테스트 계획 및 결과보고서

1. 테스트 계획

- **목적**: 리트리버 성능을 비교하기 위해 다양한 임베딩 모델을 사용하여 검색 정확도를 테스트

1.1 테스트 대상 임베딩 모델 및 선정 기준

1. 선정 기준

- 벤치마크 성능이 우수한 모델을 우선적으로 선정함. 특히 NDCG@k, MRR@k 등 주요 평가 지표에서 상위권 성능을 기록한 모델들을 중심으로 테스트를 진행함.(출처:Kor-IR: 한국어 검색을 위한 임베딩 벤치마크)
- 접근성과 사용성을 고려하여 Hugging Face에서 제공하는 모델을 주로 선택했으며 OpenAI 모델도 함께 포함하여 비교.

2. 테스트 대상

- Hugging Face 모델:
  - `intfloat/multilingual-e5-large-instruct`
  - `intfloat/multilingual-e5-large`
  - `BAAI/bge-m3`
  - `intfloat/multilingual-e5-base`
  - `intfloat/multilingual-e5-small`
- OpenAI 모델:
  - `text-embedding-3-small`

3. 평가 지표:

- **NDCG@k** (Normalized Discounted Cumulative Gain): 검색 결과의 적합성 평가
- **MRR@k** (Mean Reciprocal Rank): 검색 결과의 순위 정확도
- **MAP@k** (Mean Average Precision): 상위 검색 문서들의 평균 정밀도
- **Recall@k**: 실제 답변 문서가 검색된 비율

4. 테스트 방법

- 동일한 법령 문서 데이터셋을 사용.
- 각 모델별로 벡터화를 수행하여 ChromaDB에 저장.
- 테스트 질문으로는 `test_data`의 **'질의요지'** 컬럼에서 발췌한 데이터를 사용.
- 각 리트리버가 반환한 Top-10 검색 결과를 분석하여 성능 평가.

test_data.csv 예시


```
{
        "년도":2021,
        "분류":"해양수산",
        "제목":"(「선박직원법 시행령」 제16조제3항제3호 등 관련)",
        "질의요지":"「선박직원법 시행령」 제16조제3항 및 같은 항 제3호에 따르면, 1년 이상의 통신에 의한 교육과정을 이수한 자
        "회답":"이 사안에서 「선박직원법 시행령」 제16조제3항제3호에 따른 “통신에 의한 교육과정”은 “통신매체 등을 활용한 교육
        "이유":"일반적으로 “의한, 의하여, 의하다”라는 표현은 “무엇에 의거하거나 기초하다”를 의미1)1) 국립국어원 표준국어대사
        "관계법령":"선박직원법 시행령제16조(지정교육기관의 교육과정 이수자 등에 대한 특례) ① · ② (생  략)③지정교육기관중 다
        "법령":"선박직원법 시행령",
        "법":"제1조(목적)  이 영은 「선박직원법」에서 위임된 사항과 그 시행에 관하여 필요한 사항을 정함을 목적으로 한다. <
    },
    {
        "년도":2020,
        "분류":"국토개발",
        "제목":"(「도시 및 주거환경정비법」 제22조제1항 본문 등 관련)",
        "질의요지":"「도시 및 주거환경정비법」(이하 “도시정비법”이라 함)에 따른 주거환경개선구역1)1) 도시정비법 제2조제2호가
        "회답":"이 사안의 경우 도시정비법 제20조제1항제3호에 따라 재개발구역이 해제되더라도 같은 법 제22조제1항 본문을 적용하
        "이유":"도시정비법 제22조제1항 본문에서는 같은 법 제20조 및 제21조에 따라 정비구역등3)3) 정비예정구역 또는 정비구역
        "관계법령":"도시 및 주거환경정비법제20조(정비구역등의 해제) ① 정비구역의 지정권자는 다음 각 호의 어느 하나에 해당하
        "법령":"도시 및 주거환경정비법",
        "법":"제1조(목적)  이 법은 도시기능의 회복이 필요하거나 주거환경이 불량한 지역을 계획적으로 정비하고 노후 · 불량건축
    },
    {
        "년도":2021,
        "분류":"주택관리부동산",
        "제목":"(「공동주택관리법」 제52조제3항 등 관련)",
        "질의요지":"「공동주택관리법」 제52조제1항에서는 주택관리업을 하려는 자는 시장 · 군수 · 구청장1)1) 특별자치시장 · 특별자
        "회답":"이 사안에서 「공동주택관리법」 제52조제3항 각 호 외의 부분 전단은 주택관리업의 등록 요건을 정한 규정입니다."
        "이유":"일반적으로 “신청”이란 행정청의 처분을 구하기 위한 의사 표시를 의미하고, 그러한 의사 표시를 할 수 있는 자는 법
        "관계법령":"공동주택관리법제52조(주택관리업의 등록) ① 주택관리업을 하려는 자는 대통령령으로 정하는 바에 따라 시장 · 군
        "법령":"공동주택관리법",
        "법":"제1조(목적)  이 법은 공동주택의 관리에 관한 사항을 정함으로써 공동주택을 투명하고 안전하며 효율적으로 관리할
    },

```


5. 목표 기준

- **최적 모델 선정 기준:**
  - OpenAI의 text-embedding-3-small 모델을 중심으로 평가하며, 다른 모델들과 성능을 비교하여 결과를 분석.
  - OpenAI 모델이 NDCG@10, MRR@10, MAP@10에서 경쟁 모델보다 높은 성능을 보일 경우 최종 선정.
  - Recall@10 점수가 일정 수준 이상 유지되는지 확인하며, 부족할 경우 추가 데이터 확보 및 모델 개선 방안을 함께 검토.

---

2. 테스트 진행

2.1 데이터 준비

- **법령 문서**:
  - 법제처에서 제공하는 법령해석 문서를 변환 (각 조항의 제목과 본문을 포함)
  - 예시:

테스트 데이터

```
[Document(metadata={'source': '선박직원법 시행령 제1조(목적)'}, page_content='선박직원법 시행령 제1조(목적)'),
 Document(metadata={'source': '도시 및 주거환경정비법 제1조(목적)'}, page_content='도시 및 주거환경정비법 제1조(목적)'
 Document(metadata={'source': '공동주택관리법 제1조(목적)'}, page_content='공동주택관리법 제1조(목적)'),
 Document(metadata={'source': '환경영향평가법 시행령 제1조(목적)'}, page_content='환경영향평가법 시행령 제1조(목적)'),
 Document(metadata={'source': '국가유공자 등 예우 및 지원에 관한 법률 제1조(목적)'}, page_content='국가유공자 등 예우
 Document(metadata={'source': '공인중개사법 제1조(목적)'}, page_content='공인중개사법 제1조(목적)'),
 Document(metadata={'source': '공유수면 관리 및 매립에 관한 법률 제1조(목적)'}, page_content='공유수면 관리 및 매립에
 Document(metadata={'source': '건축법 시행령 제1조(목적)'}, page_content='건축법 시행령 제1조(목적)'),
 Document(metadata={'source': '대기관리권역의 대기환경개선에 관한 특별법 제1조(목적)'}, page_content='대기관리권역의 다
 Document(metadata={'source': '도시 및 주거환경정비법 제1조(목적)'}, page_content='도시 및 주거환경정비법 제1조(목적)'
 Document(metadata={'source': '환경영향평가법 제1조(목적)'}, page_content='환경영향평가법 제1조(목적)'),
 Document(metadata={'source': '동물 약국 및 동물용 의약품등의 제조업ㆍ수입자와 판매업의 시설 기준령 제1조(목적)'}, page_
 Document(metadata={'source': '재건축초과이익 환수에 관한 법률 제1조(목적)'}, page_content='재건축초과이익 환수에 관한
```

2.2 테스트 프로세스

1. 각 모델로 문서를 벡터화하여 ChromaDB에 저장.
2. 각 질문을 입력하여 리트리버가 반환한 **Top-10 검색 결과**를 가져옴.
3. 검색된 문서와 실제 답변 문서(ground truth)를 비교하여 **평가 지표**를 계산.

---

3. 테스트 결과

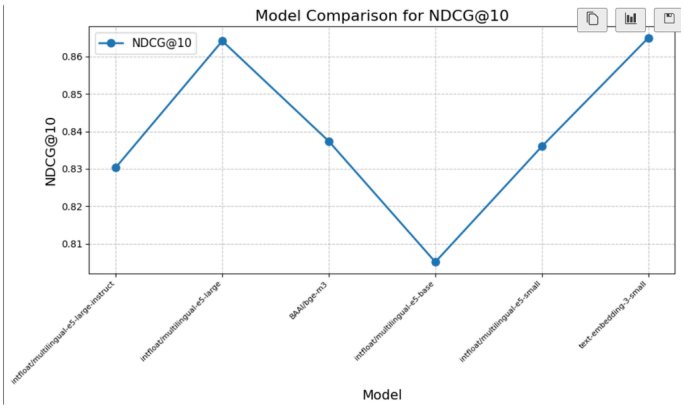
3.1 평가 결과 요약

Model	NDCG@10	MRR@10	MAP@10	Recall@10	Average
intfloat/multilingual-e5-large	0.8641	0.8002	0.8249	0.1659	0.6638
text-embedding-3-small	0.8650	0.8074	0.8245	0.0845	0.6454
intfloat/multilingual-e5-small	0.8360	0.7647	0.8049	0.1448	0.6376
BAAI/bge-m3	0.8374	0.7557	0.7895	0.1487	0.6328
intfloat/multilingual-e5-	0.8304	0.7413	0.7926	0.1443	0.6272

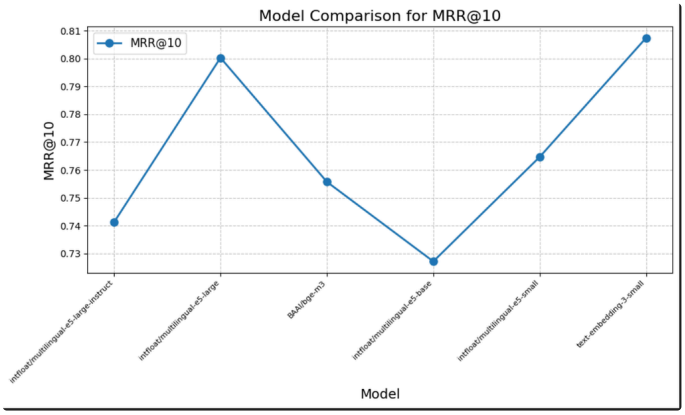
large-instruct					
intfloat/multilingual-e5-base	0.8051	0.7272	0.7685	0.1394	0.6101

3.2 각 평가지표별 그래프

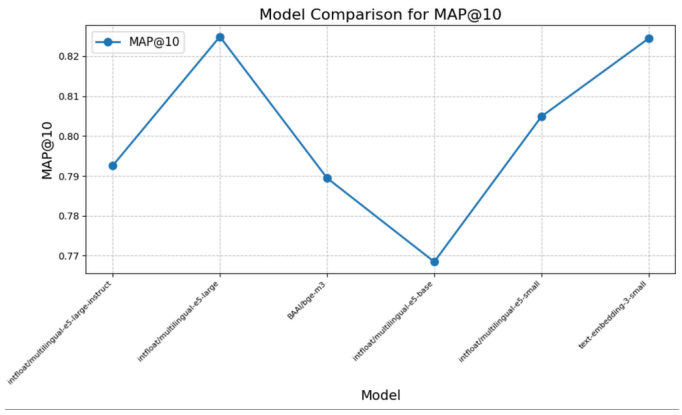
NDCG@k



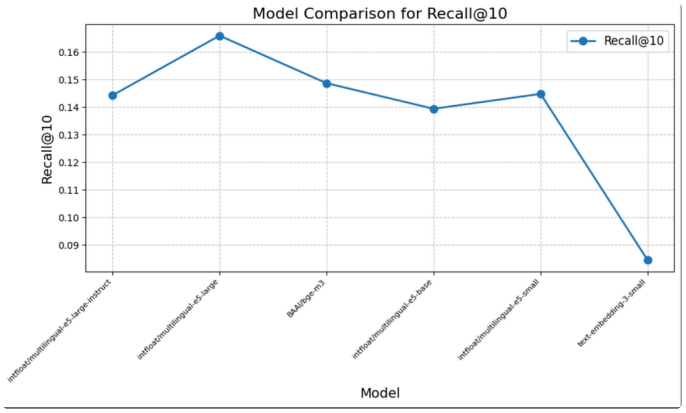
MRR@k



MAP@k



Recall@k



3.3 결과 분석

1. **OpenAI 모델**:
   - **text-embedding-3-small**는 NDCG@10, MRR@10, MAP@10에서 높은 성능을 보였으나 Recall@10 점수가 낮음.
2. **Hugging Face 모델**:
   - **intfloat/multilingual-e5-large**는 Average 점수가 가장 높아 전체적으로 안정적인 성능을 보임.
3. **Recall 성능**:
   - 전반적으로 모든 모델에서 Recall 성능이 낮아 추가 개선이 필요.

---

4. 결론 및 계획

4.1 결론

- **OpenAI의** `text-embedding-3-small` 모델은 검색 적합성과 순위 정확도(**NDCG, MRR**)에서 강력한 성능을 보였으며, API 기반 접근성 덕분에 실제 시스템에 쉽게 통합 가능할것으로 보임.
- **intfloat/multilingual-e5-large** 모델은 높은 Recall 점수와 전체적인 균형 성능을 제공.

4.2 계획

- **최적 모델**:
   - OpenAI의 `text-embedding-3-small` 을 최종 선정하여 적용예정.
- **성능 개선**:
   - Recall 성능 향상을 위해 추가 데이터셋 확보 및 fine-tuning 수행예정.
   - 키워드 기반 검색인 BM25 리트리버와 앙상블하여 Recall 향상을 기대.
- **추가 비교 테스트**:
   - OpenAI 모델의 최신 버전과 Hugging Face의 다른 경쟁 모델을 비교하여 성능 검증예정.