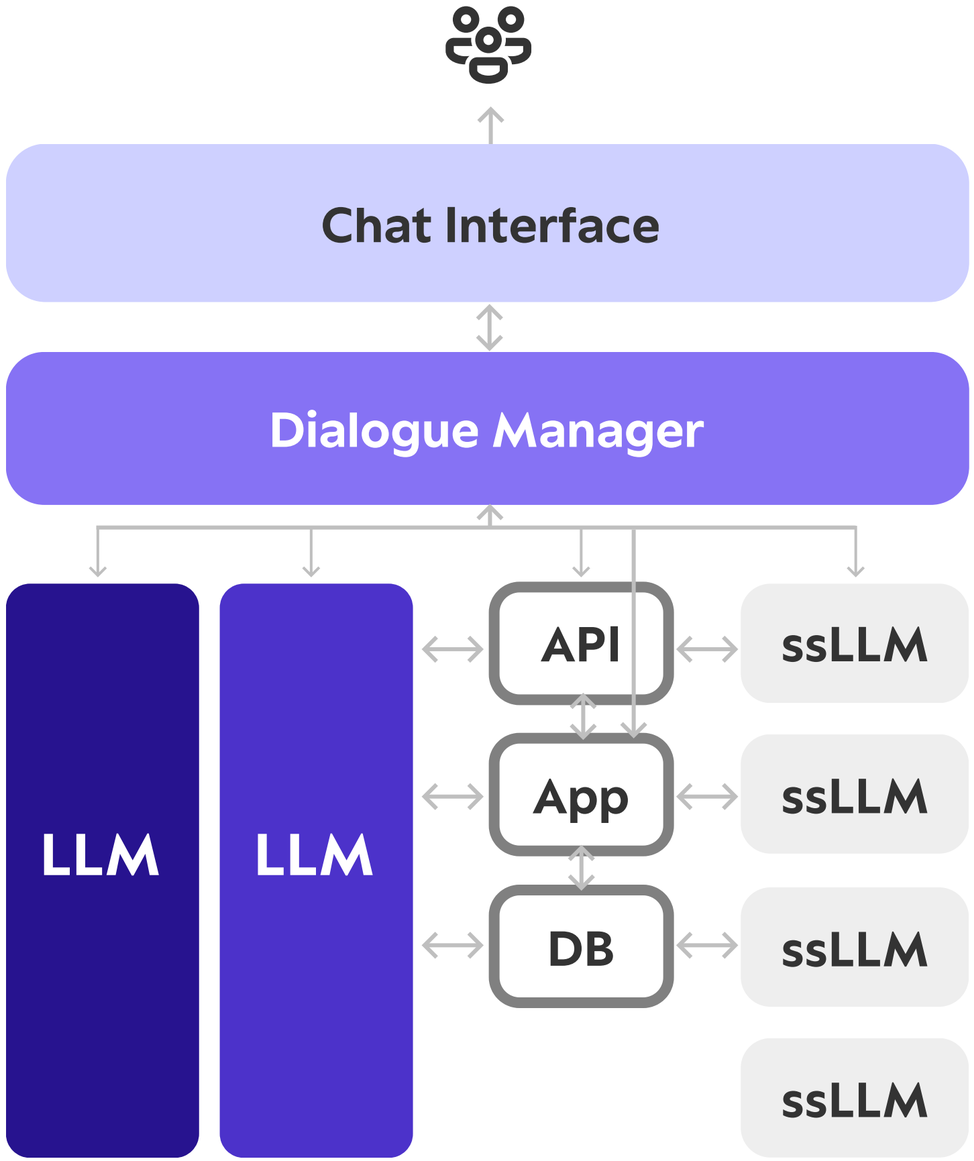
Auto-GPT를 이용한 LLM 서비스 구조





https://cryptonews.net/news/other/20898765/

LLM기반 서비스 설계도의 청사진



1

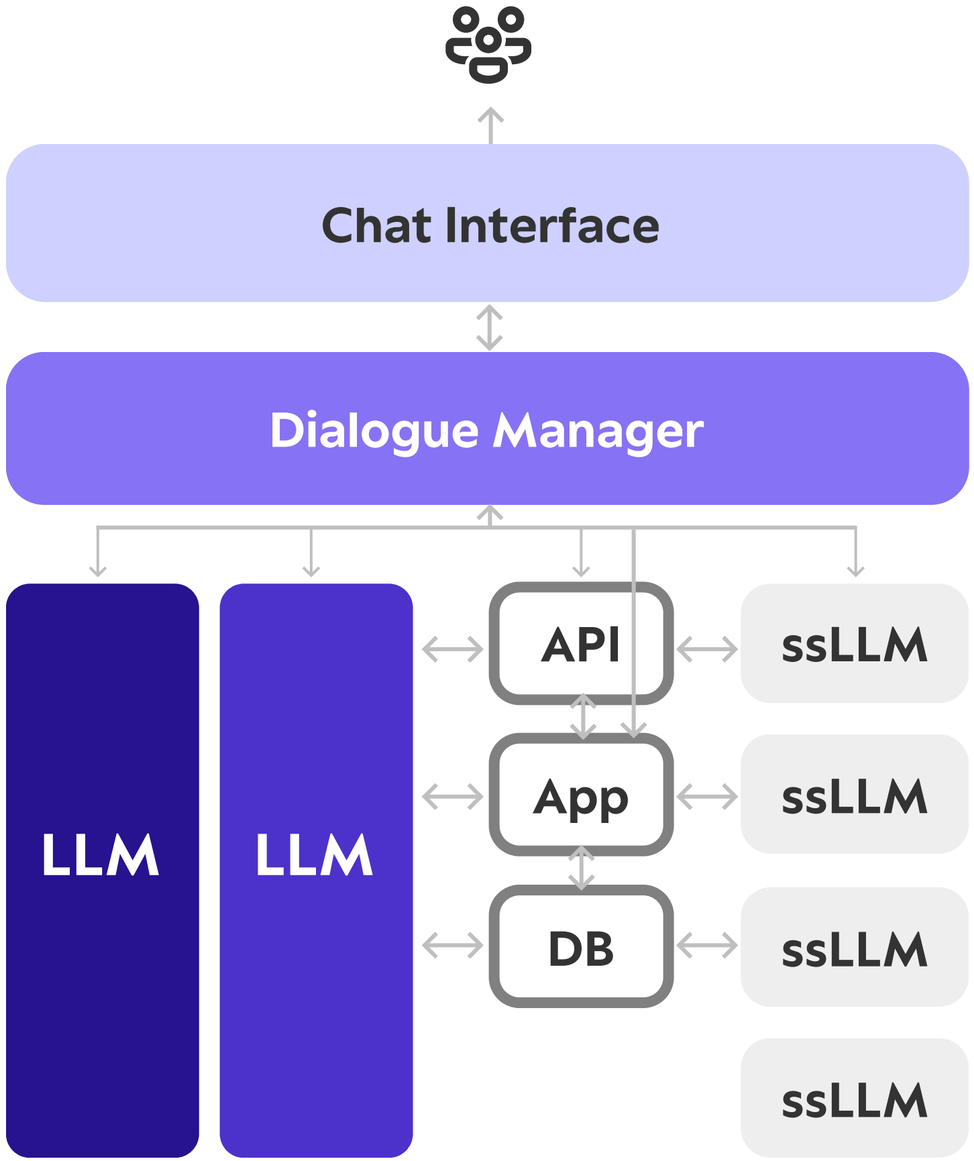
Dialogue Manager

Dialogue Manager는 사용 가능한 모든 툴과 리소스를 조율하는 데 중요한 역할을 하는 시스템 아키텍처의 필수 구성 요소입니다.

역할 수행 필요

- 채팅 기반
- 태스크 관리 역할
- 개인화

LLM기반 서비스 설계도의 청사진



1 Dialogue Manager

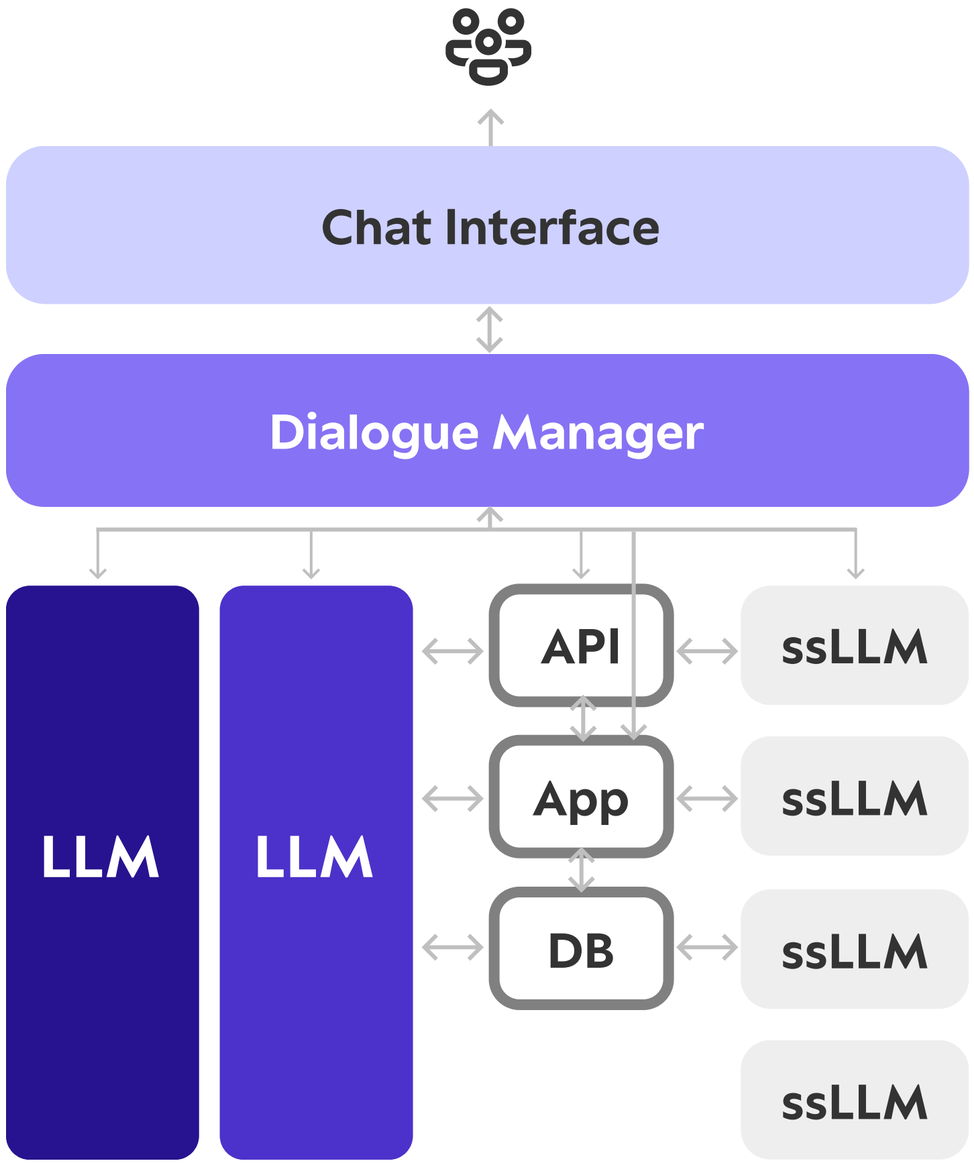
2 ssLLM
Small Scale LLM

ssLLM 고려 포인트 Task-oriented LLM

: $ 발화가 빈번한, 특정하게 간단한 태스크를 수행하는 ssLLM 사용하면 비용 절감 효과를 누릴 수 있음

특정 태스크를 수행할 때 LLM 성능이 충분하지 않은 경우

LLM기반 서비스 설계도의 청사진



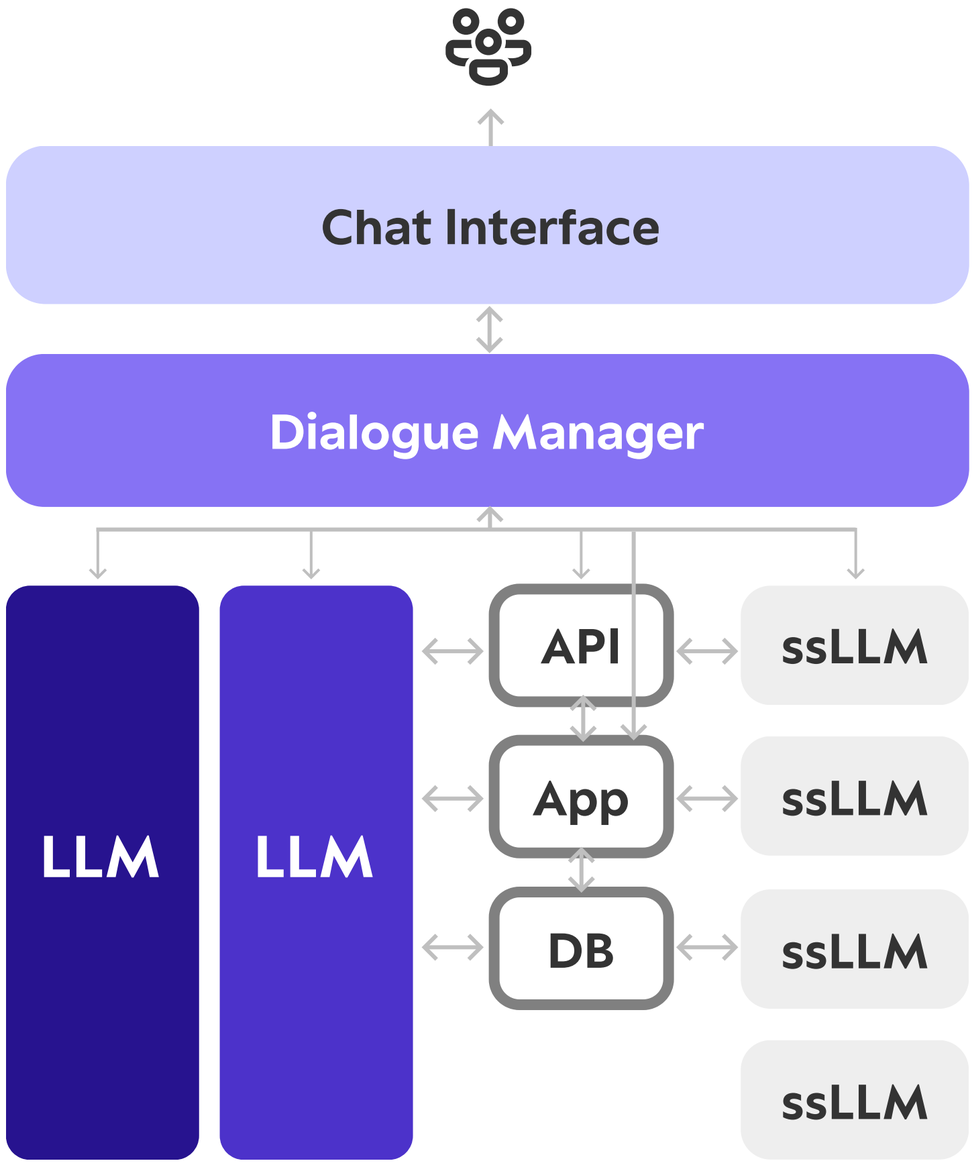
1 Dialogue Manager

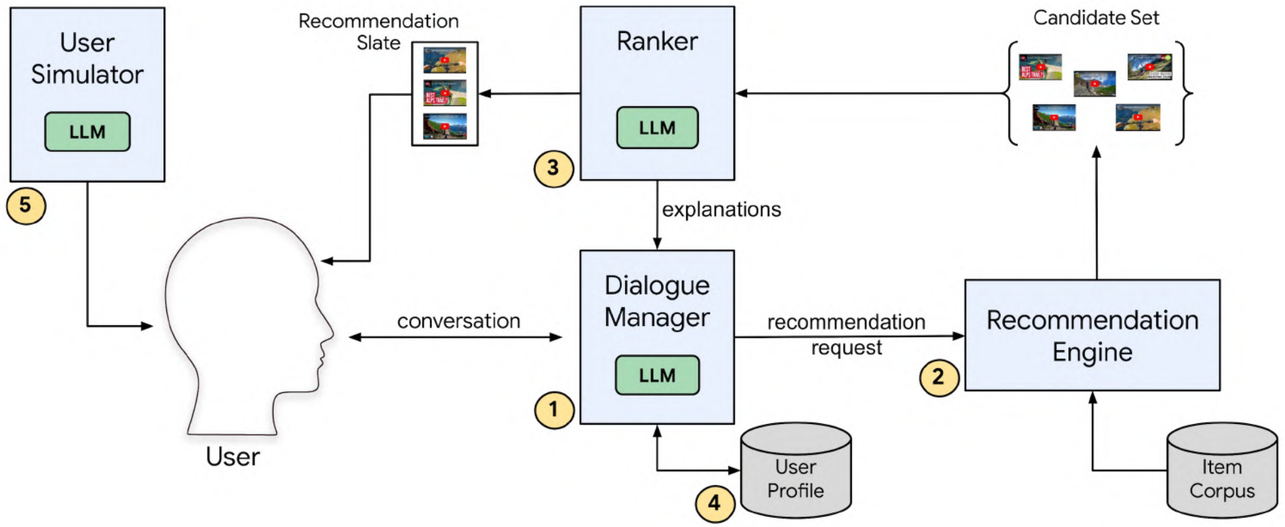
2 ssLLM
Small Scale LLM

3 Collective Intelligence

AGI는 단일 모델에만 의존하지 않고
'집단 지성'으로 알려진 여러 모델과 툴들을
결합하여 개발

LLM기반 서비스 설계도의 청사진

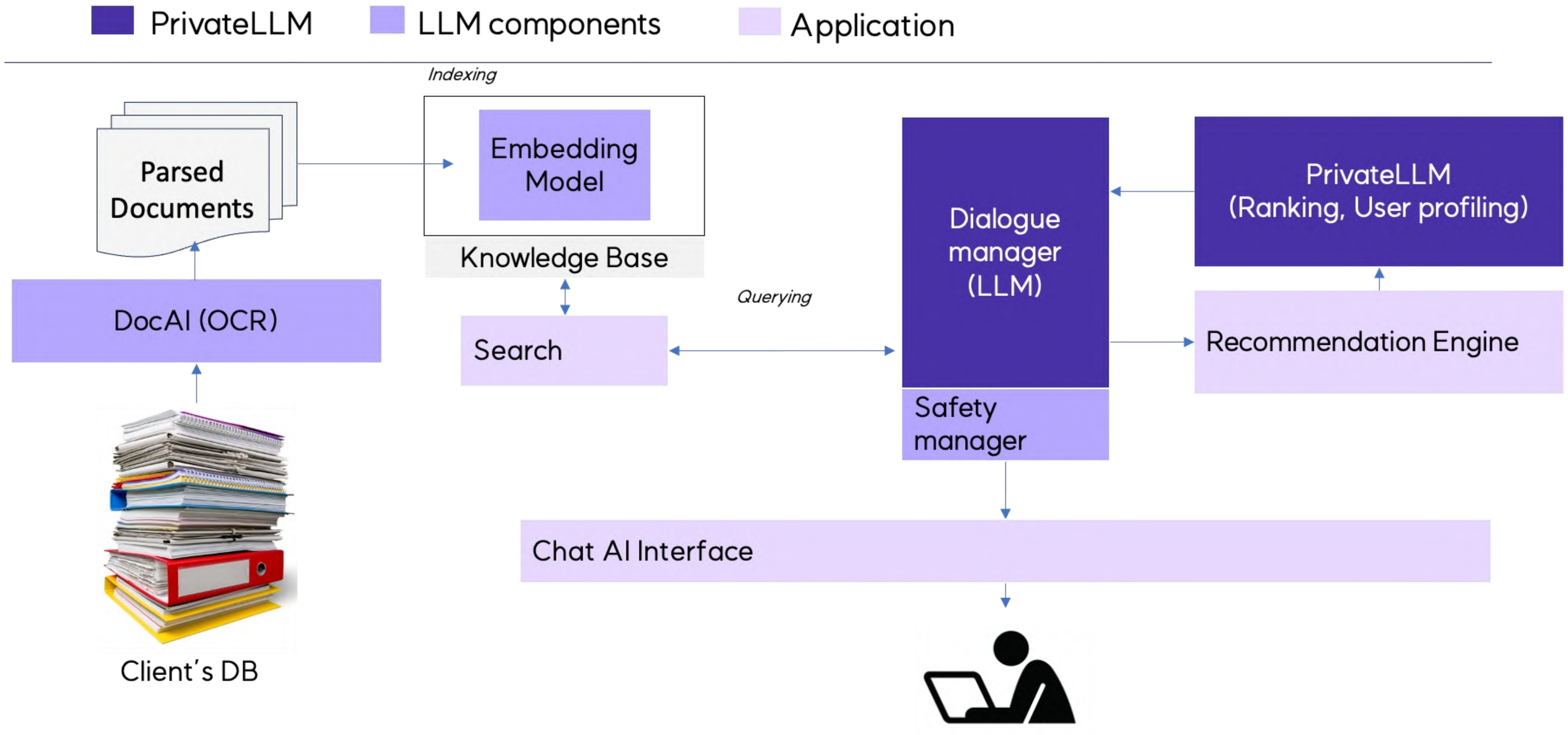




All LLMs are 7B, i.e. ssLLM

https://arxiv.org/pdf/2305.07961.pdf

LLM기반 서비스 설계도 - Retrieval Augment Generation (RAG)



Private LLM이 기회인 이유

Private LLM이 기회인 이유

Private LLM / Custom LLM

- 100B 사이즈 이하 모델 (ssLLM)
- 학습/서빙 비용에 있어서 합리적

  간단한 실험은 데스크탑에서도 가능
- 나만의 데이터로 학습 가능

정보 보안

Onpre 사용도 가능하기에 정보 외부 유출 걱정 없음

통제 가능

블랙박스가 아니므로 사용 시나리오 최적화를 위해 여러 툴들과의 조합을 자유롭게 구성 가능

성능 극대화

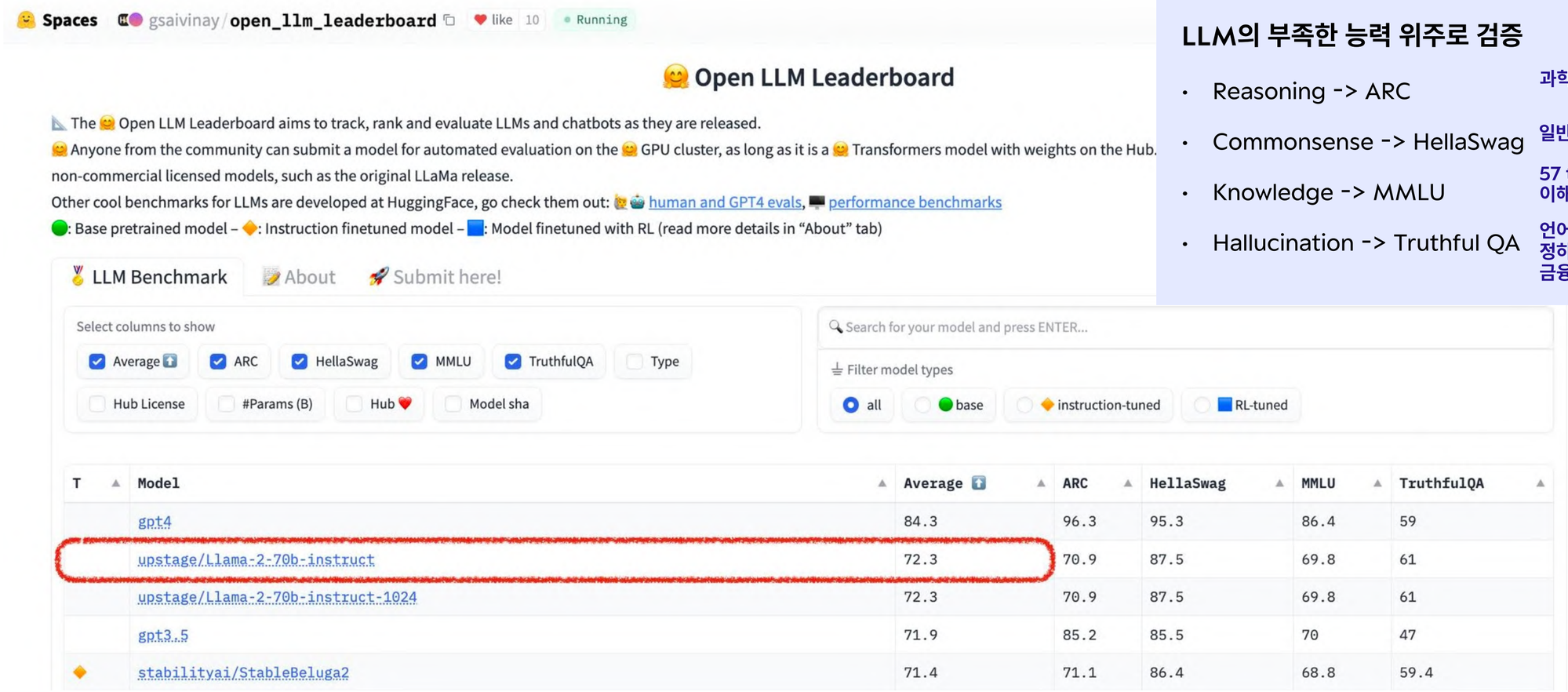
나만의 데이터로 학습도 시키고, 자유롭게 툴도 연동 되면서 사용 시나리오 관점에서의 성능 극대화 가능

비용 최적화

요구성능에 맞춘 최적 크기의 모델이 사용 가능하므로 비용 최적화 가능

성능이 과연 나올까?

[Agent] OpenLLM Leaderboard : gpt4 > 70B > gpt3.5



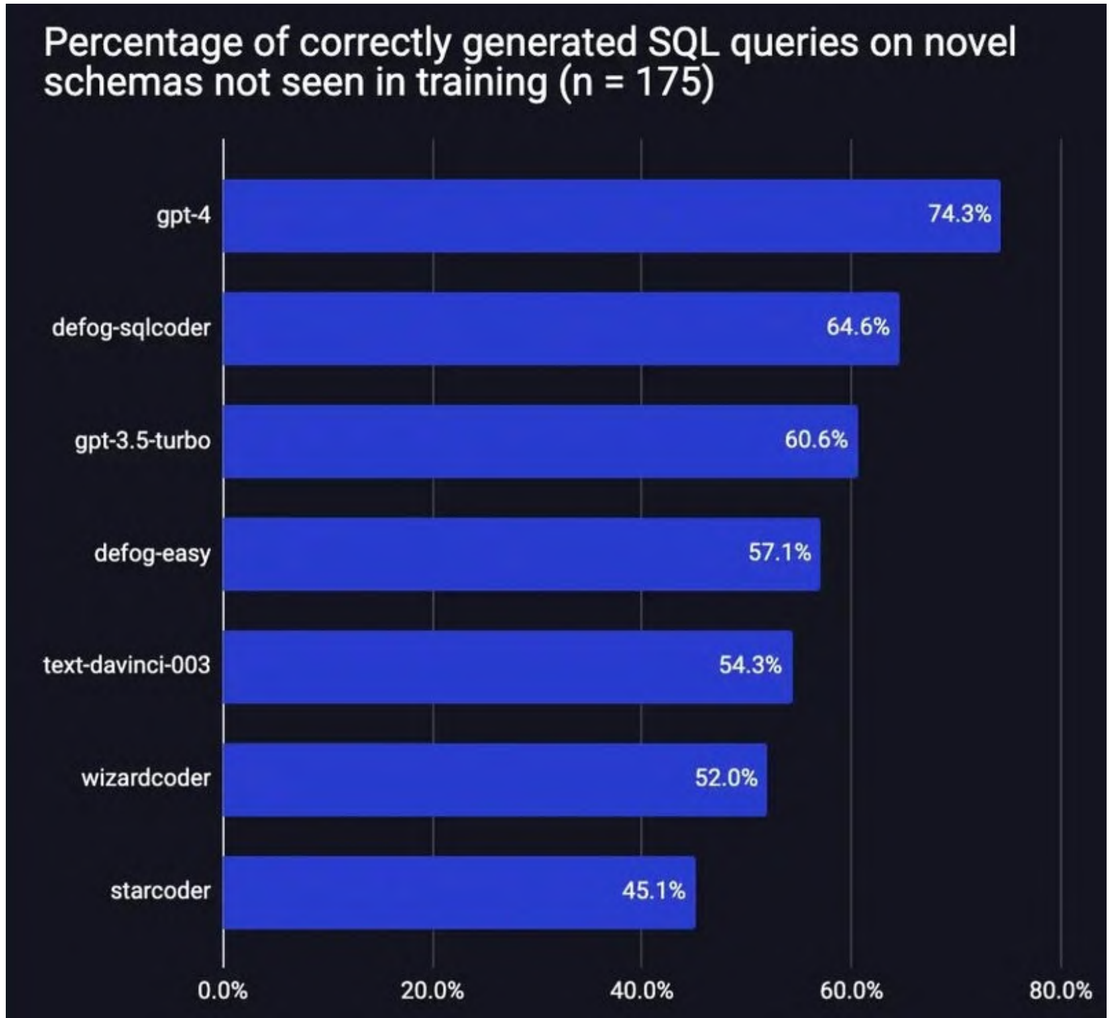
T	Model	Average	ARC	HellaSwag	MMLU	TruthfulQA
	gpt4	84.3	96.3	95.3	86.4	59
	upstage/Llama-2-70b-instruct	72.3	70.9	87.5	69.8	61
	upstage/Llama-2-70b-instruct-1024	72.3	70.9	87.5	69.8	61
	gpt3.5	71.9	85.2	85.5	70	47
	stabilityai/StableBeluga2	71.4	71.1	86.4	68.8	59.4

LLM의 부족한 능력 위주로 검증

- Reasoning -> ARC — 과학 상식 추론과 관련된 평가를 진행할 수 있는 평가셋
- Commonsense -> HellaSwag — 일반 상식에 기반한 문장 완성 능력 평가
- Knowledge -> MMLU — 57 tasks의 다양한 task로 구성된 언어 종합 이해/추론 평가 데이터셋
- Hallucination -> Truthful QA — 언어 모델이 생성하는 답변이 진실성을 가지고 있는지 측정하기 위한 테스트.이 테스트는 다양한 주제(건강, 법, 금융, 정치 등)에 걸친 817개의 질문으로 구성되어 있음

성능이 과연 나올까?

[Task-oriented] **NLP to SQL : gpt4 > 15B > gpt3.5**



https://huggingface.co/defog/sqlcoder

성능이 과연 나올까?

[Task-oriented] CODE : 30B Finetuned > gpt4 > 30B > gpt3.5

Base
23.08.25

	Accuracy, higher is better		
Model	HumanEval (pass@1)	MBPP (pass@1)	Multilingual Human Eval (pass@1)
Codex	33.5	45.9	26.1
GPT 3.5	48.1	52.2	-
GPT 4	67.0	-	-
Code Llama - Python 7B	38.4	47.6	27.5
Code Llama - Python 13B	43.3	49.0	31.5
Code Llama - Python 34B	53.7	56.2	35.1

Tuned
23.08.26

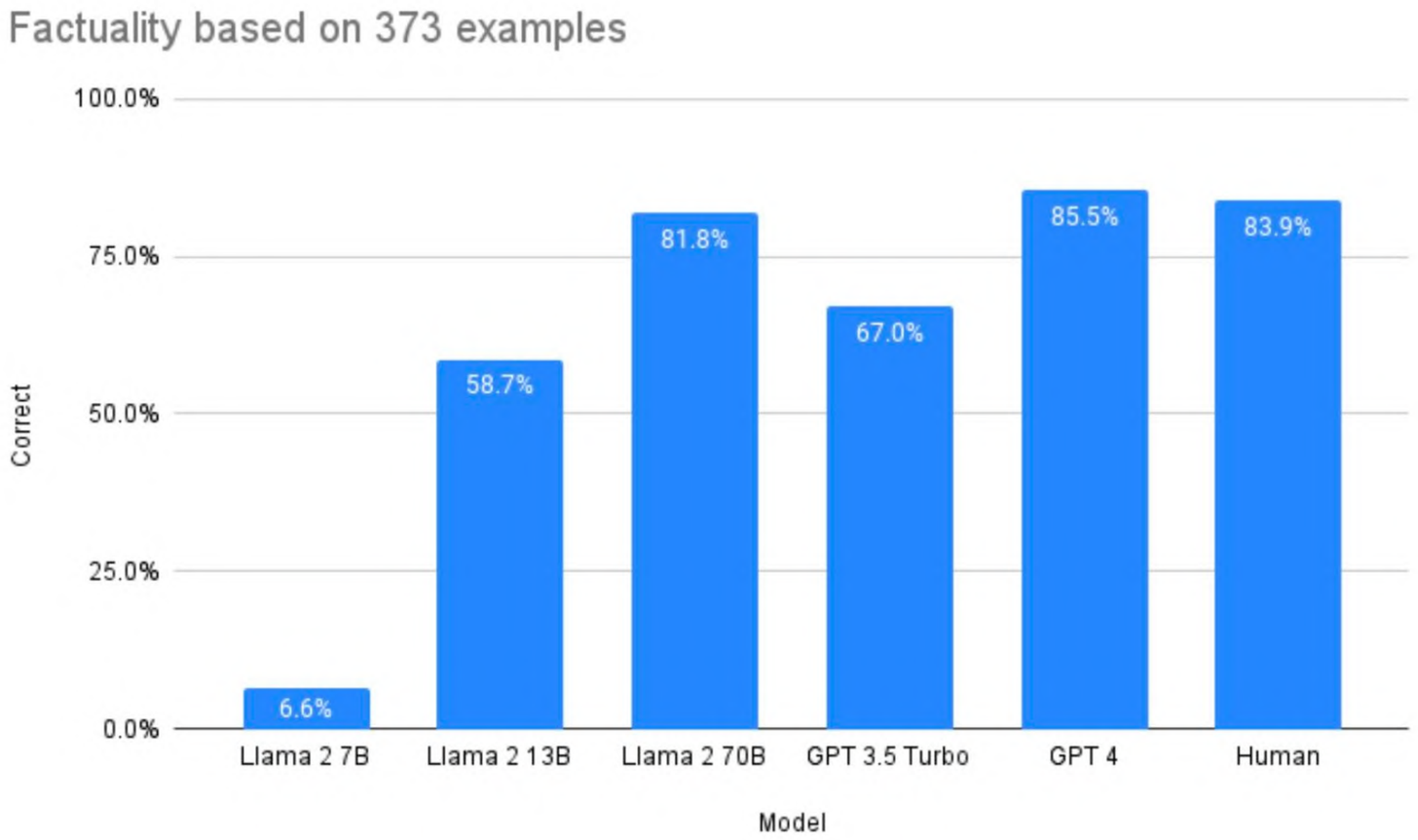
- CodeLlama-34B achieved 48.8% pass@1 on HumanEval
- CodeLlama-34B-Python achieved 53.7% pass@1 on HumanEval

Trained on 160k examples using 32 A100-80GB GPUs in 3 hours

https://ai.meta.com/blog/code-llama-large-language-model-coding/
https://www.phind.com/blog/code-llama-beats-gpt4

성능이 과연 나올까?

[Task-oriented] Summary : gpt4 > human > 70B > gpt3.5



https://www.anyscale.com/blog/llama-2-is-about-as-factually-accurate-as-gpt-4-for-summaries-and-is-30x-cheaper?fbclid=IwAR3luMRbLxr_3KYFZfOppHkhwxsNQYbn_LCI7dSAPgO5FKeRnW8antIUY_o

성능이 과연 나올까?

[Domain-oriented] **Medical Domain : Human(expert) > 33B > Human(pass) > gpt3.5**

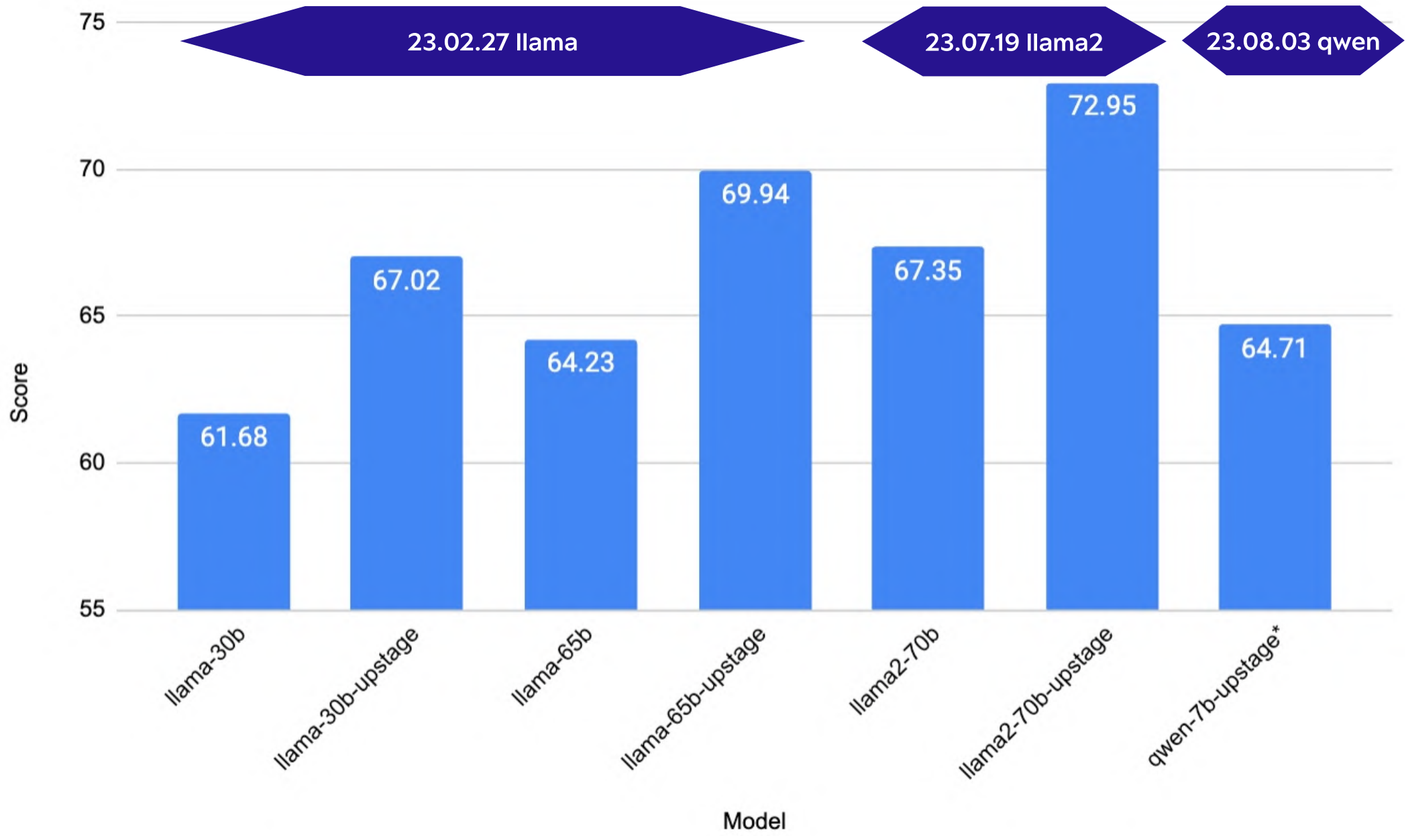
Medical Performance

	PubMedQA (ID)	MedQA-USMLE (OOD)	MedMCQA (ID)	Average
Human (pass)	60.0	50.0		
Human (expert)	78.0	87.0	90.0	85.0
InstructGPT 175B	73.2	46.0	44.0	54.4
ChatGPT	63.9	**57.0**	44.7	55.2
LLaMA 7B	5.2	27.1	24.3	18.9
LLaMA 33B	1.8	43.4	30.3	25.2
Task-tuned LLaMA 7B (Full)	**75.1**	44.5	49.9	56.5
Task-tuned LLaMA 33B (LoRA)	74.0	51.3	**50.2**	**58.5**

The LLaMA 33B (LoRA) performance is achieved with only **~16h** finetuning on the training split of PubMedQA and MedMCQA with a single 8 * A100 server. For more performance, including instruction tuning results, please refer to our Documentation.

성능이 과연 나올까?

OpenLLM 발전 속도 : 5개월 만에 동일 성능을 보장하며 1/10 다운사이징

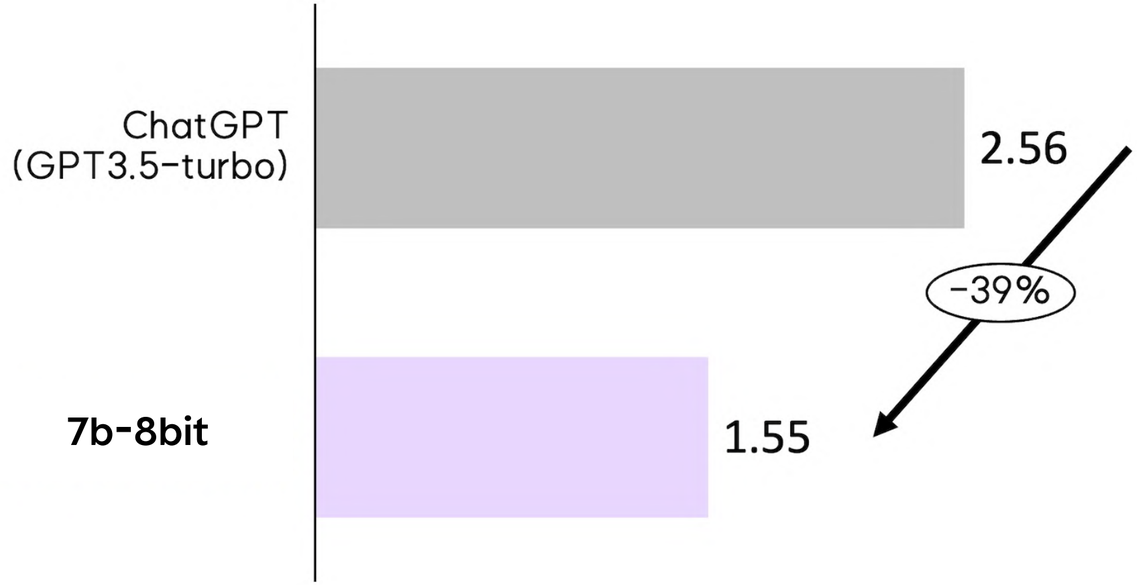


Qwen 7B (23.08.03)
- 알리바바가 공개한 Foundation Model
- 상업적으로 이용 가능
- 한국어 토큰 효율이 llama2의 두 배

얼마나 비용 절감할 수 있을까?

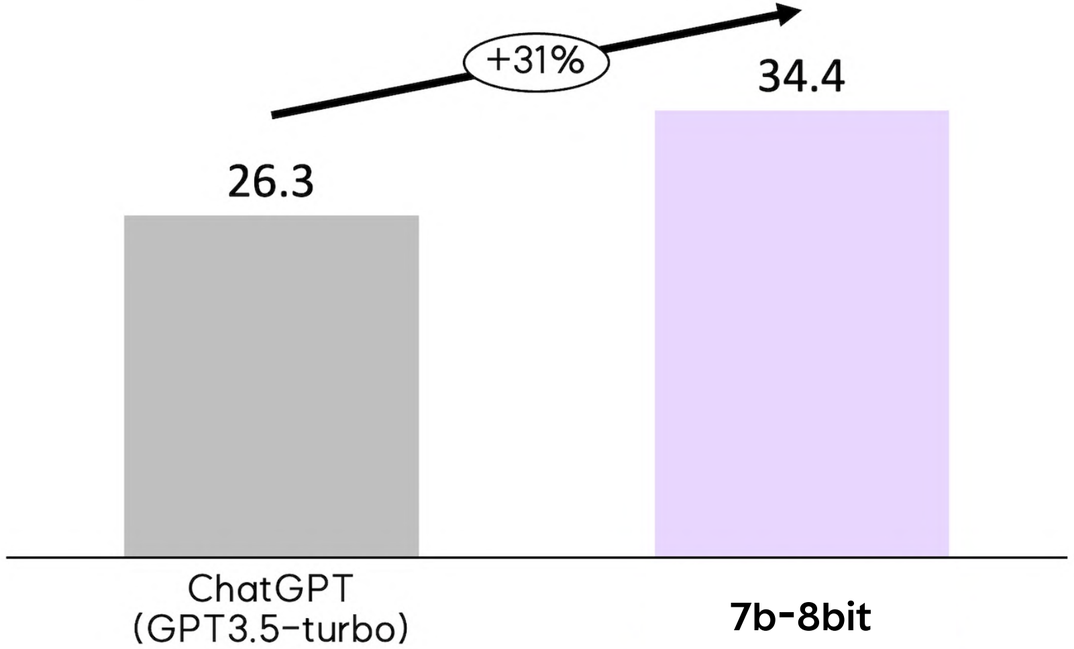
Cost saving[1]

KRW per 1000 tokens



Service speed enhancement[2]

Tokens per second



In actual service environments, the amount saved gap will be even greater by APE, continuous training, FMOps and etc.

1) Based on A100 GPU, 3 years TCO

2) Comparing ChatGPT API vs. Upstage IDC (A100 GPU)

“It's easier to invent the future than to predict it.”

Alan Kay

감사합니다