

RAG 를 사용해야 하는 이유



RAG(Retrieval-Augmented Generation) - 검색, 증강, 생성

RAG 기술이 주목받고 있는 이유

AS-IS

먼저, 현재의 ChatGPT 가 가질 수 있는 문제점에 대해 정리하면 다음과 같습니다.

1. **최신 정보**에 대하여 학습되어 있지 않다.
2. 나(개인) 혹은 우리 회사에 제한되어 있는 **내부데이터**에 대한 학습이 되어 있지 않다.
3. 따라서, 특정 도메인(나의 개인정보, 회사의 내부 정보)에 대한 질문을 하면 기대하는 답변을 얻을 수 없다.
4. 문서화 시켜 업로드를 ChatGPT 에서 질의할 수 있지만, 기대하는 답변을 받을 수 없거나, **할루시네이션(환각)** 현상 이 발생한다. 게다가 문서의 양이 많아지면 더욱 더 이러한 현상은 심해진다.

TO-BE

적합한 RAG를 적용했을 때는,

1. **최신 정보**를 기반으로 답변할 수 있으며, 정보를 찾을 수 없는 경우 "검색" 기능을 활용하여 답변할 수 있다.
2. 나(개인) 혹은 우리 회사에 제한되어 있는 **내부데이터**를 참고하여 답변할 수 있다.
3. 문서를 내부 DB 에 저장할 수 있고, DB 에 내용을 축적해 나갈 수 있으며, 저장된 DB 에서 원하는 정보를 검색하여 검색된 정보를 바탕으로 답변할 수 있

다.

4. 답변에 대한 출처를 역으로 저장되어 있는 DB 에서 검색 후 검증하는 방식으로 할루시네이션 현상을 줄일 수 있다.

궁극적으로 더 나은 답변 품질을 기대할 수 있으며, 방대한 지식 기반으로 답변하는 도메인 특화 챗봇을 생성하는 것이 가능합니다.

예시)

서울특별시에 사는 "테디" 의 아버지 이름은 뭐야?

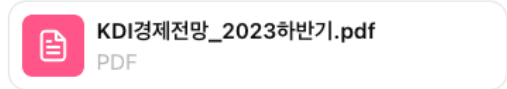
가족관계증명서 (상세)			
등록기준지	서울특별시		
구분	성명	출생연월일	주민등록번호
본인	테디 ()	년 월 일	-
가족사항			
구분	성명	출생연월일	주민등록번호
부	폴		
모	설리		
자녀	()		
자녀	()		

ChatGPT 에 내장되어 있는 RAG?

ChatGPT 에 문서를 업로드 하여 질문을 하게 되면, 업로드 된 문서를 기반으로 답변을 하게 됩니다.

하지만, ChatGPT 는 RAG 의 전반적인 과정을 블랙박스 공개하고 있지 않기 때문에 어떠한 과정으로 RAG 가 일어나는지 알 수 없습니다.

아래는 하나의 사례입니다.



이 문서에 대한 내용을 알려줘. 한글로 작성해줘

< 2/2 >



이 문서는 한국개발연구원(KDI)에서 발간한 2023년 하반기 경제전망 보고서입니다. 이 보고서는 국내외 경제 동향을 분석하고 중단기 경제 전망을 제공하며, 경제 전망에 대한 불확실성 요인과 대책을 연구하여 정부의 경제 정책 수립과 기업의 경영, 학계의 연구 활동에 도움을 주기 위해 작성되었습니다.

보고서의 주요 내용은 다음과 같습니다:

1. 현재의 경제 상황과 내년의 경제 전망을 다루며, 고금리 기조의 영향, 수출 회복세, 내수 둔화, 물가 상황, 취업자 수 변화 등에 대해 다룹니다.
2. 2024년 국내경제 전망에서는 완만한 회복세를 예측하며, 성장률, 소비자 물가, 취업자 수, 경상수지 등 다양한 경제 지표에 대해 예측합니다.
3. 재정정책, 통화정책, 금융정책 등의 정책 방향을 제시합니다.
4. 위험 요인으로는 국제유가 변동, 중국의 경제 상황 변화 등 대외 불확실성이 높은 점을 지적하며, 이에 대한 대응 방안을 모색합니다.

이 보고서는 정부 및 기업의 정책 결정에 중요한 정보를 제공하며, 경제 상황에 대한 깊은 이해와 미래 전망에 대한 지침을 담고 있습니다.

하지만, 문서 내부의 구체적인 질문에 대해서는 제대로된 답변을 못하거나, 관련 정보를 찾지 못하는 현상이 발생

3/4분기 유로존의 전체 및 청년실업률을 알려줘. 문서의 내용을 바탕으로 알려줘.

< 2/2 >



보고서에서는 유로존의 전체 및 청년 실업률에 대한 구체적인 데이터를 제공하지 않습니다. 이 문서는 한국의 경제 상황과 전망에 초점을 맞추고 있어, 유로존의 실업률과 같은 특정 국제 경제 지표에 대한 세부 수치는 포함되어 있지 않습니다. 유로존의 실업률 정보가 필요하시다면, 유럽연합 통계국(Eurostat)의 공식 웹사이트에서 최신 데이터를 찾으실 수 있을 것입니다.

🔊 📄 ↺ 🔒 ✨

II. 세계경제 동향

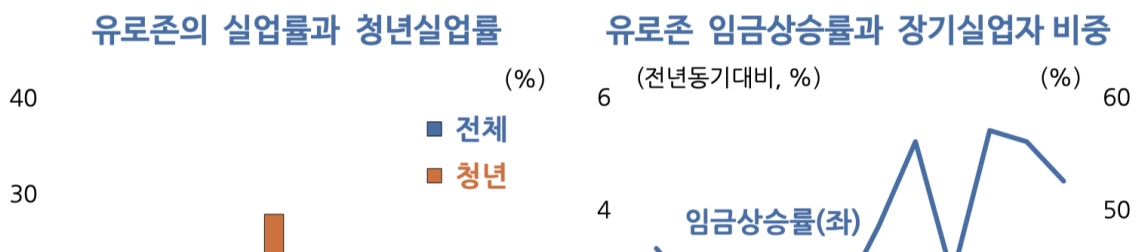
■ 실업률이 낮은 수준을 지속하는 등 노동시장 관련 지표는 양호한 흐름을 유지

- 3/4분기 유로존의 전체 및 청년실업률은 전분기에 이어 각각 6.5%, 14.0%의 낮은 수준을 기록

- 경기둔화에도 불구하고 노동 공급 부족 현상이 지속되면서 실업률은 주요국에서 큰 변동 없이 낮은 수준을 유지함.

- 2/4분기 임금상승률이 높은 수준을 지속하는 가운데, 장기실업자(전체 실업자 대비 12개월 이상 장기실업자) 비중은 하향 안정세를 유지

- 명목 임금상승률은 전분기(5.2%)보다 하락(4.5%)하였으나, 지정학적 위험에 따른 국제유가 상승 등 물가 상방 압력도 존재

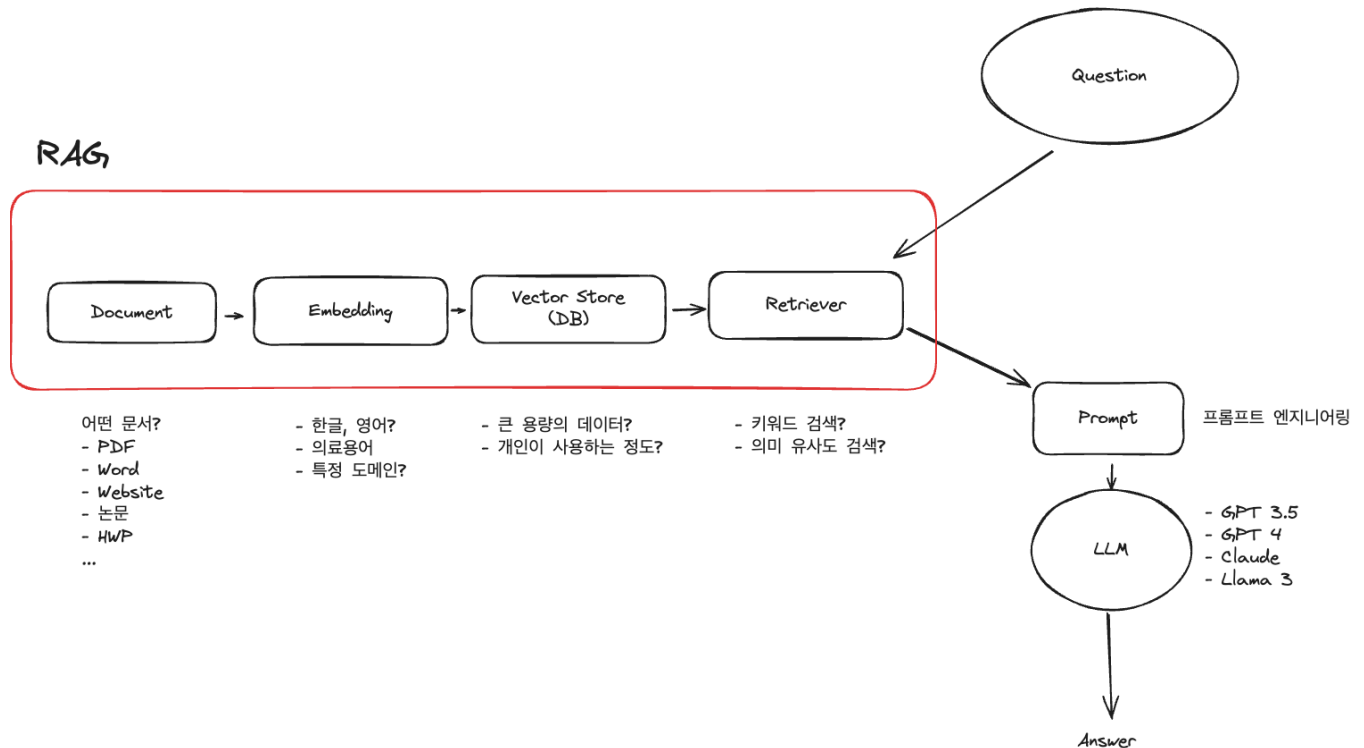


ChatGPT의 RAG 과정은 공개되어 있지 않고, 우리가 컨트롤할 수 없는 부분이기 때문에
우리가 할 수 있는 유일한 최선은 "문서"를 ChatGPT가 잘 검색할 수 있는 형태로
변경하는 것입니다.

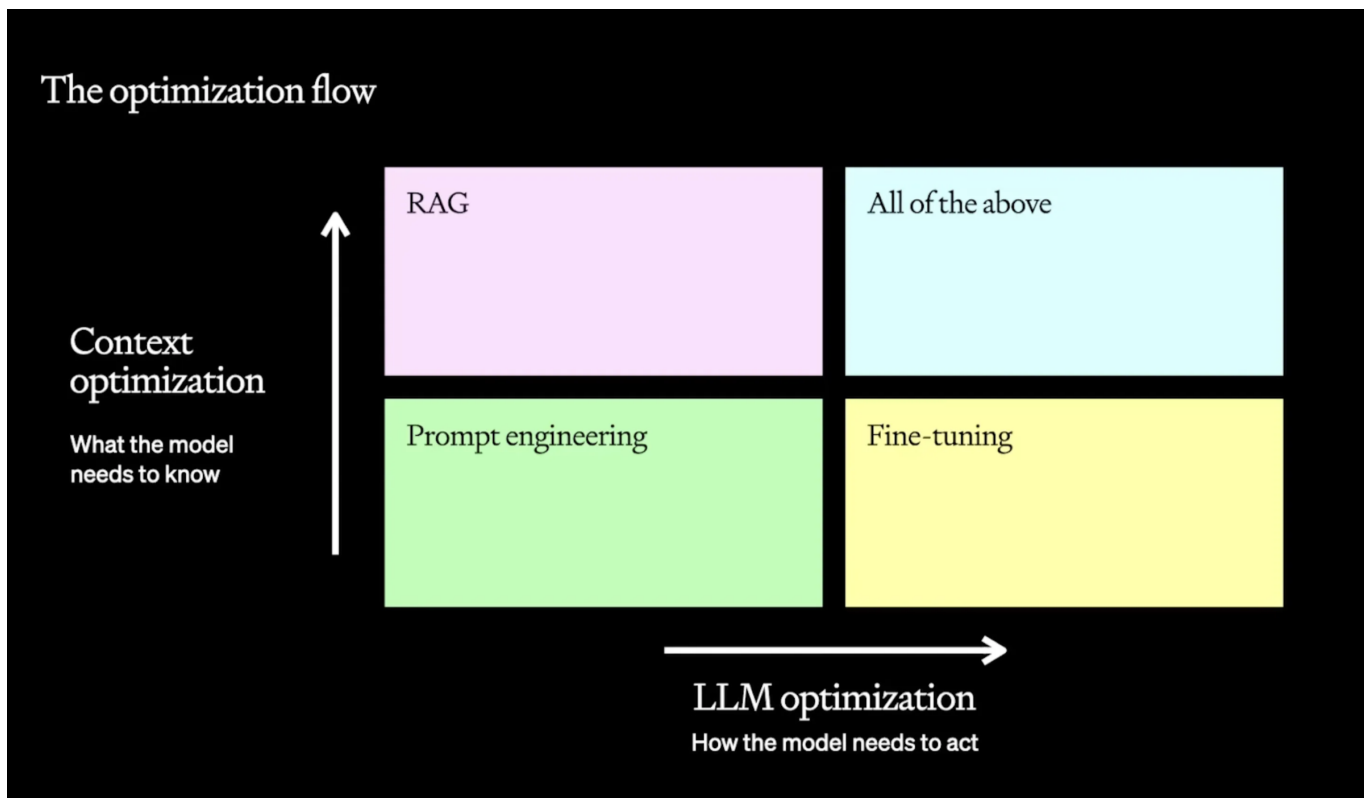
하지만, ChatGPT 가 문서 검색을 잘 하도록 우리가 가지고 있는 문서의 형태를 모두 변경하는 것은 사실상 어려운 일입니다.

RAG 프로세스

RAG 프로세스 이해

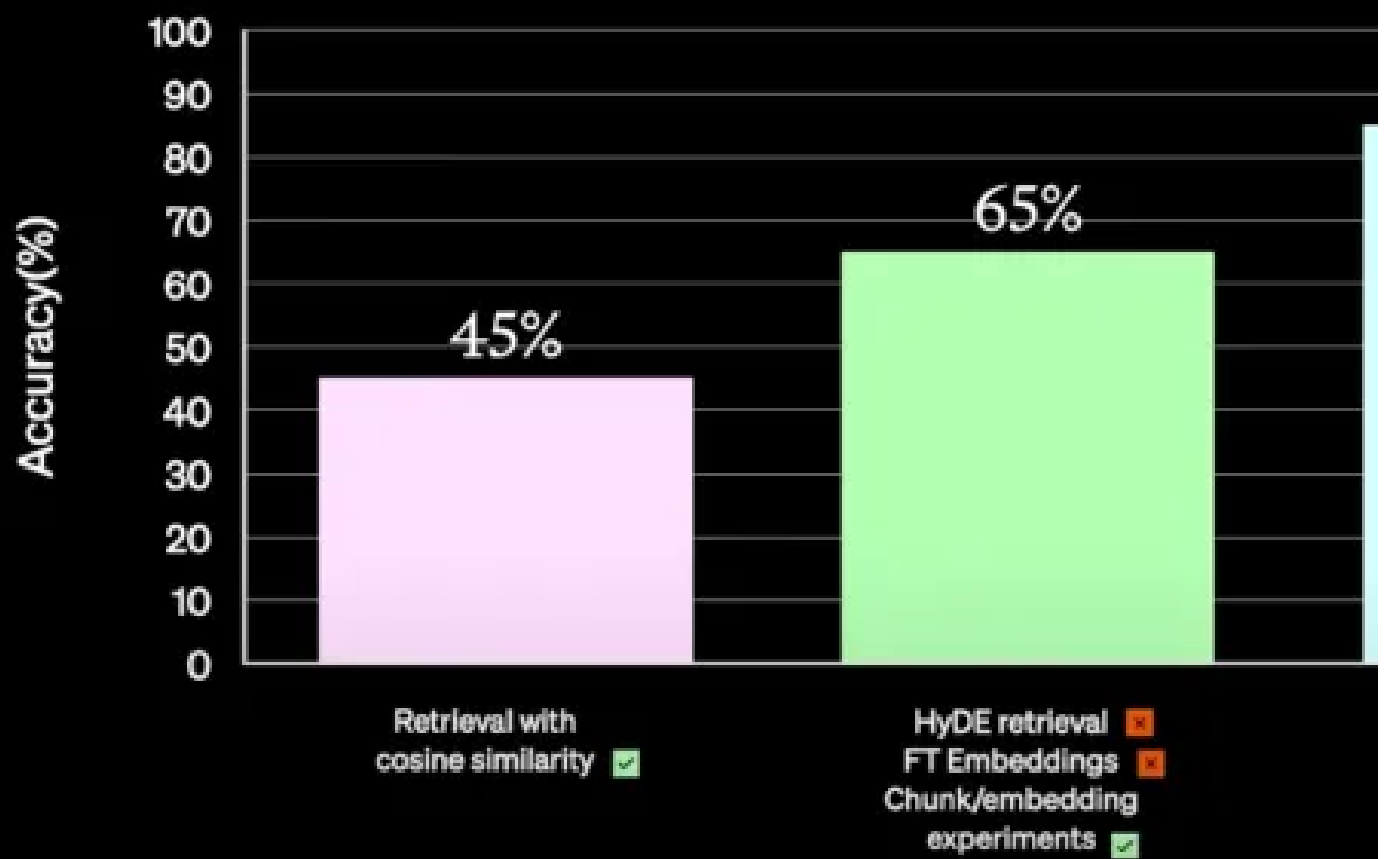


출처: OpenAI Dev Day

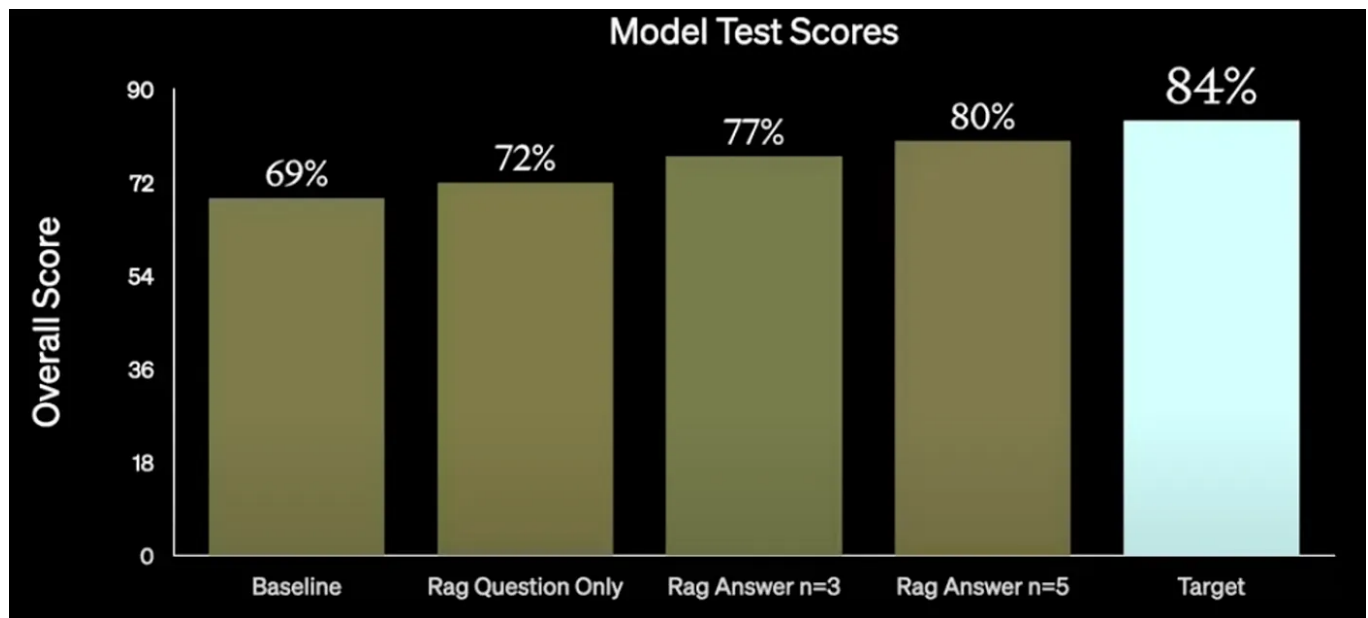


RAG 에 다양한 기법 적용시 점진적 성능 향상

출처: OpenAI Dev Day



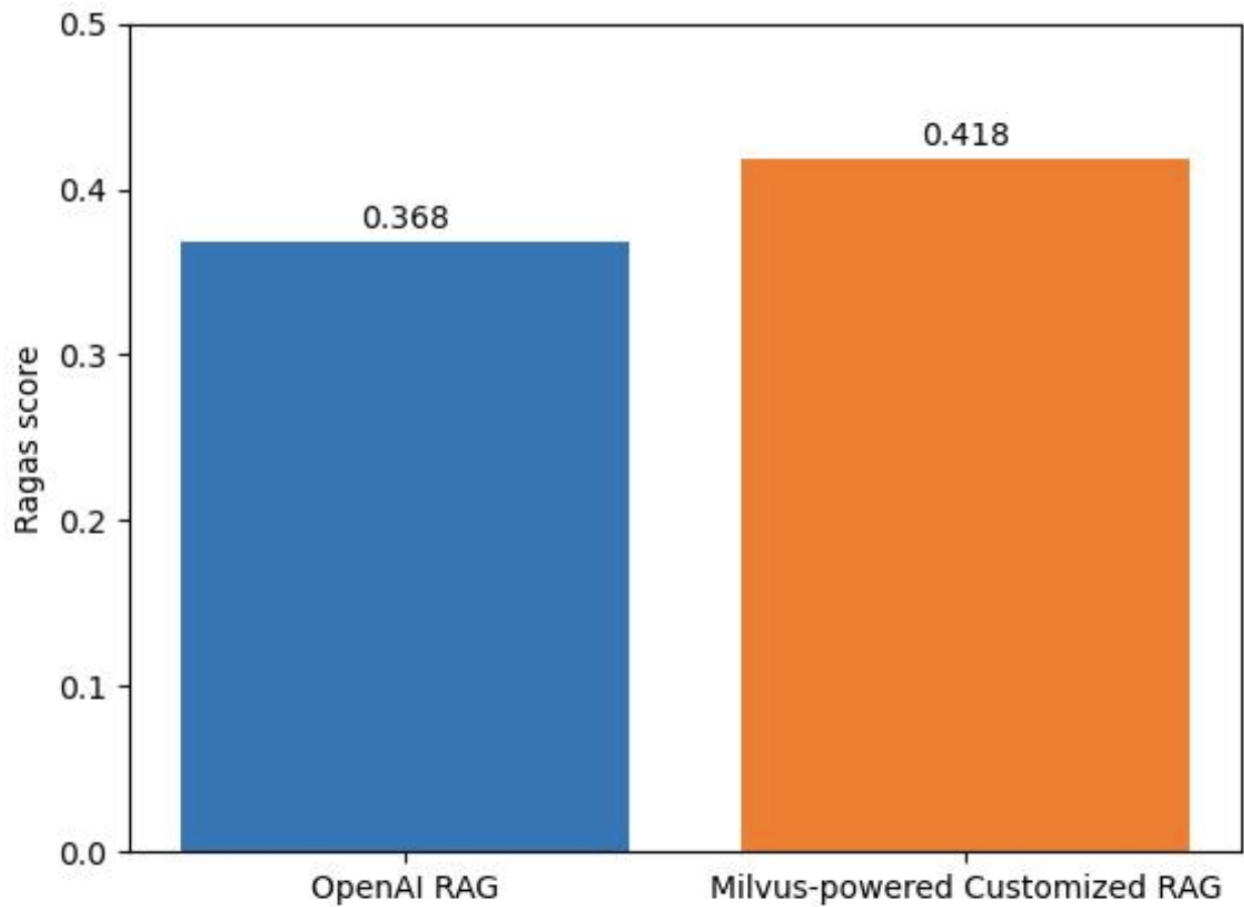
What we add



	OpenAI RAG	Customized RAG
LLM model	gpt-4-1106-preview	gpt-4-1106-preview
Vector DB	Not Disclosed	milvus
Embedding model	Not Disclosed	BAAI/bge-base-en
Chunk size	Not Disclosed	1000
Chunk overlap	Not Disclosed	40
topk	Not Disclosed	5
Use Agent	Yes	Yes

출처: <https://thenewstack.io/openai-rag-vs-your-customized-rag-which->

[one-is-better/](#)

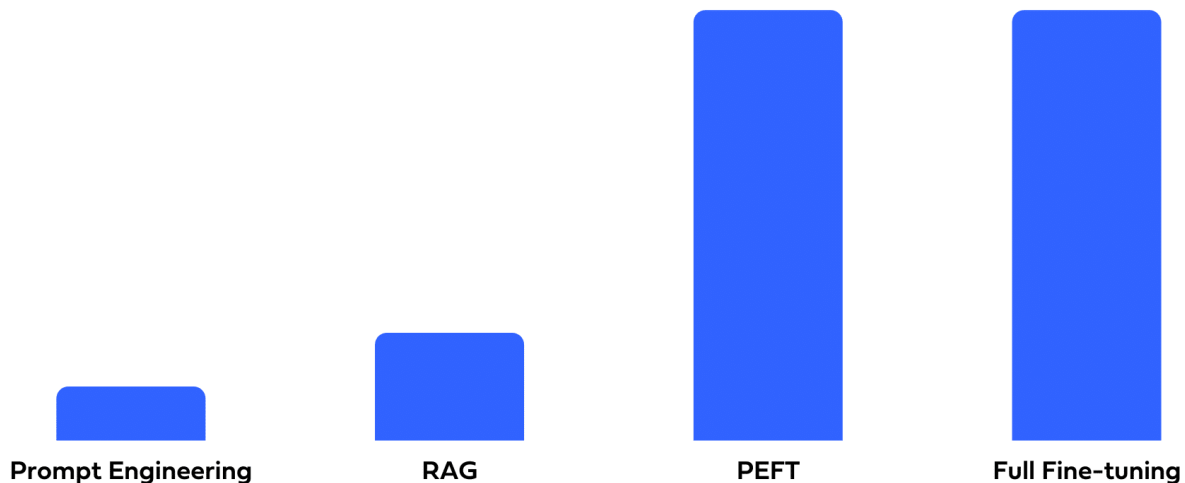


OpenAI RAG vs. Milvus-powered customized RAG on Ragas Scores



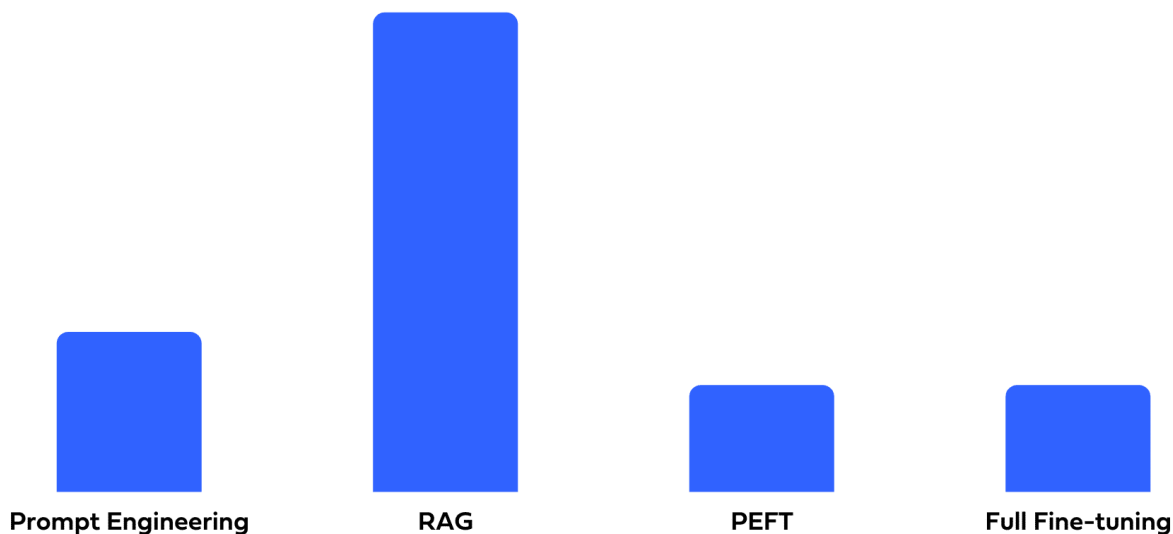
구현의 난이도가 높지 않음

Complexity of Implementation



최신 정보를 기반으로 답변

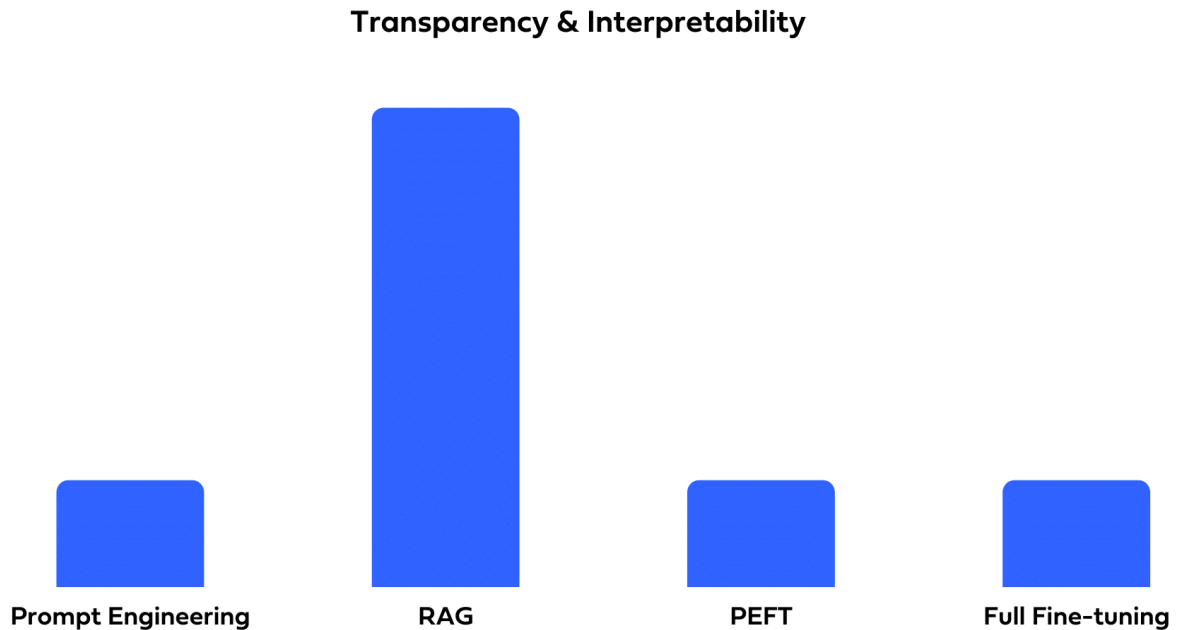
Up to Date Responses



답변에 대한 과정을 확인(투명성), 해석

우리는 RAG 의 모든 과정을 직접설계 하기 때문에 하나하나의 과정을 분석하고 컨트롤 할 수 있습니다.

잘 나온 답변에 대하여 왜 잘 나왔는지와/ 제대로된 답변을 얻지 못한다면 왜 못 얻어 왔는지 확인할 수 있습니다.



추적 예시

TRACE

Collapse Stats Most relevant ▾

RunnableSequence

🕒 1.05s 💬 1,454

Retriever 0.21s

ChatOpenAI gpt-4o 0.83s

ⓘ Some runs have been hidden. [Show 4 hidden runs](#)

RunnableSequence

Run Feedback Metadata

Input ▾

1	input: 삼성전자에서 자체 개발한 AI 의 이름은?
---	--------------------------------

YAML ⚡

Output ▾

삼성전자에서 자체 개발한 AI의 이름은 '삼성 가우스'입니다.

검색한 결과를 확인할 수 있습니다.

Retriever

Run ID
Trace ID

Run
Feedback
Metadata

Input

1
query: 삼성전자에서 자체 개발한 AI 의 이름은?

Output

DOCUMENTS 3

SPRi AI Brief | 2023-12 월호 10삼성전자, 자체 개발 생성 AI ‘삼성 가우스’ 공개 n삼성전자가 온디바이스에서 작동 가능하며...
12
data/SPRi_AI_Brief_2023년12월호_F.pdf

2024 년부터 가우스를 탑재한 삼성 스마트폰이 메타의 라마(Llama)2 를 탑재한 퀄컴 기기 및 구글 어시스턴트를 적용한 구글 ...
12
data/SPRi_AI_Brief_2023년12월호_F.pdf

2023 년 12월호 I. 인공지능 산업 동향 브리프 1. 정책/법제 > 미국, 안전하고 신뢰할 수 있는 AI 개발과 사용에 관한 행정명령 발...
1
data/SPRi_AI_Brief_2023년12월호_F.pdf

각 문서의 세부 내용을 확인할 수 있습니다.

TRACE

Collapse
Stats
Most relevant

RunnableSequence
1.05s
1,454

Retriever
0.21s

ChatOpenAI
gpt-4o
0.83s

Some runs have been hidden. [Show 4 hidden runs](#)

Retriever

Run ID
Trace ID

Run
Feedback
Metadata

1
query: 삼성전자에서 자체 개발한 AI 의 이름은?

Output

DOCUMENTS 3

SPRi AI Brief |
2023-12 월호

10삼성전자, 자체 개발 생성 AI ‘삼성 가우스’ 공개
n삼성전자가 온디바이스에서 작동 가능하며 언어, 코드, 이미지의 3개 모델로 구성된 자체 개발 생성 AI 모델 ‘삼성 가우스’를 공개
n삼성전자는 삼성 가우스를 다양한 제품에 단계적으로 탑재할 계획으로 온디바이스 작동이 가능한 삼성 가우스는 외부로 사용자 정보가 유출될 위험이 없다는 장점을 보유KEY Contents
£언어, 코드, 이미지의 3개 모델로 구성된 삼성 가우스, 온디바이스 작동 지원
n삼성전자가 2023 년 11월 8일 열린 ‘삼성 AI 포럼 2023’ 행사에서 자체 개발한 생성 AI 모델 ‘삼성 가우스’를 최초 공개
·정류분포 이론을 정립한 천재 수학자 가우스 (Gauss) 의 이름을 본뜬 삼성 가우스는 다양한 상황에 최적화된 크기의 모델 선택이 가능
·삼성 가우스는 라이선스나 개인정보를 침해하지 않는 안전한 데이터를 통해 학습되었으며, 온디바이스에서 작동하도록 설계되어 외부로 사용자의 정보가 유출되지 않는 장점을 보유
·삼성전자는 삼성 가우스를 활용한 온디바이스 AI 기술도 소개했으며, 생성 AI 모델을 다양한 제품에 단계적으로 탑재할 계획
n삼성 가우스는 △텍스트를 생성하는 언어모델 △코드를 생성하는 코드 모델 △이미지를 생성하는

RAG 는 할루시네이션을 줄이는 방법

- 유효한 정보 기반 으로 답변을 강제하거나, 답변에 대한 출처를 다시 주어진 문서로부터 찾게 하는 방식을 사용하여 할루시네이션을 줄이는데 도움을 줄 수 있습니다.

