Journal of the KNST

2026; Vol.9, No.2; pp.329-342

ISSN: 2635-4926 https://doi.org/10.31818/JKNST.2026.6.9.2.329

Check for updates

Received: 2026/03/11
Revised: 2026/03/23
Accepted: 2026/05/27
Published: 2026/06/30

*Corresponding Author:
Hyun-Ho Na
264-60, Sanho-daero, Gumi-si, Gyeongsangbuk-do, 39370, Republic of Korea
Tel: +82-54-460-5230
Fax: +82-54-460-8519
E-mail: na.hh1@hanwha.com

Abstract

본 논문에서는 LLM과 RAG를 활용한 잠수함 전투체계 교범 질의응답 시스템 구축 방안을 제시한다. 문서 임베딩, 벡터 데이터베이스, fine-tuning을 포함한 파이프라인을 구성하고 객관식, 주관식 각 100문항의 벤치마크 데이터로 성능을 비교 평가하였다. 실험 결과 rag와 fine-tuning 적용 시 기본 모델 대비 응답 정확도와 문맥 적합성이 향상되었으며, 잠수함 전투체계 적용 가능성을 확인하였다.

This paper proposes a method for large language model- and retrieval-augmented generation (RAG)-based question answering for submarine combat system manuals. A pipeline consisting of document embedding, a vector database, and fine-tuning was implemented, and the performance was comparatively evaluated using a benchmark dataset composed of 100 multiple-choice and 100 open-ended questions. The experimental results indicate that applying RAG and fine-tuning improves response accuracy and contextual relevance compared with the baseline model and confirm the applicability of the proposed approach to submarine combat systems.

Keywords

대형언어모델(Large Language Models),
검색증강생성(Retrieval-Augmented Generation),
미세조정(Fine-Tuning),
국방 인공지능(Military AI),
잠수함 전투체계(Submarine Combat Systems)

LLM, RAG 기반 잠수함 전투체계 교범 시스템 구축 방안 연구

Development of LLM- and RAG-Based Question Answering for Submarine Combat System Manuals

나현호[1*], 임진규[2], 이재근[1], 서창원[3]

[1]한화시스템 해양연구소 전문연구원
[2]한화시스템 해양연구소 연구원
[3]한화시스템 해양연구소 선임연구원

Hyun-Ho Na[1*], Jin-Kyu Lim[2], Jae-Geun Lee[1], Chang-Won Seo[3]

[1]Principal engineer, Naval R&D Center, Hanwha Systems
[2]Engineer, Naval R&D Center, Hanwha Systems
[3]Senior engineer, Naval R&D Center, Hanwha Systems

1. 서론

최근 인공지능(AI) 기술의 비약적인 발전은 다양한 산업 및 공공 영역에 큰 변화를 일으키고 있으며, 특히 국방 분야에서도 그 응용 가능성이 주목받고 있다. 그중에서도 대형언어모델(large language model, LLM)은 자연어 처리 및 생성 능력을 기반으로, 인간의 의사소통 방식을 모사할 수 있는 기술로 자리 잡았다. 이러한 LLM은 기존의 규칙 기반 시스템이나 고정된 절차서 중심의 정보 제공 방식과 비교해, 훨씬 더 직관적이고 유연한 정보 활용을 가능케 한다.

대한민국 국방부는 최근 '국방혁신 4.0' 정책을 추진하면서, 첨단 과학기술을 기반으로 한 지능형 국방 역량 강화를 핵심 과제로 설정하고 있다. 국방혁신 4.0은 AI, 빅데이터, 클라우드, 디지털 트윈 등 민간 주도의 기술을 국방에 선제적으로 도입하여 작전 수행 능력과 무기체계 운용 효율성을 극대화하는 것을 목표로 한다. 또한, 국방부는 '국방정책방향' 세부 추진 과제 이행 방안에서 AI 기술 수준과 발전단계를 고려하여 Fig. 1과 같이 '국방 AI 발전 모델'을 정립하였으며, 이에 따라 우리 군에 대한 AI 기술 적용을 단계적으로 확대해 나갈 것이라고 밝혔다. 이 중에서도 지능형 정보처리

및 전장 관리 자동화 기술은 실시간 상황 인식과 전술적 의사결정 지원 분야에서 매우 중요한 축을 이루고 있으며, LLM과 같은 자연어 기반 AI 기술은 미래 지휘통제체계(C4I)의 핵심 구성 요소로 부상하고 있다 [1,2].

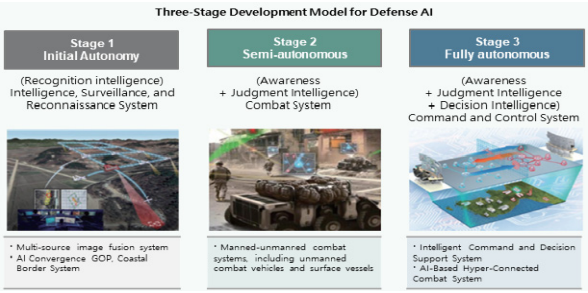

Fig. 1. Defense AI 3-stage development model

잠수함 전투체계는 고도의 기밀성과 복잡성을 지닌 군사 시스템 중 하나로, 수중작전 수행 시 다양한 센서와 무기체계를 통합적으로 운용해야 한다. 특히, 교전 상황에서는 신속하고 정확한 판단이 요구되며, 이를 위한 전술 교범의 역할은 매우 중요하다. 하지만 기존의 전투체계 교범은 방대한 문서 형태로 제공되며, 소요군 또는 시험 담당자가 실시간으로 필요한 정보를 효율적으로 파악하기에는 한계가 존재한다. 이러한 문제를 해결하는 방안으로, 본 논문에서는 LLM과 RAG(retrieval-augmented generation) 기술을 기반으로 하는 지능형 전투체계 교범 시스템을 제안한다. RAG는 대규모 문서 집합에서 필요한 정보를 검색하고, 이를 기반으로 LLM이 응답을 생성하는 방식으로 작동함으로써, LLM의 지식 한계를 보완하고 정확도를 향상시킨다. 이러한 구조는 실시간 전술 판단, 교범 해석, 사용자 질의응답 등 다양한 군사적 상황에 적용 가능성을 지닌다.

본 논문에서는 LLM과 RAG 기술의 개요를 설명하고, 이를 활용한 잠수함 전투체계 교범 시스템의 구현 방안을 제시하며, 해당 시스템의 프로토타입 테스트 결과를 통해 그 가능성을 검토한다. 또한, 국방 시스템에 적용하기 위한 보안 고려 사항 및 실전적 유용성에 대해서도 고찰한다. 논문의 구성은 다음과 같다. 2장에서는 본 연구의 기술적 배경으로 LLM, RAG, 그리고 잠수함 전투체계의 특성에 관해 설명한다. 3장에서는 제안하는 시스템의 전체 구조 및 주요 구성 요소를 제시하고, 4장에서는 구현된 시스템의 테스트 결과와 성능 평가를 수행한다. 마지막으로 5장에서는 본 연구의 결론 및 향후 발전 방향에 대해 논의한다.

2. 본론

본 장에서는 잠수함 전투체계 교범 시스템 구축에 필요한 주요 이론적 배경과 핵심 기술들을 설명한다. 특히, 본 연구에 사용한 기술인 함정 전투체계, LLM, RAG, LangChain 등 최신 자연어 처리 기반 프레임워크 및 구성 요소에 대해 상세히 기술함으로써, 이후 제안하는 시스템의 구성 근거와 기술적 타당성을 제시하고자 한다.

최근 LLM의 발전과 함께 다양한 분야에서 자연어 기반 지식 검색 및 질의응답 시스템에 대한 연구가 활발히 진행되고 있다. 특히 기술 매뉴얼이나 절차 중심 문서를 대상으로 하는 질의응답 시스템에서는 RAG 기법이 널리 활용되고 있으며, 문서 검색 기반 응답 생성 방식이 높은 정확도를 보이는 것으로 보고되고 있다. 또한, 군사 및 국방 분야에서도 인공지능 기술을 활용하여 작전 정보 검색, 의사결정 지원, 지식 관리 체계를 구축하려는 연구가 점차 증가하고 있다. 일부 연구에서는 연계된 자연어 인터페이스를 통해 운용자가 필요한 정보를 빠르게 검색할 수 있도록 하는 방법이 제안되었다. 그러나 기존 연구의 대부분은 일반 문서나 공개 데이터셋을 대상으로 수행되었으며, 실제 군사 교범과 같은 보안 문서를 기반으로 LLM과 RAG를 적용한 연구는 제한적인 상황이다. 따라서, 본 연구에서는 잠수함 전투체계 교범 문서를 대상으로 LLM, RAG, fine-tuning 기법을 결합한 질의응답 시스템을 구축하고 각 접근 방식의 성능을 비교 분석함으로써 국방 도메인에서의 적용 가능성을 검증하고자 한다.

2.1 Naval Combat Management System

함정 전투체계는 인간의 두뇌와 같은 역할을 수행한다. 함정에 탑재된 레이더, 소나와 같은 센서, 함포, 유도탄과 같은 무장과 기타 장비들을 통합해 작전에 필요한 모든 정보를 종합적으로 수집, 전술 상황 평가·지휘 결심·위협평가·무기 할당·교전 등의 기능을 효율적으로 수행할 수 있도록 지원하는 일련의 자동

화된 무기체계를 함정 전투체계라고 한다. 현재 함정 전투체계는 Fig. 2와 같이 지휘무장통제체계(combat fire command system/CFCS), 센서, 무장 및 데이터링크로 구성되며 함정의 임무와 특성에 따라 다양한 구성이 가능하다. 이 가운데에서 다양한 체계를 통합해 최대의 성능을 발휘할 수 있도록 하는 핵심적인 역할은 CFCS가 수행한다. CFCS는 데이터를 처리하는 단일기판컴퓨터(single board computer/SBC)가 내장된 정보처리장치(information processing node/IPN), 운용자의 명령을 입력받고 전술 상황을 전시하는 다기능 콘솔(multi function console/MFC), 센서 및 무장을 연동시키는 연동단(interface control unit/ICU), 연동되는 체계 상호 간 데이터를 전송하는 전투체계 통합 네트워크, 레이더 비디오 및 TV/IR 비디오를 분배·전송하는 영상 분배 장치(radar TV video distributor unit/RTVDU), 그리고 각종 전시기 등으로 구성된다[3].

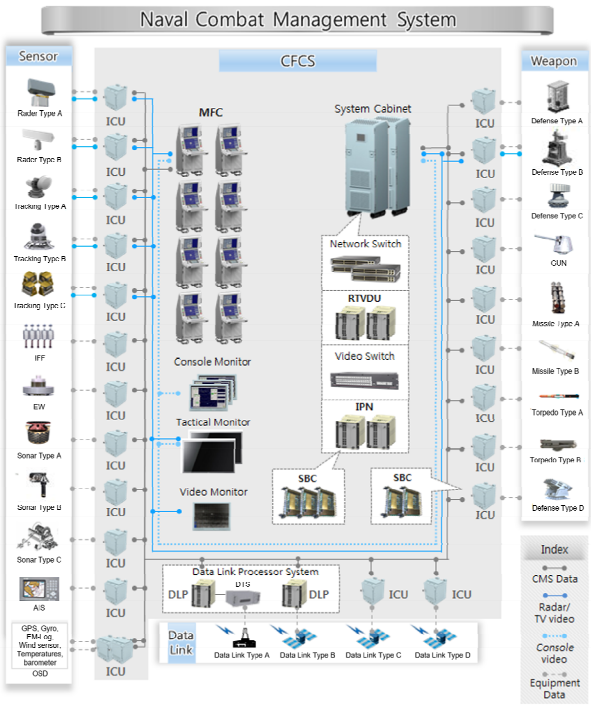

Fig. 2. Naval combat management system

2.2 LLM(Large Language Model), sLLM(Small LLM)

LLM은 트랜스포머 아키텍처를 기반으로 하는 인공지능 모델로, 방대한 텍스트 데이터 학습을 통해 자연어 이해 및 생성 능력을 갖추었다. 수십억 개 이상의 매개변수를 통해 복잡한 언어 패턴을 학습하며, 인간과 유사한 문맥 이해 및 응답 생성이 가능하다. LLM은 질문 응답, 텍스트 요약, 번역, 창의적 글쓰기 등 다양한 자연어 처리 작업에 활용되며, 지속적인 연구를 통해 성능이 향상되고 있다. 그 핵심은 입력 시퀀스를 기반으로 다음 단어의 확률 분포를 예측하는 방식으로 작동하여, 일관성 있고 유창한 텍스트를 생성하는 것이다. 군사 작전 환경에서는 대화형 인터페이스를 통해 전술 교범에 기반한 조언을 제공하거나, 복잡한 절차를 요약하여 제공하는 등의 활용이 가능하다. sLLM은 LLM을 압축하거나 경량화한 모델로, LLM의 성능을 유지하면서도 경량화된 형태로 제공된다. 보통 LLM은 수십억 개 이상의 파라미터를 가지고, sLLM은 수억~수십억개의 파라미터를 가진다. 본 연구에서는 단일 GPU에서 구동할 수 있어 자원을 효율적으로 사용하여 전투체계 내에서 사용하기에 적합하고 한글 지원이 가능한 공개 모델 중 sLLM 기반의 Gemma3 4B 모델을 이용 및 RAG를 적용, 해당 모델을 관련 도메인 데이터를 이용하여 fine-tuning 하여 시험을 진행하였다.

2.3 RAG(Retrieval-Augmented Generation)

RAG는 생성형 AI가 지닌 한계를 보완하기 위해 고안된 개념으로, 생성 모델 내부의 통계적 패턴만으로 답변을 생성하는 방식에서 벗어나 외부 지식의 실시간 참조를 가능하게 한다[4]. RAG 시스템은 사용자의 질문이 들어왔을 때, 먼저 관련성이 높은 문서를 외부 지식 저장소(예: Vector 데이터베이스)에서 검색한다. 검색된 문서는 LLM에게 제공되는 맥락 정보로 활용되며, LLM은 이 맥락을 바탕으로 답변을 생성한다. 이 방식은 LLM이 학습 시점에 포함되지 않은 최신 정보나 도메인 특화된 지식을 활용할 수 있게 하여, 답변의 정확성과 신뢰성을 높이고 허위 정보 생성(hallucination) 문제를 완화하는 데 효과적이다.

2.4 LangChain

LangChain은 LLM 기반 애플리케이션 개발을 간소화하는 오픈 소스 프레임워크이다[6]. 다양한 LLM, 데이터 로더, Vector 저장소, 에이전트 등 모듈화된 구

성 요소를 제공하여 개발자가 복잡한 LLM 파이프라인을 쉽게 구축할 수 있도록 지원한다. LangChain을 활용하면 RAG 시스템, 챗봇, 데이터 분석 도구 등 LLM을 활용한 다양한 애플리케이션을 효율적으로 설계하고 구현할 수 있으며 구성 요소 간의 유연한 조합과 확장성을 특징으로 한다. LangChain을 이용하면 잠수함 전투체계 교범 시스템에 필요한 검색기, 프롬프트 템플릿, 생성기, 보안 필터 등 다양한 구성 요소를 쉽게 결합할 수 있다.

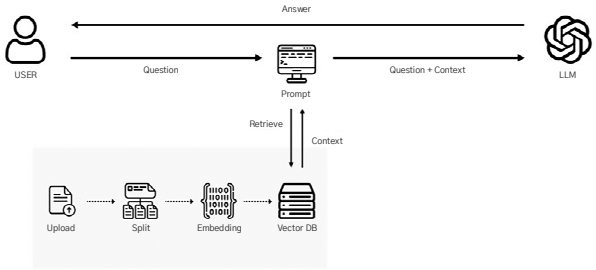

Fig. 3. RAG architecture[5]

2.5 Prompt Engineering

Prompt engineering은 LLM이 제공하는 답변의 품질을 결정하는 중요 기술로서, 모델의 사전학습 및 미세조정 없이 추론단계에서 prompt의 구조와 내용만을 재조정하여 모델의 출력을 최적화하는 기법이다[7]. 효과적인 prompt engineering은 모델의 잠재력을 최대한 활용하며, 명확하고 구체적인 지침과 충분한 맥락 정보를 포함하는 것이 중요하다. 다양한 프롬프트 기법이 연구 및 활용되고 있다.

2.6 Embedding

Embedding은 단어나 문장과 같은 이산적인 텍스트 데이터를 저차원의 연속적인 실수 vector로 변환하는 과정이다. Word2Vec 모델을 통해 CBOW (continuous bag of words)와 skip-gram 구조를 제안하여, 단어 간 의미적 유사성을 반영하는 embedding vector를 효율적으로 학습할 수 있도록 하였다[8]. 이 vector 공간에서 의미상으로 유사한 단어나 문장은 서로 가깝게 위치하게 된다. embedding은 단어의 의미, 문맥, 관계 등을 수치로 표현하며, 이를 통해 컴퓨터가 텍스트 데이터를 효율적으로 처리하고 분석할 수 있게 한다. 또한, embedding은 정보 검색, 텍스트 분류, 군집화, 추천 시스템 등 다양한 자연어 처리 및 기계 학습 응용 분야의 핵심 기술이다. 교범 문서를 사전에 vector화하여 vector DB에 저장하면, 이후 사용자의 질의와 의미상으로 유사한 교범 내용을 빠르게 검색할 수 있다.

2.7 Vector Database

Vector database는 고차원 vector 데이터의 효율적인 저장 및 검색에 특화된 데이터베이스 시스템이다. 텍스트 임베딩, 이미지 특징 vector 등 다양한 형태의 vector 데이터를 저장하고, vector 간의 유사도를 기반으로 가장 유사한 vector를 빠르게 찾아내는 기능을 제공한다. RAG 시스템에서는 검색 단계에서 외부 문서의 임베딩을 저장하고, 사용자 질문의 임베딩과 비교하여 관련 문서를 검색하는 데 필수적으로 사용된다. 대규모 데이터셋에서도 빠른 검색 속도를 제공하는 것이 주요 특징이다. 군사 시스템에서는 검색 속도와 보안, 로컬 운영 가능성 등을 고려해 오픈소스 기반의 Chroma DB를 사용하였다. Chroma DB는 임베딩된 데이터를 효율적으로 관리할 수 있도록 설계된 경량 데이터베이스로, 빠른 근접 이웃 탐색을 지원하고, 검색 단계에서는 코사인 유사도를 기준으로 질의와 가장 유사한 vector를 탐색한다. 이와 같은 설정은 전투체계 교범 질의응답 환경에서 불필요한 과잉 검색을 방지하고, 사용자가 필요로 하는 핵심 응답을 확보하는 데 최적화시킬 수 있다.

2.8 Fine-Tuning

Fine-tuning은 사전 학습된 대규모 언어 모델을 특정 하위 태스크나 도메인에 맞게 추가 학습시키는 과정이다. 비교적 소량의 레이블링된 도메인 특화 데이터를 사용하여 모델의 가중치를 미세 조정함으로써, 해당 태스크에 대한 모델의 성능을 크게 향상할 수 있다. Fine-tuning은 일반적인 LLM을 특정 산업의 전문 용어나 특정 형식의 응답에 더 잘 적응시키는 데 효과적이다. 이는 모델을 처음부터 학습시키는 것보다 계산 비용이 훨씬 적게 들면서도 뛰어난 성능 향상을 가져온다. 본 연구에서는 공개 모델 중 sLLM 기반의 Gemma-3-4b-it-unsloth-bnb-4bit 모델을 기

반으로 fine-tuning한 모델을 이용하여 시험을 진행하였다.

3. 제안방안

본 연구는 다음과 같은 절차로 진행하였다.

1. 전체 시험 개요
2. RAG 구축을 위한 데이터 수집 및 전처리
3. 문서 임베딩 및 vector DB 저장
4. 벤치마크 데이터 생성
5. Model fine-tuning을 위한 데이터 추출
6. Fine-tuning 진행 과정
7. 최종 시험 계획

본 연구에서는 실제 잠수함 전투체계 교범 데이터를 기반으로 실험을 수행하였다. 다만 군사 보안상의 이유로 교범 원문을 논문에 직접 공개하는 것은 제한되므로, 정량적 성능 평가는 보안 환경에서 구축된 실제 교범 데이터셋을 이용하여 수행하였다. 반면, 시스템의 동작 과정을 설명하기 위한 정성적 예시 및 질의응답 결과는 공개 가능한 대체 문서를 활용하여 제시하였다. 본 연구에서는 이러한 대체 문서로 차량 매뉴얼 문서를 활용하였으며, 해당 문서 역시 동일한 전처리 과정, 문서 임베딩, 벡터 데이터베이스 구축 및 RAG 기반 질의응답 파이프라인을 적용하였다. 이를 통해 보안 데이터를 직접 공개하지 않으면서도 시스템 구조와 응답 생성 과정을 이해할 수 있도록 구성하였다.

본 장에서는 LLM과 RAG를 기반으로 한 잠수함 전투체계 교범 시스템의 구축 절차와 시험 과정을 기술하였다. 전체 시스템은 RAG 기반 응답 시스템을 중심으로, 데이터수집 → 임베딩 처리 → vector 저장소 구축 → 단계별 벤치마크 시험으로 구성되며, 각 단계는 성능 향상 여부를 확인하고 최적화 방안을 도출하는 데 목적이 있다. 실험은 세 가지 조건에서 동일한 질의 셋(benchmark dataset)에 대해 응답 결과를 비교함으로써 RAG와 fine-tuning의 도입에 대해 체계적으로 분석하였다. 또한 Table 1은 각 구현 요소별 개발 환경이다. LLM과 RAG를 기반으로 한 잠수함 전투체계 교범 시스템 구축은 LangChain framework를 이용하여 Python으로 개발이 진행되었다. Fine-tuning을 진행하고, 추가 학습을 위한 하드웨어 스펙에 좀 더 유연함을 가지기 위하여 Unsloth 라이브러리와 기존 Gemma3-4B에 양자화를 더한 모델인 “Gemma-3-4b-it-unsloth-bnb-4bit” 모델을 사용하였다. Embedding 모델의 경우 고려대학교에서 개발 및 배포한 ”KoE5”를 사용하여 한글 기반 특화된 모델을 사용하였으며, vector DB로는 Chroma를 사용하였다. 모든 개발 환경의 구성은 보안상 로컬 및 폐쇄망 환경에서 진행하였다.

Table 1. Implementation environment

Implementation elements	Development environment and tools
Python Devenv	Visual Studio Code
RAG Framework	LangChain
LLM	Gemma-3-4b-it-unsloth-bnb-4bit
Embedding Model	KoE5
Fine-Tuning Library	Unsolth
Vector DB	Chroma

추가로 관련 시험을 진행하기 위한 하드웨어 및 소프트웨어 환경 및 버전 정보는 Table 2와 같다.

Table 2. Experimental environments

Item	Specification / Verison
CPU	AMD Ryzen 5 7600
GPU	Nvidia RTS 2080 Ti11Gb
RAM	32 GB
OS	Window 10
PyTorch	2.6.0+cu126
Python	3.11.10
Unsloth	2025.4.7
LangChain	0.3.24
ChromaDB	1.0.16

3.1 전체 시험 개요

본 실험은 잠수함 전투체계 교범에서 발췌한 약 74장의 문장을 대상으로 진행하였으며, 각 문장은 무장 운용에 필요한 교범 내용을 기반으로 하고 있다. 질의응답 테스트를 위해 도출된 벤치마크 질의 셋은 객관식, 주관식 100개로, 실제 잠수함 운용 시 발생 가능한 시나리오에 기반한 문항들로 구성되었다. 본 연구에서의 시험절차는 Fig. 4와 같이 구성되어 있으며, 기존 LLM 모델 단독 응답, LLM+RAG 통합 응답,

LLM+fine-tuning 모델 응답을 토대로 평가를 진행하였다.

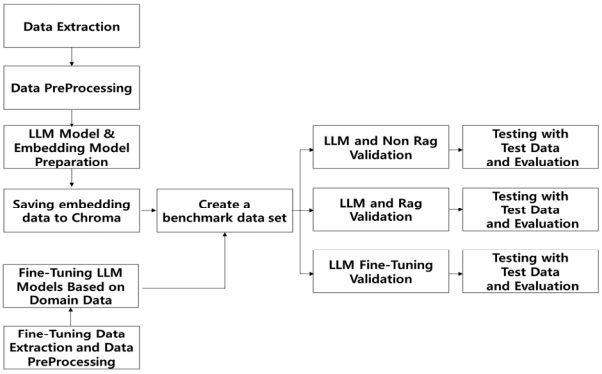

Fig. 4. Research process

3.2 RAG 구축을 위한 데이터 수집 및 전처리

본 연구는 보안 요구사항을 충족하기 위해 로컬 폐쇄망 구조에서 진행하였다. 개발 중인 잠수함 전투체계 교범 중 무장 분야에 해당하는 데이터를 사용하였으며, 확보한 데이터는 본 시험용으로 사용 이후 삭제 처리하였다.

무장 분야는 전투체계에서 사용하는 무장들에 대해 운용하는 방법이 작성되고 있어, 작전 시 매우 중요한 부분이므로 질의응답 시스템의 사례로 적합하다. 데이터 전처리 과정에서는 원문 자료를 RAG 기반 검색, 응답 구조에 최적화하기 위해 여러 정제 단계를 수행하였다. 먼저 불필요한 특수 문자와 개행을 제거하고 텍스트의 일관성을 확보하였다. 이어서 장, 절 번호와 표제를 통일된 체계로 재정렬하여 검색 시 계층적 문맥을 유지할 수 있도록 하였으며, 교범 내 표 형식 자료는 마크다운(markdown) 구조로 변환하여 향후 질의응답 과정에서 표 데이터를 그대로 활용할 수 있도록 하였다. 또한 각 문단은 100~200단어 수준으로 분할되어 임베딩 및 검색 효율성을 높였으며, 반복적으로 기술된 절차나 의미 없는 구문은 제거하여 데이터셋의 정밀성을 강화하였다.

이와 같은 전처리를 통해 구축된 데이터셋은 전투체계 교범 시스템에 최적화된 형태로 정제하였다. 이후 정제된 데이터는 문서 임베딩을 거쳐 vector 데이터베이스에 저장되어 RAG 검색 구조의 기반이 되었고, fine-tuning 학습용 데이터셋으로 재구성되어 모델의 도메인 특화 학습에 활용되었으며, 시험 및 평가를 위한 벤치마크 데이터셋으로도 구성되어 성능 비교 분석에 사용하였다. 전투체계 교범은 일반 문서와 달리 다층적 절차 구조, 전문 용어, 약어와 기호가 혼재되어 있기 때문에 단순한 텍스트 정제만으로는 한계가 존재하였다. 본 연구에서는 이러한 특성을 고려하여 군사 용어의 표준화, 절차적 서술의 유지, 계층적 문맥의 보존을 전처리 과정의 핵심 원칙으로 설정하였다. 이러한 접근은 검색 효율성과 도메인 적합성을 동시에 보장하여 이후 단계의 임베딩, vector 검색, fine-tuning 및 벤치마크 시험에서 안정적이고 신뢰성 높은 결과를 도출할 수 있었다.

3.3 문서 임베딩 및 Vector DB 저장

RAG 기반 질의응답 시스템의 핵심은 사용자의 질의를 단순히 언어모델에 입력하는 것이 아니라, 관련 문서에서 참조한 내역을 함께 제공하여 보다 정확하고 맥락에 맞는 응답을 생성하는 데 있다. 이를 위해 본 연구에서는 전처리 된 교범 데이터를 embedding 과정을 거쳐 vector로 변환하고, 이를 vector DB에 저장하여 활용하였다.

임베딩 과정에서는 고려대학교에서 개발한 한국어 특화 임베딩 모델인 “KoE5”를 사용하였다. “KoE5”는 한국어 문장 간 의미 유사도를 정밀하게 반영할 수 있도록 설계된 모델로, 교범과 같이 구사적 특수 용어와 절차적 기술이 포함된 문서를 처리하는 데 적합하며 한글 임베딩을 진행함에 따라 선정하였다[9]. 텍스트는 평균 500토큰 단위로 chunking 하여 분할 처리하였으며, 이는 문서의 문맥을 유지하면서도 검색 효율성을 높이기 위한 전략으로 지나치게 큰 단위로 나누면 검색 정확도가 저하될 수 있고, 반대로 지나치게 작은 단위로 나누면 맥락이 단절되는 문제가 발생할 수 있기 때문에, 본 연구에서는 실험을 통해 500토큰 단위가 최적의 균형을 제공한다고 판단하였다.

생성된 vector는 오픈 소스 기반의 vector DB인 Chroma DB에 저장하였다. 질의당 평균 10개의 유사 문단을 반환하도록 설정하였다. 이와 같은 설정은 LLM, RAG 기반 전투체계 교범 시스템 환경에서 불필요한 과잉 검색을 방지하고, 사용자가 필요로 하는 핵심 응답을 확보하는 데 최적화시킬 수 있다.

저장 환경은 군사적 보안 요구사항을 충족하기 위

해 로컬 폐쇄망 구조로 진행하였다. 모든 데이터는 오프라인 환경에서 관리되었으며, 보안이 중요한 문서이다 보니, 외부 네트워크와의 연결은 철저히 차단하였다. 이는 잠수함 전투체계와 같은 고도의 보안성이 요구되는 시스템에서 필수적인 조건으로, 정보 유출 가능성을 원천적으로 차단하고, 모델 학습 및 검색 과정의 안전성을 보장할 수 있었다.

3.4 벤치마크 데이터 생성

시험절차에서 사용하기 위한 벤치마크 데이터셋은 잠수함 전투체계 교범 중 무장 운용 분야를 중심으로 구축하였다. 무장 운용 분야는 무장 운용 및 발사관 운용 절차 및 작전 수행에서 핵심적 역할을 담당하므로, LLM, RAG 기반 잠수함 전투체계 교범 시스템의 성능을 검증하는데 가장 적합한 영역이라 판단하였다. 이를 위해 기존 매뉴얼 내 기술된 내용을 기반으로 주관식 100문항과 5지선다형 객관식 형식의 100문항으로 질문, 정답 데이터셋으로 구성하였다. 이러한 접근은 모델의 단순 정의형 질문뿐만 아니라 상황 맥락적 질문에도 대응할 수 있도록 설계되었다.

데이터 생성 과정에서는 로컬 환경에 배치된 LLM인 “Gemma3 27B”를 활용하였다. 이 모델은 다국어 및 멀티모달 입력을 지원하는 최신 구조를 기반으로 하고 있으며, 특히 절차적 문맥이 포함된 텍스트를 처리하는 데 있어 높은 적합성을 보였다. 학습 기반 데이터는 앞서 정제된 교범 데이터였으며, LLM이 이를 참조하여 초기 질의응답 데이터셋을 자동 생성하도록 하였다. 그러나 LLM을 통해 자동으로 생성된 데이터만으로는 정확성과 실효성을 보장하기 어려웠다. 따라서 생성된 200개(주관식 100개, 객관식 100개)의 문항은 해당 분야 SW 설계 및 개발을 진행한 전문가를 통해 수기 검증을 거쳤다. 검증 절차에서는 질문의 적절성, 응답의 정확성, 용어 사용의 일관성, 절차적 정합성 등을 세부적으로 평가하였다. 이 과정을 통해 잘못 생성된 질문이나 모호한 답변은 수정하였으며, 실제 교범의 내용과 일치하지 않는 응답은 교정하거나 제거하였다. 결과적으로 벤치마크 데이터셋의 정확성은 초기 LLM 생성 단계 대비 현저히 향상되었으며, 이는 벤치마크 데이터셋을 이용한 실험에서 신뢰성을 보장하는 중요한 기반이 되었다. 특히 해당 데이터셋은 다음과 같은 특징을 가진다. 첫째, 다양한 질의 유형(정의형, 절차형, 상황 대응형, 조건부 질문 등)을 포함하여 모델이 다양한 질문 패턴을 학습하고 평가받을 수 있도록 설계되었다. 둘째, 객관식과 주관식 응답 구조를 병행하여, 단순 사실 확인뿐만 아니라 응답의 논리적 정합성과 설명 능력을 평가할 수 있도록 하였다. 셋째, 실제 전문가 검증을 통해 데이터셋의 품질을 보증함으로써, AI 모델이 생성한 데이터의 잠재적 한계를 보완하였다.

3.5 Model Fine-Tuning을 위한 데이터 추출

본 연구에서는 LLM을 도메인 특화 영역에 맞게 최적화하기 위하여 fine-tuning 데이터셋을 별도로 구축하였다. 데이터셋은 기존 RAG 구축 과정에서 활용한 교범 문서 중 무장 운용 분야 텍스트를 기반으로 하였으며, 벤치마크 시험에 사용된 데이터는 중복성을 배제하기 위해 제외하였다. 이 과정을 통해 총 400개의 질문, 답변 쌍을 생성하였다.

모델을 fine-tuning하기 위한 데이터셋 생성 방식은 벤치마크 데이터셋과는 구분되는 특징이 있다. 벤치마크 데이터셋이 5지선다형 객관식과 주관식을 혼합하여 모델의 전반적 성능을 평가하는 데 초점을 맞췄다면, fine-tuning용 데이터셋은 전적으로 주관식 질의응답 형식으로 구성하였다. 이는 모델이 단순 선택형 응답이 아니라 자연어 기반 질문 및 응답을 학습하고자 하였다. 각 질문은 실제 잠수함 무장 운용 절차에서 발생할 수 있는 상황을 반영하여 설계되었으며, 응답은 교범 내 서술을 기반으로 정답 데이터를 매칭하였다. 생성된 데이터셋은 학습과 검증 데이터셋으로 분리하여 구성하였다. 일반적으로 학습용 데이터셋은 전체 중 300개로 구성하였으며, 나머지 100개는 검증용으로 사용하였다. 이러한 분리 방식은 모델이 학습 과정에서 과적합되는 것을 방지하고, 검증 데이터를 통해 성능을 정량적으로 평가하고자 하였다.

학습 및 검증 데이터셋의 구조는 JSON 포맷으로 구조화하였으며, Table 3와 같은 형태로 구성하였다.

Table 3. Fine-tuning dataset template

Benchmark dataset template
{ "Question":"질문, "Answer":"답변" }

JSON 포맷은 향후 fine-tuning 프로세스에서 손쉽게 불러올 수 있으며, 데이터셋의 재활용성과 확장성을 높이는 장점이 있다.

3.6 Fine-Tuning 진행 과정

베이스 LLM 모델인 “Gemma-3-4b-it-unsloth-bnb-4bit”에 fine-tuning을 진행하기 위해 3.5절에서 생성한 학습용 데이터셋 400개를 300개는 학습용, 100개는 검증용 데이터셋으로 분류하였다. Fine-tuning을 위하여 먼저 JSON 파일로 저장된 학습용 데이터셋을 읽은 후 본 연구에 사용된 모델 “Gemma-3-4b-it-unsloth-bnb-4bit”의 Chat template Table 4에 맞게 변환을 진행하였다.

Table 4. Chat template

Chat Template
<bos><start_of_turn>user 질문<end_of_turn> <start_of_tern>model 답변<end_of_tern><eos>

변환된 데이터는 SFT trainer(supervised fine-tuning trainer)를 사용하여 fine-tuning을 진행하였다. SFT trainer는 HuggingFace에서 제공하는 학습 기반 클래스이다. 사전 학습된 모델과 데이터를 빠르게 미세조정을 수행할 때 사용한다[10].

Fine-tuning 진행 시에는 Unsloth 라이브러리를 활용하였다. Unsloth 라이브러리는 fine-tuning 효율화를 목표로 하는 라이브러리로 fine-tuning에 필요한 메모리 사용량을 50% 줄이고 학습 속도를 크게 향상시켜 주는 라이브러리이다. Unsloth로 모델을 불러와 fine-tuning에 필요한 VRAM을 현저하게 줄여서 학습할 수 있었다[11].

전처리된 데이터를 이용하여, batch size 4, 10 Epoch만큼 학습을 진행하였다. 학습 시작 이후 training loss는 감소하기 시작하여, 최종 0.1488로 수렴하는 것을 볼 수 있었다, 시험용 데이터셋 100개로 검증 시, 기존 LLM으로는 답하지 못하였던 답변들에 대해 학습된 내용을 기반으로 답변이 나오는 부분을 확인할 수 있었다. 해당 모델을 이용해 검증용 데이터셋의 시험 결과는 93%의 정확도가 나왔다.

3.7 최종 시험 계획

앞서 생성된 RAG와 base 모델 기반 fine-tuning을 진행한 모델과 생성한 benchmark 데이터셋을 이용한 시험을 계획하였다.

첫 번째, base 모델을 그대로 사용하여 시험을 진행하였다. 두 번째, 기존 모델과 RAG를 결합하여 우리가 구축한 지식 데이터 기반으로 답변을 생성하는지 확인하였다. 세 번째, RAG를 제외한 base 모델에 잠수함 전투체계 무장 분야 교범의 내용에서 추출한 데이터를 기반으로 fine-tuning을 진행한 모델을 이용하여 우리가 원하는 답변이 나오는지에 대한 시험을 진행하였다. 또한, 최종 평가는 accuracy, precision, recall, answer correctness, context relevance, context faithfulness, context recall을 통하여 각각의 시험 과정을 평가하였다.

4. 실험 결과 및 평가

본 장에서는 base LLM 단독, LLM + RAG, LLM fine-tuning의 세 가지 접근 방식을 적용하여 잠수함 전투체계 교범 시스템의 성능을 비교, 평가하였다. 시험은 3장에서 구축한 벤치마크 데이터셋을 활용하여 진행하였으며, 각 접근 방식이 응답 품질에 미치는 영향을 체계적으로 분석하였다. 본 논문에서 사용한 74장의 교범은 보안 항목으로 외부에 공개할 수 없는 항목들이다. 따라서 본 논문에서 제시하는 시험 결과는 실제 잠수함 전투체계 교범을 활용한 모델들에서 나온 값으로 평가를 진행하고 예시는 보안을 위하여 잠수함 전투체계의 교범과 유사성이 높은 기아 타스만 차량의 매뉴얼을 선정하여 테스트 환경을 구성하였다. 기아 타스만 차량의 매뉴얼을 기반으로 한 테스트 환경은 동일하게 LLM, LLM+RAG, fine-tuning의 3개 모델로 구성되며, 해당 모델들에서 생성한 답변을 예시로 작성하였다.

4.1 기존 모델을 이용한 벤치마크 데이터 시험 결과

우선 비교 기준을 마련하기 위하여, 사전 구축된 RAG나 fine-tuning을 적용하지 않은 기존 LLM의 성능을 평가하였다. 시험에 사용된 모델은 Google에서 공개한 다국어, 멀티모달 기반 경량화 모델인

“Gemma-3-4b-it-unsloth-bnb-4bit”이며, 본 모델은 LoRA(low-rank adaptation) 기반 경량화와 4bit 양자화가 적용된 버전으로 제한된 하드웨어 환경에서도 실행 가능하다는 장점이 있다. 시험은 총 200개(주관식 100개, 객관식 100개) 문항으로 구성된 벤치마크 질의 셋을 대상으로 진행되었다. 이 질의 셋은 잠수함 전투체계 교범, 특히 무장 운용 절차를 중심으로 설계되어 있으며, 단순 사실 확인형 질문부터 절차 설명형, 상황 대응형 질문까지 다양하게 포함하였다. 그러나 기존 모델의 응답은 외부 문서 참조 없이 사전 학습된 일반 지식에 기반하여 생성되었기 때문에, 특수 목적 도메인인 잠수함 전투체계 교범과의 정합성이 제한적이었다. 실제 응답에서는 교범 관련 맥락을 부분적으로 반영하였으나, 절차적 세부 사항이 빠지거나 모호한 설명을 제시하는 경우가 다수 관찰되었다. 보안상 전투체계 교범의 예시를 사용하지 못하여 기아 타스만의 매뉴얼을 이용한 예시 결과로 대체하여 사용하였다. 예컨대, “지문 인증 기능에 대해 설명해 주시고, 기아 페이(전자 결제)와 관련된 내용을 알려주세요.”라는 질문에 대해, 해당 모델은 “기아 페이(전자 결제) 관련: 인포테인먼트 시스템이나 간편 설명서의 웹 매뉴얼을 통해 프로필 연동, 발렛 모드 설정 및 해제, 기아 페이 설정 방법을 확인하실 수 있습니다.”라는 식의 일반적인 응답을 생성하여 교범 내의 구체적인 절차와는 거리가 있었다.

평가 결과, 총 200개 질의에서 평균 객관식 문항 100개에 대해 정확도 61%를 기록하였다. 이는 모델이 기본적인 응답 생성 능력은 보유하고 있으나, 교범 문서 기반의 특화 지식 응답에는 한계가 있음을 보여준다. 주관식 문항 100개에 대해 답변 정확도를 1~5점 척도로 측정한 결과 1.67점이 나오며, 주관식 답변의 정확성에서도 낮은 점수를 기록하여, 군사적 응용에 있어 LLM 단독 모델의 활용은 제한적임을 확인하였다.

따라서 본 시험 결과는 LLM 단독 접근 방식이 전술 교범 기반 응답 품질을 보장하기에는 불충분하다는 사실을 입증하며, 이후 RAG와 fine-tuning 기법의 적용 필요성을 뒷받침할 수 있었다.

4.2 기존 모델 + RAG 벤치마크 데이터 시험 결과

기존 LLM인 “Gemma-3-4b-it-unsloth-bnb-4bit”에 RAG 구조를 통합하여 동일한 벤치마크 데이터셋을 대상으로 시험을 시행하였다. 이 접근 방식에서는 LLM이 단독으로 응답을 생성하는 대신, 사용자의 질의와 유사도가 높은 문단을 vector 데이터베이스에서 검색한 후 해당 내용을 프롬프트 입력에 포함했다. 이를 통해 모델은 일반적 추론보다는 교범 문서에 기반한 응답 생성을 수행하게 될 것으로 추정하였다. 시험 결과, 응답에는 실제 교범 문단의 내용을 직접 반영하거나 인용하는 경우가 다수 관찰되었다. 예컨대, “기아 타스만 차량에서 무선 충전 패드는 몇 대까지 지원하나요?”라는 질문에 대해 일반 모델에서는 “정확한 지원 대수는 차량 사양에 따라 다릅니다.”라는 모호한 답변을 생성했지만, RAG를 통합한 모델은 ”무선 충전 패드 2개가 적용되어 2개의 스마트폰을 동시에 충전할 수 있습니다.”와 같이 매뉴얼의 구체적 내용을 포함한 응답을 생성하였다. 이는 단순한 생성형 응답이 아니라 원문 문서 참조에 기반한 정확성 강화 효과를 볼 수 있었다.

전투체계 교범에서 추린 객관식 100개 문항에 대해 정확도 82%를 기록하여, base 모델 단독 대비 약 21% 향상된 성능을 보여 응답 품질 전반에서 유의미한 개선을 확인할 수 있었다. 주관식 100개 문항의 답변에서도 응답이 문서 기반 근거를 포함하는 경우가 많았으며, 1~5점 척도의 응답 정확도는 3.41, 컨텍스트 관련성 4.11, 컨텍스트 충실성 4.62, 컨텍스트 충분성 3.76으로 사용자가 답변의 출처를 신뢰할 수 있는 구조로 진화했음을 알 수 있다. 이러한 결과는 RAG 기법이 특수 도메인 문서 기반 질의응답 환경에서 가지는 장점을 잘 보여주었다. 특히 LLM 단독 모델의 한계였던 모호성, 불완전한 절차 설명, 비전문적 표현 등이 RAG 통합 모델에서는 크게 줄어들었으며, 교범 문서와 직접적으로 연결된 정확하고 맥락에 맞는 응답이 가능해졌다. 다만, 검색된 문단 품질에 따라 응답의 구체성이 좌우되는 경향이 관찰되었으므로, 향후 성능 최적화를 위해서는 문서 분할 방식과 임베딩 품질 개선이 병행되어야 할 것이다.

결론적으로, 본 시험은 RAG 구조의 도입이 잠수함 전투체계 교범 시스템에서 응답 신뢰성과 정확도를 실질적으로 개선할 수 있음을 실험적으로 입증하였다. 이는 후속 단계에서 fine-tuning과의 결합을 통해 추가적인 성능 향상 가능성을 뒷받침하는 중요한 결과를 의미한다.

4.3 기존 모델 + Fine-Tuning 데이터 시험 결과

본 연구에서는 기존 LLM인 “Gemma-3-4b-it-unsloth-bnb-4bit”에 대해 잠수함 전투체계 무장 관련 교범 데이터를 활용하여 fine-tuning을 수행하였다. 데이터셋은 총 400개의 질의응답 쌍으로 구성되었으며, 이 중 300개는 학습용으로 100개는 검증용으로 분류하였다. 학습용 데이터셋은 주관식 질의응답 형식으로 모델이 절차적 지식과 군사적 문체를 학습할 수 있도록 구성되었으며, 검증용 데이터셋은 객관식 형식을 포함해 모델의 응답 일관성과 적합성을 더욱 체계적으로 평가할 수 있도록 설계하였다.

Fine-tuning을 수행한 결과, 모델의 응답 품질은 여러 측면에서 개선되었다. 우선 용어 선택의 해당 도메인과의 유사성이 강화되었으며, 교범에 등장하는 절차적 표현과 문장 구조가 보다 명확하게 재현되었다. 예를 들어, LLM 단독 모델이 “정확한 지원 대수는 차량 사양에 따라 다릅니다.”와 같이 일반적인 응답을 제시한 것과 달리, fine-tuning 모델은 “센터 콘솔에 2개의 무선 충전 패드가 있어 동시에 2대까지 충전할 수 있습니다.”와 같은 정형화된 문체로 응답하였다. 이러한 변화는 교범 문체를 그대로 반영할 수 있다는 점에서 기존 교범의 데이터를 이용해 모델을 도메인에 특화된 버전으로 튜닝한 것과 같은 효과를 가져왔다. 성능 측정에서도 fine-tuning 모델은 뚜렷한 개선을 보였다. 총 100개의 벤치마크 객관식 질의에 대해 정확도 89%를 기록하였으며, 이는 base 모델 단독 대비 28%, RAG 결합 모델 대비 약 7% 이상 높은 수치였다. 주관식 100개 문항에 대해서도, 응답 정확도 3.51의 기존 LLM, RAG 결합 대비 좋은 성능을 보여주었다. 특히 주목할 점은 기존 LLM이 대답하지 못하거나 모호하게 처리하던 질문에 대해서도 교범 내 내용을 근거로 한 구체적이고 정확한 응답을 생성하였다는 것이다.

Fine-tuning 모델은 RAG를 사용하지 않고도 높은 정확도를 달성하였다. 이는 모델이 특정 도메인 데이터셋을 이용하여 fine-tuning을 한 경우, 외부 문서 참조 없이도 교범 지식을 내재화할 수 있음을 의미한다. 하지만, 새로운 교범 개정이나 미학습 영역에 대한 질의에서는 최신성 측면에서 RAG 기반 모델이 여전히 장점을 가질 수 있다고 볼 수 있다. Fine-tuning을 적용한 LLM은 잠수함 전투체계 교범 시스템 환경에서 객관식 질의에 대한 정확도, 주관식 질의에 대한 응답 정확도에서 가장 우수한 성능을 보였다.

4.4 평가 항목

시험 결과 평가는 다음 네 가지 핵심 지표를 기준으로 수행되었다. 질의에 대해 실제 교범에서 제시한 절차와 동일 또는 근접한 응답을 생성하였는지에 대해 accuracy(정확도), precision(정밀도), recall(재현율), F1-score를 통해 평가하였다. 성능 지표들은 Table 5의 수식으로 정리할 수 있다. 클래스별 5개의 클래스의 다중 분류를 하기 위해 클래스별 성능지표

Table 5. Performance indicator formula

Metric	Per-class	Macro average	Micro average
Accuracy	-	$\frac{\sum_{i=1}^{5} C_{ii}}{N}$	$\frac{\sum_{i=1}^{5} C_{ii}}{N}$
Precision	$\frac{TP_i}{TP_i + FP_i}$	$\frac{1}{5}\sum i \frac{TP_i}{TP_i + FP_i}$	$\frac{\sum_{i=1}^{5} TP_i}{\sum_{i=1}^{5} (TP_i + FP_i)}$
Recall	$\frac{TP_i}{TP_i + FN_i}$	$\frac{1}{5}\sum i \frac{TP_i}{TP_i + FN_i}$	$\frac{\sum_{i=1}^{5} TP_i}{\sum_{i=1}^{5} (TP_i + FN_i)}$
F1-Score	$2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$	$\frac{1}{5}\sum i\, F1_i$	$2 \times \frac{\text{Precision}-\text{micro} \times \text{Recall}-\text{micro}}{\text{Precision}-\text{micro} + \text{Recall}-\text{micro}}$

뿐 아니라 다중 평균을 낼 수 있는 성능지표 수식을 이용하여 성능을 평가할 수 있었다. 추가로, LLM-as-a-Judge 방식을 추가 평가 방법으로 활용하였다. 이 방식은 인간의 주관적인 평가를 대체하여 생성된 응답의 품질을 좀 더 좋은 성능의 LLM이 직접 점수화하거나 비교 판단하게 하는 평가 방법으로 대규모 자동 평가를 가능하게 한다[12]. 해당 평가 방법을 이용하여 각 시험의 응답 정확도, LLM 모델에 RAG를 병합한 방식의 컨텍스트 관련성, 컨텍스트 충실성, 컨텍스트 충분성을 분석했다.

세 가지 시험 과정을 통해, 각각 기존 정답과 세 시험 과정을 거쳐 생성된 정답 데이터의 confusion matrix를 아래와 같이 작성하였다. 해당 사항은 객관식 문항에 관한 내용에 대해서 confusion matrix를 구축한 내용이며, 각각 Fig. 5는 기존 LLM 모델 단독 사용, Fig. 6는 기존 LLM 모델+RAG 사용, Fig. 7은 fine-tuning LLM 모델의 confusion matrix 생성 결과이다. 실제 정답과 예측 시 정답 값과 동일한 예측을 수행하는 비율이 기존 LLM 모델 단독 사용, RAG 추가, fine-tuning 수행을 진행하며 높아지는 것을 볼 수 있다.

4.5 정략적 성능 비교 결과

Table 6는 기존 LLM 단독, LLM+RAG, fine- tuning LLM의 세 가지 조건에 대한 성능 비교를 정리한 것이다. RAG와 fine-tuning을 진행한 LLM 모델의 성능이 일반 LLM을 사용한 것보다 월등히 우수하다는 것이 확인되었다. 벤치마크 객관식 데이터 100개에 대해 정답을 예측한 결과, accuracy의 경우 RAG는 약 0.21, fine-tuning LLM 모델의 경우는 0.28 정도의 성능이 우수하였으며, precision, recall, F1-score의 경우도 RAG와 fine-tuning LLM을 진행한 것이 월등히 앞선 것을 볼 수 있었다. 시험별 주관식 100개 문항의 응답 정확도의 경우 1~5점을 척도로 하였을 때, Table 7과 같이 base LLM 모델은 1.67, LLM과 RAG를 병합하였을 때 3.41로 base LLM 모델 대비 약 2배의 성능을 보였으며, fine-tuning LLM의 경우 LLM과 RAG를 병합한 것보다 약 3% 정도의 높은 정확도를 보여주었다.

Table 8은 RAG를 병합한 모델에 대해, vector 데이터베이스 내의 문서를 참조하는 부분에 대해

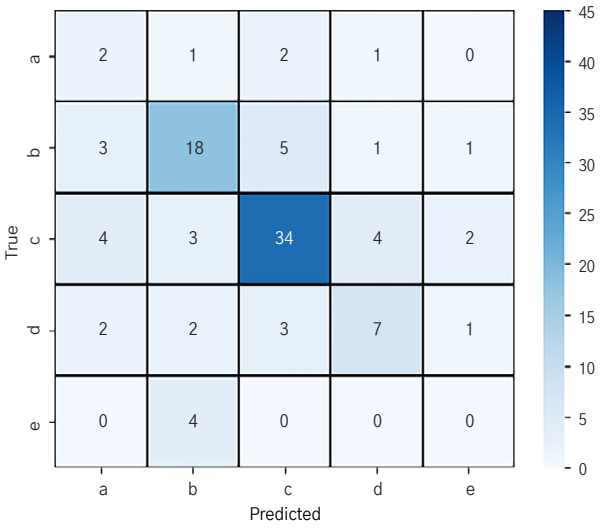

Fig. 5. LLM model exclusive confusion matrix

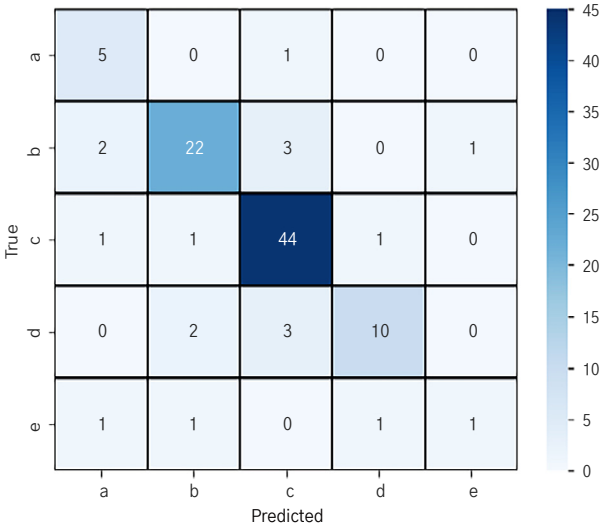

Fig. 6. LLM model + RAG confusion matrix

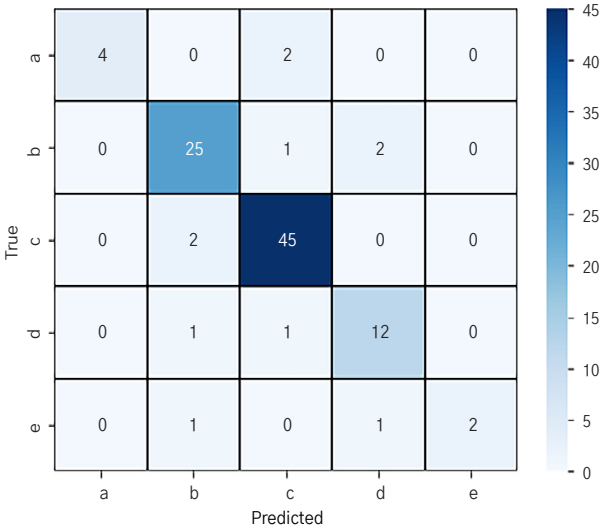

Fig. 7. Fine-tuning LLM model confusion matrix

LLM-as-a-Judge 기법을 활용하여 결과를 냈다. 결과를 내기 위해 LLM은 “Gemma3 27B”를 사용하였으며, RAG를 수행하였을 시에 전투체계 교범을 참조하여 결과를 생성하는 것을 볼 수 있었다.

Table 6. Performance quantitative evaluation

Condition	Accuracy	Precision	Recall	F1-score
LLM	0.61	0.427	0.433	0.425
LLM + RAG	0.82	0.719	0.694	0.690
Fine-Tuning LLM	0.89	0.916	0.763	0.816

Table 7. Performance answer correctness

Condition	Answer Correctness
LLM	1.67
LLM + RAG	3.41
Fine-Tuning LLM	3.51

Table 8. Performance LLM+RAG

Condition	Context Relevance	Context Faithfulness	Context Recall
LLM + RAG	4.1	4.76	3.76

4.6 정성 평가: 응답 품질 사례 비교

Table 9은 기존 LLM 단독, LLM+RAG, fine-tuning LLM의 세 가지 조건에 대한 정성 평가를 진행한 것이다. 본 논문에서 사용한 74장의 교범은 보안 항목으로 외부에 공개할 수 없는 항목들이어서, 예시는 전투체계 교범과 비슷한 기아자동차 타스만 매뉴얼을 이용하여 작성하였다. LLM 기본, LLM+RAG, fine-tuning 적용 LLM 모델 순으로 모호하게 처리하던 질문에 대해서도 타스만 자동차의 매뉴얼 내용을 근거로 한 구체적이고 정확한 응답을 생성하는 것을 볼 수 있다.

4.7 분석 및 논의

시험 결과를 종합적으로 살펴보면, LLM 단독 모델은 질의응답 과정에서 일반적이고 추상적인 응답을 생성하는 경향을 보였다. 이는 모델이 보편적 언어 이해 능력은 확보하고 있으나, 잠수함 전투체계 교범과 같이 고도로 특수화된 도메인 지식을 충분히 반영하지 못했기 때문으로 해석된다. 따라서 단일 LLM 접근법만으로는 실제 교범 절차와 일치하는 정밀한 응답을 제공하는 데 한계가 존재했다.

반면, RAG 기반 모델은 관련 문서를 vector 데이터베이스에서 실시간으로 참조함으로써 응답의 구체성과 정합성을 높일 수 있었다. 이를 통해 단순한 일반 지식수준을 넘어, 실제 교범 문장과 절차에 근거한 응답을 출력할 수 있었으며, 결과적으로 문맥 적합성 또한 크게 향상되었다. 이는 실제 운용 환경에서 신뢰할 수 있는 질의응답 체계를 구축하는 데 RAG의 도입이 필수적임을 볼 수 있었다.

또한, fine-tuning 모델은 전투체계 교범의 질의응

Table 9. Performance qualitative evaluation

Question	Model	Model Response
기아 타스만에서 스마트폰 무선 충전은 몇 대까지 지원하나요?	LLM	“기아 타스만은 스마트폰 무선 충전 기능을 지원하며, 최대 2대의 스마트폰을 동시에 충전할 수 있습니다. 1열: 운전석 및 조수석 앞쪽에 무선 충전 패드가 각각 1개씩 위치하여, 각자 스마트폰을 충전할 수 있습니다. 2열: 2열 시트에도 무선 충전 패드가 장착되어 있어, 뒷좌석 탑승객도 편리하게 충전할 수 있습니다. 다만, 무선 충전 기능은 스마트폰 모델에 따라 호환 여부가 다를 수 있으니, 차량 구매 시 또는 사용 전에 호환성을 확인하는 것이 좋습니다. 또한, 무선 충전 패드에는 스마트폰 크기 제한이 있을 수 있으므로, 큰 스마트폰이나 케이스가 두꺼운 스마트폰은 충전이 어려울 수 있습니다.”
	LLM + RAG	“기아 타스만은 최대 2대의 스마트폰 무선 충전을 지원합니다. 차량 중앙 콘솔 앞쪽에 무선 충전 패드가 2개 장착되어 있어, 운전자와 동승석 탑승객 모두 동시에 스마트폰을 충전할 수 있습니다. 주의사항: 무선 충전은 Qi표준을 지원하는 스마트폰에서만 가능합니다. 스마트폰을 무선 충전 패드 중앙에 정확히 위치시켜야 정상적으로 충전됩니다. 충전 패드 위에 다른 물건이 있으면 충전이 되지 않을 수 있습니다. 차량 전원이 꺼진 상태에서 도어를 열면 경고 메시지와 알림음이 울립니다.”
	Fine-Tuning LLM	“기아 타스만은 센터 콘솔에 2개의 무선 충전 패드가 있어 동시에 2대까지 충전할 수 있습니다.”

답 데이터셋을 학습함으로써 용어 사용의 전문성과 문장 구조의 정확성을 보완하였다. 이에 따라 응답이 교범 문체에 더 가깝게 생성되었으며, 군사적 표현력과 응답 일관성이 강화되었다. 특히 교범의 절차적 지식이 반영되어, 단순한 사실 전달을 넘어 실제적인 전술 운용 지원 기능을 수행할 수 있음을 확인하였다. 한편, 검색 성능의 품질은 vector화된 교범 문서의 구성 방식과 전처리 수준에 크게 의존하였다. 문서 분할 기준(chunk size)의 설정, 임베딩 모델의 품질, 불필요한 중복 구절 제거 여부 등이 최종 응답의 정확성과 신뢰성에 직접적인 영향을 미쳤다. 따라서 문서 분할 최적화와 임베딩 품질 개선은 향후 시스템 성능 고도화의 핵심 요소라 할 수 있다. 결론적으로, 본 연구의 시험은 LLM 단독 모델, RAG 기반 모델, fine- tuning 모델의 성능 차이를 명확히 보여주었으며, RAG는 도메인 특화된 내용을 문서 기반으로 보완하고, fine-tuning은 도메인의 내용을 추가 학습하여 base 모델을 업데이트하는 상호보완적 관계임을 확인하였다. 이는 잠수함 전투체계 교범과 같은 특수 도메인에서 AI 질의응답 시스템을 설계할 때, 단일 접근이 아닌 복합적 접근이 중요하다는 점을 알 수 있다.

5. 결론

본 논문에서는 LLM과 RAG 기반 기술을 활용하여 잠수함 전투체계 교범 질의응답 시스템을 구축하고, base LLM, LLM+RAG, fine-tuning 모델의 성능을 비교 분석하였다. 실험 결과 fine-tuning 기반 모델이 객관식 정확도와 주관식 응답 정확도 측면에서 가장 높은 성능을 보였으며, 이는 도메인 특화 데이터를 통해 모델이 교범 지식을 내부적으로 학습한 결과로 해석된다. 반면 RAG 기반 모델은 교범 문서를 직접 검색하여 응답 생성에 활용함으로써 문서 근거 기반의 응답을 제공할 수 있다는 장점을 확인하였다. 이러한 결과는 도메인 지식 내재화 방식(fine-tuning)과 문서 검색 기반 접근 방식(RAG)이 서로 다른 장점을 가지며 상호 보완적으로 활용될 수 있음을 시사한다.

또한, 본 연구를 통해 LLM 기반 질의응답 시스템이 전투체계 교범과 같은 기술 문서를 효율적으로 검색하고 활용하는 데 적용될 수 있음을 확인하였다. 이러한 시스템은 운용자가 필요한 정보를 신속하게 확인할 수 있도록 지원함으로써 전투체계 운용 효율성을 향상시키고 교육 및 훈련 환경에서도 활용될 수 있을 것으로 기대된다. 특히, 국방 혁신 4.0 정책과 연계하여 다양한 무기체계 운용 매뉴얼 및 기술 문서에 인공지능 기반 지식 검색 시스템을 적용할 수 있는 가능성을 제시하였다. 다만, 본 연구는 잠수함 전투체계 교범 중 일부 영역을 대상으로 수행되었으며, 보안상의 이유로 데이터 공개가 제한된다는 한계가 존재한다. 향후 연구에서는 다양한 전투체계 교범 영역으로 데이터 범위를 확장하고, fine-tuning과 RAG를 결합한 하이브리드 모델 구조를 적용하여 보다 높은 성능의 지식 검색 시스템을 구축할 필요가 있다.

참고문헌

[1] Department of Defense Editorial Board, 'Announcement of Defence Innovation 4.0 Framework Plan: Redesigning Defence at the Second Army Level, Fostering AI Science and Technology Force,' Defense & Technology, VOL. 530, 2023, pp. 10-11.
[2] Department of Defense Editorial Board, 'Report on the Implementation Plan for the Detailed Implementation of the National Defense Policy Direction: Establishment of a Three-Axis System, Strengthening of ROK-US Joint Exercises, and Creation of a Defense AI Center,' Defense & Technology, VOL. 522, 2022, pp. 11-13.
[3] Sang-Min Kwon & Seung-Mo Jung, 'Virtualization Based High Efficiency Naval Combat Management System Design and Performance Analysis,' Journal of the Korea Society of Computer and Information, VOL. 23, NO. 11, 2018, pp. 9-15.
[4] Yeo Chan Yoon & Soo Kyun Kim, 'Trends and Prospects of Retrieval-Augmented Generation (RAG) for Generative AI,' The Journal of Korean Association of Computer Education, VOL. 28, NO. 2, 2025, pp. 69-80.
[5] Jae-Hak Jeong, 'A Study on the Evaluation of RAG Using LLMs: Focusing on Inter-Firm ESG Reports,' Master's Thesis, Sogang University, 2025. http://www.dcollection.net/handler/sogang/000000079781
[6] LangChain. https://www.langchain.com/ (accessed 2026.03.10.)
[7] Min-Jun Son & Sung-Jin Lee, 'Research on Prompt Engineering Techniques in Large Language Models,' The Journal of Korean Institute of Communications and Information Sciences, VOL. 50, NO. 1, 2025, pp. 9-21. https://doi.org/10.7840/kics.2025.50.1.9
[8] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, & Jeffrey Dean, 'Distributed Representations of Words and Phrases and Their Compositionality,' arXiv:1310.4546, 2013. https://doi.org/10.48550/arXiv.1310.4546
[9] Hugging Face, KoE5.

https://huggingface.co/nlpai-lab/KoE5 (accessed 2026.03.10.)
[10] Hugging Face, SFTTrainer. https://huggingface.co/docs/trl/sft_trainer (accessed 2026.03.10.)
[11] unsloth. https://unsloth.ai/ (accessed 2026.03.10.)
[12] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing et al., 'Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena,' arXiv:2306.05685, 2023. https://doi.org/10.48550/arXiv.2306.05685