JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 29, No. 5: 626~634, May. 2025

검색 증강 생성을 이용한 LangChain 프레임워크 기반 보험 FAQ 고객 상담 챗봇 설계 및 구현

이태우[1] · 김도연[2] · 심재정[2] · 한혜원[2] · 김태용[3*]

Design and Implementation of an Insurance FAQ Customer Service Chatbot Using a LangChain-Based Retrieval-Augmented Generation Framework

Taewoo Lee[1] · Doyeon Kim[2] · Jaejeong Shim[2] · Hyewon Han[2] · Taeyong Kim[3*]

[1]Director, Placa, Inc., Seoul, 07327 Korea
[2]Researcher, Placa, Inc., Seoul, 07327 Korea
[3]*Ph.D. Student, Department of Electrical and Electronic Engineering, Yonsei University, Seoul, 03722 Korea

요 약

본 연구는 LangChain을 활용하여 검색 증강 생성(RAG) 기반 보험 고객 상담을 위한 도메인 특화 챗봇 시스템의 설계 및 구현 방법을 제시한다. GPT-4o와 같은 대규모 언어 모델(LLM)은 일반적인 자연어처리에서는 강력한 성능을 보이나, 보험과 같은 특수 분야에서는 전문 용어 이해 부족과 새로운 지식을 반영하는 데에 어려움을 겪는다. 이를 해결하기 위해, OpenAI 임베딩, 하이브리드 검색기(BM25 + FAISS), 구조화된 프롬프트 엔지니어링을 통합한 RAG 기반 실시간 FAQ 챗봇을 개발하였다. 6개 보험 카테고리에서 총 1,456개의 실제 QA 데이터로 평가한 결과, RAG 미적용 모델 대비 BLEU와 METEOR 지표에서 20% 이상의 성능 향상을 달성하였다. 특히 자동차보험과 운전자보험 영역에서 높은 정확도를 보였으며, 암보험과 건강보험은 데이터 부족과 복잡성으로 성능이 다소 낮았다. 또한 실무 적용을 위한 Streamlit 기반 웹 애플리케이션을 구현하였다. 본 연구는 높은 신뢰성을 요구하는 도메인에서 RAG 시스템의 효과를 입증하고, 향후 도메인 적응 및 프라이버시 강화 설계의 필요성을 제시한다.

ABSTRACT

This study presents the design and implementation of a domain-specific chatbot system for insurance customer service using a Retrieval-Augmented Generation (RAG) framework integrated with LangChain. While large language models (LLMs) exhibit strong general NLP performance, they struggle with specialized domains such as insurance due to limited understanding of terminology and outdated knowledge. To overcome these challenges, we developed a real-time FAQ chatbot that combines OpenAIEmbeddings, hybrid search (BM25 + FAISS), and structured prompt engineering. The system was evaluated on 1,456 real insurance QA pairs across six categories, achieving over 20% performance improvements in BLEU and METEOR scores compared to non-RAG models. Notably, automobile and driver insurance queries demonstrated superior accuracy, while cancer and health insurance showed lower performance due to data sparsity and complexity. A Streamlit-based web application was also implemented for practical deployment. This research highlights the effectiveness of RAG-based systems in high-reliability domains and identifies future directions such as domain adaptation and privacy-preserving designs for enhanced chatbot performance.

키워드 : 검색 증강 생성, 랭체인, 대규모 언어 모델, 보험 챗봇, 고객 상담 서비스

Keywords : Retrieval-Augmented Generation, LangChain, Large Language Model, Insurance Chatbot, Customer Service

Received 26 February 2025, **Revised** 12 March 2025, **Accepted** 30 April 2025

*** Corresponding Author** Taeyong Kim (E-mail:taeyongkim@yonsei.ac.kr)
Ph.D. Student, Department of Electrical and Electronic Engineering, Yonsei University, Seoul, 03722 Korea

Open Access http://doi.org/10.6109/jkiice.2025.29.5.626 pISSN:2234-4772

Ⅰ. 서 론

대규모 언어 모델(Large Language Models, LLM)은 최근 자연어 처리(NLP) 및 챗봇 개발 분야에서 괄목할 만한 진보를 이루며, 텍스트 생성, 질의응답, 감정 분석 등 다양한 과업에서 전통적인 규칙 기반 또는 통계 기반 방법론을 대체할 수 있는 기술적 가능성을 입증하고 있다[1]. 이러한 LLM은 일반적인 언어 표현 처리에 뛰어난 성능을 보이며 여러 산업 분야로의 확장을 이끌어냈지만, 특정 도메인에 특화된 환경에서는 여전히 뚜렷한 한계가 존재한다. 예컨대, LLM은 보험이나 의료와 같은 도메인에서 사용되는 전문 용어나 맥락을 정확히 이해하지 못하고, 최신 정보를 반영하지 못한 채 과거 학습 데이터에 기반한 환각(hallucination) 현상이 발생하는 경우가 많다[2]. 또한, 민감하거나 부정확한 정보를 무분별하게 생성할 가능성이 있어, 특히 개인정보 보호가 중요한 보험 도메인에서는 LLM 기반 챗봇의 실무 도입이 제한적이다.

이러한 문제를 해결하고자 등장한 기술은 검색 증강 생성(Retrieval-Augmented Generation, RAG)이다. RAG는 외부 지식 소스에서 관련 문서를 검색한 뒤, 이를 기반으로 LLM이 응답을 생성하도록 유도함으로써, 보다 정확하고 신뢰할 수 있는 응답을 생성할 수 있도록 한다[3]. 이는 LLM의 내부 지식 한계를 보완하고, 최신 정보 반영과 도메인 적합성을 확보하는 데 효과적이다. 특히, RAG는 정보 검색(retrieval)과 생성(generation)의 장점을 결합하여, 보험, 법률, 의료와 같은 고신뢰성 기반 산업 분야에서 실질적으로 활용될 수 있는 기술로 주목받고 있다[4-6].

RAG 기반 챗봇 시스템의 구현을 위한 핵심 구성 요소로는 LangChain 프레임워크가 있다. LangChain은 대형 언어 모델과 외부 정보원의 통합을 체계화하고, 검색, 프롬프트 설계, 단계적 체인 구성 등 다양한 모듈을 통해 복잡한 언어 기반 작업의 효율성과 확장성을 높이는 오픈소스 프레임워크이다[7-8]. 본 연구는 이 LangChain과 RAG 기술을 접목하여, 민감한 개인정보를 포함하는 보험 도메인에 특화된 FAQ 챗봇 시스템을 설계하고 구현한다.

기존 연구들은 RAG를 의료, 법률 등 일부 분야에 적용하여 일정 수준의 성능 향상을 보고한 바 있으나[4-6], 실시간 상호작용, 보험 도메인의 규제 특성, 사용자 친화적 구현 측면에서 구체적인 실현 사례는 부족한 실정이다. 특히, 보험 상담 데이터는 개인정보 보호와 의학 전문성 등의 특수한 제약이 존재하며, 문서 구조와 표현 방식이 도메인에 따라 이질적이라는 점에서 기술 적용상의 도전 과제가 존재한다. 이에 따라 본 연구는 단순히 RAG 모델을 적용하는 것을 넘어서, 실제 환경에 적합한 데이터 구성, 성능 편차 분석, 사용자 경험 중심의 실시간 시스템 구현까지 포괄적으로 다룬다.

기존의 RAG 기반 챗봇 연구들은 주로 의료, 법률, 일반 지식 응답 등 특정 고정 문서에 기반한 질의응답 성능 개선에 초점을 맞춰왔다. 하지만 보험 상담 도메인에서는 개인정보 보호, 전문용어 처리, 상담 카테고리 다양성 등 복합적인 요인으로 인해 단순 RAG 구조만으로는 실질적인 서비스 구현이 어렵다는 한계가 존재한다.

본 연구는 이러한 현실적인 제약을 고려하여, 보험 상담에 최적화된 문서 구성과 쿼리 기반 문서 검색, 그리고 LangChain 기반의 처리 파이프라인을 통합적으로 구성함으로써 기존 연구들과 차별화된 접근을 시도한다. 특히 자동차보험, 운전자보험 등 6개 세부 카테고리를 고려한 설계와 앙상블 검색기(BM25 + FAISS)를 활용한 고정밀 검색 전략, 그리고 사용자 친화적 웹 응용 구현은 기존 단순 기술 검증 수준의 RAG 챗봇과 구별되는 핵심적인 실용적 차별점이다. 또한, 본 연구는 카테고리별 성능 분석을 통해 실제 응답 품질의 세부 편차를 도출함으로써, 도메인 적용 가능성에 대한 정량적·정성적 평가를 동시에 수행한다.

Ⅱ. 관련 연구

2.1 검색 증강 생성(RAG)

RAG는 대규모 언어 모델(Large Language Model, LLM)의 고질적인 문제로 지적되는 환각(hallucination) 현상, 즉 근거 없는 응답 생성 문제를 해결하기 위해 제안된 프레임워크이다[3]. RAG는 입력 문장에 대해 LLM이 직접 응답을 생성하는 방식이 아니라, 외부 문서 저장소에서 관련 정보를 검색하고 해당 정보를 바탕으로 응답을 생성함으로써, 생성 내용의 정확성과 최신성을 확보한다. 특히 RAG는 도메인 특화 지식이 요구되는 분야에서 활용도가 높으며, 복잡한 질의에 대해 정보 기반의 정확한 응답을 제공할 수 있다는 점에서 최근 많은 연구들이 의료, 법률, 금융, 비즈니스 등 실전

응용 분야에 집중되고 있다[4-6].

손지웅 외 7인은 의료 질의응답에서 RAG의 적용을 통해 임상적 추론 기반 검색 결과를 활용하여 응답의 신뢰도를 향상시켰고[4], Wiratunga 외 8인은 법률 도메인에서 사례 기반 추론과 RAG를 결합하여 정답률을 크게 높이는 구조를 설계하였다[5]. 이광우 외 1인의 연구는 국내 기업 문서 응답 시스템에 RAG를 적용하여 정확도뿐 아니라 사용자 만족도도 향상시킨 사례를 보고하였다[6]. 그러나 이들 대부분은 보험 도메인에 적용한 사례는 없으며, 도메인의 특수성과 문서 구조, 표현의 비표준성 등으로 인해 RAG의 적용이 쉽지 않은 것으로 알려져 있다. 본 연구는 이러한 공백을 메우고자 보험 특화 FAQ 데이터셋을 활용한 RAG 기반 시스템을 설계하여 해당 분야에서 실질적인 적용 가능성을 입증하였다.

2.2 LangChain 프레임워크

LangChain은 LLM과 외부 구성요소(문서 저장소, 검색기, 프롬프트 템플릿 등)를 연결하여 다단계 작업 흐름을 체계적으로 구성할 수 있는 오픈소스 프레임워크이다[7-8]. 단일 질의-응답 방식의 한계를 넘어, 질의 임베딩, 문서 검색, 프롬프트 구성, 답변 생성 등 전체 파이프라인을 명시적으로 구성하고 디버깅할 수 있는 구조를 제공한다.

기존에는 단순한 API 기반 질의응답 구조가 대부분이었으나, LangChain은 체인(chain) 단위의 처리 흐름 설계와 다양한 검색기 및 응답기 조합, 프롬프트 템플릿화 등의 기능을 통해, 챗봇의 도메인 적합성과 확장성을 크게 높일 수 있다는 장점이 있다. 특히 LangChain은 RAG 모델의 구현을 위한 핵심 구성요소들(OpenAIEmbeddings, FAISS, PromptTemplate 등)을 기본적으로 지원하므로, 도메인 지식 응답 시스템 개발에 매우 적합하다.

정승욱 외 2인은 LangChain을 활용하여 악성코드 탐지 시스템을 구성하고 환각 문제를 완화한 바 있으며[11], 서재영 외 3인은 역사 유적지 정보를 제공하는 챗봇에 LangChain을 적용하여 문맥 기반의 정확한 응답 생성을 가능하게 하였다[12]. 본 연구는 이러한 장점을 활용하여 보험 상담 분야에 특화된 LangChain 기반 응답 파이프라인을 구현하였다. 특히, 문서 청크 처리, 벡터 임베딩, 앙상블 검색기 조합, 프롬프트 설계까지 일관된 처리 흐름을 구성하였으며, 이는 기존 LangChain 응용 연구들보다 실무 적용 수준에서 더 높은 완성도를 지닌다.

2.3 RAG 기반 챗봇 시스템 및 실제 응용 사례

최근 RAG 기반 챗봇 시스템은 정보 정확성과 실시간 응답성 측면에서 기존 LLM 단독 시스템보다 탁월한 성능을 보이며, 실제 산업현장에서의 응용 사례가 점차 증가하고 있다. 의료 상담, 법률 자문, 기업 문서 응답 등에서 도메인 지식 기반 검색과 자연어 생성의 결합은 응답 품질을 크게 개선하는 핵심 요소로 작용한다[4-6].

기존 연구들에서는 대부분 단일 검색기(BM25 또는 FAISS 단독 사용) 기반의 문서 검색을 사용하고 있으며, 사용자 인터페이스나 실제 상담 시나리오에 대한 고려가 부족한 경우가 많다. 이에 반해 본 연구는 희소 기반(BM25)과 밀집 기반(FAISS) 검색기를 결합한 앙상블 검색기 구조, 보험 카테고리 분류 기반 문서 매핑, 웹 인터페이스를 통한 실시간 상호작용 등 현실적인 사용 환경을 고려한 설계를 통해 차별성을 확보하였다.

또한, 본 연구는 기존 문헌과 달리 단순 전체 평균 성능(BLEU, METEOR)에 그치지 않고, 보험 유형별 성능 편차를 세부적으로 정량 분석함으로써 특정 도메인 내의 응답 한계 및 개선 필요 영역을 도출하였다는 점에서도 차별화된다. 이는 향후 카테고리별 도메인 적응(domain adaptation) 또는 동적 프롬프트 튜닝(dynamic prompt tuning)과 같은 후속 연구의 기반을 제공할 수 있다.

Ⅲ. RAG 기반 LangChain을 이용한 보험 FAQ 고객 상담 서비스 챗봇 시스템

[그림 1]은 LangChain 프레임워크를 활용하여 RAG 기술을 적용한 보험 FAQ 고객 상담 챗봇의 구조도이다. 먼저, (Step 1) 사용자로부터 입력된 질의는 임베딩 모델을 통해 쿼리 임베딩으로 변환된다. 문서 처리 단계에서는, 보험 상담 관련 FAQ 문서들이 문서 분할기를 통해 의미 단위로 분할된다. (Step 2) 각 분할된 문서 청크는 동일한 임베딩 모델을 통해 청크 임베딩으로 변환되어 벡터 데이터베이스에 저장된다. (Step 3) 검색 단계에서는, 쿼리 임베딩과 벡터 데이터베이스 내의 청크 임베딩 간 유사도를 계산하여 가장 관련성이 높은 문서 청크를 검색한다. 검색된 문서 청크들은 문맥 정보로 통합되어 사용

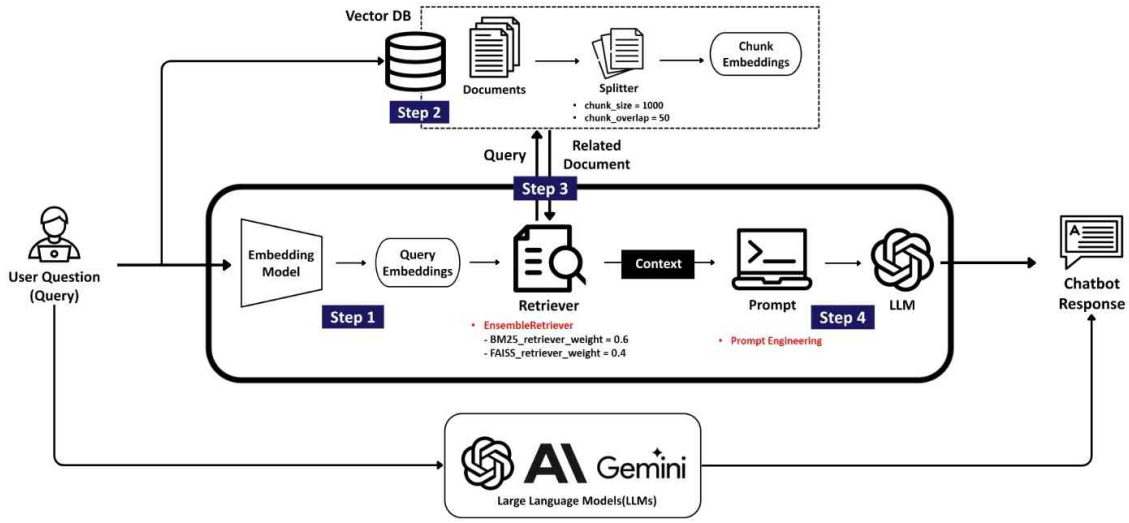

Fig. 1 End-to-end workflow of the proposed RAG-based insurance chatbot system. User queries are embedded and matched against a vector database of domain documents. Retrieved documents are combined with the query to form a structured prompt, which is then passed to the LLM to generate accurate and context-aware responses.

자의 질의와 함께 프롬프트 템플릿에 따라 구조화된다. 최종적으로, (Step 4) 구조화된 프롬프트는 LLM에 입력되어 자연스러운 응답을 생성한다.

3.1 사용자 질의 입력 및 임베딩 변환 단계

(Step 1) 사용자의 입력 질의는 임베딩 모델을 통해 수치화된 벡터로 변환된다. 이때 LangChain 프레임워크에서 제공하는 *OpenAIEmbeddings* 클래스를 사용하여 질문 문장의 의미론적 특성을 효과적으로 포착할 수 있도록 하였다. *OpenAIEmbeddings* 클래스는 OpenAI사의 *text-embedding-ada-002* 임베딩 모델을 사용하여 토큰화된 문서들의 임베딩 벡터를 구한다. 생성된 쿼리 임베딩은 벡터 데이터베이스에 저장된 문서들과의 유사도 계산을 위한 기준점으로 활용되며, 이러한 임베딩 기반의 검색 방식은 의미적 유사도를 고려하여 관련 문서 검색의 정확도를 높이는 데 핵심적인 역할을 한다.

3.2 벡터 데이터베이스 단계

(Step 2) 벡터 데이터베이스는 보험 FAQ 문서들의 벡터 표현을 저장하고 검색하기 위한 저장소로, 본 연구에서는 Facebook에서 개발한 FAISS(Facebook AI Similarity Search)를 구현하였다[13]. FAISS는 고차원 벡터의 효율적인 유사도 검색과 최적화를 위한 라이브러리로, 대규모 문서 컬렉션에서 빠른 검색이 가능하다는 장점이 있다.

보험 FAQ 문서 기반 벡터 데이터베이스 구축 과정은 다음과 같다. 첫째, 카테고리별 보험 상담 관련 질의-응답 쌍 형태의 문서들을 수집하였다. 둘째, LangChain 프레임워크의 *RecursiveCharacterTextSplitter* 도구를 활용하여 최대 1,000자를 기준으로 문서를 청크 단위로 분할하였다. 이는 검색 정확도와 처리 효율성의 균형을 위해 설정된 값이다. 셋째, 각 청크는 OpenAI의 *text-embedding-ada-002* 임베딩 모델을 통해 1,536차원의 밀집 벡터로 변환되었다. 이때 문서 임베딩과 쿼리 임베딩의 일관성을 위해 동일한 임베딩 모델을 사용하였다. 마지막으로, 생성된 청크 임베딩 벡터와 원본 문서 참조 정보를 함께 FAISS 인덱스에 저장하였다. 이러한 벡터화된 문서 저장 방식은 키워드 매칭이 아닌 의미적 유사도를 기반으로 하여 보험 상담 질의에 대해 보다 정확한 FAQ 검색 결과를 제공할 수 있다.

3.3 검색 단계

(Step 3) 검색기(retriever)는 사용자 질문의 쿼리 임베딩과 벡터 데이터베이스에 저장된 문서 벡터들 간의 유사도를 계산하여 가장 관련성이 높은 하나의 문서를 검색한다. 본 연구에서는 여러 검색기들을 결합하여 검색

성능을 극대화하기 위해 앙상블 검색기를 사용하였다. 본 연구의 앙상블 검색 프로세스는 다음과 같이 동작한다. 첫째, 사용자의 질문은 두 가지 검색 방식으로 병렬 처리된다. 하나는 키워드 기반의 희소 검색기 BM25(Best Match 25) 이며[14], 다른 하나는 의미적 유사성 기반의 밀집 검색기 FAISS이다. BM25는 단어 빈도와 역문서 빈도를 기반으로 문서와 질문 간의 관련성을 계산하며, FAISS는 임베딩 벡터 간의 코사인 유사도를 활용한다. 둘째, 각 검색기는 독립적으로 관련성 점수를 산출한 후, 각 검색 방식에 가중치를 적용하는데, BM25에는 60%, FAISS에는 40%를 부여한다. 이 가중치 설정은 보험 도메인의 특성을 반영하여, 전문 용어의 정확한 매칭과 의미적 유사성을 균형 있게 반영하기 위함이다. 셋째, 가중 평균된 점수를 기준으로 문서를 정렬하고, 상위 k개 문서(k=5)를 최종 결과로 선택한다. 이 문서들은 생성 모델의 입력 컨텍스트로 사용되어 최종 응답을 생성하는 데 활용된다. 이러한 방식은 키워드 기반 검색의 정확성과 의미 기반 검색의 유연성을 동시에 활용하여 단일 검색기 대비 더 포괄적이고 정밀한 검색 성능을 제공한다.

3.4 대규모 언어 모델 단계

(Step 4) 최종적으로 LLM은 사용자의 질의와 검색된 문맥(context)을 종합하여 응답을 생성한다. 본 연구에서는 OpenAI사의 GPT-4o 모델을 활용하였으며, 체계적인 프롬프트 엔지니어링을 통해 보험 상담에 최적화된 응답 생성 체계를 구축하였다.

본 연구의 프롬프트 설계는 다음과 같이 구성되었다. 우선 시스템 프롬프트 부분에서 “보험사의 전문 상담원”이라는 명확한 역할 정의로 시작하여 모델의 응답 정체성을 확립하였다. 이어서 응답 생성을 위한 구체적인 가이드라인을 ‘[규칙]’ 섹션으로 구조화하였다. 답변 길이 규칙에서는 "질문과 일치하는 핵심 정보만 포함하여 3~5문장으로 작성"하도록 명시하였고, 자료 활용 측면에서는 "제공된 자료를 바탕으로 작성하며, 질문에서 언급된 상황과 관련된 정확한 정보를 제공"하도록 지시하였다.

또한 문장 구조와 관련하여 "질문과 유사한 문장 구조를 사용하여 일치도를 높이고, 질문의 주요 키워드를 자연스럽게 답변에 포함"하도록 하였으며, 표현 방식에서는 "고객이 쉽게 이해할 수 있도록 간결하고 명확한 문장을 사용"하도록 구체적으로 지시하였다. 불확실한 상황에 대비해 "정보가 부족하거나 보험사 정책에 따라 달라질 경우, 고객이 추가 자료를 확인할 수 있도록 안내"하는 지침을 포함하였다. 이러한 정교한 프롬프트 엔지니어링을 통해 검색된 보험 FAQ 문서를 바탕으로 일관성 있고 사용자 친화적인 응답을 생성하도록 하였다.

Ⅳ. 실험 및 결과

본 장에서는 제안한 RAG 기반 챗봇 시스템의 성능을 정량적으로 평가하기 위한 실험 설계, 데이터 구성, 평가 지표, 실험 결과 및 그에 대한 분석을 상세히 기술한다. 특히 보험 도메인의 특성을 고려한 세부 카테고리 분석을 통해 시스템의 실제 적용 가능성과 한계점을 규명하고자 하였다.

4.1 데이터셋

실험에 사용된 데이터셋은 국내 보험 전문 상담 기업인 InsunetFC에서 수집한 실제 FAQ 데이터를 기반으로 구축되었다. 총 1,456개의 질의-응답 쌍으로 구성되어 있으며, 이는 다음의 여섯 개 보험 유형으로 분류된다: 자동차보험(574건), 운전자보험(153건), 건강보험(176건), 암보험(131건), 어린이보험(182건), 기타문의(240건). 각 질문에 대한 응답은 보험 분야 실무자가 직접 작성하였으며, 도메인 전문성이 반영된 고품질 정답 데이터를 기준 응답으로 사용하였다.

4.2 평가 방법

본 연구에서는 제안한 챗봇 시스템의 응답 품질을 평가하기 위해 BLEU(BiLingual Evaluation Understudy) [15]와 METEOR(Metric for Evaluation of Translation with Explicit ORdering) [16] 지표를 사용하였다.

BLEU는 n-gram 기반 정밀도(precision)를 통해 생성 응답과 기준 응답 간의 일치율을 측정한다. BLEU 점수는 문법적 정확성과 표현 일치도를 중심으로 응답 품질을 정량화한다. 반면 METEOR는 정밀도와 재현율(recall)의 조화 평균(F1-score)을 기반으로, 어휘 유사성, 어형 변화, 동의어 등을 고려해 의미적 자연스러움을 평가한다. 두 지표를 병행함으로써 챗봇 응답의 형식적 정확도와 의미적 유창성을 함께 정량 분석하였다. 성능 평가는 전체 테스트 데이터에 대해 모델이 생성한 응답과 기준

응답 간의 BLEU 및 METEOR 점수를 계산하고, 보험 유형별로 세부 점수를 도출하였다. 이처럼 세분화된 분석은 단순 평균 기반 평가가 놓치는 카테고리별 편차와 응답 취약점을 파악하는 데 목적이 있다.

Table. 1 Performance Comparison of Insurance FAQ Chatbot Systems with Different Model Configurations

RAG	LLM	Category	BLEU		METEOR	
O	**GPT-4o**	Car	**0.62**	**0.52**	**0.65**	**0.59**
		Driver	**0.61**		**0.70**	
		Health	**0.49**		**0.57**	
		Cancer	0.27		0.34	
		Children	**0.60**		**0.70**	
	Claude -3.5 -Sonnet	Car	0.48	0.41	0.47	0.49
		Driver	0.43		0.55	
		Health	0.38		0.45	
		Cancer	**0.33**		**0.44**	
		Children	0.42		0.55	
	Gemini -1.5 -Flash	Car	0.46	0.41	0.44	0.48
		Driver	0.45		0.55	
		Health	0.38		0.44	
		Cancer	**0.33**		**0.44**	
		Children	0.42		0.53	
X	GPT-4o	Car	0.30	0.29	0.26	0.32
		Driver	0.30		0.35	
		Health	0.28		0.31	
		Cancer	0.26		0.34	
		Children	0.30		0.36	
	Claude -3.5 -Sonnet	Car	0.33	0.28	0.28	0.34
		Driver	0.29		0.36	
		Health	0.28		0.34	
		Cancer	0.24		0.34	
		Children	0.27		0.36	
	Gemini -1.5 -Flash	Car	0.30	0.28	0.26	0.33
		Driver	0.29		0.36	
		Health	0.28		0.32	
		Cancer	0.25		0.35	
		Children	0.27		0.35	

4.3 실험 설정

본 연구에서는 LangChain 프레임워크를 활용하여 RAG를 적용한 보험 상담 챗봇을 구현하고, 이를 평가하였다. 이를 위해 Huggingface의 Transformers 라이브러리를 이용하여 질문 분류기(classifier)를 구현하고[17], 사용자가 직접 질문을 입력하고 답변을 확인할 수 있는 사용자 친화적인 Streamlit 기반의 웹 애플리케이션을 개발하였다[18].

질문 분류기는 사용자가 입력한 질문을 자동차보험, 운전자보험, 건강보험, 암보험, 어린이보험, 기타문의의 6개 카테고리 중 하나로 자동 분류하는 역할을 수행한다. 해당 분류기의 학습을 위해 Huggingface의 *tokenizer.encode_plus* 함수를 사용하여 입력 데이터를 전처리하였다. 주요 설정은 특별 토큰 추가, 최대 토큰 길이 128, 패딩 및 어텐션 마스크 적용 방식으로 구성되었다. 분류기 학습 시 사용한 하이퍼파라미터는 배치 크기 16, 학습률 2e-5, 총 3 에포크였으며, 최적화 알고리즘으로는 AdamW를 사용하였다. 분류기를 통과한 질문은 RAG 방식으로 처리되었다. 먼저, 질문이 특정 보험 카테고리로 분류되면, 해당 카테고리에 맞는 문서를 검색하여 관련 정보를 수집한다. 이후 LangChain의 생성 모델을 사용하여 검색된 문서를 기반으로 자연어 형태의 답변을 생성하는 방식으로 구성되었다.

4.4 실험 결과

제안한 RAG 기반 보험 FAQ 고객 상담 챗봇 시스템의 성능을 평가하기 위해, GPT-4o를 기반으로 한 RAG 적용 모델과 다른 비교군들의 성능을 분석하였다. [표 1]은 각 모델 구성과 보험 카테고리별 BLEU와 METEOR 점수를 보여준다. [그림 2]는 가장 우수한 LLM 세팅인 GPT-4o 환경에서 제안하는 방법과 기존 방법의 성능 결과를 시각적으로 나타낸 그래프이다.

[표 1]과 [그림 2]에서 확인할 수 있듯, 제안한 RAG-O(GPT-4o + RAG) 모델은 전체 평균 BLEU 0.52, METEOR 0.59로 가장 우수한 성능을 기록하였다. 특히 자동차보험(BLEU: 0.62, METEOR: 0.65)과 운전자보험(BLEU: 0.61, METEOR: 0.70) 영역에서 높은 정밀도와 자연스러움을 함께 달성하였다. 이는 해당 영역의 문서 구조가 비교적 표준화되어 있고, 반복적인 질의 패턴이 많아 검색기와 LLM 간의 연계가 효과적으로 작동했기 때문으로 분석된다.

반면, 암보험과 건강보험 영역에서는 상대적으로 낮은 점수를 기록하였다(BLEU 약 0.46, METEOR 약 0.52). 이는 해당 분야에서 문서 수가 적고, 전문 용어 사용이 복잡하며, 표현 방식이 비정형적인 경향이 많기 때문이다.

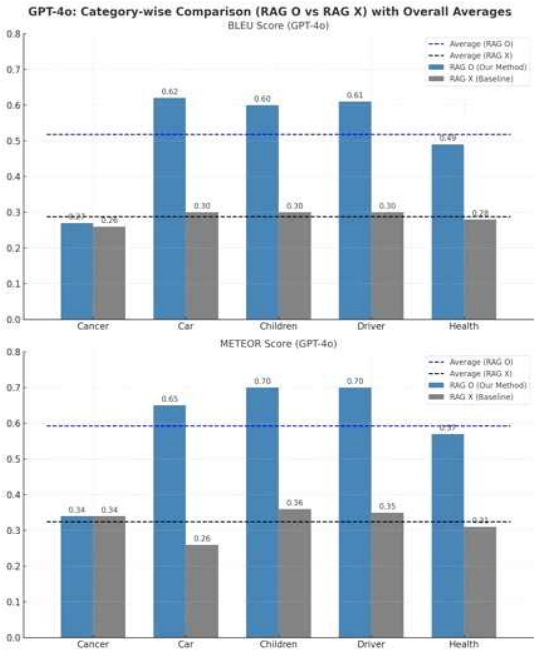

Fig. 2 Category-wise Comparison (RAG-O vs RAG-X) with GPT-4o

이러한 결과는 서론에서 제기한 바와 같이, 데이터의 편향성과 도메인 복잡도가 응답 품질에 큰 영향을 미친다는 점을 실험적으로 확인한 것이다.

또한, 동일한 RAG 구조를 Claude 3.5 Sonnet 및 Gemini 1.5 Flash에 적용한 경우에도 제안 모델 대비 낮은 성능을 보였다. 이는 단순히 RAG를 도입하는 것뿐 아니라, 고성능 LLM 선택과 프롬프트 설계의 정교함이 시스템 전체 품질에 중대한 영향을 준다는 점을 시사한다.

본 연구는 기존 챗봇 연구와 달리, 응답의 전반적 품질뿐 아니라 카테고리별 성능 편차에 대한 정량적 분석을 수행하였다는 점에서 차별성을 가진다. 실험 결과는 RAG 기술이 보험과 같은 도메인 특화 환경에서 LLM의 한계를 보완하고, 정확도와 신뢰성 측면 모두에서 응답 품질을 개선할 수 있음을 실증적으로 입증하였다. 또한 카테고리별 성능 차이를 기반으로, 후속 연구에서 세부 도메인 특화 학습, 프롬프트 튜닝, 데이터 다양성 확보 등이 필요함을 도출하였다. 이는 결론에서 제시한 향후 연구 과제와도 밀접하게 연결되며, 본 연구의 실질적인 기여를 강화한다.

Ⅴ. 응용 프로그램

본 연구는 보험 상담에 최적화된 실시간 챗봇 시스템을 구현하기 위해 사용자 친화적인 사용자 인터페이스(User Interface)를 포함한 웹 애플리케이션(Web Application)을 개발하였다. 이 응용 프로그램은 데이터 기반 개발 프레임워크인 Streamlit 기반으로 구현되었으며[18], 사용자가 입력한 질문을 자동으로 처리하고 RAG를 활용하여 신속하고 정확한 답변을 제공한다. 시스템의 주요 흐름은 다음과 같다. 사용자가 질문을 입력

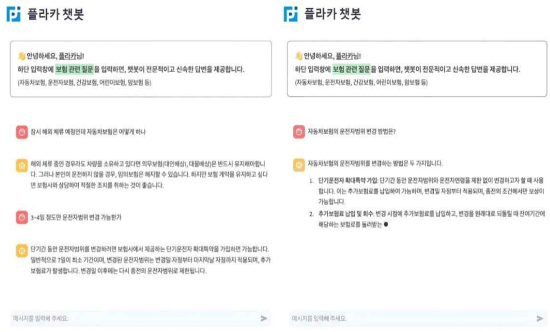

Fig. 3 User Interface Screenshots of Insurance Counseling Chatbot's Query Response System

하면, 해당 질문은 Huggingface 기반 질문 분류기를 통해 6가지 보험 카테고리(자동차보험, 운전자보험, 건강보험, 암보험, 어린이보험, 기타문의) 중 하나로 분류된다. 이후 LangChain 프레임워크를 사용하여 적합한 문서를 검색하고, 검색된 정보를 바탕으로 RAG 모델이 자연스러운 답변을 생성한다.

[그림 3]는 본 웹 애플리케이션의 대화형 인터페이스를 보여주는 화면 캡처이다. 본 웹 애플리케이션은 자동차보험, 운전자보험, 건강보험을 포함한 다양한 보험 카테고리들에 대한 전문적인 상담 기능을 제공한다. 사용자가 보험 관련 질문을 입력하면 실시간 스트리밍 방식으로 단계별 답변이 제공된다. 이를 통해 지연 시간을 최소화하고 사용자와 챗봇 간의 자연스러운 대화 흐름을 가능하게 한다. 본 시스템은 직관적인 사용자 인터페이스와 높은 응답 정확도를 바탕으로, 실시간 보험 상담 환경에서 요구되는 전문적이고 효율적인 상담 서비스를 제공한다는 점에서 중요한 의의를 가진다.

Ⅵ. 결 론

본 연구에서는 검색 증강 생성(RAG) 기술과 LangChain 프레임워크를 결합하여 보험 도메인에 특화된 고성능 FAQ 챗봇 시스템을 설계 및 구현하였다. 기존 대규모 언어 모델(LLM)이 도메인 특화 용어 처리와 최신 정보 반영에서 한계를 보이고, 민감한 정보를 다루는 분야에서 신뢰성 문제가 발생할 수 있다는 문제의식에서 출발하여, 본 연구는 외부 지식 기반 검색을 결합해 보다 정확하고 신뢰성 있는 응답을 생성할 수 있는 시스템을 제안했다.

제안한 시스템은 GPT-4o를 응답 생성 모델로 사용하고, OpenAI 임베딩 모델과 BM25 및 FAISS 기반 하이브리드 검색기를 통합하여 질의 임베딩, 문서 분할, 검색, 프롬프트 구성, 응답 생성의 전 과정을 LangChain으로 유기적으로 연결했다. 실제 보험 상담 QA 데이터를 활용한 정량 평가 실험 결과, BLEU 및 METEOR 지표에서 기존 RAG 미적용 모델 및 다른 RAG 기반 모델(Claude, Gemini 등)보다 우수한 성능을 기록했으며, 특히 자동차 보험 및 운전자보험 카테고리에서 높은 응답 품질을 보였다. 다만, 암보험 및 건강보험 영역에서는 데이터 부족과 높은 도메인 난이도로 인해 상대적으로 낮은 성능을 보였으며, 이는 카테고리별 도메인 적응의 필요성을 시사한다.

본 연구의 주요 기여는 다음과 같다:

(1) 보험 QA 도메인 특화 RAG 기반 챗봇의 설계·구현 및 보험 유형별 정량적 성능 분석; (2) BM25와 FAISS를 결합한 하이브리드 검색기 설계 및 그 효과 실험적 검증; (3) 특정 카테고리에서의 성능 저하 원인 분석 및 데이터·문서 편향 문제 확인; (4) Streamlit 기반 실시간 챗봇 UI 구현 및 실무 적용 가능성 검증.

향후 연구에서는 암보험 등 고난도 카테고리의 성능 개선을 위한 도메인 적응 및 데이터 증강, 다언어 확장, 실시간 대화 최적화 및 UI/UX 개선, 개인정보 보호를 고려한 프라이버시 보존 설계 등이 필요할 것이다.

결론적으로, 본 연구는 RAG 기반 챗봇이 보험과 같은 고신뢰성이 요구되는 분야에서도 정확성, 신뢰성, 실용성을 동시에 충족할 수 있음을 보여주었으며, 향후 다양한 산업 분야로의 확장 가능성을 제시하는 기반을 마련했다.

REFERENCES

[1] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., “Language models are few-shot learners,” *arXiv preprint arXiv:2005.14165*, May 2020. DOI: 10.48550/arXiv.2005.14165.

[2] E. M. Bender, T. Gebru, A. M. Major, and S. Shmitchell, “On the dangers of stochastic parrots: Can language models be too big?🦜,” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, New York: USA, pp. 610-623, 2021. DOI: 10.1145/3442188.3445922.

[3] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Kuttler, M. Lewis, W. T. Yih, T. Rocktaschel, S. Riedel, and D. Kiela, “Retrieval-augmented generation for knowledge-intensive nlp tasks,” in *Proceedings of the 34th Advances in Neural Information Processing Systems*, Red Hook: USA, pp. 9459-9474, 2020.

[4] J. Sohn, Y. Park, C. Yoon, S. Park, H. Hwang, M. Sung, H. Kim, and J. Kang, “Rationale Guided Retrieval Augmented Generation for Medical Question Answering,” *arXiv preprint arXiv:2411.00300*, Nov. 2024. DOI: 10.48550/arXiv.2411.00300.

[5] N. Wiratunga, R. Abeyratne, L. Jayawardena, K. Martin, S. Massie, I. N. Orji, R. Weerasinghe, A. Liret, and B. Fleisch, “CBR-RAG: Case-based reasoning for retrieval augmented generation in llms for legal question Answering,” *arXiv preprint arXiv:2404.04302*, Apr. 2024. DOI: 10.48550/arXiv.2404.04302.

[6] G. W. Yi and S. K. Kim, “Design of a question-answering system based on rag model for domestic companies,” *Journal of The Korea Society of Computer and Information*, vol. 29, no. 7, pp. 81-88, Jul. 2024. DOI: 10.9708/jksci.2024.29. 07.081.

[7] Github. langchain-ai / langchain [Internet]. Available: https://github.com/langchain-ai/langchain.

[8] K. Pandya and M. Holia, “Automating customer service using langchain: building custom open-source GPT Chatbot for organizations,” *arXiv preprint arXiv:2310.05421*, Oct. 2023. DOI: 10.48550/arXiv.2310.05421.

[9] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, and H. Wang, “Retrieval-augmented generation for large language models: a survey,” *arXiv preprint arXiv:2312.10997*, Dec. 2023. DOI: 10.48550/arXiv. 2312.10997.

[10] F. Liu, Z. Kang, and X. Han, “Optimizing RAG Techniques for Automotive Industry PDF Chatbots: A Case Study with Locally Deployed Ollama Models,” *arXiv preprint arXiv: 2408.05933*, Aug. 2024. DOI: 10.48550/arXiv. 2408.05933.

[11] S. W. Jung, Y. S. Ha, and H. Lee, “Langchain-based malware detection framework to solve hallucination phenomenon,” *Journal of Convergence Security Association (KOCOSA)*, vol. 24, no. 5, pp. 171-176, Dec. 2024. DOI: 10.33778/kcsa.2024.24.5.171.

[12] J. Y. Suh, M. Kwak, S. Y. Kim, and H. Cho, “Making a prototype of seoul historical sites chatbot using langchain,” *arXiv preprint arXiv:2402.06929*, Feb. 2024. DOI: 10.48550/arXiv.2402.06929.

[13] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P. E. Mazare, M. Lomeli, L. Hosseini and H. Jegou, “The faiss library,” *arXiv preprint arXiv:2401.08281*, Jan. 2024. DOI: 10.48550/arXiv.2401.08281.

[14] S. Robertson and S. Walker, “Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval,” in *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, Dublin: IE, pp. 232-241, 1994. DOI: 10.1007/978-1-4471-2099-5_24.

[15] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, “Bleu: A method for automatic evaluation of machine translation,” in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia: USA, pp. 311-318, 2002. DOI: 10.3115/1073083.1073135.

[16] S. Banerjee and A. Lavie, “Meteor: An automatic metric for mt evaluation with improved correlation with human judgments,” in *Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor: USA, pp. 65-72, 2005.

[17] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. V. Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush, “Transformers: State-of-the-Art Natural Language Processing,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online, pp. 38-45, 2020. DOI: 10.18653/v1/2020.emnlp-demos.6.

[18] Streamlit. A faster way to build and share data apps [Internet]. Available: https://streamlit.io/.

이태우(Taewoo Lee)

2015년 홍익대학교 경영학과 학사
2021년 고려대학교 EMBA 경영대학원 석사
2022년 3월 ~ 현재 Swiss School of Management DBA in AI Data 박사과정
2015년 12월 ~ 현재 (주)인슈넷FC 대표
2020년 5월 ~ 현재 (주)플라카 이사
※관심분야 : 딥러닝, 인공지능(AI), 데이터 분석, 데이터 보안

김도연(Doyeon Kim)

2025년 홍익대학교 컴퓨터공학과 학사
2023년 ~ 현재 (주)플라카 연구원
※관심분야 : 딥러닝, 자연어처리(NLP), LLM

심재정(Jaejeong Shim)

2024년 홍익대학교 컴퓨터공학과 학사
2024년 12월 ~ 현재 (주)플라카 연구원
※관심분야 : HCI, 자연어처리(NLP), 인공지능(AI)

한혜원(Hyewon Han)

2023년 홍익대학교 컴퓨터공학과 학사
2023년 ~ 현재 (주)플라카 연구원
※관심분야 : 인공지능(AI), 딥러닝, 머신러닝

김태용(Taeyong Kim)

2024년 홍익대학교 컴퓨터공학과 학사
2022년 ~ 2024년 (주)플라카 연구원
2025년 ~ 현재 연세대학교 전기전자공학과 석박사통합과정
※관심분야 : 딥러닝, 컴퓨터비전, 멀티모달, 연합학습, 생체보안