Journal of the Korean Institute of Industrial Engineers
Vol. 50, No. 2, pp. 97-108, April 2024.
ISSN 1225-0988 | EISSN 2234-6457
https://doi.org/10.7232/JKIIE.2024.50.2.097

FT-LLM: 축구 산업의 데이터 분석 지원을 위한 Retrieval 증강 생성 기반의 언어 모델 프레임워크 개발

한승헌[1] · 강민혁[1] · 강필성[1] · 황우현[2†]

[1]고려대학교 산업경영공학과, [2]고려대학교 통계학과

FT-LLM: Development of a Retrieval Augmented Generation based Language Model Framework for Assisting Data Analysis in the Football Industry

Seung Hun Han[1] · Minhyeok Kang[1] · Pilsung Kang[1] · Woo Hyun Hwang[2]

[1]Department of Industrial & Management Engineering, Korea University
[2]Department of Statistics, Korea University

In the football industry, data analysis is gradually gaining more recognition as a key tool for developing match and club management strategies and analyzing players. Consequently, large language models (LLMs), with the ability to offer customized responses to user inquiries, could be a valuable data analysis resource. In this paper, we propose FT-LLM, a framework that enhances the language generation capability of language models by incorporating a partially cleansed, football-related documents and a retrieval. Compared to existing RAG methodologies, FT-LLM provides more suitable answer for football related queries by effectively incorporating external documents with the help of Query Refinement and Confidence Check modules. By employing a chain of thought-based prompting strategy to prevent hallucination, language models can convey more relevant and reliable football related insights to users. As a result, FT-LLM has the potential to assist data-driven decision making and operations in the football industry by formulating a match strategy or advising player acquisition.

Keywords: Natural Language Processing, Retrieval Augmented Generation, Football Data Analysis

1. 서 론

스포츠 산업의 빠른 성장과 함께 데이터 분석의 중요성이 증가하고 있으며, 언어 모델은 잠재적으론 맞춤형 업무 지원을 제공하여 업무 효율성을 높일 수 있는 중요한 자원으로 부상하고 있다(Connor and O'Neill, 2023). 예컨대, 축구 특화 언어 모델은 선수 정보 요약이나 상대팀 전략 분석 등의 업무를 자동화하여 데이터 분석을 용이하게 할 수 있다(Unlu, 2023).

Transformer(Vaswani *et al.*, 2017)의 공개 이후 대형 언어 모델(large language models) 관련 연구는 방대한 크기의 파라미터와 대량의 데이터를 활용하여 언어 모델의 생성 능력을 비약적으로 발전시켰다. 특히, ChatGPT와 Llama(Touvron *et al.*, 2023) 같은 최신 모델은 다양한 과업에서 인간 수준의 성능을 보여준다(OpenAI, 2023). 그러나, 언어 모델은 사실과 다른 답변을 제공하는 할루시네이션(Hallucination) 문제와 주기적인 미세조정 없이는 최신 정보를 반영하지 못하는 한계를 가지고 있다(McKenna *et al.*, 2023; Zhang *et al.*, 2023). 이를 해결하기 위해 언어 모델 연구 차원에선 비교적 적은 파라미터 크기임

이 논문은 2023년도 정부의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업(NRF-2022R1A2C2005455)의 성과물임. 또한, 이 논문은 싸커러리 박동현의 지원을 받아 수행되었음.
† 연락저자 : 황우현, 02841, 서울시 성북구 안암로 145 고려대학교 통계학과, Tel: 02-3290-1324, Fax: 02-926-3601, E-mail: kurosso@korea.ac.kr
2023년 11월 16일 접수; 2023년 11월 17일 수정본 접수; 2024년 2월 19일 게재 확정.

에도 준수한 성능을 기록하는 사전 학습 모델인 Llama를 각 도메인에 맞게 미세조정 하여 활용하는 연구가 활발히 수행되고 있다(Mukherjee *et al.*, 2023; Lee *et al.*, 2023; Zheng *et al.*, 2023). 그러나, 스포츠 관련 국문 데이터셋의 부재와 컴퓨팅 자원의 제한으로 인해 축구 도메인에 특화된 언어 모델 구성에는 큰 제약이 따른다. 따라서, 언어 모델 역량의 빠른 발전에도 불구, 국내 스포츠 분석 업계 종사자들이 그 효능을 온전히 누리기 힘든 상황이다.

따라서, 미세 조정 없이 언어 모델의 생성 과정을 제어하는 프롬프팅(prompting)과 Retrieval Augmented Generation(RAG, Lewis *et al.*, 2021)은 앞서 언급한 문제를 해결할 수 있는 현실적인 방법론으로 주목 받고 있다. 프롬프팅은 모델에 입력되는 지시사항을 제어하는 것으로, 언어모델이 생성하는 결과물의 품질에 큰 영향을 미친다(Zhang *et al.*, 2023; Zhou *et al.*, 2022). 따라서, 대형 언어 모델에 대한 연구가 가속화됨에 따라 프롬프팅의 중요성이 더욱 강조되고 있다(Naveed *et al.*, 2023). 최근에는 주어진 질문에 대해 모델이 스스로 추론 과정을 도출하며 문장을 생성하도록 제어하는 기법인 Chain-of-Thoughts(CoT)가 언어 모델의 자연어 생성(natural language generation) 평가(evaluation) 과업에서 좋은 성능을 달성한 것으로 알려져 있다(Wei *et al.*, 2022). 또한, RAG는 retrieval을 사용하여 주어진 쿼리(query)와 관련성이 높은 문서를 외부 데이터베이스에서 가져와 언어 모델에 참고 자료로 제공한다. 해당 방법론은 모델의 문장 생성 과정에서 최신 정보를 반영하고, 할루시네이션을 일부 완화할 수 있기 때문에 실질적으로 언어 모델의 생성 역량을 증강하는 효과를 가진다(Chen *et al.*, 2023; Jiang *et al.*, 2023; Shuster *et al.*, 2021).

본 연구는 사전 학습된 언어 모델을 기반으로 미세조정 없이 프롬프트 제어와 CoT 기법을 응용하여 축구 관련 한국어 질의에 대해 적합한 답변을 제공하는 RAG 기반의 프레임워크인 FT-LLM을 제시한다. 언어 모델이 문서 정보를 참조하여 질의(query) 답변을 위한 전략을 구성하고, 이를 기반으로 최종 답변을 생성하도록 함으로써 정제되지 않은 문서의 정보를 효율적으로 활용할 수 있게 한다. 대형 언어 모델을 축구 도메인으로 미세조정 하는 것이 현실적으로 불가한 현실에서 RAG와 CoT를 결합한 FT-LLM은 합리적인 대안을 제공한다. 구체적으로, 기존 RAG 연구와 대비하여 FT-LLM의 차이점은 다음 2가지 부분으로 정리할 수 있다.

- 기존 RAG 연구에서 제안한 질의 refinement 방법론을 개선하였다.

RAG는 retrieval의 성능에 의해 크게 영향을 받으며, 특히 입력되는 질의의 형태에 따라 크게 영향을 받을 수 있다. 가령, 비문이나 비학술적인 용어가 포함된 질의가 입력될 경우 연관성이 매우 낮은 문서가 반환되는 한계가 있다(Mao *et al.*, 2023). 이를 방지하기 위해 본 연구와 동일하게 정제되지 않은 온라인 문서를 retrieval 데이터 셋으로 활용한 RETA-LLM(Liu *et al.*, 2023) 연구를 참고하여 Query Refinement 모듈을 구성하였다. 구체적으로, RETA-LLM은 프롬프트 상으로 "의미가 명료하지 않은 질의를 수정"하도록 요청한다. 반면, FT-LLM의 Query Refinement는 언어 모델에 "축구 전문가"라는 역할 (role) 부여한 뒤 답변을 요청하는 점에서 기존 연구와 차이가 있다. 이 같은 역할 부여는 프롬프트 내에서 언어 모델에 역할을 부여할 경우 질의에 대한 모델의 답변이 품질이 향상된다는 기존 연구 결과(Zheng *et al.*, 2023)를 참고한 것이다. 또한, 프롬프트 내에서 사용자의 질의를 언어 모델에 입력하기 적합한 형태로 변환하도록 직접적으로 명시하는 점에서 RETA-LLM과 차이를 갖는다.

- 언어 모델이 제공 받는 문서를 선택적으로 활용하는 모듈을 활용하였다.

Retrieval가 연관되지 않은 정보를 제공하지 않을 경우 언어 모델의 할루시네이션 가능성은 되려 높아진다(Zhao *et al.*, 2023; Zhou *et al.*, 2021). RETA-LLM은 이를 완화하기 위해 언어모델이 생성한 답변의 사실 여부를 판단하는 Fact-Checking 모듈을 사용하였다. 해당 모듈은 retrieval가 제공한 문서를 기반으로 모델 답변의 사실 여부를 판단하고, 옳지 못한 정보가 포함되어 있는 경우 답변을 제공하지 않는다. 반면, 본 연구에서 제안하는 Confidence Check 모듈은 연관성이 낮은 외부 정보는 선택적으로 배제한다. 답변 생성 이전에 retrieval이 충분히 유용한 정보를 제공했는지 판단함으로써 RAG의 답변 품질 변동성을 감소하는 효과를 보인다.

FT-LLM에서 적용된 CoT 프롬프팅 전략은 모델이 답변 과정에서 스스로 추론하도록 하기 때문에 언어 모델의 할루시네이션을 완화하는 효과를 거둘 수 있다. 특히, 모델이 스스로 만든 답변 생성 전략을 기반으로 retrieval가 추출한 외부 문서의 내용과 쿼리와의 연관성을 각각 평가할 수 있다. 이를 통해 답변의 할루시네이션의 정도, 설명력과 신뢰도를 가늠할 수 있는 이점을 가진다. 따라서, 이러한 장치를 응용함으로써 FT-LLM 프레임워크는 미세조정 없이 축구 도메인에 대한 언어 모델의 추론 능력을 증강시키는 효과를 거두었다.

본 논문의 구성은 다음과 같다. 제2장에서는 retrieval을 활용하여 언어 모델의 생성 능력을 증강한 연구와 프롬프트 제어에 관한 기존 연구에 대해 소개한다. 제3장에서는 FT-LLM에서 활용하는 retrieval의 외부 데이터 셋을 소개하고, 프레임워크 각 단계의 프롬프트 전략의 세부 방법론에 대해 상세히 설명한다. 제4장에서는 평가에 활용할 QA 데이터 셋 구축 방식과 해당 데이터 셋을 활용한 FT-LLM의 실험 결과를 소개하고 분석한다. 제5장에선 본 연구의 결론, 의의 및 추후 연구 방향을 제시한다.

2. 관련 연구

2.1 Retrieval

Retrieval은 주어진 쿼리와 가장 유사한 k개의 문서를 반환

하는 과업이다. 오픈 도메인 질의응답 과업에서는 데이터 베이스 질문에 대한 정답을 포함하는 문서를 높은 유사도를 가진 것으로 판단한다. 이 섹션에서는 다양한 retrieval의 종류와 방법론을 소개한다.

(1) SPR (BM25)

Sparse Passage Retrieval(SPR)는 sparse representation을 활용해 문서와 쿼리의 유사도를 측정하며, 대표적으로 BM25 (Robertson and Zaragoza, 2009) 방법론이 있다. 이 방법론은 TF-IDF(Jones, 1972)를 기반으로 유사도를 측정하며, 연산량이 적어 대규모 데이터 셋에 적용하기 적합하다. 언어에 대한 제약이 없는 장점이 있지만, 단어 빈도를 기반으로 유사도를 측정하기 때문에 의미적(semantic) 유사도를 제대로 반영하지 못하는 단점 역시 있다.

(2) DPR

Karpukhin *et al.*(2020)은 Natural Language Understanding(NLU)에 탁월한 성능을 보이는 BERT(Devlin *et al.*, 2019)를 기반으로 문장의 내포된 의미를 더 잘 포착할 수 있는 Dense Passage Retrieval(DPR)를 제안하였다. 구체적으로, 문서와 쿼리를 각각 언어모델에 입력하여 representation을 얻은 후 유사도를 계산한다. DPR은 각 도메인에 특화된 retrieval를 학습할 수 있는 이점을 가지나, BM25에 비해 연산 자원이 많이 필요한 단점이 있다.

(3) Embedding Model

사전 학습된 언어 모델의 embedding을 retrieval에 직접 활용하는 방안 역시 꾸준히 시도되고 있다. 지도학습 방식으로 다양한 과업에 보편적으로 활용할 수 있는 embedding을 학습하는 연구가 초기에 시도되었으며(Cer *et al.*, 2018; Conneau *et al.*, 2017; Kalchbrenner *et al.*, 2014), 최근에는 contrastive learning 방식으로 언어 모델을 학습하는 연구가 다수 제안되고 있다(Gao *et al.*, 2021; Kim *et al.*, 2021; Ni *et al.*, 2022). 대표적인 embedding API로는 Aleph-Alpha의 luminous (https://docs.aleph-alpha.com/docs/introduction/luminous)와 OpenAI의 text-embedding-ada-002 (Ada-002; Neelakantan *et al.*, 2022)가 있다. 대량의 데이터로 학습된 embedding API 모델들은 전반적으로 passage retrieval 과업에서 높은 성능을 기록하지만, 토큰 개수를 기준으로 API 사용 비용이 부가된다는 한계가 있다(Kamallo *et al.*, 2023).

2.2 RAG 관련 연구

언어 모델의 뛰어난 문장 생성 능력을 실용적으로 활용하기 위해 외부 정보를 적절히 활용하는 연구는 꾸준히 이어지고 있다(Jiang *et al.*, 2023; Chen *et al.*, 2023; Lewis *et al.*, 2020; Liu *et al.*, 2023). 구체적으로, RAG는 retrieval에 DPR을 활용하여 입력과 연관성이 높은 외부의 정보를 언어 모델에 제공하여 BART(Lewis *et al.*, 2020) 모델의 할루시네이션을 완화하고, 생성 능력을 개선하였다. Liu *et al.* (2023)은 대학생을 위한 정신과 상담 챗봇을 개발하는 과정에서 연관성이 높은 외부 정보를 프롬프트 내에 reference로 제공하여 언어모델이 도메인 지식을 풍부하게 활용할 수 있는 프레임워크인 RETA-LLM을 제안하였다.

2.3 Prompting 관련 연구

언어 모델의 생성 능력은 꾸준히 발전하고 있지만, 생성된 문장의 품질은 여전히 쿼리와 사용자의 지시사항을 통합하는 프롬프팅에 의해 크게 좌우된다. 구체적으로, 프롬프팅은 미세조정 없이 언어 모델의 입출력 구조를 원하는 대로 제어하고 생성 과정에 크게 관여할 수 있기 때문에 최근 크게 각광을 받고 있다.

Wei *et al.*(2022)은 언어모델이 인간과 유사한 방식으로 추론할 수 있도록 하는 CoT를 제안하였다. 인간이 문제를 해결하는 과정에서 중간에 여러 풀이 과정을 거치듯, CoT는 언어 모델이 문장을 생성할 때 스스로 답변을 위한 풀이 과정을 생성하도록 하는 프롬프팅 전략이다. Zhao *et al.*(2023)은 외부 정보를 기반으로 모델이 생성한 풀이 과정을 다시 수정하여 최종적인 답변을 생성하는 프레임워크를 제안하였다. Lyu *et al.*(2023)은 기존 CoT가 모델이 답변을 생성하기까지의 추론 과정을 완벽하게 파악할 수 없다는 한계를 지적하여 이를 보완하기 위한 방법론을 제안하였다. 구체적으론, 입력된 쿼리를 reasoning chain으로 변환하는 Translation과 이를 다시 최종 답변으로 변환하는 Problem Solving 모듈로 프레임워크를 구성하여 언어 모델의 정확도와 설명력을 비약적으로 향상시켰다.

프롬프트 구성 요소의 비교적 미세한 변화만으로도 언어 모델의 생성 능력이 크게 좌우될 수 있다. 따라서, 언어 모델의 문서 요약본을 평가하는 평가 과업(evaluation task)에선, 전문가의 도움을 받아 요약본의 평가 요소인 aspect에 대한 정의와 프롬프트의 양식(template)을 구성하였다(Fabbri *et al.*, 2021; Fu *et al.*, 2023; Liu *et al.*, 2023). 따라서, 프롬프트 관련 연구 초기 단계에서는 각 도메인에 적합한, 효율적인 프롬프트 작성을 위해 도메인 지식을 충분히 갖춘 전문가의 개입이 필요하다.

2.4 스포츠 관련 언어 모델 연구

Connor and O'Neill(2023)은 ChatGPT 공개 이후 스포츠 업계에서의 언어 모델 응용 가능성과 다양한 언어 모델의 장단점에 대해 논하였다. SportsBERT(https://huggingface.co/ microsoft/SportsBERT)는 800만 건에 달하는 4년치 스포츠 관련 기사로 BERT(Devlin *et al.*, 2019)를 재학습시켰다. Unlu(2023)는 축구 통계 데이터 셋을 활용하여 언어 모델을 미세 조정함으로써 축구 통계 관련 질의에 특화된 언어 모델 프레임워크를 제안하였다. 이처럼, 스포츠를 주제로 언어 모델을 연구한 내

용이 공개된 사례는 그리 많지 않으며, 본 연구를 통해 해당 연구 주제에 대해 자세히 탐구하고자 한다.

3. 방법론

본 섹션은 retrieval 데이터베이스(Database) 구성을 위해 사용한 데이터와 FT-LLM을 구성하는 각 모듈에 대해 설명한다. FT-LLM의 각 모듈에 입력되는 프롬프트에는 언어 모델이 비정제된 입력과 참조(reference)를 처리할 수 있도록 본 과업에 대한 정보를 포함시켰다.

먼저, 사용자의 질의는 Query Refinement 모듈에 입력되어 언어 모델이 이해하기 적합한 형태로 변형된다. 이후, Passage Retrieval 모듈은 retrieval 데이터베이스에서 쿼리와 가장 연관성이 높은 k개의 문서를 추출한다. Confidence Check 모듈은 제공된 k개의 문서로 쿼리를 충분히 답변할 수 있는지 평가하며, 만약 정보가 충분하지 않다면 언어 모델이 외부 정보 없이 쿼리에 대한 답변을 바로 생성한다. Strategy Generation 모듈은 제공된 k개의 문서를 기반으로 답변 생성 전략을 도출한다. Response Generation 모듈은 도출된 답변 생성 전략을 기반으로 쿼리에 대한 최종 답변을 생성한다. 본 프레임워크의 작동 방식은 <Figure 1>에서 확인할 수 있다.

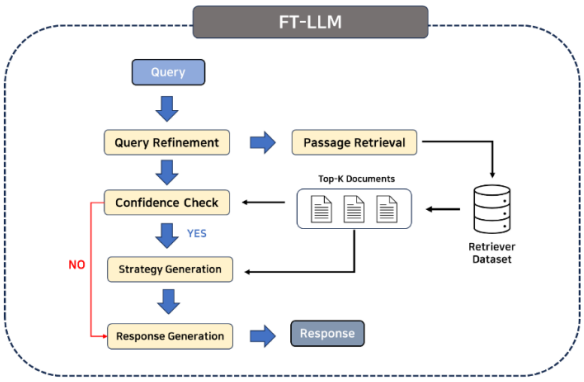

Figure 1. Overall Architecture of FT-LLM

3.1 Retrieval Dataset

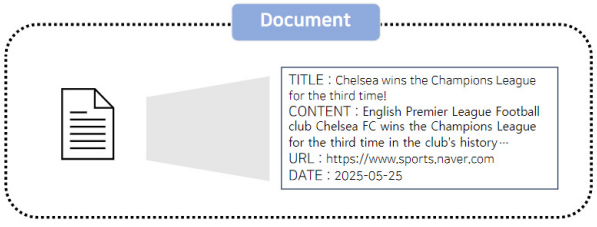

Figure 2. Example of a Retrieval Dataset Document

Retrieval 데이터베이스는 2019년 1월부터 2023년 7월까지의 기간동안 작성된 435,599건의 네이버 뉴스 해외축구 기사로 구성된다. 수집된 뉴스 기사는 문서 제목, 작성 일자, URL, 본문 내용으로 정리되어 각각 json 파일에 저장되었다.

3.2 Query Refinement

Query Refinement 모듈은 사용자의 질의를 모델이 이해하기 쉽고 정리된 형태로 변환한다. 언어 모델의 생성 결과는 프롬프트의 내용과 방식에 크게 영향 받기 때문에 질의를 언어 모델이 이해하기 적합한 구조로 구성하는 것이 매우 중요하다 (Liu *et al.*, 2023; Amplayo *et al.*, 2022). 그러나, 본 프레임워크의 실사용자인 축구 관련 산업 종사자들은 프롬프트의 중요성을 완전히 인지하기 어려운 점을 고려하여 이와 같은 모듈을 구성하였다. 또한, 사람 읽었을 때 의미상으로 명료한 문장 역시 언어 모델이 질의의 의도를 가장 파악하기 적합한 형태로 변형하는 목적을 가진다. Query Refinement에 사용하기 위해 <Table 1>과 같은 프롬프트를 구성하여 언어 모델에 입력함으로써 최초 입력된 유저의 질의를 변환하였다.

Table 1. Prompt Template for Query Refinement Module

Prompt
You are a human expert on football related knowledge. All your answers should be in Korean You will be given a football related query. The query may not be in an appropriate form to be used in LLM. Your job is to refine the query so that it can be used in LLM. **Query:** **Refined query:**

3.3 Passage Retrieval

본 모듈은 구성한 retrieval 데이터셋으로부터 쿼리와 가장 연관성이 높은 문서를 반환한다. 쿼리와 가장 연관도가 높은 문서를 탐색하기 위해 BM25를 활용하였으며, 식 (1)에 따라 주어진 쿼리와 각 문서 간의 BM25 score를 산출한다. 함수 f는 쿼리 문서 D에서 항목 q_i의 용어 빈도, $|D|$는 단어 기준 문서 D의 길이를 의미한다. $Avgdl$와 n은 각각 주어진 문서 집합 내에서 평균 문서 길이와 전체 문서 개수를 의미한다. 이때, k_1과 b는 각각 용어 빈도의 포화(saturation)와 문서 길이의 정규화 정도를 조절하는 매개변수이다.

$$score(Q,D) = \sum_{i=1}^{n} iDF(q_i) \cdot \frac{f(q_i, D) \cdot (k_1+1)}{f(q_i, D) + k_1 \cdot (1-b+b \cdot \frac{|D|}{avgdl})} \quad (1)$$

실제 코드 구현 단계에선 KLUE 벤치마크 데이터셋으로 미세 조정된 BERT(https://huggingface.co/klue/bert-base)의 토크

나이저를 활용하여 각 문서와 쿼리를 벡터화한 뒤, rank_bm25 라이브러리의 BM25Okapi(https://pypi.org/project/rank-bm25/) 함수를 활용하여 BM25 score를 산출하였다. 이후, retrieval 데이터셋 중 BM25의 점수가 가장 높은 k개의 문서를 언어 모델의 생성 능력 증강을 위한 참조로 활용하였다.

Embedding model 기반 retrieval을 위해 사전 학습된 모델로 뉴스 기사의 embedding을 취득한 뒤 이를 Faiss 라이브러리의 index로 저장하였다. 이후, 동일한 모델을 활용하여 입력된 질의의 embedding을 구한 뒤, 저장된 index에서 질의와 코사인 유사도가 가장 높은 k개의 문서를 참조로 활용하였다.

3.4 Confidence Check

Confidence Check 모듈은 프롬프트를 언어 모델에 입력하여 Passage Retrieval 모듈이 반환한 문서가 쿼리에 대한 적절한 답변을 제공할 만큼 충분하고 유의미한 정보를 포함하고 있는지 판단한다. 낮은 연관성을 가진 문서가 참조로 제공될 경, 언어 모델의 생성 능력은 되려 악화될 수 있다. 따라서, 문서에 유의미한 정보가 부재할 경우 참조를 제공하지 않게 된다. 즉, 문서의 신뢰도(confidence)가 낮다면 문장 생성 과정에서 기본 언어 모델이 활용된다.

3.5 Strategy Generation

Strategy Generation 모듈은 <Table 2>의 프롬프트를 주어진 참조와 함께 언어 모델에 입력하여 답변 생성 전략을 도출한다. BM25는 키워드 기반 유사도로 문서를 검색하기 때문에 반환된 문서가 답변 생성에 직접적으로 유용한 단서를 제공하지 않을 수 있다. 따라서, 언어 모델이 passage 내 쿼리와 관련된 핵심 정보를 스스로 파악하는 방식인 CoT(Wei *et al.*, 2022) 방법론을 참고하여 답변 생성 전략(Answer Generation Strategy)를 설계하고, 이를 최종 답변 생성 프롬프트에 제공하였다.

Table 2. Prompt Template to Generate Response Generation Strategy

Prompt
You are a human expert on football related knowledge. All your answers should be in Korean You will be given several passages and a query. Your job is to generate three step strategy that can be used to extract the most relevant fragment of the passage to the query. Each step must refer to specific sentences of the passage. You will have to incorporate relevant information from “all” passages in the strategy. **Query:** **Passage:**

3.6 Response Generation

Response Generation 모듈은 Strategy Generation 모듈으로부터 도출된 답변 생성 전략을 기반으로 언어 모델이 쿼리에 대한 최종 답변을 생성하는 과정을 포함한다. 본 모듈은 FT-LLM이 다루는 도메인과 답변 생성 전략의 작성 과정에 대한 설명을 <Table 3>과 같이 프롬프트 내에 제공하여 쿼리, 답변 생성 전략 및 답변 작성 지시사항을 통합한다.

Table 3. Prompt Template for Response Generation Module

Prompt
You are a human expert on football related knowledge. All your answers should be in Korean Based on the given strategy, Answer the query. However, do not solely rely on it. Therefore, if you think the passage does not contain enough information to answer the query, you can use your own knowledge to answer the query. Bear in mind that you are given a limited amount of knowledge and the provided passage may not contain all the information you need. The strategy may or may not contain a direct answer to the query. However, you may find semantically similar information. Therefore, think about what synonyms or paraphrases could be used to answer the query. Ensure your response is precise, detailed, and based on the information from the passages. All you answers should be in Korean **Query:** **Strategy:**

이 모듈은 결과물은 언어 모델에 의해 생성된 답변, Strategy Generation에서 참조된 외부 문서의 본문 및 원천, 그리고 답변 생성 전략을 반환한다. 이를 통해 사용자는 언어 모델의 추론 과정과 답변의 근거를 검토하며 답변의 정확도를 평가할 수 있다. 또한, 연관된 문서는 쿼리와 관련하여 유용한 정보를 추가로 제공할 수 있다.

4. Experiments

4.1 실험 환경

실험에 사용할 대형 언어 모델로는 OpenAI에서 API 서비스로 제공하는 모델인 gpt-3.5-turbo(GPT-3.5)와 gpt-4(GPT-4)를 선정하였다. 이 밖에도, 비교적 작은 모델 크기로도 좋은 언어 생성 능력을 보이는 것으로 알려진 Llama2나 Orca-13b을 후보군로 고려하였지만, 해당 모델들의 사전 학습 과정에서 한글 데이터 셋의 비중이 매우 낮아 본 연구의 과업을 위해서는 적절하지 않다 판단하여 배제하였다.

Passage Retrieval 모듈의 embedding model은 OpenAI의 text-embedding-ada-002(Neelakantan *et al.*, 2022) API를 활용하였다. Retrieval이 반환할 외부 문서 개수 인자인 k는 3으로 설정하였다(k=3). k의 크기를 바꾸어 가며 retrieval이 반환하는 문서의 내용을 정성적으로 비교했을 때 top-3 문서까지 질의와 연관성이 있다 판단하였기 때문이다. 해당 k의 크기에 따른 성능 비교 실험 결과는 부록 B에서 확인할 수 있다.

Strategy Generation 모듈에선 문서가 너무 길어 최대 토큰 개수 제한을 초과할 수 있기 때문에 문서를 토큰 제한 길이를 고려하여 분할하고, 각 부분에 대한 전략을 독립적으로 생성하였다. 이후 Response Generation 단계에선 생성된 전략들을 한 번에 프롬프트에 입력하였다. 효율적인 언어모델 추론(Inference)을 위해 Guidance(https://github.com/guidance-ai/guidance)를 활용하였다. Guidance는 탑재된 언어모델의 입출력과 하이퍼파라미터를 비교적 간편하게 확인하고 제어할 수 있는 기능을 제공하는 파이썬 라이브러리이다. 본 연구와 같이 다양한 프롬프트를 활용하여 실험을 수행할 때 효율성을 증대시켜준다. 그 외 주요한 실험 환경 설정에 대한 정보는 <Table 4>에서 확인할 수 있다.

Table 4. List of Library and Version Utilized for the Experiment

Library	Version
Guidance	0.0.64
Transformers	4.26.1
Faiss	1.7.4
torch	1.11.0

(1) QA 데이터셋 구축

실험을 위해 구축한 QA 데이터 셋은 <Table 5>에서 보이는 것과 같이 답변이 명확한 "객관적" 질의와 기반 지식을 토대로 추론을 요구하며 명확한 답이 정해져 있지 않은 "주관적" 질의로 구성하였다.

객관적 질의를 구성하기 위해 총 4가지 축구 지식의 대분류(월드컵 관련 질문, 축구 규칙 관련 질문, 축구 전술 관련 질문, 팀 및 선수 관련 질문)를 설정한 뒤, 3명의 대학 스포츠 통계 학회원로 구성된 작업자가 각 대분류에 대한 질의를 각각 10개씩 작성하였다. 또한, 모델이 retrieval을 활용하여 최신 정보를 요하는 질의에 올바르게 답변할 수 있음을 검증하기 위해 gpt-3.5-turbo와 gpt-4 API의 knowledge cutoff 시기와 동일하게 "2021년 9월 이후 발생한 내용을 파악해야 올바르게 답변을 할 수 있는 질문"을 작업자당 10개씩 구성하였다. 해당 유형의 질의는 선수의 소속, 기록 등 시간의 흐름에 따라 변동이 있는 내용을 포함한다. 주관적 질의의 경우 모델의 창의력과 추론 능력을 측정하기 위해 축구 전술, 규칙, 팀, 선수에 대한 예측을 요구하는 내용을 반드시 포함하도록 작성되었다.

수합한 질의 중 크롤링을 통해 수집한 축구 관련 위키피디아 문서 내의 정보를 토대로 정답을 직접적으로 찾거나 유추할 수 있는 질의만 남긴 다음에 각 대분류에서 동일한 개수를 층화 추출하였다. 이후, 내용상으로 서로 중복되는 질의를 제거하여 총 61개의 질의를 구성하였으며, 질의의 양식과 어조은 KorQuAD(Lim *et al.*, 2018)와 유사하도록 최종적인 수정을 거쳤다. 객관적 질의는 본 연구의 2저자가 직접 연관 있는 위키피디아 문서를 참고하여 골든 레이블(Golden Label)을 배정하였다. 주관적 질의에 대한 모델 답변의 품질 정성 평가를 위해서 축구 분석 업계 전문가의 답변을 제공 받아 Human Reference로 활용하였다.

Table 5. Example of Question and Answer Pair

Objective Question	Subjective Question
Question Which team won win the 2022 World Cup? **Answer** Argentina	**Question** Which Bundesliga team made the most efficient signings in the summer 2022 transfer window and how did they change the team's performance? **Answer** While top clubs like Munich, Dortmund, and Leipzig always make efficient signings, I would single out Freiburg's signings of Matthias Ginter, Doan Ritsu, and Michael Gregoritsch as particularly effective. After hovering around the mid-table every season, Freiburg finished the previous season, 21-22, in 6th place, and then finished the 22-23 season in 5th place, thanks to some good signings in the transfer market, which solidified Freiburg's position in the mid-table.

4.2 평가 방식

본 파트는 QA 데이터 셋에 대한 모델 답변 품질을 평가하기 위해 적용한 평가 방식에 대해 소개한다. 골든 레이블과 모델의 출력 간 의미적 유사도를 측정하기 위해 F1 지표 외에도 대학 스포츠 통계 분석 학회에서 모집한 평가자(Human annotator) 3명의 정성 평가와 언어모델 기반 지표(LLM-based Metric)로 FT-LLM의 답변 품질을 평가하였다. 또한, 객관적 질의와 주관적 질의에 대해 각각 다른 평가 방식을 적용하였다.

(1) 객관적 질문 평가 방식

QA 작업에서 F1 스코어는 답변의 정확성과 완전성을 동시에 평가하는 지표이다. 후보 답변의 토큰 중 정답과 일치하는 토큰의 비율인 정밀도(Precision)와 정답 토큰 중 후보 답변에 포함된 토큰의 비율인 재현율(Recall)의 조화 평균으로 계산된다.

Acc_human은 3명의 평가자가 직접 질의, 정답, 모델이 생성한 답변을 모두 고려하여 모델의 답변의 적중 여부를 평가하였으며, 공정성을 위해 각 답변을 생성한 모델의 종류는 알리지 않았다. 이후, 모든 평가자들의 답변을 다수 보팅(Majority

voting)하여 각 모델의 답변에 대한 최종적인 평가 결과를 0(오답) 혹은 1(정답)의 정수로 부여한다. Acc_GPT는 Table 6의 프롬프트를 GPT-4에 입력하여 객관적 질의에 대한 모델 답변의 적중 여부를 평가하는 방식의 지표로 LLM 기반 평가를 다룬 Kamallo *et al.*(2023)의 연구를 참고하였다.

Table 6. Prompt Template for Formulation of Acc_GPT metric

Prompt	**Question:** **Answer:** Is candidate correct? Answer 0 for No and 1 for Yes
Answer	1

(2) 주관적 질문 평가 방식

언어 모델이 생성한 답변은 잠재적으로 축구 산업에 종사하는 관계자들의 업무를 보조하기 위함이다. 따라서, 주관적 질의에 대해서는 절대적인 답변의 옳고 틀림을 평가하는 대신, <Table 7>에서 보이는 정의에 따른 Factuality, Relevance, Reasoning, Interestingness의 평가 요소(aspect)를 토대로 하여 세분화된 평가를 진행하였다. Factuality, Relevance는 GPTScore(Fu *et al.*, 2023)에서 제안한 정의를 QA evaluation에 적합하게 변형하였으며, 모델 답변의 할루시네이션과 창의성에 대한 평가를 반영하기 위해 Reasoning과 Interestingness 요소를 추가하였다. 각 평가 요소의 정의는 다음과 같다.

Table 7. Evaluation Aspects Used to Evaluate Response for Subjective Questions, Generated by the Language Model and Their Corresponding Definitions

Aspect	Definition
Factuality	How well is the response factually grounded?
Relevance	How well does the answer align with what's in the Golden Label?
Reasoning	How natural and logical is the rationale for the content in the response, compared to the human reference
Interestingness	Does the response provide insight or information that is as interesting or novel as the human reference?

각 평가자가 각 평가 요소에 대해 1(worst)에서 5(best) 사이의 정수로 점수를 매긴 뒤, 이들의 평균을 답변 품질 점수로 부여하였다. 이러한 평가 요소를 기반으로 언어모델 QA 답변 평가 지표인 Eval_human를 구성하여 3명의 평가자가 각 평가 요소에 대해 매긴 점수의 합산 평균(5점 만점)을 보고하였다.

4.3 실험 결과 분석

FT-LLM의 성능을 평가하기 위해 진행한 실험의 정량적 평가 결과는 <Table 8>에서 확인할 수 있다. GPT-3.5, 4는 다른 모듈이 전혀 적용되지 않은 OpenAI의 gpt-3.5-turbo와 gpt-4 API 모델을 의미한다. FT-LLM-3.5, 4는 FT-LLM 프레임워크를 gpt-3.5-turbo와 gpt-4 API 모델에 각각 적용한 방법론을 의미한다. Retrieval Type의 Ada-002와 BM25는 Passage Retrieval에서 사용한 retrieval 종류에 따른 실험 결과를 의미한다.

Confidence Check 모듈을 적용할 경우 confidence가 “No”인 질의에 대해서는 FT-LLM과 GPT의 답변이 동일하게 된다. 이 경우, 외부 정보를 활용한 CoT 기반 답변 생성 전략의 효과를 확인할 수 없다. 따라서, 두 모델의 성능을 공정하게 비교하기 위해 <Table 8>에는 Confidence Check 모듈이 적용되지 않은 FT-LLM의 실험 결과를 보고하였다. Confidence Check 모듈을 적용한 FT-LLM의 실험 결과는 4장 4절 (1)에서 확인할 수 있다.

Table 8. Main Experiment Results

Model Name	Retrieval Type	Objective Questions			Subjective Questions
		F1	Acc_ Human	Acc_ GPT	Eval_ Human
GPT-3.5	-	0.0279	0.4565	0.5000	3.8465
GPT-4		0.0352	0.5870	0.5435	3.9167
FT-LLM-3.5	Ada-002	0.0457	0.6522	0.5652	3.7916
	BM25	0.0390	0.4783	0.3913	3.5965
FT-LLM-4	Ada-002	0.0587	0.6739	0.5870	4.0833
	BM25	**0.0670**	**0.8478**	0.6957	**4.6886**

<Table 8>을 확인하면, 모델의 크기에 관계 없이 FT-LLM 프레임워크를 적용하였을 때 F1이 상승하였다. 사람이 직접 객관적 질문에 대한 정답 여부를 평가한 Acc_Human 지표의 경우, BM25를 활용한 FT-LLM-4가 가장 높은 수치인 0.8478을 기록하였다. 언어모델이 정답 여부를 평가한 Acc_GPT 지표상에서도 해당 방법론은 0.6957을 기록하여 GPT-4 대비 매우 높은 성능을 기록하였다. 또한, 주관적 질문에 대해서도 BM25 기반 FT-LLM-4는 Eval_human가 가장 높은 5점 만점에 4.6886를 기록하였다.

BM25를 retrieval에 사용했을 때 기반 모델이 바뀜에 따라 모든 지표 상에서 성능 차이가 비교적 컸다. 반면 Ada-002를 사용했을 땐 기반 모델 변경에 따른 성능 변화가 비교적 미미했다. FT-LLM-3.5를 활용할 경우 Ada-002 기반 retrieval가 BM25을 기반으로 한 경우보다 모든 지표에서 우월한 성능을 기록하였다. 보다 큰 기반 모델을 사용한 FT-LLM-4는 정반대의 결과를 기록하였다. 이는 NLG, NLU 역량이 보다 우수한 GPT-4가 BM25 retrieval이 제공하는 원문 문서의 내용을 답변 생성 과정에 더욱 효율적으로 반영하기 때문인 것으로 추측된다. 위와 같은 결과는 Lin *et al.*(2023)이 보고한 sentence embedding API 기반 passage retrieval 비교 실험 결과와 일맥상통한다. 이러한 실험 결과는 도메인과 데이터 셋에 따라 passage retrieval 방법론들은 결과적으로 각각 RAG에서 다른 영향을 끼

칠 수 있음을 보여준다.

<Table 8>에 기록된 실험 결과는 NLG, NLU 역량이 뛰어난 언어 모델을 사용한다면, 적어도 본 연구에서 활용한 retrieval 데이터 셋에 대해선 BM25가 모델의 역량을 증강하는데 있어 더욱 유용한 문서를 제공하였다 추측할 수 있다. 따라서, 연구에서 다루는 동일한 도메인과 과업에 대한 미세조정을 거치지 않은 embedding 모델을 축구 관련 질의 정보 retrieval에 활용하는 것은 BM25를 활용하는 것에 비해 좋은 성능을 보장하지 않았음을 확인할 수 있다.

결론적으로, FT-LLM은 retrieval 종류에 따라 다른 효과를 보였지만 여전히 기본 언어 모델의 기본적인 축구 배경 지식을 풍부하게 해 줄 뿐만 아니라, 답변 작성 메커니즘에 창의력을 부여하고, 생성 문장의 할루시네이션을 완화할 수 있는 효과를 거두었다.

4.4 Ablation Study

본 장에서는 FT-LLM을 구성하는 각 모듈의 모델 성능 향상 기여 여부와 그 정도를 평가하기 위한 ablation 실험 결과를 소개한다.

(1) Confidence Check 모듈의 효과

Confidence Check 모듈은 주어진 정보가 질의와 연관이 없음에도 문서를 참고할 경우, 언어 모델의 생성 문장에 할루시네이션이 포함될 수 있는 문제를 개선하는 목적을 가진다. 해당 모듈의 효과를 확인하기 위한 실험을 진행하였으며 그 결과는 <Table 9>와 <Table 10>에서 확인할 수 있다.

<Table 9>에 기록된 실험에선 BM25로 retrieval을 수행하였다. FT-LLM-3.5에 Confidence Check를 적용했을 때 모든 수치가 증가하였다. 따라서, 언어 모델이 도움이 되지 않는 문서를 참고하지 않으면 할루시네이션을 완화할 수 있음을 보여준다. 다만, FT-LLM-4의 경우, Confidence Check 적용 유무에 따른 성능 변화가 그리 크지 않았다. 이는 Confidence Check를 적용한 FT-LLM-4에 비해 3.5가 retrieval로부터 제공 받은 문서를 참고하지 않은 비율이 약 4배 많았다는 점에서 원인을 유추할 수 있다. NLG, NLU 역량이 이미 높은 모델은 문서의 내용을 파악하여 응용하는 역량 역시 높기 때문에, 제공 받은 문서를 거절할 가능성이 낮기 때문이다.

<Table 10>에서처럼, retrieval를 Ada-002의 embedding로 구성하면 Confidence Check를 적용했을 때 모든 모델 크기에서 FT-LLM의 성능이 상승하였다. 이 경우, retrieval가 제공하는 문서를 되려 사용하지 않는 것이 FT-LLM의 답변 품질을 향상시켜주는 경우가 비교적 많았음을 의미한다. 따라서, 본 ablation 실험을 통해 Confidence Check는 retrieval 종류에 따라 변동될 수 있는 RAG의 생성 능력을 보다 안정적으로 유지할 수 있도록 도와줄 수 있는 장치임을 확인할 수 있다.

(2) Query Refinement 모듈의 효과

본 연구에서 제안하는 Query Refinement 모듈은 사용자의 질의를 언어 모델이 이해하기 적합한 형태로 변환한다. Query Refinement의 온전한 효과를 확인하기 위해 본 실험에선 Confidence Check 모듈을 제거한 뒤 기존 실험에서 가장 좋은 성능을 보인 BM25 retrieval 기반 FT-LLM-3.5와 4를 활용한 실험을 진행하였다. 첫 번째로, FT-LLM의 Query Refinement를 제거한 뒤 QA 데이터 셋의 질의에 대한 모델 답변 성능 변화를 확인하였다.

<Table 11>에 제시된 실험 결과는 Query Refinement가 FT-LLM의 생성 능력 향상에 기여하였음을 입증한다. 모델의 크기와 무관하게, 다른 요인을 고정한 상태에서 Query Refinement를 적용하지 않았을 때 성능 지표가 감소하였다. 이는 본 연구에

Table 9. Comparative Experiment Result on FT-LLM-3.5 & 4 with BM25 when Confidence Check Module is Implemented

Model Name	Confidence Check	Objective Questions			Subjective Questions
		F1	Acc_Human	Acc_GPT	Eval_Human
FT-LLM-3.5	w/o Confidence Check	0.0390	0.4783	0.3913	3.5965
	w/ Confidence Check	**0.0541**	**0.5870**	**0.5652**	**3.7500**
FT-LLM-4	w/o Confidence Check	**0.0670**	0.8478	**0.6957**	**4.6886**
	w/ Confidence Check	0.0567	**0.8696**	0.6739	4.6021

Table 10. Comparative Experiment Result on FT-LLM-3.5 & 4 with Ada-002 Embedding when Confidence Check Module is Implemented

Model Name	Confidence Check	Objective Questions			Subjective Questions
		F1	Acc_Human	Acc_GPT	Eval_Human
FT-LLM-3.5	w/o Confidence Check	0.0457	0.6522	0.5652	3.7916
	w/ Confidence Check	**0.0517**	**0.6739**	**0.5652**	**4.1031**
FT-LLM-4	w/o Confidence Check	0.0587	0.6739	**0.5870**	4.0833
	w/ Confidence Check	**0.0763**	**0.7391**	**0.5870**	**4.1041**

Table 11. Comparative Experiment Result on FT-LLM-3.5 & 4 with BM25 when Query Refinement Module is not Implemented for Queries in QA Dataset

Model Name	Confidence Check	Objective Questions			Subjective Questions
		F1	Acc_Human	Acc_GPT	Eval_Human
FT-LLM-3.5	w/o Query Refinement	0.0288	0.4348	0.3696	3.5401
	w/ Query Refinement	**0.0390**	**0.4783**	**0.3913**	**3.5965**
FT-LLM-4	w/o Query Refinement	0.0430	0.7391	0.6087	4.1563
	w/ Query Refinement	**0.0670**	**0.8478**	**0.6957**	**4.6886**

서 제시된 Query Refinement가 언어 모델이 질의의 의도를 보다 효과적으로 파악할 수 있도록 수정하여 답변 품질을 개선하는데 중요한 역할을 한다는 것을 시사한다. 특히, QA 데이터셋에 포함된 질의와 같이 인간이 읽었을 때 의미적으로 명확한 문장들에 대해서도 이 기법이 효과적임을 보여준다.

5. 결 론

본 연구에서 제시한 FT-LLM 프레임워크는 사전 훈련된 언어 모델의 미세 조정 없이도 축구 관련 쿼리에 효과적으로 대응할 수 있는 잠재력을 보여준다. RAG 구조와 CoT 프롬프팅을 응용하여 축구 데이터셋이 부족하거나 도메인 특화 모델을 미세 조정할 수 없는 환경에서도 기존 언어 모델보다 정확하고 관련성 높으며, 최신 정보를 반영한 답변을 생성하는 것이 가능함을 입증하였다.

그러나, 본 연구에서 쿼리와 가장 연관성이 높은 외부 문서를 불러오기 위해 활용한 BM25의 한계는 FT-LLM의 역량은 잠재적으로 제한한다. 즉, retrieval이 유용한 정보를 제공 할 수 없는 경우에는 내용적인 측면에서 기존 언어 모델과 대비하여 큰 이점을 제공할 수 없다. 비록 유용한 정보를 제공받지 않는 경우에도 답변 생성 전략과 Confidence Check 같은 장치 덕분에 할루시네이션을 방지할 수 있었지만, 사용자에게 흥미롭고 참신한 정보를 제공할 수 없는 근본적인 한계를 극복할 수 없게 된다. 반면, 미세조정을 거치지 않은 embedding model의 출력을 그대로 Passage Retrieval에 활용한 경우 BM25을 활용했을 때보단 성능이 저조했다. 궁극적으론 축구 관련 retrieval 학습 데이터 셋을 구성하여 embedding model을 미세조정 하는 방안이 잠재적으론 가장 효과적인 방법론이 될 수 있다. 따라서, 본 연구의 결과를 기반으로 향후에는 더욱 본 과업에 적합한 retrieval을 학습하고, 이를 활용하여 축구 산업 종사자들의 업무를 효율적으로 보조할 수 있는 연구가 이어지기를 기대한다.

참고문헌

Amplayo, R. K., Webster, K., Collins, M., Das, D., and Narayan, S. (2023), Query Refinement Prompts for Closed-Book Long-Form Question Answering, *Annual Meeting of the Association for Computational Linguistics*, 7997-8012.

Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., St. John, R., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Strope, B., and Kurzweil, R. (2018), Universal Sentence Encoder for English, In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 169-174.

Chen, J., Lin, H., Han, X., and Sun, L. (2023), Benchmarking Large Language Models in Retrieval-Augmented Generation, ArXiv, abs/2309.01431.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., and Bordes, A. (2017). Supervised Learning of Universal Sentence Representations from Natural Language Inference Data, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 670-680.

Connor, M. and O'Neill, M. (2023). Large Language Models in Sport Science & Medicine: Opportunities, Risks and Considerations, ArXiv, abs/2305.03851.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019), BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv, abs/1810.04805.

Fabbri, A. R., Kryściński, W., McCann, B., Xiong, C., Socher, R., and Radev, D. (2021, 04), SummEval: Re-evaluating Summarization Evaluation, *Transactions of the Association for Computational Linguistics*, **9**, 391-409.

Fu, J., Ng, S.-K., Jiang, Z., and Liu, P. (2023), GPTScore: Evaluate as You Desire. ArXiv, abs/2302.04166.

Gao, T., Yao, X., and Chen, D. (2021), SimCSE: Simple Contrastive Learning of Sentence Embeddings, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 6894-6910.

Jiang, Z., Xu, F. F., Gao, L., Sun, Z., Liu, Q., Dwivedi-Yu, J., Yang, Y., Callan, J., and Neubig, G. (2023), Active Retrieval Augmented Generation. ArXiv, abs/2305.06983.

Jones, K. S. (1972), A statistical interpretation of term specificity and its application in retrieval, *Journal of Documentation*, **28**(1), 11-21.

Kalchbrenner, N., Grefenstette, E., and Blunsom, P. (2014). A Convolutional Neural Network for Modelling Sentences, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistic*, 655-665.

Kamalloo, E., Dziri, N., Clarke, C., and Rafiei, D. (2023). Evaluating Open-Domain Question Answering in the Era of Large Language Models, *Annual Meeting of the Association for Computational Linguistics*, 5591-5606.

Kamalloo, E., Zhang, X., Ogundepo, O., Thakur, N., Alfonso-Hermelo, D., Rezagholizadeh, M., and Lin, J. (2023), Evaluating Embedding APIs for Information Retrieval, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 518-526.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov , S., Chen, D., and Yih, W-T. (2020), Dense Passage Retrieval for Open-Domain Question Answering, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 6769-6781.

Kim, T., Yoo, K. M., and Lee, S. (2021), Self-Guided Contrastive Learning for BERT Sentence Representations, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2528-2540.

Lee, A. N., Hunter, C. J., and Ruiz, N. (2023). Platypus: Quick, cheap, and powerful refinement of llms, ArXiv, abs/2308.07317.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2019), BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *Annual Meeting of the Association for Computational Linguistics*, 7871-7880.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W-t., Rocktäschel, T., Riedel, S., and Kiela, D. (2020), Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, *Proceedings of the 34th International Conference on Neural Information Processing Systems*.

Lim, S., Kim, M., and Lee, J. (2018). KorQuAD: Korean QA Dataset for Machine Comprehension, *Proceedings of Korean Institute of Information Scientists and Engineers*.

Liu, J., Jin, J., Wang, Z., Cheng, J., Dou, Z., and Wen, J.-R. (2023). RETA-LLM: A Retrieval-Augmented Large Language Model Toolkit, ArXiv, abs/2306.05212.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. (2023), G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. ArXiv, abs/2303.16634.

Lyu, Q., Havaldar, S., Stein, A., Zhang, Li., Rao, D., Wong, E., Apidianaski, M., and Callison-Burch, C. (2023), Faithful Chain-of-Thought Reasoning. ArXiv, abs/2301.13379.

Mao, K., Dou, Z., Mo, F., Hou, J., Chen, H., and Qian, H. (2023), Large Language Models Know Your Contextual Search Intent: A Prompting Framework for Conversational Search, *Findings of the Association for Computational Linguistics: EMNLP 2023*, 1211-1225.

McKenna, N., Li, T., Cheng, L., Hosseini, M., Johnson, M., & Steedman, M. (2023), Sources of Hallucination by Large Language Models on Inference Tasks. ArXiv, abs/2305.14552.

Mukherjee, S., Mitra, A., Jawahar, G., Agarwal, S., Palangi, H., and Awadallah, A. (2023), Orca: Progressive Learning from Complex Explanation Traces of GPT-4, ArXiv, abs/2306.02707.

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Barnes, N., and Mian, A. (2023). A Comprehensive Overview of Large Language Models. ArXiv, abs/2307.06435.

Neelakantan, A., Xu, T., Puri, R., Radfɔd, A., Han, J. M., Tworek, J., Yuan, Q., Tezak, N., Kim, J. W., Hallacy, C., Heidecke, J., Shyam, P., Power, B., Nekoul, T.E., Sastry, G., Krueger, G., Schnurr, D., Such, F.P., Hsu, K., Thompson, M., Khan, T., Sherbakov, T., Jang, J., Welinder, P., and Weng, L. (2022). Text and Code Embeddings by Contrastive Pre-Training. ArXiv, abs/2201.10005.

Ni, J., Abrego, G. H., Constant, N., Ma, J., Hall, K., Cer, D., and Yang, Y. (2022), Sentence-T5: Scalable Sentence Encoders from Pre-trained Text-to-Text Models, *Findings of the Association for Computational Linguistics: ACL 2022*, 1864-1874.

OpenAI (2023). GPT-4 Technical Report. ArXiv, abs/2303.08774.

Robertson, S. and Zaragoza, H. (2009), The Probabilistic Relevance Framework: BM25 and Beyond, *Foundations and Trends in Information Retrieval*, **3**(4), 333-389.

Shuster, K., Poff, S., Chen, M., Kiela, D., and Weston, J. (2021), Retrieval Augmentation Reduces Hallucination in Conversation, *Findings of the Association for Computational Linguistics: EMNLP 2021*, 3784-3803.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023), LLaMA: Open and Efficient Foundation Language Models, ArXiv, abs/2302.13971.

Unlu, E. (2023), FootGPT : A Large Language Model Development Experiment on a Minimal Setting, Arxiv, abs/2308.08610.

Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., and Polosukhin, I. (2017). Attention is All you Need, *31st Conference on Neural Information Processing Systems*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. (2022), Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, *36th Conference on Neural Information Processing Systems*.

Zhang, B., Haddow, B., and Birch, A. (2023). Prompting Large Language Model for Machine Translation: A Case Study, *Proceedings of the 40 th International Conference on Machine Learning*.

Zhang, M., Press, O., Merrill, W., Liu, A., and Smith, N. (2023), How Language Model Hallucinations Can Snowball, ArXiv, abs/2305.13534.

Zhao, R., Li, X., Joty, S., Qin, C., and Bing, L. (2023), Verify-and-Edit: A Knowledge-Enhanced Chain-of-Thought Framework. ArXiv, abs/2305.03268.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.-Y., and Wen, J.-R. (2023), A Survey of Large Language Models. ArXiv, abs/303.18223.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., Zhang, H., Gonzalez, J., and Stoica, I. (2023), Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. ArXiv, abs/2306.05685.

Zheng, M., Pei, J., and Jurgens, D. (2023). Is "A Helpful Assistant" the Best Role for Large Language Models? A Systematic Evaluation of Social Roles in System Prompts. ArXiv, abs/2311.10054.

Zhou, C., Neubig, G., Gu, J., Diab, M., Guzmán, F., Zettlemoyer, L., and Ghazvininejad, M. (2021), Detecting Hallucinated Content in Conditional Neural Sequence Generation, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 1393-1404.

Zhou, Y., Muresanu, A., Han, Z., Paster, K., Pitis, S., Chan, H., and Ba, J. (2022), Large Language Models Are Human-Level Prompt Engineers, *11th Conference on International Conference on Learning Representations*.

저자소개

한승헌 : 고려대학교 통계학과에서 2023년 학사학위를 취득하였다. 현재는 고려대학교 산업경영공학과 석사과정에 재학 중이다. 연구 분야는 시계열 이상치 탐지 및 시계열 분석이다.

강민혁 : 고려대학교 산업경영공학과 학사과정에 재학 중이다. 연구 분야는 자연어 처리 및 Large Language Model이다.

강필성 : 서울대학교 산업공학과에서 2003년 학사, 2010년 박사 학위를 취득하였다. 이후 현대카드 과장, 서울과학기술대학교 조교수로 근무하였으며, 현재는 고려대학교 산업경영공학부 교수로 재직 중이다. 연구 분야는 정형 및 비정형 데이터를 활용한 데이터마이닝 및 기계학습 알고리즘 개발 및 제조/IT/공공분야 응용이다.

황우현 : 고려대학교 통계학과 학사과정에 재학 중이다. 연구 분야는 경영 전략 및 비즈니스 상황에서의 인공 지능의 응용이다.

<부 록>

A. Case Study

본 섹션에선 전반적으로 모든 지표에서 성능이 우수했던 FT-LLM-4을 활용해 생성한 결과물에 대해 분석한다.

<Table 12>는 retrieval이 제공한 연관 문서를 기반으로 언어 모델이 적절한 답변 생성 전략을 생성한 예시를 보여준다. 해당 전략은 질의와 관련이 높은 본문의 일부분을 잘 조명하며, 향후 답변 생성에 도움이 될 수 있는 가이드 라인을 잘 제시하였다. 이를 통해, 답변 생성 전략을 참고하여 생성된 답변은 사용자에게 관련된 정보를 정확하게 전달하는 점 역시 확인할 수 있다.

Table 12. Example of an Answer Generation Strategy

Question	Can soccer players use slogans with political, religious, or personal content?
Reference	**Title** : Kane is furious → 'clashed' with WC organizers right before the first kick-off...breaking Qatari law **Contents** : England captain Harry Kane now in Qatar challenges Qatari authorities and FIFA head-on. In Qatar, it is illegal to take action on anything remotely related to LGBTQ+. However, Kane is refusing to comply and will wear a rainbow "One Love" armband to show his support. England will kick off the World Cup against Iran at 10 p.m. on June 21. Kane has pledged to wear the rainbow armband for the match. ... omitted... FIFA Regulation 4.5, as formulated by the governing body of international football, states that equipment used by players must be free from political, religious or personal slogans, statements or images. In addition, players must not wear underwear that displays political, religious or personal slogans, statements or images, or advertising other than the manufacturer's logo.
Generated Strategy	Step 1: Find the sentence, "FIFA Regulation 4.5, as formulated by the International Football Association's Council, states that 'equipment used by players must be free of political, religious, or personal slogans, statements, or images'" to confirm that soccer players are not allowed to wear uniforms with slogans of a political, religious, or personal nature. Step 2: Find the sentence, "In addition, players must not wear underwear that displays political, religious, or personal slogans, statements, or images, or advertising other than the manufacturer's logo," to confirm that players are also not allowed to wear underwear that displays political, religious, or personal slogans, statements, or images, or advertising other than the manufacturer's logo. Step 3: Find the sentence "Violation of this will result in the player or team being sanctioned by the competition organizer, national football association, or FIFA" to confirm that violation of these regulations will result in the player or team being sanctioned.
Response	According to FIFA Regulation 4.5, as formulated by the International Football Association's Council, equipment used by players must be free of political, religious, or personal slogans, statements, or images. In addition, players must not wear underwear that displays political, religious, or personal slogans, statements, or images, or advertising other than the manufacturer's logo. Violations of these rules will subject the player or team to sanctions by the competition organizer, national football association, or FIFA.

B. Additional Experiments

Passage Retrieval top-k sensitivity analysis

가장 좋은 성능을 기록한 BM25 retrieval 기반 FT-LLM-3.5, 4를 활용하여 top-k 세팅에 따른 성능 변화를 실험하였고, 이를 <Table 13>에 기록하였다. k=3일 때 모든 정량적인 지표상 성능이 가장 우수하였음을 확인할 수 있다. 다만, k의 크기가 커질수록 FT-LLM의 답변 소요 시간과 컴퓨팅 비용이 증가하기 때문에 사용자가 목적에 맞게 크기를 조절하는 것이 적합하다.

Table 13. Sensitivity test on BM25 based FT-LLM-4 with Different top-k Hyperparameter Settings for Retrieval Module

k	Objective Questions			Subjective Questions
	F1	Acc_Human	Acc_GPT	Eval_Human
1	0.0461	0.6304	0.5870	4.1042
2	0.0498	0.6957	0.6957	4.1980
3	**0.0670**	**0.8478**	**0.6957**	**4.6886**