



Received: 2026/08/19
Revised: 2026/08/30
Accepted: 2026/09/09
Published: 2026/09/30

***Corresponding Author:**

Cho-A Song

Tel: +82-31-326-9379

E-mail: choa.song@ligdna.com

Abstract

본 연구는 기술교범 기반 검색증강생성(retrieval-augmented generation, RAG) 적용을 위한 검색단계에서 BM25, 벡터 및 하이브리드 검색방식의 정비절차 검색성능을 비교·분석을 수행하였다. 5종의 기술교범으로부터 구축한 788개의 문서 청크를 대상으로 직접질의와 표현변형질의 각 30건, 총 60건의 평가질의를 구성하였다. 벡터 검색에는 Qwen3-Embedding-8B 모델을 적용하였으며, 하이브리드 검색은 BM25와 벡터 검색의 순위를 Reciprocal Rank Fusion(RRF)으로 결합하였다. 검색성능은 Hit@1, Hit@3, Hit@5 및 Mean Reciprocal Rank(MRR)를 이용하여 평가하였다. 실험 결과 전체 질의의 MRR은 BM25 0.2535, 벡터 검색 0.6355, 하이브리드 검색 0.6415로 나타났다. 벡터 및 하이브리드 검색은 BM25보다 통계적으로 유의하게 높은 성능을 보였으며, 전체 질의에서는 벡터와 하이브리드 검색 간 유의한 차이가 확인되지 않았다. 반면 질의 유형을 구분하여 분석한 결과 직접질의에서는 하이브리드 검색이 상대적으로 높은 성능을 보였으며, 표현변형질의에서는 벡터 검색이 높은 성능을 나타냈다.

This study compared BM25, vector and hybrid retrieval methods for maintenance procedure search in the retrieval stage of RAG applied to technical manuals. A corpus comprising 788 document chunks was constructed from five manuals, and 60 evaluation queries, including 30 direct and 30 paraphrased queries, were prepared. Qwen3-Embedding-8B was used for vector retrieval, whereas a hybrid retrieval approach combined BM25 and vector rankings using reciprocal rank fusion (RRF). Performance was evaluated using Hit@1, Hit@3, Hit@5, and mean reciprocal rank (MRR). The results showed MRR values of 0.2535 for BM25, 0.6355 for vector retrieval, and 0.6415 for hybrid retrieval. Both vector and hybrid methods significantly outperformed BM25, with no significant difference between them overall. In terms of query type, hybrid retrieval performed better on direct queries, whereas vector retrieval performed better on paraphrased queries.

Keywords

기술교범(Technical Manual),
정비절차(Maintenance Procedure),
정비지원(Maintenance Support),
기술자료(Technical Data),
RAG(Retrieval Augmented Generation)

RAG 적용을 위한 기술교범 정비절차 검색방식의 성능 비교 연구: BM25·벡터·하이브리드 검색을 중심으로

Performance Comparison of Maintenance Procedure Retrieval Methods for RAG Application to Technical Manuals: Focusing on BM25, Vector, and Hybrid Retrieval

송초아^{1*}, 손은호¹, 윤연아¹, 최영원¹, 박경덕², 정준², 김경록²

¹LIG디펜스엔에어로스페이스 미래전장IPS연구소 선임연구원

²LIG디펜스엔에어로스페이스 미래전장IPS연구소 수석연구원

Cho-A Song^{1*}, Eun-Ho Son¹, Yeon-A Yoon¹, Young-Won Choi¹,
Kyoung-Deok Park², Jun Jeong², Kyungrok Kim²

¹Research engineer, Future Warfare Integrated Product Support R&D Lab,
LIG Defense&Aerospace

²Chief research engineer, Future Warfare Integrated Product Support R&D Lab,
LIG Defense&Aerospace

1. 서론

무기체계와 복합장비의 운용·정비에 사용되는 기술교범에는 장비 구성, 운용방법, 예방정비, 고장탐구, 점검 및 구성품 교환절차 등 다양한 기술정보가 포함된다. 정비자는 장비 점검이나 고장조치 과정에서 필요한 정보를 교범에서 찾아야 하지만, 문서 분량이 많고 유사 용어가 반복되면서 원하는 정비절차를 신속히 찾기 어려운 경우가 발생한다.

최근에는 대규모 언어모델(LLM)에 외부 문서를 결합하는 검색증강생성(RAG)이 문서 기반 질의응답 방법으로 활용되고 있다. RAG는 질의와 관련된 문서를 검색해 생성모델의 입력으로 제공하는 구조로, 외부 지식을 활용할 수 있다는 특징이 있다[1]. 기술교범 분야에서도 이러한 접근이 시도되고 있으며, 특히 군용 정비

매뉴얼에 RAG를 적용하여 자연어 질의 기반 정비정보 제공 가능성이 제시된 바 있다[6]. 다만 실제 활용을 위해서는 생성모델뿐 아니라 관련 정비정보를 검색단계에서 정확히 찾는 과정이 중요하다.

대표적인 문서검색 방법으로는 단어 일치 기반의 BM25와 문장 의미를 벡터로 표현하여 유사도를 계산하는 벡터 검색이 있다[2,3]. 두 방식은 서로 다른 특성을 가지며 이를 결합한 하이브리드 검색도 활용된다[4,7].

기술교범 검색에서는 구성품명처럼 정확한 용어가 중요한 경우와 자연어 기반 질의가 혼재되기 때문에 검색방식의 특성이 성능에 큰 영향을 미칠 수 있다.

이에 본 연구에서는 RAG의 전체 과정 중 검색단계에 초점을 두고 기술교범 정비절차 검색에 BM25, 벡터, 하이브리드 검색을 적용하여 성능을 비교하였다. 또한 직접질의와 표현변형질의를 구성하여 질의 형태에 따른 검색방식별 성능 변화를 분석함으로써, 향후 기술교범 기반 RAG에서 적절한 검색방식 선택을 위한 실험적 근거를 제시하고자 한다.

2. 관련 연구

2.1 RAG와 문서검색

RAG는 사용자의 질의와 관련된 외부 정보를 검색한 뒤 이를 생성모델의 입력문맥으로 제공하는 구조이다[1]. 따라서 생성모델 성능과 함께 검색 단계의 정확성이 전체 시스템 성능에 영향을 미친다.

문서검색은 어휘적 일치 기반의 희소검색과 의미적 유사도를 이용하는 밀집검색으로 구분된다. BM25는 질의어의 출현빈도와 문서집합 내 희소성을 이용해 관련도를 계산하는 대표적 희소검색 방식이다[2]. 반면 밀집검색은 질의와 문서를 벡터로 표현하여 의미적 유사도를 기반으로 검색한다. Dense passage retrieval은 이러한 벡터 기반 검색이 의미 정보를 활용해 관련 구절을 찾을 수 있음을 보인 연구이다[3].

최근에는 다국어 및 다양한 검색 태스크를 지원하는 임베딩 모델이 개발되고 있으며, Qwen3-Embedding 계열은 0.6B, 4B, 8B 모델을 제공한다[5]. 본 연구에서는 한국어 기술교범의 의미기반 검색을 위해 Qwen3-Embedding-8B를 적용하였다.

2.2 하이브리드 검색

BM25와 벡터 검색은 서로 다른 기준으로 문서 관련성을 판단한다. BM25는 단어 일치를, 벡터 검색은 의미 유사성을 반영한다. 이에 두 방식을 결합한 하이브리드 검색이 활용된다.

본 연구에서는 reciprocal rank fusion(RRF)을 적용하여 두 검색결과를 결합하였다. RRF는 각 방식의 순위를 역수 형태로 변환해 합산하여 최종 순위를 산출한다. 원 연구에서 제안된 상수($k=60$)을 그대로 사용하였으며 별도의 튜닝은 수행하지 않았다[4].

다만 하이브리드 검색이 항상 단일 검색보다 우수한 것은 아니다. 융합 방식과 모델 특성에 따라 성능 차이가 발생할 수 있으며[7], 비영어권 문서에서도 단일 검색이 더 나은 경우가 보고된다[8]. 이는 도메인별 검색 특성을 함께 고려해야 함을 시사한다.

2.3 기술교범 기반 정비정보 검색

기술교범 및 정비매뉴얼에는 점검, 시험, 고장탐구, 구성품 교환 등 절차적 정보가 포함된다. 사용자는 관련 절차가 포함된 문서 구간을 찾아야 하므로 검색 성능이 정비지원 효율에 직접적인 영향을 준다.

국내에서도 군용 정비매뉴얼에 소형언어모델과 RAG를 적용한 연구가 수행되어 제한된 환경에서의 정비지원 가능성이 제시되었다[6]. 기존 연구가 전체 시스템 적용 가능성에 초점을 둔 반면, 본 연구는 생성모델을 제외하고 검색 단계만을 대상으로 한다. 동일한 교범과 질의에 대해 BM25, 벡터, 하이브리드 검색을 적용하여 성능 차이를 정량적으로 비교한다.

3. 연구 방법

3.1 연구 절차

본 연구의 전체 절차는 Fig. 1과 같다. 먼저 기술교범 PDF로부터 텍스트를 추출하고, 텍스트 추출이 어려운 페이지는 선택적으로 OCR을 수행하였다. 이후 교범의 장·절 구조와 문장의 연속성을 고려하여 검색 대상 청크를 구성하였다.

검색성능 평가를 위해 실제 정비행위가 명확하게 포함된 정비절차를 선정하였다. 각 정비절차에는 교

범에 사용된 기술용어와 가까운 직접질의와 동일한 검색의도를 자연어 형태로 변경한 표현변형질의를 각각 구성하였다.

이후 동일한 문서집합에 대해 BM25, 벡터 및 하이브리드 검색을 수행하고 Hit@k와 MRR을 산출하였다. 마지막으로 검색방식 간 통계적 차이와 질의별 검색순위를 분석하였다.

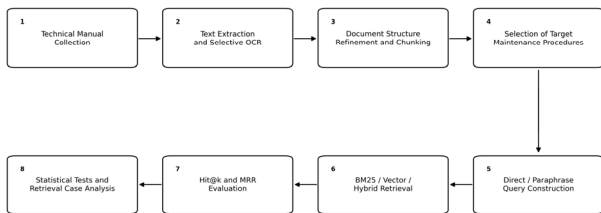


Fig. 1. Overall procedure of the proposed retrieval performance evaluation

3.2 실험 데이터 구성

실험에는 총 5종의 서로 다른 장비의 기술교범을 사용하였다. 교범 및 장비의 세부 식별정보는 연구결과에 영향을 미치는 변수가 아니므로 M1~M5로 명명화하였다.

전체 PDF 1,899페이지에 대해 텍스트 추출을 수행하였으며, 일반적인 텍스트 추출이 어려운 페이지에는 선택적으로 OCR을 적용하였다. 문서 구조와 내용의 연속성을 고려한 전처리 후 최종적으로 788개의 검색대상 청크를 구성하였다.

Table 1. Composition of technical manual corpus

Manual	No. of chunks	Ratio (%)
M1	82	10.4
M2	172	21.8
M3	95	12.1
M4	123	15.6
M5	316	40.1
Total	788	100

검색성능 평가를 위한 정답 데이터는 단순히 정비 관련 키워드가 포함된 청크를 사용하는 대신, 점검·시험·구성품 교환·고장배제 등 검색의도가 명확한 정비절차를 대상으로 선정하였다. 부록, 참고자료 및

일반적인 정비지침은 검색대상 문서에는 유지하되 평가대상 정답에서는 제외하였다.

교범별 정비절차의 내용과 수가 동일하지 않았기 때문에 교범별 개수를 인위적으로 동일하게 맞추지 않고 평가에 적합한 정비절차(gold procedure) 30건을 선정하였다. Direct query는 교범에 등장하는 주요 기술용어 및 정비작업명을 기반으로 작성하였으며, paraphrase query는 동일한 검색 의도를 유지하되 사용자가 실제로 입력할 수 있는 자연어 표현 형태로 변환하여 구성하였다. 일부 정비절차는 문서 구조상 인접한 복수의 청크에 걸쳐 존재하는 경우가 있어, 해당 경우에는 모든 관련 청크를 정답으로 인정하였다.

3.3 검색방식

3.3.1 BM25

BM25는 어휘적 일치에 기반한 검색방식으로 적용하였다. 한국어의 조사와 어미 등의 영향을 줄이기 위해 형태소 분석기를 이용하여 질의와 문서 청크를 토큰화하고 이를 이용하여 BM25 색인을 구성하였다. 동일한 788개의 문서 청크를 대상으로 각 평가질의 BM25 점수를 산출하고 점수가 높은 순서대로 검색결과를 정렬하였다.

3.3.2 벡터 검색

벡터 검색에는 Qwen3-Embedding-8B 모델을 사용하였다[5]. 각 문서 청크를 4,096차원의 임베딩 벡터로 변환하여 사전에 저장하였으며, 788개 문서의 임베딩과 각 평가질의 임베딩을 동일한 벡터 공간에서 비교하였다.

질의와 문서 사이의 유사도는 cosine similarity를 이용하여 산출하였다. 임베딩 벡터는 정규화한 후 유사도가 높은 문서부터 검색결과를 구성하였다.

3.3.3 하이브리드 검색

하이브리드 검색은 BM25의 검색결과와 벡터 검색결과를 RRF 방식으로 결합하였다. 문서 (d)의 RRF 점수는 식 (1)과 같다.

$$RRF(d) = \sum (1/(k + r_m(d))) \quad (1)$$

여기서 ($r_m(d)$)는 검색방식 (m)에서 문서 (d)가 나타난 순위이며, (k)는 상위순위 간 차이를 완화하기 위한 상수이다. 본 연구에서는 RRF의 연구에서 사용된 ($k=60$)을 적용하였으며 평가데이터를 이용한 별도의 파라미터 조정은 수행하지 않았다[4].

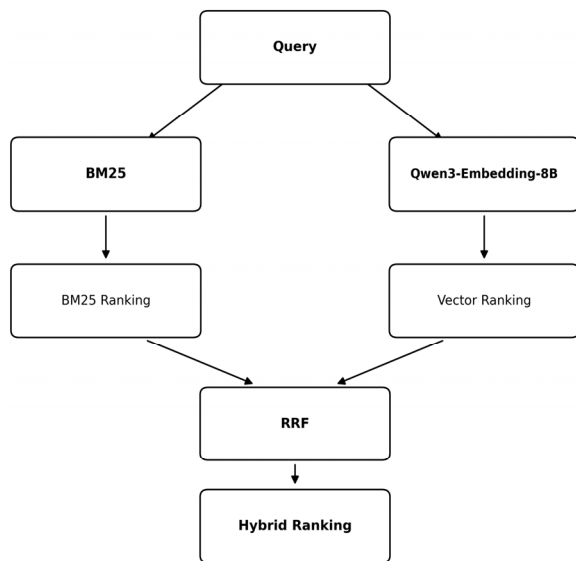


Fig. 2. Retrieval structure of BM25, vector and hybrid methods

3.4 평가지표 및 통계분석

검색성능은 Hit@1, Hit@3, Hit@5와 mean reciprocal rank(MRR)를 사용하여 평가하였다. Hit@k는 정답 청크가 상위 k개 결과 내에 포함된 비율이며, MRR은 각 질의에서 최초 정답 청크의 순위 역수(reciprocal rank)를 평균한 값으로 식 (2)와 같이 정의된다.

$$MRR = \frac{1}{N} \sum_{i=1}^N \frac{1}{rank_i} \quad (2)$$

여기서 N은 전체 질의 수, $rank_i$ 는 i번째 질의에서 최초 정답이 등장한 순위이다.

검색방식 간 차이는 각 질의의 RR을 기반으로 분석하였으며, 동일한 질의에 세 검색방식을 대응하여 평가한 점과 순위 기반 데이터 특성을 고려하여 비모수

적인 방법인 Friedman 검정을 적용하였다. 유의한 차이가 있는 경우 Wilcoxon 부호순위 검정을 적용하였으며, 다중 비교에 따른 오류를 제어하기 위해 Holm 보정을 사용하였다. 유의수준은 0.05로 설정하였다.

Table 2. Evaluation metrics

Metric	Description
Hit@1	Correct answer ranked 1st
Hit@3	Correct answer ranked within top 3
Hit@5	Correct answer ranked within top 5
MRR	Mean reciprocal rank of the first correct answer

4. 실험결과 및 분석

4.1 전체 검색성능

전체 60개 질의에 대한 검색성능은 Table 3와 같다.

Table 3. Overall retrieval performance

Method	Hit@1	Hit@3	Hit@5	MRR
BM25	0.1333	0.2667	0.4167	0.2535
Vector	0.5167	0.6667	0.8000	0.6355
Hybrid	0.5167	0.7000	0.8000	0.6415

BM25의 Hit@1은 0.1333으로 60개 질의 중 8개에서 첫 번째 검색 결과로 정답 청크가 반환되었다. 벡터 검색과 하이브리드 검색의 Hit@1은 각각 0.5167로 31개의 질의에서 첫 번째 검색결과로 정답 청크가 반환되었다.

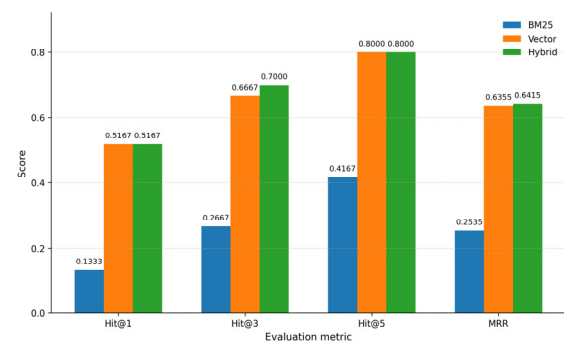


Fig. 3. Overall retrieval performance by retrieval method

Hit@5에서는 BM25가 0.4167인 반면, 벡터 검색과 하이브리드 검색은 모두 0.8000으로 나타났다. 이를 질의 수로 환산하면 BM25는 60개 질의 중 25개 질의에서 정답 청크를 상위 5개 검색 결과에 포함하였으며, 벡터 검색과 하이브리드 검색은 각각 48개 질의에서 정답 청크를 상위 5개 검색 결과에 포함하였다.

MRR은 BM25 0.2535, 벡터 0.6355, 하이브리드 0.6415로 나타났다. 하이브리드 검색의 값이 가장 높았으나 벡터 검색과의 차이는 0.0060으로 작았다.

4.2 질의 유형에 따른 성능

직접질의와 표현변형질의를 구분한 결과는 Table 4와 같다.

Table 4. Retrieval performance by query type

Query Type	Method	Hit@1	Hit@3	Hit@5	MRR
Direct	BM25	0.1333	0.3000	0.5333	0.2807
	Vector	0.5333	0.7000	0.7667	0.6546
	Hybrid	0.6333	0.7667	0.8667	0.7299
Para-phrase	BM25	0.1333	0.2333	0.3000	0.2363
	Vector	0.5000	0.6333	0.8333	0.6263
	Hybrid	0.4000	0.6333	0.7333	0.5532

직접질의에서는 하이브리드 검색이 네 개 지표 모두에서 가장 높은 값을 보였다. 특히 Hit@1은 0.6333으로 벡터 검색의 0.5333보다 0.1000 높았고, MRR 역시 하이브리드 0.7299, 벡터 0.6546, BM25 0.2807 순으로 나타났다.

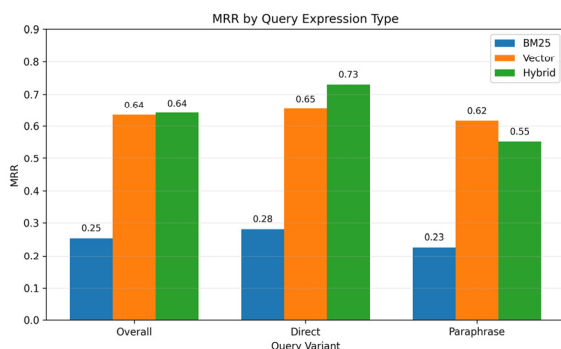


Fig. 4. Comparison of MRR according to query type

표현변형질의에서는 다른 경향이 나타났다. 벡터 검색의 Hit@1은 0.5000, Hit@5는 0.8333, MRR은 0.6263으로 가장 높았다. 하이브리드 검색의 MRR은 0.5532로 벡터 검색보다 낮았다.

이는 하이브리드 검색의 효과가 질의 표현에 따라 다르게 나타날 수 있음을 보여준다. 교범의 기술용어와 정비작업명이 직접 포함된 질의에서는 BM25의 어휘정보와 벡터 검색의 의미정보가 함께 반영되면서 정답의 순위가 상승할 수 있다. 반면 교범 원문과 표현 차이가 큰 질의에서는 BM25의 순위정보가 결합되면서 벡터 검색에서 높은 순위를 갖던 정답 청크가 일부 낮아질 가능성이 있다.

4.3 통계적 차이 분석

검색방식 간 성능 차이의 통계적 유의성을 확인하기 위해 각 질의에서 산출된 reciprocal rank(RR)를 기준으로 Friedman 검정을 실시하였다. 사후검정에는 Wilcoxon 부호순위 검정을 적용하였으며, 다중비교에 따른 오류를 통제하기 위해 p-value를 Holm 방법으로 보정하였다. Holm 보정 후 $p < 0.05$ 인 경우 통계적으로 유의한 차이가 있는 것으로 판단하였다.

전체 60개 질의에서는 BM25, 벡터 및 하이브리드 검색 간에 통계적으로 유의한 차이가 확인되었다($p < 0.001$). 사후검정 결과 BM25와 벡터 검색, BM25와 하이브리드 검색 간의 차이는 모두 유의하였다. 반면 벡터와 하이브리드 검색 간 차이는 유의하지 않았

Table 5. Statistical comparison of retrieval methods

Query Set	Comparison	Holm-adjusted p	Result
Overall	BM25-Vector	< 0.001	Significant
	BM25-Hybrid	< 0.001	Significant
	Vector-Hybrid	0.726	Not Significant
Direct	BM25-Vector	< 0.001	Significant
	BM25-Hybrid	< 0.001	Significant
	Vector-Hybrid	0.034	Significant
Para-phrase	BM25-Vector	< 0.001	Significant
	BM25-Hybrid	< 0.001	Significant
	Vector-Hybrid	0.295	Not Significant

Note: Pairwise p-values were adjusted using the Holm method. Statistical significance was determined at $\alpha=0.05$

다(Holm-adjusted $p=0.726$).

질의 유형별 분석에서도 직접질의와 표현변형질의 모두 세 검색방식 간 유의한 차이가 나타났다($p < 0.001$). 직접질의의 경우 벡터와 하이브리드 검색 간 차이도 유의하게 나타났으며(Holm-adjusted $p=0.034$), 하이브리드 검색의 MRR이 더 높았다. 반면 표현변형질의에서는 벡터 검색의 MRR이 하이브리드 검색보다 높았으나 두 방법 간 차이는 통계적으로 유의하지 않았다(Holm-adjusted $p=0.295$).

4.4 질의별 검색결과 분석

평균적인 검색성과 함께 개별 질의에서 나타나는 검색순위의 차이를 확인하기 위해 각 질의의 RR을 검색방식별로 비교하였다. 두 검색방식 가운데 표의 왼쪽에 제시된 방법을 기준으로 RR이 높은 경우, 동일한 경우, 낮은 경우를 각각 구분하였다.

전체 60개 질의에서 벡터 검색은 BM25보다 높은 RR을 나타낸 경우가 47건이었으며, 동일한 경우는 6건, 낮은 경우는 7건이었다. 하이브리드 검색은 BM25보다 높은 경우가 49건으로 나타났으며, 동일한 경우 8건, 낮은 경우 3건이었다. 이를 통해 대부분의 질의에서 벡터 또는 하이브리드 검색이 BM25보다 정답 체크를 높은 순위에 배치하는 경향을 확인하였다.

벡터와 하이브리드 검색의 비교에서는 질의 유형에 따라 서로 다른 양상이 나타났다. 직접질의에서 하이브리드 검색의 RR이 벡터 검색보다 높은 경우는 9건, 동일한 경우는 18건, 낮은 경우는 3건이었다. 반면 표현변형질의에서는 하이브리드 검색이 높은 경우 7건, 동일한 경우 12건, 낮은 경우 11건으로 나타났다.

Table 6. Per-query reciprocal-rank comparison

Comparison	Higher	Equal	Lower
Vector vs. BM25	47	6	7
Hybrid vs. BM25	49	8	3
Hybrid vs. Vector (Direct)	9	18	3
Hybrid vs. Vector (ParaPhrase)	7	12	11

개별 검색사례를 검토한 결과 BM25에서는 질의와 정답 체크 사이의 직접적인 어휘 일치 부족할 때 정

답 순위가 낮아지는 사례가 나타났다. 반면 벡터 검색은 질의의 표현이 교범 원문과 다르더라도 검색의도가 유사한 정비절차를 상대적으로 높은 순위에서 검색하는 경우가 확인되었다.

하이브리드 검색에서는 직접질의의 경우 BM25의 어휘적 검색결과와 벡터 검색의 의미적 검색결과가 함께 반영되면서 정답 순위가 상승하는 사례가 나타났다. 반면 일부 표현변형질의에서는 BM25에서 높은 순위를 가진 비정답 체크가 RRF 결과에 반영되면서 벡터 검색 단독 결과보다 정답 순위가 낮아지는 경우도 확인되었다.

5. 고찰

본 연구의 실험 결과, 기술교범 정비절차 검색에서는 BM25에 비해 벡터 검색과 하이브리드 검색이 전반적으로 높은 성능을 보였다. 전체 질의에서 벡터 검색과 하이브리드 검색의 MRR은 각각 0.6355와 0.6415로 BM25의 0.2535보다 높았으며, 이러한 차이는 통계적으로도 유의하였다. Hit@5 역시 벡터와 하이브리드 검색이 모두 0.8000으로 나타나, 본 연구에서 구성한 기술교범 검색환경에서는 의미적 유사성을 활용하는 검색방식이 정비절차 탐색에 효과적이었음을 확인하였다.

BM25는 직접질의에서도 벡터 및 하이브리드 검색보다 낮은 성능을 보였다. 기술교범에는 동일하거나 유사한 구성품명과 정비용어가 여러 장·절에 반복되어 나타날 수 있으므로, 질의와 문서 사이에 동일한 용어가 포함되어 있더라도 해당 문서가 반드시 사용자가 찾고자 하는 정비절차에 해당한다고 보기는 어렵다. 실제 질의별 검색결과에서도 벡터 검색의 RR이 BM25보다 높은 경우가 전체 60건 중 47건으로 확인되었다. 이러한 결과는 전문용어의 단순 일치뿐 아니라 질의와 정비절차 사이의 의미적 관련성을 함께 고려하는 것이 기술교범 검색에서 중요할 수 있음을 보여준다.

벡터와 하이브리드 검색의 관계에서는 질의 유형에 따른 차이가 확인되었다. 직접질의에서는 하이브리드 검색의 MRR이 0.7299로 벡터 검색의 0.6546보다 높았으며, 두 방법 간 차이는 Holm 보정 후에도 통계적으로 유의하였다. 또한 질의별 비교에서도 하이브리드 검색의 RR이 벡터 검색보다 높은 경우가 9

건인 반면 낮은 경우는 3건이었다. 이는 교범에 사용된 기술용어나 정비작업명이 질의에 직접 포함되는 경우 BM25의 어휘적 검색결과가 벡터 검색결과를 보완할 수 있음을 시사한다.

반면 표현변형질의에서는 벡터 검색의 MRR이 0.6263으로 하이브리드 검색의 0.5532보다 높게 나타났다. 질의별 비교에서도 벡터 검색의 RR이 높은 경우가 상대적으로 많았다. 다만 두 검색방식 간 차이는 통계적으로 유의하지 않았으므로, 본 연구 결과만으로 표현변형질의에서 벡터 검색이 하이브리드 검색보다 우수하다고 일반화하기는 어렵다. 다만 교범 원문과 다른 표현으로 작성된 질의에서는 어휘적 일치정보를 결합하는 것이 항상 검색성능 향상으로 이어지지는 않을 가능성을 확인하였다. 하이브리드 검색의 효과가 데이터와 검색모델, 융합방법에 따라 달라질 수 있다는 기존 연구[7,8]와도 유사한 맥락으로 볼 수 있다.

이와 같은 결과는 기술교범 기반 RAG를 구성할 때 검색방식을 단순히 하나로 고정하기보다 질의 특성을 고려할 필요가 있음을 보여준다. 본 실험에서는 직접질의에서 하이브리드 검색의 장점이 뚜렷하게 나타난 반면, 표현변형질의에서는 벡터 검색만으로도 높은 검색성능이 확인되었다. 따라서 향후에는 질의에 포함된 전문용어의 비율이나 원문과의 어휘적 유사성 등을 활용하여 검색방식을 선택하거나 결합비율을 조정하는 방법을 검토할 수 있다.

6. 결론

본 연구에서는 기술교범 기반 RAG 적용을 위한 검색단계에 초점을 두고 BM25, 벡터 및 하이브리드 검색방식의 정비절차 검색성능을 비교하였다. 이를 위해 5종의 기술교범을 788개의 문서 청크로 구성하고, 평가에 적합한 30개의 정비절차를 선정하여 직접질의와 표현변형질의 총 60건을 대상으로 검색성능을 평가하였다.

실험 결과 전체 질의에서 벡터 및 하이브리드 검색은 BM25보다 높은 Hit@k와 MRR을 나타냈으며, 통계분석에서도 BM25와 두 검색방식 간의 유의한 차이가 확인되었다. 반면 전체 질의에서 벡터 검색과 하이브리드 검색 간에는 통계적으로 유의한 차이가 나타나지 않았다. 따라서 본 연구의 실험범위에서는 의

미적 유사성을 활용하는 벡터 기반 검색이 기술교범 정비절차 검색에 효과적으로 적용될 수 있음을 확인하였다.

질의 유형을 구분한 결과에서는 서로 다른 특성이 나타났다. 직접질의에서는 하이브리드 검색이 벡터 검색보다 높은 성능을 보였으며 그 차이도 통계적으로 유의하였다. 반면 표현변형질의에서는 벡터 검색의 성능지표가 하이브리드 검색보다 높았으나 두 방법 간 차이는 통계적으로 유의하지 않았다. 이는 하이브리드 검색이 모든 질의에서 일관되게 벡터 검색을 개선하는 것은 아니며, 질의 표현의 특성에 따라 검색방식의 효과가 달라질 수 있음을 보여준다.

따라서 기술교범 기반 RAG를 구성할 때에는 단순히 BM25와 벡터 검색을 결합하는 것보다 검색대상 문서와 사용자의 질의 특성을 고려하여 적절한 검색방식을 적용하는 것이 필요하다. 특히 정비작업명이나 구성품명과 같은 기술용어가 명확한 질의에서는 하이브리드 검색의 활용을, 원문과 다른 자연어 표현이 사용되는 질의에서는 의미기반 벡터 검색의 활용을 우선적으로 검토할 수 있다.

향후 연구에서는 실제 정비자가 사용하는 질의를 포함하여 평가데이터의 범위를 확대하고, 다양한 기술교범과 임베딩 모델을 대상으로 검색성능을 추가 검증할 필요가 있다. 또한 검색된 정비정보를 생성모델에 제공하여 검색 정확도가 최종 답변의 정확성 및 근거성에 미치는 영향을 평가함으로써 기술교범 기반 RAG의 전체 질의응답 과정에 대한 성능을 검증할 필요가 있다.

참고문헌

- [1] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel et al., 'Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,' Advances in Neural Information Processing Systems, VOL. 33, 2020.
- [2] Stephen Robertson & Hugo Zaragoza, 'The Probabilistic Relevance Framework: BM25 and Beyond,' Foundations and Trends in Information Retrieval, VOL. 4, NO. 1-2, 2009.
- [3] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, Wen-tau Yih, 'Dense Passage Retrieval for Open-Domain Question Answering,' in Proceedings of EMNLP, November 2020, pp.

6769–6781.

[4] Gordon V. Cormack, Charles L. A. Clarke, & Stefan Büttcher, ‘Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods,’ in Proceedings of ACM SIGIR, 2009, pp. 758–759.

[5] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin et al., ‘Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models,’ arXiv preprint arXiv:2506.05176, 2025.

[6] Ye-rin Nam, Han-seon Choi, Jong-Geun Choi, Hyuk-Jin

Kwon, & Young-hoon Lee, ‘Small Language Model and RAG-Based Military Maintenance Manual Curation System,’ Journal of Internet Computing and Services, VOL. 26, NO. 1, 2025, pp. 181–192.

[7] Sebastian Bruch, Siyu Gai, & Amir Ingber, ‘An Analysis of Fusion Functions for Hybrid Retrieval,’ ACM Transactions on Information Systems, VOL. 42, NO. 1, 2023, Article 20.

[8] Antoine Louis, Gijs van Dijck, & Gerasimos Spanakis, ‘Know When to Fuse: Investigating Non-English Hybrid Retrieval in the Legal Domain,’ in Proceedings of COLING, Abu Dhabi, UAE, January 2025, pp. 4293–4312.