RAG 기반 LLM을 활용한 3GPP 표준 질의응답 시스템 구현

최희주, 채승호*

한국공학대학교

{babybird, shchae}@tukorea.ac.kr

Implementation of a RAG-Based LLM Question-Answering System for 3GPP Standards

Heeju Choi, Seong Ho Chae*

Tech University of Korea

요 약

3GPP 표준 문서는 방대한 분량과 절·표 중심의 서술로 인해 필요한 수치·조건의 근거를 빠르고 정확하게 찾기 어렵다. 이를 용이하게 하기 위해, 본 논문에서는 RAG(Retrieval-Augmented Generation) 기반 LLM(Large Language Model)을 활용한 3GPP 표준 질의응답 시스템을 구현하였다. 구체적으로, 오프라인에서는 텍스트 추출·청킹·메타데이터와 함께 규범 태깅과 표 태깅을 부여해 인덱싱하고, 온라인에서는 임베딩 검색과 크로스 인코더 재랭킹 후 질의 의도 기반 가중을 적용하여 컨텍스트를 구성하고 문서ID·릴리즈·페이지·절 인용과 함께 한국어 답변을 생성한다. 동일한 질의들에 대해 범용 LLM과 비교한 결과, 제안 시스템이 더 낮은 응답 시간을 가짐을 보임으로써 우수성을 입증하였다.

Ⅰ. 서 론

3GPP 문서는 전 세계 이동통신의 요구사항과 평가 방법을 규정하는 공식 규격서로, 릴리즈(release)별로 수백 페이지에 달하는 긴 분량과 절·하위 절 중심의 서술, 규범 어휘, 그리고 파라미터·임계값이 표 형태로 집약되어 있다는 구조적 특징을 가진다. 연구자는 특정 수치 및 조건과 근거 절을 빈번히 찾아야 하지만, 문서의 분량이 방대해 필요한 정보를 빠르고 정확하게 찾기 어렵다는 한계가 있다.

최근 LLM(Large Language Model)은 트랜스포머 기반의 대규모 사전 학습을 통해 요약, 코드 보조, 질의응답, 추론 등 다양한 작업에서 뛰어난 성능을 보이고 있다. 이와 함께 RAG(Retrieval-Augmented Generation)가 주목받고 있는데, 이는 LLM이 외부 지식원에서 관련 텍스트를 검색해 컨텍스트로 주입받은 뒤 답변을 생성하는 접근법이다. 3GPP 문서의 참고를 위해 범용 LLM을 단독으로 사용할 경우, 긴 표준 문서에 대해서는 응답 지연 측면이 길다는 한계를 가진다. 최근, 금융·의료 등 여러 분야에서 LLM과 RAG를 활용한 연구가 활발히 이루어지고 있으나, 통신 표준을 체계적으로 다루는 방법은 아직 충분히 정립되지 않았다[1]-[3].

따라서, 본 논문에서는 3GPP 표준에 특화된 RAG 기반 LLM 질의응답 시스템을 제안한다. 핵심은 규범 어휘가 포함된 청크에 가중치를 부여해 요구사항형 질의의 정밀도를 높이고, 수치·비교 질의에 대해 표 우선 검색을 적용해 근거 회수력을 강화하며, 문서ID·릴리즈·페이지·절 단위의 근거를 함께 제시해 신뢰성을 확보하는 것이다.

Ⅱ. 시스템 개요 및 구현

본 장에서는 제안 시스템의 전체 구조와 구현을 설명한다. 제안 시스템은 사전 구축형 오프라인 인덱싱 단계와 실시간 응답형 온라인 질의 단계로 이루어진 두 단계 파이프라인이다.

A. 오프라인 인덱싱 단계

오프라인 인덱싱 단계는 표준 PDF를 기계가 읽을 수 있는 검색 단위로 변환해 재사용 가능한 인덱스를 구축하는 과정이다. 우선, 질의응답 시스템이 근거로 삼는 3GPP 표준 PDF에서 텍스트 레이어를 추출하고, 문서ID·릴리즈·페이지·절과 같은 위치 정보를 함께 보존한다. 이는 이후 답변에서 출처 인용을 가능하게 하기 위함이다. 긴 본문은 토큰 기준으로 분할해 청크를 만들고, 경계에서 문맥이 끊기는 것을 줄이기 위해 중첩을 둔다. 본 연구에서는 청크 800 토큰, 중첩 120 토큰으로 설정하였다. 각 청크에는 [문서 식별자(doc_id), 릴리즈(release), 페이지(page), 절(clause)] 등의 메타데이터를 부여해 레코드 형태로 저장한다.

제안 시스템의 핵심은 표준 문서에 특화된 동작을 위한 도메인 태깅이다. 규범 태깅은 shall/should/requirement/mandatory/define 등 규범 어휘 빈도를 계산해 normative_score로 저장한다. 청크 분할기는 최대 800 토큰을 목표로 하되 문장 경계를 보존하므로 실제 청크 길이는 수 십 토큰 변동이 있다. 긴 청크가 유리해지지 않도록, 규범 어휘의 등장 빈도를 청크 길이로 나누어 '밀집도' 기반 점수로 사용하였다. 이 점수는 요구사항형 질의에서 근거 절을 위로 끌어올리는 신호로 쓰인다. 표 태깅은 'Table n' 캡션, '|' 패턴, 숫자 밀도 등 구조·패턴 단서를 이용해 표 청크를 식별하고 type="table"로 표시한다. 이는 수치·비교 질의에서 정답 표의 회수력을 높이는 데 직접 기여한다. 이렇게 정의된 태그 신호는 온라인 단계의 도메인 부스팅에 그대로 연동되어 답변의 정확도와 근거 회수율을 높인다.

각 청크 텍스트는 임베딩 모델을 통해 고차원 실수 벡터로 변환된다. 의미가 가까운 텍스트일수록 벡터 공간에서 서로 가깝게 배치되므로, 이후 코사인 유사도 기반 검색이 가능해진다. 마지막으로 임베딩 벡터와 원문 텍스트·메타데이터를 벡터 데이터베이스에 영구 저장하여, 온라인 질의 단계에서 저지연 유사도 검색이 이루어지도록 한다.

B. 온라인 질의 단계

온라인 질의 단계는 사용자의 질문에 대해 인덱스에서 관련 근거를 신속히 회수하고, LLM으로 근거 기반 답변을 생성하는 과정이다. 사용자가 질문을 입력하면, 질의 문자열을 인덱싱 단계와 동일한 임베딩 모델로 벡터화하여 쿼리 임베딩을 생성한다. 본 구현에서는 임베딩 모델로 BAAI/bge-small-en-v1.5(384-dim)를 사용하였다. 이후, 쿼리 임베딩과 벡터 데이터베이스에 저장된 청크 임베딩 간 코사인 유사도를 계산해 similarity_top_k개 후보를 회수한다. 회수된 후보는 크로스 인코더 기반 문장쌍 재랭커로 정밀 재점수화하고 도메인 부스팅으로 보정하여 상위 rerank_top_n개만 남긴다. 본 구현의 기본 설정은 similarity_top_k=12,

rerank_top_n=5이며, 이 값은 사용자 인터페이스(UI)에서 조정이 가능하다.

제안 시스템의 핵심은 질의 의도에 따라 후보 점수를 도메인 부스팅으로 보정하는 설계이다. 도메인 부스팅은 재랭크 점수 위에 소폭의 가산을 더하는 방식으로 정의하였다. 부스팅은 재랭크 이후 상위 후보들 사이에서만 작동하도록 하여, 의미적 관련성이 낮은 청크가 과도하게 상승하는 것을 방지하였다. 규범형 질의에서는 청크의 normative_score에 비례해 가중치를 부여하여 요구사항 절을 우선시한다. 수치·비교형 질의에서는 메타데이터가 type="table"인 청크에 보너스를 부여해 표 기반 수치 정보의 회수력을 높인다.

최종 선별된 청크는 프롬프트 구성기에서 중복 제거·길이 제한·정렬을 거쳐 하나의 컨텍스트로 합쳐진다. LLM은 질문과 이 컨텍스트를 입력으로 받아 답변을 생성하며, 컨텍스트가 부족한 경우에는 “문서에서 근거를 찾지 못했습니다.”라고 응답하도록 지침을 두어 잘못된 정보를 제공하지 않도록 한다. 그림 1은 오픈소스 Streamlit을 활용하여 구현한 사용자 인터페이스이다. ‘질의하기’를 클릭하면 검색 → 재랭크 → 도메인 부스팅 → 생성이 자동으로 수행된다. 결과 화면에는 최종 답변, 문서ID·릴리즈·페이지·절 단위의 출처 인용, 근거 청크 미리보기가 표시되도록 구성하였다.

그림 1. Streamlit 기반 사용자 인터페이스

Ⅲ. 시스템 성능 검증

본 장에서는 제안 시스템과 범용 LLM(ChatGPT)에 동일한 질문 4개를 제시하여, 답변과 문서ID·릴리즈·페이지·절 인용의 정확성을 확인하고 응답 완료 시간을 비교한다. 성능 평가는 3GPP TR 36.885 (Release 14)와 3GPP TR 38.885 (Release 16)을 대상으로 수행하였다[4],[5]. 질문은 3GPP TR 36.885 관련 2개, 3GPP TR 38.885 관련 2개로 구성했으며, 목록은 다음과 같다.

(1) 3GPP 표준의 고속도로 시나리오(freeway case)에서 한 도로의 폭은 몇 m로 설정되어 있는가?

(2) 3GPP 표준의 도시 시나리오(urban case)에서 macro eNB 기지국 간 거리는 어떻게 설정되어 있는가?

(3) 3GPP 표준에서 도시 격자(urban grid) 환경에서 거리가 150 m일 때, 유니캐스트 주기 트래픽에 대한 SL 자원 할동 모드 1(Resource Allocation mode 1; RA Mode 1)의 성능은 어떠한가?

(4) 3GPP 표준에서 RA Mode 1 시뮬레이션 시 대역폭은 설정값은 무엇인가?

평가에서의 응답 시간은 UI에서 측정한 질의 버튼 클릭부터 답변 표시까지의 총 소요 시간으로 정의하였다. 표 1은 두 시스템에서 위 4개 질문에 대해 측정한 응답 시간을 비교한 결과이다. 두 시스템 모두 질문에 올바른 답변과 문서ID·릴리즈·페이지·절 인용을 제공했지만, 측정 결과 제안 시스템이 범용 LLM 대비 더 낮은 지연을 보여 응답 처리 속도에서 우위를 확인하였다.

	Response time of the proposed system	Response time of a general-purpose LLM
Question 1	3.235 s	42 s
Question 2	3.419 s	55 s
Question 3	4.905 s	1 m 45 s
Question 4	4.862 s	1 m 28 s

표 1. 응답 시간 비교표

Ⅳ. 결 론

본 논문에서는 3GPP 표준 문서를 대상으로 한 RAG 기반 LLM을 활용한 질의응답 시스템을 구현하였다. 표준 문서의 서술 특성을 반영한 도메인 태깅과 도메인 부스팅을 결합하여 요구사항형·수치형 질의의 신뢰도를 높이는 방법을 제안하였고, Streamlit 기반 UI를 제공하며 실사용 편의성을 확보하였다. 동일한 4개 질의에 대해 범용 LLM과 비교한 실험에서, 제안 시스템은 보다 낮은 응답 지연을 보이면서도 근거 인용을 동반한 응답을 안정적으로 제공함을 확인하였다. 또한 동일한 청킹·태깅·임베딩·인덱싱 절차를 유지함으로써 다수의 표준 문서로 손쉽게 확장이 가능하다.

ACKNOWLEDGMENT

이 논문은 2022년도 정부(방위사업청)의 재원으로 국방기술진흥연구소의 지원을 받아 수행된 연구임 (KRIT-CT-22-047 우주계층 지능통신망 특화연구실)

참 고 문 헌

[1] N. Kim, “A study on a korean medical question-answering system using LLM-based retrieval-augmented generation,” in *Proc. The e-Business Studies*, 2025.

[2] Y. Kim, J. Lee, “Design of a real-time stock information retrieval and question-answering system based on LLM and RAG,” in *Proc. Korean Institute of Information Scientists and Engineers*, 2024.

[3] Y. Yun, H. Ahn, “Development of a korean news question answering system based on LLM using RAPTOR,” in *Journal of Intelligence and Information Systems*, 2024.

[4] *Technical specification group radio access network; study LTE-based V2X services; (Release 14)*, document 3GPP TR 36.885 V14.0.0, 3rd Generation Partnership Project, Jun. 2016.

[5] *Technical specification group radio access network; study on NR vehicle-to-everything (V2X); (Release 16)*, document 3GPP TR 38.885 V16.0.0, 3rd Generation Partnership Project, Mar. 2019.