

RAG-Enhanced small Large Language Models: Enhancing Battlefield Analysis through Knowledge Distillation of Large Language Models

Wonjun Cho*, Jaesung Yoo**, Sang-Min Kim***, Jaeun Jang****

*Team Leader, Future Technology Team, Hanwha Systems, Seongnam, Korea

**Chief Researcher, Future Technology Team, Hanwha Systems, Seongnam, Korea

***Senior Researcher, Future Technology Team, Hanwha Systems, Seongnam, Korea

****Researcher, Future Technology Team, Hanwha Systems, Seongnam, Korea

[Abstract]

The increasing complexity of modern battlefields and the importance of real-time data processing have heightened the need for effective battlefield situation analysis systems. This study proposes a battlefield analysis system utilizing Large Language Models (LLMs), specifically introducing an advanced approach that combines Retrieval-Augmented Generation (RAG) with Supervised Fine-Tuning (SFT). To address the hallucination problems and lost-in-the-middle phenomenon inherent in existing RAG systems, we introduce a triple-structured learning approach that incorporates reference documents in the SFT process. Based on synthetic battlefield datasets developed in collaboration with military experts, our experimental results demonstrate exceptional performance in source extraction accuracy and response quality evaluation. Notably, when applying triple-structured SFT to an 8B parameter model, we achieved comparable performance to a 405B parameter model, proving its practicality in actual battlefield environments. Furthermore, our lightweight model enhanced with specialized training strategies showed minimal performance degradation compared to larger models, suggesting its viability for deployment in resource-constrained environments. This research demonstrates the effective application of LLMs in battlefield situation analysis and presents a novel direction for military domains requiring real-time data processing and high reliability.

▶ **Key words:** Large Language Models, Retrieval-Augmented Generation, Supervised Fine-Tuning, Prompt Chaining, Battlefield Situation Analysis

-
- First Author: Wonjun Cho, Corresponding Author: Jaeun Jang
 - *Wonjun Cho (wonjun78.cho@hanwha.com), Future Technology Team, Hanwha Systems
 - **Jaesung Yoo (mars20@hanwha.com), Future Technology Team, Hanwha Systems
 - ***Sang-Min Kim (smkim0153@hanwha.com), Future Technology Team, Hanwha Systems
 - ****Jaeun Jang (wkdwodms0779@hanwha.com), Future Technology Team, Hanwha Systems
 - Received: 2025. 02. 10, Revised: 2025. 02. 25, Accepted: 2025. 03. 12.

[요 약]

복잡하고 다변화되는 현대 전장 환경에서 실시간 수집되는 데이터 처리 신속성과 신뢰성이 중요한 이슈가 되고 있으며, 이를 위해 효과적인 전장 상황 분석 시스템의 필요성이 대두되고 있다. 본 연구에서는 대규모 언어 모델(LLM)을 활용한 전장 상황 분석 시스템을 제안하며, 특히 검색 증강 생성(RAG)과 지도 학습 미세조정(SFT)을 결합한 고도화된 접근 방식을 제시한다. 기존 RAG 기법의 환각 문제와 중간 정보 손실 현상을 해결하기 위해, 본 연구는 3중 구조 SFT 방식을 도입하였다. 군사 전문가들과 협력하여 구축된 합성 전장 데이터를 기반으로 실험을 진행한 결과, 3중 구조 SFT를 적용한 8B의 경량화 모델이 405B 모델을 Source Extraction에서 F1 기준 0.07, Answer Quality에서는 5점 척도 기준 대략 1점 정도 능가하는 성능을 보여주어 실제 전장 환경에서의 실용성을 입증하였다. 본 연구의 결과는 전장 상황 분석에서 LLM의 효과적인 활용 가능성을 제시하며, 특히 실시간 데이터 처리와 높은 신뢰성이 요구되는 도메인에서 새로운 방향을 제시한다.

▶ **주제어:** 대규모 언어 모델, 검색 증강 생성, 지도 학습 미세조정, 프롬프트 체이닝, 전장 상황 분석

I. Introduction

현대 전장은 그 복잡성 및 빠르게 변화하는 특성으로 인해 효율적인 전장 상황 분석 시스템에 대한 필요성이 절실히 대두되고 있다. 전장 상황에서 실시간으로 축적되는 데이터는 방대한 양과 높은 속도를 특징으로 하며, 이를 신속하고 정확하게 처리하여 전략적 의사결정을 지원하는 것은 국방 분야의 핵심 과제 중 하나이다. 특히, 최근 우크라이나-러시아 전쟁은 데이터 중심의 분석 시스템이 효과적인 상황인식 및 대응에 중요한 역할을 수행할 수 있다는 가능성을 보여준 사례로 언급된다. 이 전쟁에서는 드론, 레이더, 위성 이미지, 전자전 장비 등을 통해 방대한 양의 데이터가 생성되었으며, 이를 적시에 효과적으로 처리하지 못할 경우 전장 상황인식 및 대응 지연이 발생하여 전략적 실패로 이어질 가능성이 제기된다. 실시간 정보 분석의 부재는 병력 배치 최적화 실패나 적군의 움직임에 대한 오판으로 이어질 위험이 있으며, 이는 전장 대응에 치명적인 영향을 미칠 수 있음을 시사한다. 이러한 상황은 실시간 데이터 분석을 통한 신뢰성있는 정보 생성이 전장상황에 얼마나 중요한 역할을 하는지 인식하게 되었고 이를 위한 기술 개발의 시급성이 대두되었다.

최근 대규모 언어 모델(Large Language Model, LLM)의 발전은 다양한 도메인에서 데이터 분석과 의사결정 지원 시스템의 새로운 가능성을 열어주고 있으나, 전장 상황 분석 시스템에 LLM을 적용한 사례는 거의 없으며 이 기술은 여전히 초기 단계에 머물러 있다. 물론 Knowledge Base Question Answering(KBQA) 시스템에서 개체 연결(Entity Linking)에 초점을 맞춘 연구가 일부 존재하지만, 이러한 연구들은 GPT-4, Claude, LLaMA와 같은 최신

대규모 언어 모델이 아닌 T5-base와 같은 상대적으로 작은 규모의 모델을 사용하여 단순 개체 식별 작업에 국한되어 있어, 전장 상황의 복잡한 맥락에서 요구되는 종합적인 상황 분석, 전술적 이해, 다양한 정보 소스 통합, 불확실성 처리, 전략적 결정 지원과 같은 고차원적 분석 능력을 평가하기에는 불충분하다. 그럼에도 불구하고, LLM은 전장 데이터를 자연어로 표현하고 해석할 수 있는 잠재력 덕분에 전장의 복잡한 상황 이해와 전략적 결정을 지원하는 데 효과적인 접근법으로 주목받고 있어, 최신 대규모 언어 모델의 전장 상황 분석 능력에 대한 포괄적인 연구는 국방 분야에서 인공지능 기술 활용의 새로운 지평을 열 수 있는 중요한 연구 주제라 할 수 있다.

도메인 특화(Domain-specific) 분야의 분석 시스템에서 일반적으로 사용되는 지도 학습 미세조정(Supervised Fine-Tuning, SFT) 방식은 정적 데이터에서 추출된 질의-응답(Question-Answer, QA) 쌍을 활용하여 모델을 학습시키는 방법이다. 이 방식은 학습된 지식 내에서는 높은 정확도를 보이지만, 실시간으로 업데이트되는 정보(dynamic updates)를 반영할 수 없다는 근본적인 한계를 지닌다. 전장에서 실시간으로 축적되는 정보를 효과적으로 처리하여 적용하지 못한다면, 변화와 예측이 어려운 전장 상황에서 작전수행을 위한 정보 분석의 신뢰성과 효율성이 저하될 수 있다.

이러한 제한점을 해소하기 위해 검색 증강 생성 기술(Retrieval-Augmented Generation, RAG)이 제안되었다 [1]. RAG는 LLM과 외부 지식을 결합하여 실시간으로 업데이트되는 정보에 대한 동적 분석을 가능하게 한다. 그러

나 보안 등급이 높은 국방 데이터의 특성상 상용 LLM(예: ChatGPT, Claude 등)을 API 방식으로 활용하는 방안에 대한 보안상 제한사항이 많고 위험하여, 이를 위해 오픈소스 LLM을 활용한 시스템 개발이 필수적이다. 하지만 오픈소스 LLM에 단순히 RAG를 결합하는 방식은 다음과 같은 중요한 문제점을 가진다:

1. 환각(Hallucination) 문제: RAG 시스템은 제한된 검색 단계에서 고정된 개수만큼의 문서를 선택해야 하며, 이로 인해 질문과 밀접한 관련이 없는 문서가 포함되는 경우가 발생한다 [2, 10, 11]. 이러한 노이즈(noise)는 LLM의 생성 결과에 혼입되어 응답의 품질을 저하시킬 수 있다. 특히, 정보의 높은 신뢰성이 요구되는 전장 환경에서 비관련 문서를 기반으로 한 응답은 전장 상황 인식을 위한 정보 분석 시 오판을 초래할 위험이 크다. 이는 정보 검색 오류(information retrieval errors)가 모델의 전반적인 성능에 부정적 영향을 미칠 수 있음을 보여준다.
2. 중간 정보 손실(Lost in the Middle) 문제: RAG 시스템이 질문과 관련 있는 문서를 검색하는 데 성공하더라도, 해당 정보가 LLM의 attention mechanism에서 유실되거나 간과되는 경우가 발생한다. 이는 LLM이 긴 입력에 대해 문서의 일부 정보에 집중하지 못하여 검색 결과의 특정 부분을 제대로 활용하지 못하는 정보 활용 불균형(ineffective information utilization)에서 기인한다 [3]. 결과적으로 중요한 세부 정보가 분석단계에서 누락될 수 있으며, 이는 전술적 의사결정에서 치명적인 결함을 초래할 수 있다.
3. 프롬프트 공학(Prompt Engineering)의 비효율성: RAG 기반 시스템의 성능을 극대화하기 위해서는 세밀하게 설계된 프롬프트가 필요하다 [4]. 그러나 이 작업은 반복적 실험(iterative experimentation)을 요구하며, 시간과 자원을 과도하게 소모한다. 특히, 특수성을 지닌 국방 도메인에서는 신속성과 신뢰성이 요구되기 때문에 이러한 비효율성은 실시간 분석 환경에서는 중요한 장애물로 작용한다. 복잡한 task의 경우 prompt를 자세히 작성해야 하며, 모델의 입력 길이 제한으로 인해 충분한 설명이나 예시를 포함시키지 못하는 한계가 존재한다. 특히 긴 프롬프트는 모델의 추론 속도를 현저히 저하시키며, 이는 입력 길이에 따라 기하급수적으로 증가하는 계산 복잡도에 기인한다. 이러한 제약들은 결과적으로 시스템의 성능과 응답의 신속성을 제한하는 주요 요인으로 작용하며, 무시할 수 없는 burden을 초래한다.

본 연구는 이러한 문제를 해결하기 위해 검색 증강 생성(RAG)과 지도 학습 미세조정(SFT)을 결합한 고도화된 RAG 기반 SFT 기법을 제안한다. 기존의 RAG나 SFT 단독 접근법과 달리, 본 연구는 두 방식의 장점을 결합하여 실시간 정보 처리와 높은 신뢰성을 동시에 확보하는 것을 목표로 한다. 본 연구의 주요 기여는 다음과 같다:

1. RAG와 SFT의 한계점 분석 및 극복: 오픈소스 LLM 기반 RAG 시스템의 환각 현상과 중간 정보 손실 문제, 그리고 기존의 일반적인 SFT 방식의 실시간 정보 반영 한계를 분석하고, 이를 해결하기 위해 참조 문서를 포함한 3중 구조 데이터 학습 기법(Triple-structured SFT)을 제시한다.
2. Knowledge Distillation을 통한 효율적 모델 경량화: Prompt Chaining과 Iterative Source Extraction이 적용된 대규모 Teacher 모델(405B)의 우수한 추론 능력을 통해 전장 상황 분석에 적합한 고품질의 3중 구조 합성 데이터를 생성하고, 이를 경량화된 Student 모델(8B)로 전달하는 효과적인 지식 증류 프레임워크를 제안한다. 특히, 3중 구조의 합성 데이터를 학습하여 추론 시 결합될 RAG 시스템과의 시너지를 극대화한다.
3. 최적화된 RAG 파이프라인 구축: Ensemble Retrieval과 Reranker를 결합한 고도화된 검색 시스템을 설계하여 문서 검색 정확도를 향상시키고, 이를 경량화된 도메인 특화 모델과 통합하여 실시간 전장 상황 분석이 가능한 높은 성능의 효율적 시스템을 구현한다.

본 연구는 이러한 기여를 국방과 같이 정보의 신속성과 정확성이 핵심적인 도메인에서 기존의 한계를 극복할 수 있는 실질적인 방안을 제공할 것이다.

II. Preliminaries

1. Traditional NLP Approaches vs. LLM-based Approaches for Battlefield Analysis

전통적인 NLP 기반의 전장 상황 분석 시스템은 규칙 기반 접근법과 통계적 방법론을 주로 활용해왔다. 이러한 접근 방식은 미리 정의된 패턴과 규칙을 바탕으로 정보를 추출하고 분석하지만, 전장 환경의 복잡성과 불확실성을 처리하는 데 있어 여러 한계점을 보인다. 특히 새로운 유형의 상황이나 예외적 케이스가 발생할 경우, 이를 처리하기 위해서는 지속적인 규칙 업데이트가 필요하며, 서로 다른 출처의 정보를 의미적으로 통합하고 연결하는 능력이 제

한적이다. 또한, 불완전하거나 모호한 정보에 대한 추론 능력이 부족하여, 전장의 전체적인 상황을 이해하고 분석하는 데 어려움이 있다.

반면, 사전 학습된 LLM 기반의 접근 방식은 이러한 한계를 상당 부분 극복할 수 있다. LLM은 방대한 데이터에 대한 사전 학습을 통해 자연어의 맥락과 의미를 깊이 있게 이해할 수 있으며, 비정형 텍스트에서 핵심 정보를 추출하고 복잡한 상황을 분석하는 데 큰 강점을 제공한다. 특히 Zero/Few-shot 학습 능력을 통해 최소한의 예시만으로도 새로운 유형의 분석 작업을 수행할 수 있으며, 여러 문서의 정보를 의미적으로 연결하고 통합하는 능력이 우수하다 [5]. 또한, 암묵적 관계나 누락된 정보에 대한 추론이 가능하며, 불안정한 환경에서 수집된 정보에 대해 합리적인 분석을 수행할 수 있다.

LLM의 이러한 장점들은 전장 상황 분석과 같이 높은 수준의 불확실성과 복잡성을 가진 도메인에서 특히 중요한 의미를 갖는다. 실시간으로 변화하는 전장 환경에서 새로운 상황에 대한 빠른 적응과 불안정한 정보에 대한 강력한 추론 능력은 시스템의 신뢰성과 효과성을 크게 향상시킬 수 있다. 다만, LLM의 이러한 장점을 최대한 활용하기 위해서는 적절한 프롬프트 엔지니어링과 도메인 특화 학습이 필요하며, 이는 본 연구에서 제안하는 RAG-SFT 방법론을 통해 효과적으로 달성될 수 있다.

2. Related works

2.1 Supervised Fine-Tuning (SFT)

지도 학습 미세조정(Supervised Fine-Tuning, SFT)은 사전 학습된 언어 모델의 성능을 특정 작업에 맞게 최적화하는 데 널리 사용되는 방법이다. LIMA 연구는 실제 사용자 프롬프트와 고품질 응답을 포함하는 1,000개의 예제를 사용하여 소규모 데이터로도 높은 성능을 달성할 수 있음을 보여주었다 [6]. 이 연구에서는 Reddit, Stack Exchange, WikiHow과 같은 커뮤니티에서 추출한 질의-응답 데이터를 선별했으며, 추가적으로 인간이 작성한 데이터로 모델의 응답 일관성과 스타일을 최적화했다.

SELF-INSTRUCT 연구는 인간 작성 데이터를 최소화하고, 사전 학습된 상용 언어 모델로부터 합성 데이터를 생성하는 자동화된 프레임워크를 제안했다 [7]. 이 접근법은 human annotator의 개입을 최소화하면서도 기존의 공개된 instruction 데이터셋을 사용하여 학습하는 것보다 더 우수한 zero-shot generalization 성능을 보이며, 모델 튜닝 과정에서의 효율성을 입증했다.

도메인 특화 적용(Domain-specific adaptation)의 맥

락에서, DISC-LawLLM은 법률 도메인에서의 특화된 instruction tuning 방법론을 제시했다 [8]. 이 연구는 법률 추론을 지원하기 위해 법률적 전제를 사용하여 질의-응답 데이터를 생성하고, 이를 통해 LLM을 튜닝하였다.

Microsoft의 연구에서는 대형 오픈소스 언어 모델(LLM)의 고도화된 지식을 소형 오픈소스 언어 모델(small Large Language Model, sLLM)로 증류(knowledge distillation)하여 비용과 지연 시간을 줄이는 자원 효율적인 방식을 제안했다 [9]. 동시에, 오픈소스 모델을 활용하기 때문에 데이터가 외부로 유출될 위험을 줄이며, 민감한 데이터가 포함된 환경에서도 안전하게 적용할 수 있는 점에서 장점이 있다.

그러나 기존 연구들은 전장 상황 분석과 같은 특수한 도메인에 적용하기 위해 요구되는 다음의 핵심적 요건들을 완벽히 충족시키지 못하는 한계를 보인다. 첫째, 인간의 개입을 최소화 하는 효율적인 학습 데이터 생성 프로세스가 담보되어야 한다. 역동적으로 변화하는 전장 환경에서는 지휘관의 전술적 요구사항을 충족시키는 대규모의 고품질 학습 데이터를 신속하게 구축해야 하기 때문이다. 둘째, 고도의 전문성이 요구되는 국방 분야에 적합한 특화 능력이 필요하다. 일반 도메인의 경우 LLM이 보유한 방대한 사전 학습 지식만으로도 상당한 수준의 과제 해결이 가능하나, 전문화된 영역에서는 이러한 접근만으로는 불충분하기 때문이다. 즉, 일반 도메인에서 제안된 방법론은 도메인 특화 영역에서는 효과적이지 않을 가능성이 높으며, 이에 따라 차별화된 접근법이 필요하다. 셋째, 데이터 보안성의 철저한 보장이 필수적이다. 군사 정보의 기밀성을 고려할 때, 데이터 유출 및 보안 취약성에 대한 완벽한 방어 체계가 구축되어야 하기 때문이다. 넷째, 실시간 정보 갱신에 대한 효과적인 처리 능력이 요구된다. 전장의 상황은 끊임없이 변화하며, 이러한 동적 환경에서의 신속한 상황 판단과 대응이 핵심적이기 때문이다. 현재까지의 SFT 연구들은 이러한 필수 요건들 중 일부만을 부분적으로 충족시키거나, 특정 요건에 대해서는 실효성 있는 해결책을 제시하지 못하고 있는 실정이다.

2.2 Retrieval-Augmented Generation (RAG)

검색 증강 생성(Retrieval-Augmented Generation, RAG)은 대형 언어 모델(LLM)의 정적 지식을 외부 데이터베이스의 동적 정보(retriever)와 결합하여 실시간으로 정보를 보강하며, 정보 검색 및 자연어 생성 작업에서 유망한 결과를 보여주고 있다 [1]. 특히, RAG는 사전에 학습된 모델이 최신 데이터에 접근할 수 없다는 한계를 극복하기

위해 설계되었다. 이를 통해 모델은 외부 데이터베이스에서 검색된 문서를 기반으로 답변을 생성할 수 있다. 그러나 사전 학습된 LLM을 fine-tuning하지 않은 채 단순히 RAG를 결합할 경우, 두 가지 문제가 발생한다. 첫째, 환각(hallucination) 문제로, 검색된 문서가 질문과 관련이 없을 때도 모델이 이를 무비판적으로 활용하여 잘못된 응답을 생성하는 현상이다 [2]. 둘째, 중간 정보 손실(lost in the middle) 문제로, 모델이 긴 문서의 중요한 세부 정보를 놓치거나 효과적으로 처리하지 못하는 현상이다 [3].

한편, 앞서 설명한 방식과 달리 참조 문서를 추론 시에만 보여주는 것이 아닌 학습 과정에서부터 고려한 연구도 있다. Facebook AI는 Dense Passage Retrieval(DPR)과 생성 모델(BART)을 결합하여 검색된 문서와 질문 간의 연관성을 강화했으며 [1], Google의 REALM은 사전 학습 단계에서 검색을 통합하여 모델이 학습과 동시에 검색 작업을 최적화할 수 있는 구조를 제시했다 [12]. 이러한 연구들은 RAG의 도메인 확장 가능성을 보여주었다. 그러나 이러한 접근법들은 학습-추론 불일치(Training-Inference Mismatch)로 인해 여전히 환각 문제를 해결하지 못했다 [13]. 이는 훈련 시에는 질문과 항상 관련이 높은 문서가 주어진 상태에서 학습이 이루어지기 때문에, 모델은 "검색된 문서는 항상 질문과 연관이 있다."는 가정을 기반으로 작동하게 된다. 그러나 실제 추론 환경에서는 검색기가 질문과 관련 없는 문서를 반환하는 경우가 빈번하며, 모델은 이러한 비관련 정보를 처리하거나 무시하는 방법을 배우지 못했기 때문에 이를 그대로 신뢰하고 응답을 생성하게 된다. 결과적으로, 비관련 문서에 기반한 응답에서 환각 현상이 발생하며, 이는 전장 분석과 같은 높은 신뢰성을 요구하는 도메인에서 치명적일 수 있다.

이러한 한계점을 보완하기 위해 최근에는 RAG 파이프라인의 중간 결과와 최종 결과를 자체 평가하여 생성 과정을 조정하는 방법론들이 제안되었다. Self-RAG는 검색된 문서와 모델의 생성 결과를 반복적으로 재검토하는 critic 모듈을 도입하고 [14], Corrective RAG는 질문과 검색된 문서의 관련성을 평가하고, 관련성에 따라 다른 방식으로 정보를 처리하는 메커니즘을 추가하여 환각 현상을 크게 줄이고, 생성 응답의 신뢰성을 강화하는 데 성공했다 [13]. 이 외에도 검색된 문서들 중에서 모델의 output 생성에 필요한 문서인지 판단하고 필터링하여 noise에 의해 부정확한 응답을 생성하는 문제를 해결하려는 시도들도 있다 [11, 23]. 하지만 해당 연구들은 품질 평가를 위한 별도의 복잡한 학습 메커니즘이 필요하고, 다른 도메인으로의 일반화가 보장되지 않는다는 한계가 있다. 이를 보완하기 위

해 최근에는 Agent 개념을 도입하여 추가 학습 없이 사전 학습된 LLM이 관련성 평가나 품질 평가를 직접 수행하게 하여 해당 아이디어들의 실용성을 극대화하는 방법론들이 제안되고 있다. 그러나 이러한 접근법들은 성능 면에서는 효과적이지만 추론 시간이 대폭 증가하여 실시간 처리가 필요한 환경에서는 지연 시간이 문제가 될 수 있고, 프롬프트 설계(prompt engineering) 작업에도 많은 시간과 자원이 요구된다는 문제점이 있다.

본 연구에서는 기존 연구들의 한계를 극복하기 위해 검색 증강 생성(RAG)과 지도 학습 미세조정(SFT)을 결합한 새로운 프레임워크를 제안한다. 기존 연구들은 참조 문서를 학습에 활용하거나(DPR, REALM) 추론 시 품질 평가를 강화하는 방식(Self-RAG, Corrective RAG)으로 접근했으나, 학습-추론 불일치 문제나 실시간 처리의 한계를 근본적으로 해결하지 못했다. 이에 반해 본 연구는 질문, 답변, 참조 문서를 포함한 3중 구조의 학습을 통해 모델이 문서의 관련성을 효과적으로 판단하고 활용할 수 있도록 하며, 이를 경량화된 모델에 성공적으로 전수함으로써 환각 문제를 해결하면서도 실시간 처리가 가능한 시스템을 구현한다. 특히, critic 모듈이나 복잡한 프롬프트 없이도 우수한 성능을 달성할 수 있어, 전장 상황과 같이 신속하고 정확한 분석이 요구되는 환경에서 실질적인 활용이 가능하다는 점에서 의미가 있다.

III. The Proposed Scheme

본 연구에서는 전장 상황 분석에 특화된 고성능 경량 sLLM을 구축하기 위해 Fig. 1과 같이 4단계 접근법을 제안한다. 첫째, Teacher 모델(llama3.1 405b)을 활용하여 고품질 (질문, 참조 문서, 답변)의 3중 구조 합성 데이터를 생성한다. 둘째, Parameter-Efficient Fine-Tuning (PEFT)을 통한 Student 모델(llama3.1 8b)로의 지식 증류를 수행한다. 셋째, 학습이 완료된 Student 모델을 RAG 파이프라인과 통합한다. 넷째, 사용자와의 상호작용을 위한 사용자 친화적인 대화형 인터페이스를 구축한다. 각 단계별 상세 내용은 다음과 같다.

1. Triple-structured Synthetic Data Generation

전장 상황 분석을 위한 고품질의 학습 데이터를 생성하기 위해, 우리는 대규모 언어 모델(llama3.1 405b)을 활용한 자동화된 데이터 생성 프레임워크를 구축하였다. 이 프레임워크는 다음의 5단계로 구성된다:

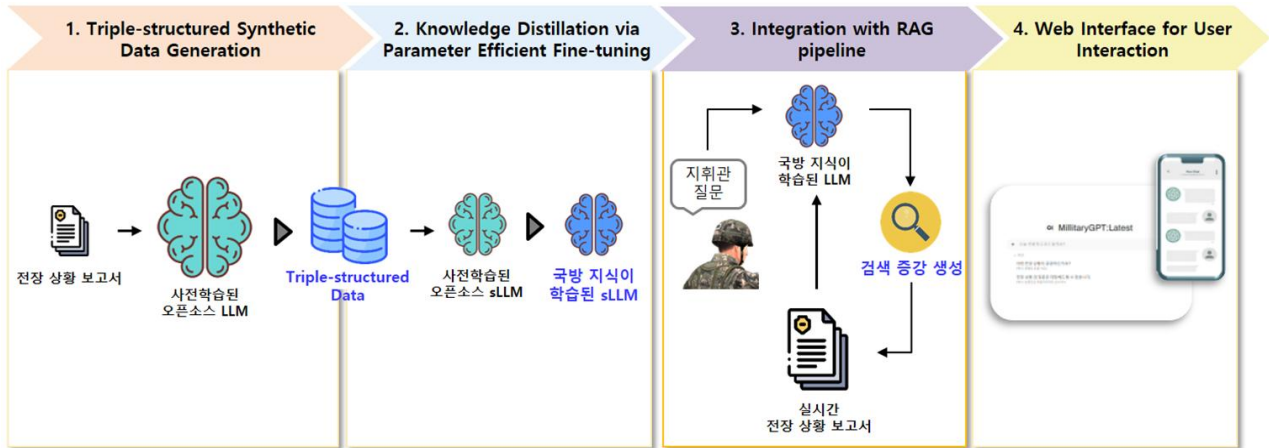


Fig. 1. Overview of the proposed methodology

1.1 Document Chunking

검색 대상이 되는 전장 보고서를 효과적으로 처리하기 위해, 각 문서를 일정 길이의 청크(chunk)로 분할한다. 문서 D 가 주어졌을 때, 청크 분할은 다음과 같이 정의된다:

$$D = \{d_1, d_2, \dots, d_M\}, C = \{c_1, c_2, \dots, c_N\}$$

여기서 D 는 원본 문서 집합이고, C 는 청크된 문서 집합이며, M 은 전체 문서 수, N 은 생성된 청크의 수를 나타낸다.

1.2 Question Generation

청크된 문서 집합을 순차적으로 Teacher(llama3.1 405b) 모델에 입력하여 해당 내용을 바탕으로 질문을 생성하였다. 생성된 질문은 입력 문서에서 답을 찾을 수 있는 질문이어야 하며, 의미론적으로 유효하고 다양한 질문을 생성하기 위해 군사 전문가들이 작성해준 전장 상황 분석 질문 리스트 E 를 LLM이 참고하도록 프롬프트 P_E 를 구성하였다.

$$q_i = \text{Teacher}(c_i; P_E)$$

여기서 c_i 는 i 번째 청크 문서, q_i 는 i 번째 청크 문서에서 생성된 질문, P_E 는 군사 전문가들이 제공한 질의 리스트 E 를 참고하여 새로운 질문을 생성하는 지시사항(instruction)이 포함된 프롬프트이다.

1.3 Dynamic Document Retrieval

생성된 질문과 관련된 문서를 Retrieval를 활용하여 k 개를 검색하였다. 검색 순위 기준은 문서와 질문과의 임베딩 벡터 기반 코사인 유사도로 계산했으며, 검색 대상이 되는 청크 문서 집합 C 의 모든 임베딩 벡터를 저장하는 벡터 데이터베이스(Vector database)로는 FAISS를 활용하였다 [16]. 검색된 문서의 수 k 는 5에서 15 사이의 값

로 랜덤 샘플링하여 설정하였다. 이를 통해 추론 시 다양한 k 값을 고려할 수 있도록 하여, 모델이 다른 검색 조건에서도 적절하게 작동할 수 있도록 하였다. 이는, 학습 과정에서 다양한 k 값에 모델이 노출되도록 하여 특정 검색 개수에 과도하게 의존하지 않도록 한다. 만약 주어진 질문에 대해 더 많은 관련 정보를 포함하고 싶다면 k 를 더 크게 설정하면 된다.

$$R_i = \text{retrieve}(q_i, C, k), k \sim U(5, 15)$$

여기서 R_i 는 질문 q_i 에 대해 검색된 청크 문서 집합이고, $U(a, b)$ 는 균등 분포를 의미한다.

이때, 평가를 위한 test set에 대해서는 질문과 관련된 모든 청크 문서가 검색된 청크 문서에 모두 포함되어 있어야 하므로 k 를 15로 고정하여 합성 데이터를 생성했으며, 이를 사람이 직접 검수했다.

1.4 Prompt Chaining: Extract sources and generate answers

llama3.1 405b를 활용하여 질문과 관련된 출처와 최종 답변을 모두 생성하도록 했다. 이는 단순히 검색된 문서를 기반으로 답변만 생성하는 방식보다 높은 신뢰성과 정확성을 확보할 수 있도록 하기 위함이다. LLM이 출처를 명시적으로 생성하도록 함으로써, 사용자가 응답의 신뢰성을 직접 검증할 수 있으며, 불확실한 정보로 인해 발생할 수 있는 판단 오류를 최소화할 수 있다.

이 과정은 Prompt Chaining이라는 기법을 적용하여 두 단계로 나누어 수행하였다. 즉, LLM이 출처를 먼저 생성하게 한 후, 추출한 출처를 기반으로 답변을 생성하는 방식이다 [15]. 실제로, 출처와 답변을 LLM이 한 번에 생성하도록 했을 때, 답변 품질이 현저히 낮았고 중요한 정보가 누락되는 문제가 발생했다. 특히, LLM이 검색된 문서에서 질문과 무관한 출처를 추출하거나, 질문 자체를 망

각하는 현상이 빈번하게 나타났다. 이러한 문제를 해결하기 위해, 출처를 먼저 추출한 후 해당 출처를 기반으로 답변을 생성하는 두 단계 접근법을 도입하였다.

1. Iterative Source Extraction: 검색된 문서 중 질문과 관련성이 높은 출처를 식별하여 추출한다. 하지만 Prompt Chaining으로 테스크를 두 단계로 분리하여 해결하더라도 검색된 문서의 길이가 길어질 경우 출처 생성 과정에서 출처 추출을 실패하는 문제가 빈번하게 발생했다. 이를 방지하기 위해, 검색된 청크 문서를 3개씩 그룹화하여 검색된 청크 문서 전체가 아닌 그룹화된 청크 문서 내에서만 출처를 추출하도록 했다. 이러한 방식을 적용한 결과, 질문에 포함된 지시 사항을 충실히 수행하면서, 출처를 보다 정확하게 추출할 수 있었으며, 최종적으로 생성된 답변의 품질과 신뢰성이 크게 향상되었다.

$$G_i = \{g_{i,1}, g_{i,2}, \dots, g_{i,m}\}, m = \left\lceil \frac{k}{3} \right\rceil$$

여기서 G_i 는 R_i 를 3개씩 그룹화하여 생성된 문서의 집합, m 은 그룹 수, $\lceil c \rceil$ 는 올림 연산을 의미한다. 즉, k 가 3의 배수가 아닐 경우, 마지막 그룹에는 3개 미만의 문서가 포함될 수 있다.

각 하위 문서 그룹 $g_{i,j}$ 에 대해, Teacher 모델 (llama3.1 405b)은 순차적으로 출처를 추출한다.

$$s_{i,j} = \text{Teacher}(q_i, g_{i,j}, P_s)$$

$$S_i = \{s_{i,1}, s_{i,2}, \dots, s_{i,m}\}$$

여기서 $s_{i,j}$ 는 주어진 질문 q_i 에 대한 출처를 j 번째 문서 그룹 $g_{i,j}$ 에서 Teacher 모델이 추출한 결과이고, P_s 는 출처 추출을 위한 지시사항(instruction)이 포함된 프롬프트이다.

2. Final Answer Generation: 앞서 추출된 출처 들을 활용하여 최종 답변을 생성한다. 모든 출처들을 활용하되, 중복된 내용은 포함하지 않고 질문과 관련성이 낮은 정보는 무시하도록 프롬프트를 구성하여, 불필요한 정보가 포함되는 것을 방지하였다.

$$a_i = \text{Teacher}(q_i, S_i, P_a)$$

여기서 a_i 는 주어진 질문 q_i 에 대한 답변을 출처 집합 S_i 를 참고하여 Teacher 모델이 생성한 결과이고, P_a 는 최종 답변 생성을 위한 지시사항(instruction)이 포함된 프롬프트이다.

1.5 Triple-structured Data Formation

위의 과정을 통해 최종적으로 (질문, 참조 문서, 답변)의 3중 구조 데이터를 생성하였다. 최종 Ground Truth는 단순히 답변 a_i 가 아니라, 출처 정보 S_i 와 최종 답변 a_i 를 결합한 형태로 구성하였다.

$$t_i = \{q_i, R_i, S_i \oplus a_i\}$$

$$T = \{t_1, t_2, \dots, t_L\}$$

여기서 \oplus 는 연결(concatenation) 연산을 의미하며, t_i 는 i 번째 청크 문서에서 생성된 3중 구조 데이터, T 는 동일한 과정을 다른 청크 문서들에 대해 L 번 반복하여 생성된 최종 학습 데이터이다.

다음은 제안하는 합성 데이터 생성 방법론을 통해 생성된 3중 구조 데이터의 예시이다. 본 예시는 retrieval 단계에서 k 값이 랜덤하게 13으로 선택되었을 때 생성된 실제 합성 데이터이다:

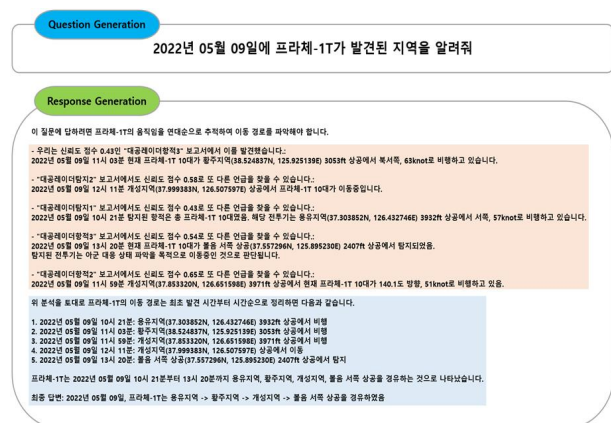


Fig. 2. An example of the triple-structured synthetic data

생성된 합성 데이터셋의 token-level 분석 결과, 평균 sequence length는 4,893 토큰이며, 최소 및 최대 토큰 수는 각각 2,849 및 6,372로 나타났다. 이는 확장된 context window를 지원하는 최신 LLM이 충분히 처리 가능한 범위로, student 모델 학습 시 truncation이나 sliding window 기법 없이도 효과적인 knowledge transfer가 가능하다. 이러한 다양한 길이의 데이터셋은 RAG 시스템의 두 가지 주요 한계점을 동시에 해결하는 데 기여한다. 첫째, "lost in the middle" 문제에 대응하여 모델이 다양한 길이의 컨텍스트에서 중요 정보를 효과적으로 식별하고 추출하는 능력을 학습함으로써, 긴 검색 결과의 중간 부분에 위치한 핵심 정보도 효과적으로 활용할 수 있게 된다. 둘째, 엄격한 verification cascade를 통과한 고품질 데이터셋으로 학습함으로써 "hallucination" 문제를 완화하여, 모델이 검색된 문서에 없는 정보나 검색된 문서들 중 질문과 관련 없는 정

보를 무분별하게 활용하는 경향이 크게 감소한다. 결과적으로, 이러한 합성 데이터를 통한 학습은 모델이 검색된 정보에 충실하면서도 다양한 길이의 컨텍스트를 효과적으로 처리할 수 있는 능력을 동시에 향상시킨다.

Prompt Chaining(프롬프트 체이닝)과 Iterative Source Extraction(단계적 출처 추출)은 계산 복잡도와 추론 시간을 증가시키는 특성을 가지고 있다. 특히, Iterative Source Extraction의 경우 그룹 수(m)만큼 추론을 반복해야 하기 때문에 검색하는 문서의 수(k)가 증가할수록 지연 시간은 선형적으로 증가하게 된다. 그러나 본 연구의 맥락에서 이러한 계산 비용의 증가는 실제 운영 환경에 영향을 미치지 않는다. 그 이유는 이러한 계산 집약적 프로세스가 오직 합성 데이터 생성 단계에서만 수행되며, 실제 서비스 환경에서는 이를 통해 획득된 지식이 경량화된 모델로 증류되어 활용되기 때문이다. 즉, 데이터 생성 단계에서의 시간적 오버헤드는 실시간 추론 성능과는 무관하며, 오히려 이러한 철저한 프로세스를 통해 생성된 고품질 학습 데이터는 경량 모델의 성능 향상에 핵심적인 역할을 한다.

2. Knowledge Distillation via Parameter Efficient Fine-tuning

앞선 단계에서 llama3.1 405b 모델(LLM)에 의해 생성된 3중 구조 데이터에 대해 llama3.1 8b 모델(sLLM)을 LoRA (Low-Rank Adaptation)를 사용하여 지도 학습 미세조정(Supervised Fine-Tuning, SFT)을 수행하였다. LoRA는 사전 학습된 모델의 가중치를 고정된 상태에서 특정 저차원 행렬을 학습하는 방식으로, 모델 크기에 비해 효과적으로 성능을 향상시킬 수 있는 기법이다 [17].

$$W = W_0 + BA$$

$$B \in R^{d \times r}, A \in R^{r \times d}$$

여기서 W_0 는 사전 학습된 가중치, B 와 A 는 저차원 행렬($r \ll d$)이다.

student 모델인 llama3.1 8b는 decoder 모델이므로, 학습 과정에서 손실 함수는 교사 강요(teacher forcing) 방식의 다음 토큰 예측(next token prediction)으로 정의된다:

$$\theta^* = \arg \min_{\theta} \sum_i^L L(f_{\theta}(q_i, R_i), S_i \oplus a_i)$$

여기서 L 은 손실 함수, θ 는 student 모델의 가중치, f_{θ} 는 미세조정될 student 모델이다. 즉, 주어진 질문 q_i 와 검색된 문서 R_i 가 입력되었을 때의 미세조정될 student 모델 f_{θ} 의 각 출력 토큰이 target인 $S_i \oplus a_i$ 의 동일 위치의 토큰에 가까워지도록 cross entropy loss L 을 최소화하도록 학습이 진행된다.

student 모델 기준에서 출처와 답변을 모두 생성하도록 학습하는 것은 일종의 CoT(Chain of Thought) 방식으로도 볼 수 있다. CoT는 대규모 언어 모델이 복잡한 문제를 해결할 때 사용하는 추론 방식으로, 모델이 최종 답변을 바로 제시하는 대신, 사람처럼 단계별로 생각하는 과정을 명시적으로 보여주는 접근법이다 [18]. CoT 방식으로 학습된 모델은 추론 능력과 일반화 성능이 향상된다고 알려져 있다. 따라서, 출처와 답변을 모두 생성하는 방식은 단순한 품질 개선을 넘어, 보다 안정적이고 신뢰성 높은 전장 분석 시스템을 구축하는 데 기여한다.

3. Integration with RAG pipeline

이전 단계에서 학습 완료된 llama3.1 8b SFT 모델을 실제 추론 과정에서 RAG 파이프라인과 결합하여 실시간 전장 상황 분석 시스템을 구현하였다. RAG 시스템은 주어진 질문에 대한 검색된 문서를 활용하여 모델이 최신 정보를 반영할 수 있도록 하며, LLM이 사전 학습된 데이터에만 의존하지 않고 동적으로 업데이트되는 정보를 효과적으로 활용할 수 있도록 한다.

$$R = \text{retrieve}(q_u, C, k_u)$$

$$\text{Final Response} = f_{\theta^*}(q_u, R)$$

여기서 q_u 는 사용자의 질의, k_u 는 사용자가 설정한 검색 문서의 수, R 은 검색된 청크 문서 집합이며, 최종 응답은 q_u 와 R 이 학습 완료된 student 모델 f_{θ^*} 에 입력되었을 때의 최종 출력이다.

4. Web Interface for User Interaction

본 연구에서는 앞선 일련의 과정을 통해 구축된 LLM-RAG 기반 전장 상황 분석 시스템의 실용성과 접근성을 향상시키기 위해 Open WebUI 기반의 대화형 인터페이스를 구현하였다. 이 인터페이스를 통해 사용자는 자연어로 질문을 입력하고, 시스템의 응답을 실시간으로 확인할 수 있다.



Fig. 3. Web Interface of the Proposed Military Situation Analysis System

이러한 웹 인터페이스는 다음과 같은 세 가지 장점을 제공한다:

1. 고빈도 질의 패턴 수집: 사용자가 자주 하는 질문들을 수집하고 학습에 반영할 수 있다.
2. 지속적 성능 개선: 시스템이 부정확한 응답을 생성하는 경우를 체계적으로 수집하고 분석하여, 모델의 약점을 파악하고 개선할 수 있다.
3. 검색 파이프라인 최적화: 실제 질의에 대한 검색 결과를 분석하여 RAG 파이프라인의 각 컴포넌트(Retrieval, Reranking 등)를 최적화할 수 있다.

IV. Performance Evaluation

본 섹션에서는 검색된 문서에서 질문과 관련된 출처를 LLM이 정확히 생성했는지를 평가하고, 이를 활용하여 적절한 답변을 생성하는 능력을 분석한다. 평가 방법은 출처 추출 평가와 답변 적절성 평가로 구분되며, 후자는 답변 품질 평가와 출처 활용 평가로 세분된다.

1. Source Data

본 연구에서 사용된 데이터셋은 실제 전장 환경을 반영한 검증된 데이터를 바탕으로 구축되었다. 실제 전장에서 수집된 보고서는 군사 작전, 정보 전략, 비밀 작전에 관한 민감한 내용을 포함하고 있어 철저한 보안 프로토콜에 따라 보호되며, 외부 연구자나 일반 대중이 접근하는 것이 극히 어렵다. 따라서, 본 연구에서는 실제 보고서의 형식과 내용을 반영한 합성 전장 보고서를 생성하여 연구를 진행하였다. 이는 현실적인 군사 작전 환경을 재현하는 동시에, 민감한 정보 보호 기준을 준수하면서도 정밀한 성능 평가를 가능하게 한다.

데이터는 다음의 3단계 과정을 거쳐 생성되었다.

1. 전문가 기반 시나리오 설계: 대한민국 군사 전문가들과 협력하여 육해공 작전을 포함한 다양한 전장 시나리오를 설계했다. 이를 통해 작전 수행 과정에서 발생할 수 있는 다양한 상황을 체계적으로 반영하였다.
2. 군 시뮬레이션 모델 활용: 대한민국 군에서 실제 훈련과 작전 계획에 활용하는 공인된 시뮬레이션 모델인 위게임 모델을 적용하여, 설계된 시나리오를 기반으로 다양한 전장 상황에 대한 시뮬레이션을 생성하였다. 이 모델들은 작전 수행의 정밀도를 높이는 데 사용되는 공식적인 시스템이다.
2. 전장 보고서 생성 및 검토: 시뮬레이션 결과를 바탕으로 군사 보고서 형식에 맞춰 데이터를 생성하였으며,

군사 전문가들이 이를 검토하여 실전에서 사용되는 보고서 형식과 일관성을 유지하도록 조정하였다.

이러한 데이터 구축 과정은 실제 작전 환경과의 높은 유사성을 유지하며, 분석 결과가 실전 적용 가능하도록 설계되었다. 본 연구에서 활용된 데이터셋은 단순한 시뮬레이션 결과가 아니라, 군사 전문가의 검토를 거쳐 실질적인 작전 수행 방식과 부합하도록 구성되었다. 이를 통해 LLM의 성능을 평가하고, 실전 환경에서의 활용 가능성을 보다 정밀하게 분석할 수 있도록 한다.

위에서 서술한 체계적 프로세스를 통해 구축된 한국어 시뮬레이션 데이터 코퍼스는 총 53,896개의 개별 보고서로 구성되어 있으며, 이들은 963개의 고유 엔터티를 포함하고 있다. 텍스트적 측면에서 분석해보면, 각 보고서는 평균 487자의 문자열로 구성되어 있으며, 최소 321자에서 최대 623자까지의 분포를 보인다. 다음은 실제 생성된 시뮬레이션 데이터이다.

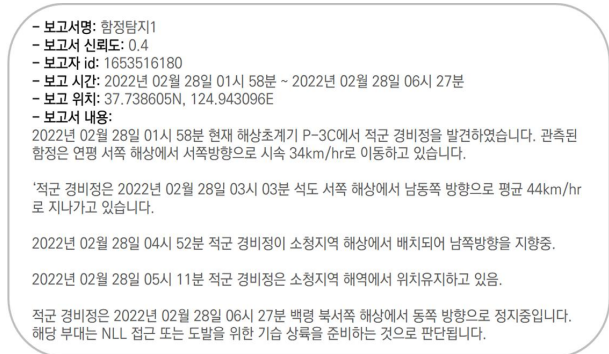


Fig. 4. Example simulation data

2. Training and Evaluation Data

본 연구에서는 train set, dev set, test set을 구성할 때, 데이터의 일반화 성능을 극대화하기 위해 보고서 내 등장하는 개체(entity)가 서로 겹치지 않도록 구성하였다. 즉, 특정 개체가 한 데이터셋에 포함되면 다른 데이터셋에서는 등장하지 않도록 설계하여, 새로운 상황에서도 적절히 대응할 수 있는지를 평가하였다.

이러한 방식은 실시간으로 업데이트되는 전장 환경에서의 성능을 측정하는 데 적합하다. 실제 작전에서는 매 순간 새로운 정보가 들어오고, 과거에 등장한 개체와 완전히 다른 상황이 전개될 수 있기 때문에 모델이 새로운 데이터를 얼마나 효과적으로 처리하는지가 중요한 평가 요소이다. 데이터셋은 source data를 기반으로 본 연구에서 제안하는 합성 데이터 생성 방법론을 통해 생성된 3중 구조 합성 데이터 3,243개에서 분할되었다.

3. Implementation Details

본 연구의 실험은 Meta-Llama-3.1-8B 모델을 기반으로 진행되었다. 모든 실험은 NVIDIA A40 GPU 3대를 활용하여 CUDA 12.0 환경에서 수행되었다. 모델의 효율적인 미세조정을 위해 LoRA를 적용하였으며, output text에 대해서만 loss를 계산하였다. 구체적으로 어텐션 및 feed-forward 계층의 주요 모듈에 LoRA를 적용하였으며, rank 16, alpha 32의 구성을 사용하였다. 학습은 learning rate 1e-4, micro batch 64, maximum sequence length 8,192 tokens, 10 epochs으로 진행하였다. Optimizer로는 AdamW를 사용하였으며, constant with warmup learning rate scheduling을 적용하였다.

데이터 생성 과정에서는 LangChain 라이브러리를 활용하였고, 모델은 모두 Ollama를 통해 로컬 환경에서 호스팅되었다. 높은 품질을 위해 Llama 3.1 405B 모델을 활용하였으며, 임베딩 및 retrieval은 각각 BAAI/bge-m3 모델과 FAISS 벡터 저장소를 사용하였다. 문서 청크 간 유사도 검색은 normalize_embeddings 옵션을 활용해 정규화된 임베딩을 기반으로 수행되었다.

4. Evaluation Metric

본 연구에서는 출처 추출 정확성과 답변 생성 신뢰성을 다각도로 평가하기 위해 출처 추출 평가(Precision, Recall, F1-score)와 답변 적절성 평가(답변 품질 평가, 출처 활용 평가)를 수행하였다. 특히, 기존의 정량적 평가 방식과 더불어 Likert 척도를 활용한 인간 평가 및 LLM 평가를 병행함으로써, 의미적 유사성과 사실적 정확성을 동시에 측정할 수 있는 정밀한 평가 체계를 구축하였다. 본 섹션에서 제시하는 평가 지표들의 신뢰성을 확보하기 위해 RAG 파이프라인의 대표적인 평가 프레임워크인 RAGAS를 참고하여 설계했다 [19].

4.1 Source Extraction Evaluation

검색된 문서에서 질문과 관련된 출처를 LLM이 얼마나 정확하게 생성했는지를 평가하기 위해 Precision, Recall, F1-score를 사용하였다.

- Precision (=P): LLM이 생성한 출처 중 질문과 실제 관련된 출처의 비율.
- Recall (=R): 질문과 관련된 모든 출처 중 LLM이 정확히 생성한 출처의 비율.
- F1-score (=F1): Precision과 Recall의 조화 평균으로, 균형 잡힌 평가 척도를 제공함.

Precision과 Recall은 출처 생성 정확성을 평가하는 데 필수적인 지표이다. Precision이 높을수록 LLM이 생성한

출처가 실제 질문과 관련이 높음을 의미하며, Recall이 높을수록 중요한 출처를 놓치지 않고 포함하였음을 의미한다. F1-score는 Precision과 Recall 간의 균형을 유지하여, 출처 누락과 불필요한 출처 포함 사이의 트레이드 오프(trade-off)를 고려하는 데 유용하다. 이는 특히 출처를 그대로 생성해야 하는 문서 검색 및 요약 시스템에서 중요한 평가 방법이다.

4.2 Answer Quality Assessment

LLM이 생성한 답변의 적절성을 평가하기 위해 답변 품질 평가와 출처 활용 평가를 수행하였다.

- Answer Correctness Evaluation (답변 품질 평가): LLM이 생성한 답변이 주어진 정답(Ground Truth)과 얼마나 일치하는지 평가하기 위해 5점 Likert 척도를 아래와 같이 적용했다.

Table 1. Answer Correctness Evaluation

Score	Criteria
1	The answer is completely wrong or has nothing to do with the correct answer.
2	Contains some similar content to the correct answer, but contains significant errors or is missing significant amounts of key information.
3	Contains about half of the key points of the correct answer, and the information included is accurate.
4	Contains most of the key points of the correct answer, with only minor omissions or minor errors.
5	Contains all key information of the correct answer and is error-free.

이러한 평가 기준을 도입한 이유는 기존의 ROUGE 또는 BLEU와 같은 전통적인 평가 지표는 n-gram 기반의 유사도 측정 방식을 사용하지만, 이는 의미적 동등성을 제대로 반영하지 못하는 한계가 있다 [24]. 질문에 대한 답변은 정답과 반드시 동일한 단어를 사용할 필요 없이 다양한 표현이 가능하므로, 의미적 유사성(Semantic Similarity)을 반영할 수 있는 평가 척도가 필요하다. 이를 위해 사람 평가와 함께 LLM 평가를 병행하여 더욱 객관적인 평가가 가능하도록 하였다. LLM 평가의 경우, 다양한 Temperature(0.0, 0.25, 0.5, 0.75, 1.0)을 통해 LLM의 응답 특성을 포괄적으로 평가하였다. Temperature 0.0은 가장 결정론적인 평가를 제공하며, 1.0에 가까울수록 점진적으로 더 높은 자유도를 허용하여 다양한 관점에서의 평가를 가능하게 한다. 최종 평가 점수는 이들의 평균값으로 산출된다.

- Source Utilization Assessment (출처 활용 평가): LLM이 스스로 생성한 출처를 적절히 활용하여 답변을 생성하는지 평가하기 위해 아래 표와 같이 5점 Likert 척도를 사용했다.

Table 2. Source Utilization Assessment

Score	Criteria
1	Creating content unrelated to the given source, or not utilizing the source at all.
2	Although some sources were used, excessive additions were made that did not appear in the sources.
3	About half of the key information in the source is used, and most of the answer is based on the source.
4	It uses most of the core content of the source, and contains almost no content that is not in the source.
5	Accurately utilizes key information from all sources provided, and there is no content that is not in the source.

LLM은 종종 사실과 다른 내용을 생성하는 환각 문제를 일으킬 수 있으므로, 출처 기반 생성이 얼마나 충실하게 이루어지는지를 평가하는 것이 필수적이다. 단순히 출처를 참조했는지 여부가 아니라 출처 정보를 어떻게 활용했는지를 분석하는 것이 중요하기 때문에, 위 테이블과 같은 Likert 척도를 활용한 사람 평가와 다양한 Temperature 설정(0.0, 0.25, 0.5, 0.75, 1.0)에서의 LLM 평가를 수행했다.

5. Final Student Model Selection

본 실험에서는 전장 환경에서의 정보 처리 성능을 비교하기 위해 10 billion 이하의 파라미터를 가진 student 모델을 평가하였다. 실험 대상 모델은 llama3.1 8b [20], qwen2.5 7b [21], gemma2 9b [22]이며, 모두 3중 구조의 train set을 활용하여 SFT를 수행했다. 평가는 test set에서의 성능을 기준으로 하였고, 평가 지표로는 출처 추출 능력 및 답변 적절성 평가를 사용하였다. 해당 실험에서는 LLM 기반 자동 평가만을 활용하였으며, 사람 평가 없이 진행하였다.

다음 표는 평가된 student 모델들의 성능 결과를 요약한 것이다.

Table 3. Student Model Performance Comparison

Model	Source Extraction			Answer Quality	
	P	R	F1	Answer Correctness (LLM)	Source Utilization (LLM)
gemma2 9b	0.95	0.93	0.93	4.25	4.57
llama3.1 8b	0.96	0.95	0.96	4.43	4.72
qwen2.5 7b	0.95	0.95	0.95	4.36	4.62

평가 결과, llama3.1 8b 모델이 출처 추출 평가와 답변 적절성 평가에서 가장 높은 성능을 보였다. 따라서, 본 연구에서는 llama3.1 8b를 최종 student 모델로 선정하였다.

6. Performance comparison with large open source LLMs

본 실험에서는 다양한 모델 구성에 따른 성능을 비교 분석하였다. 특히, 대규모 모델과 경량화 모델의 성능 차이, 그리고 SFT와 Prompt Chaining의 영향을 평가하였다. 실험은 앞서 설명한 출처 추출 평가와 답변 적절성 평가를 통해 수행되었으며, 각 모델의 특성과 성능을 종합적으로 분석하였다.

본 실험에서는 다음과 같은 다섯 가지 모델 구성을 비교하였다:

- llama3.1 405b: Meta에서 공개한 405b 파라미터의 instruction-tuned 모델이다. 이는 8b의 소규모 언어 모델인 llama3.1 8b와의 성능을 비교하기 위한 기준으로 사용되었다.
- llama3.1 405b+Prompt Chaining: 기본 llama3.1 405b 모델에 Prompt Chaining 기법을 추가로 적용한 모델이다. 복잡한 분석 작업을 여러 단계로 분해하여 순차적으로 처리함으로써, 각 단계에서의 정확성을 높이고 최종 결과의 신뢰성을 향상시키는 접근 방식을 채택하였다. 특히 전장 상황과 같이 복잡한 맥락 이해가 필요한 경우, 단계별 추론을 통해 오류 가능성을 최소화한다.
- llama3.1 8b: Meta에서 제공하는 8b 규모의 경량화 모델이다. 이는 컴퓨팅 자원이 제한된 환경에서 지도 학습 미세조정(SFT)의 필요성과 대규모 모델인 llama3.1 405b과의 추론 능력 차이를 확인하기 위해 선정되었다.
- llama3.1 8b SFT(Pair): 경량화 모델을 기반으로 질문-답변 쌍을 활용한 지도 학습 미세조정(SFT)을 수행한 모델이다. 도메인 특화 학습을 통해 전장 상황 분석에 최적화되, 기본적인 질의-응답 형태의 학습 데이터만을 활용하여 그 효과를 검증하였다.
- llama3.1 8b SFT(Triple): 경량화 모델에 질의, 응답 뿐만 아니라 참조 문서까지 포함한 3중 구조의 학습 데이터를 활용하여 SFT를 수행한 모델이다. 이는 retrieval의 검색 결과물인 참조 문서를 효과적으로 활용하는 방법까지 학습함으로써, RAG 시스템의 성능을 극대화하는 것을 목표로 하였다.

다음 표는 각 모델 구성별 성능 결과이다.

Table 4. Performance Comparison of Model Configurations

Model	Source Extraction			Answer Quality	
	P	R	F1	Answer Correctness (Human/LLM)	Source Utilization (Human/LLM)
llama3.1 405b	0.88	0.89	0.89	3.58 / 3.52	3.72 / 3.67
llama3.1 405b + Prompt Chain	0.98	0.97	0.97	4.52 / 4.57	4.77 / 4.81
llama3.1 8b	0.43	0.45	0.44	2.12 / 2.06	2.28 / 2.11
llama3.1 8b SFT(Pair)	0.47	0.48	0.49	2.88 / 2.67	3.62 / 3.59
llama3.1 8b SFT(Triple)	0.96	0.95	0.96	4.35 / 4.42	4.72 / 4.73

- Prompt Chaining의 효과성: llama3.1 405b+Prompt Chaining이 출처 추출, 답변 품질, 출처 활용도 등 모든 평가 지표에서 최고 성능을 달성했다. 이는 단계별로 분해하여 처리하는 Prompt Chaining 방식이 복잡한 추론 과정이 요구되는 작업에 효과적임을 입증한다. 그러나 405b 규모의 파라미터로 인한 높은 컴퓨팅 요구사항과 다단계 추론으로 인한 지연 시간은 실시간 분석이 필요한 전장 환경에서 심각한 제약이 된다. 따라서 본 연구에서는 이 모델을 지식 분류를 위한 teacher 모델로 활용하여, 성능은 최대한 보존하면서 실제 배포 가능한 수준의 효율성을 확보하고자 하였다.
- SFT 학습 전략의 효과: llama3.1 8b 대비 llama3.1 8b SFT(Pair)는 출처 추출과 답변 품질에서 유의미한 성능 향상을 보였다. 더 나아가 llama3.1 8b SFT(Triple)는 이를 크게 상회하는 성능을 달성했다. 이는 참조 문서를 포함한 3중 구조의 학습이 참조 문서의 구조와 활용 방식을 모델이 사전에 이해하게함으로써, 추론 단계에서 더욱 효과적으로 정보를 활용할 수 있게 해주었음을 시사한다.
- 지도학습 미세조정된 경량화 모델의 잠재력: 지도학습 미세조정을 적용하지 않은 기본 구성에서는 모델의 규모가 성능에 큰 영향을 미치는 것으로 나타났다. llama3.1 405b는 llama3.1 8b 에 비해 모든 지표에서 현저히 높은 성능을 보였다. 그러나 주목할 만한 점은 llama3.1 8b SFT(Triple)이 llama3.1 405b보다 우수한 성능을 보였다는 것이다. 이는 전문 도메인 특화된 학습 전략이 모델 크기의 한계를 극복할 수 있음을 입증한다. 50배 작은 파라미터 수에도 불구하고 더 높은 성능을 달성한 것은, 효율적인 학습 방법이 실제 배포 환경에서 매우 중요한 역할을 할 수 있음을

시사한다.

7. RAG Pipeline Optimization Experiment

본 섹션에서는 llama3.1 8b SFT(Triple) 모델의 성능을 극대화하기 위한 최적의 검색 파이프라인 구성을 탐색하였다. 특히, 다양한 검색 방식의 조합이 문서 검색 정확도에 미치는 영향을 분석하여, 전장 상황 분석에 가장 적합한 RAG 파이프라인을 도출하고자 하였다.

검색 성능을 평가하기 위해 정보 검색 분야에서 널리 사용되는 Precision@K와 Recall@K를 활용하였다. 본 실험에서는 K=10과 K=15를 사용하여, 더 많은 검색 결과에서의 정확도를 평가하였다. 이는 전장 상황에서 더 포괄적인 정보 수집이 필요할 수 있다는 점을 고려한 것이다.

- Precision@K: 상위 K개 검색 결과 중 관련 있는 문서의 비율을 측정한다. 예를 들어, Precision@10이 0.85라면, 상위 10개 문서 중 8.5개가 관련 문서임을 의미한다. 해당 metric은 검색된 문서들 중 질문과 관련 없는 문서(noise)로 인해 발생하는 환각 문제가 사용자의 의사 결정에 심각한 문제를 발생시키는 경우에 중요한 지표로 쓰인다.
- Recall@K: 전체 관련 문서 중 K개 검색 결과 내에 포함된 관련 문서의 비율을 의미한다. 특히, 전장 상황에서는 누락된 정보가 치명적인 결과를 초래할 수 있으므로, 높은 K값에서의 Recall 성능이 중요한 의미를 갖는다.

다음과 같은 5가지 검색 파이프라인 구성을 비교 실험하였다:

1. Dense Retrieval: 임베딩 기반의 의미론적 검색
2. Sparse Retrieval: 키워드 기반의 어휘적 검색
3. Ensemble Retrieval: Sparse Retrieval와 Dense Retrieval의 가중치 기반 검색 결과 통합
4. Ensemble Retrieval+Multi-query: Ensemble Retrieval에 다중 질의 변형을 통한 검색 확장
5. Ensemble Retrieval+Reranker: Ensemble Retrieval에 교차 인코더 기반의 정밀 재순위화

Table 5. Performance Comparison of Different Retrieval Configurations

Retrieval Configuration	Precision @10	Precision @15	Recall @10	Recall @15
Dense Retrieval	0.73	0.65	0.77	0.83
Sparse Retrieval	0.64	0.58	0.65	0.72
Ensemble Retrieval	0.80	0.74	0.87	0.93
Ensemble Retrieval + Multi-query	0.80	0.75	0.87	0.94
Ensemble Retrieval + Reranker	0.88	0.82	0.93	0.98

- Sparse Retrieval은 키워드 매칭에 의존하여 Dense Retrieval와 비교하여 상대적으로 낮은 성능을 보였다. 이는 전장 보고서의 특성상 동일한 상황이 다양한 표현으로 기술될 수 있어, 단순 키워드 매칭보다 의미 기반 검색이 더 효과적임을 시사한다.
- Ensemble Retrieval은 두 검색 방식의 장점을 결합하여 의미 있는 성능 향상을 달성하였다. 특히 추가적인 계산 비용이 크지 않으면서도 Dense와 Sparse 각각의 한계를 상호 보완하는 것으로 나타났다. 이는 키워드 기반의 명시적 매칭과 의미 기반의 유사도 계산이 상호 보완적으로 작용함을 보여준다.
- Ensemble Retrieval+Reranker 구성이 모든 평가 지표에서 가장 우수한 성능을 보였으며, 다른 구성들과 비교하여 유의미한 성능 향상을 보여주었다. 이는 교차 인코더 기반의 재순위화가 검색 결과의 정밀도를 효과적으로 유지함을 입증한다.
- Ensemble Retrieval+Multi-query은 기본 Ensemble Retrieval과 유사한 수준의 성능을 보였다. Multi-query 방식은 K값이 증가함에 따라 미세한 성능 향상을 보였으나, 이는 질의 변환과 다중 검색으로 인한 계산 복잡도의 증가를 고려할 때 실용적이지 않은 것으로 판단된다. 특히 실시간 처리가 중요한 전장 환경에서는 추론 시간의 증가가 시스템의 실효성을 저해할 수 있다.

이러한 실험 결과를 바탕으로, 최종 RAG 파이프라인은 Ensemble Retrieval+Reranker 구성을 채택하였다. Reranker의 경우 교차 어텐션 연산으로 인한 지연시간이 발생하지만 상당한 성능 향상을 달성하여, 비용 대비 효율성이 가장 높은 것으로 판단되었다. 이는 더 많은 수의 문서를 처리할 때도 높은 검색 정확도를 유지하며, 전장 상황 분석에서 요구되는 신뢰성과 실시간 처리 요구사항을 동시에 충족하는 것으로 판단된다.

8. Performance Analysis of SFT Model with Enhanced RAG Pipeline

학습 완료된 Student 모델(llama3.1 8b SFT(Triple))과 최적화된 RAG 파이프라인(Ensemble Retrieval + Reranker)의 결합 효과를 검증하기 위해 검색 문서 수(K)에 따른 성능을 평가했다. K=5, 10, 15, 20에서의 성능 비교 결과는 아래와 같다.

Table 6. Performance Comparison Based on Number of Retrieved Documents

Metric		K				
		5	10	15	20	
Retrieval	Precision@K	0.95	0.88	0.82	0.79	
	Recall@K	0.89	0.93	0.98	0.98	
Overall Performance	Source Extraction	Precision	1.0	0.98	0.96	0.96
		Recall	0.76	0.87	0.95	0.95
		F1-score	0.86	0.92	0.96	0.96
	Answer Quality	Answer Correctness (Human/LLM)	3.59	4.11	4.33	4.34
		Source Utilization (Human/LLM)	/	/	/	/
		Source Utilization (Human/LLM)	3.65	4.07	4.39	4.38
		Source Utilization (Human/LLM)	4.88	4.81	4.72	4.73
		Source Utilization (Human/LLM)	/	/	/	/
Source Utilization (Human/LLM)	4.83	4.77	4.70	4.71		

실험 결과, K=15일 때, recall이 크게 향상되었다. 이는 더 많은 관련 문서가 검색 결과에 포함되었음을 의미한다. 이에 반해, precision의 경우 K가 증가함에 따라 다소 감소하는 경향을 보였는데, 이는 검색된 문서에 noise(관련성이 낮은 문서)가 K가 증가할수록 더 많이 포함되었음을 의미한다.

전반적인 성능(Overall Performance)의 경우, K=15와 K=20에서 가장 높게 나왔다. 해당 결과에서 주목할 점은 두 가지이다. 첫째, K=20일 때의 성능이 K=15와 유사한 수준을 유지했다는 것이다. 이는 학습 과정에서 다양한 K값(5~15)에 노출되면서 모델이 문서 수의 변화에 강건해졌기 때문으로 분석된다. 즉, 모델이 학습하지 않은 K값에 대해서도 효과적으로 일반화되었음을 보여준다. 이러한 결과는 제안된 프레임워크가 실제 전장 환경에서 요구되는 다양한 상황에 유연하게 대응할 수 있음을 시사한다. 둘째, K=15와 K=20의 경우 모두 동일한 recall을 보이지만, K=20에서는 precision이 다소 하락했음에도 불구하고 출처 추출 평가(Source Extraction)와 답변 적절성 평가(Answer Quality)는 거의 동일한 수준을 유지했다. 이러한 현상은 Triple-structured SFT를 통해 모델이 검색된 문서들 중에서 관련성이 낮은 정보를 식별하고 필터링하는 능력을 효과적으로 학습했음을 시사한다. 즉, 더 많은 노이즈가 포함된 검색 결과에서도 중요 정보를 선별적으로 활용할 수 있는 견고성을 획득했다고 해석할 수 있다.

9. Qualitative evaluation

정량적 평가와 더불어, 본 연구에서 제안된 모델의 k=15일 때의 실제 전장 상황 질의에 대한 모델의 응답을 검토하였다. 다음은 두 가지 대표적 질의-응답 사례이며, 지면 제약으로 인해 검색된 15개의 참조 문서 전체는 생략하고, 모델이 생성한 최종 답변만을 제시한다.

QUESTION:
R1군단애호보병대의 최근 위치는?

ANSWER:
R1군단애호보병대의 최근 위치를 파악하려면, 해당 대대의 움직임을 연대순으로 추적하여 최근 위치를 파악해야 합니다.

- "지상부대탐지종합1" 보고서에서 신뢰도 점수 0.54로 이를 발견했습니다.
2022년 04월 03일 21시 23분 현재 R1군단애호보병대대가 김화지역에 남북방향으로 배치되어 있습니다.

- 신뢰도 점수 0.52인 "군단(A)탐지종합1" 보고서에서도 또 다른 언급을 찾을 수 있습니다.
2022년 04월 06일 06시 00분 현재 R1군단애호보병대대가 강릉지역에 배치되어 남북방향을 지향하고 있습니다.

- 우리는 신뢰도 점수 0.41인 "정보분석3" 보고서에서도 또 다른 언급을 찾을 수 있습니다.
2022년 04월 06일 08시 26분 현재 R1군단애호보병대대가 개성지역에 남북방향으로 배치되어 있습니다.

- 신뢰도 점수 0.49인 "지상부대탐지1" 보고서에서도 또 다른 언급을 찾을 수 있습니다.
2022년 04월 03일 07시 14분 현재 R1군단애호보병대대가 김화지역에서 남북방향으로 이동 중.

위 출처들을 기반으로 R1군단애호보병대대의 이동 현황을 시간 순으로 나열해보겠습니다.

- 2022년 04월 03일 07시 14분: 김화지역
- 2022년 04월 03일 21시 23분: 김화지역
- 2022년 04월 05일 06시 00분: 강릉지역
- 2022년 04월 06일 08시 26분: 개성지역

위 분석을 토대로 R1군단애호보병대대의 최근 위치를 파악하기 위해 연대순으로 추적한 결과, 가장 최근의 "정보분석3" 보고서에서 R1군단애호보병대대가 2022년 04월 06일 08시 26분에 개성지역(37.857634N, 126.571094E)에 남북방향으로 배치되어 있는 것으로 확인되었습니다. 하지만 해당 보고서의 신뢰도가 0.5보다 낮기 때문에 보고 내용에 대한 확인이 필요합니다.

최종 답변: R1군단애호보병대대의 최근 위치는 2022년 04월 06일 08시 26분에 개성지역입니다.

Fig. 5. Actual output of the proposed method

이러한 정성적 분석은 제안된 모델이 단순한 텍스트 생성을 넘어, 군사적 맥락에서 중요한 분석 능력을 갖추고 있음을 보여준다. 특히 출처 식별, 시간적 추론, 불확실성 인식과 같은 특성은 전장 상황 분석에서 요구되는 핵심 역량이며, 이는 제안된 방법론이 실제 전장 환경에서의 적용 가능성을 시사한다.

V. Conclusions

본 연구는 전장 상황 분석을 위한 LLM 기반 시스템의 새로운 프레임워크를 제안하였다. 특히, RAG와 SFT를 결합한 고도화된 접근 방식을 통해 실시간 데이터 처리의 신속성과 신뢰성 향상을 동시에 달성하고자 하였다. 본 연구의 주요 발견 및 기여는 다음과 같다:

첫째, 프롬프트 체이닝을 결합한 대규모 LLM(llama3.1 405b + RAG + Prompt Chaining)은 출처 추출과 답변 품질 측면에서 높은 성능을 달성하였다. 이는 복잡한 추론 과정을 단계별로 분해하는 접근이 전장 상황 분석의 정확성을 크게 향상시킬 수 있음을 입증한다. 실시간 추론 시 대규모 모델과 다단계 추론으로 인한 상당한 지연 시간이 발생하지만, 이는 학습 단계인 knowledge distillation 과정에서는 제약 사항이 되지 않는다. 따라서 본 연구에서는 이 모델을 teacher 모델로 선정하여 최고 수준의 성능을 경량화된 student 모델로 전달하는 전략을 채택하였다.

둘째, 참조 문서를 포함한 3중 구조의 SFT 학습(llama3.1 8b SFT(Triple))이 50배 이상 작은 모델 크기에도 불구하고 기본 대규모 모델(llama3.1 405b + RAG)을 상회하는 성능을 달성하였다. 이는 전문 도메인에 특화된 효율적인 학습 전략이 모델 크기의 한계를 극복할 수 있음을 보여준다.

셋째, RAG 파이프라인 최적화 실험을 통해 Ensemble Retrieval과 Reranker를 결합한 구성이 문서 검색 정확도를 크게 향상시킬 수 있음을 확인하였다. 특히 많은 수의 문서(K=15)를 처리할 때도 높은 재현율을 유지하여, 중요 정보의 누락 위험을 최소화할 수 있음을 보여주었다.

이러한 연구 결과는 LLM을 활용한 전장 상황 분석 시스템의 실현 가능성을 입증하며, 특히 경량화 모델과 특화된 학습 전략의 결합이 실용적인 해결책이 될 수 있음을 시사한다. 향후 연구에서는 다중 모달리티 데이터의 통합, 적대적 공격에 대한 강건성 향상, 그리고 설명 가능성 개선 등을 통해 시스템의 실전 적용성을 더욱 높일 수 있을 것으로 기대된다. 특히, 본 연구의 프레임워크를 대규모 언어-비전 모델(Large Multimodal Model, LMM)로 확장하여 텍스트뿐만 아니라 이미지, 영상, 센서 데이터 등 다양한 형태의 전장 정보를 통합적으로 분석할 수 있는 방향으로 발전시키는 것이 중요한 후속 연구가 될 것이다. LMM으로의 확장은 전장 환경의 복잡한 시각적 정보와 텍스트 정보를 동시에 처리함으로써 상황 인식의 정확도와 포괄성을 한층 더 향상시킬 수 있을 것으로 기대된다.

본 연구의 성과는 실시간 급변하는 전장 상황 분석이라는 특수한 도메인을 넘어, 신속성과 신뢰성이 동시에 요구되는 다양한 분야의 LLM 응용 연구에도 유용한 통찰을 제공할 것으로 기대된다.

ACKNOWLEDGEMENT

This work was supported by a grant-in-aid of HANWHA SYSTEMS.

REFERENCES

- [1] Lewis, Patrick, et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks." *Advances in Neural Information Processing Systems* 33 (2020): 9459-9474.
- [2] Zhang, Yue, et al. "Siren's song in the AI ocean: a survey on hallucination in large language models." *arXiv preprint arXiv:2309.01219* (2023).
- [3] Liu, Nelson F., et al. "Lost in the middle: How language models use long contexts." *Transactions of the Association for Computational Linguistics* 12 (2024): 157-173.
- [4] Lester, Brian, Rami Al-Rfou, and Noah Constant. "The power of scale for parameter-efficient prompt tuning." *arXiv preprint*

- arXiv:2104.08691 (2021).
- [5] Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.
- [6] Zhou, Chungting, et al. "Lima: Less is more for alignment." *Advances in Neural Information Processing Systems* 36 (2024).
- [7] Wang, Yizhong, et al. "Self-Instruct: Aligning Language Models with Self-Generated Instructions." *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2023.
- [8] Yue, Shengbin, et al. "DISC-LawLLM: Fine-tuning Large Language Models for Intelligent Legal Services." *CoRR* (2023).
- [9] Shirgaonkar, Anup, et al. "Knowledge Distillation Using Frontier Open-source LLMs: Generalizability and the Role of Synthetic Data." *arXiv preprint arXiv:2410.18588* (2024).
- [10] Mallen, Alex, et al. "When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories." *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2023.
- [11] Wang, Zhiruo, et al. "Learning to filter context for retrieval-augmented generation." *arXiv preprint arXiv:2311.08377* (2023).
- [12] Guu, Kelvin, et al. "REALM: retrieval-augmented language model pre-training." *Proceedings of the 37th International Conference on Machine Learning*. 2020.
- [13] Yan, Shi-Qi, et al. "Corrective retrieval augmented generation." *arXiv preprint arXiv:2401.15884* (2024).
- [14] Asai, Akari, et al. "Self-rag: Learning to retrieve, generate, and critique through self-reflection." *arXiv preprint arXiv:2310.11511* (2023).
- [15] Wu, Tongshuang, et al. "Promptchainer: Chaining large language model prompts through visual programming." *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 2022.
- [16] Douze, Matthijs, et al. "The faiss library." *arXiv preprint arXiv:2401.08281* (2024).
- [17] Hu, Edward J., et al. "LoRA: Low-Rank Adaptation of Large Language Models." *International Conference on Learning Representations*.
- [18] Wei, Jason, et al. "Chain-of-thought prompting elicits reasoning in large language models." *Advances in neural information processing systems* 35 (2022): 24824-24837.
- [19] Es, Shahul, et al. "RAGAs: Automated Evaluation of Retrieval Augmented Generation." *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. 2024.
- [20] Dubey, Abhimanyu, et al. "The llama 3 herd of models." *arXiv preprint arXiv:2407.21783* (2024).
- [21] Jang, An, et al. "Qwen2. 5 technical report." *arXiv preprint arXiv:2412.15115* (2024).
- [22] Rivière, Morgane, et al. "Gemma 2: Improving Open Language Models at a Practical Size." *CoRR* (2024).
- [23] Asai, Akari, Matt Gardner, and Hannaneh Hajishirzi. "Evidentiality-guided generation for knowledge-intensive NLP tasks." *arXiv preprint arXiv:2112.08688* (2021).
- [24] Liu, Yang, et al. "G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment." *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 2023.
- [25] Jang, Jaeun, et al. "imEL: Instance-level Masked Entity Linking Model." *2024 58th Annual Conference on Information Sciences and Systems (CISS)*. IEEE, 2024.

Authors



Wonjun Cho received his bachelor's degree in computer science from Kyonggi University in 2002 and his master's degree in defense digital convergence from Ajou University in 2025. Wonjun Cho currently works as a team

leader for the future technology team at Hanwha Systems. He is interested in natural language processing and computer vision.



Jaesung Yoo received the B.S. degree in Information Engineering from Korea University, Korea, in 2002. Jaesung Yoo is currently a chief engineer in Hanwha Systems.

He is interested in C4I system, SW development, and anomaly detection.



Sang-Min Kim received the B.S, M.S. degrees in Electronic Engineering from Kwangwoon University, Korea, in 2013, 2015, respectively. Sang-Min Kim is currently a senior engineer in Hanwha Systems.

He is interested in natural language processing and deep learning



Jaeun Jang received the B.S degree in Mathematics and Statistics from Sejong University, Korea, in 2019 and M.S degree in computer engineering from Korea University, Korea, in 2022.

Jaeun Jang is currently a AI researcher in Hanwha Systems. During his master's degree, he studied computer vision and explainable AI, and while at Hanwha Systems, he conducted natural language processing and LLM research.