LLM 기술 활용

24.04.15(월)

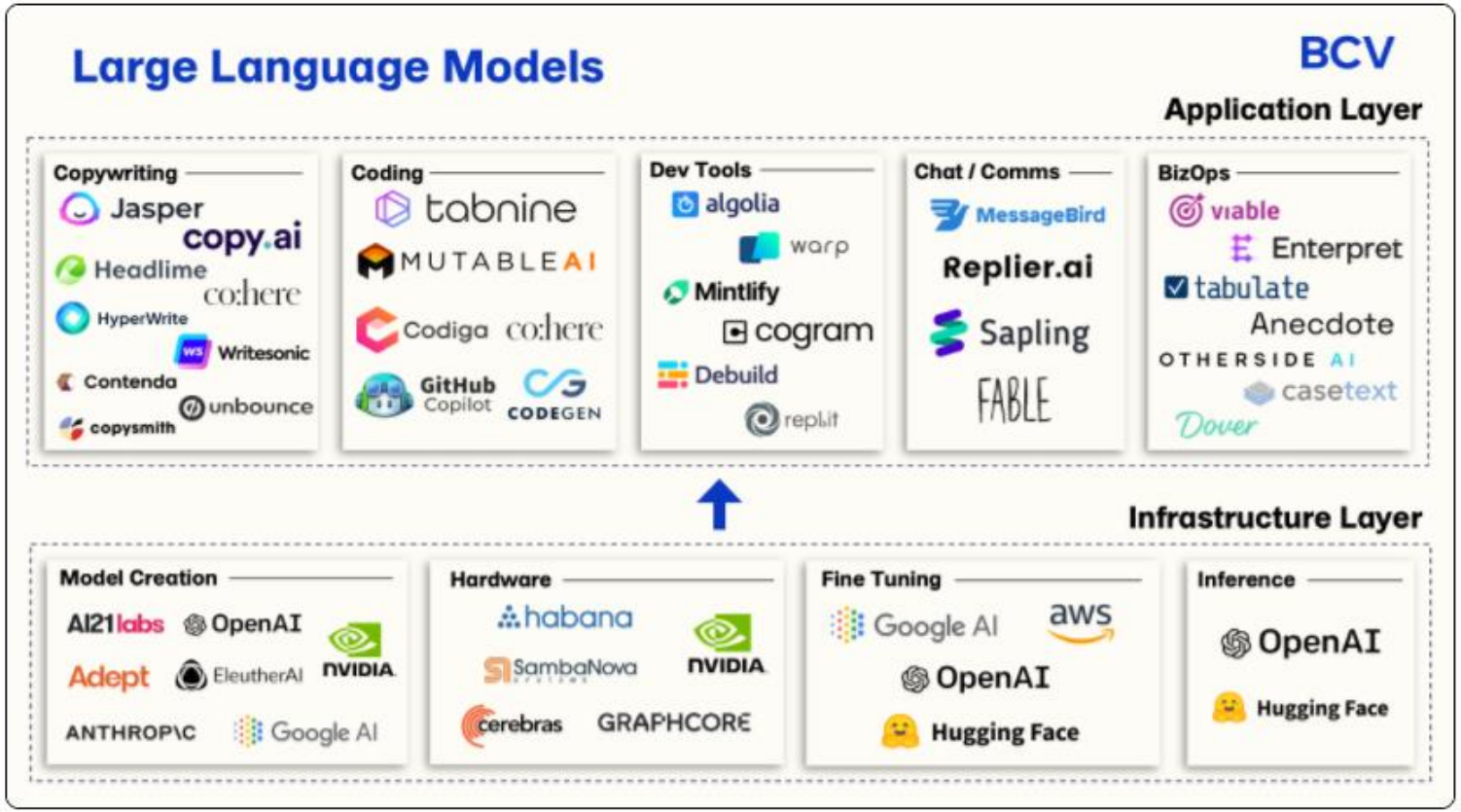
김동억

Contents

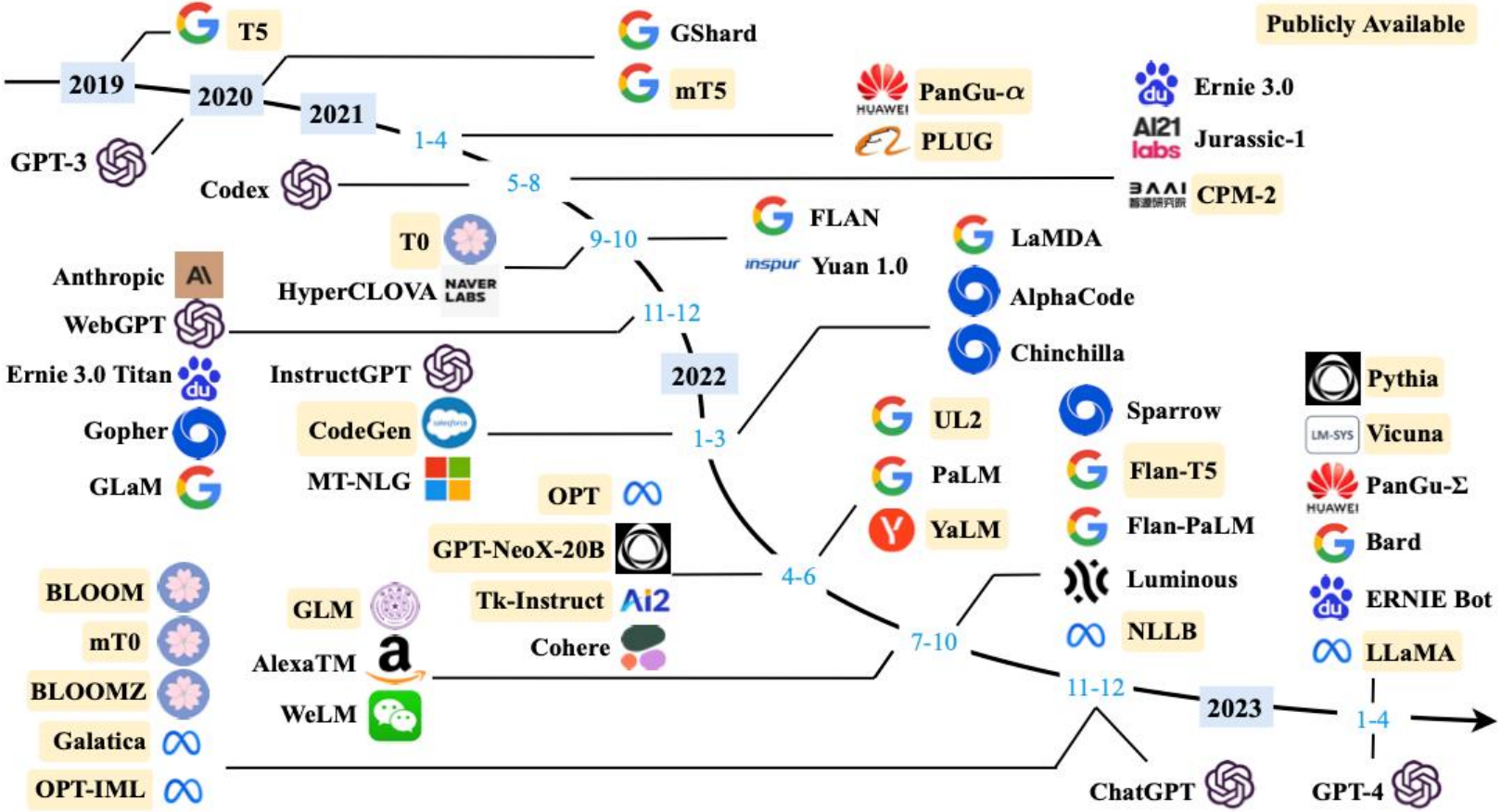
- LLM 기술 -

- AI 흐름
- 한계점
- 보완책
- 구현
- 시연

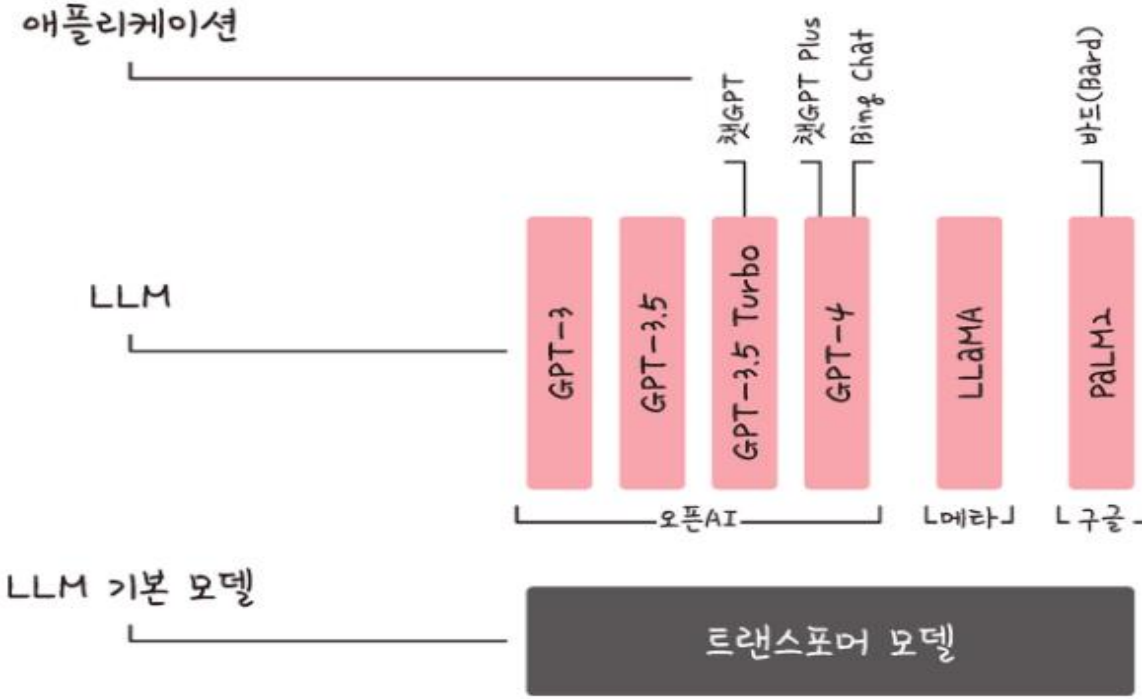
GAI(Generative Artificial Intelligence)



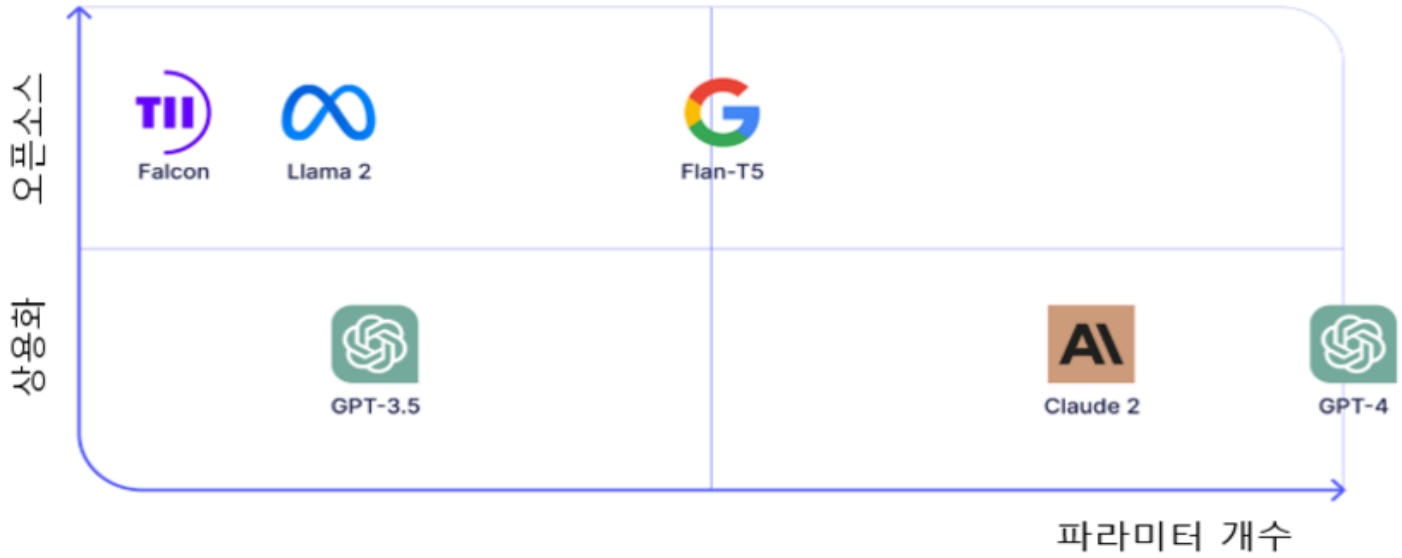
GAI History, 2019~2023



대표 LLM


애플리케이션
챗GPT
챗GPT Plus
Bing Chat
바드(Bard)
LLM
GPT-3
GPT-3.5
GPT-3.5 Turbo
GPT-4
LLaMA
PaLM2
오픈AI
메타
구글
LLM 기본 모델
트랜스포머 모델

개방성 vs 모델 크기



대표 LLM 성능 비교

	Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku	GPT-4	GPT-3.5	Gemini 1.0 Ultra	Gemini 1.0 Pro
Undergraduate level knowledge *MMLU*	**86.8%** 5 shot	79.0% 5-shot	75.2% 5-shot	86.4% 5-shot	70.0% 5-shot	83.7% 5-shot	71.8% 5-shot
Graduate level reasoning *GPQA, Diamond*	**50.4%** 0-shot CoT	40.4% 0-shot CoT	33.3% 0-shot CoT	35.7% 0-shot CoT	28.1% 0-shot CoT	—	—
Grade school math *GSM8K*	**95.0%** 0-shot CoT	92.3% 0-shot CoT	88.9% 0-shot CoT	92.0% 5-shot CoT	57.1% 5-shot	94.4% Maj1@32	86.5% Maj1@32
Math problem-solving *MATH*	**60.1%** 0-shot CoT	43.1% 0-shot CoT	38.9% 0-shot CoT	52.9% 4-shot	34.1% 4-shot	53.2% 4-shot	32.6% 4-shot
Multilingual math *MGSM*	**90.7%** 0-shot	83.5% 0-shot	75.1% 0-shot	74.5% 8-shot	—	79.0% 8-shot	63.5% 8-shot
Code *HumanEval*	**84.9%** 0-shot	73.0% 0-shot	75.9% 0-shot	67.0% 0-shot	48.1% 0-shot	74.4% 0-shot	67.7% 0-shot
Reasoning over text *DROP, F1 score*	**83.1** 3-shot	78.9 3-shot	78.4 3-shot	80.9 3-shot	64.1 3-shot	82.4 Variable shots	74.1 Variable shots
Mixed evaluations *BIG-Bench-Hard*	**86.8%** 3-shot CoT	82.9% 3-shot CoT	73.7% 3-shot CoT	83.1% 3-shot CoT	66.6% 3-shot CoT	83.6% 3-shot CoT	75.0% 3-shot CoT
Knowledge Q&A *ARC-Challenge*	**96.4%** 25-shot	93.2% 25-shot	89.2% 25-shot	96.3% 25-shot	85.2% 25-shot	—	—
Common Knowledge *HellaSwag*	**95.4%** 10-shot	89.0% 10-shot	85.9% 10-shot	95.3% 10-shot	85.5% 10-shot	87.8% 10-shot	84.7% 10-shot

LLM 서비스 시 고려해야 하는 요소

LLM의 한계

편향과 공정성

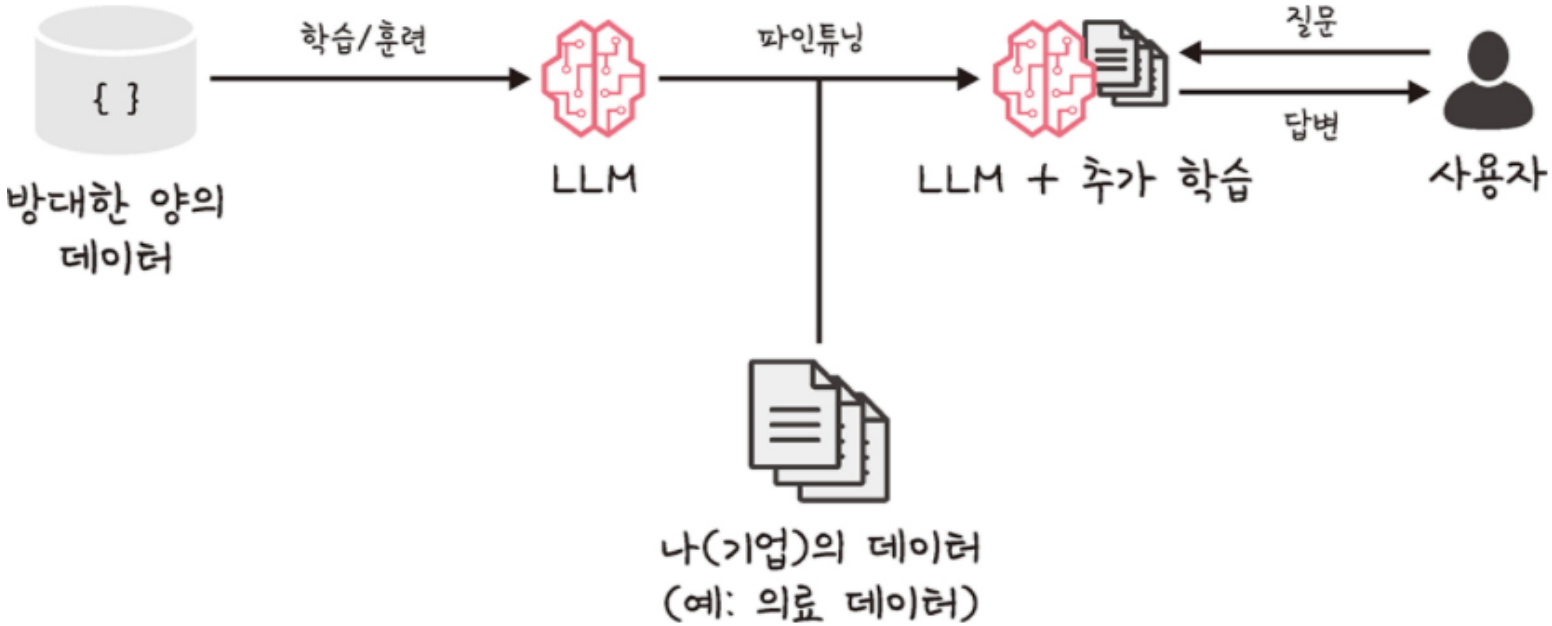
오류 가능성

개인정보 보호

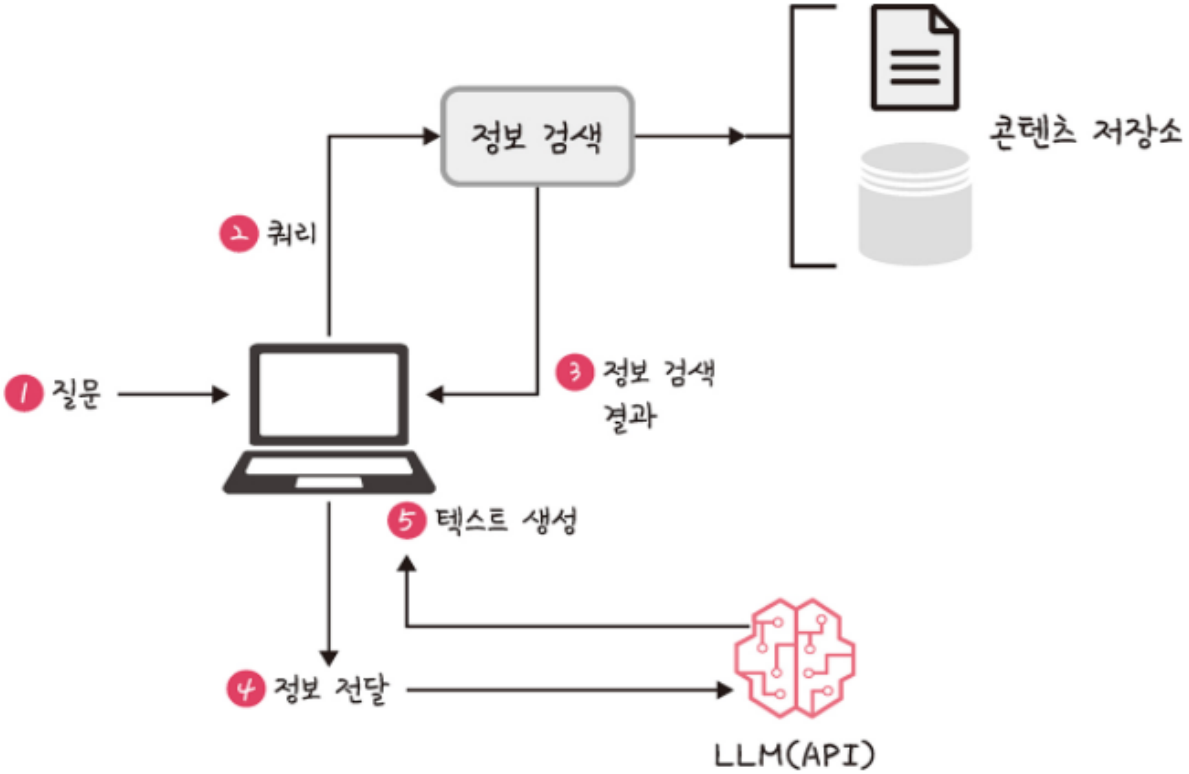
새로운 정보의 결여

기업 내 데이터 미활용

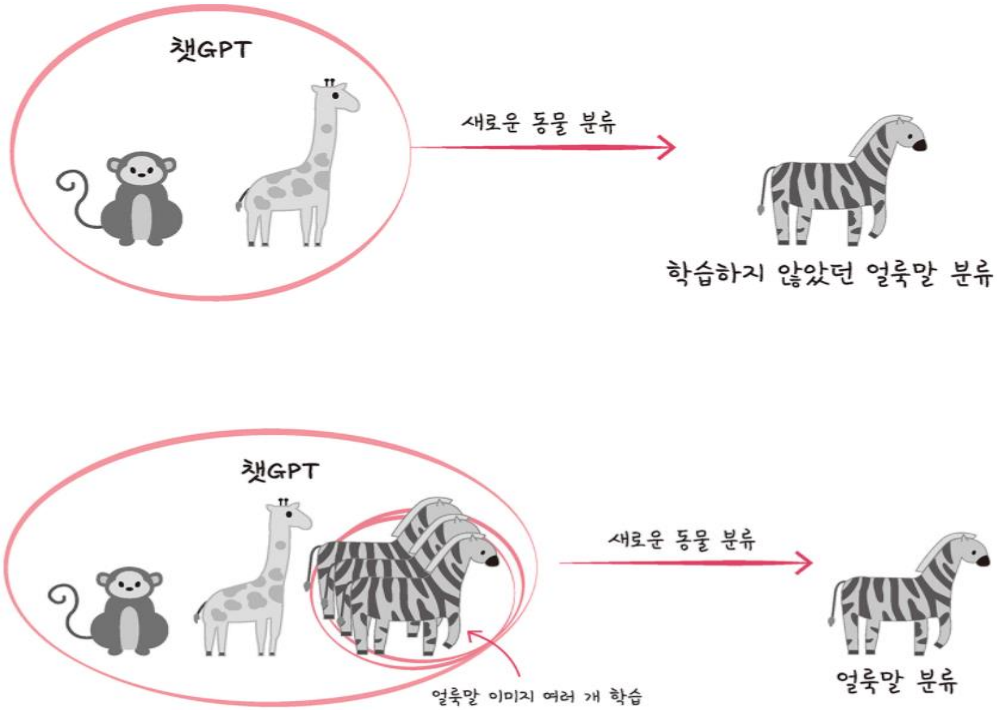
Fine Tuning



RAG(Retrieval-Augmented Generation)



Few-Shot Learning feat. Zero-Shot, One-Shot



보완책

프롬프트의 구성 요소(1)

작업지침

- 수행해야하는 작업에 대한 명확하고 간결한 지시
- 프롬프트의 작성 목적(“쉽게 요약해줘”)
- 작업의 참고 범위(“[Text]를 보고”)
- 제약조건(“5문단 이내로”)

컨텍스트

- 보다 정확하고 일관된 답변 생성의 핵심.
- 적절한 응답을 생성하는 것에 아주 큰 도움, 할루시네이션 감소
- 도메인 관련 지식, 노하우와 같은 세부 정보(“초등학생이 이해할 수 있는 쉬운 단어 사용”)
- 용어 정의 및 설명(“대미 무역: 미국을 대상으로 하는 무역”)
- 개인화 정보(“수강생 A는 ~~에 관심이 많습니다”)

보완책

프롬프트의 구성 요소(2)

페르소나

- 답변의 어조 및 스타일을 형성하는 데 중요한 역할
- 현재 상황(“너는 초등학생에게 개인 교습을 해주고 있어”)
- 직업(“너는 어려운 내용을 쉽게 알려주는 초등학교 선생님이야”)
- 어조(“친구에게 대하듯 편하게 말해줘”)

시작 단어 및 구문

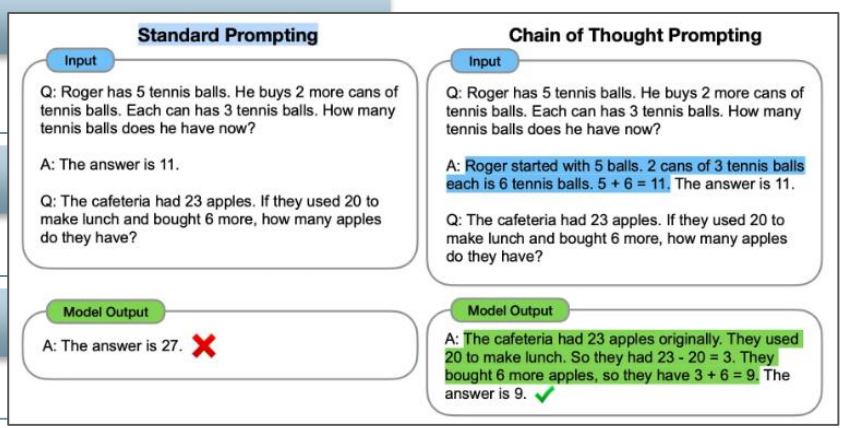
- 답변의 형식과 내용을 결정하는데 중요한 역할

예제

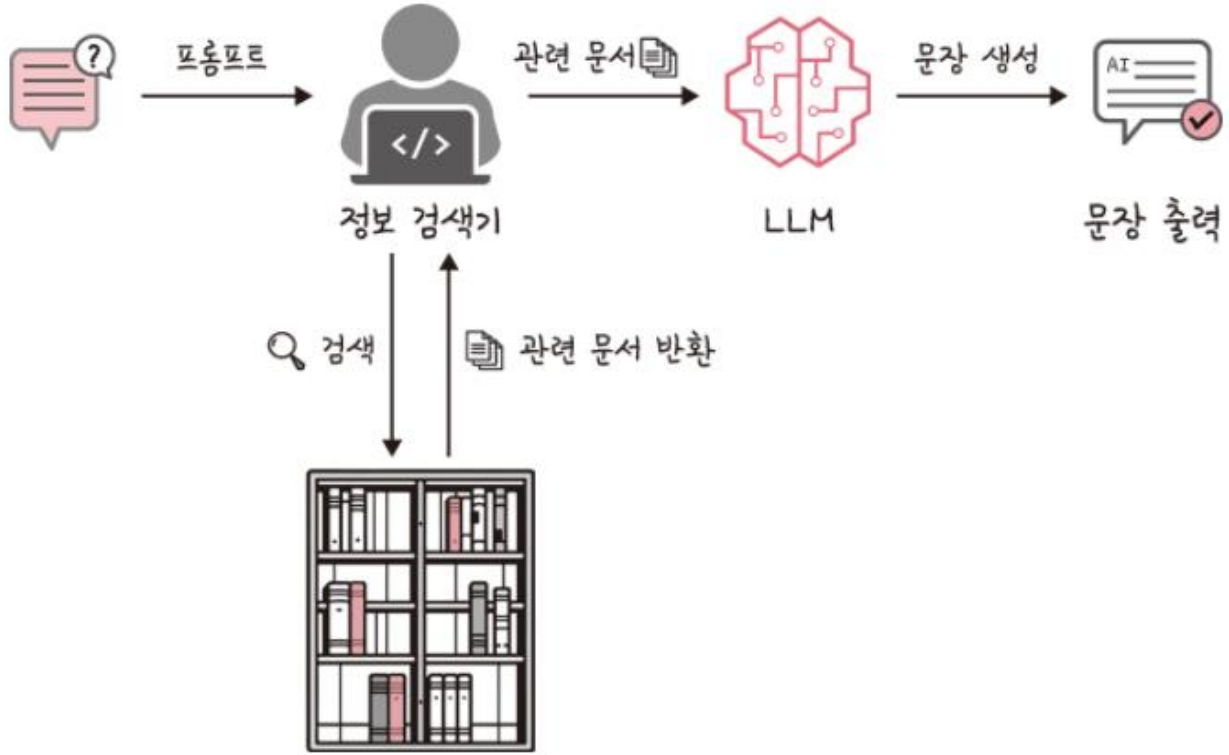
- Zero-Shot, One-Shot, Few-Shot

Structured LLM Reasoning

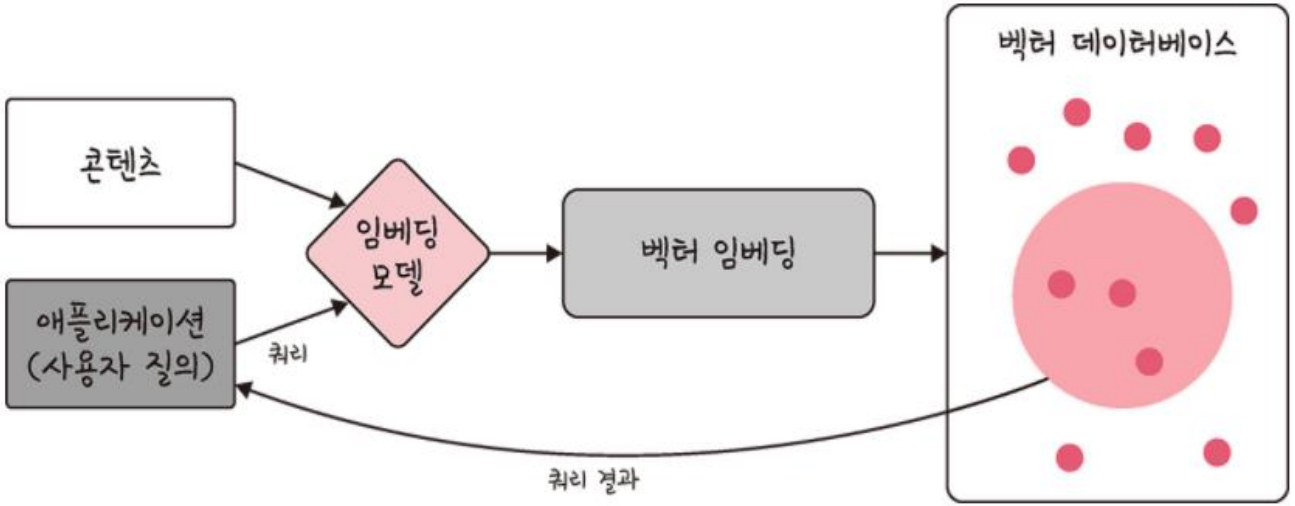
- CoT, ToT, GoT, …



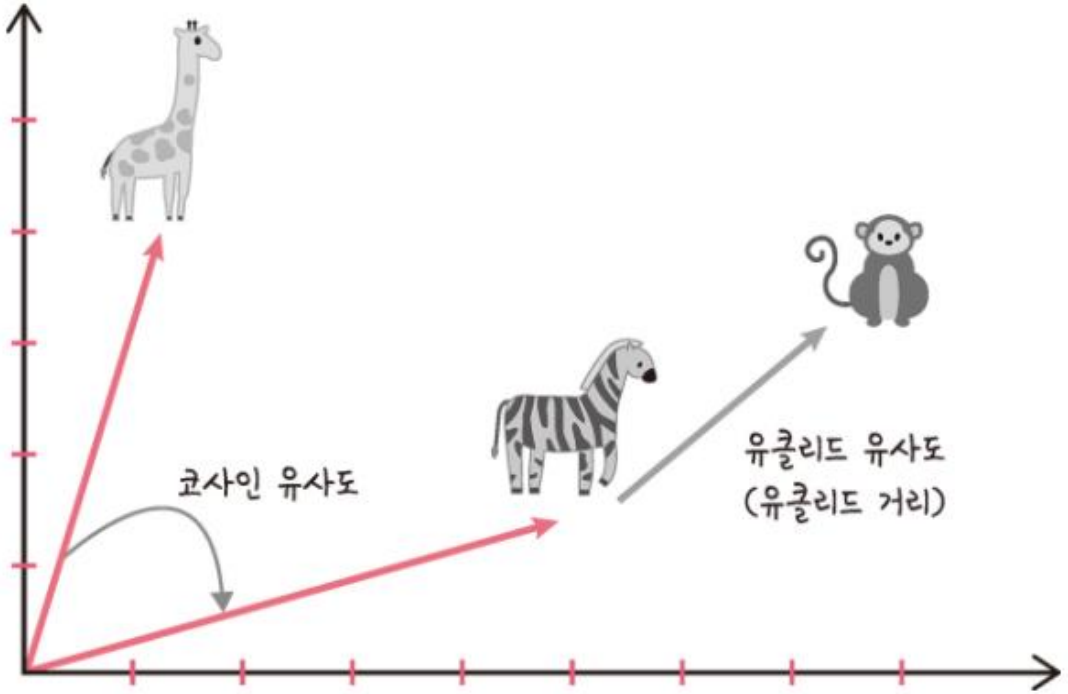
RAG 구현 방법



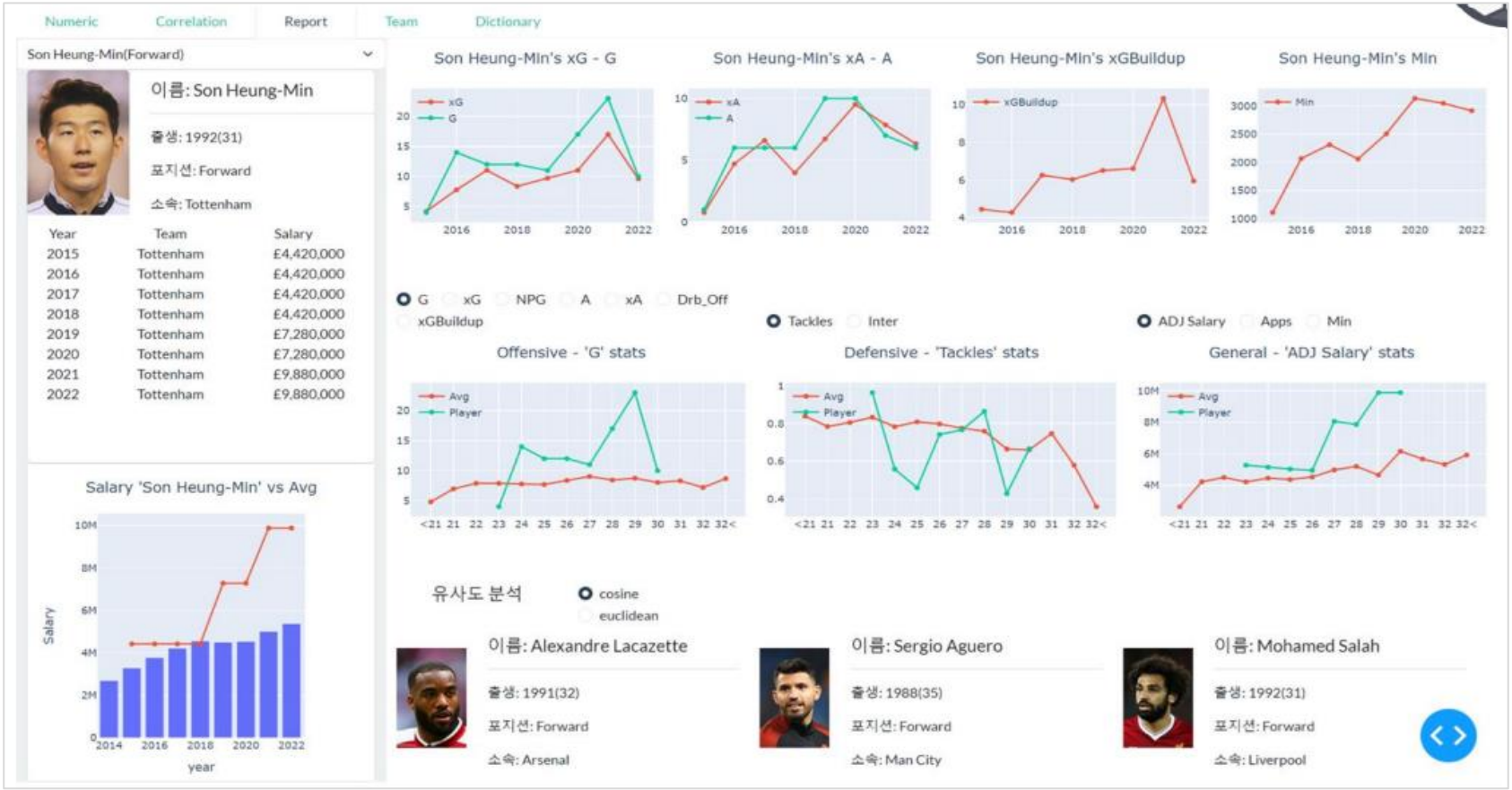
RAG 구현 상세



RAG 동작 원리

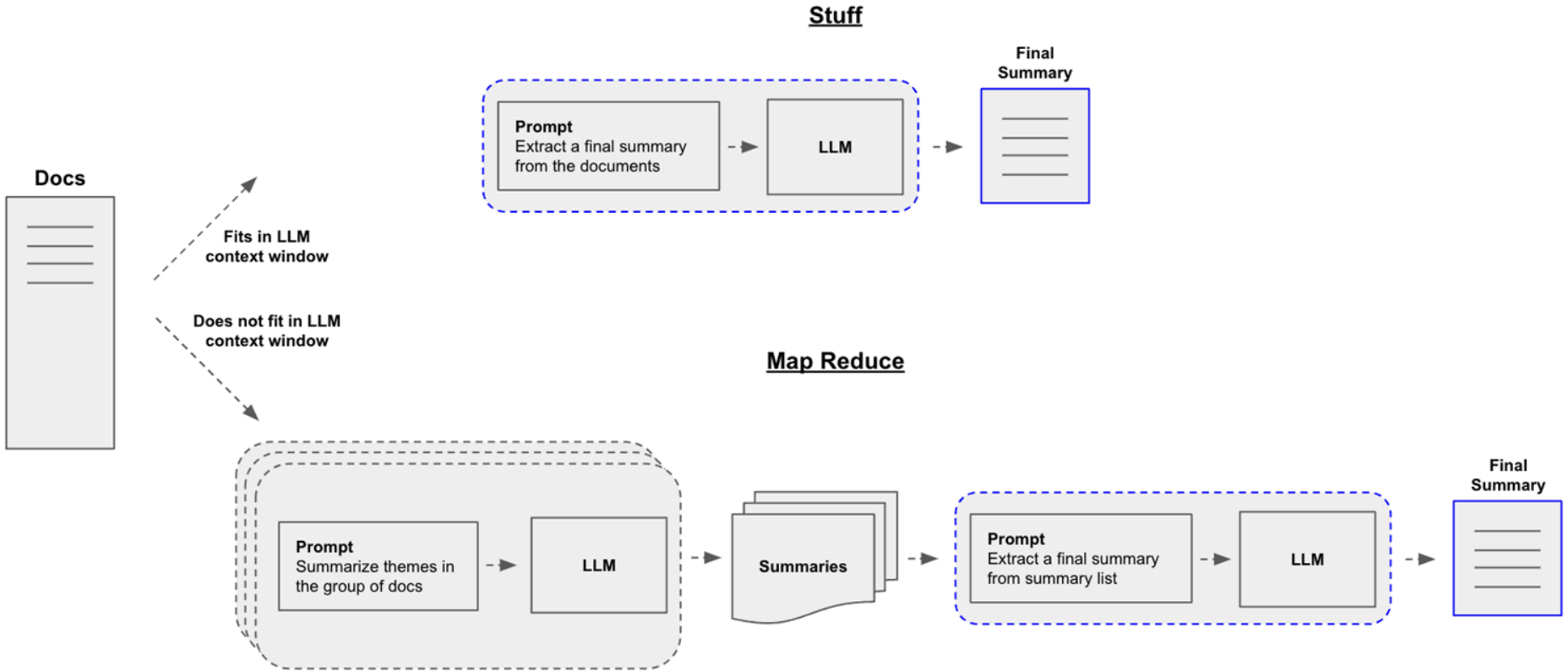


유사도 예제

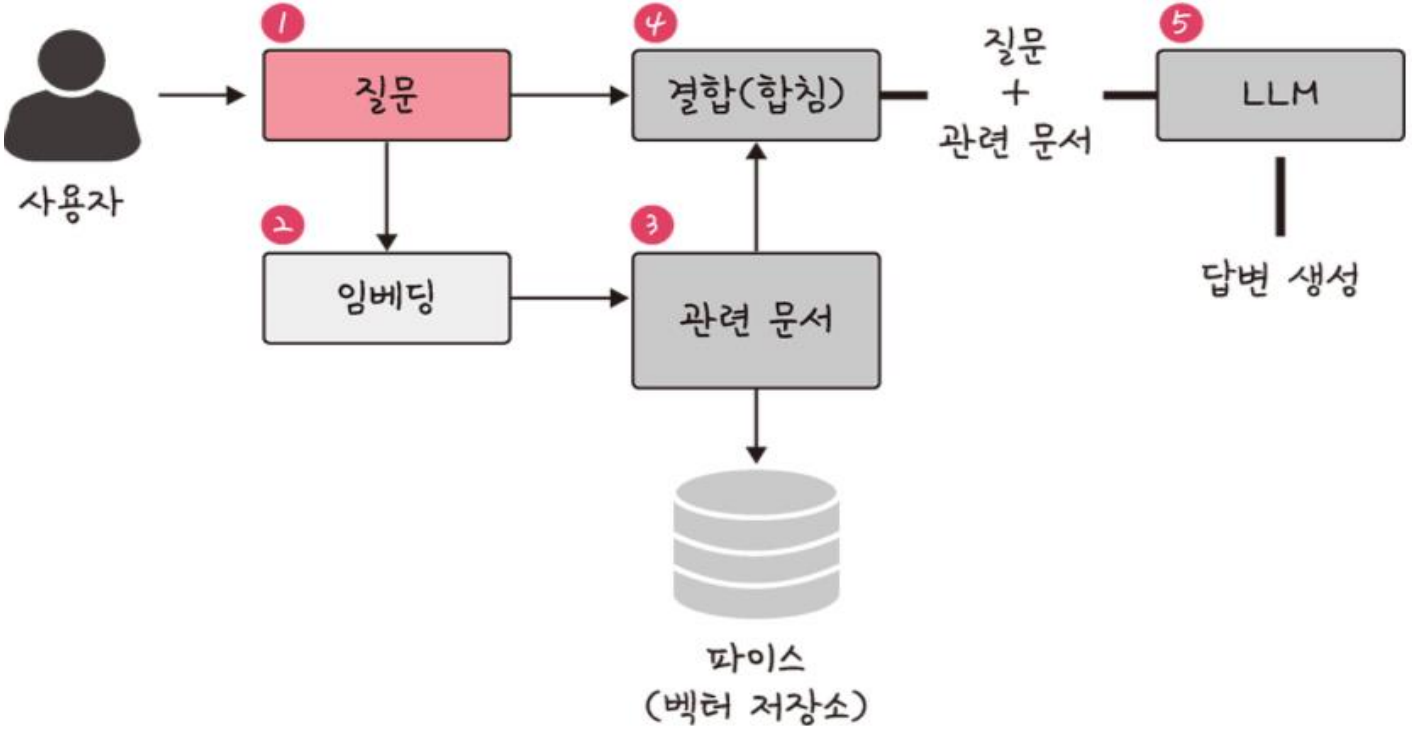

Numeric
Correlation
Report
Team
Dictionary
Son Heung-Min(Forward)
이름: Son Heung-Min
출생: 1992(31)
포지션: Forward
소속: Tottenham
Year Team Salary
2015 Tottenham £4,420,000
2016 Tottenham £4,420,000
2017 Tottenham £4,420,000
2018 Tottenham £4,420,000
2019 Tottenham £7,280,000
2020 Tottenham £7,280,000
2021 Tottenham £9,880,000
2022 Tottenham £9,880,000
Salary 'Son Heung-Min' vs Avg
Salary
year
Son Heung-Min's xG - G
Son Heung-Min's xA - A
Son Heung-Min's xGBuildup
Son Heung-Min's Min
G xG NPG A xA Drb_Off
xGBuildup
Tackles Inter
ADJ Salary Apps Min
Offensive - 'G' stats
Defensive - 'Tackles' stats
General - 'ADJ Salary' stats
Avg
Player
유사도 분석
cosine
euclidean
이름: Alexandre Lacazette
출생: 1991(32)
포지션: Forward
소속: Arsenal
이름: Sergio Aguero
출생: 1988(35)
포지션: Forward
소속: Man City
이름: Mohamed Salah
출생: 1992(31)
포지션: Forward
소속: Liverpool

LangChain 구성

Summarize, feat. Stuff vs Map Reduce



Chatbot with RAG



QnA