

RAG(검색증강생성)를 이용한 야구 기사 문답 챗봇 개발



24-2 빅데이터 자연어처리 기술 Project

목차

1. 서비스 개발 목표
2. 데이터베이스 구축
3. 성능 평가
4. 시스템 구축
5. 추가 보완 사항
6. Pain Points

1. 서비스 개발 목표

KBO 에서의 GPT의 환각, 사람들이 야구에 대한 정보를 찾아 볼 때 주로 최근에 있을 경기들에 관심을 갖는 경우가 많다. 언어모델은 최근 선수들의 부상정보·컨디션을 반영하기 어렵다

ChatGPT ▾

↑ 공유하기

촬영

kbo 어느팀이 우승할 것 같아?

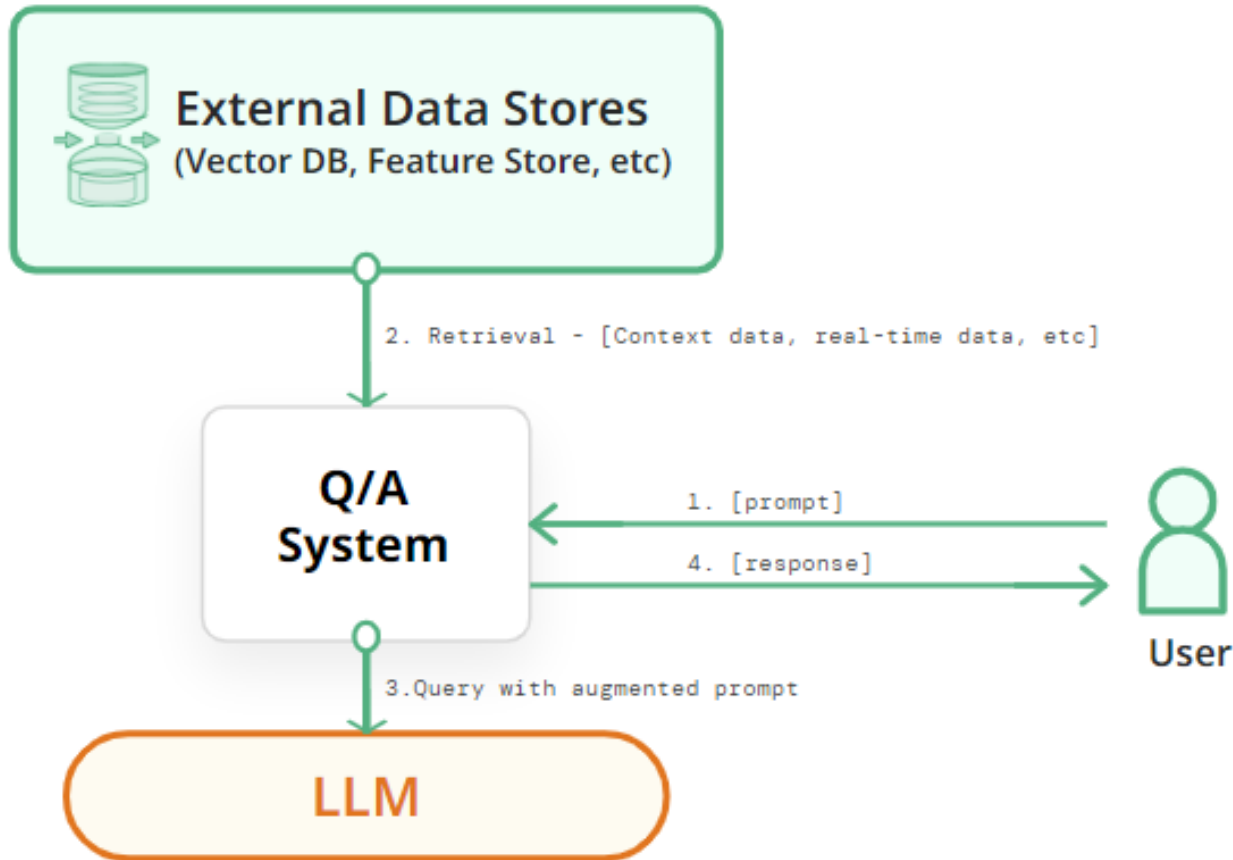


올해 KBO 한국시리즈 우승팀 예상이라면 키움 히어로즈에 기대를 거는 분들이 많겠네요! 키움이 꾸준히 강력한 모습을 보여주고 있지만, 다른 팀들도 막판에 변수로 작용할 수 있어 예측이 어렵죠. 예측해본다면 키움의 저력에 무게를 두고 싶습니다.

🔊 📄 📌 🔄 ⌵

Domain specific, hallucination free 챗봇 개발!

1. 서비스 개발 목표 > RAG(Retrieval-augmented generation)



✓. RAG

- LLM에게 족보(기사)를 주고 답을 맞춘다.
- 질문 벡터화 -> 유사도 검색 -> 관련 기사 추출
- 기사 데이터베이스 구축 필요

2. 데이터베이스 구축 > 데이터 선정

MARKET DATA

가장 핫한 KBO의 최근 경기 기록을 가져오는 데이터베이스 구축

소비자들이 가장 궁금해할 바로 다음 경기에 대한 전망 및 최근 부상 선수들을 실시간으로 가져올 수 있는 네이버 스포츠 야구 뉴스를 선택하였습니다.

sports.news.naver.com/kbaseball/news/index?isphoto=N

☆

N 스포츠

뉴스 | 연예

로그인

☰

홈

야구

해외야구

축구

해외축구

농구

배구

N골프

일반

e스포츠

뉴스

영상

일정

순위

포토

구단

크볼렉트

야구 전체

KBO

야구 일반

최신순

인기순

구단별 ▾

☑ 화보기사 제외

☰

📖



“부상 원태인 대신 구속 커브 좋아진 임찬규”

LG 임찬규 | 연합뉴스에 따르면 선발 부상 속 류중일 프리미어12 대표팀 감독의 선택은 임찬규 (32-LG)였다. 한국야구위원회(KBO)는 “전력강화위원회는 부...

스포츠경향 | 2024.10.30 01:00



“로하스, 심진 잡아야죠” 동료에서 적으로...박영현 “자신 있다” [교척 현장]

KT 위즈 마무리투수 박영현이 29일 서울 고척스카이돔에서 열린 2024 WBSC 프리미어12 대...



전체 임플란트—더이상 미루지 마세요

강남역 11번 출구

대한치과의사협회 의료심의회 제230316-중-25437호

원태이치과의원

더 알아보기 >

이 시각 많이 본 뉴스 ①

더보기 >

야구

스포츠 종합

2. 데이터베이스 구축 > 데이터 수집

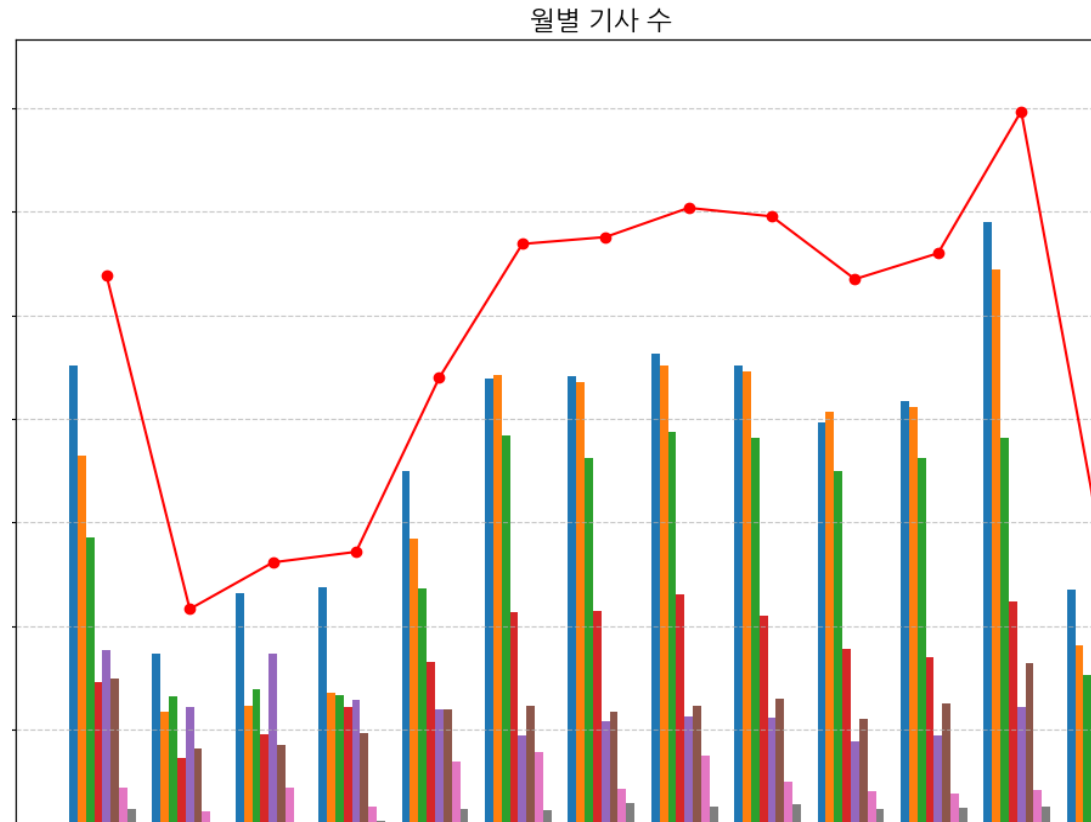
```
naver_sports_news_20231101_to_20231130.csv
naver_sports_news_20231201_to_20231231.csv
naver_sports_news_20240101_to_20240131.csv
naver_sports_news_20240201_to_20240229.csv
naver_sports_news_20240301_to_20240329.csv
naver_sports_news_20240401_to_20240430.csv
naver_sports_news_20240501_to_20240531.csv
naver_sports_news_20240601_to_20240630.csv
naver_sports_news_20240701_to_20240731.csv
naver_sports_news_20240801_to_20240831_c.csv
naver_sports_news_20240901_to_20240930.csv
naver_sports_news_20241001_to_20241031.csv
naver_sports_news_20241101_to_20241116..csv
```

```
10747 "나갈 때마다 자극 받아"...국제대회 통해 한 단계 성장하고 있는 독수리 군단 특급 우완 루키,2023.11.30 07:01,https://sports.news.naver.com/news?
10748 "[단독] 'FA 최대어' 양석환, 두산 남는다...4+2년 '70억원 후반' 계약 합의 '2024년 FA 최고 금액 예약'",2023.11.30 07:00,https://sports.news.nav
10749 "수상자 0명, 신인상 까마득...2023 무관의 롯데",2023.11.30 06:47,https://sports.news.naver.com/news?oid=241&aid=0003315954,"(부산=연합뉴
10750 "[조아제약 프로야구 대상] 최고투수상 뜨거운 삼파전, 최고구원투수상 주인공은 서진용 유력",2023.11.30 06:46,https://sports.news.naver.com/new
10751 "'방출 자청했는데...' FA 삼수생, 69억 포수, 우승반지 3개 투수, 새로운 기회가 있을까",2023.11.30 06:41,https://sports.news.naver.com/news?oid
10752 "한화에 이런 날이! 문동주의 확신 ""김서현-황준서, 신인상 당연히 받아야죠""",2023.11.30 06:31,https://sports.news.naver.com/news?oid=108&ai
10753 근조화환에 메모지 가득한 랜더스필드...1일 철거한다지만 성난 'SSG 팬심' 심각했다 [엑스케치],2023.11.30 06:31,https://sports.news.naver.com/
10754 "'또치의 귀환' 코치로 돌아온 김용의 ""영원히 LG와 함께할 운명인가 봅니다""",2023.11.30 06:30,https://sports.news.naver.com/news?oid=076&ai
10755 '불펜 보강' 성공→선발도 아쉬운데... '젊은 피' 이호성-육선엽에 쏠리는 '눈길' [SS시선집중],2023.11.30 06:10,https://sports.news.naver.com/new
10756 "새 감독 선임 일주일 만에 단장 찾기 나선 SSG ""신중 기한다""",2023.11.30 06:01,https://sports.news.naver.com/news?oid=003&aid=0012239419
10757 "GG 최대 격전지 유격수, 누가 받아도 이상하지 않다...오지환 수성이나, 박찬호 첫 수상이나",2023.11.30 06:01,https://sports.news.naver.com/news?
10758 "'현역 연장 고민' 추신수, SSG에 불어 닦친 한파 영향 받을까?",2023.11.30 06:01,https://sports.news.naver.com/news?oid=382&aid=0001090970,
10759 "만 36세에도 여전한 클래스...양의지, '포수 최다 GG' 전설 반열 오르나",2023.11.30 06:00,https://sports.news.naver.com/news?oid=421&aid=00072
10760 [IS 포커스] 피터스? 엔스? 스카우트의 눈이 '일본'으로 향한다,2023.11.30 05:30,https://sports.news.naver.com/news?oid=241&aid=0003315941,"
10761 끝판왕의 뒤는 누구,2023.11.30 05:05,https://sports.news.naver.com/news?oid=081&aid=0003412740,"고우석 美 가면... LG 마무리는 '1순위' 시속 15
10762 "'박헌 돌' 이정후 빠진 GG 외야수, '세 자리' 놓고 치열한 경쟁",2023.11.30 05:00,https://sports.news.naver.com/news?oid=421&aid=0007206482
10763 완투·완봉 급감... '선발투수' 시대 저무나,2023.11.30 04:08,https://sports.news.naver.com/news?oid=005&aid=0001655993,"토종 200이닝, 7년째
10764 김혜성 "김하성 형과 '빅리그 키스톤 콤비'... 최고의 시나리오죠",2023.11.30 03:02,https://sports.news.naver.com/news?oid=020&aid=0003534199,"
10765 [IS 시선] '사비로 아버지까지 초청' 시상식의 가치를 높인 페디,2023.11.30 00:52,https://sports.news.naver.com/news?oid=241&aid=0003315938,"
```

✓. 네이버 뉴스 수집

- 기사 등록 기간 : 2023년 11월 01일 ~ 2024년 11월 16일
- 총 111,081 건의 기사 수집

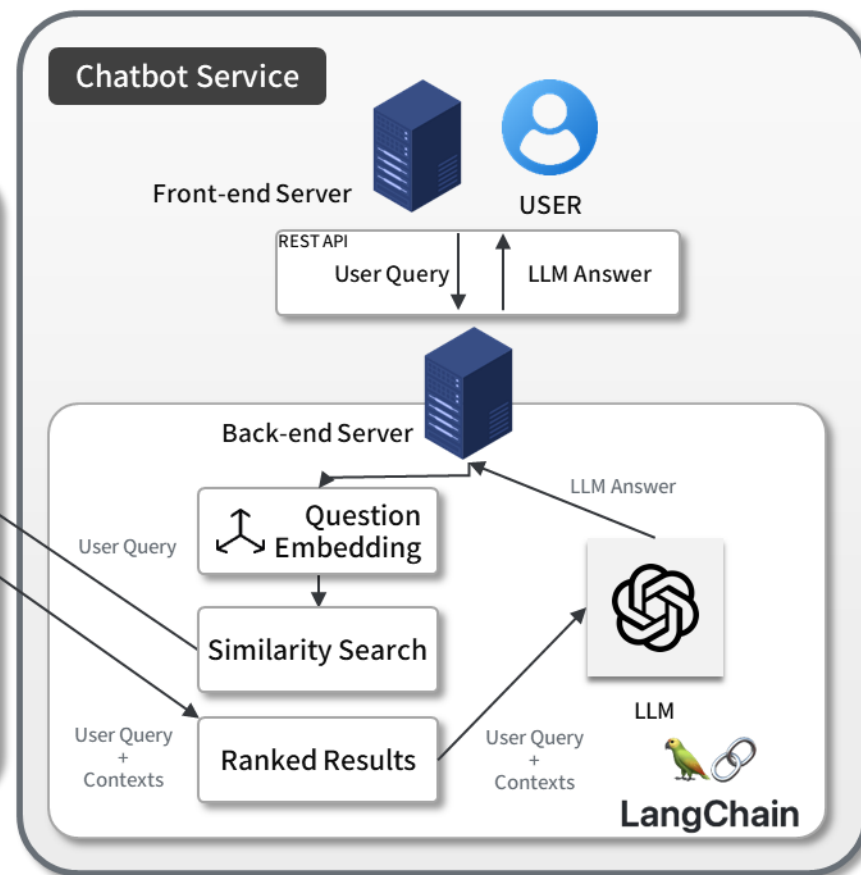
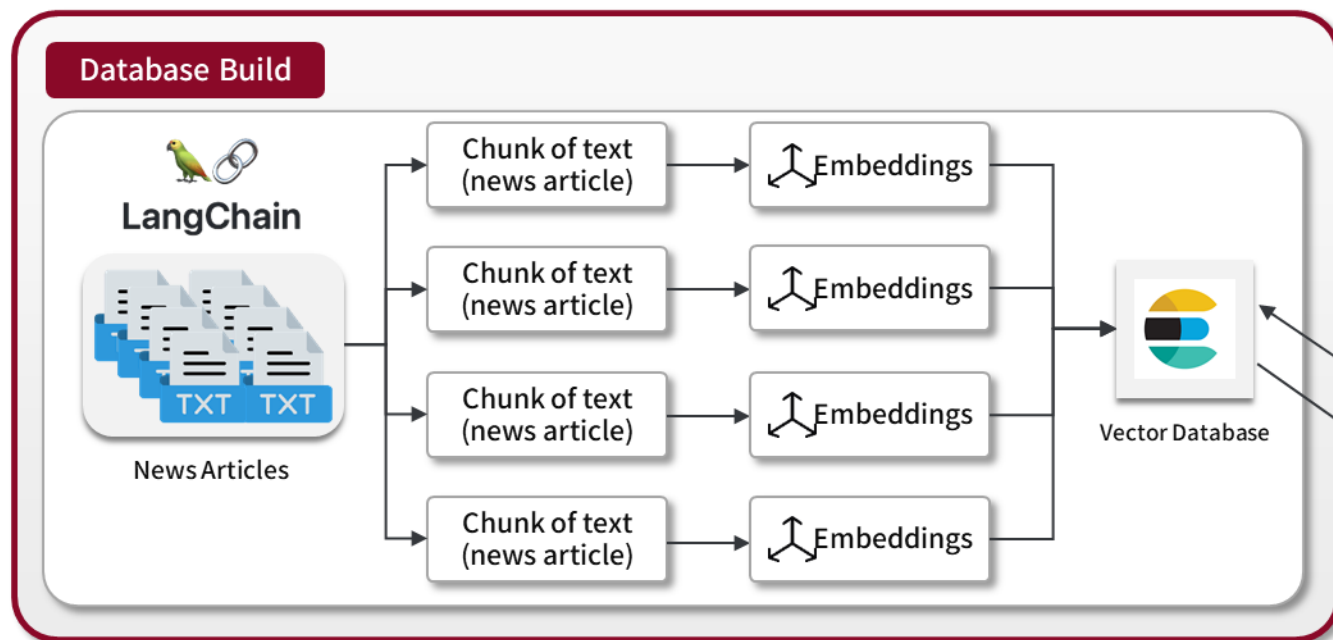
2. 데이터베이스 구축 > EDA(Exploratory Data Analysis; 데이터 분석)



✓. 구축 사항

- 3월~4월 (시즌 시작): 경기 결과와 기록/통계 기사의 급증, 부상/복귀 기사도 시즌 시작과 함께 증가
- 5월~9월 (정규 시즌): 경기 결과와 기록/통계가 지속적으로 높은 비중
- 10월 (포스트시즌): 경기 결과와 기록/통계 기사가 최고치
- 11월 (시즌 종료 후 비시즌 시작): 시즌 종료와 함께 팀 재편, 선수 계약, FA 시장 관련 기사가 다시 주요 비중

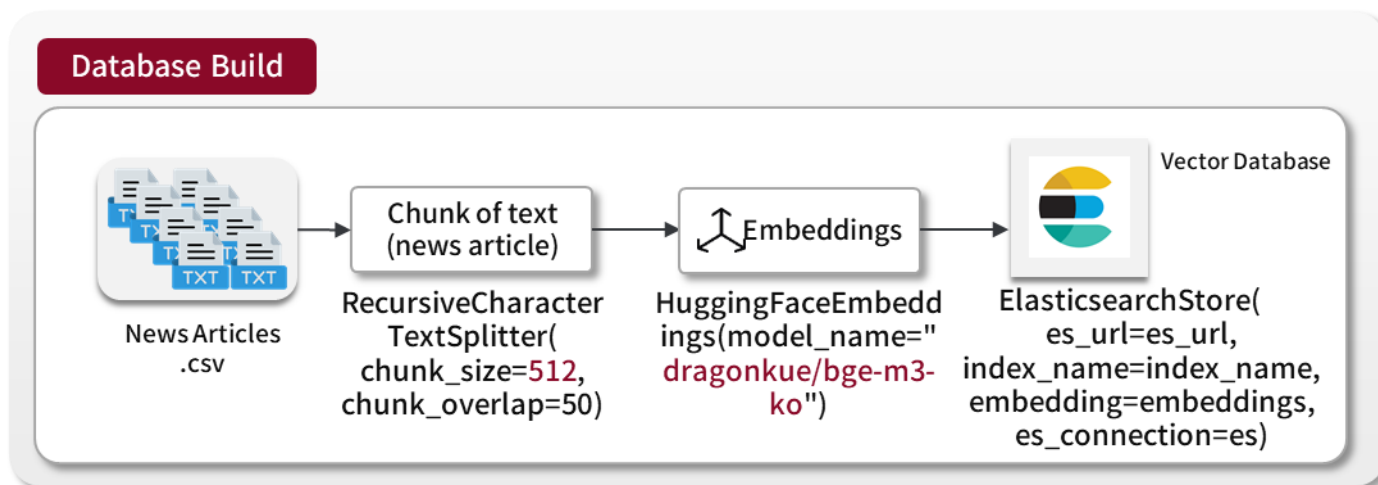
2. 데이터베이스 구축 > 전체 시스템 구성



✓. 구축 사항

- RAG 프레임워크(LangChain) 도입
- 기사 데이터베이스 구축, End-to-end 챗봇 서비스 구축

2. 데이터베이스 구축 > 모듈 선정



Chunk Size	lyft_2021	
	Average Faithfulness	Average Relevancy
2048	80.37	91.11
1024	94.26	95.56
512	97.59	97.41
256	97.22	97.78
128	95.74	97.22

Table 3: Comparison of different chunk sizes

Model name	F1	Recall	Precision	mAP	mRR	NDCG
KoSimCSE-roberta	0.4298	0.4298	0.4298	0.4298	0.4298	0.4298
Cohere embed-multilingual-v3.0	0.3596	0.3596	0.3596	0.3596	0.3596	0.3596
openai ada 002	0.4737	0.4737	0.4737	0.4737	0.4737	0.4737
multilingual-e5-large-instruct	0.4649	0.4649	0.4649	0.4649	0.4649	0.4649
Upstage Embedding	0.6579	0.6579	0.6579	0.6579	0.6579	0.6579
ko-sroberta-multitask	0.4211	0.4211	0.4211	0.4211	0.4211	0.4211
openai_embed_3_large	0.6053	0.6053	0.6053	0.6053	0.6053	0.6053
BGE-m3-ko	0.7456	0.7456	0.7456	0.7456	0.7456	0.7456

```
38 for data in datas:
39     for i in range(0, len(data)):
40         database.add_documents(text_splitter.split_documents(data[i:i+1]))
41
```

✓ 118m 55.0s

✓. 데이터베이스 선정

- 기존 lexical search 검색엔진으로 사용 중 -> 관리 포인트 ↓
- Lexical + Semantic 하이브리드 서치 연계 용이

✓. 파라미터, 모델 선정

- 최적의 청크사이즈 선정 -> 512
- 임베딩 모델 선정 -> dragonkue/bge-m3-ko
저용량(2.1GB), 고가용(8192토큰), 고성능(지표)

Document id: 0bbb5b85-bfb9-4d7a-b4de-eb6c40211109	
text	<p>"공을 던져준 서인석 매니저는 오스틴에게 뛰어올라 매달리며 함께 기뻐했다.페라자는 마지막 타석에서 아오됐고, 오스틴은 배트를 던지며 기뻐했다.LG 동료들은 오스틴에게 생수를 끼었으며 축하했다.이날 홈런더비엔 6일 올스타전에 출전하는 각 팀 선수가 모두 출석해 동료들을 응원했다.12년 만에 올스타전에 나서는 한화 이글스의 류현진은 자신의 휴대전화를 이용해 인스타그램 라이브 방송으로 홈런더비를 중계하기도 했다."</p>
metadata	<pre>{ "source": "https://sports.news.naver.com/news?oid=425&aid=0000151274", "row": 2442, "title": "'두 차례 서든데스서 승리' LG 오스틴, 올스타전 홈런더비 우승..연장서 페라자 제압", "datetime": "2024.07.06 15:29", "id": "e732c67d-6c90-437c-bc33-7101deb4968f" }</pre>

2. 데이터베이스 구축 > TEST

2023년 11월 13일 서울 잠실야구장에서 열린 KBO 한국시리즈 5차전에서 LG 트윈스가 kt 위즈를 상대로 몇 대 몇으로 승리했나요?

→ 모르겠습니다.

2023년 11월 13일 LG가 kt를 상대로 한국시리즈에서 승리한 점수는 얼마인가요?

→ 모르겠습니다.

KT 웨스 벤자민이 LG를 상대로 선발 등판한 경기에서 KT가 거둔 승리는 몇 경기인가요?

→ 모르겠습니다.

2023년 KBO리그에서 NC의 페디 선수가 MVP로 선정된 날짜는 언제인가요?

→ 모르겠습니다.

페디 선수가 2023 KBO리그에서 기록한 투수 3관왕의 성과는 무엇인가요?

→ "페디 선수는 2023 KBO리그에서 투수 3관왕을 기록하며, 평균자책점, 탈삼진, 승리 부문에서 1위를 차지했습니다. 이는 그의 뛰어난 투구 능력을 입증하는 성과입니다."

Without RAG

2023년 11월 13일 서울 잠실야구장에서 열린 KBO 한국시리즈 5차전에서 LG 트윈스가 kt 위즈를 상대로 몇 대 몇으로 승리했나요?

→ 2023년 11월 13일, 서울 잠실야구장에서 열린 KBO 한국시리즈 5차전에서 LG 트윈스는 kt 위즈를 6-2로 꺾고 시리즈 전적 4승 1패로 29년 만에 한국시리즈 우승을 차지했습니다.

2023년 11월 13일 LG가 kt를 상대로 한국시리즈에서 승리한 점수는 얼마인가요?

→ KT의 웨스 벤자민 선수가 LG를 상대로 선발 등판한 경기에서 KT가 승리한 경기는 2023년 한국시리즈 1차전으로, KT는 이 경기에서 3-2로 승리했습니다.

KT 웨스 벤자민이 LG를 상대로 선발 등판한 경기에서 KT가 거둔 승리는 몇 경기인가요?

→ KT 웨스 벤자민이 LG를 상대로 선발 등판한 경기에서 KT가 거둔 승리는 5경기입니다.

2023년 KBO리그에서 NC의 페디 선수가 MVP로 선정된 날짜는 언제인가요?

→ NC 다이노스의 에릭 페디 선수는 2023년 11월 27일에 KBO 정규시즌 MVP로 선정되었습니다.

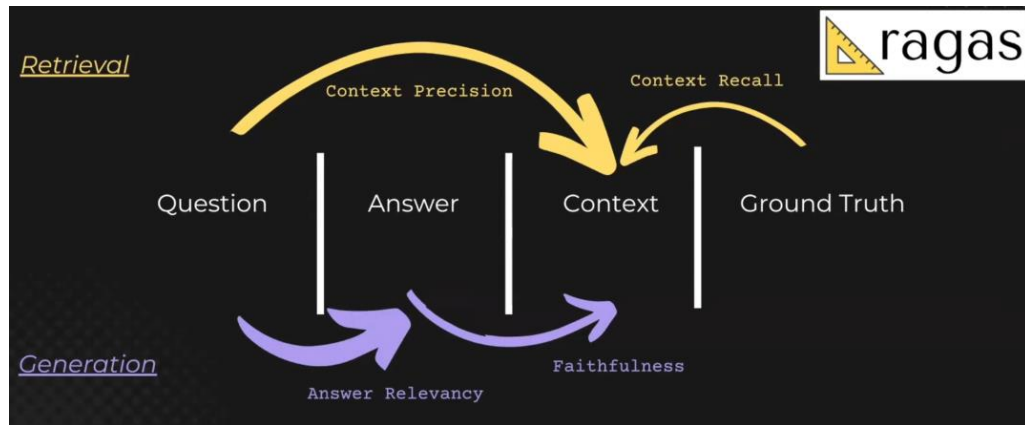
...

RAG

✓. TEST 결과

- RAG 사용 시, 검색 결과 성능이 증가한 것으로 보임
- 정량적인 성능 측정 필요

3. 성능 평가 > RAGAS(Retrieval Augmented Generation Assessment)



Question : 2023년 11월 7일 서울 잠실구장에서 열린 KBO 한국시리즈 1차전에서 문상철이 어떤 상황에서 삼중살을 기록했나요?

Context : "[""2회초 번트 시도하다 삼중살...올 PS서만 3번째 번트 실패2-2 맞선 9회초"" , '타석에 나온 문상철은 번트를 시도했다. 하지만 타구가 포수 앞에 떨어졌다. 박동원이 3루로 송구해 장성우를 잡았다. ...'"]"

Answer: "문상철은 2023년 11월 7일 서울 잠실구장에서 열린 KBO 한국시리즈 1차전에서 2회초 무사 1, 2루 상황에서 번트를 시도하다 삼중살을 기록했습니다. 그의 번트 타구가 포수 앞에 떨어지면서 3루로 송구되어 2루 주자가 아웃되고, 이어 1루에서도 아웃되었으며, 2루 주자가 3루를 노리다 아웃되면서 삼중살이 완성되었습니다."

Ground Truth : "문상철은 2023년 11월 7일 서울 잠실구장에서 열린 KBO 한국시리즈 1차전에서 2회초 무사 1, 2루 찬스에서 번트를 시도했으나 실패하여 삼중살이 발생했습니다. 그의 타구가 포수 앞에 떨어지면서 LG 포수 박동원이 3루로 던져 2루주자 장성우를 아웃시켰고, 문상철도 1루에서 아웃되었습니다. 이어서 2루로 향했던 배정대가 3루를 노리는 무리한 주루플레이를 하다 아웃되면서 삼중살이 기록되었습니다."

TEST DATA

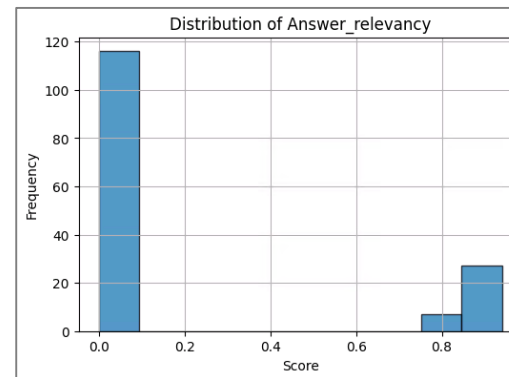
✓. RAGAS

- LLM 기반 RAG 성능 평가
- LLM으로 Question, Answer, Ground Truth 생성 -> 의존성 이슈

✓. 측정 결과

- RAG 사용 시, narrow domain 질문 대응 능력 개선

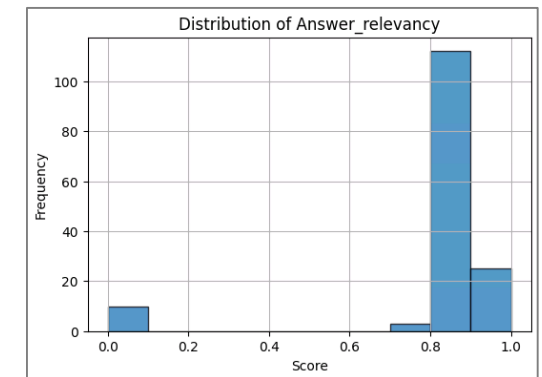
결론 : Hallucination 만 없으면 RAG 쓰는 게 낫다



Without RAG

Average Score for answer_relevancy: 0.198

Median Score for answer_relevancy: 0.0

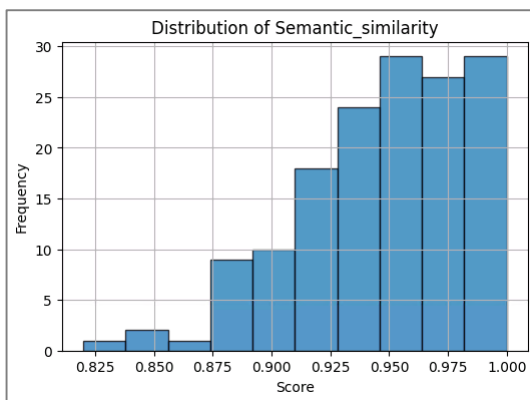
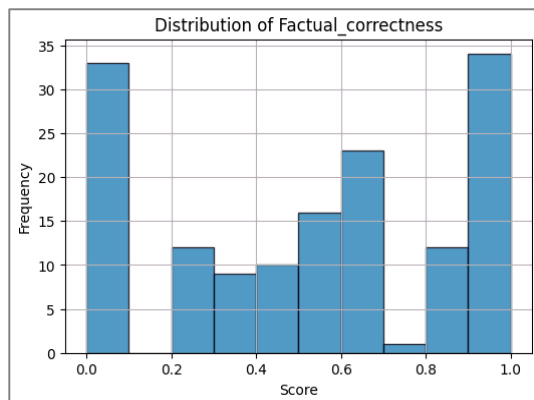
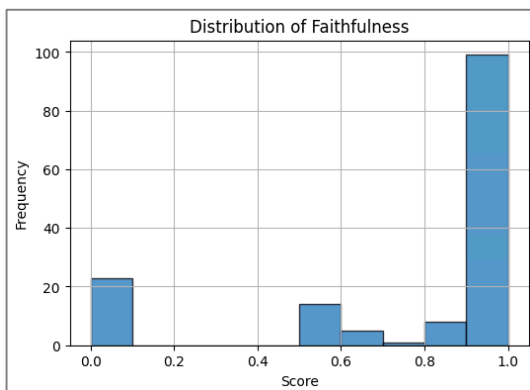
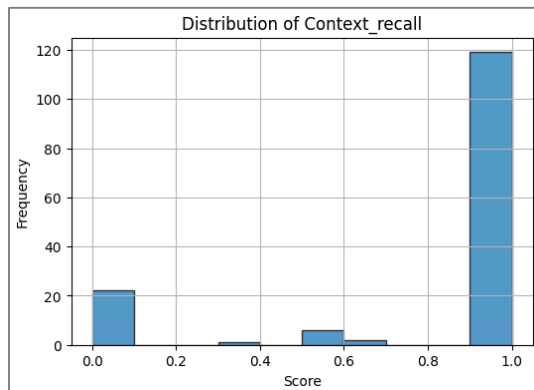


RAG

Average Score for answer_relevancy: 0.817

Median Score for answer_relevancy: 0.874

3. 성능 평가 > Hallucination diagnosis



✓. Average Scores

context_recall : 0.824444

faithfulness : 0.779063

factual_correctness : 0.520800

semantic_similarity : 0.947198

✓. Median Scores

context_recall : 1.000000

faithfulness : 1.000000

factual_correctness : 0.500000

semantic_similarity : 0.953011

Factual Correctness (0.520800)

의미: 생성된 응답의 사실적 정확성을 나타냅니다. **Answer <-> Context, Ground truth**

해석: 52%의 정확도는 응답의 절반가량이 사실과 일치하지 않을 수 있음을 의미합니다. 이는 **환각 현상이 발생할 가능성이 높다**는 것을 시사합니다.

Faithfulness (0.779063)

의미: 생성된 응답이 주어진 컨텍스트에 얼마나 충실한지를 나타냅니다.

Answer <-> Context

해석: 약 78%의 충실도는 응답의 22%가 주어진 정보와 일치하지 않을 수 있음을 의미합니다. 이는 **환각의 잠재적 위험이 존재함**을 나타냅니다.

Context Recall (0.824444)

의미: 시스템이 관련 컨텍스트를 얼마나 잘 회상하는지를 나타냅니다.

해석: 82%의 회상률은 대부분의 관련 정보를 잘 활용하고 있음을 의미하지만, 여전히 일부 중요한 정보를 놓칠 수 있음을 나타냅니다.

Semantic Similarity (0.947198)

의미: 질문과 응답 간의 의미적 유사성을 나타냅니다.

해석: 95%의 유사성은 응답이 질문과 매우 밀접하게 관련되어 있음을 의미합니다.

✓. 측정 결과

- 현 RAG 시스템은 환각 가능성이 높은 것으로 추정

3. 성능 평가 > Hallucination case study(Factual Correctness)

Question : 2024년 3월 9일 대전 한화생명이글스파크에서 열린 삼성 라이온즈와 한화 이글스의 시범경기에서 강민호가 기록한 2타점 2루타의 상황은 어땠나요?

Context : "'삼성'은 1회초 강민호의 2타점 2루타로 앞서갔지만, ...'"

Answer : "2024년 3월 9일 대전 한화생명이글스파크에서 열린 삼성 라이온즈와 한화 이글스의 시범경기에서, 삼성의 강민호는 1회초 2사 1, 2루 상황에서 2타점 2루타를 기록했습니다. 이로 인해 삼성은 초반에 앞서갔지만, 이후 추가 득점을 올리지 못해 결국 2-6으로 패배했습니다."

Ground truth : "2024년 3월 9일 대전 한화생명이글스파크에서 열린 시범경기에서 삼성 라이온즈의 강민호는 1회말 2사 1, 2루 상황에서 2타점 2루타를 기록했습니다. 이 장면에서 강민호는 타임을 요청하기도 했습니다."

context_recall: 1.0

faithfulness: 0.6

factual_correctness: 0.0

semantic_similarity: 0.973

✓. Low score 도출 이유

- 1) Ground truth 데이터 자체가 잘못 되었다.
- 2) 동일한 사건에서 Ground truth 데이터보다 더 많은 정보를 가진 데이터가 있다.

Factual Correctness (사실적 정확성): 0.0

측정 내용: 생성된 응답이 정답(ground truth)과 비교해 사실적으로 정확한지 평가합니다. Answer <-> Context, Ground truth

계산 방식: 정답에는 "1회말"이라고 되어 있지만, 생성된 응답에는 "1회초"라고 되어 있어 사실과 다릅니다.

또한, 생성된 응답은 "최종 점수 (2-6)"를 잘못 언급하고 있으며, 정답에 나와 있는 "강민호의 타임 요청"에 대한 언급도 누락되었습니다.

점수가 0.0인 이유: 생성된 응답에서 주요 사실적 요소가 잘못되었기 때문에, 0.0 (완전히 부정확함) 점수를 받았습니다.



3월 9일, 12:59 ~ 15:19 (2시간 20분), 대전 한화생명 이글스파크 12,000명 (매진)													
팀	선발	1회	2회	3회	4회	5회	6회	7회	8회	9회	R	H	E B
삼성	이호성	2	0	0	0	0	0	0	0	0	2	5	0 4
한화	산체스	1	0	0	5	0	0	0	0	-	6	10	0 4

Ground truth 참조 기사 : 2024년 3월 9일 대전 한화생명이글스파크에서 열린 시범경기에서 삼성 라이온즈의 강민호는 1회말 2사 1, 2루 상황에서 2타점 2루타를 기록했습니다...

3. 성능 평가 > Hallucination case study(Faithfulness)

Question : 2024년 3월 9일 대전 한화생명이글스파크에서 열린 삼성 라이온즈와 한화 이글스의 시범경기에서 강민호가 기록한 **2타점 2루타의 상황**은 어땠나요?

Context : "[삼성은 **1회초** 강민호의 **2타점 2루타**로 앞서갔지만, ...]"

Answer : "2024년 3월 9일 대전 한화생명이글스파크에서 열린 삼성 라이온즈와 한화 이글스의 시범경기에서, 삼성의 강민호는 **1회초** 2사 1, 2루 상황에서 2타점 2루타를 기록했습니다. 이로 인해 삼성은 초반에 앞서갔지만, 이후 추가 득점을 올리지 못해 결국 2-6으로 패배했습니다."

Ground truth : "2024년 3월 9일 대전 한화생명이글스파크에서 열린 시범경기에서 삼성 라이온즈의 강민호는 **1회말** 2사 1, 2루 상황에서 2타점 2루타를 기록했습니다. 이 장면에서 강민호는 타임을 요청하기도 했습니다."

context_recall, **faithfulness**, **factual_correctness**,
semantic_similarity
1.0, **0.6**, **0.0**, 0.973

✓. **Low score 도출 이유**

- 1) LLM 이 실수 -> Evaluation 신뢰성 의심

Ground truth, 평가 방법 자체에 문제가 있다.

Faithfulness (충실도): 0.6

측정 내용: 생성된 응답이 검색된 문맥에 얼마나 충실한지 평가합니다. **Answer <->**

Context

계산 방식: 생성된 응답에는 "1회초", "삼성의 초기 리드"와 같은 검색된 문맥에 직접적으로 뒷받침되지 않는 추가 정보가 포함되어 있습니다. 검색된 문맥에는 "강민호의 2타점 2루타"만 언급되어 있고, 회차나 점수와 같은 세부사항은 없습니다. 점수가 0.6인 이유: 생성된 응답이 문맥과 일부 일치하지만, 뒷받침되지 않는 정보가 포함되어 충실도가 부분적으로 낮아졌습니다.



생성된 응답에서 사용된 정보(1회초, 강민호의 2타점 2루타, 삼성의 초반 리드, 최종 점수 2-6)는 모두 문맥에서 확인할 수 있습니다. 따라서, 생성된 응답은 검색된 문맥에 충실하다고 볼 수 있습니다.

Faithfulness 점수 재평가:

원래 점수: 0.6 -> 수정된 점수: 1.0

이유: 생성된 응답이 검색된 문맥에 기반하고 있으며, 불필요한 정보 또는 잘못된 세부정보를 포함하지 않았습니다.

3. 성능 평가 > 결과 고찰

Faithfulness (충실도): 0.6

측정 내용: 생성된 응답이 검색된 문맥에 얼마나 충실한지 평가합니다. Answer <->

Context

✓. Low score 도출 이유

- 1) LLM 이 실수

✓. 해결방안

- 1) Better context feeding
- 2) Better LLM

Factual Correctness (사실적 정확성): 0.0

측정 내용: 생성된 응답이 정답(ground truth)과 비교해 사실적으로 정확한지 평가합니다.

Answer <-> Context, Ground truth

✓. Low score 도출 이유

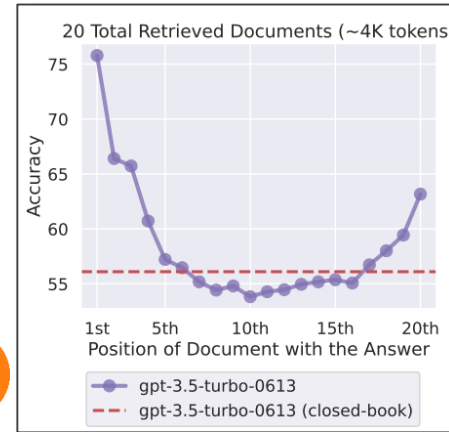
- 1) 동일한 사건에서
Ground truth 데이터보다
더 많은 정보를 가진
데이터가 있다.
- 2) Ground truth 데이터
자체가 잘못 되었다.

✓. 해결방안

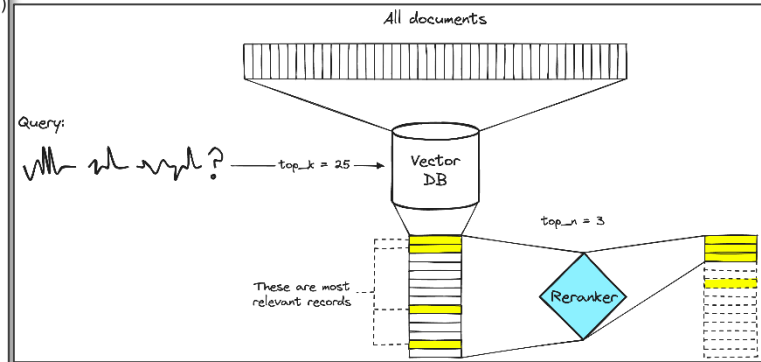
- 1) Better retrieving



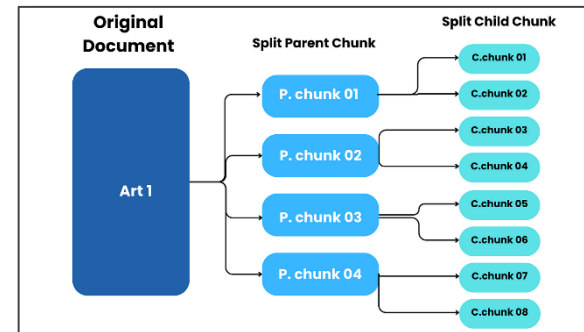
Context Reranking



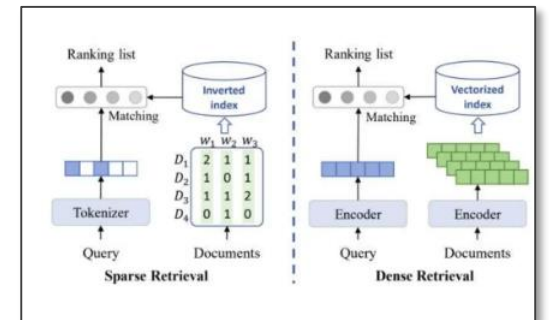
Lost in the Middle: How Language Models Use Long Contexts



Parent Document retriever



Ensemble retriever



3. 성능 평가 > 문서 상충 처리, 평가 결과

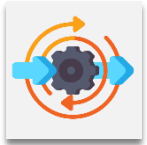
Context 1 : 삼성은 1회초 강민호의 2타점 2루타로 앞서갔지만, 이후 추가점을 올리지 못하며 시범경기 시작부터 아쉬움을 삼켰습니다.

Context 2 : 삼성 라이온즈의 강민호는 1회말 2사 1, 2루 상황에서 2타점 2루타를 기록했습니다.



문서 간 증거 분석

문서의 세부 사항을 비교하여 주장 검증:
시간 순서: 다른 사건들이 "1회초" 또는 "1회말"을 뒷받침하는가?
예: 다른 문서에서 강민호가 경기 초반 득점을 기록했다면 "1회초"가 더 설득력 있을 수 있음.
문맥적 단서: "초반에 리드를 잡았다" 같은 표현은 "1회초"를 지지할 가능성 있음.
직접적인 모순: 특정 문서가 스스로를 모순되게 기술한다면 신뢰도가 낮음.



서비스 단계

모순 자동 탐지

- 1) 문서별 NER 기반 사건별 주요 엔티티(날짜, 시간, 인물, 사건 등)를 추출 > 문서 간 또는 문서 내 주요 엔티티 간의 불일치를 탐지
- 2) 문서 내에서 서로 상반된 주장을 식별하는 규칙 기반 또는 ML 기반 접근 방식.



데이터 관리 단계

✓. 모든 문서가 진실이라면 어떻게 처리해야 할까?

- 임의 추론 금지 > 상충 사실 표기
- 주기적인 데이터 논리 정렬 > 에이전트

평가 결론 : 실제 서비스 배포 후, 사용자 피드백으로 평가

사실 기반의 균형 잡힌 답변 생성

답변 생성 시 포함해야 할 요소:
문맥적 일관성과 증거를 바탕으로 한 가장 가능성 높은 사실.
불일치가 중요한 경우 이를 명확히 언급하여 투명성을 유지.

예시 응답:

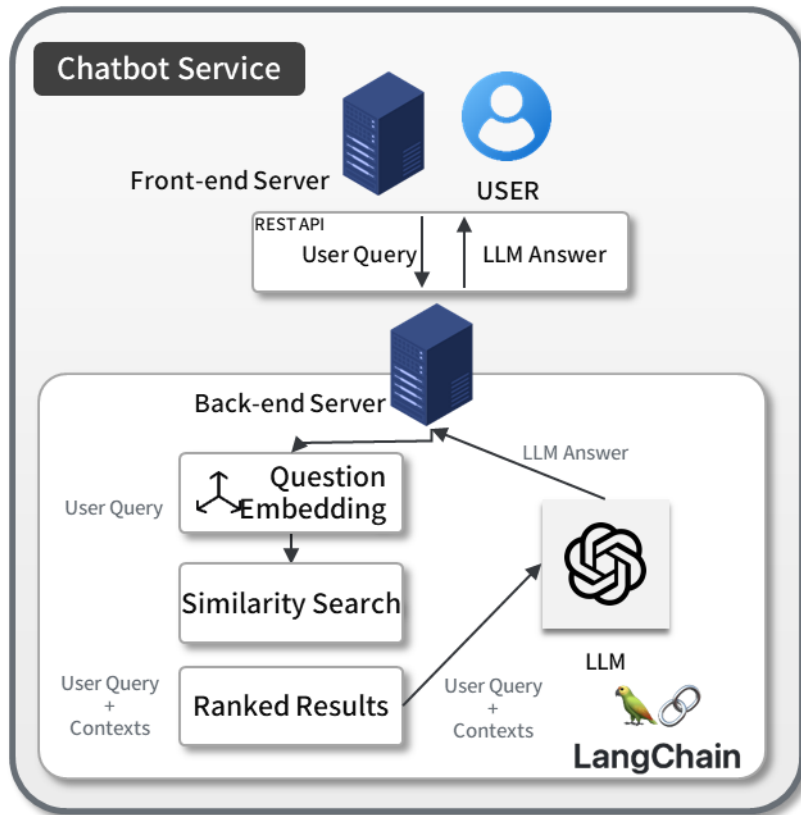
"2024년 3월 9일 대전 한화생명이글스파크에서 열린 경기에서 강민호는 2타점 2루타를 기록했습니다. 일부 문서에서는 이 장면이 1회초에 발생했다고 하고, 다른 문서에서는 1회말에 발생했다고 기록되어 있습니다. 다만, 경기를 앞서간 상황을 고려할 때, 1회초에 있었을 가능성이 높습니다."

탐지 예시

모호성 탐지 결과:

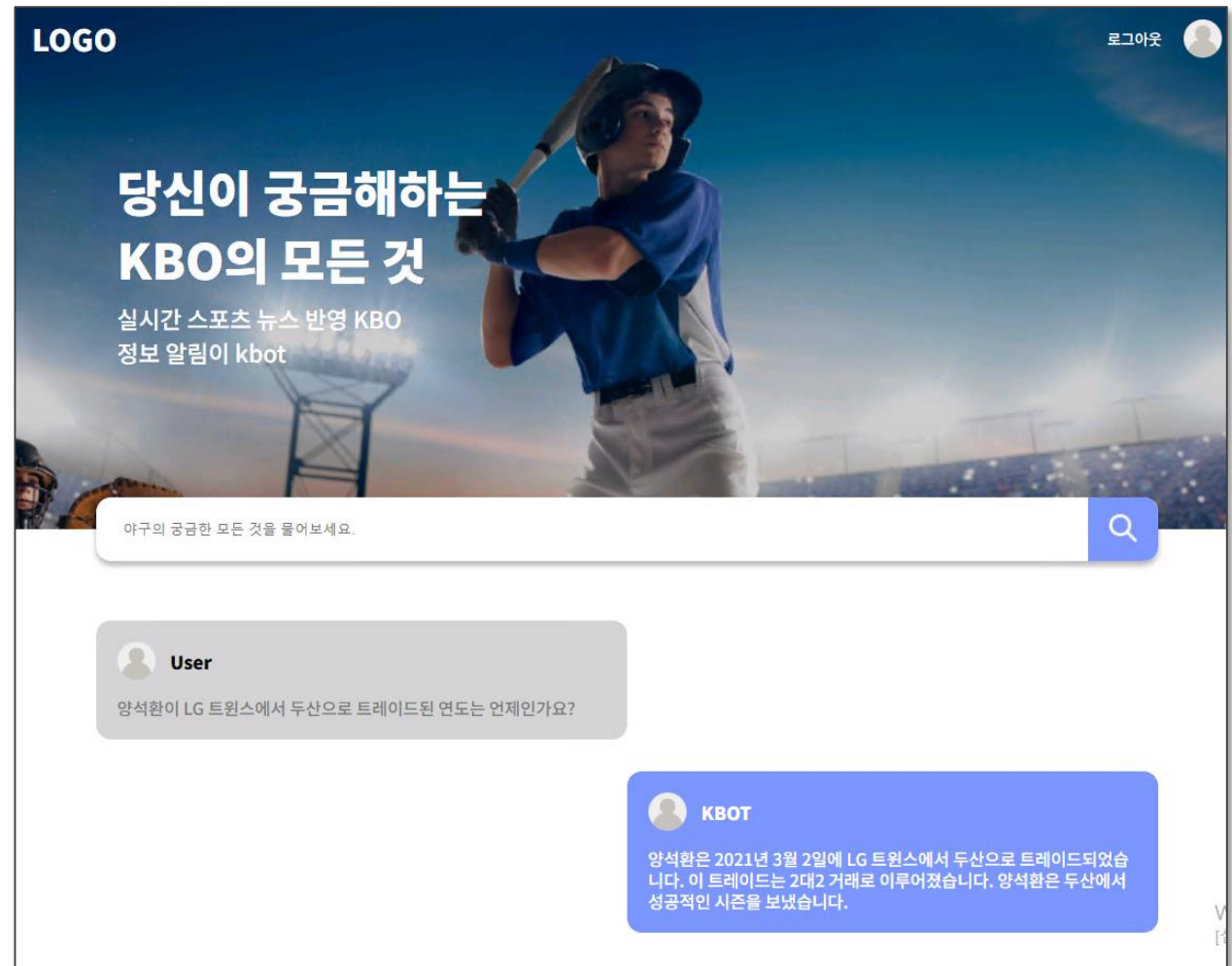
1. 사건: 강민호의 2타점 2루타
 - 문서 1: "1회초"에 발생 (출처: A 신문)
 - 문서 2: "1회말"에 발생 (출처: B 웹사이트)
 - 신뢰 점수: 문서 1 (0.85), 문서 2 (0.60)
 - 결론: 1회초 발생 가능성 높음.

4. 시스템 구축 > 서비스 페이지

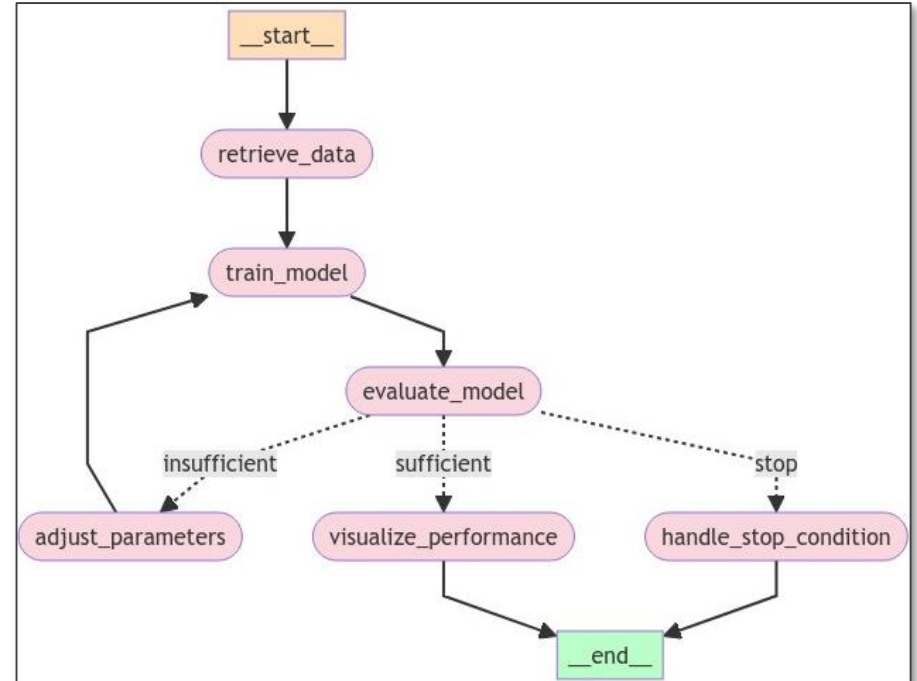
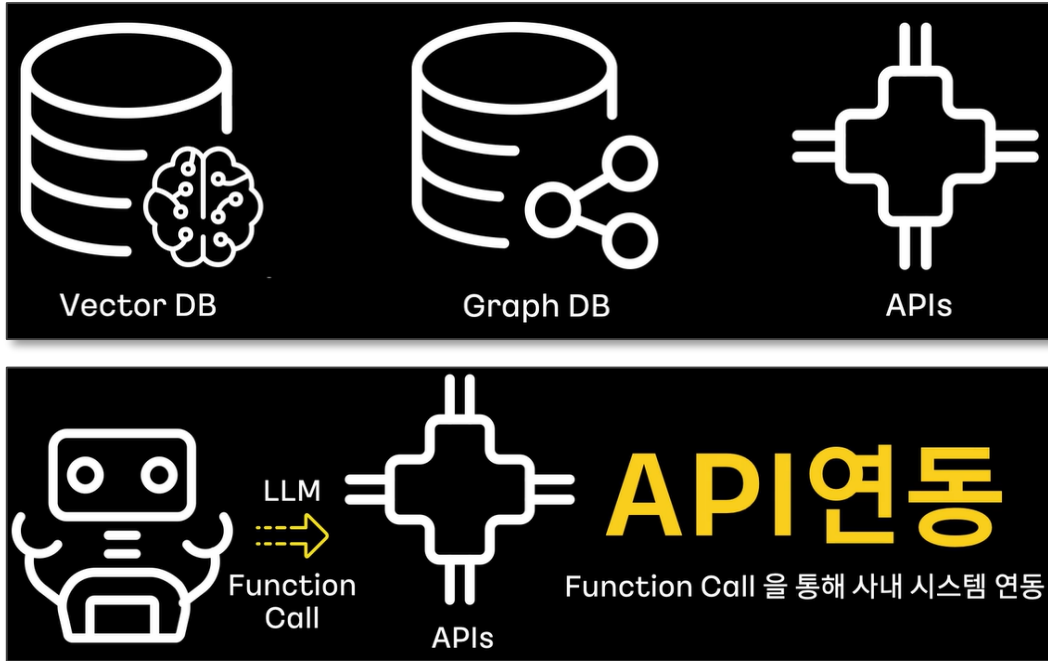


✓. 서비스 사이트 구축

- React 기반 front-end server 구축
- Fastapi 기반 back-end server 구축



4. 시스템 구축 > Agent

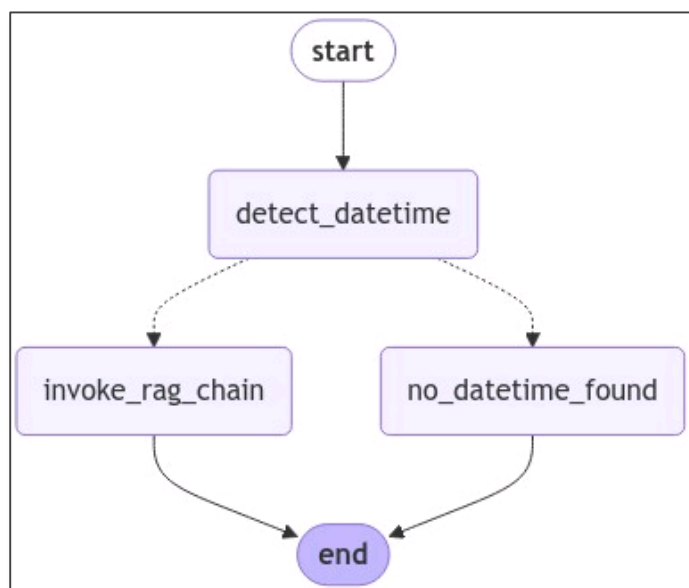


✓. 에이전트 중요도 증가

- 정교한 시스템 구축 시
- 다양한 데이터 베이스 참조 및 시스템 연동

4. 시스템 구축 > Prompting Agent

LangGraph



✓. 사용 예시

- 완벽하지 않은 사용자의 질문을 필터링

Input

```
{  
  role: "user"  
  content: "Does this question contain any date or time information? Answer 'yes' or 'no'.  
  
  Question: 한국 야구 대표팀이 대만에게 패한 점수는 얼마인가요?"  
}
```

Output

```
{  
  role: "assistant"  
  content: "no"  
}
```

Input

```
{  
  role: "user"  
  content: "Does this question contain any date or time information? Answer 'yes' or 'no'.  
  
  Question: 2024 WBSC 프리미어12 조별리그 B조 1차전에서 한국 야구 대표팀이 대만에게 패한 점수는 얼마인가요?"  
}
```

Output

```
{  
  role: "assistant"  
  content: "Yes"  
}
```

4. 시스템 구축 > DEMO

```
Run Terminal Help
final-llm-kbo-chatbot [WSL: Ubuntu]

...
csv_insert.ipynb langgraph.ipynb main.py JS Chat.js autorag_evaluation.ipynb csv_clean.ipynb extract_from_langfuse.ipynb evaluate.ipynb

UBUNTUJ
nb
mb
t_rag.ipynb
thout_rag_scores.csv
thout_rag.csv
n_results.csv
r_score_answer_relevancy.csv
r_score.csv
answer_relevance.csv
sv
ncremental_2.csv
ncremental.csv
sv
518Document.csv
926Document.csv
231101_to_20231130.csv
231201_to_20231231.csv
240101_to_20240131.csv
240201_to_20240229.csv
240301_to_20240329.csv
240401_to_20240430.csv
240501_to_20240531.csv
240601_to_20240630.csv
240701_to_20240731.csv

backend > main.py > invoke_rag_chain
64 es_url = 'http://127.0.0.1:9200'
65 index_name = 'news_articles_2'
66 ES_HOST = 'https://127.0.0.1:9200'
67
68 # Connect to Elasticsearch
69 es = Elasticsearch(
70     hosts=[ES_HOST],
71     http_auth=('elastic', 'Bj8DOyhJwM0cZL9s-_jF'),
72     verify_certs=True,
73     ca_certs='/mnt/d/certs/http_ca.crt',
74 )
75
76 # Initialize ElasticsearchStore with embeddings
77 from langchain.embeddings.huggingface import HuggingFaceEmbeddings
78 embeddings = HuggingFaceEmbeddings(model_name='dragonkue/bge-m3-ko', model_kwargs={'device': 'cuda'})
79 database = ElasticsearchStore(
80     es_url=es_url,
81     index_name=index_name,
82     embedding=embeddings,
83     es_connection=es
84 )
85
86 # Create retriever
87 retriever = database.as_retriever()
88
89 # Load prompt template
90 prompt = hub.pull('rlm/rag-prompt')
91
92 from langfuse.callback import CallbackHandler
93 langfuse_handler = CallbackHandler(
94     public_key='pk-lf-65a5c8b5-ed9-4fcf-9fcd-02758ceb7e5a',
95     secret_key='sk-lf-4e89d384-59c6-41e7-8474-a18581d4d874',
96     host='https://langfuse.com'
97 )
98
99 PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS 5 JUPYTER
100 /home/jtbcadmin/.virtualenvs/rag/lib/python3.11/site-packages/langsmith/client.py:333: LangSmithMissingAPIKeyWarning: API key must be provided when using hosted LangSmith,
101     warnings.warn(
102 /home/jtbcadmin/.virtualenvs/rag/lib/python3.11/site-packages/langsmith/client.py:543: LangChainBetaWarning: The function 'loads' is in beta. It is actively being worked
103     so the API may change.
104     prompt = loads(json.dumps(prompt_object.manifest))
105 /mnt/c/Users/JtbcAdmin/DEV/final-llm-kbo-chatbot/backend/main.py:181: LangChainDeprecationWarning: The class 'ChatOpenAI' was deprecated in LangChain 0.0.10 and will be re
106     in 1.0. An updated version of the class exists in the langchain-openai package and should be used instead. To use it run 'pip install -U langchain-openai' and import as
107     from langchain_openai import ChatOpenAI.
108     llm = ChatOpenAI(model_name='gpt-4o-mini', temperature=0, callbacks=[langfuse_handler])
109 Traceback (most recent call last):
110   File "main.py", line 181, in invoke_rag_chain
111     llm = ChatOpenAI(model_name='gpt-4o-mini', temperature=0, callbacks=[langfuse_handler])
112   File "C:\Users\JtbcAdmin\AppData\Local\Programs\Python\Python311\Lib\site-packages\langchain_openai\chat_models.py", line 100, in __init__
113     self._client = OpenAIClient(api_key=api_key, base_url=base_url, timeout=timeout, max_retries=max_retries, http_client=http_client)
```

4. 시스템 구축 > System monitoring

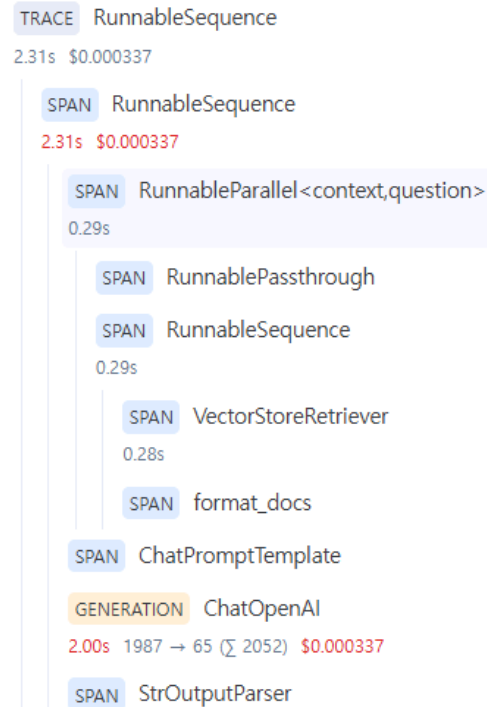


User

양석환이 LG 트윈스에서 두산으로 트레이드된 연도는 언제인가요?

KBOT

양석환은 2021년 3월 2일에 LG 트윈스에서 두산으로 트레이드되었습니다. 이 트레이드는 2대2 거래로 이루어졌습니다. 양석환은 두산에서 성공적인 시즌을 보냈습니다.



SPAN VectorStoreRetriever

2024. 12. 2. 오후 1:28:02

Latency: 0.28s

Pretty ✨ JSON

Input

양석환이 LG 트윈스에서 두산으로 트레이드된 연도는 언제인가요?

Output

[~
0: { ~
metadata: { ~
id: "807d5446-f6de-4e12-9953-c109d6e8a6c2"

GENERATION ChatOpenAI

2024. 12. 2. 오후 1:28:02

Latency: 2.00s

1987 prompt → 65 completion (Σ 2052)

gpt-4o-mini

\$0.000337

temperature: 0.0

Pretty ✨ JSON

user

You are an assistant for question-answering tasks. Use the following pieces of retrieved context to answer the question. If you don't know the answer, just say that you don't know. Use three sentences maximum and keep the answer concise.
Question: 양석환이 LG 트윈스에서 두산으로 트레이드된 연도는 언제인가요?
Context: 그런 양석환에게 두 번째 터닝포인트가 찾아왔다. 2021년 3월 2대2 트레이드를 통해 LG에서 라이벌

✓. RAG 파이프라인 모니터링

- query 요청 별, 벡터 유사도 검색 결과, LLM 프롬프트 및 LLM API 사용 요금 관제

5. 추가 보완 사항

✓. RAG pipeline 고도화

- chunking 고도화
- retriever 고도화
- reranker 도입
- lexical + semantic search 도입

✓. Evaluation 고도화

- ragas 보완 -> NER 기반 엔티티만 평가

✓. 에이전트 고도화

- 사용자 질문에 대한 exception 처리

✓. 다양한 비정형 데이터 처리

- PDF, WORD, EXCEL 처리

6. Pain Points > Data Crawling

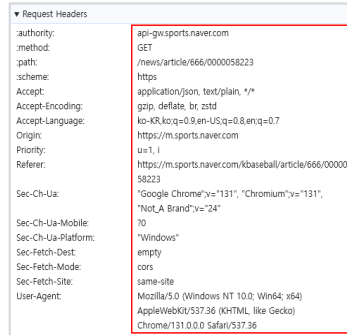
크롤러 차단 방지



VPN

```
print(ua.chrome)
# Mozilla/5.0 (Windows NT 10.0; Win64; x64)
print(ua.google)
# Mozilla/5.0 (Macintosh; Intel Mac OS X 10_
print(ua['google chrome'])
# Mozilla/5.0 (Windows NT 10.0; Win64; x64)
print(ua.firefox)
# Mozilla/5.0 (X11; Linux x86_64; rv:109.0)
print(ua.ff)
# Mozilla/5.0 (X11; Linux x86_64; rv:102.0)
print(ua.safari)
# Mozilla/5.0 (Macintosh; Intel Mac OS X 10_
```

Fake_useragent



정상 Header 활용

크롤링 속도 개선

```
def __init__(self):
    # Session 초기화
    self.session = requests.Session()
    self.session.headers.update({
        "User-Agent": UserAgent().chrome
    })
```

세션 재사용

```
# ThreadPoolExecutor를 사용한 병렬 처리
with ThreadPoolExecutor(max_workers=5) as executor:
    articles = list(executor.map(
        self.get_article_content,
        *iterables: news_list
    ))
    news_data.extend(articles)
```

병렬 처리

✓. 서비스 사이트 구축

- 크롤러가 아닌 것으로 인식하기 위해서 다양한 방법을 시도
- 속도 개선을 위해 세션 재사용 및 병렬 처리를 진행했으나, Error 코드가 발생하여 순차적으로 크롤링 진행

Q & A

감사합니다.