JDFR Journal of Data Forensics Research
데이터포렌식연구, 2권 2호, 2025년 12월, 161-174면
Journal of Data Forensics Research, Volume 2, Issue 2, December 2025, Pages 161-174.
https://doi.org/10.12972/JDFR.2025.2.2.3
Print ISSN 3058-9460
Online ISSN 3058-9479

원저

검색 증강 생성(RAG)을 활용한 보이스피싱 상담 챗봇의 성능 최적화 연구

공도현[1], 정지호[1], 최주현[2], 김경종[3]

[1]경찰대학 행정학과 학생
[2]경찰대학 행정학과 연구원
[3]경찰대학 행정학과 교수

교신저자: 김경종, leeyeongul@police.go.kr

요약

최근 보이스피싱 수법이 고도화됨에 따라 이에 대응할 수 있는 인공지능 상담 챗봇의 필요성이 대두되고 있다. 그러나 기존 LLM 기반 챗봇은 부정확한 정보를 생성하는 '환각' 문제로 인해 신뢰성이 중요한 상담 분야에 적용하는 데 한계가 있다. 본 연구는 이러한 문제를 해결하고 신뢰할 수 있는 상담 시스템을 구현하기 위해 검색 증강 생성(RAG) 기술을 적용한 챗봇 시스템을 제안하며, 성능 향상을 위한 최적화 방법론을 제시한다. 이를 위해 첫째, 문장 재진술 기법을 활용한 데이터 증강으로 학습 데이터의 품질을 개선하였다. 둘째, 5종의 최신 문장 임베딩 모델을 비교 분석하여 검색 정확도가 가장 우수한 'multilingual-e5-large' 모델을 선정하였다. 셋째, RAG 시스템의 성능을 결정짓는 핵심 요소인 프롬프트 구조와 참조 문서 정렬 순서를 최적화하였다. 성능 평가 결과, 제안된 최적화 RAG 모델은 기존 LLM 단독 모델 대비 월등한 수치 향상을 보였다. 본 연구는 실제 보이스피싱 대응 현장에서 즉시 활용 가능한 고성능 AI 상담 시스템 구축을 위한 구체적인 기술적 가능성을 제시했다는 점에서 의의가 있다.

주제어

보이스피싱, 챗봇, 검색 증강 생성, 데이터 증강, 프롬프트 최적화

Received: December 10, 2025
Revised: December 27, 2025
Accepted: December 27, 2025
Published: December 31, 2025

Original Article

Study on performance optimization of a voice phishing consultation chatbot using retrieval-augmented generation

Dohyeon Kong[1], Jiho Jung[1], Juhyun Choi[2], Kyungjong Kim[3]

[1]Undergraduate Student, Department of Police Administration, Korean Police National University, Republic of Korea
[2]Researcher, Department of Police Administration, Korean Police National University, Republic of Korea
[3]Professor, Department of Police Administration, Korean Police National University, Republic of Korea

Corresponding Author: Kyungjong Kim, leeyeongul@police.go.kr

ABSTRACT

As voice-phishing methods become more sophisticated, the need for artificial intelligence counseling chatbots capable of effective responses has increased. However, existing large language model (LLM)-based chatbots have limitations in counseling applications due to the "hallucination" problem, where false information is presented as fact. To develop a reliable counseling system, this study applied retrieval-augmented generation (RAG) technology and proposed an optimization methodology to improve chatbot performance. First, data augmentation through sentence paraphrasing was performed to improve response quality. Second, five state-of-the-art embedding models were compared to identify the optimal model for search accuracy, resulting in the adoption of "multilingual-e5-large." Finally, prompt structure and reference document ordering, which are key factors in RAG performance, were optimized through experimentation. Evaluation results show that the proposed optimized RAG model significantly outperforms the stand-alone LLM. This study demonstrates practical possibilities for building high-performance AI counseling systems applicable to real-world voice-phishing responsescenarios.

KEYWORDS

voice phishing, chatbot, retrieval-augmented generation, data augmentation, prompt optimization

I. 서론

1.1. 연구 배경

보이스피싱은 통신 매체를 이용하여 불특정 다수를 기만하고 금전적 이득을 취하는 지능형 범죄로, 그 수법이 날로 교묘해지면서 심각한 사회 문제로 자리 잡았다. 과거 특정 연령층에 집중되었던 피해는 이제 연령과 계층을 가리지 않고 전방위적으로 확산되고 있으며, 1인당 피해액 또한 급증하는 추세로 범죄의 심각성이 더욱 커지고 있다[1]. <Figure 1>에 나타난 바와 같이, 보이스피싱 범죄는 2006년 최초 발생 이후 급격한 증가 추세를 보였으며, 최근까지도 여전히 심각한 수준을 유지하고 있다. 특히 <Figure 2>와 같이 1건당 피해 금액이 급증하는 추세는 보이스피싱 범죄가 더욱 지능화되고 고액 피해를 유발하는 형태로 진화하고 있음을 보여준다.

출처: 경찰청, 「보이스피싱 통계자료」, 내부자료, 2022.6.

<Figure 1> Amount of Damage Caused by Voice Phishing

출처: 경찰청, 「보이스피싱 통계자료」, 내부자료, 2022.6.

<Figure 2> Yearly Number of Voice Phishing Occurrences

이러한 상황에서 잠재적 피해자가 신속하고 정확하게 대응할 수 있도록 돕는 즉각적인 지원 시스템의 필요성이 대두되고 있다. 현재 운영 중인 전기통신금융사기 통합대응단 등 기존의 대응 체계는 상담 인력의 한계로 인해 24시간 즉각적인 대응이 어렵다는 문제점이 있다. 이에 최신 대형 언어 모델(Large Language Model, LLM)[2]기술을 활용한 AI 챗봇은 시간과 장소에 구애받지 않고 사용자에게 즉각적인 상담을 제공할 수 있어 효과적인 1차 대응 수단으로 주목받고 있다. 그러나 기존의 LLM 기반 챗봇을 도입하더라도 모델이 사실과 다른 내용을 그럴듯하게 생성하는 '환각(Hallucination)' 현상과 외부 상용 API 사용 시 발생할 수 있는 금융 정보 및 개인 신상 등 민감 정보의 유출 가능성이 존재한다.

이러한 문제를 해결하기 위해 본 연구는 '로컬 LLM 기반의 검색 증강 생성(RAG, Retrieval-Augmented Generation)[3]' 기술을 도입할 것을 제안한다. RAG는 LLM이 답변을 생성할 때 신뢰할 수 있는 외부 지식베이스를 참조하여 환각을 억제하고, 로컬 환경에서 LLM을 직접 구동함으로써 모든 데이터 처리가 폐쇄된 환경 내에서 이루어져 민감 정보의 외부 유출을 원천적으로 차단할 수 있다[4].

1.2. 연구 목표 및 기여

본 연구의 최종 목표는 RAG 기술과 로컬 LLM을 결합하여, 보이스피싱 관련 질의에 대해 환각 현상을 최소화하고 데이터 유출 우려 없이 안전하며 정확한 답변을 생성하는 고성능 챗봇 시스템을 구축하고 그 효용성을 실증적으로 검증하는 것이다.

이를 위한 세부 목표 및 기여는 다음과 같다.

(1) 한국어 보이스피싱 데이터에 특화된 최적의 문장 임베딩 모델과 프롬프트 구성 방식을 실증적으로 제시한다.

(2) RAG 기술을 로컬 LLM에 적용한 보안 강화 파이프라인을 구체적으로 제시함으로써, 유사 시스템 개발의 기술적 토대를 마련한다.

(3) 신뢰도 높은 정보 제공과 데이터 보안이라는 두 가지 핵심 요소를 모두 만족시켜, 보이스피싱 대응에 실질적으로 기여할 수 있는 AI 시스템의 가능성을 입증한다.

이러한 일련의 최적화 과정을 통해, 실제 현장에서 즉시 활용 가능한 고성능 보이스피싱 상담 챗봇의 최적화 전략을 제안하고자 한다.

II. 관련 연구

2.1. 기존 챗봇 연구 동향 및 한계

챗봇 기술은 초기 규칙 기반(Rule-based) 방식에서 시작하여 의도 분류 모델을 거쳐 최근의 생성형 AI 기반 모델로 비약적인 발전을 거듭해왔다[5]. 초기 챗봇은 선택형 혹은 단답형 방식으로, 개발자가 사전에 정의한 특정 키워드나 규칙에 매칭되는 경우에만 정해진 답변을 출력하는 형태였다. 이러한 규칙 기반 챗봇은 설계가 단순하고 명확한 답변을 제공한다는 장점이 있으나, 사용자의 발화가 미리 설정된 시나리오를 벗어나거나 복잡한 문맥을 포함할 경우 적절히 대응하지 못하는 확장성의 한계가 뚜렷했다[6].

이후 머신러닝 기술의 발전과 함께 분류형 챗봇이 등장하였다. 이 방식은 사전에 정의된 카테고리로 분류하여 적절한 답변을 제공한다. 현재 전기통신금융사기 통합대응단 등이 이 방식을 채택하고 있으며, 규칙 기반 방식보다는 유연한 대응이 가능하다[6]. 하지만 여전히 학습 데이

터에 포함되지 않은 새로운 유형의 질문이나 모호한 상황에 대해서는 사전에 정의된 의도 카테고리로 매칭시키지 못하는 한계가 존재한다.

최근에는 대규모 언어 모델(LLM)의 등장으로 유창한 대화형 챗봇이 주목받고 있다. LLM 기반 챗봇은 방대한 데이터를 학습하여 사람과 유사한 수준의 자연스러운 대화가 가능하며, 특정 주제에 국한되지 않고 다양한 질문에 답변할 수 있는 범용성을 갖추었다. 그러나 이러한 생성형 모델은 확률적 생성 방식에 의존하기 때문에 사실과 다른 정보를 그럴듯하게 꾸며내는 '환각(Hallucination)' 현상이 발생할 수 있다[2]. 특히 정확한 사실 전달이 필수적인 보이스피싱 예방 및 법률 상담 분야에서는 이러한 정보의 오류가 사용자에게 치명적인 피해를 줄 수 있어, LLM을 단독으로 적용하기에는 신뢰성 측면에서 한계가 있다.

따라서 본 연구에서는 기존 챗봇들의 한계를 극복하고, 유창한 대화 능력과 정확한 정보 전달을 동시에 충족하기 위해 검색 증강 생성(RAG) 기술을 챗봇 시스템에 도입한다[4]. RAG는 LLM의 생성 능력에 신뢰할 수 있는 외부 지식 베이스의 정보를 기반으로 답하는, '신뢰 가능한 전문가형' 챗봇을 구현하는 데 효과적인 대안이 될 수 있다[3].

2.2. 검색 증강 생성 (RAG)

LLM은 트랜스포머(Transformer) 아키텍처를 기반으로 방대한 텍스트 데이터를 학습하여 인간의 언어를 이해하고 생성하는 능력을 갖춘 인공지능 모델이다[2]. 그러나 이러한 모델의 지식은 학습 시점에 고정되어 있어 최신 정보를 반영하지 못하는 한계가 있다. RAG는 이러한 LLM의 한계를 보완하기 위해 제안된 아키텍처로, LLM을 외부의 신뢰할 수 있는 지식 소스와 결합하는 방식으로 작동한다. 구체적으로, RAG는 LLM이 답변을 생성하기 전에 데이터 베이스에서 관련 문서를 검색하고 이를 프롬프트에 포함하여 답변을 생성한다. 이를 통해 모델 재학습 없이도 최신 정보를 반영할 수 있으며, 답변의 근거를 명확히 하여 신뢰성을 확보할 수 있다[4].

이는 크게 색인(Indexing), 검색(Retrieval), 생성(Generation)의 3단계로 구성된다[7].

(1) 색인(Indexing): 외부 지식 소스(PDF, 문서 등)를 청크(Chunk) 단위로 분할한 후, 임베딩 모델을 통해 고차원 벡터로 변환하여 벡터 데이터베이스(Vector DB)에 저장하는 과정이다.

(2) 검색(Retrieval): 사용자의 질의가 입력되면 이를 동일한 임베딩 모델로 벡터화한다. 이후 벡터 DB 내에서 코사인 유사도 등의 지표를 활용하여 질의와 의미적으로 가장 유사한 상위 k개의 문서를 추출한다.

(3) 생성(Generation): 검색된 문서와 원본 질의를 결합하여 프롬프트를 구성하고, 이를 LLM에 입력하여 최종 답변을 생성한다.

이 과정을 통해 RAG는 LLM이 마치 '오픈북 시험'을 치르듯, 검증된 자료를 참고하여 답변을 생성하도록 유도함으로써 환각을 줄이고 답변의 정확성과 신뢰성을 높인다[8]. 그러나 단순히 외부 문서를 검색하여 제공하는 것만으로 완벽한 성능을 보장하는 것은 아니다. 검색된 문서가 질의와 의미적으로 유사하더라도 부정확한 정보를 포함하고 있을 경우 오히려 모델의 환각을 악화시킬 수 있다. 반면, 관련성이 다소 떨어지는 정보가 섞여 있더라도 LLM의 고유한 추론 능력을 통해 정답을 도출할 수 있는 등, 검색된 정보의 '질'과 '맥락'이 성능에 복합적인 영향을 미친다[9]. 따라서 고성능 RAG 시스템을 구축하기 위해서는 단순한 검색을 넘어, LLM이 정답을 도출하는 데 실질적으로 기여할 수 있는 최적의 문서를 선별하고 배치하는 고도화된 전략이 필수적이다.

2.3. 문장 임베딩 모델

문장 임베딩이란 자연어로 구성된 문장을 그 의미적 특성을 반영하여 고차원 벡터 공간의 수치로 변환하는 기술을 말한다. RAG 시스템에서 생성되는 답변의 품질은 관련 정보를 얼마나 정확하게 찾아내느냐에 달려 있으므로, 고성능 문장 임베딩 모델의 사용은 시스템 성능 확보에 필수적인 요소이다.

문장 임베딩 기술의 초기 연구들은 주로 사전 학습된 언어 모델들이 가지는 표현력의 한계를 개선하는 방향으로 진행되었다. 그 대표적인 예로 Sentence-BERT(SBERT)를 들 수 있다. 이 모델은 기존 BERT 구조에 샴 네트워크(Siamese Network) 방식을 적용하여 미세 조정함으로써, 문장 간의 유사도를 코사인 유사도로 효과적으로 계산할 수 있게 만들어 문장 임베딩의 활용성을 크게 높였다[10]. 이후 발표된 SimCSE는 대조 학습(Contrastive Learning) 기법을 도입하여 성능을 한 단계 끌어올렸다. 동일한 문장에 서로 다른 드롭아웃 노이즈를 주어 긍정 쌍을 만드는 간단하지만 강력한 방식으로, 임베딩 벡터가 특정 방향으로 쏠리는 비등방성 문제를 해결하였다[11].

최근에는 대규모의 약한 감독 데이터(Weakly Supervised Data)를 활용하고 정교한 대조 학습 기법을 적용한 모델들이 의미 검색 분야를 주도하고 있다. 그중 E5 계열 모델은 SBERT나 SimCSE와 같은 기존 모델 대비 방대한 코퍼스 학습을 통해 다국어 및 제로샷(Zero-shot) 태스크에서 뛰어난 일반화 능력을 보이며, RAG 시스템의 검색 모듈로서 가장 효율적인 대안으로 평가받고 있다[12].

III. 실험 설정

3.1. 데이터셋 및 실험 환경

본 연구는 경찰청의 실제 피해자 상담 사례를 재가공하여 구축한 100개의 사용자 질의(Query)와, 이에 대한 답변(Ground Truth) 쌍을 바탕으로 실제 보이스피싱 피해 상황을 반영한 고품질 데이터셋을 구성하였다. 이를 통해 챗봇이 다양한 피싱 시나리오에 대해 얼마나 정확하고 실질적인 조언을 제공할 수 있는지 검증하고자 한다.

실험은 대규모 언어 모델의 연산 부하를 처리하기 위해 NVIDIA A100 80GB PCIe GPU 2대를 탑재한 고성능 서버 환경에서 진행되었다. 소프트웨어 환경은 Ubuntu 리눅스 운영체제 상에서 NVIDIA 드라이버 버전 575.57.08과 CUDA 12.9 버전을 기반으로 구성하였다. 특히, 본 연구는 보이스피싱 상담 과정에서 다루어지는 민감한 개인정보와 금융 데이터의 보안을 최우선으로 고려하였다. 이를 위해 외부 클라우드나 상용 API를 일절 사용하지 않고, 모든 데이터 처리와 모델 구동이 외부 망과 단절된 로컬 서버(On-Premise) 내에서 독립적으로 수행되도록 환경을 구축하여 정보 유출 가능성을 원천적으로 차단하였다. LLM의 효율적인 로컬 구동을 위해 Ollama 프레임워크를 활용하였으며, RAG 시스템의 검색 및 생성 파이프라인은 Sentence-Transformers 및 FAISS-CPU 라이브러리를 사용하여 구현하였다.

3.2. 성능 평가 지표

챗봇이 생성한 답변이 정답(Ground Truth)과 얼마나 유사한지 정량적으로 평가하기 위해, 의미적 유사도를 나타내는 코사인 유사도와 BERTScore(F1)를 주요 지표로 채택하였다.

▶ 코사인 유사도 (Cosine Similarity)

두 벡터 $\boldsymbol{A}$(정답 임베딩)와 $\boldsymbol{B}$(생성 답변 임베딩) 사이의 각도 코사인을 측정하여 의미적 방향성을 평가하는 지표이다. 두 벡터의 내적을 각 벡터의 크기(L2 Norm) 곱으로 나누어 계산한다. 수식은 다음과 같다.

$$C(A,B) = \frac{A \cdot B}{|A| \cdot |B|} \quad (1)$$

이는 RAG 시스템이 생성한 답변이 정답과 얼마나 같은 의미를 담고 있는지를 벡터 공간에서 효율적으로 평가한다[13].

▶ BERTScore (F1)

사전 학습된 언어 모델인 BERT를 활용하여, 생성 문장과 정답 문장 간의 토큰별 임베딩 유사도를 계산하여 정밀도(P)와 재현율(R)을 도출한다. 정밀도(P)는 생성된 답변의 각 토큰이 정답 문장의 토큰과 얼마나 유사한지를 측정하며, 재현율(R)은 정답 문장의 각 토큰이 생성된 답변에 얼마나 잘 반영되었는지를 측정한다. 이들을 조화 평균하여 최종 F1 점수를 산출한다. 수식은 다음과 같다.

$$R_{BERT} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x_j} \in \hat{x}} (x_i^T \hat{x_j}) \quad (2)$$

(단, x_i : 정답 단어의 임베딩 벡터, T: 전치)

$$P_{BERT} = \frac{1}{|\hat{x}|} \sum_{\hat{x_j} \in \hat{x}} \max_{x_i \in x} (x_i^T \hat{x_j}) \quad (3)$$

(단, $\hat{x_j}$: 생성 단어의 임베딩 벡터, T: 전치)

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R} \quad (4)$$

이는 RAG 시스템이 생성한 답변에 대해 문맥적 의미를 고려한 더욱 정교하고 인간적인 수준의 유사성을 평가할 수 있다[14].

3.3. 모델 선정 및 하이퍼파라미터 설정

제안하는 RAG 시스템의 핵심인 생성 모델과 검색 모델은 한국어 처리 능력과 보안성을 고려하여 선정하였다. 먼저, 답변 생성을 담당하는 LLM은 Gemma3:27b를 채택하였다. 특히, 사용자의 민감한 상담 내용이 외부로 유출되는 것을 방지하고 개인정보를 보호하기 위해, 외부 API를 호출하는 방식이 아닌 모든 실험을 로컬 환경에서 직접 모델을 구동하여 수행하였다.

검색 정확도를 결정짓는 문장 임베딩 모델은 성능 비교를 위해 비교 모델로 KR-SBERT-klueNLI, all-MiniLM-L6-v2, ko-sbert-sts, multilingual-MiniLM, multilingual-e5-large 5종을 사용하였다. SBERT 모델 계열인 KR-SBERT-klueNLI는 NLI 기반 Fine-tuning을 통해 문장 유사도 측정에 특화되었다. ko-sbert-sts는 STS(Semantic Textual Similarity) 태스크 기반 Fine-tuning을 통해 문장의 의미적 유사도 성능을 극대화한

다[10].

MiniLM 모델 계열인 all-MiniLM-L6-v2는 지식 증류(Knowledge Distillation)를 통해 원본 BERT 모델 대비 사이즈가 작고 추론 속도가 매우 빠르다는 특성을 지닌다. multilingual-MiniLM은 MiniLM 아키텍처를 다국어 코퍼스에 적용하여 효율성과 다국어 지원을 결합한 특성을 지닌다[15].

마지막으로 E5 모델 계열인 multilingual-e5-large는 규모 대조 학습(Contrastive Learning)을 기반으로 학습되어, 다국어 및 제로샷(Zero-shot) 태스크에서 SOTA 성능을 보인다.

다음으로 답변 생성을 위한 정보 검색 단계에서는 고속 유사도 검색 라이브러리인 FAISS를 활용하여 벡터 데이터베이스를 구축하였다. 검색 품질 최적화를 위해 핵심 하이퍼파라미터는 다음과 같이 설정하였다. 사용자 질의와 가장 유사한 문서를 참조하기 위한 Top-k 값은 5로 설정하여 충분한 문맥 정보를 제공하도록 했으며, 관련성이 낮은 정보가 답변에 포함되어 환각(Hallucination)을 유발하는 것을 방지하기 위해 코사인 유사도 임계값(Threshold)을 0.8로 엄격하게 설정하여 신뢰성을 확보하였다.

Ⅳ. 연구 방법론

4.1. 보이스피싱 상담 챗봇 시스템 아키텍처

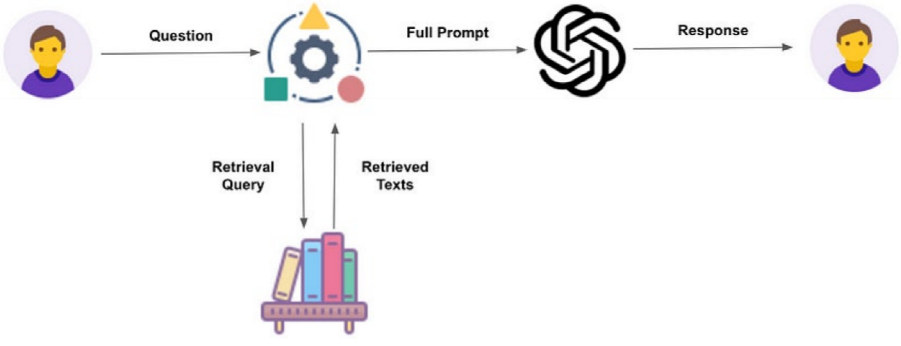
제안하는 시스템은 검색 증강 생성(RAG) 아키텍처를 따른다. 시스템의 작동 흐름은 <Figure 3>과 같으며, 주요 단계는 다음과 같다.

DB 구축: 증강된 전체 시나리오 데이터셋의 각 문장을 미리 선정한 문장 임베딩 모델(intfloat/multilingual-e5-large)을 이용하여 벡터로 변환해 벡터 데이터베이스를 구축한다. 모델 선정 이유에 대해서는 5.1 세션에서 설명한다.

벡터 변환: 사용자의 실시간 질의(Query)가 입력되면, 동일한 임베딩 모델을 통해 질의 문장을 벡터로 변환한다.

유사도 검색 및 프롬프트 구성: 질의 벡터와 데이터베이스 내 모든 벡터 간의 코사인 유사도를 계산하여, 가장 관련성 높은 상위 5개의 문서를 검색한다. 검색된 문서(Context)와 사용자 질의(Query)를 결합하여 LLM에 입력할 최종 프롬프트를 구성한다.

답변 생성: 구성된 프롬프트를 LLM에 전달하여 최종 답변을 생성한다.



<Figure 3> RAG-based Consultation Chatbot Architecture

4.2. 데이터셋 구축 및 증강

실제 보이스피싱 피해 상황을 가정한 질의와 답변을 바탕으로 모델의 일반화 성능을 확보하기 위해, 초기 데이터셋에 Paraphrasing 기반의 데이터 증강을 적용하였다. 사전에 로컬 환경에 준비한 Gemma3:27b 모델을 사용하여 원본 문장의 의미는 유지하되, 문체나 단어 표현, 어순 등을 다양하게 변형한 새로운 문장들을 생성하여 데이터셋의 크기를 확장하였다. 예를 들어, 'OK금융 담당자와 전화로 대출상담을 했는데 이후 기존 대출이 있는 신한카드 담당자한테서 계약위반을 했으니 기존 대출금을 대면상환 해야 한다면서 해지통보서 및 지급정지통보서를 보내왔어요'라는 문의 내용을 'OK금융과 대출 상담을 한 후에 신한카드로부터 기존 대출금에 대한 계약 위반으로 해지 및 지급 정지 통보를 받게되었습니다.' 등 아래 예시 사진과 같이 증강하였다. 이를 이용하여 보이스피싱 신고대응센터를 통해 받은 100개의 상담 내용을 10배 증강하여 총 1,000개의 상담 내용을 데이터로 구축하였다.

아래 <Figure 4>는 문장 증강 예시이다.

OK금융 담당자와 전화로 대출 상담을 진행한 후, 신한카드 담당자로부터 계약 위반으로 기존 대출금을 즉시 상환하라는 해지 및 지급정지 통보서를 받았습니다.
대출 상담을 OK금융과 통화했는데, 그 후 신한카드 쪽에서 계약 위반을 이유로 기존 대출금 전액을 상환하라고 하며 해지 통보서와 지급정지 통보서를 보냈어요.
OK금융에서 대출 상담을 받았는데, 이후 신한카드 담당자가 계약 위반을 주장하며 기존 대출금의 즉시 상환을 요구하는 통보서를 보내왔습니다.
저는 OK금융 담당자와 대출 상담을 했고, 그 결과 신한카드에서 기존 대출 관련 계약 위반을 이유로 해지 통보서와 함께 지급정지 통보서를 받았습니다.
OK금융과 대출 상담을 한 후에 신한카드로부터 기존 대출금에 대한 계약 위반으로 해지 및 지급 정지 통보를 받게 되었습니다.
신한카드 담당자에게서 계약 위반을 이유로 기존 대출금을 즉시 상환하라는 해지통보서와 지급정지통보서를 받았는데, 이는 OK금융과의 대출 상담 이후에 일어난 일입니다.

<Figure 4> Example of sentence augmentation

4.3. 프롬프트 구조와 참조 문서 정렬 순서

RAG 시스템의 프롬프트 구조는 LLM의 정보 처리 과정에 직접적인 영향을 미쳐 답변의 사실 충실도(Fidelity)와 정확도에 영향을 미친다[9]. 본 연구는 LLM이 RAG 기반 답변 생성 시 가장 정확한 답변을 얻기 위해 참조 문서(Context)를 질문(Query)보다 먼저 배치하는 Context_First 방식과 그 반대인 Query_First 방식의 성능을 비교하였다[16].

Context → Query 구조는 LLM이 답변 생성 전에 사실적 근거(Context)를 충분히 흡수하도록 유도하는 특징이 있다. Query → Context 구조는 LLM이 먼저 질의를 파악한 후 Context를 참고하게 한다. 프롬프트 구조 실험이 Context와 Query의 배치를 탐색했다면, 참조 문서 정렬 순서 실험은 Context 내부에서 문서 간의 영향력을 극대화하는 것을 목표로 한다. LLM은 처리 편향이 있기에 본 실험은 이러한 편향을 활용하여 가장 중요한 사실적 근거(Highest Relevance Document)를 LLM이 가장 잘 처리할 수 있는 위치에 배치하고자 한다.

이를 위해 코사인 유사도를 바탕으로 참조 문서 정렬 순서를 오름차순과 내림차순으로 달리하여 비교 실험하였다.

V. 실험 및 결과

5.1. 최적 임베딩 모델 선정

고품질의 검색 결과를 얻기 위해 5개의 주요 임베딩 모델의 성능을 비교하였다. 각 모델을 사용하여 각 쿼리에 대한 모델별 유사도 Top 10 문장 중 재진술 되지 않은 문장이 등장하기 전 마

지막 paraphrase 유사도 즉, 모델별 오답이 등장하기 직전까지 검색된 정답 문장의 최저 유사도 평균을 비교하였다. 아래 <Table 1>은 5개의 주요 임베딩 모델의 성능을 정답 문장의 최저 유사도 평균과 Top-10 검색 결과 내 재진술 된 문장 포함 개수, 두 가지 기준으로 비교 분석한 결과이다. <Table 1>에서 볼 수 있듯이 'multilingual-e5-large' 모델이 0.9575라는 가장 높은 점수를 기록하였다. 또한, Top-10 검색 결과 내 재진술 된 문장 포함 개수에서도 9.80개로 최상위권의 성능을 보였다. 따라서 이후 모든 실험에서는 'multilingual-e5-large'를 기본 임베딩 모델로 사용하였다.

<Table 1> Performance comparison of embedding models

Model name	Minimum answer similarity	Top-10 matches
KR-SBERT-klueNLI	0.8615	9.83
all-MiniLM-L6-v2	0.9369	4.83
ko-sbert-sts	0.8770	9.80
multilingual-MiniLM	0.8043	9.00
multilingual-e5-large	0.9575	9.80

5.2. 프롬프트 구조 성능 비교

본 연구에서는 LLM의 답변 품질을 최적화하기 위해, 검색된 참조 문서(Context)와 사용자 질문(Query)을 결합하여 모델에 입력하는 프롬프트의 구조적 배치에 따른 성능 변화를 심층적으로 비교 분석하였다. 이는 단순히 정보를 제공하는 것을 넘어, '어떤 순서로 정보를 제공하는가'에 따라 LLM의 처리 방식이 완전히 달라지기 때문이다. 구체적으로 Context_First 방식과 Query_First 방식의 두 가지 구조에 대한 성능을 측정하였다.

Context_First 방식은 검색된 참조 문서(Context)를 먼저 제시하고 그 뒤에 질문(Query)을 배치하는 구조([Context] → [Query])이다. 이는 모델에게 "여기 근거 자료가 있다. 이것을 먼저 읽고 질문에 답해라"라는 논리를 부여한다. 반면, Query_First 방식은 질문(Query)을 먼저 제시하고 그 뒤에 참조 문서(Context)를 배치하는 구조([Query] →[Context])이다. 아래 <Table 2>는 코사인 유사도와 BERTScore F1, 두 가지 지표를 활용하여 성능을 비교·분석한 결과이다.

<Table 2> Performance comparison by prompt structure

Prompt structure	Cosine	BERTScore F1
Context_First	0.9327	0.7124
Query_First	0.9269	0.7091

<Table 2>에서 볼 수 있듯이, 모든 평가 지표에서 Context_First 방식이 더 우수한 성능을 보였다. 이는 LLM이 충분한 참조 문맥을 먼저 인지한 후 질문을 이해하는 것이 답변 생성에 더 효과적임을 시사한다.

이러한 결과는 LLM의 '처리 편향(Processing Bias)'에 기인한다[16]. Query_First 방식은 모델이 질문을 먼저 받은 후, 내부 지식을 꺼내 답을 만들려는 반사적인 경향을 유발하며, 이때 환각이 발생할 틈이 생기기 쉽다. 반면, Context_First 방식은 답변을 생성하기 전에 모델이 강

제로 사실 정보(Context)를 먼저 읽도록 한다. 이는 모델의 주의(Attention)를 팩트에 먼저 '그라운딩(Grounding)' 시켜놓고 질문을 처리하게 함으로써 사실 기반의 답변을 효과적으로 유도해낸다[7].

5.3. 종합 성능 비교 분석

앞선 프롬프트 구조 성능 비교 결과를 바탕으로, 주요 설정에 따른 성능을 종합적으로 비교하고 분석하였다. <Table 3>은 RAG를 적용하지 않은 Baseline 모델과, 프롬프트 구조 및 코사인 유사도를 바탕으로 참조 문서 정렬 순서(desc: 내림차순, asc: 오름차순)를 달리한 주요 RAG 설정들의 최종 성능을 보여준다.

실험 결과, RAG 기술의 도입 자체가 성능 향상의 가장 결정적인 요인임을 확인할 수 있었다. RAG를 적용한 시스템은 의미적 정확성을 나타내는 코사인 유사도와 BERTScore에서 Baseline 모델 대비 각각 약 7.3%p, 15.8% 향상된 성능을 기록했다.

<Table 3> Overall performance comparison: Baseline and RAG key settings

Configuration	Cosine	BERTScore F1
Baseline (no RAG)	0.899	0.6531
RAG desc(C→Q)	0.965	0.7566
RAG asc(C→Q)	0.964	0.7435

VI. 결론

본 연구는 신뢰도 높은 보이스피싱 상담 챗봇을 구축하기 위해, 검색 증강 생성(RAG) 기술을 도입하고 그 성능을 체계적으로 최적화하는 방법론을 제시하였다. 연구 결과 5종의 임베딩 모델 비교를 통해 'multilingual-e5-large'가 가장 뛰어난 성능을 보임을 실험적으로 증명하였고, 프롬프트 구조 및 참조 문서 정렬 순서를 분석하여 RAG desc(C→Q) 구조가 본 시스템에 가장 최적화됨을 도출하였다.

본 연구에서 수행한 단계적 최적화 과정을 거친 최종 RAG 시스템은, 베이스라인인 LLM 단독 모델 대비 압도적인 성능 향상을 보여주었다. 이는 본 연구가 제안한 최적화 방법론이 실질적으로 유효함을 증명하는 결과이다. 결론적으로 본 연구는 고도화되는 보이스피싱 최신 수법에 실질적으로 대응 가능한 AI 상담 시스템 구축을 위해, 실험적으로 검증된 구체적인 설계 가능성을 제공했다는 점에서 학술적 가치와 실무적 유용성을 동시에 지닌다.

후속 연구에서는 현재의 텍스트 기반 분석이 가지는 한계를 보완하기 위해, 음성 데이터까지 함께 분석하는 멀티모달(Multi-modal) 상담 시스템으로의 확장을 제안한다[17]. 실제 보이스피싱 상황에서는 대화 내용뿐만 아니라 범인의 목소리 톤, 말의 속도, 불안정한 어조 등 음향적 특성이 사기 여부를 판단하는 결정적인 단서가 되기 때문이다. 이러한 음성 특징을 활용한다면, LLM이 범인의 대화 스타일이나 특정 말투 패턴을 학습하여 정교한 사칭형 보이스피싱 시도까지 정확하게 식별해낼 수 있을 것이다. 나아가 통화 중인 음성을 실시간으로 분석하여 위험 징후가 포착될 경우, 사용자에게 즉각적인 경고를 보냄으로써 실제 피해 발생을 사전에 차단하는 효과를 기대할 수 있다. 이를 위해 텍스트 분석에서 도출된 의미적 정보와 음성 신호 처리로 추출된 음향적 특징을 결합하는 '특징 융합(Feature Fusion)' 모델 개발이 필요하다[17]. 구체적인 방법으로는 텍스트와 음성의 임베딩 벡터를 결합하여 단일 모델에 입력하거나, 각 모달리티

를 개별 모델로 학습시킨 뒤 결과를 앙상블하는 방식 등을 고려해 볼 수 있다. 이러한 멀티모달 시스템은 대화 내용이 정상적으로 보이더라도 목소리나 어조에서 이상 징후를 포착할 수 있으므로, 훨씬 더 정교하고 강력한 보이스피싱 대응 체계를 구축하는 데 기여할 것이다.

또한, 본 연구가 RAG 최적화를 통해 답변의 정확도를 높이는 데 성공했지만, 시스템의 운영 효율성과 능동적인 대처 능력 면에서는 여전히 개선의 여지가 있다. 현재 구조는 질문의 난이도와 관계없이 동일한 고성능 LLM을 사용하기 때문에, 단순한 질문을 처리할 때도 불필요하게 높은 연산 자원이 소모되고 응답이 지연되는 비효율성이 존재한다. 따라서 비용 효율성과 성능을 동시에 잡기 위해, 질문의 복잡도나 검색 결과의 품질에 따라 사용할 LLM을 동적으로 결정하는 '에이전트 RAG(Agentic RAG)' 모델 도입을 제안한다[18]. 예를 들어, 데이터베이스 검색 결과만으로 충분히 답변 가능한 경우에는 MiniLM이나 경량화된 Gemma와 같은 가벼운 모델을 사용하여 즉각적으로 응답함으로써 시간을 절약한다. 반대로 검색 정보가 부족하거나 복잡한 추론이 필요한 고난이도 질문의 경우, GPT-OSS와 같은 고성능 모델로 자동 전환하여 심층적이고 정확한 답변을 제공하도록 설계한다. 이러한 하이브리드 에이전트 방식은 시스템 운영 비용 절감과 응답 속도 향상이라는 두 마리 토끼를 잡으면서도, 긴급 상황에서 요구되는 높은 정확성을 유지하는 유연하고 강력한 상담 시스템 구축을 가능하게 할 것이다.

참고문헌 (References)

[1] Yang J, Lee C, Kim SB. 2023. Development and utilization of voice phishing prevention service through KoBERT-based voice call analysis. KIISE Transactions on Computing Practices, 29(5), 205-213. http://doi.org/10.5626/KTCP.2023.29.5.205

[2] Jeong C. 2023. Generative AI service implementation using LLM application architecture: Based on RAG model and LangChain framework. Journal of Intelligence and Information Systems, 29(4), 129-145. https://doi.org/10.13088/jiis.2023.29.4.129

[3] Yoon Y, Kim S. 2025. Trends and prospects of retrieval-augmented generation (RAG) technology for generative AI. Journal of Korean Association of Computer Education, 28(2), 71-85. https://doi.org/10.32431/kace.2025.28.2.007

[4] Lewis P, Perez E, Piktus A, et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. Proceedings of the 34th International Conference on Neural Information Processing System, pp. 9459-9474.

[5] Park H. 2024. Future of voice phishing detection technology based on LLM and on-device AI. Journal of Data Forensics Research, 1(1), 177-183. https://doi.org/10.12972/JDFR.2024.1.1.11

[6] Lee DG. 2019. Design and implementation of AI chatbot platform based on SNS information [SNS 정보 기반의 인공지능 챗봇 플랫폼 설계 및 구현]. PhD dissertation, Chonnam National University, Gwangju, Korea.

[7] Gao Y, Xiong Y, Gao X, et al. 2023. Retrieval-augmented generation for large language models: A survey. arXiv:2312.10997. https://doi.org/10.48550/arXiv.2312.10997

[8] Lee HE, Yoon SW, Kwon JY, et al. 2024. An efficient prompt for public service using LLM+RAG. Proceedings of the Korean Society of Computer Information Conference, Jeju, pp. 37-39.

[9] Cuconasu F, Trappolini G, Siciliano F, et al. 2024. The power of noise: Redefining retrieval for RAG systems. Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, Washington DC, pp. 719-729. https://doi.org/10.1145/3626772.3657834

[10] Reimers N, Gurevych I. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, pp. 3982-3992. https://doi.org/10.18653/v1/D19-1410

[11] Gao T, Yao X, Chen D. 2021. SimCSE: Simple contrastive learning of sentence embeddings. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online, pp. 6894-6910. https://doi.org/10.18653/v1/2021.emnlp-main.552

[12] Wang L, Yang N, Huang X, et al. 2023. Text embeddings by weakly-supervised contrastive pre-training. arXiv:2212.03533. https://doi.org/10.48550/arXiv.2212.03533

[13] Jurafsky D, Martin JH. 2000. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition, 2nd ed. Prentice Hall.

[14] Zhang T, Kishore V, Wu F, et al. 2020. BERTScore: Evaluating text generation with BERT. International Conference on Learning Representations 2020, Online. https://iclr.cc/virtual_2020/poster_SkeHuCVFDr.html

[15] Wang W, Wei F, Dong L, et al. 2020. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. arXiv:2002.10957. https://doi.org/10.48550/arXiv.2002.10957

[16] Lu Y, Bartolo M, Moore A, et al. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Dublin, pp. 8086-8098. https://doi.org/10.18653/v1/2022.acl-long.556

[17] Baltrušaitis T, Ahuja C, Morency LP. 2019. Multimodal machine learning: A survey and

taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423-443. https://doi.org/10.1109/TPAMI.2018.2798607

[18] Yao S, Zhao J, Yu D, et al. 2023. React: Synergizing reasoning and acting in language models. 11th International Conference on Learning Representations (ICLR 2023), Kigali.