HTP 검사 해석 RAG pipeline 구축

박진호

목차

1. 요구사항 / 목표 설정

(1) HTP 그림 심리 검사 이해

HTP 심리 검사는 House-Tree-Person 그림을 통해 내적 심리를 분석하는 심리 투사 검사이다. 심리검사 실시로 시작되는 심리치료에서 내담자는 성인이나 아동 모두 새로운 환경에 대한 불안으로 인하여 위축되거나 회피하는 모습을 보이며, 자신의 정서나 감정을 원활하게 표현하기에 어려움을 호소하기도 한다[백원대(2019)]. 투사적 검사는 검사 시 비교적 모호한 자극에 대해 자신의 반응이 갖는 의미를 인지하지 못해 의식적인 통제를 덜 하게 되고 자유롭게 표현할 수 있다(박현주, 서명옥, 2014). 그래서 그림을 통한 검사로 인해 좀더 솔직하고 의식하지 않은 아픔과 어려움을 표현할 수 있다.

검사방법 설명

종이, 연필, 지우개를 사용해 집, 나무, 사람(2 장) 순서로 그리게 한다. 그림 솜씨 상관없이 수행할 수 있으며 이를 알린다. 그림을 통해 수행하기 때문에 문화, 언어능력, 나이 등을 초월해 수행할 수 있다. [Buck. J.N(1948)]

해석 방법

해석 방법에는 형식적 분석, 내용적 분석, 검사 시간 및 태도 등 크게 세 가지 관점에서 이루어진다. 그림에 대한 형식, 내용적 분석과 검사 대상자의 태도 등도 크게 작용하는 검사이기에 관찰과 상담사의 경험이 중요한 지점으로 작용할 수 있다.

(2) 요구사항

선행 연구 분석

제공받은 선행 연구

1. 아동 미술 심리검사를 위한 AI 기반 그림 데이터 분석 모델 연구[박성진. 2024]

    - **데이터**: AI HUB 에 구축된 집 그림 14,000 건(JPG & JSON) 활용.
    - **분석 모델**: 4 가지 딥러닝 객체 인식 모델 후보(Faster R-CNN, SSD, YOLOv5, EfficientDet) 중 YOLOv5(mAP@0.5 기준 우수) 최종 선정 및 적용.
    - **LLM(대형 언어모델) 비교**: Bing, Bard, ChatGPT4.0 의 이미지 분석 및 심리 해석 능력 실험.
    - **상담사 설문**: 실제 현업 상담사(16 명)를 대상으로 AI 분석 및 상담 활용 가능성 평가.

한계점

LLM 활용에 있어서 파운데이션 모델을 활용하였는데 단순히 상용 각 모델을 비교하여 HTP 검사가 수행 가능한지 고려하고 생성된 응답이 전문 상담가의 설문에 의존하여 제대로 응답을 생성할 수 있는지 비교하는 결과에 도달함.

2. 그림기반 아동심리검사에 있어 H(house)그림에 대한 신경망 기술을 적용한 심리검사 자동화 방법[이은정. 황세진. 2023]

- 합성곱 신경망(CNN)을 활용하여 아동이 그린 집, 나무, 사람 등 주요 객체의 위치, 크기, 개수 등 특성을 자동으로 탐지·분석한다.
- 자체적으로 수집한 손그림 데이터셋을 통해 정밀하게 튜닝된 신경망 모델을 사용한다.
- 전문가가 진단할 때 사용하는 기준을 퍼지 규칙으로 설계.
- 이미지 분석 결과를 퍼지 추론 엔진에 입력, 해당 규칙에 따라 각 그림(집·나무·사람)의 심리를 분류하고 유형(type)으로 최종 진단 결과를 도출한다.

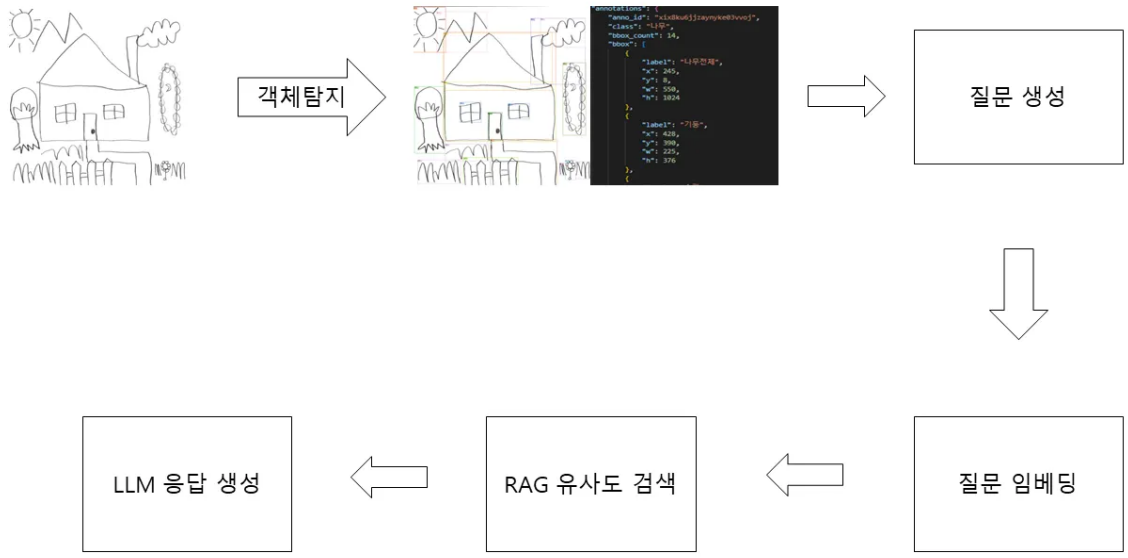
그림을 해석하기 위한 지식 DB 를 구축하는 데에 있어서 5 개의 논문 [25]~[29] 을 참고하여 공통된 해석 기준을 수립하였음. 저자가 구축한 기준에서 별도로 계량화하여 점수를 부여하였는데, 해당 점수 체계에 대한 기준이 모호하고 점수 체계를 세웠던 근거가 설명되어 있지 않음

HTP 검사 자동화

HTP 그림 심리 검사를 객체 탐지와 LLM 을 활용하여 자동화하여 접근성을 늘리고 RAG 구축을 통해 신뢰성을 확보하려고 시도하였다.

첫 번째 요구사항

1. 객체 탐지를 통해 그림의 위치 정보, 존재 여부를 파악한다.
2. 파악된 특징을, RAG 를 통해 의미 해석을 검색한다.
3. 그림 해석을 요청한 사용자는 LLM 을 통해 결과를 받는다.

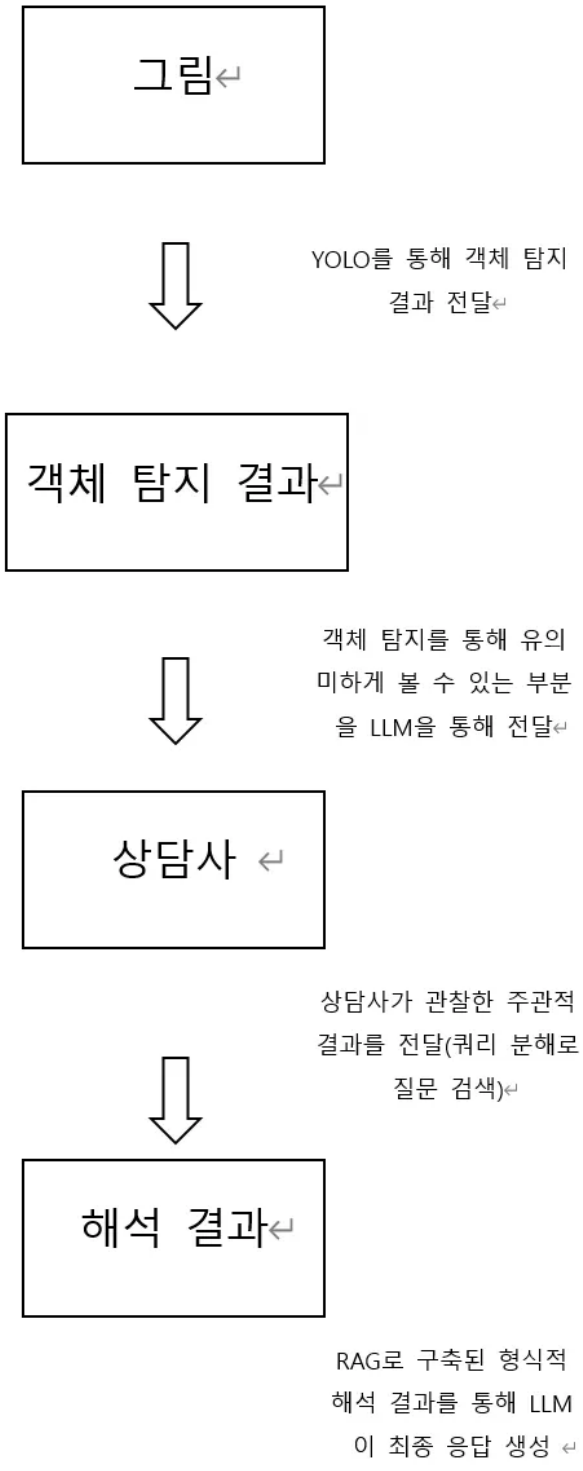


한계점

HTP 검사에서는 내용과 형식에만 의미가 있는 게 아니라 검사 대상자의 배경, 태도 등도 매우 중요한 요소로 작용한다. 그리고 내용적 해석에서도 어느 정도로 큰지 판단하고, 선의 굵기나 날림 정도, 어느 지점이 강조되어 있는지 파악하기 어려웠다. 마지막으로 HTP 검사는 결국 **임상 실험 기반의 해석 체계**이기 때문에 **상담사의 경험과 주관 등이 개입될 수밖에 없다는 점**이 있다. 그렇기 때문에 첫 번째 요구사항을 달성하더라도 **의미 있는 해석이 되기 어렵고**, 실제로 도움이 되기 어렵다는 판단에 **요구사항을 수정**하기로 하였다.

두 번째 요구사항

1. 객체 탐지를 통해 그림의 위치 정보, 존재 여부를 파악한다.
2. 파악된 정보를 상담사가 확인 후 추가 정보를 작성한다.
3. 상담사의 관찰 결과와 판단을 LLM 이 분석하여 RAG 에서 특징에 대한 해석을 검색한다.
4. 상담사가 LLM 을 통해 해석에 도움을 받고 시간을 단축, 상대적으로 주관을 배제한 결과를 받을 수 있다.



상담사의 관찰을 통해 해석을 돕고 해석에 있어 논문과 임상 자료를 통한 객관성을 확보하는 파이프라인을 구축하여 더 실용성 있는 서비스가 될 수 있도록 하였다.

(3) 목표 설정

목표 : 상담사의 관찰 결과 쿼리를 분해하여 각 쿼리에 대한 해석을 제공

검색된 정보를 종합하여 LLM 이 해석 결과를 제공하도록 한다. 여기서 가장 중요한 점이 정확하고 신뢰성 있는 정보 제공이다. RA 를 구축하는 문서에 노이즈가 최대한 없고 단순한 청킹이 아니라 의미론 적으로 분해가 가능하도록 하였다. 그리고 검색기(retriever)는 시간보다 정확성이 가장 중요하기 때문에 재현율(recall)점수가 가장 높게 나올 수 있도록 구성하는 게 목표이다.

- 의미 기반 청크 구성
- 정확도 우선 recall 0.9 이상

2. 데이터 구성 및 활용

(1) 데이터 수집

데이터는 그림의 특징에 대한 해석 정보, 임상 정보가 있는 논문과 서적을 중심으로 수집하였다. HTP 검사는 1948 년 J. N. Buck 에 의해서 처음으로 제창되었으며 Hammer 에 의해서 크게 발전되었다. J. N. Buck 과 Hammer 는 임상실험과 자료를 토대로 그림 해석 테크닉을 종합한 서적(1968)을 발표하였고, 이는 아직까지 HTP 검사의 토대가 되고 있다. 이들이 작성한 논문과 서적, 그리고 이를 활용한 기타 논문, 서적을 주요 데이터로 지정하였다. 가장 중요한 것은 그림의 특징과 해석이 연결되어 있는 자료여야 한다는 점이다.

참고 논문

[정현희 and 이화영. (2012). HTP 평가항목 간편화. 미술치료연구, 한국미술치료학회 19(2), 231-244.]

[이은정 and 황세진. (2023). 인공지능 객체검출모델 기반 집-나무-사람(HTP) 그림검사의 형식적 해석 연구. 미술치료연구, 30(5), 1241-1257.]

[Buck, J.N. (1948), The H-T-P test. J. Clin. Psychol., 4: 151-159.]

[Buck, J.N. (1948), The H-T-P technique. A qualitative and quantitative scoring manual. J. Clin. Psychol., 4: 317-396.]

참고 서적

Hammer, E. F. (1968). ***The House-Tree-Person (H-T-P) clinical research manual***. Western Psychological Services.

신민섭.(2003). 그림을 통한 아동의 진단과 이해. 서울: 학지사. 김동연, 공마리아(2002). HTP 와 KHTP 심리진단법. 서울: 동아문화사.

(2) 데이터 가공

1. 논문, 서적에서 HTP 임상 실험, 해석 정보에 대한 내용을 발췌해서 하나의 문서로 정리

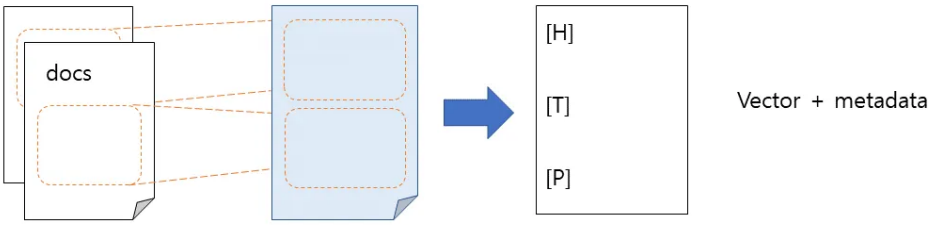
Why?

각 논문, 서적을 그대로 청킹, 임베딩을 진행하면 노이즈와 문서의 양이 너무 많아진다는 점이 존재, 그래서 각 자료에서 임상 실험, 그림에 대한 해석 정보를 하나의 문서로 만들어 필요한 정보로 압축하였다.

2. 압축한 자료를 의미론적으로 분해하여 구조화함

Why?

이미 수작업으로 자료를 정리하는 과정에서 어느 정도 구분됨 → 정확도와 metadata 정보를 정확하게 입력하기 위해 수작업으로 구분하는 과정을 통해 자료를 정리하였다. 또 해당 HTP 심리 검사에서 가짜 정보나 노이즈가 포함된 정보가 제공된다는 것은 굉장히 critical 하다고 생각하여 정확한 정보제공 및 검색을 위해 작업하였다.



3. RAG 파이프라인 설계

(1) 청크 구성

파이프라인에서 RAG 구축은 필수적, 따라서 그림에 대한 해석 자료를 임베딩화 하여 vectorDB 를 구축할 것이다. 그리고 일반적으로 문서를 사이즈로 청킹했을 때 문제가 발생할 수 있다고 생각하였다. 예상 가능한 문제점은 질문 쿼리에 적절한 응답임에도 불구하고 본문에 질문이 드러나지 않아 검색 상위에 노출되지 않는 현상이 발생할 수 있다고 생각하였다.

이러한 문제가 발생하는 이유는 다음과 같다.

- **임베딩의 특성**

  벡터 기반 검색은 질문의 의미적 벡터와 문서 또는 청크의 벡터 간 *유사도*를 계산하며, 키워드 매칭이 아닌 *의미적 근접성*에 의존

- **응답이 암묵적으로 포함된 경우**

  문서에 답변이 존재해도, 질문과의 연관성이 임베딩 모델에서 제대로 반영되지 않으면 유사도가 낮게 나오거나 컨텍스트 상위에 노출되지 않음

- **질문-답변 매핑 부족**

  답변 자체는 적합하지만 "질문" 단어나 그와 동일한 의미 구조가 본문에 명시적으로 나오지 않는다면 임베딩 유사도가 충분히 높게 평가되지 않을 수 있음

➔ 청크 단위를 의미론적으로 분해하고 메타데이터와 함께 저장하자

RAG 를 구축하기 위한 문서는 이미 의미상으로 어느 정도 분해되어 있다. 이를 적극 활용해 사이즈 별 청크 구성이 아니라 직접 문서를 확인해 의미가 변화하는 부분에 체킹하여 단위를 나누었다. 예를 들면 집에 대한 해석이 있는 부분만 묶었고, 안에서도 각 객체 문, 창문 등과 같은 요소로 해석이 바뀌는 지점을 묶어 청크를 분해할 수 있도록 미리 작업을 진행하였다. 이렇게 함으로써 메타데이터로 분류를 할 수 있게 되었고, 나중에 실 사용할 때 메타데이터 분류를 통해 검색 성능을 올릴 수 있다고 판단하였다.

검색에서 메타데이터의 필요성

- HTP 검사는 민감한 정보를 다루기 때문에 정확한 정보를 전달해야 함
- 정확한 정보를 전달하기 위해 단순히 정보 청킹이 아닌 정확한 검색이 이루어져야 함
- 생성된 질문 쿼리와 정확한 검색이 매칭될 수 있도록 메타데이터 구성

구성 청크 예시

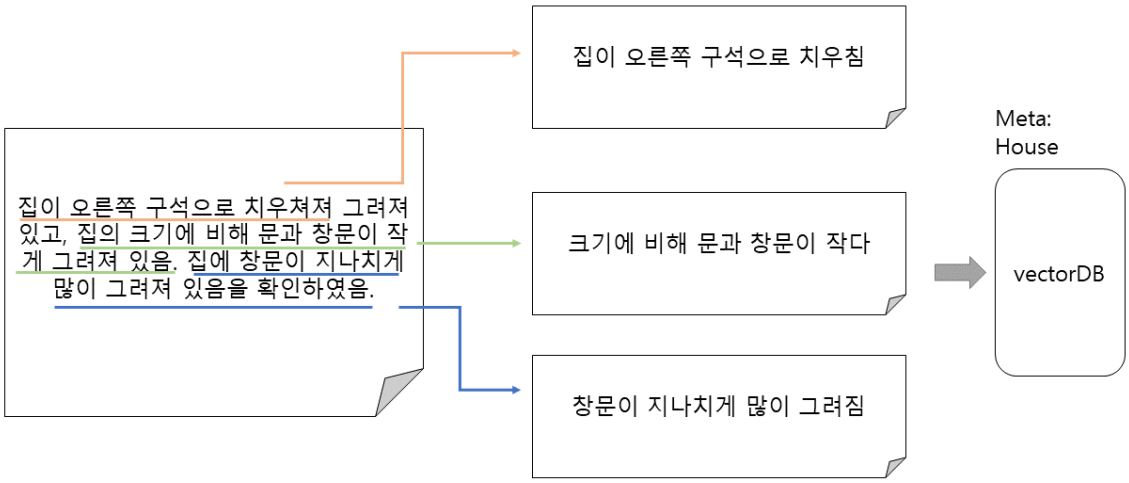
{ "content": "문이 너무 작다면 사회적 관계에 대해 접근과 회피 사이에서 갈등하거나, 대인관계 기술 자체가 부족함을 시사합니다.",

"metadata": { "main_component": "집", "sub_component": "문", } }

(2) 검색 단계

검색을 위해 상담사의 쿼리를 분해할 필요가 있다. 예를 들면 쿼리는 그림에 대한 다양한 관찰 사실을 연속적으로 입력할 것이다. 이를 의미 단위로 분해하여

각 분해된 쿼리별로 정확한 해석 정보를 검색해 오도록 구성한다.



4. RAG 검색 평가 및 결과

평가 방법 : RAG 평가 프레임워크 RAGAs

RAGAs? LLM 을 활용하여 검색과 생성 두 가지 부분을 평가해 주는 프레임워크이다.

평가 방법은 RAGAs 에서 제공해 주는 검색 평가 진행

Context Precision(정밀도) : 검색된 문맥에서 실제로 질문과 관련된 유용한 정보의 비율이 얼마나 되는지 평가. 점수가 낮으면 관련 없는 내용이 많이 검색된 것

Context Recall(재현율) : 정답을 생성하는 데 필요한 모든 관련 정보가 검색된 문맥에 포함되어 있는지 평가. 검색 시스템이 답변에 필요한 정보를 얼마나 빠짐없이 찾아냈는지 측정

임베딩 모델 순위 :

https://huggingface.co/spaces/mteb/leaderboard

vectorDB / Retriever 비교

FAISS (Facebook AI Similarity Search) 고속 벡터 검색 라이브러리

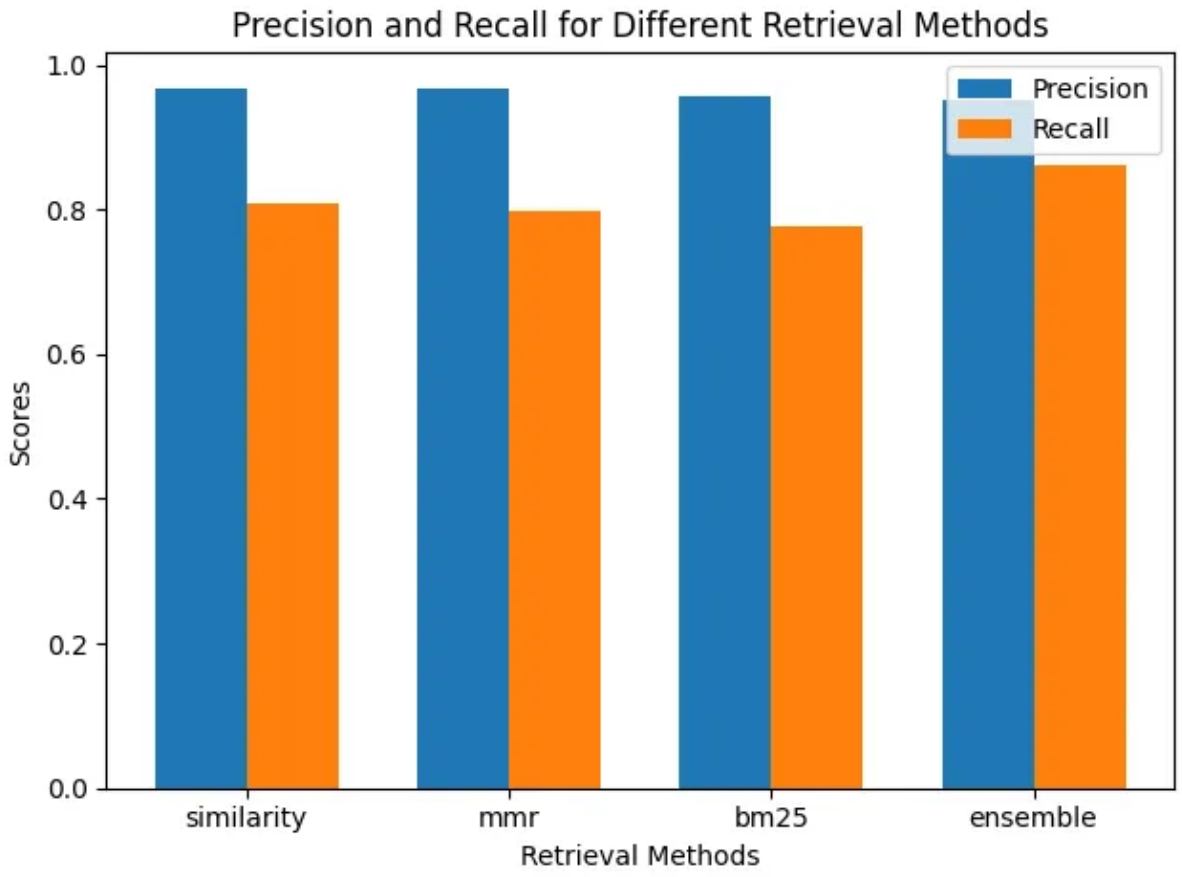
ChromaDB 사용이 용이하고 간단한 VectorDB

일반적으로 FAISS 가 빠르나 문서 특성상 정확도가 우선시되어야 하기에 비교를 진행 필요성이 있다.

검색기(Retriever)는 의미 검색(dense)에서 similarity, mmr 검색 알고리즘을 적용해 보고 키워드 검색(sparse)에서 Bm25 알고리즘을 활용하여 검색 성능을 진행해 보았다. 그리고 두 개의 검색기를 ensemble 한 성능을 최종적으로 적용해서 Hybrid 검색 성능도 비교해 보았다.

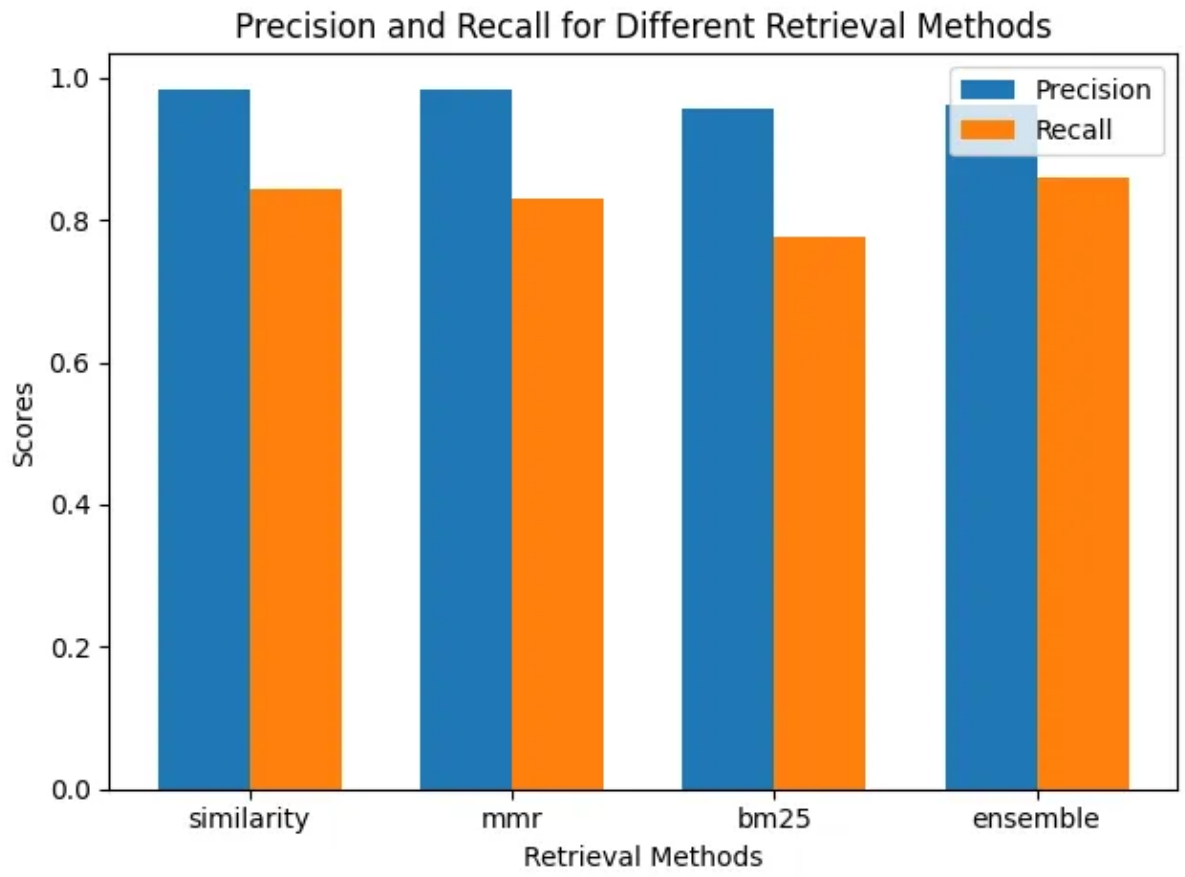
K=3 일 때

FAISS 결과



	Similarity	Mmr	Bm25	Ensemble
Precision	0.9680	0.9671	0.9574	0.9503
Recall	0.8097	0.7983	0.7758	0.8614

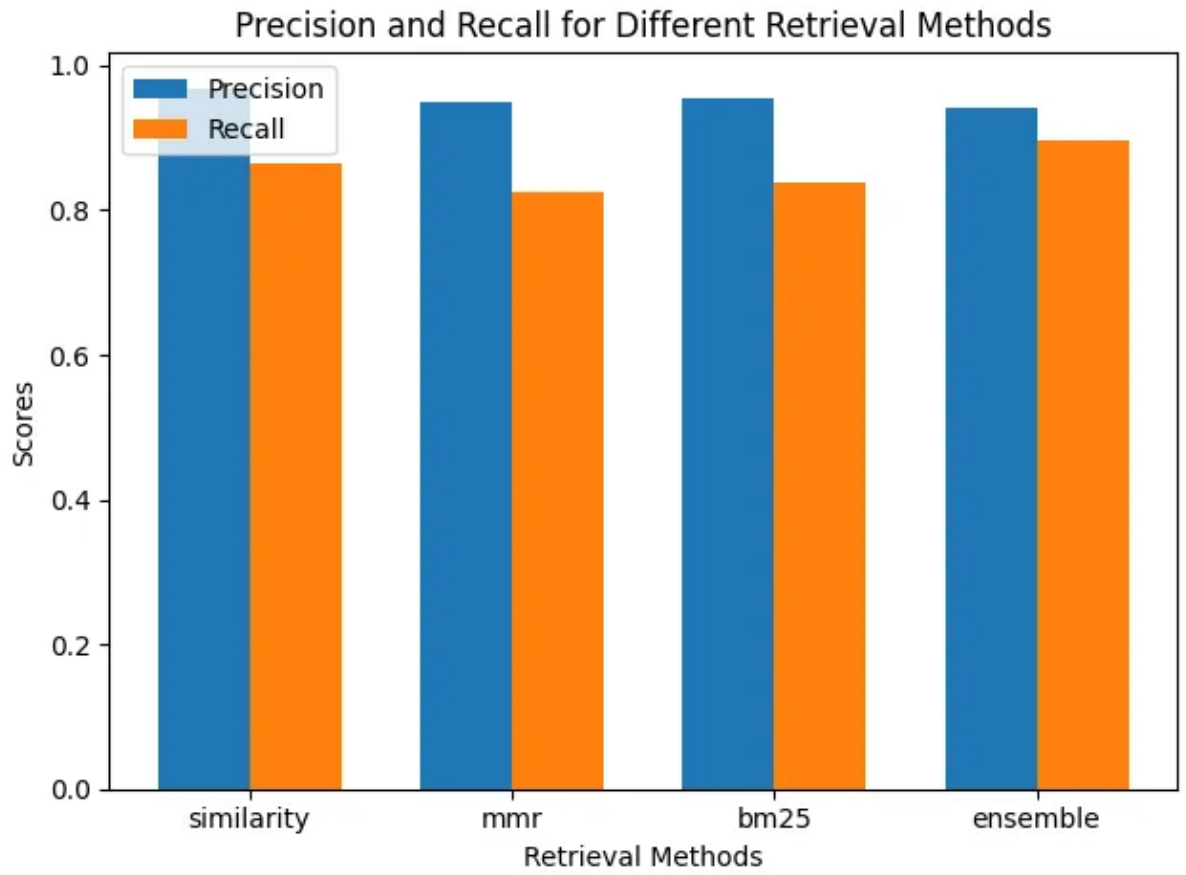
ChromaDB



	Similarity	Mmr	Bm25	Ensemble
Precision	0.9845	0.9853	0.9574	0.9613
Recall	0.8440	0.8297	0.7758	0.8585

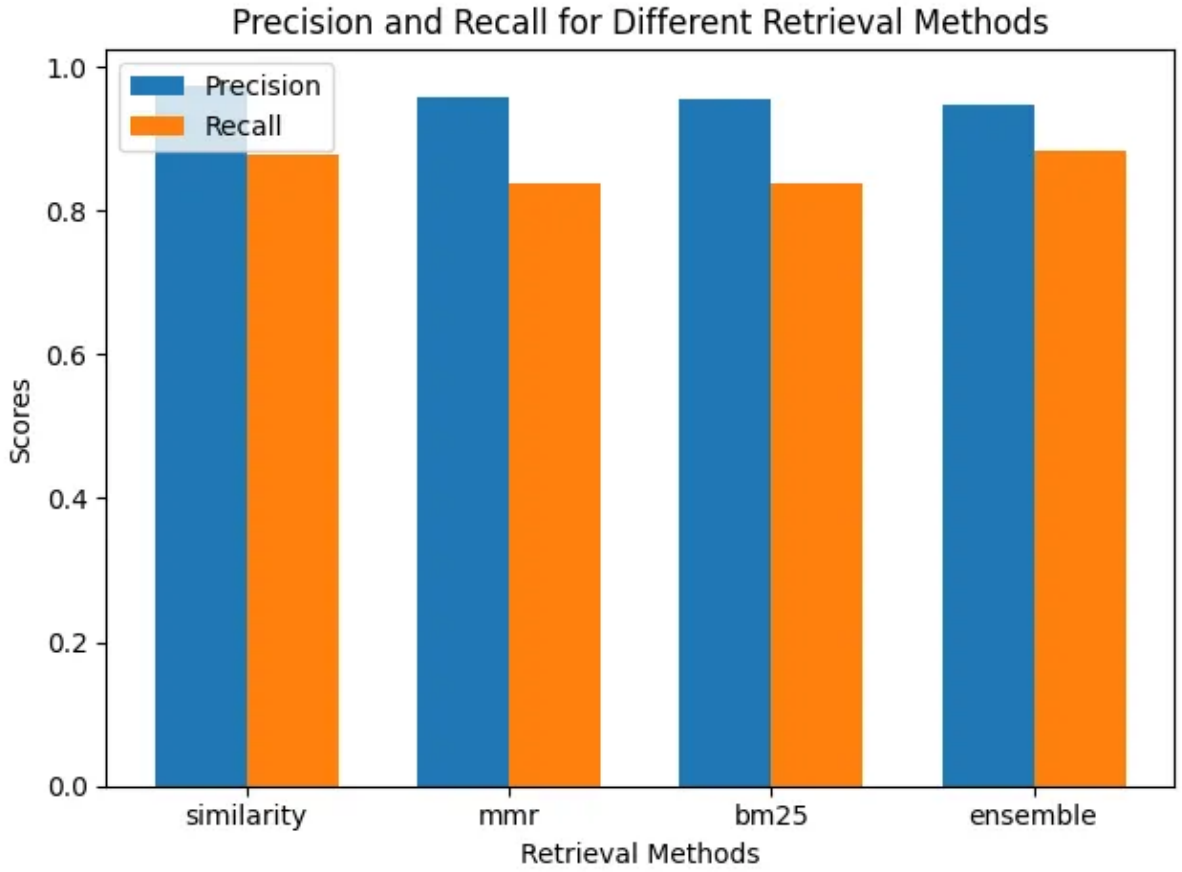
K= 5 일 때

FAISS 결과



	Similarity	Mmr	Bm25	Ensemble
Precision	0.9690	0.9501	0.9545	0.9424
Recall	0.8636	0.8244	0.8372	0.8959

ChromaDB 결과



	Similarity	Mmr	Bm25	Ensemble
Precision	0.9745	0.9578	0.9545	0.9470
Recall	0.8779	0.8384	0.8372	0.8837

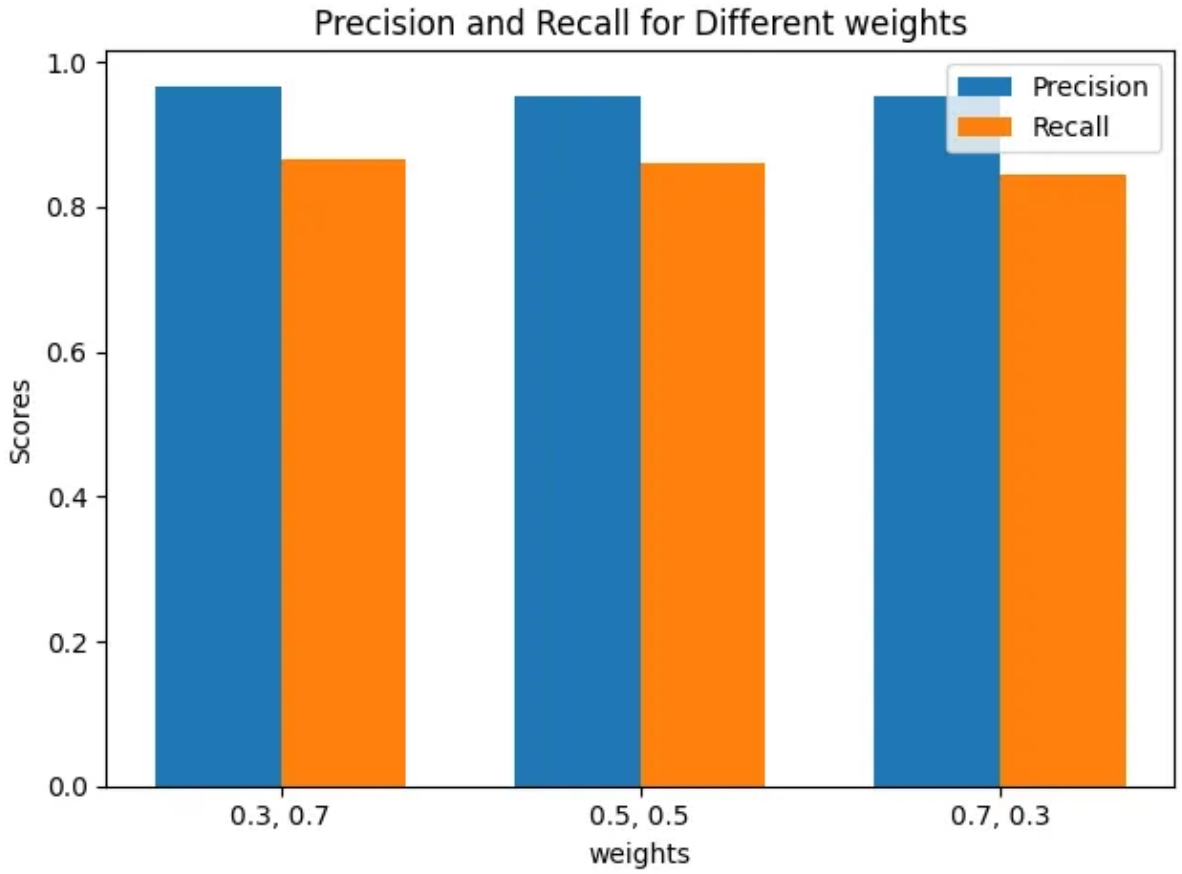
성능 평가 결과

- 유사도 검색 dense 는 Chroma 가 높게 나왔음
- ensemble 결과는 K= 3, K=5 둘 다 FAISS 에서 높게 나왔음
- K=5 에서 FAISS ensemble retriever 사용시 recall 이 0.9 에 근접

최종 선택 결과 vectorDB 는 FAISS retriever 는 ensemble 로 결정

추가적으로 ensemble 가중치 테스트 수행

bm25, similarity (0.3, 0.7) (0.5, 0.5), (0.7, 0.3), K=3



	0.3 / 0.7	0.5 / 0.5	0.7 / 0.3
Precision	0.9672	0.9535	0.9542
Recall	0.8653	0.8601	0.8459

가중치 평과 결과

K = 3 일 때 Bm25 가중치 0.3 similarity 0.7 가중치가 가장 높게 나왔으며 추가적인 평가를 진행하여 더 최적화된 가중치를 찾을 수 있겠지만 이미 평가 진행하면서 GMS 크레딧을 다수 사용했기 때문에 최종 가중치로 선택하기로 하였다.

추가 개선 전략 rerank 적용

CrossEncoder 를 활용하여 rerank 전략을 적용하여 평가

Why? 문서의 개수가 많아질 수록 recall 상승, precision 하락

→ 많은 문서에서 가져와서 CrossEncoder 활용하여 ReRanking 전략 실험 해보기

GPU 환경이 아닌 CPU 환경이기에

경량 모델

model = HuggingFaceCrossEncoder(model_name="BAAI/bge-reranker-base")

결과

K = 3 / K = 10, top_n = 3

	Ensemble	Rerank
Precision	0.9680	0.9855
Recall	0.8446	0.8295
time	150s	525s

- CPU 를 통해 rerank 를 진행하여 86 개 GT 의 평가 수행 시간이 약 3.5 배 차이 났음
- recall 이 하락한 결과를 확인함
- k=3 후보군으로 평가한 ensemble 보다 훨씬 많은 후보를 탐색하지만, 결과를 top_n=3 으로 줄이면서 일부 좋은 후보를 놓칠 수 있기에 recall 이 하락한 것으로 보임

추가적인 test 나 cpu 로 rerank 하는 건 cost 적 손해가 있다고 생각하여 ensemble retriever 로만 사용하기로 하였음

5. 결론 및 향후 과제

VectorDB : FAISS (성능, 속도 모두 충족)

Retriever : ensemble retriever 선택 (목표 recall 0.9 에 가장 근접)

목표 설정치인 재현율을 달성하진 못했지만, 현재 검색 성능 평가에는 메타데이터를 포함하지 않았다. 앞으로 LLM 과 연동하여 메타데이터를 추출하여 좀 더 검색 정확도를 높일 수 있을 것으로 기대하고 있다. 할루시네이션을 최대한 줄이고 검색 성능을 위해 임베딩 모델도 좀 더 높은 비용의 모델도 적용해 볼 예정이다. LLM 을 활용하여 검색과 결합하여 최종 결과를 만들어 낼 때에도 추가적인 테스트를 진행하고 최적화된 파이프라인을 구성하도록 할 것이다.

테스트 코드 작성의 중요성

이번에 평가를 진행하면서 테스트 코드를 작성하고 로그를 위해 print 를 많이 해보는 것의 중요성을 느꼈다. RAGAs 평가에는 LLM 을 활용하는데 평가를 위해 사용되는 크래딧이 저가형 모델인 gpt-4o-mini 를 사용했음에도 불구하고 사용량이 3 천 크레딧에서 33000 크레딧까지 상승하였다. 첫 테스트를 시도 했을 때, 완벽하지 않은 코드로 진행하게 되어서 다시 작성해야 했는데, 이미 평가가 진행되어 버린 경우 API 비용이 청구되기 때문에 이를 막아야 할 필요성이 있다고 느꼈다.

```
    if test:
        print(f"{retr_name} test")
        fill_data(_data_frame, QA_set[0]["Q"], retr, QA_set[0]["A"])

    else:
        for idx, qa in enumerate(QA_set):
            fill_data(_data_frame, qa["Q"], retr, qa["A"])
            print(f"✅{idx + 1}/{len(QA_set)}")
```

```
def evaluate_retr(all_retrievers_map, score, test=False):
```

단순히 코드 몇 줄 추가함으로 써 함수를 변형하고 수정하는데 문제 없이 테스트를 진행할 수 있었다. API 요청 비용에 대해서도 cost 를 줄이기 위해 고민을 해야함을 느끼게 되었다.

참고 논문, 서적

[정현희 and 이화영. (2012). HTP 평가항목 간편화. 미술치료연구, 한국미술치료학회 19(2), 231-244.]

[이은정 and 황세진. (2023). 인공지능 객체검출모델 기반 집-나무-사람(HTP) 그림검사의 형식적 해석 연구. 미술치료연구, 30(5), 1241-1257.]

[Buck, J.N. (1948), The H-T-P test. J. Clin. Psychol., 4: 151-159.]

[Buck, J.N. (1948), The H-T-P technique. A qualitative and quantitative scoring manual. J. Clin. Psychol., 4: 317-396.]

[김영호. (2022). 그림기반 아동심리검사에 있어 H(house)그림에 대한 신경망 기술을 적용한 심리검사 자동화 방법. 충남대학교. 석사학위논문]

[박성진. (2024). 아동 미술심리검사를 위한 AI 기반 그림 데이터 분석 모델 연구. 동의대학교. 석사학위논문]

[백원대. (2019). 검사 해석체계 구축 및 타당성 제고. 삼육대학교]

[박현주, 서명옥(2014) 투사적 심리검사자료를 활용 한 초보상담자와 경력상담자의 사례개념화 비교연구. 미술치료연구, 21(6), 1283-1304.]

Hammer, E. F. (1968). ***The House-Tree-Person (H-T-P) clinical research manual***. Western Psychological Services.

신민섭.(2003). 그림을 통한 아동의 진단과 이해. 서울: 학지사. 김동연, 공마리아(2002). HTP 와 KHTP 심리진단법. 서울: 동아문화사.