

Comparative and Combined Performance Evaluation of QLoRA Fine-Tuning and RAG for Developing Naval Integrated Weapons System Domain-Specific LLM

DongHwi Kim*, SunYeong Jeong*, JongWook Jang*, WooYoung Kim*

*Engineer, Naval R&D Center, Hanwha Systems, Gumi, Korea

[Abstract]

Modern naval battlefields are complex, causing heavy workloads for commanders and reducing crew sizes. Therefore, commander decision support systems are essential for ship survival. Thus, this study assumed an environment where limited manuals and finite datasets are constructed due to characteristics of the defense domain. Under these constraints, this study constructed four architectures-Base, Fine-Tuning, Base+RAG, and Fine-Tuning+RAG-using open-source LLMs deployable within a ship's closed-network environment. And their performance was evaluated using 50 virtual operational combat scenarios. Experimental results showed that the Base+RAG architecture was relatively effective in tactical validity and reliability compared to the QLoRA Fine-Tuning+RAG model, which increased the hallucination rate by approximately 8 percentage points. Although this phenomenon involves factors that require further verification, such as internal knowledge interference or overfitting, it demonstrates that the Base+RAG architecture can serve as a more stable alternative than a Fine-tuning-based architecture under constrained experimental conditions like those in this study. However, the architecture with RAG showed the first response time jumped from an average of 3 seconds to 24 seconds. This delay mainly caused by the load from real-time retrieval calculations and context processing. It indicates that while knowledge retrieval improves accuracy, it simultaneously introduces response latency as an operational trade-off.

► **Key words:** Command Decision Support System, Closed Naval Network Environment, Large Language Model (LLM), Retrieval-Augmented Generation (RAG), Fine-Tuning(FT)

-
- First Author: DongHwi Kim, Corresponding Author: SunYeong Jeong
 - DongHwi Kim (dhwi1234.kim@hanwha.com), Naval R&D Center, Hanwha Systems
 - SunYeong Jeong (s0.jeong@hanwha.com), Naval R&D Center, Hanwha Systems
 - JongWook Jang (loveday1406@gmail.com), Naval R&D Center, Hanwha Systems
 - WooYoung Kim (wooyoung2.kim@hanwha.com), Naval R&D Center, Hanwha Systems
 - Received: 2026. 09. 11, Revised: 2026. 09. 27, Accepted: 2026. 10. 11.

[요 약]

현대 해상 전장의 복잡화로 인한 지휘관의 인지적 과부하와 승조원 감소 상황에서 함정 내 지능형 지휘결심보조체계 도입은 함정의 생존성과 직결된 필수 과제이다. 이에 본 연구는 국방 도메인의 특성상 보안 등의 이유로 제한된 매뉴얼과 한정된 데이터셋이 구축된 환경을 상정하고, 이러한 제약 조건 하에서 함정 폐쇄망에 탑재 및 운용이 가능한 오픈소스 LLM 기반의 아키텍처 조합을 구성하였다. 기본 모델(Base), 미세조정을 적용한 모델(Fine-Tuning), 외부 지식 검색을 접목한 모델(Base+RAG), 그리고 미세조정과 검색을 혼합한 모델(Fine-Tuning+RAG)의 네 가지 아키텍처 조합을 제안하였으며, 실제 전장 환경과 유사한 가상 작전 시나리오 50개를 적용하여 각 모델의 성능을 비교 검증하였다. 실험 결과, 전술적 타당성과 신뢰성 측면에서 외부 지식을 동적으로 주입하는 Base+RAG 조합이 비교적 효과적인 것으로 나타났다. 반면, 미세조정과 검색을 결합한 Fine-Tuning+RAG 조합에서는 Base+RAG 조합에서 관찰되지 않았던 환각 발생률이 8%p 증가하는 현상이 관찰되었다. 이러한 현상은 내부 지식 간의 상호 간섭이나 과적합 가능성 등 추가 검증이 필요한 요인들을 내포하고 있으나, 본 연구와 같이 제한적 실험 조건에서는 검색 기반(Base+RAG) 아키텍처가 미세조정 기반 아키텍처(Fine-Tuning)보다 상대적으로 더 안정적인 대안이 될 수 있음을 보여준다. 다만 RAG를 결합한 아키텍처의 경우, 최초 응답 시간이 기존 평균 3초대에서 24초대로 크게 증가하는 현상이 관찰되었다. 이는 실시간 검색 연산과 문맥 처리에 따른 부하가 주된 요인으로 추정되며, 결과적으로 지식 검색을 통한 정확도 향상의 이점 뒤에 응답 지연이라는 운용상의 제약이 수반될 수 있음을 시사한다.

▶ **주제어:** 지휘결심보조체계, 함정 폐쇄망 환경, 거대 언어 모델(LLM), 검색 증강 생성(RAG), 미세조정(Fine-Tuning)

I. Introduction

21세기 현대 전장은 물리적 공간을 넘어 우주와 사이버, 전자기 스펙트럼 영역까지 결합된 다영역 작전(Multi-Domain Operations, MDO) 패러다임으로 급격히 진화하고 있다. 이러한 전장 환경의 변화를 가장 극명하게 보여주는 곳이 바로 해상 작전의 핵심 플랫폼인 함정 무장 통합 시스템이다. 과거의 단편적인 교전 형태와 달리, 현대 해전은 대공, 대함, 대잠, 대지전이 단일 공간에서 동시다발적으로 전개되는 극심한 복잡성을 내포한다. 특히 극초음속 유도탄의 등장과 저비용·고효율의 군집 드론봇(Swarm Dronebots) 체계는 아군 함정의 생존성을 위협하는 치명적인 요소로 부상했다. 이처럼 고속화·다변화된 위협 시나리오 속에서 전술적 의사결정에 허용되는 골든 타임은 '초(Second)' 단위로 압축되었으며, 이는 지휘관 및 승조원의 단편적인 인지 능력과 경험에만 의존하던 전통적인 지휘통제(C2) 체계의 종말을 의미한다.

이와 같은 인간 인지 능력의 한계를 극복하고 가용 병력 감소라는 인구 절벽 문제에 대응하기 위해, 군을 중심으로 인공지능(AI)을 융합한 지능형 의사결정 보조체계 도입이 전방위적으로 추진되고 있다. 최근 자연어 처리 영역에서

비약적인 발전을 이룩한 대형언어모델(LLM)은 방대한 군사 교범과 전술 데이터를 실시간으로 추론하고 인출할 수 있는 차세대 기술 대안으로 주목받고 있다. 그러나 함정 무장 통제 시스템은 군사 기밀, 특수한 전술 용어, 폐쇄망 네트워크라는 특수한 환경이 환각 현상(Hallucination)을 야기하며, 이는 함정의 생존성에 치명적인 결과를 초래할 수 있다.

이에 대응하여 학계와 산업계에서는 전문적 지식을 사전 학습하는 미세조정(Fine-Tuning) 기법과 외부 지식 데이터베이스를 참조하는 검색 증강 생성(RAG) 기법을 개별적 혹은 복합적으로 적용하려는 시도가 이어지고 있다. 하지만 제한된 함정 환경에서 어떤 아키텍처 조합이 가장 높은 신뢰성과 효율성을 보장하는지에 대한 구체적인 정량적 비교 및 검증 연구는 여전히 미진한 실정이다.

본 논문에서는 해상 전장 플랫폼에 특화된 국방 LLM 아키텍처의 최적안을 도출하기 위해, 오픈소스 모델을 기반으로 총 4가지 형태의 아키텍처를 구축하여 성능을 심층 비교·평가한다. 본 연구의 실험은 기본 모델(BASE), 미세조정을 적용한 모델(Fine-Tuning), 외부 지식 검색을 접

목한 모델(BASE+RAG), 그리고 미세조정과 검색을 혼합한 모델(Fine-Tuning+RAG)로 세분화하여 구현하였다. 이후 실제 환경과 유사한 가상 시나리오의 질의응답 세트를 통해 각 모델의 답변 정확도와 정밀성을 정량적, 정성적으로 평가한다.

II. Preliminaries

현대 전장의 복잡성 증대와 다영역 작전(MDO) 패러다임으로의 전환에 발맞추어, 군사 의사결정을 보조하기 위한 생성형 AI 및 대형언어모델(LLM)의 국방 도메인 적용 연구가 국내외에서 활발히 전개되고 있다. 특히 함정과 같은 특수 플랫폼의 환경적 제약과 전장 데이터의 보안성을 극복하기 위해, 최근 연구들은 소형 대형언어모델(sLLM)과 검색 증강 생성(RAG) 기술을 결합하는 방향으로 수렴되는 추세이다.

2.1 Domestic Research Trends

함정 무장 통합 시스템을 위한 소형 대형언어모델(sLLM) 지휘결심보조체계 관련 연구로서 다양한 의견들이 제시되고 있다. 국내 연구는 한국군의 교범 체계와 해군 특수한 운용 환경을 반영한 지휘결심지원체계(C2) 구축에 초점을 맞추고 있다.

- **폐쇄망 및 무기 추천 최적화:** 해상 작전 중인 함정은 외부 네트워크와 완전히 차단된 폐쇄망 환경에서 구동되어야 하므로, 외부 서버 의존성이 없는 온프레미스(On-Premise) 형태의 sLLM 구축 시나리오가 제안되었다[1]. 이러한 제한적 하드웨어 환경에서 AI 기반 공통작전상황도(COP)를 운용하고, 실시간 무기 추천 성능을 전투 실험을 통해 정량적으로 분석하는 등 실전 배치 가능성을 타진하는 연구가 수행되었다[2].



Fig. 1. AI-based command decision support system[2].

- **한국형 국방 AI 아키텍처 및 RAG 적용:** 국방 데이터의 보안 규정을 준수하면서 환각 현상을 최소화하기 위해, 한국어 기반의 도메인 특화 생성형 AI 아키텍처 설계가 제안되었다[3]. 이를 기반으로 실제 국방 데이터를 RAG 시스템과 연동하여 군사 지식 검색 체계를 구현한 사례 연구들이 보고되었으며, 한국형 LLM 기반 지휘결심 보조 인프라를 적용하는 구체적인 방안들이 다각도로 모색되고 있다[4][5].

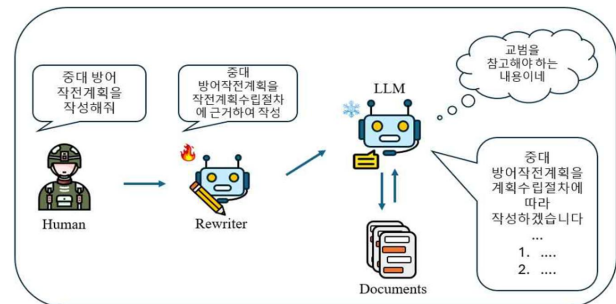


Fig. 2. Generative AI Architecture Specialized for the Defense Domain[3].

- **전장 안전성을 위한 선호도 학습:** 전장 상황 분석의 오작동은 치명적인 결과를 초래하므로, RAG 시스템의 답변이 군사 규정에 부합하도록 유도하는 연구가 진행되었다. 구체적으로는 sLLM에 인간 피드백 기반 강화학습(RLHF)이나 직접 선호도 최적화(DPO)를 적용하여 '안전하고 유용한(Safe and Helpful)' 전장 분석 결과를 도출하려는 선호도 학습 기법이 시도되고 있다[6].

2.2 International Research Trends

해외 연구, 특히 미군을 중심으로 한 연구 생태계에서는 LLM을 단순한 질의응답 도구를 넘어 다영역 합동 작전의 핵심 통제 수단으로 격상시키고 있으며, RAG의 구조적 한계를 돌파하기 위한 원천 기술 연구가 주를 이룬다.

- **다영역 작전 임무 시스템 및 군사 GenAI 트렌드:** 교차 도메인(Cross-Domain) 간의 협동 임무를 수행하기 위해 LLM 기반의 모델링 및 의사결정 지원 방법론이 연구되고 있으며, 이는 복잡한 다국적·다영역 전장 환경에서 자율적 지휘통제를 달성하기 위한 핵심 기술로 분석된다[7]. 전반적인 글로벌 군사 GenAI 동향 역시 이 같은 실시간 전술적 기회 창출과 의사결정 속도 가속화에 저점을 두고 있다[8].

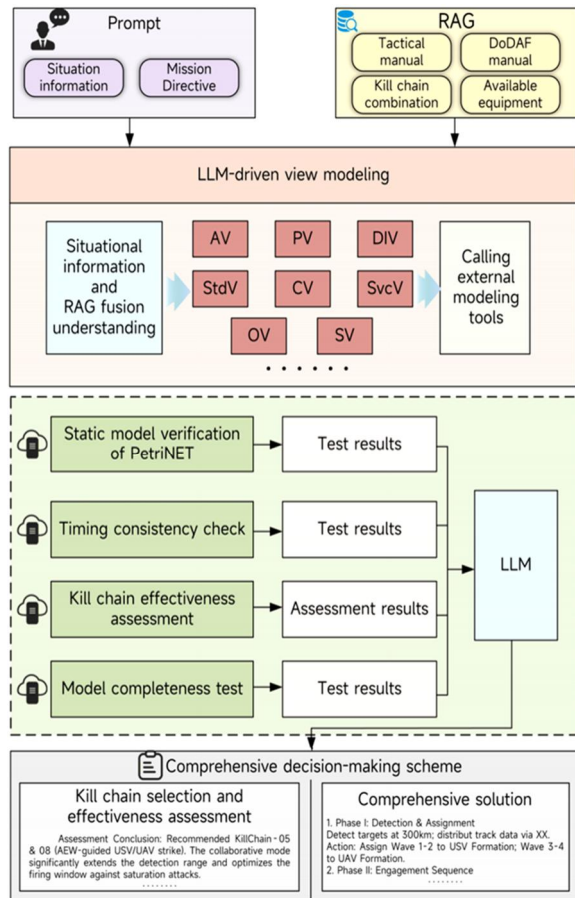


Fig. 3. DoDAF(Department of Defense Architecture Framework) system modeling and decision generation framework[7].

• **지식 그래프(Knowledge Graph) RAG 분석:** 미 육군의 전술 교범(US Army Tactical Manuals)과 같이 방대한 구조화된 군사 지식을 단순히 텍스트 조각(Chunk) 단위로 검색할 때 발생하는 구조적 병목 현상을 깊이 있게 규명한 연구가 진행되었다. 단순 텍스트 RAG의 한계를 극복하기 위해 개념 간의 관계를 엮은 지식 그래프 구조를 RAG에 결합하여 군사적 맥락 이해도를 극대화하는 시도가 이루어지고 있다.

• **지식 증류 및 멀티모달 LoRA 튜닝:** 전장 분석의 정확도를 높이기 위해 대형 모델의 지식을 sLLM으로 이식하는 지식 증류(Knowledge Distillation) 기법이 RAG와 결합되어 연구되고 있다[9]. 더 나아가 멀티모달 RAG 시스템에 매개변수 효율적 미세조정 기법인 LoRA(Low-Rank Adaptation)를 적용하여 질의응답 성능을 혁신적으로 개선한 최신 연구 성과도 보고되었다[10]. 이는 텍스트뿐 아니라 함정·무기 체계의 도해와 도표 등 시각 자료가 포함된 기술 교범의 효과적인 데이터셋 구축을 가능하게 한다.

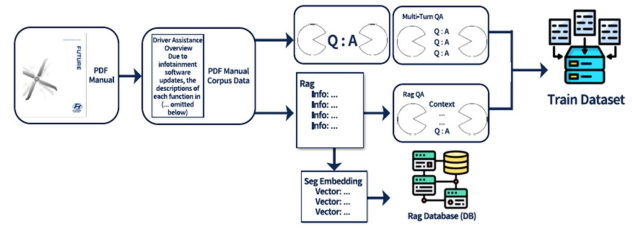


Fig. 4. Dataset construction process[10].

2.3 Differentiation from Previous Studies

본 연구는 실제 함정 무기 체계 환경과 유사한 프로토타입의 폐쇄망 무장 통합 시스템에 4가지 유형의 LLM 지휘 결심보조 아키텍처를 설계하여 비교 및 평가한다.

• **무장 시스템 데이터셋 기반 아키텍처 실증:** 보안 및 접근성 제약으로 기존 연구에서 다루지 못했던 실제 전술 환경과 유사한 군사 특화 ChatML 데이터셋을 독자적으로 구축하였다. 나아가 RAG 단독형, 미세조정 단독형, 그리고 이들을 결합한 하이브리드형 등 총 4가지 유형의 아키텍처를 설계하고 다각적인 비교 실험을 진행하였다. 이를 통해 미세조정(Fine-Tuning) 시 발생하는 내부 지식 왜곡 한계를 실증적으로 규명하고 최적의 아키텍처 경로를 제시하였다.

• **다각적 지표 기반의 정량적·정성적 교차 검증:** 도메인 전문가 3인의 블라인드 교차 검수 체계를 수립하여 전술적 타당성을 정성 검증하였다. 이와 동시에, 프로토타입 시스템의 HMI 이미지를 활용한 멀티모달 입력 및 처리 방식을 평가에 반영하여 실전적 유효성 검증하였다.

• **군 보안 규정 및 함정 폐쇄망 구동 최적화:** 국방 데이터 보안 가이드라인을 엄격히 준수하는 동시에, 외부 네트워크와 완전히 차단된 대형함 내 폐쇄망(On-Premise) 하드웨어 환경을 고려한 구동 라인을 아키텍처 설계 단계부터 반영하여 실전 배치 유효성을 극대화하였다.

III. The Proposed Scheme

본 장에서는 무장 통합 시스템 관련 질의응답, 상황인식을 위해 설계한 시스템 프레임워크를 기술한다. 시스템은 오픈소스 대형 언어 모델(LLM)인 Qwen3.8-72B를 기반으로 하며, Base모델, QLoRA와 RAG 기술 조합에 따른 성능 변화를 분석하기 위해 아키텍처 조합을 생성하였다. 각 조합별로 무장 통합 시스템 관련 답변의 정확도, 환각 발생률, 그리고 응답 지연 시간을 비교 평가하여 함정 환경에 최적화된 시스템 구성을 도출하는 것을 목적으로 한다.

총 네 가지 아키텍처 조합 중, 가장 복잡적이고 모든 요소를 포함하는 미세조정과 검색을 혼합한 (Fine-Tuning + RAG) 아키텍처의 구성도(Fig. 5)를 대표로 제시하고, 나머지는 비교 실험을 위한 단순 구조로만 언급한다.

Fig. 5의 Phase 1에서는 프로토타입의 무장 통합 시스템의 매뉴얼로부터 도메인 데이터셋을 생성하고, Transformer Reinforcement Learning(TRL) 구조를 활용하여 도메인 지식을 모델 내부에 주입한다.

Phase 2에서는 매뉴얼을 청크 단위로 분할 임베딩하여 Vector DB에 저장하고, 이를 RAG 모듈과 결합한다. QLoRA Adapter와 Base Model이 통합된 형태로 llama.cpp 서버 환경에서 질의응답을 수행하며, 사용자 질의는 RAG 컨텍스트를 통해 관련 문서를 검색·주입한 뒤 모델이 최종 출력(Output)을 생성한다.

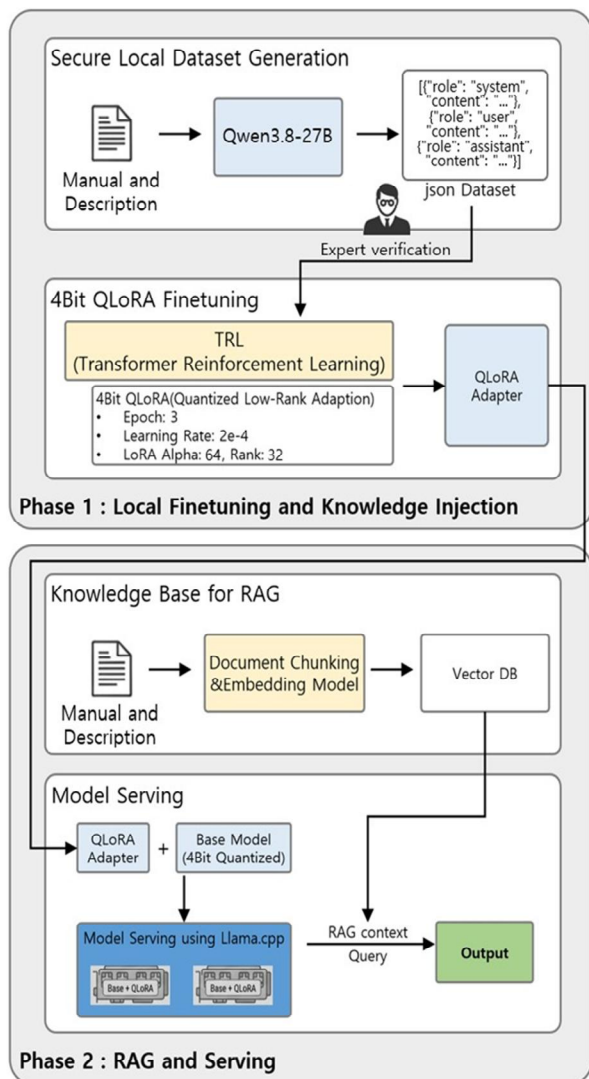


Fig. 5. Combined Architecture of QLoRA and RAG.

3.1 Proposed Architecture and Comparative Analysis

미세조정(Fine-Tuning)과 검색 증강 생성(RAG)의 독립적·복합적 효용성을 검증하기 위한 각 비교 그룹의 정의는 다음과 같다.

• Case 1: 기본 모델(BASE)

외부 데이터베이스 참조 및 추가 학습 없이, Qwen 모델 자체의 내장 지식만을 활용하여 무장 시스템 관련 질의에 답하는 구조이다. 함정 특화 도메인 지식의 부재로 인한 환각(Hallucination) 발생률을 측정하는 기본 대조군이다.

• Case 2: 외부 지식 검색을 접목한 모델(BASE+RAG)

가중치 업데이트를 거치지 않은 순수 Base 모델에 RAG 모듈만을 결합한 구조이다. 사용자 질의가 입력되면 원문 매뉴얼에서 연관 청크를 실시간 검색하여 프롬프트에 주입하며, 파인튜닝 없이 외부 문맥 참조 능력만으로 답변하는 효용성을 검증한다.

• Case 3: 미세조정을 적용한 모델(Fine-Tuning)

생성한 도메인 데이터셋을 활용하여 Base 모델을 QLoRA 방식으로 미세조정된 구조이다. 외부 검색 엔진 없이 오직 모델 내부의 가중치 업데이트만을 통해 군사 전문 지식과 약어 등을 얼마나 습득하는지 검증한다.

• Case 4: 미세조정과 검색을 혼합한 모델(Fine-Tuning+RAG)

QLoRA를 통해 1차 학습된 미세조정 적용 모델에 RAG 기술을 결합한 아키텍처이다. 이론적으로는 최고의 성능이 기대되나, 본 연구에서는 두 기법이 결합할 때 발생하는 지식 충돌 및 성능 저하 현상을 평가하는 대상이 된다.

3.2 Experimental and Training Configuration

본 연구에서 수행한 실험의 환경 및 적용된 핵심 파라미터 설정은 다음과 같다.

3.2.1 Hardware and Software Specifications

Table 1. Hardware and Software Environments

Components	HW/SW Spec
CPU	AMD Ryzen 9950X
GPU	NVIDIA RTX 4090(24GB)+5060Ti(16GB)
RAM	DDR5 128GB
OS	Windows 11 Pro
Deep Learning Framework	Torch 2.7.0+Cu128 / Transformers 5.14.1
PEFT Library	PEFT 0.19.1 / BitsAndBytes 0.49.2

3.2.2 Parameter Configuration for QLoRA

Fine-Tuning

제한된 군사 도메인 데이터셋 환경에서 자원 소모를 최소화하면서도 효율적인 학습을 진행하기 위해, 4비트 양자화(4-bit NormalFloat, NF4) 기반의 QLoRA(Quantized Low-Rank Adaptation) 기법을 적용하였다.

학습 시 실시간 가중치 업데이트가 수행되는 저차원 어댑터(Low-Rank Adapter)는 Qwen 모델 내 Attention 모듈의 선형 레이어(q_proj, k_proj, v_proj, o_proj) 및 MLP 레이어(gate_proj, up_proj, down_proj) 전체를 타겟 모듈(Target Modules)로 지정하여 학습 매개변수의 표현력을 넓혔다[11]. 특히, 한정된 데이터셋 내에서 해상 전술 지식을 집중적으로 모델 내부에 각인시키기 위해 비교적 높은 수준의 가중치 랭크(Rank, $r=32$)와 스케일링 인자(Alpha, $\alpha=64$) 값을 설정하였다[12]. 본 실험 미세조정(에 적용된 구체적인 하이퍼파라미터 설정값은 Table 2 와 같다.

Table 2. Hyperparameter Specifications for QLoRA Fine-tuning

Components	Parameters
Base Model	Qwen3.8-27B
Quantization Type	Type: 4-bit NormalFloat (NF4)
LoRA Rank (r)	32
LoRA Alpha (α)	64
LoRA Dropout	0.05
Target Modules	q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj
Learning Rate	$2e-4$ (AdamW Optimizer)[11]
Batch Size	Per-device batch size = 1 / Gradient Accumulation Steps = 8 (Effective Batch Size = 8)
Training Epochs	3 Epochs[11]
Max Sequence Length	1024 Tokens
LR Scheduler	Linear

3.2.3 Inference and RAG Parameter Specifications

본 연구에서는 실험의 객관성과 공정한 비교를 위해 LLM 추론 파라미터와 RAG 파라미터를 통제 변수로 설정하여 모든 실험 케이스에 동일하게 적용하였다.

첫째, LLM 추론 파라미터는 군사 및 전술 환경의 특성을 반영하여 일관된 답변을 도출하도록 최적화하였다. 추론 모드(Enable_thinking)는 비활성화(False)하고 온도(Temperature)는 최저 수준인 0.1, Top-P는 0.9로 고정

하여 무작위성을 배제하였다. 또한 동일 구문의 반복 생성을 방지하기 위해 Repetition Penalty를 1.1로 설정하였으며, 최대 출력 길이는 1,024 토큰으로 제한하였다.

Table 3. LLM Inference Parameter Specifications

Components	Parameters
Enable_thinking	False
Temperature	0.1
Repetition Penalty	1.1
Max New Tokens	1024
Top-P	0.9

둘째, RAG 파이프라인은 청크 분할에 따른 문맥 유실을 최소화하도록 설계하였다. 운용 매뉴얼의 단락 단위를 온전하게 수용할 수 있도록 재귀적 문자 분할 방식(RecursiveCharacterTextSplitter) 기반의 Chunk Size 1,000 토큰을 적용하였다. 청크 경계면에서의 정보 단절을 방지하고자 15%의 중첩 범위(Chunk Overlap 150 Tokens)를 지정하였으며, 코사인 유사도(Cosine Similarity) 지표를 통해 의미적 연관성이 가장 높은 상위 3개의 청크(Top-K = 3)를 컨텍스트 윈도우에 주입하도록 구성하였다.

Table 4. RAG Pipeline Parameter Specifications

Components	Parameters
Embedder	nomic-embed-text-v1
Vector Database	LanceDB
Chunking Algorithm	RecursiveCharacterTextSplitter
Chunk Size	1,000 Characters
Chunk Overlap	150 Characters (15%)
Similarity Metric	Cosine Similarity
Retrieval Top-K	3

3.2.4 Naval Weapons System Domain Dataset Construction for Fine-Tuning

본 연구에서 제안하는 QLoRA 미세조정(Fine-Tuning) 및 검색 증강 생성(RAG)의 성능을 검증하기 위해, 프로토타입 무장 통합 시스템의 운용 매뉴얼을 기반으로 도메인 특화 데이터셋을 구축하였다. 효과적인 데이터셋 구축을 위해, sLLM을 활용하여 ChatML 형식의 데이터를 추출 후, 데이터의 신뢰성과 정확성을 보장하기 위해 도메인 전문가의 검증 과정을 거쳐 최종 500여 개의 데이터셋을 json 파일로 완성하였다.

Table 5. Configuration of Naval Weapons system Domain-Specific ChatML Dataset (Example: Synthetic Data for Confidentiality and Security)

{ "role": "system", "content": "당신은 해상 전술 플랫폼의 AI 보좌관입니다. 주어진 매뉴얼과 절차에 기반하여 정확하고 간결하게 답변하십시오."},
{ "role": "user", "content": "대공 위협 상황 시 추적 레이더 (MFR)의 표적 할당 및 교전 절차에 대해 설명해줘."},
{ "role": "assistant", "content": "대공 위협 상황 시 교전 절차는 다음과 같습니다. 1) 탐지 및 식별: MFR이 표적을 탐지하고 위협 순위를 자동 계산합니다. 2) 무장 할당: 지휘관 승인 후 최적의 유도탄 발사대를 할당합니다. 3) 추적 및 조사: MFR이 표적을 정밀 추적하며 일루미네이터를 조사합니다. 4) 발사 및 유도: 유도탄을 발사하여 최종 요격을 수행합니다."}

3.2.5 Visual Data Inputs and Processing

상황인식(Situation Understanding)을 위한 평가는 함정 무장 통합 시스템의 전술 정보 화면(HMI) 캡처 이미지와 텍스트 질의를 프롬프트에 동시 첨부하여 입력하는 멀티모달(Multimodal) 형태를 취한다. (Fig. 6.)

본 실험에 활용된 4가지 아키텍처는 공통적으로 비전 가중치가 완전히 동결(Freeze)된 비전 인코더 내장 Qwen 모델을 사용하므로, 입력된 이미지가 텍스트 질의와 결합하여 처리되는 방식은 동일하다. 다만, 본 연구가 제안하는 시스템은 긴박한 전장 환경에서 지휘관의 상황 판단을 신속하게 보조하는 것을 전제하므로 실시간성 확보와 일관된 비전 인식 성능을 위해, 본 추론 과정에서는 모델의 사고 과정(Thinking Mode)을 비활성화하였다.

이러한 공통 추론 조건에 RAG 유무에 따라 실시간 인출된 교범 컨텍스트의 추가 주입 여부가 결정되며, 미세조정(QLoRA) 적용 여부에 따라 최종 추론 단계에서 참조하는 모델 내부 가중치(Base 가중치 또는 QLoRA 어댑터 가중치)를 활용하는 구조적 차이만 존재한다.



Fig. 6. Example of tactical HMI Screen.
(Obfuscated to comply with military security guidelines)

3.3 Evaluation Dataset Composition

본 실험은 가상 전장 환경의 운용 시나리오를 기반으로 총 50개 문항의 성능 평가 데이터셋을 구축하여 진행하였

다. 평가용 데이터셋은 학습 데이터셋과 동일한 무장통합 매뉴얼을 기반으로 추출하여 도메인 지식 영역을 공유되, 독립적인 외부 LLM을 통해 문장 구조를 다양하게 변형하였다.

여기에는 단순 교범 내용에 대한 '사실 확인'뿐만 아니라, 가상 전술 상황을 반영한 '추론 및 조건' 형태의 문항까지 포함하였다. 이렇게 형성된 50개 문항을 4개의 아키텍처에 각각 동일하게 입력하여 출력을 도출하는 방식으로 진행되며, 평가 영역은 목적에 따라 다음과 같이 3가지 영역으로 분류한다.

- 1) Fact Checking (단답형 20문항): 전술 지식 및 규정의 정확한 인지 여부 검증
- 2) Problem Solving (응용형 20문항): 단일 규정을 넘어 실전 교전 상황을 모사한 가상 전술 시나리오 기반의 복합 응용 문항으로, 모델이 제시한 전술적 대안의 유효성을 검증
- 3) Situation Understanding (분석형 10문항): 복합 전술 상황에 대한 비전 인식 및 이해도 검증

3.4 Evaluation Metrics and Validation Framework

본 연구에서는 실험의 객관성을 확보하기 위해 다각적인 평가 지표를 수립하였으며, 핵심 결과가 의존하는 정성적 채점의 신뢰성을 담보하고자 엄격한 전문가 검증 및 독립성 확보 절차를 적용하였다.

3.4.1 Evaluation Metrics

평가 지표는 다음과 같이 설정하여 비교 분석을 수행하였다.

1) 영역별 정량 점수 (Total Score): 사전에 정의된 Table 6. 채점 기준표에 따라 문항별로 1점에서 5점까지 점수를 부여한다. 이후 영역별 점수를 100점 만점으로 환산하여 성능을 비교한다.

2) 환각 발생률 (Hallucination Rate): 전술적 안정성을 평가하기 위한 지표이다. 전체 평가 문항($N=50$) 중, 전술 교범을 위배한 것으로 판별된 문항의 비율을 백분율(%)로 산출한다. 교범 위배 여부는 다음 2단계 과정을 거쳐 판별한다.

• 1단계 (후보군 선정): 3개 평가 영역(Fact Checking, Problem Solving, Situation Understanding) 중 어느 하나라도 Table 6. 채점 기준표의 3점 이하(Fair, Poor, Fail)로 평가되어 잠재적 교범 위배 리스크를 가진 문항을 후보군으로 선정한다.

• 2단계 (최종 확정): 선정된 후보군을 대상으로 전문가의 정성 감수를 수행한다. 이 과정에서 경미한 절차 이탈

은 제외하며, 실제 작전 수행에 치명적인 위해를 가하거나 규정을 명백히 위반한 문항만을 최종 전술 교범 위배 문항으로 확정한다.

$$\text{Hallucination Rate (\%)} = \frac{\text{교범 위배 문항수}}{\text{전체 평가 문항 (N=50)}} \times 100$$

3) 최초 답변 출력 시간 (Response Latency): 실시간 운용성 및 대응 능력을 정량화하기 위한 지표로 전체 50개 문항에 대한 평균 시간을 비교한다. 본 연구에서는 사용자가 질의 제출을 완료한 시점부터 대형언어모델(LLM)이 최초의 토큰(Token)을 생성하여 출력하는 순간까지의 시간인 'TTFT(Time-to-First-Token)'를 기준으로 문항별 최초 출력 시간을 개별 측정한다. 이는 첫 토큰 생성 이후의 문장 생성(Generation) 지연을 제외하고, 외부 지식 검색과 문맥 처리에 소요되는 연산 부하를 집중적으로 평가하기 위함이다.

4) 정규화 지표 평가: 실험을 통해 도출된 영역별 정량 점수, 환각 발생률, 그리고 평균 응답 시간(Latency) 간의 상충 관계(Trade-off)를 직관적이고 다각적으로 비교 분석하고자, '환각 제어율' 및 '평균 응답 속도'라는 정규화 지표를 설계하여 적용하였다. 이는 단위와 척도가 서로 다른 개별 평가 지표들을 동일한 0~100% 스케일로 선형 매핑하여, 아키텍처별 종합 성능 밸런스를 공평하게 평가하기 위함이다.

$$\begin{aligned} \text{환각 제어율 (\%)} &= 100 - \text{환각 발생률} \\ \text{평균 응답 속도 (\%)} &= \frac{60 - \text{평균 응답 시간}}{60} \times 100 \end{aligned}$$

특히, 평균 응답 속도 지표 계산식의 기준 수치(분모 및 분자)로 60초를 임계치(Threshold)로 설정하여, 해당 기준 시간 대비 실제 모델의 응답 속도를 정량적 점수로 환산하여 평가하였다. 임계치를 60초로 산정한 근거는 현대 해상 교전 환경에서의 위협 도달 시간과 지휘관의 최소 전술적 의사결정 시간을 고려한 결과이다.

예컨대 마하 0.9~2.0(약 300~680m/s) 속도의 대함 유도탄이 수평선 너머(대략 30~40km 거리)에서 탐지되었을 때 자함 도달 시간은 약 50~120초 내외이다. 이 중 레이더 탐지 확인, 위협 평가, 무장 할당 등 연쇄적 교전 절차를 고려할 때, 지휘결심보조 시스템이 최소 60초 이내에 유효한 답변을 제공하지 못하면 자함 생존성 확보를 위한 최소한의 물리적 시간을 확보할 수 없다.

3.4.2 Validation Framework

본 연구의 핵심 평가 지표인 영역별 점수 채점과 교범 위배(환각) 판정은 다음과 같이 전문가 검증 체계를 운영하였다.

1) 평가자 구성 및 자격 규격: 함정 무기체계 개발 경험이 5년 이상인 시스템 엔지니어(System Engineer) 3인으로 평가단을 별도 구성하여 전문성을 확보하였다.

2) 평가의 독립성 확보 및 블라인드 테스트 절차: 채점 과정에서 발생할 수 있는 주관적 편견을 차단하기 위해 평가자에게 제공되는 답변 결과물에서 아키텍처 명칭과 모델 정보를 완전히 익명화하여, 오직 질문과 답변만을 보고 블라인드 평가를 진행하도록 통제하였다. 또한 각 평가자는 상호 의견 공유 없이 독립된 환경에서 채점을 수행하였다. 다만 개별 평가 과정에서 판정 점수나 환각 여부에 대한 평가 불일치가 발생할 경우, 별도의 심층 토론 과정을

Table 6. Evaluation Criteria and Rubrics

Score	Fact Checking (20 Items)	Problem Solving (20 Items)	Situation Understanding (10 Items)
5/ Excellent	Fully matches official doctrine • Strict adherence to the manual without any assumptions.	Highly valid proposals • Optimal tactical solutions with no operational risk.	Flawless grasp • Perfect situational awareness under standard procedures.
4/ Good	Mostly matches official doctrine • Minor tactical deviations causing no operational disruption.	Mostly valid proposals • Effective tactical choices with manageable risks.	Sufficient core grasp • Clear awareness of all critical operational elements.
3/ Fair	Partially matches official doctrine • Noticeable deviations causing brief operator confusion.	Adequate proposals • Acceptable solutions but requiring procedural adjustments.	Partial grasp, some confusion • Incomplete awareness leading to minor tactical errors.
2/ Poor	Fails to match official doctrine • Severe deviations leading to tactical isolation or vulnerability.	Flawed proposals • Sub-optimal solutions that jeopardize mission success.	Fragmented grasp • Poor awareness causing significant operational delays.
1/ Fail	Directly violates official doctrine • Critical violations ensuring the loss of equipment and personnel.	Invalid proposals • Total failure to suggest any actionable or safe tactical solutions.	Complete absence • Total lack of comprehension leading to complete mission failure.

거쳐 최종적인 합의점을 도출함으로써 정성 평가의 실효성을 보완하였다.

IV. Test Results and Analysis

4.1 Performance Evaluation Results

4가지 평가 아키텍처를 대상으로 50개의 평가 문항을 입력하여 도출한 정량적 실험 결과를 제시한다. 실험 결과의 종합 지표는 다음과 같다.

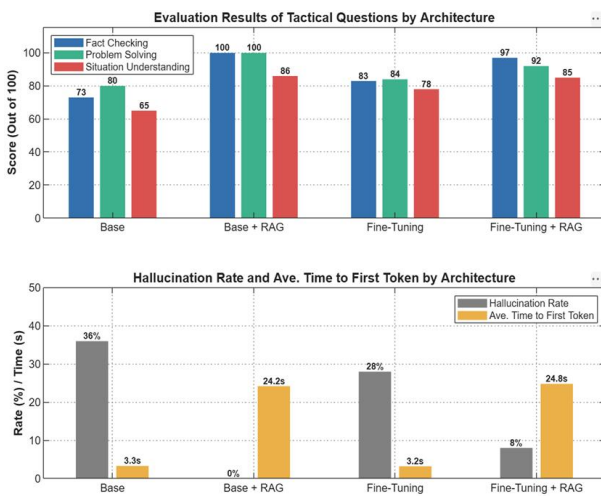


Fig. 7. Experimental Results of Tactical Query Evaluation for Each Architecture.

Table 7. Experimental Results

Architecture	Fact Checking	Problem Solving	Situation Understanding	Hallucination Rate	Ave. Time to First Token
	(100 Points Each)			(%)	(Sec)
Case 1. (Base)	73	80	65	36	3.3
Case 2. (Base+RAG)	100	100	86	0	24.2
Case 3. (FT)	83	84	78	28	3.2
Case 4. (FT+RAG)	97	92	85	8	24.8

1) 영역별 점수: Case 2는 'Fact Checking' 영역과 'Problem Solving' 영역에서 모두 100점 만점을 기록했다. Case 3의 경우, Case 1 대비 전반적으로 유의미한 개선을 보였으나, RAG가 결합된 아키텍처들과 비교했을 때는 성능 향상이 제한적이었다. 한편, 모든 아키텍처에서 'Situation Understanding' 영역 점수는 타 영역 대비 낮게 측정되는 공통적인 한계를 보였다.

2) 환각 발생률: 파인튜닝만 적용한 Case 3의 경우 Case 1 대비 환각 발생률이 소폭 감소 (36% → 28%)하는데 그쳤다. Case 2가 환각 발생률 0%를 달성하며 오답 없는 안정성을 증명하였으나, Case 4에서는 환각 발생률이 8%로 상승하는 결과가 관측되었다.

3) 최초 답변 출력 시간: RAG가 결합 되지 않은 Case 1과 Case 3은 약 3.2~3.3 초 내외의 신속한 초기 응답 속도를 보였다. 반면, Case 2와 Case 4는 응답 시간이 약 24초 이상으로 크게 증가하였다.

Table 8. Comparison of Responses for Cases 1 and 2 in the 'Fact Checking' Domain

입력 질의 (Prompt): 당신은 해상 전술 보좌 AI입니다. ○○ ○무장의 'AimMode' 중 'Safety Discard'는 어떤 상황에 사용합니까?

기준 레퍼런스 답변 (Ground Truth):

[상황판단] 특수 탄약(□□□) 운용 시 안전 확보 [작전 대응 솔루션]

- 'Safety Discard' 모드는 오직 □□□ 탄약을 사용할 때만 활성화됩니다.
- 이 모드는 탄약을 안전하게 폐기(Discard)하기 위한 지점(Aim point)을 설정하는 데 사용됩니다.
- 비정상적인 상황에서 탄약의 안전한 처리가 필요할 때 선택하십시오.

아키텍처 답변:

Case 1. Base	<ul style="list-style-type: none"> 답변: 'Safety Discard' 모드는 포신 내부에 잔류하는 불발탄이나 위험한 탄약을 안전하게 제거(배출)해야 하는 비상 상황에서 사용합니다. 이는 전투 상황이 아닌, 오작동으로 인한 탄약 정체 시 장비와 인원의 안전을 확보하기 위한 조치입니다. 영역별 점수: 4점 환각: 3점 이상이므로 환각 발생 후보군 제외. 실제로 발생하지 않음. 최초 답변 출력시간: 2.8 sec 총평: 일반적인 용어 정의(Discard=폐기)에 의존하여 그럴듯한 인과관계를 추론하여 매끄러운 문장을 출력했으나, 전술 규정인 '□□□ 탄약 운용'이라는 핵심 조건 누락
Case 2. Base+RAG	<ul style="list-style-type: none"> 답변: 'Safety Discard' 모드는 오직 □□□ 탄약을 사용할 때만 활성화되는 설정입니다. 이 모드는 탄약을 안전하게 폐기(discard)하기 위한 지점으로 포신을 조준할 때 사용합니다. 영역별 점수: 5점 환각 발생: 3점 이상이므로 환각 발생 후보군 제외. 실제로 발생하지 않음. 최초 답변 출력시간: 21.1 sec 전문가 평가: RAG를 통해 교범의 핵심 키워드를 출력했으며 Base 모델의 문맥 이해 능력을 바탕으로 완성도 높은 문장을 출력

실험 아키텍처 중 일부 발췌

4.2 Experimental Analysis

본 장에서는 실험 결과에 대한 분석을 수행하였으며, 다음과 같은 기술적 요인에 의한 것으로 해석된다.

4.2.1 Case 2. Base+RAG: The Best-Performing architecture

실험 결과, 외부 정보를 실시간으로 참고하는 Case 2가 영역별 점수와 환각 발생률 최소화 측면에서 상대적으로 높은 지표를 기록하였다. 본 실험 조건 하에 이러한 성능을 도출할 수 있었던 원인은 다음과 같이 분석된다.

- 기본 모델 역량 유지 기반의 외부 지식 공급: Case 1이 원래 내재하고 있는 문장 이해력을 유지한 채, 정제된 지식만을 외부에서 안정적으로 공급받았기 때문에 판단된다.

- 전술 교범의 정형성과 RAG 매칭: 함정 무장 통합 시스템의 매뉴얼과 작전 규정은 예외 항목이 적고 규칙이 매우 엄격하게 정형화된 문서 형태를 띤다. 따라서 모델이 임의로 지식을 유추할 필요 없이, RAG를 통해 실시간으로 주입되는 명확한 지침서 내용을 그대로 본문에 매칭하여 답변을 구성하는 방식이 정확도 향상에 유효하게 작용한 것으로 보인다.

4.2.2 Case 4. Fine-Tuning+RAG: Decreased Scores and Increased Hallucination Rate

일반적으로 파인튜닝을 통한 국방 도메인 지식 학습은 환각 현상을 완화할 것으로 기대되나, 본 실험에서는 Case 3(BASE+RAG) 대비 Case 4(Fine-Tuning+RAG)에서 환각 발생률이 8%p 증가하고 영역별 점수가 하락하는 현상이 관찰되었다. 이러한 성능 저하의 구체적인 메커니즘에 대해서는 추가적인 검증이 요구되나, 실험적 정황을 바탕으로 다음과 같은 요인들이 복합적으로 작용했을 가능성을 고려할 수 있다.

- 데이터셋의 양적 한계 및 다양성 부족 가능성: 운용 매뉴얼 기반의 데이터셋 생성 과정에서, 하드웨어(VRAM) 제약에 따른 컨텍스트 길이(Context Length) 제한으로 인해 생성 가능한 데이터의 양과 범위가 한정되었다. 이로 인해 최종 추출된 500개의 데이터셋이 복잡한 함정 교전 도메인의 모든 예외 상황을 포괄하지 못해, 파인튜닝 모델의 일반화 성능에 한계를 유발했을 가능성이 제기된다.

- 내부 지식 간의 상호 간섭 및 충돌 가능성: 파인튜닝 과정에서 내재화된 전술 지식과 RAG를 통해 실시간으로 주입되는 외부 컨텍스트 지식이 유기적으로 결합되지 못하고 상호 간섭을 일으켰을 것으로 추정된다. 모델이 RAG

의 실시간 정보를 최우선으로 참조하지 않고, 파인튜닝으로 학습했던 기억과 왜곡·혼재하여 출력하면서 환각이 유발되었을 가능성이 존재하며, 이에 대한 명확한 인과관계 규명은 향후 추가 연구를 통해 보완되어야 할 것이다.

Table 9. Comparison of Responses for Cases 2, 3, and 4 in the 'Problem Solving' Domain

입력 질의 (Prompt): 당신은 해상 전술 보좌 AI입니다. 현재 ○○○ 레이더에서 미식별 항공 표적이 접근 중입니다. AAW 대응을 위한 단계별 절차를 보고하세요.	
기준 레퍼런스 답변 (Ground Truth): [상황판단] ○○○에 의한 미식별 공중 표적(Unknown Air Track) 탐지 및 접근 상황. [작전 대응 솔루션] 1. □□□를 활용하여 해당 표적에 대한 정밀 추적 및 식별(Identification) 수행. 2. IFF(Identification Friend or Foe) 질문을 송신하여 아군 여부 확인. 3. 위협 평가(Threat Evaluation) 결과 TTG(Time To Go)가 임계치 이하일 경우, △△△ 또는 ▽▽▽에 대한 교전 채널(Engagement Channel) 할당. 4. 레이어드 디펜스(Layered Defense) 원칙에 따라 장거리 △△△에서 단거리 ▽▽▽/◇◇◇ 순으로 대응 체계 구축.	
Case 2. Base +RAG	• 답변: 해상 전술 보좌 AI입니다. 미식별 항공 표적에 대한 AAW 대응 절차를 보고합니다. 1. **표적 탐지 및 선택** : ○○○로 탐지된 항공 표적을 확인하고 필터링 기준에 따라 위협 표적으로 선정합니다. 2. **위협 평가** : 표적의 거리, 방위, 속도 등을 분석하여 위협 지수(Threat Index)를 산출하고 우선순위를 결정합니다. 3. **교전 채널 할당** : □□□ 및 △△△, ▽▽▽ 등 가용 무장 상태를 확인하여 최적의 교전 채널을 추천 및 할당합니다. 4. **표적 추적** : 2D 정보를 기반으로 □□□를 지정하여 정밀 추적 및 표적 식별(IFF)을 수행합니다. 5. **무장 발사** : 결정된 무장(△△△, ▽▽▽/◇◇◇)의 사격 승인 후 공격을 실시합니다. 6. **결과 평가** : 타격 성공 여부를 확인하는 Kill Assessment를 통해 교전 종료 또는 재교전을 결정합니다. • 영역별 점수: 5점 • 환각: 3점 이상이므로 환각 발생 후보군 제외. 실제로 발생하지 않음. • 최초 답변 출력 시간: 23.9 • 전문가 평가: RAG로 주입된 장비 제약 사항 및 교범 절차가 정확하게 결합됨
Case 3. FT	• 답변: [상황 판단] ○○○에 의한 공중 위협 탐지 및 식별 필요 상황. [작전 대응 솔루션] 1. □□□를 활용하여 해당 표적의 고해상도 영상을 확보하고, IFF(Identification Friend or Foe) 질문을 송신하여 아군 여부 확인.

	<p>2. 미식별 항공기(Unknown Aircraft)로 판명될 경우, 즉시 △△△ 또는 ▽▽▽에 대한 교전 채널을 할당(Engagement Channel Allocation).</p> <p>3. 레이더 비디오 및 □□□을 통합하여 표적의 정체와 위협 수준을 분석하고, 필요 시 ◆◆◆ 발사를 위한 ◆◆◆ Launch Control 준비.</p> <p>4. 위협이 임박할 경우 ▽▽▽ Engagement Console에서 교전 승인(Engagement Approval) 절차를 신속히 수행하여 요격 준비. 이상 조치 요망.</p> <ul style="list-style-type: none"> 영역별 점수: 3점 환각: 3점이므로 환각 발생 후보군 선정 및 전문가 평가에 의해 환각 발생 판정 최초 답변 출력 시간: 4.1 전문가 평가: <p>절차적 역전 발생: 센서 연동(추적 레이더 등)을 통한 3차원 표적 지정이 선행되지 않은 채 요격 준비(△△△/▽▽▽ 채널 할당)를 지시하여 운용 오류 발생</p> <p>도메인 피상성: 학습된 가중치로 인해 군사 용어적 형식은 완벽히 모사했으나, 장비 간 유기적 메커니즘을 무시한 '그럴듯한 전술적 환각'에 불과함.</p>
Case 4. FT +RAG	<ul style="list-style-type: none"> 답변: <ol style="list-style-type: none"> □□□를 사용하여 해당 트랙을 직접 추적하십시오. 'Engagement Channel Recommend' 창에서 [△△△-FC1(□□□)] 또는 [◆◆◆] 채널을 택하여 할당(Alloc)하십시오. △△△ 사용 시, 'Fire Authorization' 창에서 사격 승인(Fire Authorize)을 완료하십시오. ▽▽▽ 대응이 필요한 경우 ▽▽▽ 콘솔에서 직접 Lock-on 및 발사를 수행하십시오. <p>이상 조치 요망.</p> <ul style="list-style-type: none"> 영역별 점수: 3점 환각: 3점이므로 환각 발생 후보군 선정 및 전문가 평가에 의해 환각 발생 판정 최초 답변 출력 시간: 22.4 전문가 평가: RAG를 통해 ○○○의 제약 사항(고도 부재)과 선행 절차(□□□ 추적) 교범이 주입되었음에도, 파인튜닝 단계에서 굳어진 지식이 압도하여 왜곡된 절차 생산

실험 아키텍처 중 일부 발췌

4.2.3 Underperformance in Vision-Based Situation Understanding

모든 아키텍처에서 복합적 추론 능력이 요구되는 'Situation Understanding' 부분의 성능이 상대적으로 낮게 관측되었다. 이러한 현상의 요인은 다음으로 유추된다.

- 핵심 전술 정보에 대한 식별력 부족: 아키텍처 전반에서 입력으로 주어진 전술 화면 내의 핵심 콘텐츠를 명확히 식별하지 못하는 한계가 드러났다. 이로 인해 모델이 전술

적 우선순위가 낮은 부차적인 정보를 위주로 답변을 출력하는 현상이 발생하였으며, 이는 전술적 핵심을 정확히 포착하는 선별력이 미흡했음을 시사한다.

- 추론형 학습의 부재: 전술 상황 판단은 단순한 시각적 인식을 넘어 전술적 인과 관계를 이해해야 하는 영역이다. 그러나 본 실험의 아키텍처는 이러한 고차원적 맥락을 파악할 수 있는 추론형 사전 학습 기법이 충분히 적용되지 못해 성능 한계를 나타낸 것으로 판단된다.

4.2.4 Delay in Time to First Token Due to RAG Implementation

RAG 결합 모델은 정확도 향상이라는 성과와 달리, 최초 답변 출력 시간이 평균 3초대에서 24초대로 지연되는 한계가 관측되었다. 이러한 응답 지연 현상은 다음과 같은 복합적 요인이 작용한 결과로 추정된다.

- 다단계 인출 알고리즘의 복잡성: RAG 모델은 [질의 임베딩 → 벡터 검색 → 유사 문서 추출 → 본문 재정렬 및 프롬프트 재구성]의 전처리 과정을 필수적으로 거친다. 내부 파라미터(Parametric Memory)만을 활용하여 즉각적으로 답변을 생성하는 Case 1 및 Case 3과 달리, 외부 데이터에 접근하고 연산을 수행하는 과정에서 물리적인 지연 시간이 추가된 결과로 풀이된다.

- 컨텍스트 확장으로 인한 연산 부하: 검색된 대용량 매뉴얼 문헌이 입력 프롬프트에 추가됨에 따라 전체 토큰(Token) 수가 급격히 증가하였다. 입력 토큰의 확장은 최초 응답 생성을 위한 연산량을 기하급수적으로 증가시켜, Fig. 8에서 환각 제어율과 평균 응답 속도 간의 명확한 트레이드오프(Trade-off) 관계가 존재함을 시사한다.

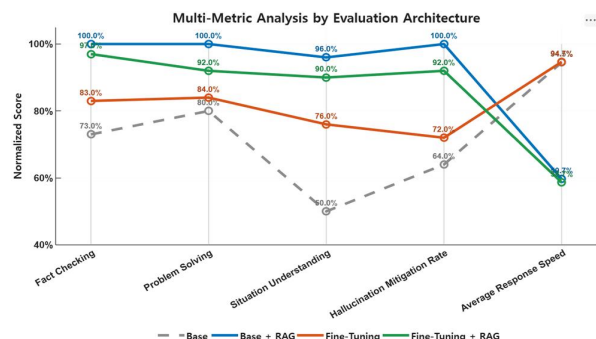


Fig. 8. Evaluation Results of Normalized Metrics by Architecture.

V. Conclusions

본 연구는 보안이 유지되는 폐쇄망 환경에서 오픈소스 소형 언어 모델(sLLM)인 Qwen3.8-27B 모델을 기반으로 4가지 아키텍처 조합(Base, Base+RAG, Fine-Tuning, Fine-Tuning+RAG)을 구축하고 성능을 비교 실증하였다.

실험 결과, 본 연구와 같이 제한된 도메인 매뉴얼과 한정된 데이터셋 환경에서는 RAG를 도입한 아키텍처가 외부 매뉴얼을 정밀하게 탐색하여 상대적으로 전술적 안정성과 신뢰성이 높은 답변을 생성하였다. 반면, Fine-Tuning+RAG 조합에서는 내부 지식 간의 상호 간섭 등의 요인으로 인해 오히려 환각 발생률이 약 8%p 상승하는 부작용이 관찰되었다. 이는 본 실험에 사용된 한정된 데이터 규모 내에서는 Fine-Tuning 적용에 의한 개선 효과가 검색 기반 방식에 비해 제한적일 수 있음을 시사한다. 물론 RAG 결합 아키텍처는 실시간 검색 연산 및 대용량 문맥 처리에 따른 부하 등의 요인으로 최초 답변 출력 시간(TTFT)이 증가하는 시간적 페널티가 나타났지만 그럼에도 국방 특화 모델의 자체 추론 능력이 정교화되기 전까지는 정보의 타당성을 보장하는 RAG 메커니즘을 결합하는 것이 현실적인 대안이 될 수 있다.

다만, 장기적인 관점에서 해상 전투는 단 몇 초의 지연으로도 함정의 생존을 위협받을 수 있다. 따라서 향후 실전 배치 및 전력화를 고려했을 때, 검색 지연 없는 고속 추론 실현이 필요하다. 그러기 위해서 시스템 부하의 정량적 분석이 필요하고 정교한 전술 데이터셋을 구축해야 한다. 이를 바탕으로 모델 스스로 논리적 사고를 빠르게 수행하도록 돕는 '추론 파인튜닝(Reasoning Fine-Tuning)' 기술과 RAG 메커니즘을 유기적으로 융합·발전시켜, 즉각적이고 깊이 있는 전술 판단 능력을 내재화한 고신뢰성 모델을 확보하는 연구가 필요하다.

REFERENCES

- [1] J. W. Ok, J. S. Lee, and I. S. Han, "Design and Operational Scenarios of LLM Systems for Closed-network Naval Platforms," *Journal of Naval Science and Technology*, Vol. 8, No. 3, pp. 474-480, September 2025. DOI: 10.31818/JKNST.2025.9.8.3.474
- [2] D. H. Kim, J. H. Kim, and J. H. Jeon, "Combat experiment analysis of common situational understanding operation and weapon recommendation performance in AI command decision support systems," *Advanced Defense Studies*, Vol. 9, No. 1, pp. 1-21, April 2026. DOI: 10.37944/jams.v9i1.296
- [3] D. H. Kim, S. J. Kim, H. B. Kim, J. H. Park, S. E. Yuk, Y. H. Chae, and O. S. Hong, "Domain-Specific (Defense) Generative AI Framework," *Proceedings of the Korea Information Science Society Conference*, pp. 1420-1422, Jeju, Korea, December 2024.
- [4] S. H. Jeong and H. J. Kim, "A Study on the Application Plan of Korean LLM-Based Command Decision Support System for Mounting on Large Ships," *Journal of the Korea IT Policy and Management Society*, Vol. 17, No. 3, pp. 120-135, September 2025.
- [5] S. W. Jeong, "A Case Study on Building a Defense Data Retrieval AI Using RAG," *Journal of Convergence Security*, Vol. 26, No. 2, pp. 135-142, March 2026. DOI: 10.33778/kcsa.2026.26.2.135
- [6] H. S. Lee, D. H. Kim, H. S. Kim, and J. S. Yoo, "Preference-Aligned sLLM for Safe and Helpful RAG-Based Battlefield Analysis System," *Journal of the Korea Society of Computer and Information*, Vol. 30, No. 6, pp. 31-46, June 2025. DOI: 10.9708/jksci.2025.30.06.031
- [7] H. Li, D. Li, Y. Liu, J. MA, G. Wang, and J. Ai, "LLM-Driven Modeling and Decision Support Methods for Cross-Domain Collaborative Mission Systems," *Applied System Innovation*, vol. 9, no. 4, p. 80, Apr. 2026. DOI: 10.3390/asi9040080
- [8] L. Vasankari and A. Koski, "GenAI in the Military: Trends and Opportunities," *Scandinavian Journal of Military Studies*, vol. 8, no. 1, pp. 416-434, Nov. 2025, DOI: 10.31374/sjms.415.
- [9] J. Chen et al., "DRAG: Distilling RAG for SLMs from LLMs to Transfer Knowledge and Mitigate Hallucination via Evidence and Graph-based Distillation," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025)*, 2025. DOI: 10.48550/arXiv.2506.01954.
- [10] W. Cho, J. Yoo, S. M. Kim, and J. Jang, "RAG-Enhanced small Large Language Models: Enhancing Battlefield Analysis through Knowledge Distillation," *Journal of The Korea Society of Computer and Information*, vol. 30, no. 3, pp. 43-57, Mar. 2025, DOI: 10.9708/jksci.2025.30.03.043.
- [11] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient Finetuning of Quantized LLMs," *arXiv preprint arXiv:2305.14314*, 2023, DOI: 10.48550/arXiv.2305.14314.
- [12] D. Kalajdziewski, "A Rank Stabilization Scaling Factor for Fine-Tuning with LoRA," *arXiv preprint arXiv:2312.03732*, 2023. DOI: 10.48550/arXiv.2312.03732.

Authors



Donghwi Kim received the B.S. degree in Electronic Engineering from Inha University, Incheon, Korea, in 2014. He currently works for the Naval R&D Naval System 4 Team at Hanwha Systems, Korea.



SunYeong Jeong received the B.S. degree in Electronic Engineering from Kyungpook National University, Daegu, Korea, in 2014. She currently works for Naval R&D Naval System 4 Team at Hanwha Systems, Korea.



JongWook Jang received the B.S. degree in Electrical and Electronic Engineering from Yonsei University, Seoul, Korea, in 2021. He received the M.S. degree in Electrical and Electronic Engineering from Yonsei

University, Seoul, Korea, in 2025. He currently works for Naval R&D Naval System 4 Team at Hanwha Systems, Korea.



WooYoung Kim received the B.S. degree in Electronic Engineering from InJe University, Gimhae, Korea, in 2003. He currently works for Naval R&D Naval System 4 Team at Hanwha Systems, Korea.