

폐쇄망 환경에서의 항공·방산 기술문서 대상 BM25+FAISS
하이브리드 검색을 이용한 RAG 검색 파라미터 최적화 연구안희석*¹, 이호성*², 김건우**A Study on RAG Retrieval Parameter Optimization using
BM25+FAISS Hybrid Retrieval for Aerospace and Defense
Technical Documents in a Closed Network EnvironmentHui-Seok Ahn*¹, Ho-Seong Lee*², and 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

항공·방산 기술문서는 약어·수치·표 구조가 복잡하고 외부망 접근이 제한되어, 기존 클라우드 기반 RAG 시스템을 그대로 적용하기 어렵다. 이에 본 연구에서는 KIPRIS 특허 500건·NASA NTRS 보고서 700건으로 도메인 데이터셋을 구축하고, BM25와 FAISS를 결합한 완전 오프라인 하이브리드 검색 구조를 구현하여 하이브리드 가중치(α)·전처리·Top-k·유사도 임계값 조합에 따른 Recall@k, nDCG@k, MRR 변화를 정량 분석하였다. 그 결과, Dense 중심 가중치($\alpha=0.2$)에서 BM25 단일 대비 64.5%, FAISS 단일 대비 26.6% 향상된 성능을 확인하였으며(nDCG@5=0.9948, $p<0.05$), $k=8$ 에서는 모든 α 조건에서 완전 재현율을 달성하였다. 본 연구는 클라우드 의존 없이도 하이브리드 RAG 구조가 실용적 수준의 검색 성능을 낼 수 있음을 보였으며, 이를 통해 폐쇄망 환경에 적합한 파라미터 조합의 실증적 근거를 제시하였다.

Abstract

Closed-network deployments of RAG systems cannot rely on cloud-based retrieval, yet most existing work assumes unrestricted API access. We built a fully offline BM25+FAISS hybrid retrieval pipeline and tested it on 500 KIPRIS patents and 700 NASA NTRS reports (~39,900 chunks). Varying hybrid weight (α), preprocessing, top-k, and similarity threshold, we measured Recall@k, nDCG@k, and MRR across configurations. The dense-favoring setting ($\alpha=0.2$) reached nDCG@5 = 0.9948 and Recall@5 = 0.905 — a 64.5% gain over standalone BM25 and 26.6% over FAISS ($p<0.05$) — while $k=8$ yielded perfect recall across all α values. This work provides a reproducible offline retrieval framework, domain-specific validation, and practical parameter guidance for security-constrained RAG deployments.

Keywords

hybrid retrieval, BM25, FAISS, retrieval-augmented generation, technical document retrieval, closed network

* 경상국립대학교 AI융합공학과
- ORCID¹: <https://orcid.org/0009-0008-2668-5542>
- ORCID²: <https://orcid.org/0009-0005-6168-6450>
** 경상국립대학교 컴퓨터공학과 교수(교신저자)
- ORCID: <https://orcid.org/0000-0001-5643-4797>

• Received: Mar. 30, 2026 Revised: Apr. 15, 2026, Accepted: Apr. 18, 2026
• Corresponding Author: Gun-Woo Kim
Dept. of Computer Science and Engineering, College of IT Engineering,
Gyeongsang National University, Jinju, Korea
Tel.: +82-55-772-3323, Email: gunwoo.kim@gnu.ac.kr

I. 서 론

LLM(Large Language Model) 기반 응용이 확산되면서 외부 문서를 실시간으로 참조하는 RAG(Retrieval-Augmented Generation) 구조가 다양한 분야에 도입되고 있다[1]. RAG는 검색 단계에서 관련 청크를 추출한 뒤 생성 단계에 제공함으로써 환각(Hallucination) 문제를 줄이고 도메인 특화 질의응답의 신뢰성을 높인다. 기술문서·특허·연구보고서처럼 명시적 근거가 요구되는 분야에서 특히 수요가 크며, 최근에는 FAISS(Facebook AI Similarity Search)·Chroma·Milvus 등 벡터 검색엔진과 BM25(Best Matching 25) 계열 키워드 검색을 결합한 하이브리드 구조가 주목받고 있다.

그런데 이러한 연구의 대부분은 클라우드 접근이 자유로운 환경과 공개 QA 데이터셋을 전제로 한다. 항공·방산 현장처럼 외부망이 차단된 폐쇄망 환경에서는 API 호출 자체가 불가능하며, 도메인 문서에 빈번한 약어·수치·단위·표 구조는 의미 검색(밀집 검색)만으로 포착하기 어렵다. 밀집 검색은 자연어 질의에는 강하지만, "T-700 엔진 압축비 17.0:1" 같은 수치·약어 중심 질의에서는 임베딩 공간의 유사도가 의미를 제대로 반영하지 못한다. 반대로 BM25는 정확한 키워드 매칭에는 유리하나, 표현이 조금만 달라져도 관련 문서를 놓친다. 결국 두 방식을 결합한 하이브리드 접근이 현실적인 대안이 되지만, 이 결합이 도메인 특화 문서에서 실제로 얼마나 효과적인지, 어떤 파라미터 조합에서 최적 성능이 나오는지 실증한 연구는 드물다. 특히 α 가중치·전처리·Top-k·유사도 임계값 각각이 검색 품질에 미치는 영향을 체계적으로 분석한 사례는 찾기 어렵다.

이에 본 연구는 두 가지를 목표로 삼았다. 첫째, 외부 API 없이 로컬에서만 동작하는 BM25+FAISS 하이브리드 검색 구조를 Python으로 직접 구현하는 것이다. 둘째, 실제 항공·방산 도메인 문서 KIPRIS(Korea Intellectual Property Rights Information Service) 특허 500건, NASA NTRS(NASA Technical Reports Server) 보고서 700건을 대상으로 α 가중치·전처리·Top-k·유사도 임계값 조합에 따른 Recall@k, nDCG(normalized Discounted Cumulative Gain)@k,

MRR(Mean Reciprocal Rank) 변화를 측정하여, 폐쇄망 환경에 적합한 최적 파라미터 기준을 도출하는 것이다. 기존 하이브리드 RAG 연구와 비교했을 때 본 연구의 차별점은 Whoosh·FAISS·BGE-M3(BAAI General Embedding-M3) 등 오픈소스 도구만으로 완전 오프라인 파이프라인을 구성하고, 표·단위·약어가 혼재하는 실제 산업 문서 기반 데이터셋을 직접 구축했으며, 네 변수 조합의 영향을 통계적으로 확인했다는 데 있다. 세부 결과는 IV장에서 다룬다.

II. 관련 연구

RAG 시스템에서 검색 단계의 품질은 최종 응답의 정확도를 직접 결정한다. 특히 기술문서·특허·연구보고서처럼 수치와 명시적 근거가 요구되는 영역에서는 이 의존성이 더욱 두드러진다. Lee와 Bae[2]는 RAG 아키텍처 전반을 정리하며 LLM 환각 완화를 위한 검색 최적화의 중요성을 강조한 바 있다. 최근에는 키워드 기반 희소 검색과 임베딩 기반 밀집 검색을 결합한 하이브리드 구조가 검색 품질 향상의 핵심 수단으로 부상하고 있다. 이하에서는 하이브리드 검색의 발전 동향, 가중치 융합 전략, 평가 지표 영향, 폐쇄망·도메인 특화 연구 순으로 검토한다.

2.1 하이브리드 검색의 발전 동향

하이브리드 검색은 희소 검색과 밀집 검색을 결합하여 검색 정확도와 재현율의 균형을 동시에 높이려는 접근이다. Doan et al.[3]은 텍스트 임베딩과 지식 그래프 임베딩을 결합한 하이브리드 검색기를 제안하여 파인튜닝된 밀집 검색(BGEit) 대비 HitRate 최대 8.1%, MRR 최대 8.7%의 검색 성능 향상을 보고하였다. Sun et al.[4]은 BM25와 FAISS를 결합한 하이브리드 RAG 시스템을 적용하여 비 RAG 기준 모델의 F1 점수 5.45%를 42.21%로 향상시켰으며, 이는 도메인 특화 질의응답에서 하이브리드 검색의 효용을 표 1과 같이 실증적으로 보여준다. Stuhlmann et al.[5]은 Biomedical QA 실험에서 BM25와 MedCPT를 결합하여 정확도 0.90, 재현율

0.90을 달성함으로써 밀집/희소 융합의 효용을 입증하였다. 국내에서도 Yang과 Choi[6]이 BM25의 문맥 이해 한계를 ResNet(Residual Network)-Transformer 결합으로 보완하여 nDCG를 9.38% 향상시킨 바 있다.

표 1. RAG 구성별 성능 비교

Table 1. Performance comparison by RAG configuration

Configuration	EM(%)	Precision (%)	Recall (%)	F1(%)
Non-RAG	0.00	3.11	33.59	5.45
RAG+Re-ranker +Few-shot +Ensemble	20.25	47.29	56.18	42.21

2.2 하이브리드 검색의 가중치 조정 및 융합 기법

하이브리드 검색의 핵심 변수는 두 검색기의 점수를 결합하는 방식과 α 가중치의 조정 전략이다. Hsu 와 Tzeng[7]은 질의별로 α 값을 자동 조정하는 DAT(Dynamic Alpha Tuning) 프레임워크를 제안하여 Precision@1과 MRR을 개선하였고, Lu et al.[8]은 BM25와 T5(Text-to-Text Transfer Transformer) Dual-Encoder 기반 검색기의 결과를 cross-encoder reranker로 재정렬하는 HYRR(Hybrid Infused Reranking) 방식을 통해 일관된 성능 향상을 보고하였다. 이러한 기법들은 대규모 코퍼스와 학습 기반 모델을 전제하므로, 폐쇄망 환경에서의 적용에는 제약이 따른다. 본 연구에서는 재현성 확보를 우선시하여, 외부 학습 모델이나 동적 조정을 배제하고 z-score 정규화와 고정 α 구조($\alpha \in \{0.2, 0.5, 0.8\}$)를 채택하였다.

2.3 RAG 성능 평가 지표 및 하이퍼파라미터 영향

RAG의 검색 단계는 보통 Recall@k, nDCG@k, MRR 지표를 기준으로 평가되며, 검색 후보 수 (Top-k), 유사도 임계값, 전처리 적용 여부 등이 검색 품질에 직접적인 영향을 미친다. Ammar et al.[9]은 Chroma 및 FAISS 기반 RAG 시스템을 대상으로 다양한 청킹 전략과 하이퍼파라미터가 성능과 효율성에 미치는 영향을 체계적으로 분석하였다. 그 결

과, 표 2와 같이 고정 길이 청킹 중 1024토큰·128오버랩 구성이 문맥 정확도 0.904, 문맥 재현율 0.935, 충실도 0.947로 가장 높은 성능을 기록하였으며, 실행시간도 6.02초로 가장 짧아 품질과 효율성의 균형점으로 평가되었다. 반면, 청크 크기 확대(2048, 4096)나 의미 기반 청킹 방식은 성능 저하 또는 실행시간 증가로 이어지는 경향을 보였다. 본 연구는 Ammar et al.[9]의 청킹 전략별 성능 분석 및 2.2절의 α 조정 기법 연구를 바탕으로, $\alpha(0.2, 0.5, 0.8)$, Top-k(3, 5, 8), 임계값(0.20, 0.25, 0.30), 전처리 적용 여부를 실험 변수로 설정하였다.

표 2. 청킹 전략별 성능 비교

Table 2. Performance comparison by chunking strategy

Chunking strategy	Context precision	Context recall	Faithful-ness	Time (s)
Naive(1024,128)	0.904	0.935	0.947	6.02
Naive(2048,128)	0.897	0.890	0.930	10.7
Naive(4096,128)	0.838	0.827	0.892	9.28
Semantic	0.62-0.90	0.64-0.80	0.72-0.87	8-9

2.4 폐쇄망 환경 및 도메인 특화 RAG 연구 동향

보안 제약이 강한 폐쇄망 환경에서의 RAG 연구는 상대적으로 드물지만, 국방·항공·의료 등 고신뢰 분야에서 그 필요성이 높아지고 있다. Cheng et al.[11]은 지식 지향 RAG 서버를 통해 검색·생성·통합 구조를 체계화하고, 재현성 중심 설계가 지식형 RAG의 핵심임을 제시하였다. Yuan et al.[10]은 CRAG(Comprehensive RAG) 워크숍에서 웹 데이터 정제, 속성 예측, 지식 그래프 결합을 포함한 하이브리드 RAG 프레임워크를 제시하여 복합 질의 환경에서 우수한 성능을 보고하였다.

국내에서는 Yi와 Kim[12]이 폐쇄망 환경에서 동작 가능한 RAG 기반 질의응답시스템을 설계하여, 한국어 문장 임베딩을 활용한 지식 데이터베이스 구축과 기밀 정보 유출 방지 구조를 제시하였다. 그러나 검색 단계에서 희소-밀집 결합 구조나 파라미터 최적화 실험은 수행되지 않아, 본 연구와 상호보완적 관계에 있다. Nam과 Kim[13]은 군사 영어 약어·두문자어 5,006건을 대상으로 외부 참고자료의

문장화 추출 기법을 적용하여 RAG 모델의 도메인 특화 성능을 향상시켰으며, 이는 기술문서에서 약어와 전문 용어의 정확한 검색이 중요함을 뒷받침한다. Baek과 Kim[14]은 한국어 규제 문서의 구조적 복잡성을 지식 그래프 기반 문서 변환 기법으로 해소하는 수정된 RAG 프레임워크를 제안하여, 도메인 특화 문서 처리에서 RAG의 적용 가능성을 확장하였다. 최근 Jeon et al.[15]는 국방 법령·훈령 데이터를 대상으로 19개 공개 한국어 LLM을 파인튜닝하여, 제한된 GPU 환경에서도 실용적 성능을 확보할 수 있음을 확인하고, RAG 도입을 향후 연구 방향으로 제시하였다. 이는 보안 제약 환경에서의 도메인 특화 AI 시스템 구축 수요가 증가하고 있음을 보여준다.

본 연구는 이러한 연구 방향성과 궤를 같이하면서도, 공개 데이터(KIPRIS, NTRS)를 이용하여 보안 친화적 로컬 실험 환경을 구현하였다. 이를 통해 실제 산업 보안 조건에서도 완전 재현 가능한 하이브리드 RAG 구조를 실증적으로 검증하였다.

2.5 종합 비교 및 본 연구의 위치

기존 연구들은 하이브리드 검색이 RAG 성능 향상에 효과적임을 일관되게 보고하였으나, 적용 환경과 데이터 구성 측면에서 한계를 지닌다. 대부분의 실험은 클라우드 접근이 가능한 환경에서 공개 벤치마크를 대상으로 수행되었으며, 폐쇄망 환경이나 도메인 특화 기술문서에 대한 체계적 실험은 부재하다. 본 연구는 로컬 기반의 완전 오프라인 하이브리드 RAG 구조를 구현하고, 항공·방산 기술문서를 대상으로 독자적인 도메인 데이터셋을 구축하였으며, $\alpha \cdot \text{Top-k} \cdot \text{Threshold}$ 전처리 변수의 상호작용을 정량적으로 분석하여 밀집 검색 중심 가중에서의 최적 조합을 도출하였다는 점에서 기존 연구와 차별화된다.

III. 폐쇄망 환경을 위한 BM25+FAISS 하이브리드 검색 구조 설계

본 장에서는 본 연구의 실험 환경, 데이터 구축

및 전처리 과정, 하이브리드 검색 구조 설계, 평가 지표, 실험 변수 및 절차를 구체적으로 기술한다.

3.1 연구 환경 및 시스템 구성

실험은 Windows 11 기반 로컬 워크스테이션(NVIDIA GeForce RTX 4080 12GB, Intel i9-13950HX, 64GB RAM)에서 수행되었으며, Python 3.11과 CUDA 12.1로 소프트웨어 환경을 구성하였다. 외부 API 호출, 클라우드 연결 및 상용 LLM 사용이 불가능한 보안 제약 환경, 즉 폐쇄망 환경을 가정하여 모든 실험을 100% 로컬에서 수행하였다. 검색 단계의 검색 구조는 표 3과 같이 세 가지 방식으로 구성하였다.

표 3. 검색 단계의 검색 방식 구성

Table 3. Retrieval stage search configurations

No.	Method	Features	Library
(1)	BM25	Sparse keyword search	whoosh
(2)	FAISS	Dense embedding similarity	faiss, bge-m3
(3)	Hybrid	Z-score + α weighted fusion	custom module

3.2 데이터셋 구축 및 전처리

데이터셋은 표 4와 같이 항공·방산 분야 기술문서를 대상으로 구축하였다. KIPRIS에서 항공기 구조 및 장비 관련 IPC(International Patent Classification) 코드(B64C, B64D, B64F)에 해당하는 한글 특허 500건을 수집하고[16], NASA NTRS에서 기술 보고서 700건을 확보하여[17] 총 1,200건의 문서를 확보하였다.

표 4. 데이터셋 구성 요약

Table 4. Dataset composition summary

Source	Documents	Language	Chunks	Note
KIPRIS patent	500	ko	2,969	xlsx metadata
NASA NTRS	700	en	36,933	PyMuPDF +OCR
Total	1,200	ko/en	~39,900	Avg 49.67 pages

처리 과정은 메타데이터 노이즈 제거, 구조 보존, 단위 통일, 다국어 용어 정규화, 청킹의 5단계로 표 5와 같이 구성하였으며, 전처리를 통해 전체 청크 수가 표 6과 같이 약 43,000개에서 39,900개로 약 7.2% 감소하였다.

표 5. 전처리 단계별 규칙

Table 5. Preprocessing rules by stage

Stage	Rule	Example	Purpose
(a) Meta removal	Remove doc no., footnotes	STD-STR-002 → (removed)	Noise reduction
(b) Structure	Preserve table/section tags	Table 3 → <TABLE>	Context preservation
(c) Unit norm.	Unify units (in→mm)	5 in. → 127 mm	Search consistency
(d) Language norm.	Merge bilingual terms	KO/EN term → English	Multilingual QA
(e) Chunking	800 tok, 150 overlap	2400 tok → 3 chunks	Context continuity

표 6. 전처리 전후 데이터셋 통계

Table 6. Dataset statistics before/after preprocessing

Item	Before	After	Change (%)	Note
Total documents	1,200	1,200	0	KIPRIS +NTRS
Total chunks	~43,000	39,900	-7.2	800tok sliding
Avg chunks/doc	~36	33.3	-7.6	Per 1,200 docs
Structure	Partial	Complete	++	Table/section preserved

3.3 하이브리드 검색 구조

최소 검색에는 순수 Python으로 구현되어 별도의 서버 프로세스가 불필요한 Whoosh 라이브러리를 표 7과 같이 채택하였다.

검색 구조는 세 가지 요소로 구성된다. 최소 검색은 BM25 알고리즘($k_1=1.2$, $b=0.75$)을 Whoosh로 구현하였고, 밀집 검색은 BGE-M3 임베딩과 Cosine Similarity 기반의 FAISS IndexFlatIP를 사용하였다.

하이브리드 융합은 BM25와 FAISS의 상위 Top-k 후보 점수를 z-score로 정규화한 뒤, α 가중치로 선형 결합($\alpha \times \text{BM25} + (1-\alpha) \times \text{FAISS}$)하여 최종 하이브리드 점수를 산출하는 방식이다. α 는 0.2, 0.5, 0.8의 세 가지 값을 실험적으로 비교하였다. 하이브리드 검색의 구현 구조는 LangChain의 설계 패턴을 참고하였다[19].

표 7. 주요 BM25 구현체 비교

Table 7. Comparison of major BM25 implementations

Feature	Whoosh	Pyserini	Elasticsearch
Language	Python	Java (Lucene)	Java (Server)
Operation	Local embedded	CLI/Server	Server process
Dependencies	None	JRE required	JRE + REST
Platform	Win/Linux	Linux-focused	Linux-focused
Air-gap ready	Very easy	Limited	Difficult
Scale	$\leq 1e5$ docs	$\geq 1e6$ docs	Large distributed

3.4 평가 지표 및 실험 변수

검색 성능 평가는 Recall@k, nDCG@k, MRR 세 지표를 표 8과 같이 사용하였다. 여기서 @k는 상위 k개의 검색 결과만을 대상으로 지표를 산출함을 의미하며, k 값에 따라 검색 시스템의 정밀도와 재현율 간 균형이 달라진다. 실험 변수는 표 9와 같이 α 가중치, Top-k, 유사도 임계값, 전처리 적용 여부 of 네 가지로 설정하였다.

표 8. 평가 지표의 정의

Table 8. Definition of evaluation metrics

Metric	Definition	Purpose
Recall@k	Ratio of relevant docs in top-k	Retrieval coverage
nDCG@k	Normalized discounted cumulative gain	Ranking precision
MRR	Mean reciprocal rank of first hit	Response efficiency

표 9. 실험 변수 설정

Table 9. Experimental variable settings

No.	Variable	Values	Description
(1)	α weight	{0.2, 0.5, 0.8}	Sparse-Dense ratio
(2)	Top-k	{3, 5, 8}	Candidate count
(3)	Threshold	{0.20, 0.25, 0.30}	Cosine similarity min
(4)	Preprocessing	{Applied, Not}	Structure preservation

3.5 실험 절차

실험은 표 10과 같이 데이터 수집, 정제, 청크 분할, 임베딩 생성, 인덱스 구축, 하이브리드 검색, 성능 평가의 7단계 파이프라인으로 수행하였다.

표 10. 실험 파이프라인 (7단계)

Table 10. Experiment pipeline (7 stages)

Stage	Task	Tool	Output
①	Data collection	requests, pdfplumber	raw_texts/
②	Data cleaning	Python preprocessing	cleaned_texts/
③	Chunk splitting	custom chunker	chunks.jsonl
④	Embedding	sentence-transformers	embeddings.npy
⑤	Index build	faiss, whoosh	faiss.index, whoosh/
⑥	Hybrid search	hybrid_search.py	search_results.jsonl
⑦	Evaluation	eval_retrieval.py	results.csv, graphs/

3.6 정답 라벨링 및 관련성 판단 기준

평가 라벨은 총 40개 쿼리를 대상으로, 하이브리드 검색($\alpha=0.5$, $k=20$, $\text{threshold}=0.25$)을 수행하여 상위 20개 청크를 수집하고 자동으로 부여하였다. 각 쿼리별로 상위 1~5위를 관련($\text{rel}=1$), 6~20위를 비관련($\text{rel}=0$)으로 분류하여 표 11과 같이 총 800개의 이진 라벨을 구성하였다. 평가 지표는 이진 관련성 기반으로 계산되며, nDCG의 이득 함수는 이진 라벨

특성에 따라 정답이 상위권에 존재할 경우 이론적으로 1.0에 근접하는 값이 산출된다.

표 11. 이진 분류 기준

Table 11. Binary classification criteria

Grade	rel	Definition
Relevant	1	Hybrid result rank 1-5
Not relevant	0	Hybrid result rank 6-20

다만, 자동 라벨은 하이브리드 검색($\alpha=0.5$) 결과에 기반하여 생성되었으므로, 평가 대상인 하이브리드 검색 구조와 라벨 생성 기준 간에 순환논리의 가능성이 존재한다. 이를 검증하기 위해, BM25 단독($\alpha=1.0$)과 FAISS 단독($\alpha=0.0$)으로 동일한 라벨링 절차를 독립적으로 수행하여 표 12와 같이 3자 간 라벨 일치도를 분석하였다.

표 12. 자동 라벨 안정성 분석

Table 12. Auto-label stability analysis (40 queries average)

Comparison	Cohen's κ	Top-5 overlap	Interpretation
Hybrid vs FAISS	0.68	72.0%	Substantial
Hybrid vs BM25	0.21	33.5%	Fair
BM25 vs FAISS	-0.11	5.5%	Poor
3-way common	—	5.5%	Few common hits

분석 결과, Hybrid 라벨은 FAISS 단독 라벨과 높은 일치도($\kappa=0.68$)를 보인 반면, BM25 단독 라벨과는 낮은 일치도($\kappa=0.21$)를 보였다. 이는 $\alpha=0.5$ 조건의 하이브리드 검색에서 밀집 검색(FAISS)의 기여가 상대적으로 크게 반영되었음을 의미한다. 한편, 희소 검색과 밀집 검색 간 상위 5개 문서 겹침률이 5.5%로 매우 낮아, 두 검색 방식이 상이한 관련 문서를 포착하고 있음이 확인되었다. 이러한 결과는 자동 라벨이 밀집 검색 관점에 편향되어 있을 가능성을 시사하나, 검색 방식 간 상대적 성능 비교에서는 동일한 라벨 기준이 적용되었으므로 경향성의 유효성은 유지된다. 순환논리에 따른 절대적 성능 수치의 한계는 V절에서 상세히 논의한다.

또한, 라벨 품질의 독립적 검증을 위해 5개 쿼리(카테고리당 1~2개)를 표본 추출하여 각 쿼리의 상위 20개 청크(총 73건, 텍스트 미확인 27건 제외)에 대해 도메인 전문가가 수작업으로 관련성을 판단하였다. 자동 라벨과 수작업 라벨 간 일치도를 Cohen's κ 로 산출한 결과 $\kappa = 0.15$ 로 나타났다.

다만, 자동 라벨의 Precision은 88.9%로 높아 '관련'으로 판정된 문서의 신뢰성은 확보된 반면, Recall은 31.4%에 그쳐 상위 5개 이외의 관련 문서를 충분히 포착하지 못하는 한계가 확인되었다. 이는 상위 5개만을 관련으로 분류하는 이진 라벨링 방식의 구조적 특성에 기인하며, 절대적 관련성 평가에는 한계가 있으나 검색 방식 간 상대적 성능 비교의 유효성은 유지된다.

3.7 재현성 확보 전략

폐쇄망 환경에서도 동일한 결과를 재현할 수 있도록, 표 12와 같이 실험 환경 고정, 데이터 분할 규칙 명시, 실행 명령 공개, 3자 라벨 일치도 분석 및 수작업 표본 검증(5개 쿼리, Cohen's κ)을 통한 자동 라벨 타당성의 정량적 검증, 그리고 데이터 무결성 검증(SHA-256 checksum) 등의 항목을 명시적으로 관리하였다.

IV. 실험 및 결과

4.1 실험 개요

실험은 3개 α 값 $\times$ 3개 Top-k 값의 총 9가지 조합으로 구성하였으며, 40개 질의에 대해 각 질의별 상위 20개 청크를 평가 대상으로 하여 그림 1과 같

이 총 800개의 이진 라벨을 기반으로 수행하였다.

4.2 α 가중치 변화에 따른 성능 비교

하이브리드 검색의 핵심 변수인 α 는 희소 검색과 밀집 검색의 기여도를 결정한다. 하이브리드 점수는 $\alpha \times \text{BM25} + (1-\alpha) \times \text{FAISS}$ 로 계산되므로, α 가 커질수록 BM25의 비중이 높아지고, 작을수록 FAISS의 비중이 커진다.

표 13에서 확인되듯, $\alpha=0.2$ 에서 $\text{nDCG}@5$ 가 가장 높았으며, α 가 증가할수록 nDCG 는 감소하는 경향을 보였다. α 값에 따른 상세 트레이드오프 분석은 4.6절에서 기술한다.

표 13. α 가중치별 검색 성능 비교 ($k=5$)

Table 13. Search performance by α weight ($k=5$)

α	Recall@5	nDCG@5	MRR	Feature
0.2	0.905	0.9948	1.0	Semantic query advantage
0.5	0.940	0.9903	1.0	Balanced (default)
0.8	0.940	0.9836	0.9875	Keyword/numeric advantage

4.3 전처리 적용 여부에 따른 성능 변화

전처리 적용이 검색 품질에 미치는 영향을 확인하기 위해, 표 14와 같이 $\alpha=0.5$, $k=5$ 의 기본 설정 조건에서 전처리 적용·미적용 조건을 비교하였다. 전처리 적용 시 $\text{nDCG}@k$ 가 15.3%, MRR이 13.8% 향상되었으며, 이는 메타데이터 노이즈 제거와 구조 보존이 검색 정밀도에 유의미한 영향을 미침을 보여준다.

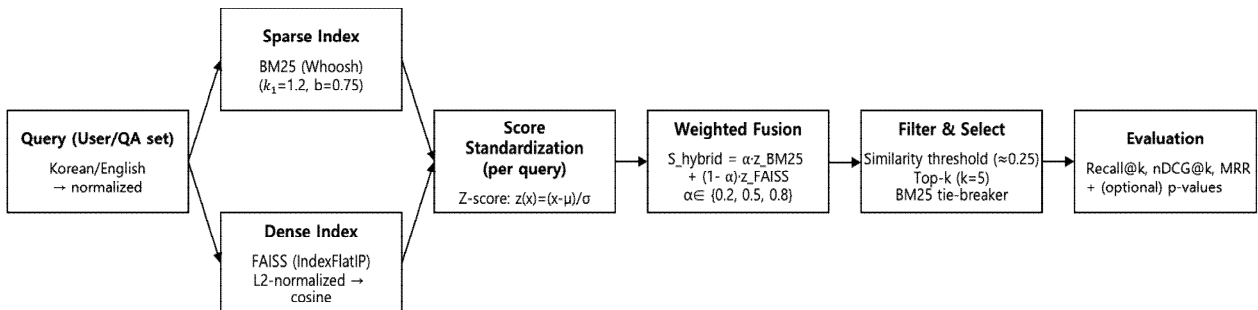


그림 1. 질의-검색-평가 단계의 하이브리드 RAG 처리 흐름
Fig. 1. Hybrid RAG processing flow for query-search-evaluation

표 14. 전처리 적용 여부에 따른 성능 비교 ($\alpha=0.5$, $k=5$)

Table 14. Performance with/without preprocessing

Preprocessing	Recall@5	nDCG@5	MRR	Impact
Not applied	0.68	0.59	0.58	Metadata noise
Applied	0.72	0.68	0.66	Context preserved

다만, 본 실험은 전처리 미적용 조건을 포함하므로, 전처리가 적용된 상태에서 수행한 다른 절의 실험 결과와 절대 수치를 직접 비교할 수 없다.

4.4 Top-k 변화에 따른 성능 비교

검색 후보 수(Top-k)의 변화가 검색 성능에 미치는 영향을 분석한 결과는 표 15와 같다. $k=3$ 에서는 높은 정밀도($nDCG=0.9960$)를 보이나 재현율이 56.0%에 그치며, $k=5$ 에서는 정밀도와 재현율의 균형을 달성한다. $k=8$ 에서는 모든 관련 문서가 검색 결과에 포함되어 완전 재현율($Recall=1.0$)을 그림 2와 같이 달성하였다.

표 15. Top-k 값에 따른 검색 성능 비교 ($\alpha=0.2$)Table 15. Search performance by Top-k ($\alpha=0.2$)

Top-k	Recall@k	nDCG@k	MRR	Feature
3	0.560	0.9960	1.0	High precision, limited recall
5	0.905	0.9948	1.0	Balanced performance
8	1.000	0.9932	1.0	All relevant docs returned

4.5 유사도 임계값에 따른 영향 분석

세 가지 임계값(0.20, 0.25, 0.30) 조건에서 동일한 검색 성능이 표 16과 같이 관측되었다. 이는 z-score 정규화 후 선형 α 가중치 결합 과정에서 임계값의 영향이 상쇄되었기 때문으로 해석된다.

표 16. Threshold에 따른 검색 성능 비교

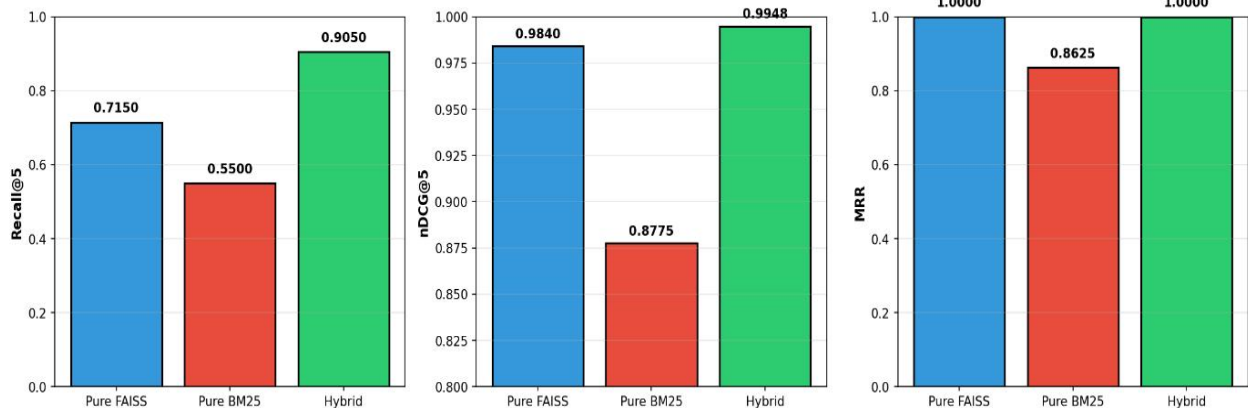
Table 16. Search performance by threshold

Threshold	Recall@5	nDCG@5	Note
0.20	0.905	0.9948	Identical results (irrelevant)
0.25	0.905	0.9948	
0.30	0.905	0.9948	

4.6 α 파라미터 시각화 분석 및 최적화

$\alpha=0.2$ (BM25 20%, FAISS 80%), $\alpha=0.5$ (BM25 50%, FAISS 50%), $\alpha=0.8$ (BM25 80%, FAISS 20%)의 세 가지 구성에서 $nDCG@k$ 와 $Recall@k$ 를 측정하였다. $k=5$ 기준으로, $nDCG$ 는 α 값이 증가함에 따라 일관되게 감소한다($\alpha=0.2$: 0.9948 $\rightarrow$ $\alpha=0.5$: 0.9903 $\rightarrow$ $\alpha=0.8$: 0.9836). $nDCG$ 성능 차이는 $\alpha=0.2$ 와 $\alpha=0.8$ 간 약 1.1%로 통계적으로 유의한 수준(paired t-test, $p<0.05$)을 보였다. 이는 항공·방산 기술문서 환경에서 BGE-M3 기반 밀집 검색의 기여가 클수록 순위 정확도가 향상됨을 시사한다.

Baseline vs Hybrid Search Performance (k=5)

그림 2. Top-k 값에 따른 검색 성능 비교 ($\alpha=0.2$)Fig. 2. Search performance by Top-k ($\alpha=0.2$)

한편, $\alpha=0.2$ 에서 $\text{Recall}@5(0.905)$ 가 $\alpha=0.5(0.940)$ 보다 낮은 것은, FAISS 비중이 높아지면서 BM25가 포착하던 키워드 일치 문서가 상위권에서 밀려난 결과로 해석되며, 이는 nDCG와 Recall 간 상충 관계를 보여준다.

4.7 검색 방식별 종합 비교

세 가지 검색 구조(BM25, FAISS, Hybrid)의 종합 비교 결과를 표 17에 요약하였다. 실험은 $\alpha=0.2$ (BM25 20%, FAISS 80%), Top-k=5 기준으로 수행하였다.

표 17. 검색 방식별 검색 성능 비교 (Top-k=5)
Table 17. Search performance comparison by method (Top-k=5)

Method	Recall@5	nDCG@5	MRR	Feature
BM25 ($\alpha=1.0$)	0.550	0.8775	0.8625	Keyword-focused
FAISS ($\alpha=0.0$)	0.715	0.9840	1.0	Semantic-focused
Hybrid ($\alpha=0.2$)	0.905	0.9948	1.0	Best overall

표 17에 따르면, 하이브리드 검색($\alpha=0.2$, BM25 20%+FAISS 80%)은 BM25 및 FAISS 대비 모든 평가 지표에서 우수한 성능을 보였다(그림 3). Recall@5 기준으로는 FAISS 대비 26.6%(0.715 $\rightarrow$ 0.905), BM25 대비 64.5%(0.550 $\rightarrow$ 0.905)의 향상을 확인하였다(paired t-test, $p<0.05$). nDCG@5와 MRR 또한 단일 검색 방식보다 높은 값을 기록하였으며,

이는 하이브리드 구조가 키워드 기반 검색의 명시적 매칭 장점과 의미적 문맥 기반 검색의 유연성을 동시에 확보함을 의미한다.

4.8 결과 해석 및 시사점

실험 결과는 네 가지 관점에서 시사점을 제공한다. 첫째, 폐쇄망 환경에서도 오픈소스 도구만으로 독립적 하이브리드 검색 구조의 구현이 가능함을 실증하였다. 둘째, 전처리를 통한 구조 보존이 nDCG를 15.3%, MRR을 13.8% 향상시켰으므로, 도메인 특화 기술문서의 형식적 특성을 반영하는 것이 검색 품질 확보의 전제 조건임을 확인하였다. 셋째, $\alpha=0.2$ 에서의 최적 성능은 BGE-M3 임베딩이 항공·방산 기술문서의 의미적 유사성을 효과적으로 포착되, BM25의 키워드 매칭이 이를 보완하는 역할을 수행함을 보여준다. 넷째, 이상의 결과를 종합하면 항공·방산 기술문서 기반 폐쇄망 RAG 시스템의 권장 파라미터 조합으로 $\alpha=0.2$, Top-k=5, Threshold=0.25를 제시할 수 있다. 다만, nDCG@5가 0.99 이상으로 매우 높게 나타난 것은 평가 라벨의 구성 방식과 관련이 있다. 본 실험에서는 상위 5개 문서를 관련, 나머지를 비관련으로 분류하는 이진 방식을 사용하였으므로, 관련 문서가 상위권에 포함되기만 하면 nDCG가 높게 산출되는 구조적 특성이 있다. 이에 따라 보고된 절대 수치는 라벨 방식의 영향을 받은 상한 추정값으로 해석해야 하며, 본 실험 결과는 검색 방식 간 상대적 성능 차이의 경향성에 초점을 두고 해석하는 것이 타당하다.

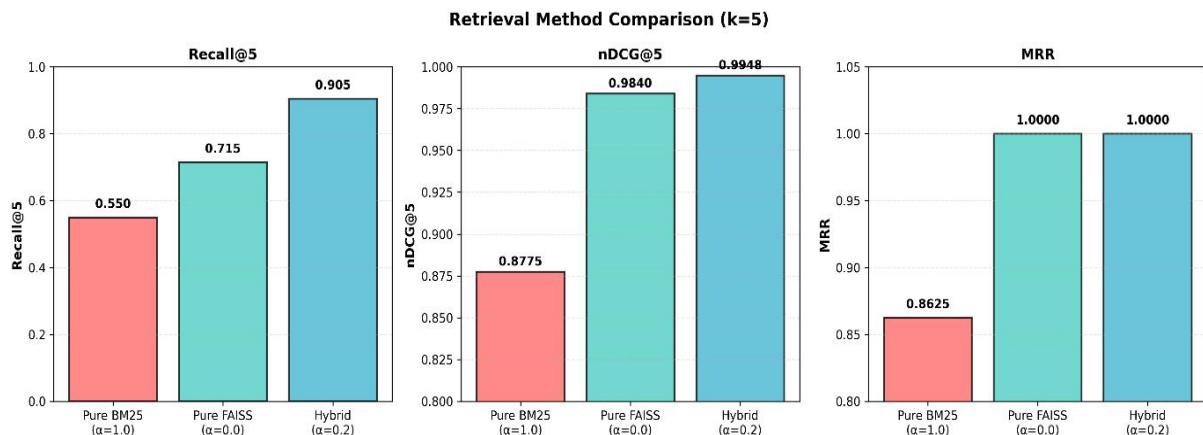


그림 3. 순수 검색 방식 대비 하이브리드 검색의 성능 비교 (k=5)

Fig. 3. Performance comparison: standalone vs. hybrid retrieval (k=5)

4.9 요약

본 장의 실험 결과를 종합하면 다음과 같다. 하이브리드 검색 구조($\alpha=0.2$)는 BM25 단일 대비 Recall@5를 64.5%, FAISS 단일 대비 26.6% 향상시켰으며, nDCG@5=0.9948로 최고 성능을 기록하였다. 전처리 적용은 nDCG를 15.3%, MRR을 13.8% 개선하였고 Top-k 증가는 재현율 향상에 직접적 효과를 보였다. 유사도 임계값은 z-score 정규화 구조에서 실질적 영향을 미치지 않았다. 이러한 결과는 폐쇄망 환경에서 회소 검색과 밀집 검색을 적절히 조합하는 전략이 단일 검색 방식 대비 높은 검색 성능을 확보할 수 있음을 실증적으로 보여준다.

V. 결론 및 향후 과제

항공·방산 기술문서를 대상으로 BM25+FAISS 하이브리드 검색 구조를 구현하고, 폐쇄망 환경에서 재현 가능한 RAG 검색 품질 향상 방안을 실험으로 확인하였다. $\alpha=0.2$ 조건에서 nDCG@5=0.9948을 기록했으며, BM25 단일 대비 Recall@5 64.5%, FAISS 단일 대비 26.6% 향상이 확인되었다($p<0.05$). Top-k를 8로 늘리면 모든 α 조건에서 완전 재현율에 도달했고, 유사도 임계값은 z-score 정규화 구조에서 성능에 실질적 영향을 미치지 않았다.

본 연구의 기여는 네 가지다. 첫째, 오픈소스 도구만으로 API 호출 없이 동작하는 완전 오프라인 하이브리드 검색 파이프라인을 구현했다. 둘째, 표·약어·수치가 혼재하는 실제 산업 문서 기반 데이터셋을 직접 구축하고 구조 보존형 전처리를 적용하여 도메인 특화 평가 체계를 갖췄다. 셋째, α ·전처리·Top-k·Threshold 네 변수 조합의 영향을 통계적으로 분석해 최적 설정($\alpha=0.2$, Top-k=5, Threshold=0.25)을 도출했다. 넷째, 이 결과를 토대로 보안 제약 환경에서 RAG 시스템을 구축하는 실무자를 위한 파라미터 선택 기준을 제시하였다.

한계도 분명하다. 공개 데이터(KIPRIS, NASA NTRS)만으로 구성된 데이터셋이 실제 사내 보안문서를 완전히 대변하기는 어렵고, QA 라벨을 하이브리드 검색 결과로 자동 생성했기 때문에 순환논리

가능성이 남는다. 실제로 III장의 3자 라벨 일치도 분석에서 자동 라벨이 FAISS 단독($k=0.68$)에는 가깝고 BM25 단독($k=0.21$)과는 멀어, 보고된 수치는 상한 추정값으로 봐야 한다. BGE-M3 단일 모델·800토큰 고정 청크·40개 쿼리라는 조건도 일반화의 제약이다. 향후에는 폐쇄망 환경에서 문서가 지속적으로 축적되는 운용 조건을 고려하여, 신규 문서가 추가될 때 기존 BM25 및 FAISS 인덱스를 전체 재구축하지 않고 변경분만 반영하는 증분 갱신 방안을 마련하고, 다단계 검색 구조, 검색·생성 통합 평가, 임베딩 모델 비교, 동적 최적화, 산업 현장 실증으로 연구를 확장할 계획이다.

요컨대 본 연구는 클라우드 없이도 하이브리드 RAG가 실용적 검색 성능을 낼 수 있음을 실제 도메인 데이터로 보였으며, 항공·방산 도메인의 보안 제약 환경에서 RAG 시스템을 설계하는 실무자에게 구체적인 파라미터 기준을 제공한다는 점에서 의의가 있다.

References

- [1] P. Lewis, et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks", *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 33, pp. 9459-9474, Dec. 2020, <https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html>.
- [2] E.-B. Lee and H. Bae, "A Survey on Recent Research Trends of Retrieval-Augmented Generation Technology", *Journal of KIPS*, Vol. 13, No. 9, pp. 429-436, Sep. 2024. <https://doi.org/10.3745/TKIPS.2024.13.9.429>.
- [3] N. N. Doan, A. Härmä, R. Celebi, and V. Gottardo, "A Hybrid Retrieval Approach for Advancing Retrieval-Augmented Generation Systems", *Proc. of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, Trento, Italy, pp. 397-409, Oct. 2024. <https://aclanthology.org/2024.icnls-1.41>.

- [4] H. Sun, Y. Wang, and S. Zhang, "Retrieval-Augmented Generation for Domain-Specific Question Answering: A Case Study on Pittsburgh and CMU", arXiv preprint arXiv:2411.13691, Nov. 2024. <https://doi.org/10.48550/arXiv.2411.13691>.
- [5] L. Stuhlmann, M. A. Saxer, and J. Fürst, "Efficient and Reproducible Biomedical Question Answering using Retrieval-Augmented Generation", Proc. of the 2025 IEEE Swiss Conference on Data Science (SDS), Zürich, Switzerland, pp. 154-157, Jun. 2025, <https://doi.org/10.1109/SDS66131.2025.00029>.
- [6] H.-J. Yang and I.-Y. Choi, "Improving BM25 Ranking Algorithm Using ResNet-Transformer Model", Journal of KSCI, Vol. 29, No. 11, pp. 67-77, Nov. 2024. <https://doi.org/10.9708/jksci.2024.29.11.067>.
- [7] H.-L. Hsu and J. Tzeng, "DAT: Dynamic Alpha Tuning for Hybrid Retrieval in Retrieval-Augmented Generation", arXiv preprint arXiv:2503.23013, Mar. 2025. <https://doi.org/10.48550/arXiv.2503.23013>.
- [8] J. Lu, K. Hall, J. Ma, and J. Ni, "HYRR: Hybrid Infused Reranking for Passage Retrieval", Proc. of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Torino, Italia, pp. 8528-8534, May 2024. <https://aclanthology.org/2024.lrec-main.748>.
- [9] A. Ammar, A. Koubaa, O. Nacar, and W. Boulila, "Optimizing Retrieval-Augmented Generation: Analysis of Hyperparameter Impact on Performance and Efficiency", arXiv preprint arXiv:2505.08445, May 2025. <https://doi.org/10.48550/arXiv.2505.08445>.
- [10] Y. Yuan, C. Liu, J. Yuan, G. Sun, S. Li, and M. Zhang, "A Hybrid RAG System with Comprehensive Enhancement on Complex Reasoning", Proc. of the 2024 KDD Cup Workshop on Retrieval-Augmented Generation, Barcelona, Spain, Aug. 2024. <https://doi.org/10.48550/arXiv.2408.05141>.
- [11] M. Cheng, et al., "A Survey on Knowledge-Oriented Retrieval-Augmented Generation", arXiv preprint, arXiv:2503.10677, Mar. 2025. <https://doi.org/10.48550/arXiv.2503.10677>.
- [12] G.-W. Yi and S.-K. Kim, "Design of a Question-Answering System based on RAG Model for Domestic Companies", Journal of KSCI, Vol. 29, No. 7, pp. 81-88, Jul. 2024. <https://doi.org/10.9708/jksci.2024.29.07.081>.
- [13] M.-G. Nam and N.-G. Kim, "Sentencification Extraction Methodology for Improving RAG-based Model Performance", Journal of KSCI, Vol. 29, No. 12, pp. 29-39, Dec. 2024. <https://doi.org/10.9708/jksci.2024.29.12.029>.
- [14] J.-H. Baek and J.-R. Kim, "Optimizing LLM Performance for Regulatory Text Comprehension Based on an Advanced RAG Framework", Journal of KIIS, Vol. 35, No. 1, pp. 65-71, Feb. 2025. <https://www.kci.go.kr/kciportal/ci/sereArticleSearch/ciSereArtiView.kci?sereArticleSearchBean.artiId=ART003173995>.
- [15] J.-H. Jeon, et al., "A Study on the Fine-Tuning Feasibility of Public Korean Large Language Models for the Defense Domain", Defense and Security, Vol. 7, No. 2, pp. 234-256, Dec. 2025. <https://doi.org/10.23425/defsec.2025.7.2.234>.
- [16] KIPRIS, Patent Database: B64C, B64D, B64F Codes, <https://www.kipris.or.kr>. [accessed: Mar. 10, 2026]
- [17] NASA STI Repository, Technical Reports Collection, <https://ntrs.nasa.gov>. [accessed: Mar. 10, 2026]
- [18] Meta AI / PyTorch, Sentence-Transformers and FAISS Integration Guide, <https://pytorch.org>, <https://github.com/facebookresearch/faiss>. [accessed: Mar. 10, 2026]
- [19] LangChain Documentation, Hybrid Retrieval Implementation, <https://python.langchain.com>. [accessed: Mar. 10, 2026]

저자소개

안 희 석 (Hui-Seok Ahn)



2009년 2월 : 경상대학교
기계항공공학부(공학사)
2018년 6월 ~ 현재 :
(주)한국항공우주산업
디지털엔지니어링팀 책임연구원
관심분야 : 항공우주 분야 AI 응용,
산업용 지능형 시스템 설계, AI
기반 항공기 설계 및 운용 최적화

이 호 성 (Ho-Seong Lee)



2017년 2월 : 경상대학교
항공우주시스템공학과(공학사)
2026년 2월 : 경상대학교
AI융합공학과(공학석사)
2017년 1월 ~ 현재 :
(주)한국항공우주산업
디지털엔지니어링팀 선임연구원

관심분야 : Aerospace Digital Twin, Model-Based System
Engineering, Reinforcement Learning

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 컴퓨터공학과 부교수
관심분야 : 인공지능, 시멘틱 헬스케어, 데이터마이닝