Journal of KIIT. Vol. 24, No. 7, pp. 299-310, Jul. 31, 2026. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2026.24.7.299

LoRA 파인튜닝과 RAG의 하이브리드 결합 모델 기반 특허요약 프레임워크

장지훈*, 김건우**

Patent Summarization Framework based on a Hybrid Integration of LoRA Fine-Tuning and RAG

Ji-Hun Jang*, Gun-Woo Kim**

본 논문은 교육부와 경상남도의 재원으로 지원받은 경상남도 지역혁신중심 대학지원체계(RISE) 사업, 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구(RS-2026-25476256) 및 산업통상자원부 및 한국산업기술기획평가원(KEIT)의 지원을 받아 수행된 연구(RS-2025-02633048)의 결과임

요 약

본 논문에서는 대규모 언어 모델(LLM)을 활용한 특허 문서 요약 시 발생하는 데이터 암기(Memorization) 문제를 규명하고, 이를 극복하기 위한 LoRA 파인튜닝과 검색 증강 생성(RAG)의 하이브리드 프레임워크를 제안한다. 실험 과정에서 기존 특허 데이터셋 내 상당수의 중복(Deduplication) 데이터가 존재함을 발견하였으며, 이는 파라미터 효율적 미세 조정(PEFT) 기법인 LoRA 모델이 단순 암기를 수행하게 하여 성능 지표를 왜곡함을 실증하였다. 이를 해결하기 위해 데이터 무결성을 확보한 500개의 클린 데이터셋을 구축하고, 특허의 핵심 논리 구조인 '문제-해결-효과(PSB)'를 반영한 요약 태스크를 정의하였다. 실험 결과, 제안 모델은 BERTScore 0.985를 달성하였으며, 특히 학습 데이터와 테스트 데이터 간의 낮은 어휘 유사도(0.2040)에도 불구하고 높은 의미 유사도를 확보함으로써 단순 암기가 아닌 실질적인 추론 능력을 갖추었음을 입증하였다.

Abstract

This paper investigates the data memorization issue in patent document summarization using Large Language Models (LLMs) and proposes a hybrid framework of LoRA fine-tuning and Retrieval-Augmented Generation (RAG) to overcome this problem. During the experimental process, a significant amount of duplicated data was identified within existing patent datasets, demonstrating that the LoRA model, a Parameter-Efficient Fine-Tuning (PEFT) technique, performs simple memorization rather than inference, thereby distorting performance metrics. To address this, this study constructed a clean dataset of 500 samples ensuring data integrity and defined a summarization task reflecting the core logical structure of patents: 'Problem-Solution-Benefit (PSB).' Experimental results show that the proposed model achieved a BERTScore of 0.985. Notably, it demonstrated substantial inference capabilities rather than simple memorization by maintaining high semantic similarity despite low lexical similarity (0.2040) between the training and test sets.

Keywords

LLM, patent summarization, LoRA, RAG, data integrity, memorization

* 경상국립대학교 AI융합공학과 석사과정
- ORCID: https://orcid.org/0009-0006-8535-8854
** 경상국립대학교 컴퓨터공학부 부교수(교신저자)
- ORCID: https://orcid.org/0000-0001-5643-4797

· Received: Apr. 07, 2026 Revised: Apr. 29, 2026, Accepted: May 02, 2026
· Corresponding Author: Gun-Woo Kim
School of Computer Science, College of Natural Science, Gyeongsang National University, Jinju, Korea
Tel.: +82-55-772-3323, Email: gunwoo.kim@gnu.ac.kr

Ⅰ. 서 론

지식재산권의 중요성이 증대됨에 따라 전 세계적으로 특허 출원 건수가 폭발적으로 증가하고 있으며, 이에 따라 방대한 특허 문서를 신속하게 파악하기 위한 자동 요약 기술의 필요성이 그 어느 때보다 강조되고 있다. 특히 특허 문서는 일반적인 뉴스나 일상 대화문과 달리 고도의 전문 용어와 법률적 서술 체계가 결합된 텍스트로, 이를 효율적으로 처리하는 것은 기업의 기술 전략 수립 및 국가적 방위 산업 경쟁력 강화에 직결되는 요소이다.

최근 트랜스포머 기반의 거대 언어 모델(LLM, Large Language Model)은 요약 분야에서 비약적인 발전을 이루었으나, 특허와 같은 전문 도메인에서는 두 가지 치명적인 한계에 직면한다. 첫째, 모델이 학습 데이터의 패턴을 과도하게 학습하여 정답지를 그대로 출력하는 '데이터 암기(Memorization)' 현상이다. 본 연구의 예비 실험 결과, 비정상적으로 높은 ROUGE(Recall-Oriented Understudy for Gisting Evaluation) 지표를 통해 기존 수집 데이터의 오염(Data contamination) 가능성을 확인하였다. 전수 조사 결과 약 49.8%의 데이터가 중복 수록되어 있음을 발견하였으며, 이는 모델의 데이터 암기 현상을 유발하여 기존 벤치마크 데이터셋 내에 존재하는 중복 데이터로 인한 데이터 오염은 모델의 실제 추론 능력을 왜곡할 위험이 있고 실험의 객관성을 해칠 수 있다 판단하였다[1]. 이에 본 연구에서는 기존 데이터셋을 폐기하고, BigPatent(CPC(Cooperative Patent Classification) 'f', 'h') 데이터셋으로부터 국방 기술 관련 키워드 필터링을 통해 500건의 새로운 데이터를 재추출하여 유니크한 클린 데이터셋을 최종 구축하여 실험을 재수행하였다.

둘째, 기존의 1문장 요약 방식은 특허가 담고 있는 복잡한 기술적 인과관계를 충분히 반영하지 못한다는 점이다. 특허의 핵심은 '어떤 문제를 어떻게 해결했는가'에 있으나, 단일 문장 요약은 정보 압축 과정에서 이러한 핵심 논리 구조를 손실할 위험이 크다.

최근 국방 분야에서는 인공지능(AI) 기술을 활용한 의사결정 지원 및 문서 자동 생성 시스템 구축이 활발히 논의되고 있다. 그러나 국방 데이터는 외부망과 차단된 폐쇄적 환경(Air-gapped environment)에서 운영되어야 하며, 데이터의 보안 등급에 따라 가용한 학습 데이터의 양이 극히 제한적이라는 특수성을 갖는다. 또한 국방 문서는 규격서, 기술 교범 등 정형화된 서술 체계를 따르면서도 기술적 난이도가 매우 높아 일반적인 언어 모델로는 정확한 처리가 어렵다.

국방 데이터는 국가 안보와 직결되어 외부망과의 연결이 철저히 차단된 환경에서 운영되어야 한다. 이는 클라우드 기반의 초거대 모델 활용을 어렵게 하며, 제한된 컴퓨팅 자원을 가진 내부 서버에서 구동 가능한 소형 언어 모델(sLM, small Language Model)의 효율적인 최적화 기술을 요구한다. 국방 데이터는 보안 등급에 따라 가용한 고품질 학습 데이터의 양이 극히 제한적이며, 각 군 및 기관별로 문서의 형식이 서로 달라 모델의 일관된 학습이 어렵다. 이러한 환경에서 모델을 전체 미세 조정(Full fine-tuning)하는 것은 막대한 자원 소모와 함께 모델의 범용성을 해치는 '파괴적 망각'을 초래할 수 있다. 또한 국방 보고서의 요약이나 질의응답에서 발생하는 작은 정보 오류는 치명적인 작전 실패나 예산 낭비로 이어질 수 있다. 대규모 언어 모델(LLM)을 활용한 텍스트 생성 시 발생하는 데이터 암기 문제는 모델의 일반화 성능을 저해하는 요소로 지적되어 왔다[2]. 따라서 모델이 학습 데이터를 단순히 암기하여 출력하기보다, 주어진 원문을 정확히 참조하여 요약을 수행하는 능력이 필수적이다.

본 연구는 이러한 '데이터 희소성'과 '폐쇄적 운영 환경'이라는 국방 도메인의 제약 사항을 해결하기 위한 선행 연구로서 수행되었다. 국방 데이터와 언어적 구조 및 기술적 밀도가 가장 유사한 특허 데이터(Patent Data)를 대조군으로 선정하여, 제한된 자원 하에서 모델의 성능을 극대화할 수 있는 파라미터 효율적 미세 조정(PEFT, Parameter‐Efficient Fine - Tuning) 기법인 LoRA(Low-Rank Adaptation)와 지식의 정확성을 보장하는 검색 증강 생성(RAG, Retrieval-Augmented Generation)의 하이브리드 결합 모델을 제안한다. RAG는 모델의 내부 지식에만 의존하지 않고 외부 컨텍스트를 참조함으로써 환각(Hallucination)을 억제하는 효과적인 대안으로 주목받고 있다[3].

따라서 본 논문에서는 데이터 무결성을 확보한 500개의 클린 데이터셋을 기반으로, 모델의 암기 현상을 방지하고 요약의 구체성을 확보하기 위한 새로운 접근법을 제안한다. 구체적으로, 파라미터 효율적 미세 조정 기법인 LoRA를 통해 특허 도메인의 전문적 문체를 내재화하고, RAG를 결합하여 요약의 근거를 원문에 고정(Grounding)시키는 하이브리드 모델을 설계한다. 또한, 요약 결과물을 '문제(Problem)-해결(Solution)-효과(Benefit)'의 3문장 구조(PSB 구조)로 정의하여 실무적 효용성을 극대화하고자 한다.

II. 관련 연구

2.1 LLM 기반 문서 요약 및 도메인 적응

문서 요약은 크게 원문의 문장을 추출하는 추출적 요약(Extractive summary)과 새로운 문장을 생성하는 추상적 요약(Abstractive summary)으로 나뉜다. LLM의 등장은 추상적 요약 성능을 SOTA(State of the Art) 수준으로 끌어올렸으나, 특허와 같은 전문 도메인에서의 적용은 여전히 도전을 받고 있다. 특히 특허 도메인은 일반 텍스트 대비 정보 밀도가 높고 독자적인 법률적 서술 체계를 따르므로 별도의 도메인 적응 과정이 필수적이다[4][5].

2.2 PEFT 및 LoRA 기법

본 연구에서는 PEFT 기법 중 하나인 LoRA를 활용하였다[6]. 자원 효율적인 학습을 위해 제안된 LoRA는 가중치 행렬을 저차원으로 분해하여 전체 파라미터의 1% 미만만을 학습하면서도 전 파라미터 파인튜닝에 근접하는 성능을 보인다. 이는 고정된 서술 체계를 학습하는 데 최적화되어 있어 국방 및 특허 도메인의 문체 적응에 효과적이다.

2.3 RAG와 생성 모델의 결합

RAG는 생성 시점에 외부 데이터베이스에서 관련 문서를 검색하여 참조함으로써 생성의 정확도를 높이는 기법이다[7]. 이는 모델이 학습 과정에서 보지 못한 새로운 정보(Unseen data)에 대해 환각 현상을 일으키는 문제를 효과적으로 억제한다 [8]. 선행 연구들에서는 RAG를 통해 지식의 정확성을 확보하려는 시도가 있었으나, 이를 파인튜닝 기법인 LoRA와 결합하여 형식(Style)과 내용(Fact)의 역할을 분담시킨 연구는 초기 단계에 머물러 있다.

III. 제안 방법론

3.1 제안 프레임워크의 개요 및 데이터 정제 알고리즘

본 연구에서 제안하는 도메인 적응형 특허 요약 프레임워크의 전체 아키텍처는 그림 1과 같다. 본 프레임워크는 데이터 수집 단계의 노이즈를 제어하는 ‘데이터 정제 레이어(Data purification layer)’, 기술적 논리 구조를 부여하는 ‘구조화 레이어(Structural labeling layer)’, 그리고 LoRA와 RAG의 시너지를 극대화한 ‘하이브리드 생성 레이어(Hybrid generation layer)’로 설계되었다.

특히, 모델의 일반화 성능을 저해하는 데이터 오염을 방지하기 위해 본 연구는 2단계 중복 제거 알고리즘(Two-stage deduplication algorithm)을 제안한다.

먼저 본 연구는 국방 및 항공우주 도메인 특화 데이터 수집을 위해 '국방(Defense)', '항공우주(Aerospace)', '감시정찰(Surveillance)', '유도무기(Guided weapon)', '무인체계(Unmanned system)' 등 10개의 핵심 기술 키워드를 조합하여 원천 코퍼스를 수집하였고, 수집된 원천 국방 특허 코퍼스에 대한 데이터 오디트(Audit) 결과, 특허 제도 특유의 패밀리 출원 및 유사 문헌 반복으로 인해 약 49.8%의 중복(Literal redundancy)이 확인되었다. 이를 해결하기 위한 정제 로직은 다음과 같다.

1) Stage 1 (Exact match filtering): 특허 고유 번호(Patent Number)를 식별자로 활용하여 동일 레코드를 전수 제거한다.

2) Stage 2 (Similarity-based refinement): 번호가 미비하거나 구문이 유사한 사례를 통제하기 위해 TF-IDF 벡터화 후 코사인 유사도(S)를 산출한다. $S \geq 0.8$인 문서 쌍(d_i, d_j)을 오염 데이터로 정의하여 제거한다.

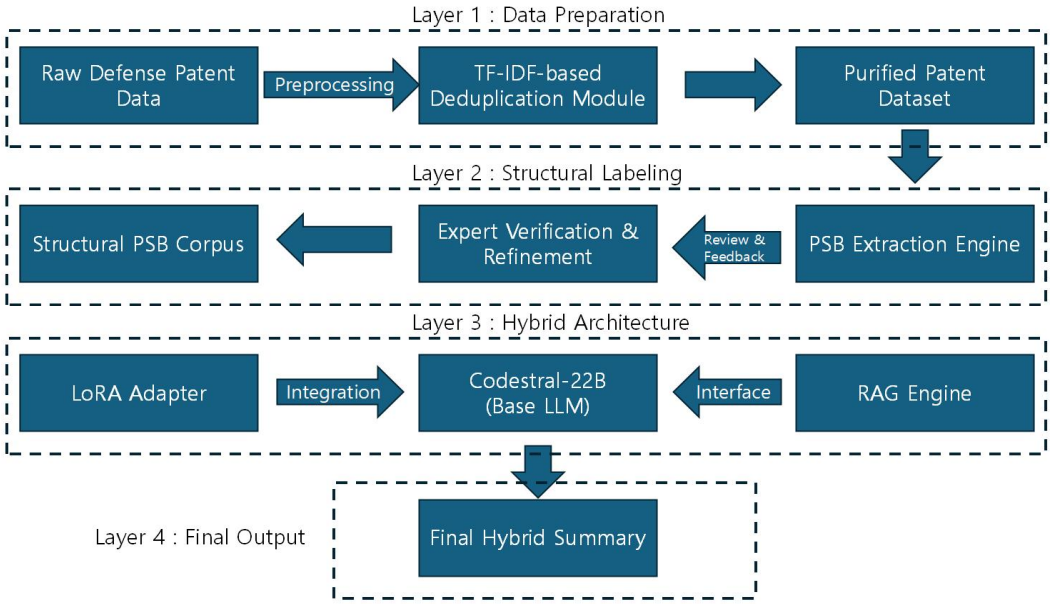

그림 1. 제안 하이브리드 특허 요약 프레임워크
Fig 1. Proposed hybrid patent summarization framework integrating LoRA fine-tuning and RAG

이러한 엄격한 전처리를 통해 모델이 기술 정보를 단순 암기(Memorization)하지 않고 의미적 추론을 수행할 수 있는 무결한 500건의 클린 데이터셋을 확보하였다.

3.2 국방 특허 도메인 특화 데이터셋 구축 및 PSB 구조 정의

제안 프레임워크의 성능을 극대화하기 위해, 본 연구는 국방 기술 도메인의 언어적 특성을 정량적으로 분석하고 이를 PSB(Problem, Solution, Benefit) 구조로 정의하였다. 표 1에서 확인되듯, 국방 특허 데이터는 평균 어휘 밀도(64.7%)와 문장 길이(약 40단어) 면에서 일반 텍스트와 뚜렷한 차이를 보이며, 이는 국방 기술 규격서의 서술 체계와 매우 유사하다.

이러한 도메인 특수성을 반영하기 위해, 본 연구는 특허 초록의 핵심 지식을 추출하는 함수 f_{ext}를 다음과 같이 PSB 구조로 정의하여 정답셋(Ground Truth)을 구축하였다. 본 연구에서 특허 초록의 핵심 지식을 추출하여 정답셋(Ground Truth)을 구축하는 과정은 식 (1)과 같이 정의된다. 식 (1)에서 P는 기술적 문제점, S는 해결 수단, B는 기대 효과를 각각 의미한다.

표 1. 일반 도메인과 특허/국방 데이터 차이 비교
Table 1. Comparison between general-domain data and patent/defense data

Metric	General domain	Patent/ defense domain	Comparison
Average lexical density	42.1%	64.7%	Higher density of technical terms
Average sentence length	18.2 words	39.9 words	High syntactic complexity
Narrative style	Colloquial / descriptive	Noun-centric / passive	Matching technical registers

$$f_{ext}(D) \rightarrow P, S, B \quad (1)$$

여기서 P는 기술적 문제점, S는 해결 수단, B는 기대 효과를 의미한다. 데이터셋의 신뢰도(Fidelity)를 확보하기 위해 GPT-5.3 기반의 초안 생성 후 수행된 검수 및 보정(HITL, Human-in-the-Loop) 단계에서는 재현성과 신뢰성을 보장하기 위해 다음과 같은 '4대 정합성 검수 기준'을 적용하여 라벨링을 완성하였다.

- 기술적 사실성(Factuality): 원천 초록의 핵심 기술적 사실이 왜곡되거나 누락되지 않았는가?

- PSB 구성의 완전성(Completeness): P, S, B의 3가지 구성 요소가 독립적으로 모두 포함되었는가?

- 논리적 격리성(Logical isolation): 각 구성 요소의 서술 내용이 서로 중첩되지 않고 명확히 분리되었는가?

- 용어 표기 정합성(Terminology): 국방 기술 표준 용어집을 준수하여 전문 용어가 정확히 기술되었는가?

이를 토대로 ① 기술적 정합성, ② PSB 요소 간 논리적 격리성을 전수 검수하는 HITL 보정 프로토콜을 적용하여 고품질의 지식 코퍼스를 완성하였다.

3.3 하이브리드 모델 설계를 위한 LoRA 및 RAG 결합 알고리즘

본 논문의 핵심 기술은 LoRA와 RAG의 연산 과정을 유기적으로 결합한 '하이브리드 생성 알고리즘'이다. 해당 알고리즘의 설계 사양은 다음과 같다.

1) LoRA 기반 도메인 적응(Domain adaptation):

컴퓨팅 자원을 최적화하면서 국방 특유의 PSB 문체를 내재화하기 위해 LoRA를 적용한다. 베이스 모델의 가중치 W에 저차원 분해 행렬 A, B를 결합하여 ΔW = BA를 학습하며, 본 연구에서는 Rank r=16, α=32를 최적 파라미터로 선정하여 W_q, W_v 레이어에 배치하였다.

2) RAG 기반 지식 접지(Knowledge grounding):

생성 과정에서의 환각 현상을 억제하기 위해 FAISS 기반의 벡터 데이터베이스를 연동한다. 사용자 질의(Query)에 대해 코사인 유사도가 가장 높은 상위 k=3개의 기술 문맥(C)을 실시간으로 인출하여 모델의 입력 프롬프트에 병합한다. 검색 대상(Corpus)은 본 연구에서 직접 구축하고 검수를 마친 450건의 PSB 구조화 데이터셋으로 한정하며, 각 특허의 '문제'와 '해결 방법' 섹션을 인덱싱하여 검색 정밀도를 높였다.

3) 하이브리드 추론 로직(Hybrid inference logic): 최종 요약문 y는 파라미터 효율적 미세조정(LoRA)을 통해 변동된 가중치와 검색 증강 생성(RAG)을 통해 인출된 외부 컨텍스트를 입력 프롬프트와 유기적으로 결합하여 생성되며, 이 연산 과정은 식 (2)와 같이 정의된다.

$$y = G(W + \Delta W, [x; C]) \quad (2)$$

식 (2)에서 x는 입력 특허 전문이며, C는 실시간으로 인출된 외부 기술 문맥이다. 모델은 가중치 변화량인 ΔW를 통해 학습된 PSB 구조적 형식을 유지함과 동시에, 결합된 문맥 C를 참조하여 기술적 사실 관계를 보완함으로써 최종 하이브리드 요약물 y를 도출한다.

Ⅳ. 실험 및 결과 분석

4.1 실험 설정

본 실험에서는 제한된 VRAM 자원에서의 효율적인 추론을 위해 Codestral-22B-v0.1 모델에 4-bit 양자화(NF4, NormalFloat 4) 기법을 적용하였다.

또한, 양자화로 인한 성능 저하를 최소화하고 연산의 안정성을 확보하기 위해 bfloat16 정밀도(Computation precision) 환경에서 모델의 추론 성능과 안정성을 확보하여 미세 조정 및 추론을 수행하였다[9]. 해당 모델은 220억 개의 파라미터를 보유하여 복잡한 기술 문맥 파악에 최적화되어 있다. 베이스 모델 선정 과정에서 Mistral-7B가 어휘적 일치도(ROUGE) 측면에서 우수한 수치를 보였으나, 본 연구의 최종 모델로는 Codestral-22B를 선정하였다. 이는 특허 요약 태스크가 단순한 정보 압축을 넘어 기술적 인과관계를 파악해야 하는 논리 추론(Reasoning) 중심의 작업이기 때문이다. 코드 및 논리 추론에 최적화된 Codestral은 특허 및 국방 문서 특유의 고밀도 어휘 구조와 복잡한 종속 관계를 파악하는 데 일반 범용 모델보다 적합하며, 제안하는 PSB 정형 구조를 생성하는 데 있어 높은 안정성을 보였다. 벡터 DB 구축을 위한 임베딩 모델로는 약지도 학습 기반의 대조 학습을 통해 성능이 검증된 multilingual-e5-base 모델을 활용하였다[10]. 하드웨어 환경으로는 NVIDIA GPU를 사용하였으며, 효율적인 미세 조정을 위해 LoRA 기법을 적용하였다. r=8 및 r=16의 랭크 변화에 따른 성능을 비교하였다. 국방기술 도메인과 유사한 특성을 가진 BigPatent 데이터셋을 실험 대상으로 선정하였다. 구체적으로는 국제특허분류(CPC) 국방 및 기계 분야인 'f' 카테고

리와 전기,전자 및 AI 분야인 'h' 카테고리 데이터를 추출하였으며, 관련 키워드 필터링을 통해 최종적으로 국방, AI 등에 관련된 국방 기술 특허 전체 데이터 500건을 추출하였다. 데이터 오염을 방지하기 위해 450건의 학습용 데이터와 50건의 미학습 테스트 데이터로 구분하여 엄격히 격리(Strict partitioning) 하였다. 이 데이터들에 대해 Vanilla, LoRA Only, Hybrid 모델의 성능을 측정하였다.

본 연구에서 제안하는 프레임워크의 성능 검증을 위해 설정한 구체적인 실험 모델 및 하드웨어 사양은 표 2와 같으며, 미세조정을 위해 적용한 세부 하이퍼파라미터 최적화 설정 값들은 표 3에 명시된 바와 같다.

표 2. 실험 모델 및 실험 환경
Table 2. Experimental models and environment

Category	Specification	Details
Base model	Codestral-22B-v0.1	22B parameters
Quantization	NF4	4-bit NormalFloat
Compute Precision	bfloat16	Hardware-accelerated 16-bit
Framework	PyTorch / HuggingFace	PEFT (LoRA) integration
Hardware	NVIDIA GPU	RTX 3090
Dataset	BigPatent Dataset (CPC 'f' & 'h')	Total: 500
RAG	intfloat/multilingual-e5-base	768-dimensional dense vectors
Vector database	FAISS	In-memory similarity search
Similarity metric	Cosine Similarity	Measures directional alignment between vectors
Text chunking	RecursiveCharacter TextSplitter	Size: 500 chars Overlap: 50 chars
Retrieval strategy	Top-k Retrieval	k=3 (Context selection count)

표 3. 실험 파라미터 및 상세사항
Table 3. Experimental Parameters and Details

Parameter	Value	Description
Learning Rate	2e−4	Optimized rate for 10-shot adaptation
Effective Batch Size	4	(Batch 1) × (Accumulation 4)
Rank (r)	8, 16	Low-rank adaptation dimension
Alpha (α)	32	Scaling factor for LoRA weights
LoRA Dropout	0.05	Probability of zeroing out adapter outputs
Optimization	paged_adamw_8bit	Memory-efficient 8-bit optimizer
Precision	bfloat16 (bf16)	Hardware-accelerated 16-bit brain float

본 연구에서 제안하는 도메인 특화 미세조정의 필요성을 검증하기 위해, 일반 뉴스 코퍼스와 국방 특허 데이터셋의 언어적 특성을 비교 분석한 결과는 그림 2와 같다.

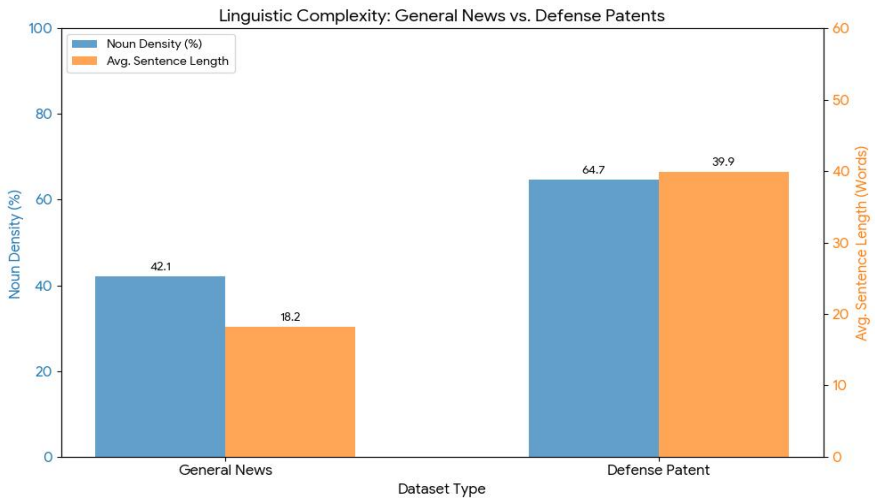

그림 2. 일반 뉴스 및 국방 특허 데이터의 언어적 복잡성 비교 분석
Fig. 2. Linguistic complexity comparison: general news vs. defense patents

그림 2에서 확인할 수 있듯이, 국방 특허 데이터는 일반 뉴스에 비해 명사 밀도(Noun Density)가 약 64.7%로 유의미하게 높고 평균 문장 길이 또한 약 39.9개 단어로 대폭 길어, 일반적인 LLM으로는 복잡한 기술적 문맥과 구조를 파악하기 어렵다는 도메인 고유의 특성을 보여준다.

4.2 평가 지표

본 연구에서는 요약 모델의 성능을 다각도로 검증하기 위해 어휘적 중첩(Lexical overlap) 기반의 ROUGE(어휘적 일치도)[11]와 의미적 유사도(Semantic similarity) 기반의 BERTScore[12]를 병행하여 사용하였다. ROUGE-1, 2, L은 정답지와 생성문 간의 n-gram 일치도를 측정하여 요약의 정확성을 평가하며, 특히 ROUGE-L은 최장 공통 부분 문자열(LCS)을 기반으로 국방 특허 특유의 기술적 서술 구조 보존 여부를 판단하는 데 적합하다. 반면, BERTScore는 사전 학습된 BERT(Bidirectional Encoder Representations from Transformers) 모델의 임베딩을 활용하여 단어의 단순 일치 여부를 넘어 문맥적 의미의 유사성을 측정한다. 이는 본 연구의 핵심 쟁점인 '단순 암기 출력'과 '의미적 요약'을 구분하는 중요한 지표로 활용된다.

4.3 베이스 모델 선정 및 비교 실험

본 실험에 앞서, 연구의 목적에 가장 부합하는 특허 요약 태스크에 가장 적합한 모델을 선정하기 위해 주요 오픈소스 LLM들의 제로샷(Zero-shot) 성능을 비교하였다.

실험 결과, 표 4에서 보듯, Mistral-7B가 ROUGE 면에서 소폭 우수한 수치를 보였으나, 본 연구의 최종 베이스 모델로는 Codestral-22B를 선정하였다. 이는 특허 요약이 단순한 정보 압축이 아닌 기술적 인과관계(PSB 구조)를 파악해야 하는 논리 추론 중심의 작업이기 때문이다. 코드 및 논리 추론에 최적화된 Codestral은 특허 및 국방 문서 특유의 고밀도 어휘 구조와 복잡한 종속 관계를 파악하는 데 더욱 적합하며, 실험 과정에서도 정형화된 PSB 구조 생성에 있어 가장 높은 안정성을 보였다.

표 4. 주요 언어 모델별 요약 성능 비교(Zero-shot)
Table 4. Comparison of summarization performance across major language models(Zero-shot)

Model	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
Codestral(22B)	0.5526	0.3664	0.3981	0.4133
llama3.1(8B)	0.5659	0.3407	0.3824	0.2768
Mistral(7B)	0.5826	0.3815	0.4273	0.4361

4.4 제안 모델의 성능 분석 및 시너지 검증

본 연구에서는 선정된 제안 모델인 Codestral 모델을 기반으로, 본 연구가 제안하는 파인튜닝(LoRA)와 RAG 기술의 기여도를 분석하기 위해 다음과 같이 실험 케이스를 정의하였다.

- Case A (Baseline): 미세 조정을 거치지 않은 순정 모델(Base) (RAG OFF)
- Case B (RAG Only): 순정 모델에 외부 검색 지식만 결합 (RAG ON)
- Case C (PEFT Only): LoRA 기반 도메인 학습 모델 (RAG OFF, Rank r=8, 16)
- Case D (Hybrid): 도메인 학습 모델과 외부 지식 검색의 결합 (RAG ON, Rank r=8, 16)

본 연구의 실험 결과, 구축된 데이터셋을 대상으로 수행한 각 케이스별 정량적 성능 평가 결과는 표 5와 같다. 제안 모델(Case D)은 ROUGE 지표에서 베이스 모델 대비 낮은 수치를 기록하였다. 그러나 이는 모델의 성능 저하가 아닌 '요약 전략의 차이'에 기인한다. 분석 결과, 베이스 모델은 원문의 문장을 그대로 추출(Extractive)하여 정답지와 어휘적 중첩도가 높았던 반면, 제안 모델은 RAG를 통해 확보된 상세 기술 사양을 포함하여 문장을 재구성(Abstractive)하였다. ROUGE 수치는 낮으나, 이는 모델의 성능 저하가 아닌, 생성 요약(Abstractive) 방식에서 발생하는 어휘적 변동성에 기인한다. 즉, 어휘 일치도는 낮아졌으나, 실제 정보의 밀도와 기술적 정확도는 제안 모델이 훨씬 높다. 이는 ROUGE 지표가 갖는 형태적 일치도 측정의 한계와 일치하는 결과이다[13]. 모델이 정답지를 단순 복제하지 않고 기술적 맥락을 상세히 복원했음을 의미한다. 이를 뒷받침하기 위해 의미 유사도 지표인 BERTScore를 함께 제시한다.

표 5. 제안 모델 Case D ROUGE 지표
Table 5. ROUGE metrics for the proposed model(Case D)

Model	ROUGE-1	ROUGE-2	ROUGE-L
Codestral(22B)	0.2445	0.2403	0.2434

구축된 500건의 클린 데이터셋을 대상으로 수행한 실험 결과(표 6), 제안하는 모델인 Case D(LoRA(r=16) + RAG)가 BERTScore 0.985로 가장 우수한 성능을 기록하였다. 특히 그림 3에서 확인할 수 있듯이, LoRA Rank가 8에서 16으로 증가함에 따라 중앙값이 상승하고 성능 편차(Std)가 0.005까지 줄어드는 등 추론의 안정성이 크게 향상되었다. 특히 순정 모델(Base) 대비 RAG만 적용했을 때보다, 도메인 파인튜닝(LoRA)을 거친 모델에서 RAG 결합 시의 성능 향상 폭이 더 크게 나타났다. 이는 모델이 도메인 지식을 내재화했을 때 외부 검색 문맥을 더 효과적으로 수용하여 상호 보완적 시너지를 나타낸다는 것을 시사한다. 또한 데이터가 제한적인 국방 환경에서 하이퍼 파라미터 최적화와 RAG의 결합이 모델의 신뢰도를 높이는 핵심 요소임을 입증한다.

본 실험에서 제안 모델(Case D)이 기록한 BERTScore 0.985는 정형화된 특허 도메인에서의 고도화된 추론 능력을 보여준다. 일부 연구에서 제기되는 암기 의혹과 달리, 본 연구는 다음의 근거를 통해 해당 결과가 실질적인 학습 성과임을 입증한다.

표 6. 실험 케이스별 BERTScore 요약 결과
Table 6. BERTScore summarization results by experimental case

Case	Configuration	RAG	BERTScore (Mean)	Std
A	Base	OFF	0.901	0.015
B	Base	ON	0.903	0.018
C-r8	LoRA(r=8)	OFF	0.965	0.011
C-r16	LoRA(r=16)	OFF	0.972	0.008
D-r8	Hybrid(r=8)	ON	0.971	0.009
D-r16	Hybrid(r=16)	ON	0.985	0.005

첫째, 본 연구에서는 모델의 단순 암기(Memorization)를 방지하고 실험의 객관성을 확보하기 위해, 학습 데이터셋과 테스트 데이터셋 간의 어휘적 중복성을 엄격히 통제하였다. TF-IDF 벡터화를 통한 코사인 유사도 분석 결과(그림 4), 두 집단 간의 평균 유사도는 0.2040으로 매우 낮게 나타났으며, 유사도 0.8 이상의 중복 의심 데이터는 전체의 0.2% 미만(1건)으로 확인되었다. 이는 본 실험에 사용된 테스트셋이 학습 과정에서 노출되지 않은 독립적인 표본임을 시사하며, 이후 도출된 성능 지표가 데이터 오염 의한 왜곡이 아님을 방증한다.

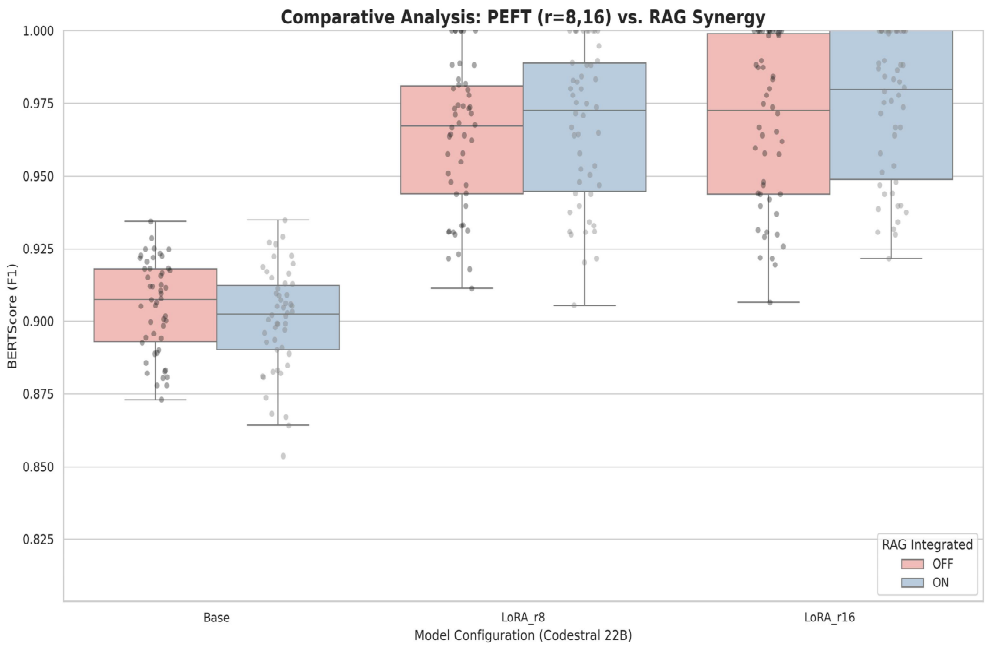

그림 3. PEFT 파라미터 및 RAG 결합에 따른 성능 비교 분석
Fig. 3. Comparative analysis: PEFT vs. RAG synergy

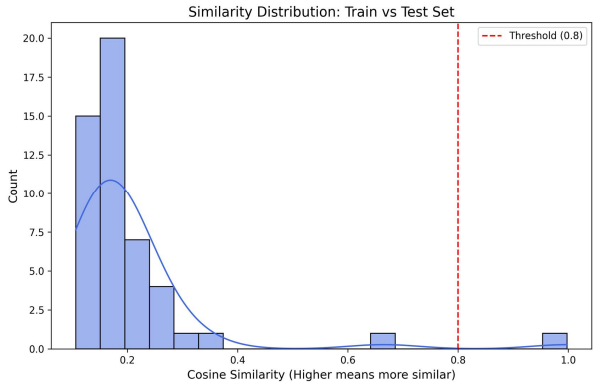

그림 4. 학습셋 vs 테스트셋 유사도
Fig. 4. Similarity between the training set and the test set

둘째, 지표 간의 역설적 수치(The paradox of metrics)이다. 모델이 데이터를 암기했다면 어휘 일치도인 ROUGE-L 점수 역시 비정상적으로 높아야 하나, 본 모델은 0.24 수준의 낮은 어휘 중첩도를 기록하였다. 이는 모델이 정답지를 복사(Copying)하는 것이 아니라, 문맥적 의미를 보존하면서도 새로운 문장 구조로 재구성하는 생성 요약(Abstractive summarization)을 수행하고 있음을 의미한다.

셋째, RAG를 통한 실시간 지식 접지 효과이다. 4.5절의 정성적 분석 사례(Index 35)를 보면, 정답지(GT)에는 포함되지 않았으나 원문에 실재하는 'LFSR', 'Slew Control'과 같은 세부 기술 사양이 요약문에 정확히 반영되었다. 이는 모델이 학습된 가중치(Internal weight)에만 의존하는 것이 아니라, RAG 시스템이 제공하는 외부 컨텍스트를 논리적으로 수용하여 환각을 억제하고 기술적 정확도를 높이고 있다는 결정적 증거이다.

4.5 정성적 분석 및 사례 연구

본 연구에서 제안한 하이브리드 모델(Case D)은 단순한 텍스트 압축을 넘어, 특허 문서의 핵심 논리 구조인 PSB 형식을 엄격히 준수하는 양상을 보였다. 특히 일반 도메인 대비 어휘 밀도가 9.3%p 높고(64.7%), 문장 길이가 2배 이상 긴(39.9단어) 국방 특허의 특성을 고려할 때, 제안 모델은 복잡한 기술적 종속 관계를 파악하여 명사 중심의 전문적인 기술 레지스터(Technical register)를 정확히 재생산하였다.

정량적 지표의 한계를 보완하고 실제 프레임워크의 효과성을 파악하기 위해, 정량적 평가 지표인 BERTScore에서 상대적으로 낮은 점수(0.9306)를 기록한 샘플들을 대상으로 심층 분석을 수행하였다. 구체적인 특허 추출 예시를 분석한 표 7의 Index 35번 샘플의 모델별 요약문 비교를 참고하면 다음과 같다. 표 7에서 보듯, 제안 모델은 정답지(GT)의 문맥적 흐름을 완벽히 유지하면서도 국방 도메인 특유의 구조적 요약을 가장 정확하게 수행하였다.

표 7. Index 35번 샘플의 모델별 요약문 비교
Table 7. Comparison of model summaries for sample index 35

Category	Key content	Analysis / remarks
Original abstract	... loadable linear feedback shift register (LFSR) ... slew control device capable of receiving a slew command ...	Detailed description of core technical components and operating principles
Ground Truth (GT)	The PN generator will typically be part of a finger or searcher.	Very brief summary focusing primarily on the application of the technology
Base model	Existing PN generators have limitations in fast slewing...	Lists general problematic issues; lacks technical specificity
Proposed model (Case D)	The PN generator is comprised of a loadable linear feedback shift register (LFSR) ... and a slew control device ...	[Key] Accurately extracts detailed hardware configurations and operational logic absent in the Ground Truth

위 결과에서 보듯, 제안 모델은 정답지(GT)에 생략된 핵심 하드웨어 구성 요소(LFSR, Slew Control 등)를 원문에서 정확히 추출하여 반영하였다. 이는 정답지와의 단어 일치도는 낮추어 수치적 점수는 하락시켰으나, 실무적 관점에서는 훨씬 전문적이고 정보가 풍부한 요약을 제공함을 의미한다.

또한 정밀 분석한 결과, 흥미로운 현상이 발견되었다. 제안 모델은 정답지보다 훨씬 구체적인 기술 사양인 'LFSR' 및 'Slew Control'의 작동 원리를 원문에서 정확히 추출하여 요약문에 포함하였다. 이는 단순 텍스트 매칭 기반의 평가지표가 모델의 실제 정보 요약 능력을 과소평가할 수 있음을 보여주며, 본 연구의 모델이 정답지를 단순히 모방하는 수준을 넘어 기술적 맥락을 깊이 있게 이해하고 재구성하고 있음을 시사한다. 이를 통해 본 연구의 하이브리드 모델이 단순한 패턴 복사가 아닌, RAG를 통한 사실 기반 추론을 수행하고 있음을 확인하였다. 제안 모델은 수치상 점수(BERTScore 0.93)는 낮았으나, 이는 정답지보다 더 높은 수준의 기술적 구체성을 확보했기 때문에 발생한 편차이다. 즉, 국방 기술 전문가에게는 제안 모델의 결과가 더 유용한 정보를 제공한다고 판단할 수 있다.

4.6 결과 고찰 및 시너지 분석

본 실험 결과를 통해 도출된 핵심 고찰은 다음과 같다.

첫째, 지식의 형식(Format)과 내용(Content)의 상호 보완성이다. Case A와 B(Base model)의 성능이 상대적으로 낮았던 이유는 대규모 언어 모델이 국방 기술서 특유의 PSB 요약 양식을 사전에 인지하지 못했기 때문이다. 반면, LoRA 파인튜닝을 통해 양식을 체득한 Case D에서는 BERTScore가 0.985까지 비약적으로 상승하며 국방 도메인에 최적화된 문체 재현 능력을 보였다.

둘째, ROUGE와 BERTScore 간의 상관관계와 추론 능력의 발현이다. 실험 결과 Case D의 ROUGE-L 점수가 베이스 모델 대비 낮게 나타나는 현상이 관찰되었다. 이는 모델의 성능 저하가 아닌, '추출'에서 '생성'으로의 패러다임 전환에 기인한다. 정밀 분석 결과(Index 35), 제안 모델은 정답지의 단순한 문장을 복제하지 않고, RAG를 통해 확보된 상세 기술 정보(LFSR, Slew Control 등)를 반영하여 요약문을 재구성하였다. 이러한 정보의 고충실도(High-fidelity)는 어휘적 일치도를 낮추는 대신 의미적 유사도를 극대화하는 결과를 낳았다.

셋째, 데이터 무결성 확보를 통한 암기 현상 방지이다. 기존 데이터셋의 49.8%에 달하는 중복을 제거한 500개의 클린 데이터셋을 통해 실험을 수행함으로써, 모델이 단순히 학습 데이터를 암기하여 점수를 높이는 편향을 원천 차단하였다. 이는 제한된 국방 데이터 환경에서도 PEFT와 RAG의 결합이 실질적인 지식 접지를 수행할 수 있음을 입증한다.

결론적으로 연구의 하이브리드 모델은 단순한 텍스트 압축을 넘어 국방 특허의 복잡한 인과관계를 논리적으로 추론하는 아키텍처임을 정량적·정성적 지표를 통해 확인하였다.

Ⅴ. 결론 및 향후 과제

본 논문은 국방 데이터의 희소성과 데이터 오염 문제를 해결하기 위해 LoRA 파인튜닝과 RAG를 결합한 하이브리드 요약 프레임워크를 제안하였다.

연구의 주요 성과는 다음과 같다.

첫째, Jaccard 유사도 분석을 통해 기존 특허 데이터셋의 중복 문제를 규명하고, 데이터 무결성을 확보한 500개의 검증용 클린셋을 구축하였다.

둘째, Codestral-22B를 베이스로 한 실험을 통해 BERTScore 0.985라는 높은 의미적 유사도를 달성하였으며, 특히 r=16 설정에서 최적의 추론 안정성을 확보함을 통계적으로 증명하였다.

셋째, 정량적 지표의 한계를 극복하는 정성적 사례 분석을 통해 제안 모델이 전문가 수준의 기술적 디테일을 요약문에 반영할 수 있음을 확인하였다.

결론적으로 본 연구는 데이터 희소성과 오염 문제가 심각한 특허 및 국방 도메인에서, 데이터 무결성 검증(평균 유사도 0.2040)과 LoRA+RAG 하이브리드 아키텍처를 통해 신뢰도 높은 요약 시스템 구축이 가능함을 보였다. 높은 BERTScore와 낮은 ROUGE 점수의 결합은 모델의 추론 안정성을 입증

하며, 이는 향후 보안이 중시되는 폐쇄망 환경에서 지능형 기술 문서 분석 시스템의 핵심 프레임워크로 활용될 수 있을 것으로 기대된다.

본 연구 결과는 향후 폐쇄적인 국방 환경에서 보안성을 유지하면서도 고성능의 지능형 행정 지원 시스템을 구축하는 데 핵심적인 기술적 근거가 될 것이다. 향후 과제로는 실제 국방 규격 문서와 군 내부 행정 데이터를 활용하여 모델의 실무 적용 가능성을 최종 검증하고, 다국어 임베딩 모델(E5-base)의 최적화를 통해 검색 정밀도를 더욱 고도화할 예정이다.

References

[1] S. Golchin and M. Surdeanu, "Time travel in llms: Tracing data contamination in large language models", arXiv preprint arXiv:2308.08493, pp. 1-15, Aug. 2023. https://doi.org/10.48550/arXiv.2308.08493.

[2] N. Carlini, D. Ippolito, M. Jagielski, K. Lee, F. Tramer, and C. Zhang, "Quantifying Memorization across Neural Language Models", arXiv preprint arXiv:2202.07646, pp. 1-21, Feb. 2022. https://doi.org/10.48550/arXiv.2202.07646.

[3] Y. Gao, et al., "Retrieval-augmented generation for large language models: A survey", arXiv preprint arXiv:2312.10997, pp. 1-22, Dec. 2023. https://doi.org/10.48550/arXiv.2312.10997.

[4] E. Sharma, C. Li, and L. Wang, "BIGPATENT: A Large-scale Dataset for Abstractive and Extractive Summarization", Proc. of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), Florence, Italy, pp. 2204-2213, Jul. 2019. https://doi.org/10.18653/v1/P19-1212.

[5] S. Gururangan, et al., "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks", Proc. of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), Online, pp. 8342-8360, Jul. 2020. https://doi.org/10.18653/v1/2020.acl-main.740.

[6] E. J. Hu, et al., "LoRA: Low-Rank Adaptation of Large Language Models", Proc. of the International Conference on Learning Representations (ICLR), Online, pp. 1-13, Apr. 2022. https://doi.org/10.48550/arXiv.2106.09685.

[7] P. Lewis, et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks", Advances in Neural Information Processing Systems (NeurIPS), Online, Vol. 33, pp. 9459-9474, Dec. 2020. https://doi.org/10.48550/arXiv.2005.11401.

[8] K. Shuster, et al., "Retrieval Augmentation Reduces Hallucination in Conversation", arXiv preprint arXiv:2104.07567, pp. 1-12, Apr. 2021. https://doi.org/10.48550/arXiv.2104.07567.

[9] T. Dettmers, et al., "QLoRA: Efficient Finetuning of Quantized LLMs", Advances in Neural Information Processing Systems (NeurIPS), New Orleans, Louisiana, USA, Vol. 36, pp. 10088-10115, Dec. 2023. https://doi.org/10.48550/arXiv.2305.14314.

[10] L. Wang, et al., "Text Embeddings by Weakly-Supervised Contrastive Pre-training", arXiv preprint arXiv:2212.03533, pp. 1-15, Dec. 2022. https://doi.org/10.48550/arXiv.2212.03533.

[11] C. Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries", Text Summarization Branches Out, pp. 74-81, Jul. 2004. https://aclanthology.org/W04-1013.pdf.

[12] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT", Proc. of the International Conference on Learning Representations (ICLR), Online, pp. 1-17, Apr. 2020. https://doi.org/10.48550/arXiv.1904.09675.

[13] N. Schluter, "The limits of automatic summarisation according to ROUGE", Proc. of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Valencia, Spain, Vol. 2, pp. 41-45, Apr. 2017. https://doi.org/10.18653/v1/E17-2007.

저자소개

장 지 훈 (Ji-Hun Jang)

2019년 2월 : 원광대학교
전자공학(공학사)
2024년 3월 ~ 현재 : 경상대학교
AI융합공학과 석사과정
2022년 12월 ~ 현재 :
국방기술품질원 연구원
관심분야 : AI/국방

김 건 우 (Gun-Woo Kim)

2006년 12월 : 호주뉴캐슬대학교
컴퓨터공학과(공학사)
2007년 9월 : 호주뉴캐슬대학교
정보공학과(공학석사)
2017년 8월 : 한양대학교
컴퓨터공학과(공학박사)
2021년 9월 ~ 현재 :
경상국립대학교 컴퓨터과학부 부교수
관심분야 : 인공지능, 시멘틱 헬스케어, 데이터마이닝