Journal of the KNST

2025; Vol.8, No.3; pp.474-480

ISSN: 2635-4926 https://doi.org/10.31818/JKNST.2025.9.8.3.474

Check for updates

Received: 2025/07/10
Revised: 2025/07/24
Accepted: 2025/09/10
Published: 2025/09/30

*Corresponding Author:
Jiwon Ock
Tel: +82-2-3400-2585
Fax: +82-2-403-3512
E-mail: jwock@add.re.kr

Abstract

본 논문에서는 해군 함정과 같은 폐쇄망 환경에서 운용 가능한 LLM 시스템 구축 방안을 제안한다. 제한된 연산 자원과 네트워크 단절 조건을 고려하여, MCP(model context protocol)와 RAG(retrieval-augmented generation) 기술을 통합한 아키텍처를 설계하고, 실시간 문서 기반 질의응답이 가능한 프로토타입 시스템을 구현하였다. 또한, 해군 내 운용 시나리오를 기반으로 LLM의 적용 가능성을 확인하였다. 오픈소스 모델 활용을 위한 GPU 자원 분석으로 제한된 하드웨어를 탑재한 함정에서 LLM을 도입하게 되었을 때 참고지표로 활용할 수 있다.

This study proposes an LLM system for closed network environments like Navy ships, addressing limited hardware and no-network conditions. By combining MCP and RAG techniques, a prototype for real-time, document-based Q&A was developed and validated through a Navy operational scenario. Additionally, by analyzing GPU resource requirements of open-source LLMs, we provide reference benchmarks for implementing such systems on ships with constrained hardware capabilities.

Keywords

함정 폐쇄망(Closed-Network of Battleship),
거대 언어모델(Large Language Model),
모델 컨텍스트 프로토콜(Model Context Protocol),
검색 증강 생성(Retrieval-augmented Generation),
해군 운용 시나리오(Navy's Operational Scenarios)

Acknowledgement

이 논문은 2025년도 한국해군과학기술학회 하계학술대회 발표 논문임.

함정 폐쇄망 조건을 고려한 LLM 시스템 구축과 운용 시나리오

Design and Operational Scenarios of LLM Systems for Closed-network Naval Platforms

옥지원[1*], 이정식[2], 한인성[2]

[1]국방과학연구소 제3기술연구원 연구원
[2]국방과학연구소 제3기술연구원 책임연구원

Jiwon Ock[1*], Jungsik Lee[2], Insung Han[2]

[1]Researcher, 3rd R&D Institute, Agency for Defense Development
[2]Principal researcher, 3rd R&D Institute, Agency for Defense Development

1. 서론

최근 거대 언어모델(large language model, LLM)의 비약적인 발전은 자연어 처리(natural language processing, NLP) 기술의 수준을 획기적으로 향상시키며, 인간과 유사하게 텍스트 생성, 이해, 요약, 번역 기능을 가능하게 한다. 이러한 기술적 진보는 민간 및 군사 영역에서도 그 응용 가능성을 제기하고 있다. 특히, 해군 함정과 같이 외부와 물리적으로 단절된 폐쇄망 환경에서도 LLM 기술은 작전정보 분석, 기술문서 질의응답, 다국어 통신 지원 등 다양한 지능형 임무 수행에 기여할 수 있는 잠재력이 크다[1].

미국 해군은 2023년 8월부터 아멜리아(Amelia)라는 인공지능 챗봇 서비스를 시범 배치하였다[2]. C4I 체계와 인사, 교육 등 미 해군과 해병대의 업무용 IT 체계를 지원하는 챗봇이며, 인가된 장병들은 스마트폰, 태블릿, 함선 컴퓨터를 통해 상호작용할 수 있다. 각종 무기체계의 운용 매뉴얼과 정비에 대한 데이터베이스와 연결되어 장병들이 육성, 메시지, 이메일 등으로 질문을 하면 답변을 받는 형태이다. 전체 서비스가 아마존 웹 서비스에서 호스팅하는 정부 클라우드 환경에 위치하여 장소와 시간에 구애받지 않고 액세스할 수 있지만, 대부분의 LLM 기반 시스템은 대규모 클라우드 인프라와 인터넷 연결을 전제로 설계되어 있다. 즉, 해군 함정과 같은 제한적 환경에 직접 적용하기에는 기술적 제약이 따른다. 해군 함정은 작전 중 외부와 단절된 전술망 또는 완전한 폐쇄망 상태로 운

용되며, 네트워크 연결 불가, 전력 및 자원 제한, 보안 위협에 대한 엄격한 통제 등의 제약을 가진다. 따라서 LLM 기반 시스템 도입 시 상술한 내용을 포함한 여러 사항을 고려해야 한다.

실제 미 해군 함정 콜롬비아호(SSN-771)와 토페카호(SSN-754)에 승선한 해군 장교의 연구에서는 군함이라는 제한적인 환경의 한계점을 제시하고 있다[3]. 함정의 경우 독립적인 LLM 운용을 위한 GPU 서버의 물리적 탑재 공간이 부족할 수 있다. 따라서 실제 운용 가능한 GPU 서버에 따라 LLM 모델의 파라미터 규모와 시스템 구성에 대한 연구가 필요하다. 이러한 필요성뿐만 아니라 폐쇄망 내에서 군사 기밀 데이터를 다루는 LLM 시스템은 높은 수준의 보안성과 자원 효율성을 모두 확보해야 한다.

본 논문에서는 해군 폐쇄망 환경에 적합한 LLM 시스템을 제안한다. 이를 위하여 활용 가능한 GPU 요구사항을 정의하고, 자원 효율적이고 보안성이 강화된 운용 방안을 제안한다. 특히, RAG(retrieval-augmented generation, 검색 증강 생성)와 MCP(model context protocol, 모델 컨텍스트 프로토콜) 프레임워크 구조를 접목하여 실시간 문서 검색 및 질의응답 기능을 폐쇄망 내부에서 구현할 수 있는 지능형 정보처리 시스템의 프로토타입 아키텍처를 설계하였다. 해군에서 LLM 시스템 도입으로 발생할 수 있는 시나리오에 대해 예상하며 본 연구가 도움이 될 수 있음을 시사한다.

2. 배경 지식

본 장에서는 해군 폐쇄망 환경에서 활용될 수 있는 LLM 관련 최신 기술인 MCP 및 RAG에 관련하여 설명한다.

2.1 모델 컨텍스트 프로토콜(MCP)

MCP는 다양한 인공지능 모델들이 서로 정보를 주고받는 규칙이다[4]. LLM을 사용할 때 대화 모델, 추천 모델, 이미지 분석 모델 등이 동시에 운용될 수 있다. 이때 각 인공지능 모델이 데이터베이스, 파일시스템과 같은 외부 기능과 페어링할 때마다 $m \times n$개의 맞춤형 통합 환경을 만들어야 한다는 한계점을 가진다. MCP를 사용한다면 인공지능 모델들은 MCP의 클라이언트 측을 1번, 외부 데이터 소스들은 MCP의 서버 측을 1번 구현하여 $m+n$개의 방법으로 기존의 복잡성과 유지보수 문제를 해결할 수 있다. 즉, 모델 A가 사용자와의 대화에서 얻은 정보를 모델 B, 모델 C와 공유하여 사용자에 대한 정보, 사용자의 세션 상태 등의 컨텍스트 데이터를 여러 모델에 전달할 수 있다. 결과적으로 MCP는 여러 모델 간 협업이 원활해지도록 도와주는 통신 규칙 역할을 하며 멀티 에이전트 시스템이나 다양한 인공지능 기능이 연결된 복합 서비스에서 유용하게 사용되고 있다.

Fig. 1에서 볼 수 있듯이, MCP는 서버-클라이언트 아키텍처를 따른다. MCP 호스트가 필요한 도구가 있을 때마다 MCP 서버에 연결된다. 그다음 MCP 서버는 데이터 소스들에 연결되며, MCP 호스트와 서버는 MCP 프로토콜을 통해 서로 연결된다.

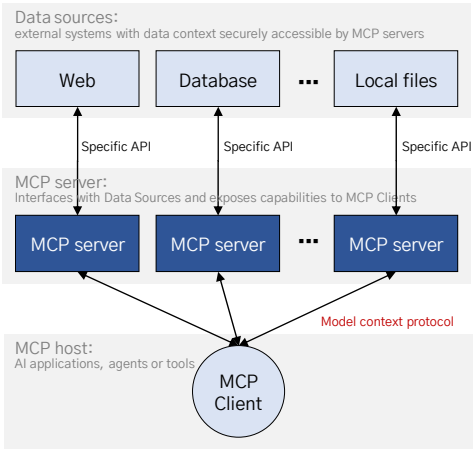

Fig. 1. Architecture about model context protocol

주요 컴포넌트는 MCP 호스트, MCP 클라이언트, MCP 서버, MCP 프로토콜로 구성되어 있으며 이들에 대한 설명은 Table 1에 정리하였다.

Table 1. MCP components

Components	Descriptions
MCP host	LLM application that require access to external features
MCP client	Protocol client within the host application that maintains a 1:1 connection to the server
MCP server	Programs that expose specific features and provide context, tools, and prompts to clients, respectively.
MCP protocol	Intermediate transport layer (JSON-RPC2.0)

2.2 검색 증강 생성(RAG)

기존 LLM은 거짓 정보를 진짜인 것처럼 제공하는 할루시네이션(hallucination) 문제가 존재하며, 폐쇄적 환경에서는 최신 지식을 반영하지 못한다는 한계점이 있다. 이를 극복하기 위해 RAG 구조는 모델 자체의 지식과 신뢰할 수 있는 외부 데이터를 결합하여 보다 정확한 결과를 도출해낼 수 있다[5].

Fig. 2는 RAG의 동작 과정을 요약하여 보여준다. ① 사용자의 질문이 벡터로 인코딩되어 vector DB에 전송된다. ② Vector DB는 내부 콘텐츠에서 관련 정보를 검색한다. ③ 검색된 결과 중 관련도가 높은 정보가 생성형 AI의 입력으로 전달된다. ④ 생성형 AI는 검색된 지식을 활용하여 사용자 질문의 답변을 생성한다.

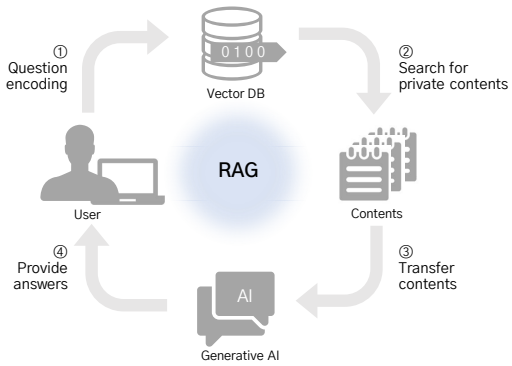

Fig. 2. RAG operation process

3. 폐쇄망 해군 함정 환경에서의 LLM 기반 지능형 정보처리 시스템 구축 방안

폐쇄망 상태의 함정 내 LLM 기반 시스템을 도입하기 위해 다음 세 가지 사항을 고려해야 한다. (1) 인터넷 기반 API 및 클라우드 서빙이 불가능하다. (2) 외부로부터의 모델 및 데이터 이관은 반드시 오프라인 방식이어야 한다. (3) GPU 서버 탑재 시 공간 및 전력 효율성을 고려하여야 한다.

3.1 폐쇄망 함정 환경에서의 LLM 구축 방안

제안하는 LLM 시스템은 크게 승조원 질의 인터페이스, 검색 기반 문서 처리 모듈(RAG), LLM 추론 서버, 결과 응답 시스템으로 구성된다. 모든 구성요소는 로컬 폐쇄망 내부에서 동작하며, 문서 데이터는 별도의 NAS 또는 데이터베이스에 저장된다.

MCP 기반 에이전트는 각기 다른 기능(예: 기술문서 분석, 함정 운용지침 요약, 다국어 번역, 로그 분석 등)을 담당하며, 승조원 질의에 따라 역할을 자동 배정받는다. 이를 통해 복합 명령 처리 및 임무별 정보 응답이 가능하다.

안전한 운용을 위해 보안성 측면도 고려해야 한다. 외부에서 확보한 LLM 모델 및 문서는 SHA-256 해시 검증을 통해 무결성을 확인해야 하며, 바이러스 검사, 승인된 USB 장비를 통한 오프라인 이관 절차를 통해 유입할 수 있다. 모든 이관 기록은 감사 로그로 저장되며, 모델 및 데이터 이관 절차는 자동화된 검증 프로세스를 통해 이루어져야 한다. 시스템 사용자는 역할 기반 접근 제어(roll-based access control, RBAC)로 구분되며, 중요 작전 자료에 대한 질의는 인증된 권한 하에서만 허용한다. 사용자 요청 로그는 중앙 감사 서버에 주기적으로 백업되어야 한다.

Fig. 3는 폐쇄망 환경에서 운용되는 LLM 기반 질의응답 시스템의 구조를 나타낸다. 시스템은 사용자 인터페이스를 통해 해군 담당자가 질의를 입력하는 것으로 시작된다. 입력된 질의는 MCP 서버를 통해 프롬프트 파서와 컨텍스트 빌더 모듈에서 전처리되며, 이후 관련 문서를 검색한다. 사전에 임베딩된 문서를 보유한 안전한 저장소와 연동되며 retrieval 모듈을 통해 질의에 적합한 정보를 추출한다. 이렇게 구축된 컨텍스트는 폐쇄망 내에서 운용되는 LLM에 전달되어 응답을 생성하게 되며 담당자에게 반환된다.

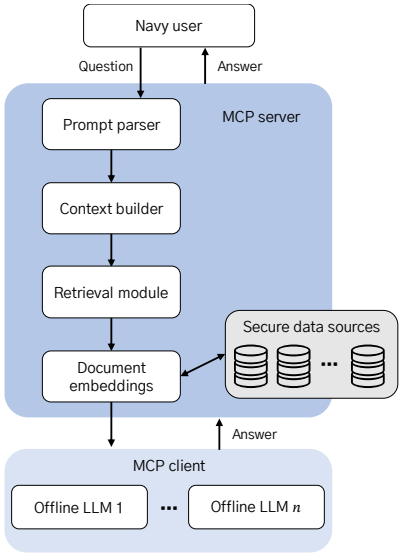

Fig. 3. The proposed navy's LLM system

3.2 GPU 자원 사용량 분석

함정 전력 운용 관점에서 GPU의 개수에 따라 효율적인 추론이 수행되어야 한다. 효율적인 LLM 추론은 GPU 자원을 요구하므로, 모델 크기와 하드웨어 자원 간의 균형이 중요하다. 이에 따라 최근에는 인공지능 모델들의 파라미터 수가 다양해지고, 허깅페이스(Hugging Face)[6]라는 인공지능 플랫폼을 통해 베이스 모델을 공유받아 파인튜닝에 이용하는 패턴이 보편화되고 있다. 최신 GPU는 수요가 높아 비용이 크게 들거나 구매하여 입수하기까지 오랜 시간이 걸린다[7]. 따라서 GPU의 클러스터 최적화를 통해 성능을 향상시킬 수 있어야 하므로, 각 모델별 GPU 요구사항에 대하여 파악할 필요가 있다. 파인튜닝이나 추론 시 어떤 GPU를 사용하느냐에 따라 필요한 시간과 전력 소비량을 도출하는 데에 도움을 줄 수 있다.

다음으로 LLama, Gemma, Mistral, Qwen 등 2025년 6월 시점에서 주로 활용되는 오픈소스 모델의 최신 버전에 대하여 다룬다. Llama는 Meta에서, Gemma는 구글 딥마인드에서, Mistral은 MistralAI에서, Qwen은 알리바바에서 개발한 최신 LLM 모델이다[8-11]. Table 2는 파라미터별로 필요한 GPU를 가장 많이 사용되는 엔비디아 GPU 기준으로 나타낸 것으로, 사용 분야 및 가용 비용에 따라 기재된 GPU 뿐만 아니라 비슷한 사양의 다른 GPU 제품을 선택할 수 있다. 함정 내 탑재 가능한 GPU 개수와 전력에 맞추어 모델을 선택하여 LLM 시스템에 탑재하여 활용할 수 있다. 모델을 8비트, 4비트 등으로 양자화한다면, VRAM 사용량이 크게 줄어들어 더 작은 GPU에서도 실행이 가능하다. 하지만 양자화는 메모리 사용량을 줄이지만 특정 프로세스에서 속도나 정확도 저하를 발생시킬 수 있으므로 주의해야 한다.

Table 2. GPU requirements based on model parameters

Models	Parameters*	VRAM requirements	Recommended GPU
Llama 3.3	70B	161 GB	A100 80 GB×2
Gemma 3	1B	23 GB	GTX 1650 4 GB
	4B	92 GB	RTX 3060 12 GB
	12B	27.6 GB	RTX 5090 32 GB
	27B	62.1 GB	RTX 4090 24 GB×3
Mistral	7B	24 GB	RTX 3090
Qwen 3	1.7B	18 GB	RTX 3060/A5000
	4B	32 GB	A100 40GB/RTX 3090
	8B	60 GB	A100 80GB/H100
	14B	80 GB	A100 80GB×2/H100
	32B	160 GB	A100×4/H100×2

*B: billion

3.3 폐쇄망 LLM 시스템 구축 실험

해군 함정 내 LLM 시스템의 적용 가능성을 확인하기 위해 폐쇄망에 LLM 시스템을 구축하였다. Nvidia GeForce RTX 3090 8개의 GPU가 탑재된 환경에서 구축하였으며, ollama에서 제공하는 오픈소스 라이브러리를 활용하여 폐쇄망 LLM 서버와 클라이언트를 파이썬 언어로 개발하였다. 허깅페이스 오픈소스 플랫폼에서 사용할 모델 관련 파일들(safetensors, config 등)을 다운로드하여 폐쇄망으로 안전하게 압축 전송하고, 모델을 GGUF 형식으로 변환하여 폐쇄망에 LLM 시스템을 구축할 수 있다. 다만, 실제 군용 LLM 시스템을 구축하기 위해서는 오픈소스의 라이센스나 사용 정책 등을 고려하여 LLM을 선택해야 한다[1].

하나의 예시로 개발한 폐쇄망 LLM 시스템에 구글에서 개발한 gemma-3-27b 모델을 포팅하였다. 해당 모델을 사용하였을 때 Fig. 4는 LLM 서버의 UI를, Figs. 5-6은 LLM 클라이언트의 CLI(command line interface) 및 WebUI(Web user interface) 방식 UI를 각각 보여준다. 폐쇄망에서의 LLM 시스템 구축을 통해 제안한 LLM 시스템의 가능성을 보여주며, 해군의 데이터로 파인튜닝된 MCP와 RAG 기술 기반 LLM 시스템이 구축된다면 향상된 성능을 기대할 수 있다.

4. 해군 함정 내 LLM 시스템 운용 시나리오

제안한 LLM 시스템이 해군 함정 내에서 실제 운용될 수 있는 시나리오 사례를 구성하였다. 본 장에서는 네 가지 상황 중심으로 LLM 시스템의 입력 질의와 동작 과정을 설명한다. 전투정보실(CIC), 기관부, 작전장교, 통합사이버방호체계 등의 실제 업무 흐름을 반영한다면 실효성과 적용 가능성이 향상될 것이다.

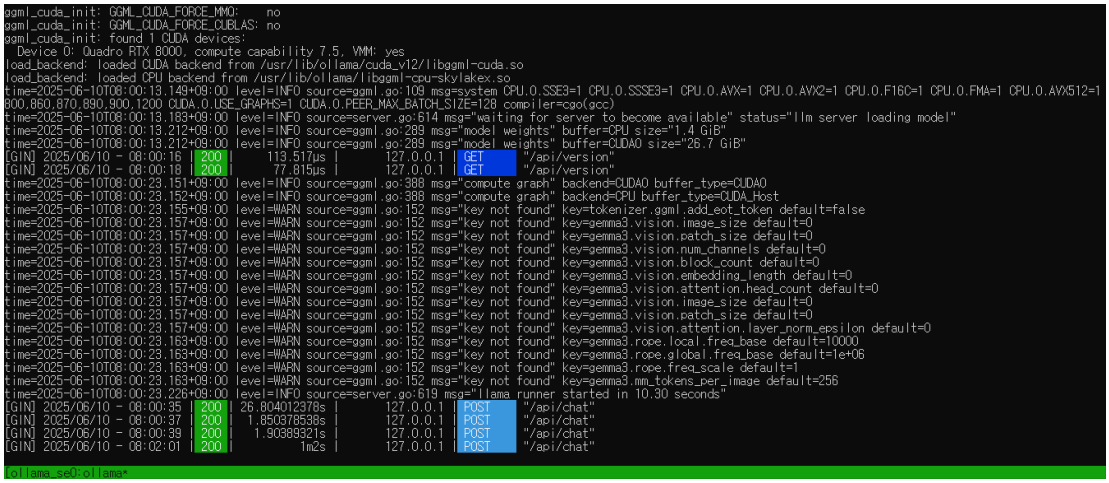

Fig. 4. LLM server

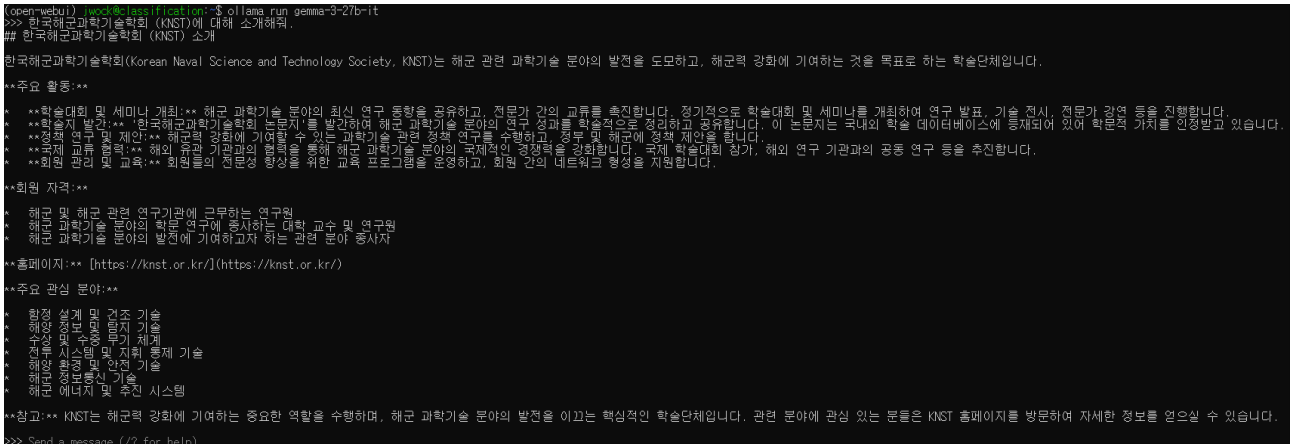

Fig. 5. LLM client using the CLI

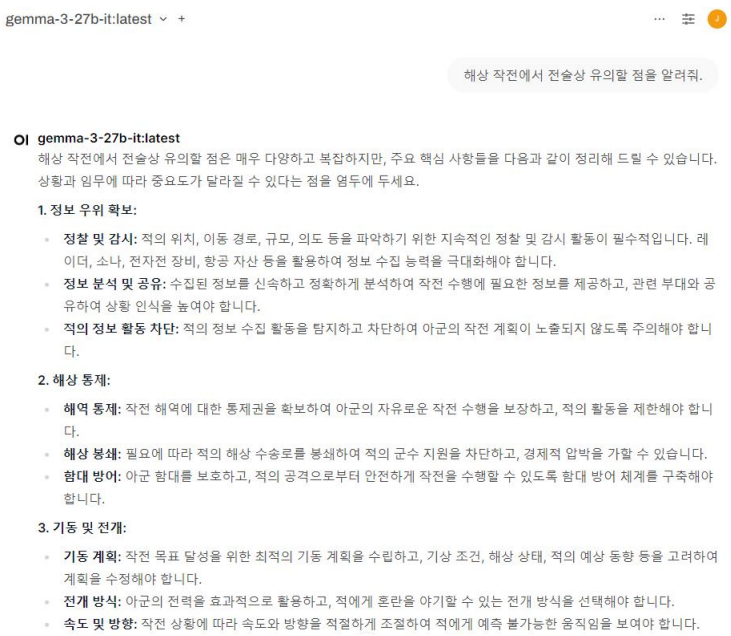

Fig. 6. LLM client using the WebUI

첫 번째 시나리오는 전투정보실이 적 함정 신호 식별 및 대응 전술에 대해 질의하는 상황이다.

> 질의: "지금 수신한 □ Hz 대역 대기 신호가 2020년 이후 해군 ○○단 통신 사례와 유사한지 분석해줘. 비슷한 사례가 있다면 대응 전술도 알려줘."

질의를 접수한 LLM 시스템의 RAG가 전술 데이터베이스, 관련 보고서, 작전 매뉴얼에서 관련 문서를 검색한다. 검색된 유사 사례(예: 2021년 ○○도 해역 통신 패턴)와 매칭 분석을 수행한 다음, 결과를 바탕으로 위협 수준과 표준 대응 절차(예: SOP-25)에 따른 전파탐지 대응법 등을 요약 및 제시할 수 있다.

두 번째 시나리오는 함정 기관부의 사관이 장비 이상을 진단하고 부품 수급을 확인하는 질의이다.

> 질의: "가스터빈 3번 축이 비정상 진동을 보여. 과거 진동 주기 180 Hz 문제가 있었던 고장 사례를 찾아줘."

LLM 시스템은 과거 정비 기록, 국방 기술보고서, 제조사 매뉴얼을 검색한다. 유사 진동 주파수 고장 사례를 매칭하고 원인을 진단한다. 다음으로 가능성 높은 원인을 바탕으로 대체 부품 재고 위치, 수리 방법 및 소요 시간을 제시한다.

세 번째 시나리오는 작전장교가 다국어 연합작전 상황 브리핑 요약에 대한 질의이다.

> 질의: "□□에서 온 브리핑 파일 내용을 한국어로 요약해줘. 전술상 유의할 점은 따로 알려줘."

다국어로 파인튜닝된 LLM은 각 나라 언어에 맞추어 문서를 자동 요약 및 번역할 수 있다. 이때 전술 절차, ROE 등의 작전 관련 단어 중심으로 요약하고, 위협 요소 및 권고사항을 강조하여 표시해서 결과를 출력한다.

네 번째 시나리오는 함정 사이버 담당관이 랜섬웨어 경고에 대응하는 질의이다.

> 질의: "통합함교체계 단말에 랜섬웨어가 설치되어 있다고 방호체계에서 경고가 발생했어. 랜섬웨어라고 판단한 근거를 설명해 주고, 통합함교체계 운영에 영향이 최소화될 수 있는 대응 절차를 단계적으로 상세한 설명과 함께 제시해줘."

LLM 시스템은 통합 사이버방호체계의 경고 근거가 되는 단말 로그를 분석하여, 과거 랜섬웨어의 유사도와 체계에 미치는 영향을 바탕으로 통합사이버방호체계가 랜섬웨어로 탐지한 이유를 설명한다. 이때 과거 대응 방안, 국방 보안업무 훈령지침 등을 검색하여 체계 운영에 영향이 적은 대응 절차를 상세한 설명과 함께 제시한다.

이처럼 제안한 시스템은 실전 작전 중 실시간 대응이 가능하고, 공식 작전자료 기반의 추론으로 오판 위험을 최소화할 수 있다는 전술적 이점을 가진다. 또한 다기능 agent 운용으로 반복 질의를 자동화하며 정보를 효율적으로 요약할 수 있다. 또한 본 시스템으로 외부 연결 없이 함정 내 단독 운용이 가능하여 보안을 내재화한 폐쇄망 운영에 최적화할 수 있을 것이다.

5. 결론

본 논문에서는 해군 함정과 같은 폐쇄망 환경에서도 운용 가능한 LLM 시스템의 요구사항을 도출하고, MCP 및 RAG 기반의 아키텍처를 통해 보안성과 GPU 자원 효율성을 포함한 운용 방안을 제안하였다. 제안된 LLM 시스템은 전술적인 상황에 따라 문서 기반 실시간 질의응답을 가능하게 하여 군사 정보 처리를 자동화·지능화할 수 있다. 또한 승조원의 역할을 대체하여 인원 부족 문제에 기여할 수 있다. 향후 인공지능 최신 기술이 해군 내에 안전하고 효과적으로 활용될 수 있도록 추가적인 연구가 필요하다.

참고문헌

[1] Kapiamba, S., Fouad, H., and Moskowitz, I. S., "Responsible Integration of Large Language Models (LLMs) in Navy Operational Plan Generation," Proceedings of the AAAI Symposium Series, Vol. 3, No. 1, 2024, pp. 50–53.
[2] Lee, S. U., Lee, S. H., Yu, H. C., Park, S. S., and Kim, U. H., "Analyzing and Addressing Challenges of LLM Applications in Defense," Journal of Defense and Security, Vol. 6, No. 1, 2024, pp. 302–328.
[3] Jim Halsell, "Generative Large Language Models of Navy Crews," U.S. Naval Institute Proceeding, Vol. 151, No. 468, 2025.

[4] Xinyi, H., Yanjie, Z., Shenao, W., Haoyu, W., "Model Context Protocol (MCP): Landscape, Security Threats, and Future Research Directions," arXiv preprint arXiv:2503.23278, 2025.
[5] Lewis, Patrick, et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks," Advances in neural information processing systems, Vol. 33, 2020, pp. 9459-9474.
[6] Hugging Face, https://huggingface.co/, Accessed June 9, 2025.
[7] L. Tuggener, P. Sager, Y. Taoudi-Benchekroun, B. F. Grewe and T. Stadelmann, "So You Want Your Private LLM at Home?: A Survey and Benchmark of Methods for Efficient GPTs," 2024 11th IEEE Swiss Conference on Data Science (SDS), Zurich, Switzerland, 2024, pp. 205-212.
[8] Meta, "The llama 3 hred of models," arXiv preprint arXiv:2407.21783, 2024.
[9] Google Team, "Gemma3 technical report," arXiv prepring arXiv:2503.19786, 2025.
[10] Mistral Ai, "Mistral 7B," arXiv preprint arXiv:2310.06825, 2023.
[11] Qwen Team, "Qwen technical report," arXiv preprint arXiv:2309.16609, 2023.