심화 프로젝트 - STUBO

Deep Learning

Natural Language Processing

수험생을 위한
RAG 기반
AI 튜터링 시스템 개발

15기 | 딥러닝 김현주 김규연 박지수 이채원 정보경

목차

프로젝트 배경

문제 상황

- 수험생은 방대한 자료 속에서 정확한 개념과 기출 문제를 스스로 탐색해야 함

기존 AI는 국어 과목의 맥락 이해와 정확한 해설 제공에 한계를 보임

LLM과 RAG 기술을 활용하면 개념 설명부터 유사 기출 추천까지 자동화할 수 있어
국어 과목에 최적화된 AI 튜터링이 지금 꼭 필요한 시점

수험생을 위한
RAG 기반 AI 튜터링 시스템 개발

주제 RAG 기반 국어 문제 풀이 및 유사 기출 추천 챗봇 개발

목표 질문에 대한 설명 + 유사 기출 문제 + 출처 정보를 함께 제공

핵심 기능

국어 문제 해설 (LLM 기반)	유사 기출 추천 (RAG 기반)

기대 효과 수험생이 혼자서도 정확한 해설과 기출 연계를 통해 효율적인 학습 가능

데이터 소개

Done

2022~2025학년도 3월, 6월, 9월 모의고사 및 수능 - 문제, 정답, 해설 PDF

- 출처 기관: 대한민국 교육부 및 각 지역 교육청 (서울시교육청, 인천시교육청, 경기도교육청 등)
- 시험 주관:
  - 수능: 한국교육과정평가원 (KICE, www.kice.re.kr] (https://www.kice.re.kr/))
  - 모의고사: 서울특별시교육청, 인천광역시교육청, 경기도교육청 등에서 주관
- 공개 경로: 평가원 또는 교육청 공식 웹사이트, EBSi
- 해당 자료는 대한민국 정부 및 지방자치단체 산하 공공기관에서 제공한 **공공저작물**로,
- **공공누리 제1유형(출처 표시, 자유 이용)** 또는 **제4유형(출처 표시, 비영리·변경 금지)** 기준에 따라 활용하였습니다.

데이터 소개 (문학)

Done

수능특강 문학 교사용 PDF

- 제작 및 배포 기관: EBS(한국교육방송공사), EBSi 고등 사이트
- 접근 경로: EBSi 홈페이지 → 교재 자료실 → 수능특강 → 과목 선택 → 교사용 자료 다운로드
- 자료 출처: https://www.ebsi.co.kr
- 본 자료는 한국교육방송공사(EBS)에서 제작한 『수능특강 문학』 교사용 자료를 기반으로 하였으며, 비영리 교육 목적에 따라 원본 그대로 활용하였습니다.

문학 교과서가 사랑한 작품500 (고전산문, 고전운문, 현대산문, 현대시) PDF

- 본 자료는 네이버 블로그 'kkongsstore' (https://blog.naver.com/kkongsstore/221551221637)에서 제공한, 미래엔 교과서에 수록된 문학작품 요약본을 기반으로 하였습니다.
- 원 저작물인 미래엔 교과서 및 작품에 대한 저작권은 출판사 및 원저작자에게 있습니다.
- 본 프로젝트에서는 해당 자료를 직접 배포하거나 수정하지 않았으며, 교육 및 연구 목적으로 참고하여 활용하였음을 명확히 밝힙니다.

데이터 전처리 과정

텍스트 추출 + 전처리

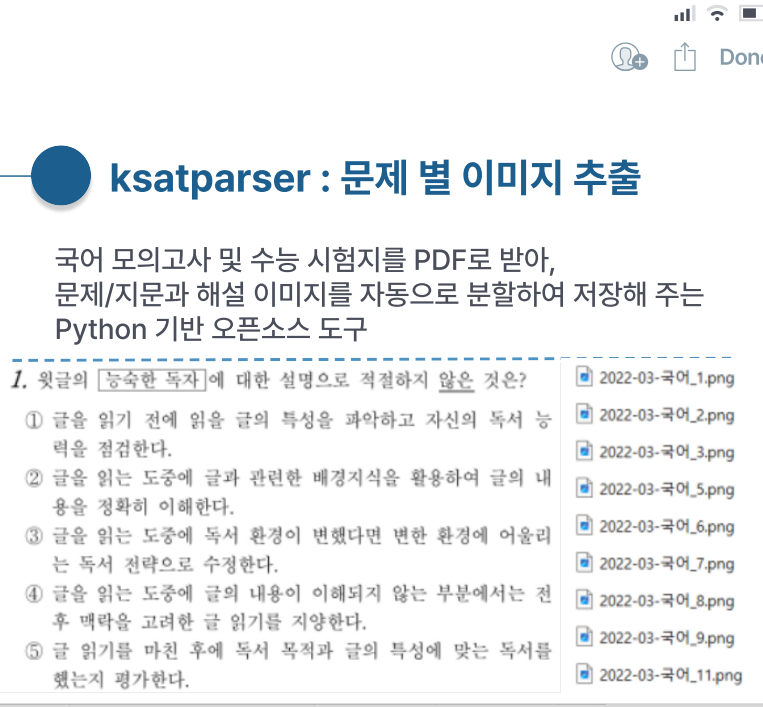
PyPDF 라이브러리의 PdfReader

텍스트 추출 후 줄바꿈 등 전처리

```
텍스트 저장 완료: /content/drive/MyDrive/save_text/2022-06-국어.txt
텍스트 저장 완료: /content/drive/MyDrive/save_text/2022-09-국어.txt
텍스트 저장 완료: /content/drive/MyDrive/save_text/2022-수능-국어.txt
텍스트 저장 완료: /content/drive/MyDrive/save_text/2023-03-국어.txt
텍스트 저장 완료: /content/drive/MyDrive/save_text/2023-06-국어.txt
텍스트 저장 완료: /content/drive/MyDrive/save_text/2023-09-국어.txt
텍스트 저장 완료: /content/drive/MyDrive/save_text/2023-수능-국어.txt
텍스트 저장 완료: /content/drive/MyDrive/save_text/2024-03-국어.txt
텍스트 저장 완료: /content/drive/MyDrive/save_text/2024-06-국어.txt
텍스트 저장 완료: /content/drive/MyDrive/save_text/2024-09-국어.txt
텍스트 저장 완료: /content/drive/MyDrive/save_text/2024-수능-국어.txt
텍스트 저장 완료: /content/drive/MyDrive/save_text/2025-03-국어.txt
텍스트 저장 완료: /content/drive/MyDrive/save_text/2025-06-국어.txt
텍스트 저장 완료: /content/drive/MyDrive/save_text/2025-09-국어.txt
텍스트 저장 완료: /content/drive/MyDrive/save_text/2025-수능-국어.txt
```

ksatparser : 문제 별 이미지 추출

국어 모의고사 및 수능 시험지를 PDF로 받아,
문제/지문과 해설 이미지를 자동으로 분할하여 저장해 주는
Python 기반 오픈소스 도구



데이터 전처리 과정

Parsing

지문/문제/해설 parsing해서 json 파일에 저장

예시)


```
{
  "question": "35.윗글을 읽고 이해한 내용으로 적절하지 않
  "type": "언어와매체",
  "med": false,
  "tag": []
},
{
  "question": "36.윗글을 바탕으로 <보기 >의 ⓐ ~ⓔ를 통
  "type": "언어와매체",
  "med": false,
  "tag": []
},
```


Tagging

Done

prompt 설계 (최대 6개까지 추출하도록)

GPT-4o 사용해서 자동으로 tagging

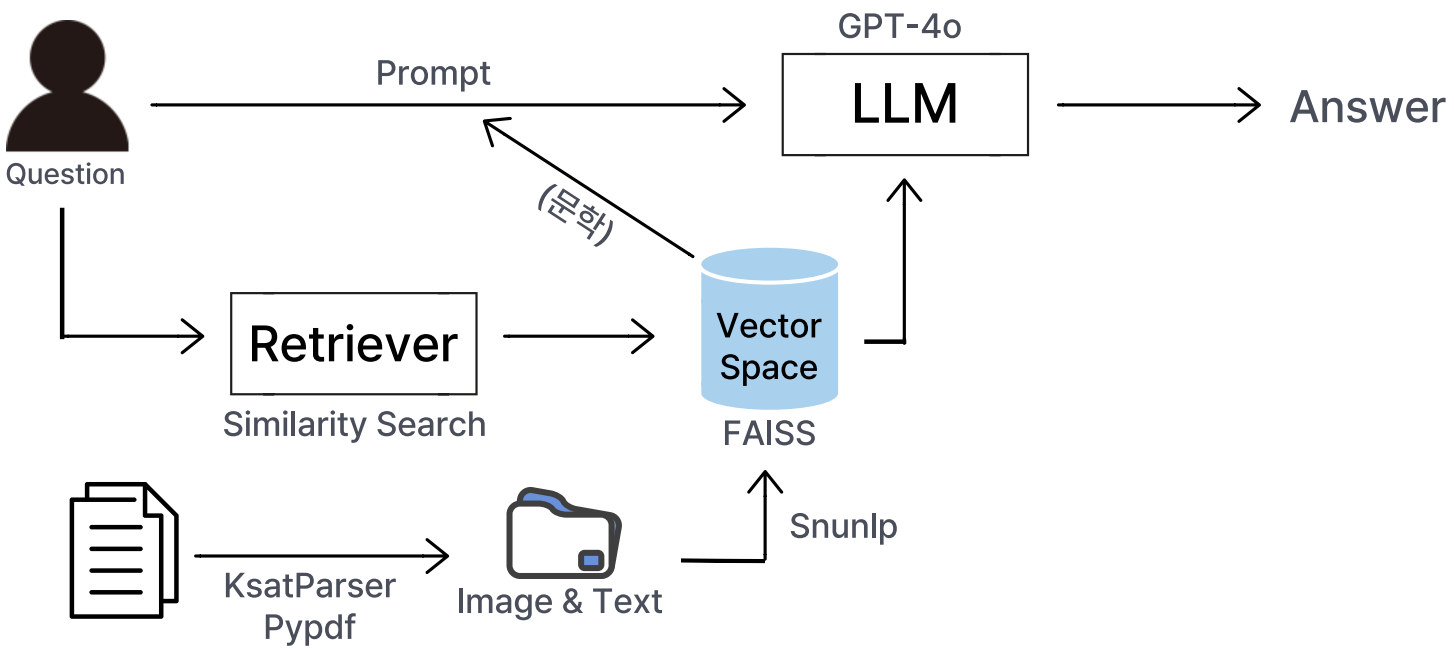
json 파일에 저장


```
{
  "pNum": 2,
  "year": "2022",
  "month": "03",
  "passage": "(가)Mn식품처럼 개인 차원에서
  "question_type": "복합문제",
  "type": "비문학",
  "genre": "사회-경제",
  "keywords": [
    "공공재",
    "정책 딜레마",
    "지방 정부 재정 지원"
  ],
  "start_Qnum": 4,
  "end_Qnum": 9
},
```


과목별 json 통합

→ Embedding Input ready!

모델 아키텍처


Done
GPT-4o
Prompt
LLM
Answer
Question
(문학)
Retriever
Similarity Search
Vector Space
FAISS
KsatParser
Pypdf
Image & Text
SnunIp

임베딩 & 검색기 실험

Done

- **Embedding 실험**

	Top-1 점수	Top-2 점수	Top-3 점수	전체 점수
jhgan/ko-sbert-nli	1	**15**	1	17
snunlp/KR-SBERT-V40K-klueNLI-augSTS	**25**	1	**5**	**31**
BM-K/KoSimCSE-roberta	1	1	1	3

- **Retrieval 실험**

	Top-1 점수	Top-2 점수	Top-3 점수	전체 점수
BM25	25	1	1	27
FAISS-Flat	**25**	1	**5**	**31**

임베딩 & 검색기 실험

Done

- **Embedding 실험**

	Top-1 점수	Top-2 점수	Top-3 점수	전체 점수
jhgan/ko-sbert-nli	1	**15**	1	17
snunlp/KR-SBERT-V40K-klueNLI-augSTS	**25**	1	**5**	**31**
BM-K/KoSimCSE-roberta	1	1	1	3

- **Retrieval 실험**

	Top-1 점수	Top-2 점수	Top-3 점수	전체 점수
BM25	25	1	1	27
FAISS-Flat	**25**	1	**5**	**31**

추천된 유사 문제들은 문제 유형은 유사하지만, 지문 내용의 유사성은 낮은 한계가 있었습니다.

해결 방안 - Tagging

Done

문제	해결
문제 유형은 비슷하나 지문 내용은 다름	GPT API를 활용해 문항별 핵심 키워드 추출
의미 기반 연결이 부족함	태그 기반 정확한 의미 유사도 추천 도입
실질적으로 비슷한 문제로 느껴지지 않음	수험생 입장에서 정말 유사한 문제 확인 가능

GPT vs OCR 실험

Done

- 이미지에서의 텍스트 추출 성능 비교 실험

	문항 구분 가능 여부	보기 항목 정확도	텍스트 신뢰도	전체 점수
EasyOCR	9	4	7	20
PaddleOCR	8	6	5	19
GPT-4o	**10**	**10**	**10**	**30**

GPT vs OCR 실험

- 이미지에서의 텍스트 추출 성능 비교 실험

	문항 구분 가능 여부	보기 항목 정확도	텍스트 신뢰도	전체 점수
EasyOCR	9	4	7	20
PaddleOCR	8	6	5	19
GPT-4o	**10**	**10**	**10**	**30**

GPT-4o는 가장 우수한
성능을 보였으나,
비용 부담으로 인해
텍스트 추출의 정확도가
특히 중요한 과목에만 적용하고,
그 외 과목은 성능 2위인
EasyOCR을 활용

깨끗한 텍스트 vs OCR 사용해서 답변 (비문학)

Done

< 깨끗한 텍스트 >

2025-06.txt → ✅ **오답 개수: 1**

- 오답 문제는 <보기> 문제였는데, ChatGPT와 Perplexity에게 물어봤을 때도 오답
- 조금 더 개입해서 힌트를 주니 ✅ 오답 개수: 0

2025-09.txt → ✅ **오답 개수: 2**

- 어휘 문제, 관계 역전 문장을 제대로 분별하지 못함.
- 조금 더 개입해서 힌트를 주니 ✅ 오답 개수: 0

2025-수능.txt → ✅ **오답 개수: 3**

- <보기>와 비교 문제
- 힌트 추가 시에도 성능이 안 좋았음.

< OCR로 추출한 텍스트 >

항목	수치	비율 (%)
정답 수	33	**64.7%**
오답 수	18	**35.3%**
총 문항 수	51	100%

전반적인 오답 경향

1. 지문 논지 파악 부족
2. 보기 - 질문 연계 약함
3. 질문 제한 조건 누락
4. 지시어나 대상 파악 오류

* 깨끗한 텍스트: 사람이 인위적으로 개입한 텍스트

깨끗한 텍스트 vs OCR 사용해서 답변 (문학)

Done

Test Data : 2025학년도 6월, 9월, 수능

< 깨끗한 텍스트 >

❌ 오답 문항 수

- 2025년 6월: 3개 (20, 26, 34)
- 2025년 9월: 2개 (18, 27)
- 2025년 수능: 4개 (20, 21, 26, 32)
- 총 오답: 9문항

요약

- 총 문항 수: 51
- 정답 수: 42
- 오답 수: 9

정답률 : 약 82.4%
오답 중 정답률이 50% 아래인 문제들이 많으며
두번째로 높은 선택비율의 정답을 선택한 것이 많음
→ LLM 추론 능력 부족
오답 중 대부분이 <보기>가 포함된 문제
보기의 조건을 지문에 정확히 적용하는 문제에 약함

< OCR로 추출한 텍스트 >

❌ 오답 문항 수

- 2025년 6월: 18, 20, 21, 26, 27, 29, 32 → 7문항
- 2025년 9월: 24, 27 → 2문항
- 2025년 수능: 18, 19, 20, 21, 26, 32 → 6문항
- 총 오답: 15문항

요약

- 총 문항 수: 51
- 정답 수: 36
- 오답 수: 15

정답률 : 약 70.59%
깔끔한 텍스트에서 실험했을 때 틀렸던 문항들이
이번에도 틀리는 모습을 확인
지문 이미지의 텍스트를 잘못 추출했을 때
그 지문의 문항들이 틀려 정답률이 낮아짐

* 깨끗한 텍스트: 사람이 인위적으로 개입한 텍스트

인사이트 (Insight)

Done

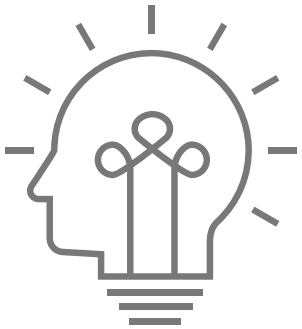
- 정답률이 낮은 문항일수록, 모델이 두 번째로 많이 선택된 선택지를 고르는 경향이 나타남
  - → LLM이 선택지 간 미세한 의미 차이를 구분하는 데 어려움을 겪고 있음을 시사
- <보기> 조건을 적용해야 하는 문항에서 오답 비율이 높음
  - → 보기-지문 간 조건 연계 추론에서 취약점을 보였으며, 이는 LLM의 복합적 지문 이해 능력 한계로 해석 가능
- OCR 텍스트 품질이 모델 성능에 직접적인 영향을 미침
  - → 깔끔한 텍스트의 경우 정답을 맞췄지만, 왜곡되거나 누락이 발생하면 같은 문항에서도 오답을 보임

한계점 및 추후 보완사항

Done

<보기>가 포함된 문항에서 유의미하게 오답률이 높게 나타남

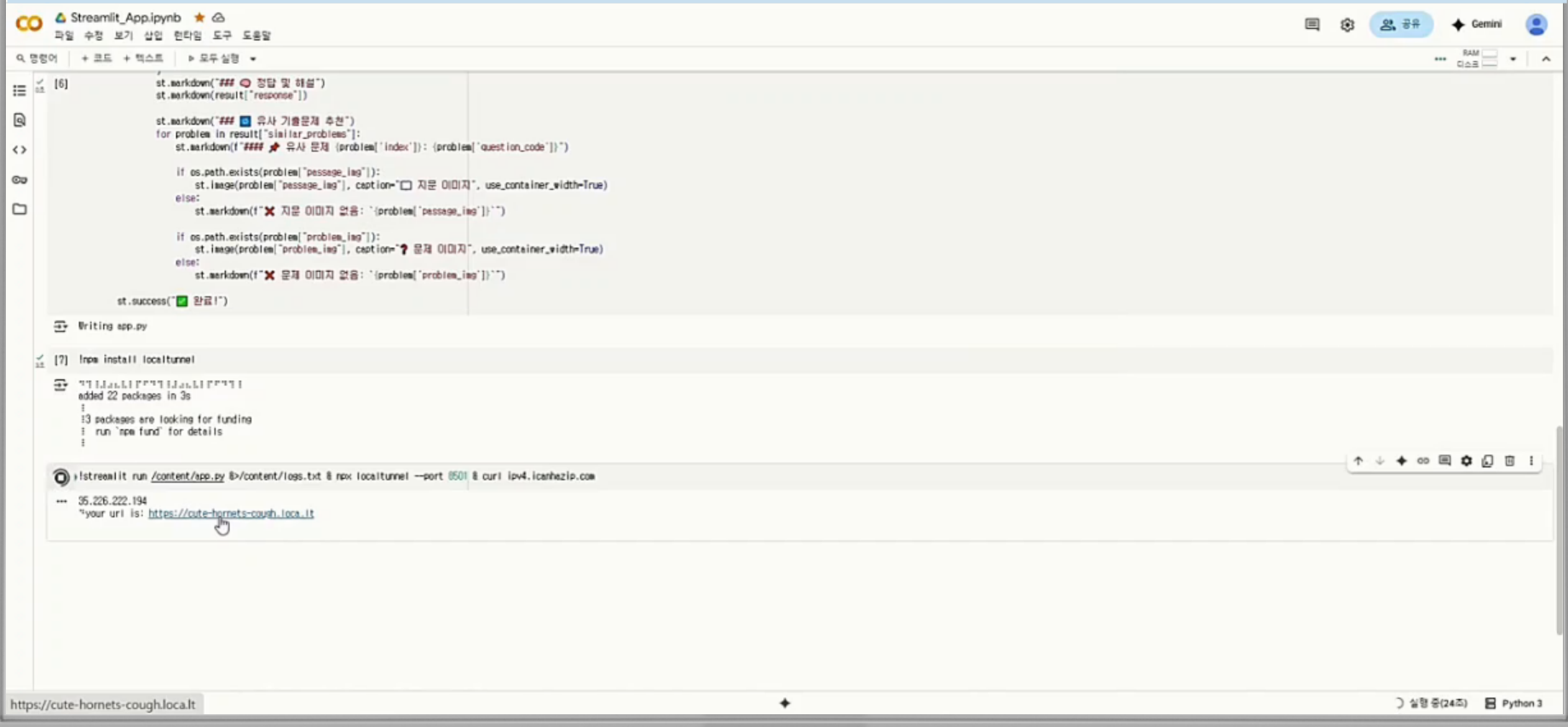
<보기> 포함 문항에 특화된 입력 구조 및 설명 유도 전략 설계
→ CoT 방식 프롬프트 설계를 통해 복합 추론을 유도하고
입력 텍스트 내 구조를 명확히 하여 모델이 보기 조건을 놓치지 않게 함



OCR이 부정확하게 추출된 지문에 포함된 문항들은 정답률이 떨어짐

OCR confidence score 기반 필터링 도입
→ OCR이 불완전할 경우 재업로드 요청 메시지 출력

시연 영상


Streamlit_App.ipynb
파일 수정 보기 삽입 런타임 도구 도움말
공유
Gemini
st.markdown("## 정답 및 해설")
st.markdown(result["response"])
st.markdown("## 유사 기출문제 추천")
for problem in result["similar_problems"]:
st.markdown(f"### 유사 문제 {problem['index']}: {problem['question_code']}")
if os.path.exists(problem["passage_img"]):
st.image(problem["passage_img"], caption="지문 이미지", use_container_width=True)
else:
st.markdown(f"지문 이미지 없음: `{problem['passage_img']}`")
if os.path.exists(problem["problem_img"]):
st.image(problem["problem_img"], caption="문제 이미지", use_container_width=True)
else:
st.markdown(f"문제 이미지 없음: `{problem['problem_img']}`")
st.success("완료!")
Writing app.py
[7] !npm install localtunnel
added 22 packages in 3s
3 packages are looking for funding
run `npm fund` for details
!streamlit run /content/app.py &>/content/logs.txt & npx localtunnel --port 8501 & curl ipv4.icanhazip.com
35.226.222.194
your url is: https://cute-hornets-cough.loca.lt
https://cute-hornets-cough.loca.lt
실행 중(24초)
Python 3

TAVE 최종 컨퍼런스

감사합니다

발표자 박지수

수험생을 위한 RAG 기반 AI 튜터링 시스템 개발