Journal of KIIT. Vol. 23, No. 1, pp. 13-20, Jan. 31, 2025. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2025.23.1.13

검색 증강 생성 기법과 소규모 대형 언어 모델을 활용한 문서 요약 기반 챗봇 시스템

우민식*, 시종욱**, 김성영***

Document Summary-based Chatbot System Utilizing Retrieval-Augmented Generation and Small Large Language Models

Minsik Woo*, Jongwook Si**, and Sungyoung Kim***

본 연구는 2023년도 중소벤처기업부의 기술개발사업 지원에 의한 연구임 [S3344882]

요 약

최근 챗봇 기술은 다양한 산업에서 작업 자동화, 고객 경험 향상, 운영 효율성 증대에 중요한 역할을 하고 있다. 하지만 대형 언어 모델(LLM) 기반 챗봇은 운영 비용이 높다는 문제가 있다. 이를 해결하기 위해 본 논문에서는 비용을 최소화하면서도 정확하고 신뢰할 수 있는 응답을 제공하기 위해 검색 증강 생성(RAG)과 파인 튜닝한 sLLM을 활용한 챗봇 시스템을 제안한다. 제안된 시스템은 모델 파인 튜닝, 검색 쿼리 생성, 관련 문서 검색, 응답 생성의 네 가지 구성 요소로 이루어지며, RDASS, ROUGE, BLEU, METEOR 지표를 사용해 성능을 평가하였다. 그 결과 제안 방법의 챗봇 시스템은 보다 간결하고 관련성 높은 요약을 통해 답변을 제공할 수 있음을 확인하였다.

Abstract

Recent advancements in chatbot technology have played a significant role in automating tasks, enhancing customer experience, and improving operational efficiency across various industries. However, Large Language Model (LLM)-based chatbots face the challenge of high operational costs. To address this issue, this paper proposes a chatbot system utilizing a fine-tuned small LLM(sLLM) incorporating Retrieval-Augmented Generation(RAG) aiming to minimize costs while providing accurate and reliable responses. The proposed system consists of four key components: model fine-tuning, query generation, relevant document retrieval, and response generation. Performance was evaluated using RDASS, ROUGE, BLEU, and METEOR metrics, demonstrating that the proposed chatbot system can deliver concise and contextually relevant summaries as responses.

Keywords

chatbot, documnet summary, small large language model, retrieval-augmented generation, query

* 국립금오공과대학교 컴퓨터공학과 학사과정
- ORCID: https://orcid.org/0009-0003-2798-250X
** 국립금오공과대학교 컴퓨터·AI융합공학과 박사과정
- ORCID: http://orcid.org/0000-0003-2092-2769
*** 국립금오공과대학교 컴퓨터공학과 교수(교신저자)
- ORCID: http://orcid.org/0000-0002-7722-6759

· Received: Oct. 14, 2024, Revised: Nov. 12, 2024, Accepted: Nov. 15, 2024

· Corresponding Author: Sungyoung Kim
Dept. of Computer Engineering, Kumoh National Institute of Technology, 61 Daehak-ro (yangho-dong), Gumi, Gyeongbuk, [39177] Korea
Tel.: +82-54-478-7530, Email: sykim@kumoh.ac.kr

Ⅰ. 서 론

다양한 산업 분야에서 챗봇 기술은 단순한 대화형 인터페이스 이상의 다양한 업무를 자동화하며, 고객 경험을 개선하는 중요한 수단으로 부상하고 있다. 챗봇은 24시간 운영 가능성, 대규모 고객 요청 처리 능력 등을 통해 기업의 고객 서비스 효율성을 높이고, 운영 비용을 절감하는 데 기여하고 있다[1-2]. 하지만, ChatGPT[3], Gemini[4], Claude[5]와 같은 대형 언어 모델을 기반으로 한 챗봇은 높은 운영 비용이라는 한계를 지니고 있다. 이러한 문제를 해결하기 위해 다양한 방법으로 챗봇을 개발하는 연구가 진행되어 왔다[6-8].

sLLM(Small Large Language Model)은 대형 언어 모델(LLM)에 비해 상대적으로 작은 크기의 매개변수를 가지는 언어 모델을 의미한다. sLLM은 메모리와 연산 자원을 절약하면서도 특정 작업에 맞게 최적화되어 효율적인 성능을 발휘할 수 있다. sLLM은 주로 특정 도메인 또는 작업에 맞춰 파인튜닝되어 사용되며, 비용 효율성이 뛰어나다. 또한, 양자화 기법을 활용해 성능을 유지하면서도 메모리 사용을 더욱 줄일 수 있다. 따라서 이러한 소규모 모델은 자원 절감이 중요한 상황에서 최적의 솔루션을 제공할 수 있으며, 특히 챗봇 시스템 개발에서 중요한 역할을 할 수 있다.

본 논문에서는 소규모 대형 언어 모델(sLLM)을 활용한 챗봇 시스템을 제안한다. 이는 기존 대형 언어 모델 기반 챗봇의 한계를 극복하는 동시에 비용 효율적이라는 점에서 유의미한 접근이라 할 수 있다. 제안하는 연구는 여러 단계의 과정을 통해 구현되며, 비용측면에서 경제적이면서도 정확한 답변을 제공할 수 있다는 장점이 있다.

2절에서는 챗봇과 관련한 연구를 소개하고 3절에서는 sLLM을 활용한 챗봇 시스템의 구현 방안을 논의한다. 4절에서는 제안한 시스템에 대한 실험과 결과에 대한 분석을 진행하고, 5절에서는 결론과 추후 연구에 대하여 소개한다.

Ⅱ. 관련 연구

2.1 질문-응답 유사도 기반 챗봇

질문-응답 유사도 기반 챗봇은 대규모 질문-쌍 데이터베이스에서 사용자가 입력한 질문과 가장 유사한 질문을 찾아 그에 맞는 답변을 제공하는 방식이다. 이를 위해 코사인 유사도와 같은 텍스트 유사도 계산 알고리즘을 사용하여 정확한 질문 매칭이 없을 때도 유사한 질문을 기반으로 적절한 답변제공한다. 코사인 유사도는 주로 BERT[9]와 같은 텍스트 임베딩 모델을 통해 계산하며, 해당 모델들은 자연어 이해 능력이 높아 코사인 유사도 계산의 정확성을 높이는데 기여한다.

해당 방식은 고객 서비스, 헬스케어, 금융 등에서 널리 활용되고 있으며 고객의 자주 묻는 질문(FAQ)을 자동으로 처리하는데 효과적이다. 아마존[10]은 이러한 방식의 챗봇을 통해 제품 배송, 반품 절차 등 자주 묻는 질문에 신속하게 대응한다. 또한, Azure AI[11]에서도 유사도 기반 챗봇을 통해 고객 문의에 대한 응답을 자동화하여 제공하고 있다.

2.2 사전 정의된 질의 - 답변 기반 챗봇

사전 정의된 질의-답변 기반 챗봇은 간단한 절차나 정보 제공에 적합하며, 사용자가 선택한 질의에 따라 고정된 답변을 제공하는 방식이다. 해당 방식은 예약 시스템, 주문 처리, 결제 안내 등에 널리 사용되며, 명확한 응답이 요구되는 작업에 적합하다. 도미노 피자[12]의 경우, 주문 과정에서 사용자가 메뉴를 선택하면 이에 따른 고정된 응답을 제공하는 방식의 챗봇을 사용한다. 해당 챗봇은 주문 내역 확인, 배달 시간 안내 등 명확한 정보를 실시간으로 제공하며, 고객의 편의성을 높이고 있다. 또한, 델타 항공[13]은 항공편 예약 및 변경과 관련된 정보를 사전에 정의된 응답을 통해 제공하여 고객 경험을 개선하고 있다.

2.3 대규모 언어 모델 기반 챗봇

대규모 언어 모델(LLM, Large Language Model)은 수십억에서 수천억 개의 파라미터를 기반으로 훈련된 자연어 처리 모델로, 자연스러운 대화 생성을 비롯해 텍스트 생성, 요약, 번역 등 다양한 작업이 가능하다.

대규모 언어 모델을 활용한 챗봇 연구로는 DialoGPT[14], BlenderBot[15], LaMDA[16] 등이 있으며 챗봇의 기반이 되는 언어 모델에 중점을 둔 강화 학습과 인간의 피드백을 결합한 RLHF (Reinforcement Learning with Human Feedback)와 같은 강화학습 기반의 방법[17], RAG(Retrieval Augmented Generation)[18] 기법을 활용한 검색 증강형 파인 튜닝 방법인 RAFT(Retrieval Augmented FineTuning)[19] 등 여러 파인 튜닝 접근법과 같이 다양한 방식을 통해 모델의 성능을 높이는 연구가 진행 중에 있다.

대규모 언어 모델을 특정 작업에 맞게 전체 파인 튜닝하는 방식은 높은 계산 비용과 메모리 자원을 필요로 한다. 이러한 제한점을 극복하기 위해 파라미터 효율적 파인 튜닝(PEFT, Parameter-Efficient Fine-Tuning)기법이 제안되었다. PEFT는 전체 파라미터를 갱신하는 대신, 소수의 추가 파라미터만 학습함으로써 효율적인 파인 튜닝을 가능하게 하였다. PEFT 기법의 대표적인 방법인 LoRA(Low-Rank Adaptation)[20]은 모델의 가중치 행렬을 저랭크 근사로 분해하여 학습해야 할 파라미터 수를 크게 줄이는 방식이다. 이를 통해 메모리 사용량과 계산 비용을 절감함으로써 기존 전체 파인 튜닝의 한계를 극복하였다.

기존 대규모 언어 모델 기반 챗봇의 경우 대부분 프롬프트를 통해 상황을 설정하고, 사용자의 입력으로부터 프롬프트를 바탕으로 답변을 출력하는 형식의 챗봇이었다. 본 논문은 기존과는 다른 각 역할을 지닌 여러 모델을 활용하여 답변 생성을 위한 총 4단계를 거치는 챗봇 시스템을 제안한다.

III. 제안하는 챗봇 시스템

본 논문에서 제안하는 챗봇 시스템의 전체적인 구조는 언어 모델 파인 튜닝, 검색 쿼리 생성, 유사 답변 탐색, 그리고 답변 생성의 4단계로 구성된다. 언어 모델의 경우 소규모 대형 언어 모델(sLLM)을 대상으로 사용자 요구에 맞는 모델을 구축하고, 사용자 입력을 분석해 검색 쿼리를 생성한다. 이후, RAG[18]을 활용하여 관련 데이터를 임베딩하고, 생성된 쿼리를 바탕으로 유사한 정보를 검색한다. 마지막으로, 검색된 데이터를 요약해 사용자에게 정확하고 신뢰성 있는 답변을 제공한다. 그림 1은 제안 방법의 전체적인 구조와 순서를 나타낸다.

3.1 언어 모델 파인 튜닝

본 연구에서는 sLLM을 기반으로 한 파인튜닝을 수행하였으며 이 과정에서 QLoRA[21] 기법을 적용한다. QLoRA[21]는 LoRA[20]에 양자화를 결합하여 메모리 사용량을 더욱 줄이고, 학습 속도를 향상시키는 장점을 지닌다. 이를 통해 소규모 대형 언어 모델(sLLM)을 효율적으로 학습할 수 있도록 설계하여 자원이 제한된 환경에서도 대규모 언어 모델의 성능을 최적화할 수 있다.

3.2 검색 쿼리 생성

검색 쿼리 생성 단계에서는 google/gemma-2-2b-it [22] 모델을 활용하여 사용자의 입력을 분석하고 적절한 검색 쿼리를 생성한다. 이 과정은 모델이 사용자의 질문을 정확히 이해하여, 데이터베이스에서 관련 정보를 효과적으로 검색하는 데 중요한 역할을 한다. 검색 쿼리 생성의 주요 목표는 사용자의 입력이 빠르고 정확하게 관련 데이터를 검색하여 답변을 제공할 수 있도록 돕는 것이다. 이를 위해 모델은 자연어 처리 기술을 통해 입력된 질문의 핵심 요소를 추출하고, 해당 정보를 검색 가능한 형태로 변환한다.

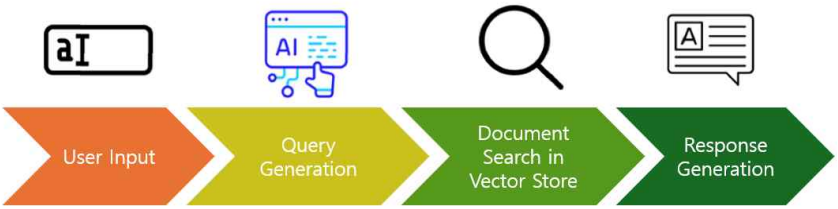


그림 1. 제안하는 챗봇 시스템 흐름도
Fig. 1. Flowchart of proposed chatbot system

생성된 검색 쿼리는 RAG[18] 단계에서 활용되어 데이터베이스 내 적합한 자료를 검색하며, 검색된 정보는 이후 응답 생성에 반영된다. 이 과정을 통해 사용자는 보다 구체적이고 관련성 높은 답변을 제공받을 수 있다.

3.3 유사 답변 탐색

RAG는 언어 모델이 외부 지식을 통합하여 더 풍부하고 정확한 응답을 생성하는 방식이다. 이는 대형 언어 모델의 생성 능력을 바탕으로 정보 검색 시스템의 지식 획득 능력을 결합하여 모델이 자체 파라미터에 지닌 지식의 한계를 극복하도록 돕는 역할을 한다. 사용자가 입력한 질문에 대하여 RAG는 해당 데이터베이스에서 관련 자료를 검색하고, 이를 하나의 컨텍스트로 활용하여 응답을 생성하는 방식이다. 이러한 방식은 추가적인 파인 튜닝이 필요 없고, 최신 정보나 훈련 데이터에 포함되지 않은 지식에 대해 접근할 수 있다. 이는 모델의 지식이 쉽게 최신 상태로 유지할 수 있으며, 언어 모델이 생성하는 응답의 정확성과 신뢰성을 향상시킨다는 장점이 있다.

RAG는 특히 챗봇 시스템에서 매우 중요한 역할을 하며, 검색된 정보를 바탕으로 답변을 생성하는 과정이다. 본 시스템에서는 FAISS[23] 벡터 스토어를 이용해 문서들을 임베딩하고, 임베딩 모델로 BAAI/bge-m3[24]를 활용한다. 3.2에서 생성된 검색 쿼리를 입력 데이터로 사용하여, 해당 벡터 스토어에 입력되어 데이터를 검색한 후 google/gemma-2-9b-it[25] 모델을 이용해 최종 답변을 생성한다.

3.3.1 문서 임베딩 생성 및 벡터 스토어 구축

본 시스템에서 챗봇은 대용량 문서 데이터를 사용해 답변을 생성한다. 먼저, PDF파일에서 문서를 추출한 후, LangChain[26]의 HuggingFaceEmbeddings를 이용해 각 문서의 임베딩 데이터를 생성한다. 생성된 임베딩 데이터는 FAISS 벡터 스토어에 저장된다. 해당 벡터 스토어는 검색 쿼리와 유사도를 계산해 빠르게 관련 문서를 검색하는데 사용된다. FAISS 벡터 스토어는 고속 유사도 검색을 위한 라이브러리로, 특히 대규모 문서 컬렉션에서 효율적인 검색을 가능하게 한다.

LangChain[26]은 다양한 자연어 처리 도구와 프레임워크를 연결하여 대규모 언어 모델 작업을 단순화하는 라이브러리로, 임베딩 생성 및 문서 검색 과정에서 효율성을 높인다.

3.3.2. 문서 검색

3.2에서 생성된 검색 쿼리는 FAISS 벡터 스토어에 입력되어, 코사인 유사도 계산을 통해 유사한 내용을 가진 문서를 검색한다. 이때 사용되는 k값은 K-최근접 이웃(K-Nearest Neighbor) 알고리즘에서 사용하는 값으로, 쿼리와 가장 유사한 k개의 문서를 반환하도록 설정하는 파라미터이다. 따라서, 사용자는 원하는 k값에 따라 검색 결과의 문서 수를 조정할 수 있다. 검색 결과는 상위 k개의 문서로 반환되며, 해당 문서들은 사용자의 질문에 대한 최적의 답변을 생성하는데 활용된다. 코사인 유사도는 식 (1)과 같이 계산한다.

$$CosineSimilarity(A,B) = \frac{A \circ B}{\| A \| \| B \|} \qquad (1)$$

여기서 A와 B는 각각 검색 쿼리 벡터와 선택된 문서 벡터이며, 두 벡터의 내적을 벡터 크기의 곱으로 나눈 값이 유사도를 나타낸다. 이 유사도를 바탕으로 가장 관련성이 높은 문서들이 검색되고, 최종 답변 생성에 사용된다. 해당 방식은 빠르고 효율적인 문서 검색을 가능하게 하며, 검색된 정보는 사용자의 질문에 대해 보다 구체적이고 신뢰성 있는 답변을 제공하는데 기여한다.

3.4 최종 답변 생성

최종 답변 생성 단계는 FAISS 벡터 스토어에서 검색된 문서와 생성형 언어 모델을 이용해 사용자에게 답변을 생성하여 제공하는 중요한 과정이다. 이 단계에서는 3.1에서 파인 튜닝된 google/gemma-2-9b-it 모델을 활용하여, 대규모 문서에서 검색된 내용을 사용자의 입력을 바탕으로 요약하고, 이를 사용자에게 제공할 수 있는 정제된 형태로 응답을 생성한다.

RAG 시스템은 검색된 문서와 그에 대한 답변을 생성하지만, 이러한 결과물은 종종 길고 복잡한 경우가 많다. 따라서 최종 응답 생성 단계에서는 불필요한 정보를 제거하고, 사용자 질문에 맞는 핵심 정보만을 담은 답변을 제공한다.

검색된 문서의 중요한 내용을 기반으로 간결하고 핵심적인 정보를 사용자에게 전달한다. 이 과정을 통해 RAG는 보다 정확하고 신뢰성 있는 답변을 제공하여, 사용자 경험을 향상시킬 수 있다.

Ⅳ. 결과 및 성능 평가

sLLM의 미세조정은 AIHUB의 문서 요약 데이터[27]를 사용하였으며 문서를 요약할 수 있는 능력을 학습시키기 위해 context: response쌍으로 데이터를 전처리하여 학습 데이터로 사용하였다. 다국어 데이터에 대해 대량의 파라미터로 훈련되어 한국어 성능이 뛰어나 별도의 한국어 파인튜닝이 필요 없을 것으로 생각되는 google/gemma-2-9b-it[22] 모델을 채택하였다. 총 48만개의 context:response쌍을 train: validation 데이터셋의 비율을 8:2로 나누어 데이터를 구축하였다. 실험 환경은 GeForce RTX 3090 두 대를 사용해 Ubuntu 18.04.5 LTS 운영체제를 기반으로 한다. 모델 학습은 학습률 0.004로 설정하였으며, 3 epochs 동안 약 60시간 동안 학습을 진행하였다.

평가 데이터로는 AIHUB 문서 요약 데이터셋 중 학습에 사용하지 않았던 신문 학습 데이터의 일부를 사용한다. 평가를 위해 최근 성능 지표로 활용되는 RDASS[28]와 해당 분야에서 자주 사용하는 성능 지표인 ROUGE-U, ROUGE-SU, BLEU, METEOR을 사용하여 제안 방법의 성능을 평가한다.

RDASS는 임베딩 모델을 사용하여 텍스트 간의 의미적 유사성을 측정하는 지표로, 요약의 의미적 일관성을 정밀하게 평가한다. RDASS 점수가 높을수록 요약과 원본 간의 의미적 일관성이 높다는 의미이다. ROUGE-U와 ROUGE- SU는 요약에서의 단어 및 n-그램 일치도를 바탕으로 정확성을 평가한다. ROUGE-U 점수가 높을수록 원본 텍스트의 중요한 단어들이 잘 보존되었다는 의미이다. 또한, ROUGE-SU 점수가 높을수록 요약에서 중요한 어구들이 잘 보존됨을 나타낸다. BLEU는 생성된 문장과 원본 문장의 일치 정도를 측정하는 지표이다. BLEU 점수가 높을수록 생성된 문장과 원본 문장이 높은 유사성을 가진다는 것을 의미한다. 마지막으로 METEOR는 동의어와 어형 변화를 고려해 생성된 문장의 의미적 유사성을 평가하는 지표로, 높을수록 의미와 형태가 일치한다는 의미로 해석할 수 있다.

표 1은 요약을 위한 파인튜닝의 효과를 나타내기 위한 실험 결과이다. 파인튜닝을 거친 Gemma, Llama-3.1[29], Qwen2[30] 모델들은 파인튜닝을 하지 않은 버전과 비교했을 때, RDASS 점수가 각각 2.4%, 2.5%, 1% 정도 낮아지는 경향을 보인다. 하지만, 나머지 네 가지 성능 지표에서는 모두 개선된 결과를 나타낸다. 구체적으로, ROUGE-U 점수는 각각 18%, 4.9%, 11.6% 상승하였고, ROUGE-SU 점수는 13.4%, 5.6%, 5.3% 증가하였다. BLEU 점수는 12.1%, 3.4%, 5.6%씩 향상되었으며, METEOR 점수는 15%, 0.3% 상승하고, Qwen2의 경우는 0.5% 하락된 결과를 보였다.

이를 통해 다른 모델들과 비교했을 때 Gemma 모델을 파인튜닝한 경우 가장 좋은 성능을 보인다는 것을 알 수 있다. 비록 RDASS에서 Gemma는 0.869로, Llama-3.1의 파인튜닝 모델인 0.886보다 낮은 성능을 보이지만, 나머지 지표에서는 모두 우수한 결과를 나타낸다. 특히, ROUGE-U에서 0.408, ROUGE-SU에서 0.178, BLEU에서 0.173, METEOR에서 0.389로 모든 비교에서 가장 높은 성능을 달성했음을 확인할 수 있다.

이를 종합해보면, RDASS에서의 소폭 성능 저하에도 불구하고, 다른 중요한 지표에서 성능이 크게 향상된 점을 감안하면, Gemma 모델을 한국어로 파인튜닝하는 것이 매우 효과적임을 알 수 있다. 특히 ROUGE 지표의 상승은 Gemma 모델이 정보의 정확성 측면에서 우수한 능력을 보유하고 있음을 시사한다. 또한, BLEU와 METEOR의 향상은 텍스트의 유사성에 있어서도 탁월한 결과를 보여준다는 의미로 해석할 수 있다. 따라서, 텍스트 생성 및 이해 측면에서 Gemma 모델이 가장 뛰어난 성능을 보여주며, 한국어 환경에서 높은 활용 가능성을 보인다.

검증을 위해 요약 학습 데이터와는 다른 데이터를 사용하여 추가 실험을 진행한다.

표 1. sLLM의 파인튜닝 전후에 따른 성능 평가
Table 1. Performance evaluation before and after fine-tuning of sLLM

Metrics	Fine-tuning	RDASS	ROUGE-U	ROUGE-SU	BLEU	METEOR
Gemma-2-9b-it[22]	×	0.885	0.228	0.044	0.052	0.239
	○	0.869	**0.408**	**0.178**	**0.173**	**0.389**
Llama-3.1-8B-Instruct[29]	×	**0.886**	0.233	0.043	0.070	0.261
	○	0.861	0.284	0.099	0.104	0.264
Qwen2-7B-Instruct[30]	×	0.879	0.130	0.008	0.025	0.266
	○	0.869	0.246	0.061	0.081	0.261

이는 학습 데이터가 단순 요약 목적에 맞춰져 있어 실제 사용 사례에서 챗봇의 성능을 평가하는 데 한계가 있기 때문이다. 요약 학습 데이터는 문서에서 중요한 정보를 추출하는 것에 중점을 두기 때문에, 챗봇이 실제 데이터를 바탕으로 다양한 질문에 대해 의미 있는 응답을 생성할 수 있는지를 평가하기에는 적합하지 않다. 따라서 학사 규정집과 같은 실제 데이터를 활용하여, 챗봇이 실제 상황에서 다양한 질문을 처리하고 신뢰성 있는 답변을 제공할 수 있는지를 검증하는 것이 중요하다. 그림 2는 제안된 방법을 국립금오공과대학교 학사 규정집(98페이지, 878,045개의 단어)을 사용하여 실험한 응답의 예시를 보인다.

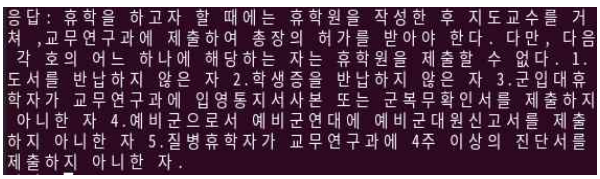


그림 2. 챗봇 시스템 테스트 결과
Fig. 2. Test result of chatbot system

표 2는 FAISS 벡터 스토어를 이용한 RAG의 성능을 포함하여 챗봇의 성능을 비교하기 위해 “휴학 처리에 대해 알려주세요”라는 동일한 질문에 대해 다양한 k로 설정하여 챗봇이 생성한 답변을 제시한다. 챗봇의 답변을 분석해보면, k가 작을수록 챗봇이 휴학 처리와 같은 특정 주제에 대해 관련 문서에 접근하지 못하는 경향이 나타남을 알 수 있다. 이는 검색된 문서의 수가 적기 때문에 사용자 질문에 맞는 정보를 포함한 문서가 선택되지 않을 가능성이 높아지기 때문이다. 반면 k를 크게 설정한 경우, 관련 정보가 과도하게 많아져 요약 과정에서 핵심 정보를 정확히 추출하지 못하고 불필요한 정보가 포함되어 요약이 비효율적으로 진행되는 경향을 보인다.

표 2. k값의 변화에 따른 출력 예시
Table 2. Examples of output based on changes in k

k	Output
1	전부 전과한 자는 전입 학부(과)의 교육과정에 따르며, 전입 학부(과)전 학년 과정의 필수과목을 모두 이수하여야 하며, 재학연한은 종전의 재학연한을 통산한다. 휴학하고자 하는 자는 당해 학기 지정된 기간에 휴학원을 제출하여야 하며, 휴학을 하고자 할 때에는 휴학원을 작성한 후 지도교수를 거쳐 학부(과)장에게 제출하여야 한다.
3	전부전과한 자는 전입학부(과)의 교육과정에 따르며, 전입학부(과)전 학년 과정의 필수과목을 모두 이수하여야 하며, 전부전과한 자의 재학연한은 종전의 재학연한을 통산한다. 휴학을 하고자 하는 자는 이를 인정할 수 있는 증명서 또는 확인서를 제출하여야 하며, 휴학기간은 재학기간에 산입하지 아니한다. 휴학기간 중 군입대휴학으로 변경한 경우의 일반휴학은 군입대휴학으로 본다. 휴학기간 중이라도 총장의 허가를 받아 역학기에 복학할 수 있으며, 군입대휴학자가 개강 이후 수업일수의 4분의 1이내에 전역하는 경우에는 등록기간 내에 전역 예정증명서를 제출하고 복학할 수 있다.
5	휴학 신청시 유의사항 1번은 휴학자의 경우 관련서류를 제출하여야 휴학처리 됨을 의미하고, 2번은 휴학 후 귀향 조치 학생은 즉시 학사서비스센터(☎478-7066) 로 연락 바람을 의미하며, 3번은 휴학 신청 후 일주일내에 원스톱서비스에서 휴학처리여부 꼭 확인 바람을 의미하고, 4번은 휴학 만료시 휴학연장 또는 복학신청 미 이행시 제적 대상자가 되므로 주의 요망을 의미한다.
10	휴학원을 제출할 수 없다. 1.예비군대원신고서를 제출하지 아니한 자. (예비군은 주민등록초본 1부,군제대자는 전역증 사본 1부) 2.군제대 복학자로서 교무연구과에 전역증 사본 1부을 제출하지 아니한 자 3.질병휴학자로서 교무연구과에 진단서 사본 1부을 제출하지 아니한 자 4.군입대휴학자로서 교무연구과에 입영통지서 사본 1부을 제출하지 아니한 자 5.예비군으로서 예비군대원신고서 [별지 제59호 서식]를 제출하지 아니한 자 6.질병휴학자로서 교무연구과에 4주 이상의 진단서를 제출하지 아니한 자 7.휴학기간이 만료된 후 수업일수 4분의 1이내에 이유 없이 복학하지 아니한 자

따라서, k의 설정은 챗봇의 성능에 중요한 영향을 미치며 적절한 k를 선택하는 것이 매우 중요함을 알 수 있다. 일반적으로 k는 사용자가 질문한 주제와 관련된 충분한 정보를 확보하면서도 과도한 정보가 되지 않도록 적절한 중간 수준으로 값을 설정하는 것이 좋은 것을 알 수 있다. 따라서, k를 5로 설정하였을 때 가장 일관되게 신뢰할 수 있는 답변을 제공할 수 있다. 이는 검색된 문서의 양이 적절하여 요약이 명확하고 질문과 관련성이 높기 때문이다.

Ⅴ. 결론 및 향후 과제

본 논문에서는 sLLM을 파인 튜닝하여 요약에 성능이 높은 모델을 제작하고, 이를 검색 쿼리와 RAG를 활용하는 챗봇 시스템을 제안하였다. 요약을 위한 파인튜닝의 경우 RDASS 점수가 소폭 감소하였지만 다른 지표들의 경우 점수가 매우 크게 증가한 것을 보아, 제안 방법에서 채택하여 파인 튜닝한 Gemma는 높은 요약 성능을 지니고 있음을 알 수 있다. 특히, k를 5로 설정하였을 때 관련된 질문을 잘 요약하여 답변할 수 있음을 확인하였다. 하지만, 최대 토큰 생성 제한으로 인해 모든 관련 내용을 대답하지 못하는 문제를 해결해야만 한다.

추후 연구로 문서 검색 후 추가적인 처리를 통해 보다 효율적인 요약을 수행하고, 그 이후 답변을 생성하는 방식의 연구가 필요하다. 검색된 문서의 양이 많을 경우 이를 분할하고 중요도를 평가하여 핵심적인 내용을 우선적으로 요약하는 방식과 최대 생성 토큰 제한 문제를 해결하기 위해 요약된 결과를 단계별로 제공하는 방식을 통해 성능을 향상할 수 있을 것이다.

References

[1] S. Park, "A Study on the Impact of Chatbot Service Quality on Customer Loyalty and Satisfaction in Online Shopping", The e-Business Studies, Vol. 24, No. 3, pp. 19-28, Jun. 2023. https://doi.org/10.20462/tebs.2023.6.24.3.19.

[2] Y. Noh and K. G. Lee, "A Study on Factors Affecting User Satisfaction in Chatbot", Journal of Customer Satisfaction Management, Vol. 24, No. 4, pp. 107-124, Oct. 2022. https://doi.org/10.34183/KCSMA.24.4.7.

[3] https://chatgpt.com. [accessed: Oct. 02, 2024]

[4] https://claude.ai. [accessed: Oct. 02, 2024]

[5] https://gemini.google.com. [accessed: Oct. 02, 2024]

[6] M. Woo, J. Si, J. Jo, and S. Kim, "Development of a Chatbot Prototype for Customer Intent Classification in O2O Stores", Proc. of KIIT Conference, Jeju, Korea, pp. 442-443, Nov. 2023.

[7] J. Park, "Development of Chatbot for Dental Consultation using AI Agent", Journal of The Institute of Internet, Broadcasting and Communication, Vol. 24, No. 6, pp. 217-222, Dec. 2024. https://doi.org/10.7236/JIIBC.2024.24.6.217.

[8] Y. Jeong, et al., "Voice Recognition Chatbot System for an Aging Society: Technology Development and Customized UI/UX Design", Journal of The Institute of Internet, Broadcasting and Communication, Vol. 24, No. 4, pp. 9-14, Aug. 2024. https://doi.org/10.7236/JIIBC.2024.24.4.9.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding", arXiv preprint arXiv:1810.04805, Oct. 2018. https://doi.org/10.48550/arXiv.1810.04805.

[10] https://aws.amazon.com/ko/kendra. [accessed: Oct. 02, 2024]

[11] https://azure.microsoft.com/en-us/products/ai-services/question-answering. [accessed: Oct. 02, 2024]

[12] https://web.dominos.co.kr. [accessed: Oct. 2, 2024]

[13] https://delta.com. [accessed: Oct. 02, 2024]

[14] Y. Zhang, et al., "DIALOGPT : Large-Scale Generative Pre- training for Conversational Response Generation", Proc. of the Annual Meeting of the Association for Computational Linguistics, pp. 270-278, Jul. 2020. https://doi.org/10.18653/v1/2020.acl-demos.30

[15] S. Roller, et al., "Recipes for building an open-domain chatbot", Proceedings of the 16th Conference of the European Chapter of the

Association for Computational Linguistics, pp. 300-325, Apr. 2021. https://doi.org/10.18653/v1/2021.eacl-main.24.

[16] R. Thoppilan, et al., "LaMDA: Language Models for Dialog Applications", arXiv preprint arXiv:2201.08239, Jan. 2022. https://doi.org/10.48550/arXiv.2201.08239.

[17] L. Ouyang, et al., "Training language models to follow instructions with human feedback", Advances in Neural Information Processing systems, pp. 27730-27744, Nov. 2022.

[18] P. Lewis, et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks", NIPS'20: Proc. of the 34th International Conference on Neural Information Processing Systems, Vancouver BC Canada, pp. 9459-9471, Dec. 2020.

[19] T. Zhang, et al., "RAFT: Adapting Language Model to Domain Specific RAG", arXiv preprint arXiv:2403.10131, Mar. 2024. https://doi.org/10.48550/arXiv.2403.10131.

[20] E. Hu, et al., "LoRA: Low-Rank Adaptation of Large Language Models", International Conference on Learning Representations, 2022.

[21] T. Dettmers, M. Lewis, Y. Belkada, and L. Zettlemoyer, "QLoRA: Efficient Finetuning of Quantized LLMs", NIPS '23: Proc. of the 37th International Conference on Neural Information Processing Systems, New Orleans LA USA, pp. 10088-1015, Dec. 2023.

[22] https://huggingface.co/google/gemma-2-2b-it. [accessed: Oct. 02, 2024].

[23] https://ai.meta.com/tools/faiss. [accessed: Oct. 02, 2024]

[24] https://huggingface.co/BAAI/bge-m3. [accessed: Oct. 02, 2024]

[25] https://huggingface.co/google/gemma-2-9b-it. [accessed: Oct. 02, 2024]

[26] https://www.langchain.com. [accessed: Oct. 02, 2024]

[27] https://aihub.or.kr/aihubdata/data/view.do?currMenu=&topMenu=&aihubDataSe=data&dataSetSn=97. [accessed: Oct. 02, 2024]

[28] D. Lee, et al., "Reference and document aware semantic evaluation methods for Korean language summarization", Proc. of the International Conference on Computational Linguistics, Apr. 2020.

[29] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct. [accessed: Oct. 02, 2024]

[30] https://huggingface.co/Qwen/Qwen2-7B-Instruct. [accessed: Oct. 02, 2024]

저자소개

우 민 식 (Minsik Woo)

2020년 3월 ~ 현재 : 국립금오공과대학교 컴퓨터공학과 학사과정

관심분야 : 자연어처리, 대규모 언어 모델, 인공지능

시 종 욱 (Jongwook SI)

2020년 8월 : 국립금오공과대학교 컴퓨터공학과(공학사)

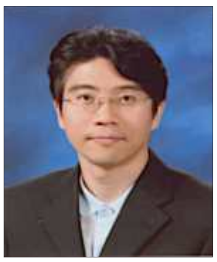
2022년 2월 : 국립금오공과대학교 컴퓨터공학과(공학석사)

2022년 3월 ~ 현재 : 국립금오공과대학교 컴퓨터·AI 융합공학과 대학원 박사과정

2023년 9월 ~ 현재 : 국립금오공과대학교 인공지능공학과 강사

관심분야 : 영상처리, 컴퓨터비전, 디지털트윈, 생성형 AI

김 성 영 (Sungyoung Kim)



1994년 2월 : 부산대학교 컴퓨터공학과(공학사)

1996년 2월 : 부산대학교 컴퓨터공학과(공학석사)

2003년 8월 : 부산대학교 컴퓨터공학과(공학박사)

2004년 ~ 현재 : 국립금오공과대학교 컴퓨터공학과 교수

관심분야 : 영상처리, 컴퓨터비전, 기계학습, 메타버스