JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 29 No. 12, pp. 29-39, December 2024
https://doi.org/10.9708/jksci.2024.29.12.029

Sentence-Based Extraction Methodology from External References to Enhance Performance in RAG

Myoungkuk Nam*, Namgyu Kim**

*PhD Student, Graduate School of Business IT, Kookmin University, Seoul, Korea
**Professor, Graduate School of Business IT, Kookmin University, Seoul, Korea

[Abstract]

The reliability of Large Language Models(LLMs) can be compromised by limitations in up-to-date information and gaps in specific domain knowledge, often leading to issues like hallucination and decreased trustworthiness. To address these challenges, Retrieval-Augmented Generation(RAG) models are increasingly utilized, allowing LLMs to provide relevant answers by leveraging external data without additional training. While much research has demonstrated the potential of RAG models to enhance the reliability of LLMs, there has been limited investigation into how best to utilize external resources to improve RAG model performance. In this study, we propose a methodology to enhance RAG model performance through sentence-based extraction of external reference materials. To evaluate our proposed methodology, we conducted a Q&A task in a specialized domain (Military English) using 5,006 abbreviations and acronyms. We compared the accuracy of an LLM and two types of RAG models (simple text extraction and sentence-based extraction), finding that our proposed approach outperformed the other models.

▸ **Key words**: Large Language Model, Retrieval-Augmented Generation, LangChain, Abbreviation, Acronym

[요 약]

LLM은 학습 데이터에 포함되지 않은 최신 정보나 특정 도메인 지식 부족으로 인해 신뢰성 문제와 환각 현상이 발생할 수 있으며, 이를 보완하기 위해 외부 자료를 활용하여 추가 학습 없이도 적절한 답변을 제공하는 RAG 모델의 활용도가 높아지고 있다. 이에 따라 RAG 모델이 LLM의 신뢰성을 향상시킬 수 있다는 많은 연구 결과들이 보고되고 있으나, 외부 자료를 어떠한 형태로 활용하면 RAG 모델의 성능을 더욱 향상시킬 수 있을지에 대한 연구는 상대적으로 미진하다. 이에 본 연구에서는 외부 참고자료의 문장화 추출을 통해 RAG 모델의 성능을 향상시킬 수 있는 방법론을 제안하고자 한다. 제안 방법론의 성능 평가를 위해 특정 도메인(군사영어)에서 사용하는 약어와 두문자어 5,006개로 LLM과 두 가지 RAG 모델(텍스트 단순 추출, 문장화 추출)에 대한 Q&A 태스크를 수행하여 정확도를 비교하였으며, 그 결과 제안 방법론인 문장화 추출 RAG 모델이 다른 모델보다 우수한 성능을 보임을 확인하였다.

▸ **주제어**: 거대 언어 모델, 검색-증강 생성, 랭체인, 약어, 두문자어

• First Author: Myoungkuk Nam, Corresponding Author: Namgyu Kim
*Myoungkuk Nam (nmk10068@kookmin.ac.kr), Graduate School of Business IT, Kookmin University
**Namgyu Kim (ngkim@kookmin.ac.kr), Graduate School of Business IT, Kookmin University
• Received: 2024. 11. 06, Revised: 2024. 12. 02, Accepted: 2024. 12. 04.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

대규모 언어 모델(Large Language Model, LLM)은 자연어처리, 이미지인식, 그리고 의사결정 등에서 인간 수준의 복잡한 작업을 수행하기 위하여 방대한 양의 파라미터와 데이터로 학습한 대규모 AI 시스템 혹은 서비스를 말한다[1]. LLM은 일반적으로 수천억개 이상의 파라미터를 가지는 대규모 인공신경망을 사용하며, GPT-3, GPT-4, BERT, LaMDA, 그리고 CLIP 등이 대표적이다[1]. LLM은 텍스트 생성 및 요약, 언어 번역, 그리고 질의응답 등 매우 다양한 분야에서 뛰어난 능력을 입증해 왔으며, 복잡한 언어 패턴을 분석하고 맥락에 따라 적절한 응답을 제공하는 능력을 갖추고 있다[2].

그러나 LLM에도 한계점이 존재하는데 바로 응답에 대한 '신뢰성'과 정보의 '최신화'를 만족하기 어렵다는 것이다. 구체적으로 LLM은 주어진 학습 데이터를 활용하여 만들어진 모델이기에 학습에 포함되지 않은 내용에 대해서는 답을 하지 못하거나 잘못된 답을 생성하는 환각현상(Hallucination)을 초래하기도 한다[3-4]. 예를 들어 LLM을 활용한 ChatGPT도 특정 연도까지의 공개 데이터만으로 학습되었기 때문에, 그 이후의 데이터나 비공개 데이터를 포함하는 특정 도메인에서의 질의응답에는 좋은 답변을 하기 어렵다.

이러한 문제점을 해결하기 위한 방법으로 미세 조정(Fine-tuning)과 검색-증강 생성(Retrieval-Augmented Generation, RAG) 방법이 활용되고 있다. 미세 조정은 사전 학습된 LLM을 특정 목적에 적합하도록 파라미터 값을 조정하거나 성능을 향상시키기 위해 특정 데이터셋에 대한 추가적인 학습을 거치는 반복적인 과정을 의미한다[5]. 그러나 미세 조정은 부가적인 학습 비용 발생, 특정 작업 또는 도메인에 대한 라벨링된 학습 데이터의 필요성, 동적 데이터 환경에서의 재훈련 빈도 증가, 그리고 학습된 모델이 블랙박스처럼 작동하여 모델의 응답을 이해하기 어렵다는 문제점을 가지고 있다[6]. 또 하나의 방법인 RAG는 검색 기반 모델을 LLM에 결합하여 응답의 질과 정확성을 향상시키는 기술로, 외부에서 가져온 정보를 활용하여 맥락에 맞는 정확한 응답을 생성한다[7]. RAG 기술은 외부 데이터 소스를 활용하여 결과물을 생성하기 때문에 응답의 정확도를 높일 뿐만 아니라, 미세 조정과 달리 기존의 LLM을 재학습시키지 않기 때문에 비용과 시간 면에서도 효율성을 제공한다. 특히 RAG는 외부 데이터의 변화를 적극적으로 반영할 수 있으므로 최신 정보를 활용할 수 있는 장점이 있다.

이처럼 RAG 모델은 사용자의 질의에 더 정확하고 적절한 답변을 제공하기 때문에, 다양한 도메인에 적용되어 LLM의 성능을 향상시키는 방법으로 활용되고 있다. 구체적으로 외부의 추가적인 정보를 활용하여 질문과 유사한 내용을 추천하는 QA시스템[8], 마이크로소프트의 Bing과 같이 웹의 검색을 통해 추가적인 정보를 제공하는 WebGPT[9], 그리고 클라우드 기술을 통해 정형 데이터 전처리 과정과 RAG기법의 벡터 임베딩(Vector Embedding) 과정을 자동화하여 손쉽게 생성형 AI의 LLM 학습을 가능토록 하는 국내 민간 공공클라우드 인프라(CSAP) 기반의 Chat서비스 관련 연구[10] 등 다양한 분야에서 활발한 활용이 이루어지고 있다. 또한 시스템 보안 유지와 공개가 제한되는 특정한 도메인의 환경에서도 LLM을 사용하기 위해 RAG 모델을 이용하는 방법도 활발하게 연구되고 있으며, 특히 우리 군은 국방혁신 4.0을 위한 AI 과학기술강군 육성을 위하여 RAG 기반 군사용 소형 언어모델(mil-sLLM)을 활용한 국방 AI플랫폼의 발전방향을 모색하고 있다[11-12].

RAG 모델의 성능에는 LLM 외부에서 어떤 자료를 참고하는지 뿐 아니라, 해당 자료들을 어떤 형태로 활용하는지가 영향을 끼치는 것으로 알려져 있다. 참고하는 자료들은 PDF, CSV, TXT, EXCEL 파일 또는 DB 등의 다양한 형식으로 존재하며, 그 내부는 텍스트, 표, 그리고 그림 등의 다양한 형태로 구성이 된다. 특히 표 형태는 행과 열의 관계를 통해서 각각의 데이터가 특별한 의미를 가지게 되는데, 단순히 텍스트만 추출한다면 이 과정에서 문서가 나타내고자 하는 문맥적 의미가 소실될 우려가 있다. 따라서 문서 내부의 문맥적 의미를 잘 추출해 낼 수 있어야 성능이 좋은 RAG 모델이라 할 수 있을 것이다.

이에 본 연구에서는 RAG 모델을 활용하면 LLM의 신뢰성을 향상시킬 수 있다는 기존 연구의 발견에서 더 나아가, 외부 참고자료의 문장화 추출을 통해 RAG 모델의 성능을 향상시킬 수 있는 방안을 제안하고자 한다. 구체적으로 본 연구에서는 특정 도메인(군사영어)에서 사용하는 약어에 대해 LLM과 두 가지 RAG 모델(텍스트 단순 추출, 문장화 추출)을 활용하여 Q&A 태스크(Task)를 수행하고, 각 모델의 정확도를 비교하여 제안 모델의 유용성을 확인하고자 한다.

본 논문의 이후 구성은 다음과 같다. 2장에서는 LLM, LangChain, RAG, 그리고 FAISS(Facebook AI Similarity Search)에 대한 기존연구와 군사영어의 특성 등을 소개하고, 3장에서는 본 연구에서 제안하는 문장형식의 참고자료를 활용하는 방법론을 소개한다. 4장에서는 제

안 방법론의 성능 평가 결과를 제시하고, 5장에서는 본 연구의 기여와 한계를 요약한다.

II. Preliminaries

1. LLM

통계를 기반으로 한 초기 언어 모델에서 출발한 LLM은 방대한 훈련 데이터와 사전 학습 방법을 활용한 트랜스포머의 등장으로 인해 인간의 언어로 표현된 텍스트를 이해하고 새로운 내용을 생성할 수 있는 모델로 변모하였다. 이를 위해 일반적으로 LLM은 수천억 개 이상의 파라미터를 활용하여 방대한 텍스트 데이터를 처리하면서 그 성능을 향상시켰다[13]. LLM은 자연어 처리(NLP) 영역에서 큰 발전을 이끌었으며, 위험 평가, 프로그래밍, 취약점 탐지, 의료 텍스트 분석, 그리고 검색 엔진 최적화 등 다양한 분야에서 활용되고 있다[14-19].

일반적으로 LLM은 최소한 다음 네 가지 주요 특징을 가져야 한다[20]. 첫째, 자연어 텍스트에 대한 깊은 이해와 해석 능력을 보여주며, 텍스트 추출 및 번역과 같은 다양한 언어 관련 작업을 수행할 수 있어야 한다. 둘째, 주어진 질문에 따라 인간처럼 텍스트를 생성할 수 있어야 하며, 문장 완성, 단락 작성, 그리고 기사 작성까지 가능해야 한다. 셋째, 도메인 전문성을 고려한 맥락 인식을 보여주어야 한다. 넷째, 문제 해결 및 의사결정에서 뛰어나야 하며, 텍스트 내 정보를 활용하여 정보 검색 및 Q&A 시스템 같은 작업에 적절한 답변을 제공해야 한다.

이러한 LLM은 학습에 사용된 소스의 공개 여부에 따라 개방형 모델과 API와 같은 기능을 통해 활용만 할 수 있는 폐쇄형 모델로 나눌 수 있으며, Table 1에 나타난 바와 같이 다양한 모델들이 있다[20].

Table 1. Comparison of popular LLMs

Model	Provider	Params	Open -Source	Tunability
GPT-4	OpenAI	1.7T	×	×
GPT-3.5 turbo	OpenAI	175B	×	×
GPT-3	OpenAI	175B	×	×
BERT	Google	340M	○	○
PaLM	Google	540B	○	○
LLaMA	Meta AI	65B	○	○

개방형 모델은 개발자가 파라미터의 조정과 추가 학습을 통해 더 나은 성능을 발휘할 수 있으므로, 내부 자료를 보호하고 특정 목적을 달성하기 위한 언어 모델이 필요한 기업이나 공공기관에서 사용하기에 좀 더 적합한 것으로 알려져 있다[21].

2. LangChain

LangChain은 LLM을 기반으로 애플리케이션을 구축하기 위한 오픈 소스 프레임워크이다[22]. LangChain은 LLM 모델이 생성하는 정보의 정확성 및 연관성을 개선하기 위해 새 프롬프트 체인을 구축하거나, 기존 템플릿을 맞춤화하여 LLM이 추가 학습이나 미세 조정 없이 새로운 데이터에 접근할 수 있도록 도와준다.

LangChain 프레임워크는 Fig. 1에서와 같이 언어모델이 어떤 순서로 어떤 동작을 취할지 결정하는 Agents 모듈, 모델에 정확한 응답을 위한 저장 메커니즘을 제공하는 Memory 모듈, LLM 애플리케이션의 다양한 단계에 연결할 수 있는 Callbacks 모듈, PDF 또는 엑셀 파일과 같은 외부 문서를 로드 및 임베딩하고 검색하는 Data Connection 모듈, 한 모듈의 출력이 다른 모듈에 입력으로 전송되는 Chains 모듈, 그리고 모델에서 보낸 응답을 애플리케이션에서 사용할 수 있도록 원하는 형식으로 해석하는 작업을 돕는 Model I/O 모듈로 구성되어 있다[23].

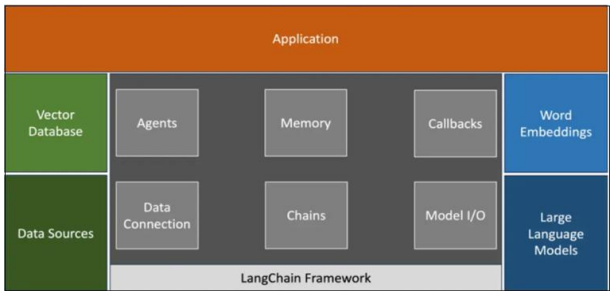


Fig. 1. LangChain Framework[23]

3. RAG

RAG는 LLM이 응답을 생성하기 전에 학습 데이터 소스 외부의 신뢰할 수 있는 지식 베이스를 참조하도록 하는 프로세스로서, 주어진 외부 데이터로부터 정보를 검색하고 활용하여 LLM이 맥락에 맞는 정확한 응답을 생성하도록 도움을 준다[7]. RAG 사용을 위한 외부 데이터는 청크(Chunk) 단위의 작은 조각으로 나뉜다. 이렇게 나누어진 조각들은 텍스트 데이터를 숫자인 벡터로 전환하는 임베딩(Embedding) 과정을 거치고, 임베딩 결과는 벡터 저장소에 저장한다[24].

LLM에 RAG를 활용하는 개념적 흐름은 Fig. 2에서와 같이 나타낼 수 있다. 사용자 질의가 발생하면, 이를 벡터

로 변환한 뒤, 외부 데이터가 저장된 벡터 데이터베이스와 매칭 작업을 거쳐 유사한 정보를 검색한다. 검색된 데이터는 사용자 질의에 추가하여 프롬프트를 보강하는데 사용되며, 이를 활용하여 LLM은 사용자 질의에 대해 더 정확한 답변을 생성할 수 있다[24].

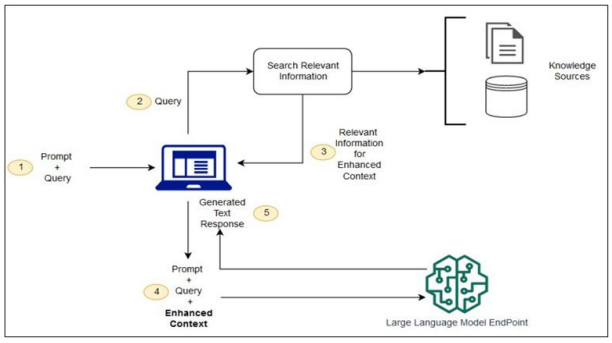


Fig. 2. Conceptual Flow of Using RAG[24]

LangChain을 사용하면 저장된 내부 문서에서 데이터를 읽고, 이를 대화형 응답으로 요약하는 애플리케이션을 구축할 수 있는 RAG 모델을 만들 수 있다. 이렇게 구축된 RAG 모델은 LLM의 할루시네이션을 줄이고, 응답의 정확도를 향상시킬 수 있다[22].

4. FAISS

FAISS는 벡터 형태로 표현된 데이터를 효율적으로 저장하고 검색하기 위해 Facebook AI Research에서 개발한 라이브러리로, 로컬 환경에서 인덱스를 생성하고 유지할 수 있는 기능을 제공한다[25]. 데이터를 외부 서버로 전송하지 않고 저장 및 검색이 가능하므로 민감한 정보가 외부 네트워크를 통해 유출될 위험을 줄일 수 있으며, 로컬 메모리와 디스크에서 직접 검색을 수행하므로 실시간 검색이 용이하다. 또한 벡터를 압축하여 저장하는 양자화 기법(Quantization)을 통해 대규모 데이터셋에서도 메모리 사용량을 최소화할 수 있으며, HNSW(Hierarchical Navigable Small World)와 같은 그래프 기반 인덱스를 활용하여 고속 검색을 지원한다. 이러한 특징으로 인해 FAISS는 웹 기반 벡터 저장소에 비해 데이터 보안성과 실시간 처리 성능에서 강점을 가지고 있다[25].

5. English for Specific Purpose and Military English

군에서 사용하는 영어는 일반영어 뿐만 아니라 군이라는 전문분야에 관련된 영어가 많다. 따라서 군에서 원활한 의사소통을 위해서는, 일반영어(English for General Purpose, EGP)뿐 아니라 군대라는 특수 조건과 환경에서 사용하는 특수목적영어(English for Specific Purpose, ESP) 역시 명확히 이해해야 한다[26]. 특히 한국군의 군 간부들은 향후 한·미간 전작권 전환을 대비하기 위해 전쟁을 주도적으로 계획하고 시행할 수 있는 능력을 기본적으로 갖추어야 하고, 이를 통해 우수한 미군을 작전통제할 수 있어야 한다. 이처럼 한미 연합방위체계에서 주도적 역할을 수행하기 위해서는, 연합작전 기획 및 지휘, 작전계획 작성, 작전수행능력, 그리고 연합전력의 전략적 전술적 운용 능력을 보유할 수 있는 군사영어의 교육 강화가 요구된다[27].

특수목적영어는 아래 Table 2와 같이 학문적 목적을 위한 영어(English for Academic Purpose, EAP)와 직업 교육을 위한 영어(English for Occupational Purpose, EOP)로 구분할 수 있다[28]. EAP는 전공 관련 과학 및 기술 영어, 의학영어, 법률영어, 그리고 금융영어 등으로 세분화할 수 있으며, EOP는 직업적인 관점에서 다루는 직무영어로서 의학영어, 비즈니스 영어, 그리고 군사영어 등의 전문직 목적 영어(English for Professional Purposes, EPP)와 직업 및 취업을 위한 직업 목적 영어(English for Vocational Purposes, EVP)로 세분화된다. 본 논문에서 다룰 군사영어는 EOP의 한 종류인 EPP의 하나로 일반영어와 명확히 구분된다.

Table 2. Category for ESP

ESP	EAP		Science & Technology English
			Medical English
			Legal English
			Financial English
	EOP	EPP	Medical English
			Business English
			Military English
		EVP	Pre-vocational English
			Vocational English

군사영어는 신속 정확한 군사작전에 사용되기 때문에, 객관적 사실을 바탕으로 문장이 매우 간결할 뿐 아니라 일반영어와 다르게 해석이 되는 경우가 많다는 특징을 갖는다. 예를 들어 군사영어에서 “Engineer Unit”은 기술부서가 아니라 공병부대를 지칭하는 의미로 사용된다. 또한 효율적인 의사전달을 위해 두문자어(Acronym)와 약어(Abbreviation)를 빈번하게 사용한다. 예를 들어 합동참모본부를 뜻하는 “Joint Chiefs of Staff”는 두문자어인 “JCS”로, 사단을 뜻하는 “Division”은 약어인 “Div.”로 기재하여 문장을 간결하게 만드는 경향이 있다. 이러한 군사영어의 특수성을 고려하지 않은 문장 번역은 원래의 의미를 왜곡시킬 우려가 있다. 예를 들어, “The Birdfarm

requests IPR on BDA of the Blue Force."라는 문장은 일반영어로 해석을 하면 "새농장이 청군의 BDA에 대한 IPR을 요청했다."로 번역될 수 있다. 하지만, 군사영어에서 Birdfarm은 항공모함을, IPR은 중간보고 또는 추진현황보고(In-Progress Report)를, BDA는 전투피해평가를, 그리고 Blue Force는 아군을 의미한다. 따라서 해당 문장은 "항공모함에서 우군 전투피해평가에 대한 중간진행보고를 요청했다"고 번역해야 한다[29].

III. The Proposed Scheme

1. Research Process

본 논문은 군사 용어의 특수성을 반영하고 있는 RAG를 활용하여 약어 또는 두문자에 대한 Full Form을 반환해 주는 시스템을 구현하였는 바, 본 장에서는 외부 데이터를 문장 형태로 참조하여 RAG 모델의 성능을 향상시키기 위한 방법론을 소개하고, 단계별 구체적인 프로세스를 설명한다. 제안 방법론의 전체적인 과정은 Fig. 3과 같다.

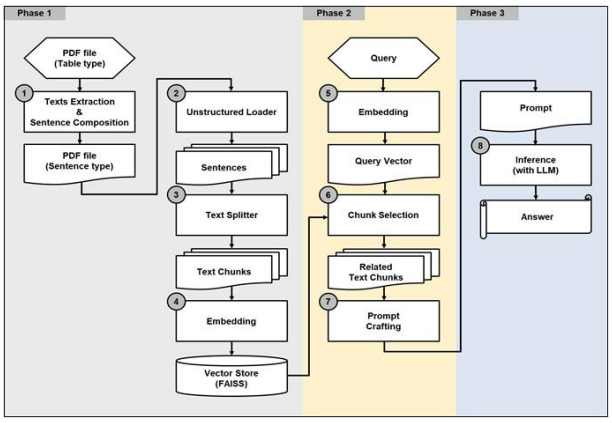


Fig. 3. Overall Research Process

먼저 Phase 1은 기존 표 형태의 외부 데이터를 문장 형태의 데이터로 변경하고(①), 문장 형태로 구성된 PDF 문서에서 텍스트를 추출하여 자연어 처리에 사용할 수 있는 형태로 만든다(②). 이후 추출된 문장들을 모델이 처리하기 용이한 적절한 크기의 청크(Chunk)로 분할한 뒤(③), 각 청크들의 텍스트 데이터를 수치 벡터로 변환하여 벡터 저장소에 저장한다(④). Phase 2에서는 질의 문장을 임베딩하여 벡터로 변환하고(⑤), 이를 벡터 저장소에 담긴 외부 데이터와 비교하여 유사한 청크들을 추출한 뒤(⑥), LLM이 답변을 생성하기 위해 필요한 적절한 프롬프트를 작성한다(⑦). 이후 Phase 3에서는 생성된 프롬프트를 LLM에 적용하여 답변을 생성한다(⑧). 각 단계에 대한 세부적 프로세스는 다음의 각 절에서 설명하며, 실제 데이터를 적용한 제안 방법론의 성능 평가는 4장에서 제시한다.

2. Sentence Composition from Table Type Data

본 절에서는 Fig. 3의 Phase 1에서 이루어지는 과정 중, 표 형태의 데이터를 문장 형태의 데이터로 재구성하는 과정을 제시한다(①). 본 논문에서 사용하는 외부 데이터인 군사용어 약어집은 Fig. 4에서 보는 바와 같이 약어와 의미가 표의 형태로 되어 있으며, 텍스트만 추출하는 경우 이 두 열의 관계성이 손실된다. 따라서 본 단계에서는 해당 두 열을 그대로 추출하는 대신, 약어와 의미 사이에 'stands for'를 추가함으로써 표에 제시된 의미가 손실없이 유지될 수 있도록 외부 데이터를 재구성하였다. 또한 문서 각 페이지의 상하 여백에 있는 불필요한 텍스트 제거 등의 전처리 역시 본 단계에서 수행하였다.

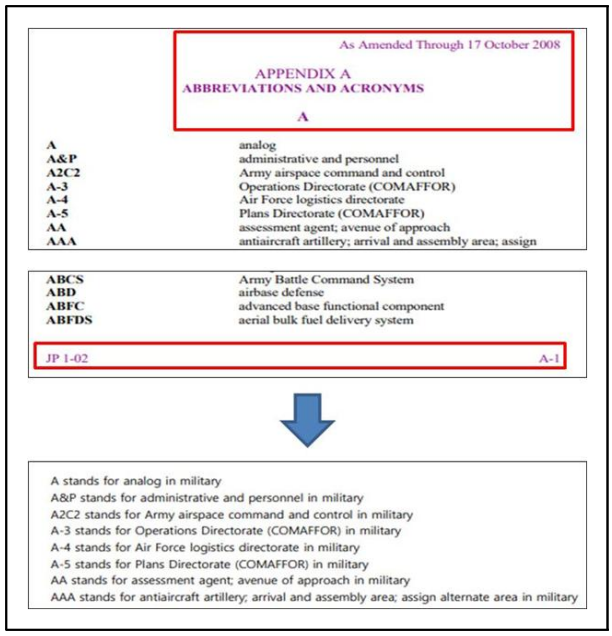


Fig. 4. Example of Sentence Composition

3. Chunking, Embedding, and Vector Store

본 절에서는, 참고하는 외부 데이터에서 텍스트를 추출하여(②), 자연어 처리가 용이한 적절한 크기로 분할하고(③), 이를 질의어와 비교할 수 있도록 벡터 저장소에 저장하는 과정을 소개한다(④). LangChain 프레임워크에서 텍스트를 추출하고 분할하는 방법은 청크 크기에 따라 균일하게 분할하는 'CharacterTextSplitter' 방법과 문장의 문맥을 유지하기 위해 문장 형태에 따라 분할하는 'RecursiveCharacterTextSplitter'방법 등이 대표적이

며, 본 논문에서는 문맥의 특성을 잘 간직할 수 있다고 알려진 후자의 방법을 사용한다. Fig. 5는 청크 분할을 보여주는 사례로, 'A-5'에 대한 설명이 전자의 방법에서는 두 개의 청크로 분할되어 의미 보전이 어렵지만, 후자의 경우는 두 번째 청크에서 완전한 문장을 이루고 있음을 확인할 수 있다.

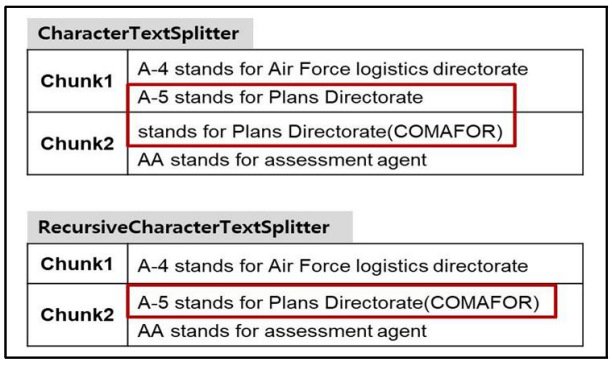


Fig. 5. Compare of TextSplitters

이렇게 분할된 청크들은 Fig. 6과 같이 임베딩 과정을 거쳐서 수치로 표현되고, 수치화된 자료들은 FAISS를 활용하여 벡터 저장소에 저장하였다. 본 연구는 내부 자료의 보호와 보안이 매우 중요한 도메인에서의 RAG 사용을 목표로 진행되므로, 벡터 저장소로 로컬 환경에서 자료의 저장과 검색이 이루어지는 FAISS를 사용하였다[25].

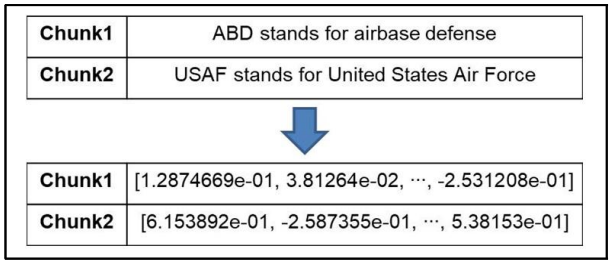


Fig. 6. Example of Embedding

4. Chunk Selection and Prompt Crafting

본 절에서는 Fig. 3의 Phase 2에서 이루어지는 과정 중, 질의문을 임베딩한 쿼리 벡터(⑤)를 벡터 저장소에 저장된 자료들과 비교하여 유사도가 높은 청크를 선택하고(⑥), 검색된 청크들의 문장들과 질의의를 포함한 프롬프트를 작성한 뒤(⑦), Phase 3에서 이를 활용하여 LLM을 통해 결과값을 생성하는 과정을 소개한다(⑧). 질문자의 쿼리가 입력되면 이를 외부 데이터를 임베딩한 것과 동일한 방식으로 수치화한다. 다음으로 Fig. 7에서 보는 바와 같이 수치화된 쿼리의 벡터값과 벡터 저장소에 저장되어 진 각 청크의 벡터값들을 비교하여 유사도가 높은 청크를 선택한다.

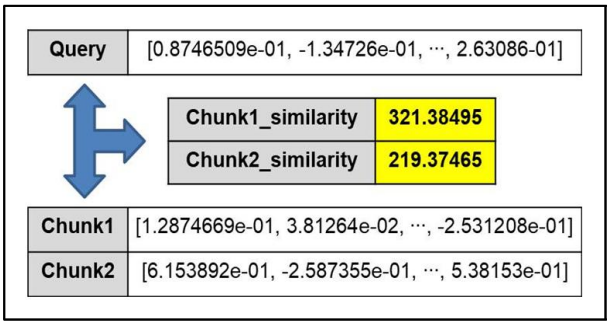


Fig. 7. Example of Chunk Similarity Score

LLM에서 프롬프트는 모델이 응답을 생성하는 데 사용하는 입력 텍스트를 의미하며, 원하는 출력 형태 및 질의자의 요구사항을 알려줌으로써 모델이 생성하는 결과에 영향을 미친다[30]. LLM은 진술형 문장보다 명령형이나 질문형의 지시 사항에 더 효과적으로 반응하므로[31], 명확하고 구체적인 지시와 예시를 사용하여 효과적인 프롬프트를 작성하는 것이 중요하다. 답변은 Fig. 8과 같이 예시가 없는 경우인 Zero-Shot, 예시가 1개인 One-shot, 그리고 2개 이상인 Few-shot Learning으로 구분하며, 본 논문에서는 참고자료의 문장형태와 유사하게 답변을 생성하기 위해 정확한 답변을 생성해야 하므로 One-shot 방법을 활용한다[9].

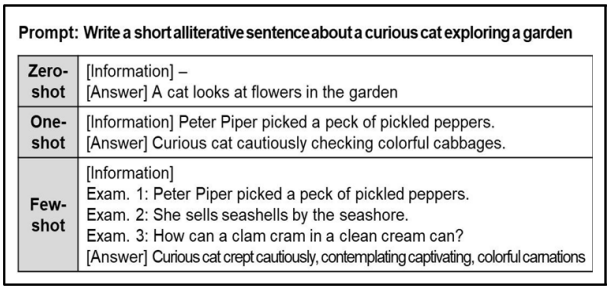


Fig. 8. Example of Zero/One/Few-shot learning

마지막으로, 최종 작성된 프롬프트를 질의로 활용하여 LLM을 통해 결과값을 얻을 수 있는데, 본 논문에서는 여러 LLM 중 로컬환경에서 활용할 수 있는 Mistral 7B 모델을 선택하였다. Mistral 7B 모델은 매개변수가 73억개임에도 불구하고 벤치마크 전 분야에서 메타의 '라마 2 13B'를 능가했으며, 매개변수 340억개인 '라마 1 34B'까지 상당 분야에서 뛰어넘는 등 뛰어난 성능을 갖춘 것으로 알려져 있다. 또한, Mistral 7B 모델은 파인튜닝 등을 통해 특

정 도메인에 부합된 모델을 만드는데 용이한 장점을 가지고 있다[32]. 위의 전체 과정을 통해 구현된 시스템의 질의 응답은 Fig. 9와 같이 정확한 답변, 가능성 있는 답변, 그리고 답변할 수 없음의 세 가지 형태를 나타낸다.

Correct Answer	**Question: What does the "NSL" stand for in military?** - NSL stands for "no-strike list."
Available Answer	**Question: What does the "AAAS" stand for in military?** - I can't provide an exact answer without additional context - It could potentially stand for different thing * American Aviation Academy and School(US Army) * Asiatic-Pacific Area Service Command(US Navy)
Can't Answer	**Question: What does the "NFESC" stand for in military?** - I cannot directly answer that question - I'm unable to provide an exact answer - I need more information to provide an accurate answer

Fig. 9. Example of Answers

IV. Experiment

1. Experiment Overview

본 장에서는 앞서 소개한 제안 방법론을 실제 데이터에 적용한 실험 결과를 소개한다. 실험에는 미 국방부 발행 군사용어 사전에서 수집한 5,006개의 약어 및 두문자어를 데이터로 사용하였다. 본 데이터 셋은 표 형태의 PDF 파일로 구성되어 있으며, 이를 'stands for'로 연결한 문장화 형태의 PDF 파일을 추가로 생성하여 두 가지 형태의 파일을 외부 데이터로 활용하였다. 실험 환경은 Python을 통해 구축하였으며, 구체적인 H/W 및 S/W 환경은 Table 3과 같다. 또한 성능 비교 실험의 전체 프로세스는 Fig. 10과 같다.

Table 3. System Environments

H/W	CPU	AMD Ryzen 7
	GPU	Radeon 780M
	RAM	32GB
S/W	OS	Windows 11
	Python	3.12.3
	LLM	Mistral 7b
	Ollama	0.3.13
	LangChain	0.2.1
	FAISS	1.8.0

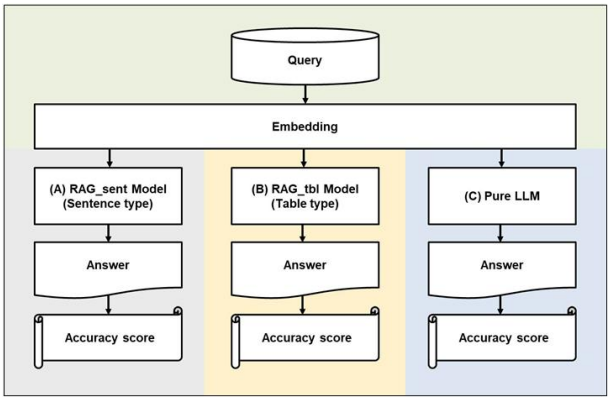


Fig. 10. Overall Process of Performance Evaluation

Fig. 10의 (A)는 본 연구에서 제안하는 방법론을 적용한 RAG_sent 모델로, 문장화된 외부 데이터를 참조하여 약어와 두문자어에 대한 Full Form을 응답하고 그 정확도를 평가한다. 한편, Fig. 10의 (B)와 (C)는 제안 방법론과의 성능 비교를 위해 수행한 실험이다. (B)는 표 형태의 외부 데이터에서 텍스트만을 추출하고 원형 그대로 참조하여 응답하는 RAG_tbl 모델이고, (C)는 참조하는 외부 데이터 없이 순수하게 Pure LLM으로부터 응답을 생성하는 모델이다.

본 논문의 실험 데이터는 Fig. 11과 같이, 군사용어의 약어와 두문자어를 질의(Question) 데이터로 구성하고, 각각의 해설은 모델들이 추론한 답변 내용의 정확성을 평가하기 위한 정답(Answer) 데이터로 사용하였다. 군사용어는 활용시 정확히 하나의 의미만을 지칭하므로, 중의어나 다의어로 구성된 약어 및 두문자어는 실험 데이터에서 배제하였다.

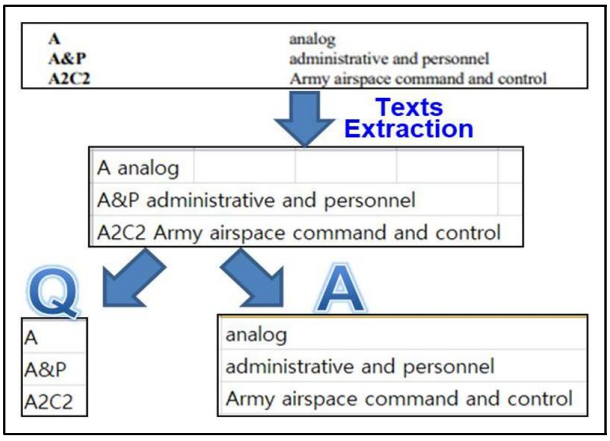
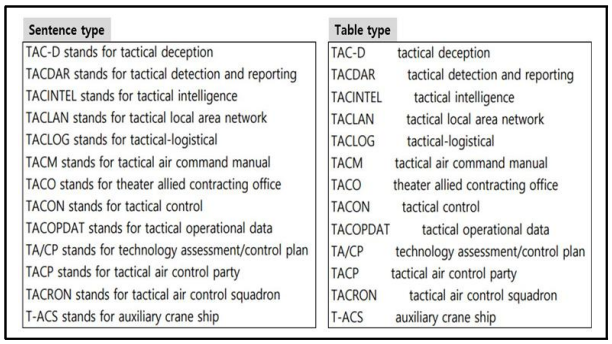


Fig. 11. Example of Q&A Data

또한, 외부 데이터는 Fig. 12와 같이 두 가지 유형으로 구성하였다. 본 논문에서 제안하는 RAG_sent 모델은 문장화된 자료(Sentence type)를 활용하고, 비교 대상 중 하나인 RAG_tbl 모델은 텍스트로만 구성된 표 형식의 자료(Table type)를 활용하였다.



Sentence type
TAC-D stands for tactical deception
TACDAR stands for tactical detection and reporting
TACINTEL stands for tactical intelligence
TACLAN stands for tactical local area network
TACLOG stands for tactical-logistical
TACM stands for tactical air command manual
TACO stands for theater allied contracting office
TACON stands for tactical control
TACOPDAT stands for tactical operational data
TA/CP stands for technology assessment/control plan
TACP stands for tactical air control party
TACRON stands for tactical air control squadron
T-ACS stands for auxiliary crane ship

Table type	
TAC-D	tactical deception
TACDAR	tactical detection and reporting
TACINTEL	tactical intelligence
TACLAN	tactical local area network
TACLOG	tactical-logistical
TACM	tactical air command manual
TACO	theater allied contracting office
TACON	tactical control
TACOPDAT	tactical operational data
TA/CP	technology assessment/control plan
TACP	tactical air control party
TACRON	tactical air control squadron
T-ACS	auxiliary crane ship

Fig. 12. Example of PDF files

2. One-shot Prompt Crafting

본 절에서는 모델에 최종 프롬프트를 적용하기 위한 One-shot 프롬프트 작성 과정을 설명한다. 하나의 질의응답 답변 예시를 One-shot 템플릿 형태로 구성하고, 외부 데이터를 활용하여 구축된 벡터 저장소 자료 중 질의어와 가장 유사한 자료를 검색한 결과와 벡터화된 질의를 이용하여 모델에 적용할 체인을 Fig. 13과 같이 구성한다.

```
Template = """
  Q: What does the acronym or abbreviation "{acronym}"
     stand for in military?
  A: "{acronym}" stands for "{definition}"  """

Prompt =PromptTemplate(input_variables=["acronym", "definition"],
          template=template)

Answer = chain({"input_documents":docs, "question":Prompt})
          *docs: retrieved chunks of Vector Store
```

Fig. 13. Example of Prompt & Chain

3. Results

본 절에서는 최종적으로 생성된 프롬프트를 모델에 적용하여 생성된 결과를 소개한다. 질의어는 임베딩 과정을 거쳐 벡터화되고, 외부 데이터가 구축되어 있는 벡터 저장소의 자료들과 유사도를 비교하여 수 개의 가장 유사도가 높은 자료들이 검색 및 추출된다. Fig. 14는 참고자료의 형태에 따라 하나의 질의에 대한 유사도가 가장 높은 상위 세 개의 청크들이 추출된 예로서, 상위 청크들이 두 가지 모델에서 서로 다르게 형성됨을 확인할 수 있다.

Question: What does the acronym or abbreviation "A&P" stand for in military?

Data type	Table	Sentence
Rank 1	A analog A&P administrative and personnel A2C2 Army airspace command and control	A stands for analog in military A&P stands for administrative and personnel in military A2C2 stands for Army airspace command and control in military
Sim_score	309.0539	240.76529
Rank 2	ACOCC air combat operations command center ACOS assistant chief of staff AAMDC US Army Air and Missile Defense Command	AA&E stands for arms, ammunition, and explosives in military AAEC stands for aeromedical evacuation control team in military AAFES stands for Army and Air Force Exchange Service in military
Sim_score	343.11017	262.425
Rank 3	AAOE arrival and assembly operations element AAOG arrival and assembly operations group AA&E arms, ammunition, and explosives	AAP stands for Allied administrative publication AAR stands for after action report
Sim_score	359.25125	285.23712

Fig. 14. Example of Question & Chunk Similarity Score

이렇게 검색된 자료 및 질의어를 활용하여 생성한 최종 프롬프트와 LLM을 통해 최종 결과를 생성해 낼 수 있다. Fig. 15는 각 모델별로 질의에 대한 답변을 나타낸 예시로, 답변은 파란색과 같이 실험 데이터의 질의에 대한 정답인 'Answer 데이터'와 일치하는 정확한 답변, 적색으로 표현된 'several other acronym such as, various things depending on, several meanings depending on'과 같이 여러 가지 가능성 있는 자료들을 활용하여 LLM이 생성해 낸 답변, 그리고 'does not offer an answer, answer cannot be given, cannot provide an answer'와 같이 응답이 불가한 답변으로 구분된다.

RAG_sent	-The acronym "AFI" in military stands for "Air Force Instruction." -"AIS" appears to be a part of several other acronyms such as ... · Automatic Identification System(maritime navigation and tracking) · Airborne Information System -The context provided does not offer an answer.
RAG_tbl	- In the context provided, "BLOS" stands for "beyond line of sight," -"ALCG" could potentially stand for various things depending on ... · AACG(Arrival Airfield Control Group) · ARL-M(Airborne Reconnaissance Low-Multifunction) -Without additional context, an answer cannot be given with certainty
Pure LLM	-"BVR" stands for Beyond Visual Range, In military context. -"CAR" can have several meanings depending on the context. · Combat Arms Rifleman - A soldier who is trained to engage the enemy · Combat Aircraft Rescue - aircraft that are used to rescue -I apologize, but I cannot provide an answer without additional context

Fig. 15. Example of Model Answers

4. Performance Evaluation

본 절에서는 제안하는 방법론인 문장화된 외부 참고자료를 활용한 RAG_sent 모델, 텍스트만을 추출하여 활용하는 RAG_tbl 모델, 그리고 RAG를 사용하지 않는 Pure LLM 모델의 성능을 비교한 결과를 소개한다. 성능 척도는 정확도를 사용하였는데, Fig. 16과 같이 Full Form과 정확히 일치하는 'Clear Answer', 그리고 일부 다른 표현이 있으나 의미가 동일한 'Unclear Answer', 이렇게 각 상황을 정답으로 간주하는 두 가지 실험을 수행하였다. 후자의 경우 전문가의 오답 검토과정을 통해 정답 여부를 평가하였다.

Abbreviation	Q-ship
Clear Answer	Decoy ship
Unclear Answer	Quasi-Ship or Queens's Ship. In the military context, a Q-ship is a decoy vessel used to lure submarines into revealing their position by simulating the appearance of a real warship.

Fig. 16. Example of Model Answers

실험 결과 Table 4 및 Fig. 17에서와 같이 제안 방법론인 RAG_sent 모델이 텍스트만을 추출하여 활용한 RAG_tbl 모델 및 단순히 LLM만을 사용한 Pure LLM 모델보다 정확도 측면에서 우수한 성능을 나타냄을 확인하였다. 이는 RAG 활용시 외부 참고자료의 형태를 문장화하여 추출함으로써 LLM 답변의 정확도를 향상시킬 수 있음을 나타낸다.

Table 4. Performance Comparision

Model	RAG_sent (A)	RAG_tbl (B)	Pure LLM (C)
Accuracy	83.4%	60.4%	14.8%

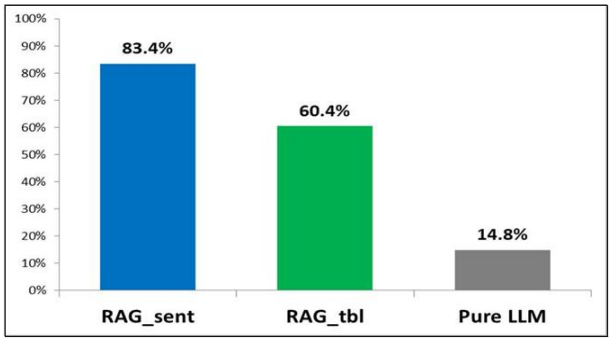


Fig. 17. Performance Comparison

V. Conclusions

최근 LLM은 다양한 분야에서 다양한 형태로 활용되어 우수한 성과를 거두고 있다. 하지만 범용 LLM이 할루시네이션 현상과 최신 정보의 적시적인 반영이 제한되는 한계를 보임에 따라, 실제 현장에서는 내부자료를 보호하면서도 정확한 정보를 제공하는 모델에 대한 수요가 급증하고 있다. 이러한 범용 LLM의 제한점을 극복하기 위해 RAG 모델을 이용하는 방법이 활발하게 연구되고 있지만, 외부 참고자료의 특성에 따른 RAG의 성능을 분석한 사례는 상대적으로 찾아보기 어렵다. 이러한 배경에서 본 연구에서는, 외부 참고자료가 내포하는 의미를 잘 보존하고 있는 문장 형태의 자료를 활용하여 RAG 모델의 성능을 향상시키는 방법론을 제시하였다. 제안 방법론의 성능 평가를 위해 특정 도메인(군사영어)에서 사용하는 약어와 두문자어 5,006개로 LLM과 두 가지 RAG 모델(텍스트 단순 추출, 문장화 추출)에 대한 Q&A 태스크를 수행하여 정확도를 비교하였으며, 그 결과 제안 방법론인 문장화 추출 RAG 모델이 다른 모델보다 우수한 성능을 보임을 확인하였다.

본 연구는 단어와 단어간의 의미가 내포되어 있는 문장을 참고자료로 활용하는 RAG 모델이 텍스트만 나열되어 있는 자료를 참고하는 기존의 RAG 모델보다 더 효과적이라는 점을 실험을 통해 입증했다는 점에서 학술적 기여를 인정받을 수 있다. 즉, 신뢰할 수 있는 정보를 효율적으로 추출 및 가공하여 활용하는 하나의 방식을 제안하였으며, 향후 여러 가지 형태의 외부 참고자료에서 효과적으로 자료를 추출하고 활용하는 다양한 후속 연구가 이어질 것으로 기대한다. 또한, 제안 방법론은 군사분야와 같이 약어 및 두문자어와 같은 특별한 형태가 매우 중요하게 활용되는 특정 도메인에서 우수한 성능을 나타냄을 확인하였으며, 이러한 성능 향상 측면에서 본 연구의 실무적 기여를 찾을 수 있을 것이다. 특히, 해당 방법론은 데이터 추출 형태에 기반을 두고 있으므로, 다양한 유형의 데이터에 대한 추출 방법의 정의를 통해 RAG 모델의 지속적 성능 향상에 기여할 수 있을 것으로 기대한다.

본 연구에서 제안 방법론의 성능 평가는 군사용어의 약어 및 두문자어라는 특수한 분야에 대해서 이루어졌다. 하지만 문장화를 통한 성능 향상이라는 제안 방법의 안정성과 견고성을 확인하기 위해서는, 다양한 조건과 환경에서 성능을 평가하는 후속 연구가 수행될 필요가 있다. 한편 본 연구에서 군사용어는 문장에서 정확하게 하나의 의미만을 갖는 것으로 가정하여 중의어 및 다의어 데이터를 분석에서 배제하였다. 하지만 실제 업무에서는 동일한 향후 연구에서는 문장 속에서 맥락을 파악하여 각 용어의 정확한 의미를 찾아내는 방식으로 연구가 확장될 필요가 있다. 또한 본 연구에서는 영어 데이터만을 활용하여 실험을 수행하였다. 하지만 향후 연구에서는 한국어를 대상으로 한 분석을 통해, 분석 대상 언어의 차이에 따라 제안 방법론의 성능 향상이 상이하게 나타나는지 여부를 확인할 필요가 있다.

REFERENCES

[1] S. P. Shin, "The Concept and Standardization Trends of Foundation Models in LLM," Information and Communications

Magazine(KICS), Vol. 40, No. 6, pp. 12-21, May. 2023.

[2] B. Ch. Das, M. H. Amini, and Y. Wu, "Security and Privacy Challenges of Large Language Models: A Survey", 30 Jan. 2024. arXiv:2402.00888

[3] N. Kandpal, H. Deng, A. Roberts, E. Wallace, and C. Raffel, "Large Language Models Struggle to Learn Long-tail Knowledge," arXiv preprint arXiv: 2211.08411, Jul. 2023. DOI: 10.48550/arXiv.2211.08411

[4] Y. Zhang, Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, X. Huang, E. Zhao, Y. Zhang, Y. Chen, L. Wang, A. T. Luu, W. Bi, F. Shi, and S. Shi. "Siren's Song in The AI Ocean: A Survey on Hallucination in Large Language Models," arXiv preprint arXiv: 2309.01219, Sep. 2023. DOI: 10.48550/arXiv. 2309.01219

[5] Wikipedia, "Fine-Tuning (Deep Learning)," https://en.wikipedia.org/wiki/Fine-tuning (deep_learning), 2023.

[6] T. B. Brown et al., "Language Models are Few-Shot Learners," arXiv:2005.14165, May. 2020. DOI: 10.48550/arXiv.2005.14165

[7] K. Shuster, and Kurt, et al., "Retrieval Augmentation Reduces Hallucination in Conversation," 2021. arXiv preprint arXiv:2104.07567

[8] 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.

[9] C. S. Jeong, "Generative AI Service Implementation Using LLM Application Architecture: based on RAG Model and LangChain Framework," Journal of Intelligent Informagion System, Vol. 29, No. 4, pp. 129-164, Dec. 2023.

[10] H. S. Kim, and J. J. Lee, "Development of Chat Web Service Functions for Administrative and Public Institutions in The Cloud Environment Using the Implementation of RAG Technology Data Learning Automation," KCIS Winter Confrrence 2024, pp. 416-417, Jan. 2024.

[11] M. Ch. Kim, Ch. W. Lee, and S. H. Yeom, "A Comparative Study of Performance between Large Language Model(LLM) of Open SourceModel and Closed Source Model Implementations for the Retrieval-Augmented Generation (RAG) System," KCIS Summer Confrrence 2024, pp. 1,354-1,355, Jun. 2024.

[12] J. H. Jeon, K. B. Kim, J. S. Kim, and S. T. Park, "Research on The Development of a Defense AI Platform Using a RAG-based mil-sLLM," Korea Computer Congress 2024, pp. 44-46, Jun. 2024.

[13] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, and Z. Dong, et al., "A Survey of Large Language Models," 2023. arXiv preprint arXiv:2303.18223

[14] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, and E. Hovy, "A Survey of Data Augmentation Approaches for NLP," 2021. arXiv preprint arXiv:2105.03075

[15] C. Novelli, F. Casolari, A. Rotolo, M. Taddeo, and L. Floridi, "Taking AI Risks Seriously: A New Assessment Model for The AI Act," AI & Society, pp. 1–5, 2023.

[16] Y. Cai, S. Mao, W. Wu, Z. Wang, Y. Liang, T. Ge, C. Wu, W. You, T. Song, and Y. Xia, et al., "Low-code LLM: Visual Programming over LLMs," 2023. arXiv preprint arXiv:2304.08103.

[17] R. Jain, N. Gervasoni, M. Ndhlovu, and S. Rawat, "A Code Centric Evaluation of C/C++ Vulnerability Datasets for Deep Learning based Vulnerability Detection Techniques," Proceedings of The 16th Innovations in Software Engineering Conference, pp. 1–10, 2023.

[18] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, "Large Language Models in Medicine," Nature Medicine Vol. 29, No. 8, pp. 1930–1940, 2023.

[19] B. B. Arcila, "Is It a Platform? Is It a Search Engine? It's ChatGPT! The European Liability Regime for Large Language Models," J. Free Speech L. 3, 455, 2023.

[20] J. Yang, H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, B. Yin, and X. Hu, "Harnessing The Power of LLMs in Practice: A Survey on ChatGPT and Beyond," 2023. arXiv preprint arXiv:2304.13712.

[21] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang, "A Survey on Large Language Model (LLM) Security and Privacy: The Good, The Bad, and The Ugly," High-Confidence Computing 4, 100211, Mar. 2024.

[22] AWS, What is LangChain?, https://aws.amazon.com/ko/what-is/langchain

[23] Janakiram MSV, LangChain Guide for Developers, https://www.ciokorea.com/column/305341#csidx973f1264e8a2e758d10e50c3f1541b5

[24] AWS, What is RAG?, https://aws.amazon.com/ko/what-is/retrieval-augmented-generation/

[25] ProjectPro, FAISS Vector Database: A High-Performance AI Similarity Search, https://www.projectpro.io/article/faiss-vector-database/1009

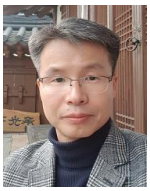
[26] H. Lee, and N. Kim, "Exploring the Relationship among Language Learning Strategies, Motivation, and Achievement in ESP Learning: A Case of Learning Military English," Korean Journal of Military Art and Science Vol. 6, No. 2, pp. 283-302, Jun. 2020. DOI: http://doi.org/10.31066/kjmas.2020.76.2.012

[27] G. Song, "A Study on the Improvement of ELT Materials to Enhance the Military English Communication Skills of the Korean Armed Forces," Korean Journal of Military Art and Science, Vol. 78, No. 1, Feb. 2022. DOI: http://doi.org/10.31066/kjmas.2022.78.1.014

[28] D. Evans, and St. John, "Development in English for Specific Purpose: A Multi-disciplinary Approach,". Cambridge, Cambridge Univ. Press, 1998.

[29] J. S. Lee, “Interpretation Officer Guide” Chaek-maru, 2011.
[30] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. S. Smith, and D. C. Schmidt, “A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT,” Feb. 2023. DOI: 10.48550/arXiv.2302.11382
[31] A. Leidinger, R. V. Rooij, and Ekaterina Shutova, “The Language of Prompting: What Linguistic Properties Make a Prompt Successful?,” Association for Computational Linguistics: EMNLP2023, Singapore, Nov. 2023. DOI: 10.18653/v1/2023.findings-emnlp.618
[32] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Banford, D. S. Chaplot, D. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saullnier, L. R. Lavaud, M. A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, “Mistral 7B,” Oct. 2023. DOI: 10.48550/arXiv.2310.06825

Authors



Myoungkuk Nam received the B.S. degree in Information Engineering from Korea Military Academy in 1997, M.S. degree in Computer Engineering from Korea National Defense University in 2009, and currently enrolled in Graduate School of Business IT, Kookmin University. He is interested in natural language processing, text mining, and RAG

Namgyu Kim received the B.S. in Computer Engineering from Seoul National University in 1998, M.S. and Ph.D. degrees in Management Engineering from KAIST, Korea, in 2000 and 2007, respectively. Dr. Kim joined the faculty of the School of Management Information Systems at Kookmin University, Seoul, Korea, in 2007. He served as the Dean of the Graduate School of Business IT at Kookmin University and is currently a professor at the Business IT. He is interested in deep learning, text mining, and data modeling.