

랙(RAG) 걸린 팀

TEAM 3 | 서주하 강동우 김철현 조민정 조용원



목차

01

Chapter 1. 서론

1. 철현씨의 약속
2. EDA
3. 프로젝트 구조
4. System Architecture

02

Chapter 2. 실험과정

1. Version 1.0
2. Version 2.0
3. Version 3.0

03

Chapter 3. 결론

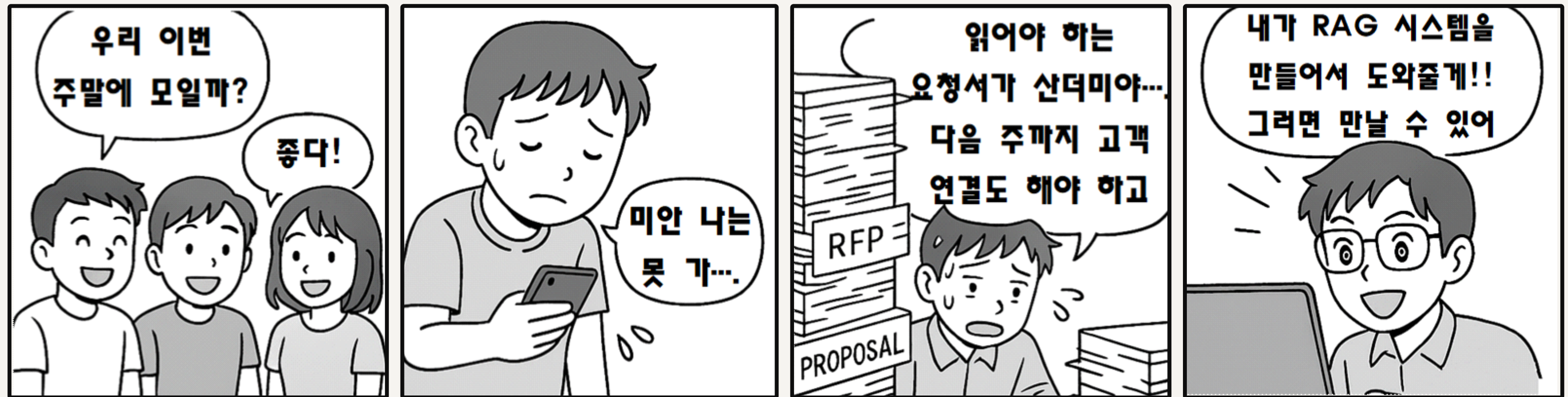
1. OPENAI Model
2. Hugging Face Model
3. 어느 모델이 더 성능이 좋을까?

01

Chapter 1. 서론

1. 철현씨의 약속
2. EDA
3. 프로젝트 구조
4. System Architecture

철현씨의 약속



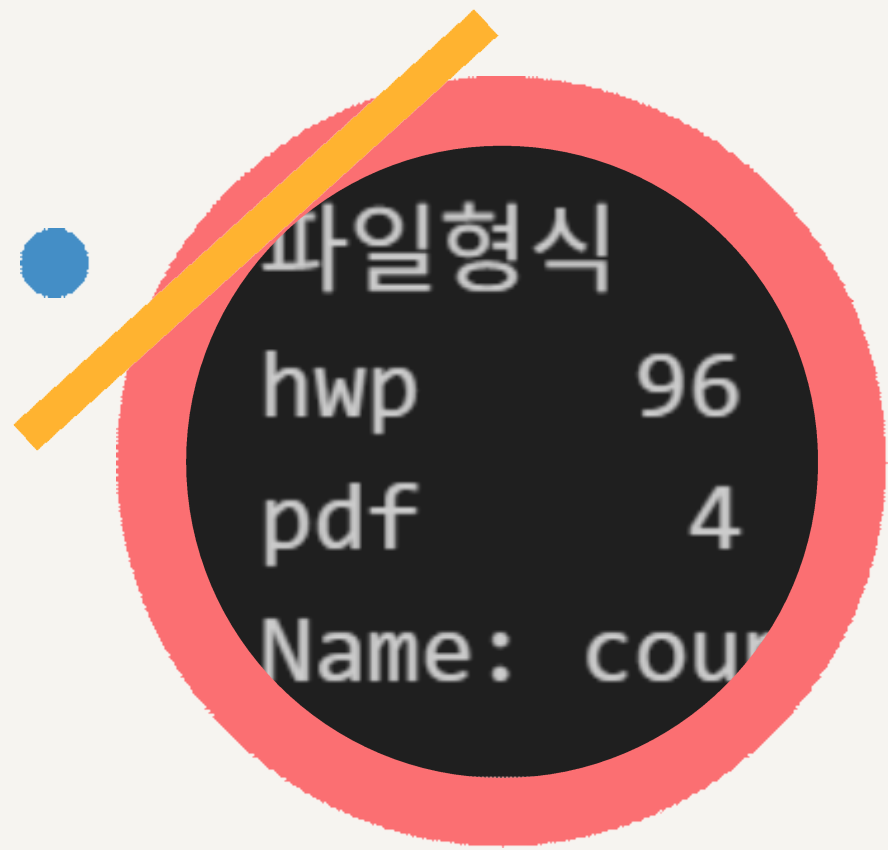
철현씨의 친구는 밀린 제안서를 처리해야 해서 약속에 참석하지 못 하는데...
이 모습이 안타까운 철현씨는 친구를 위한 RAG 시스템을 만들어 주게 되었습니다.

RAG

OPEN AI

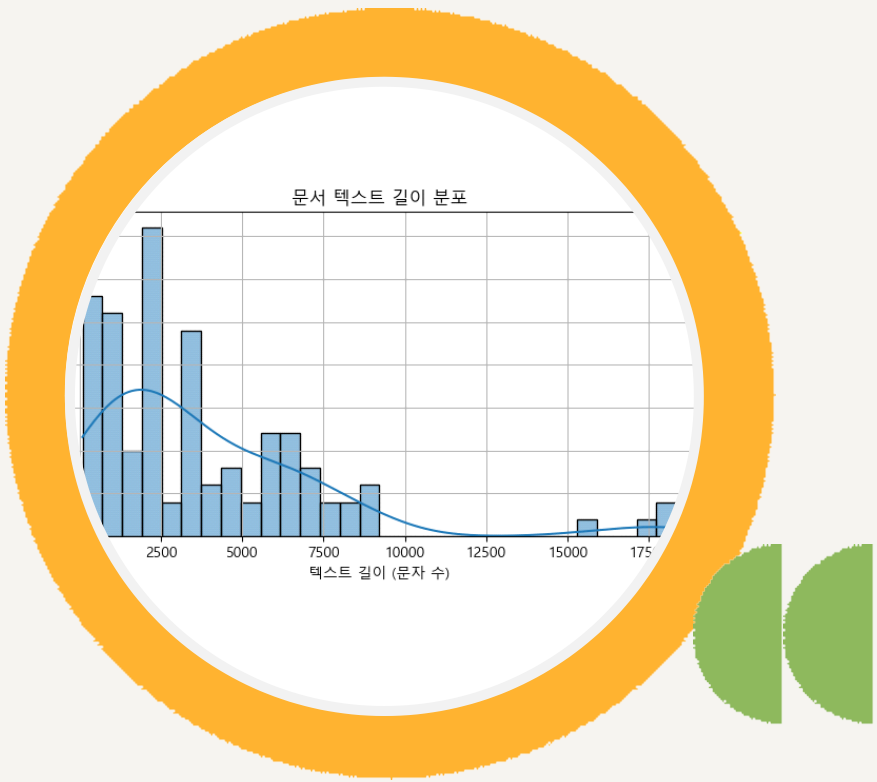
HUGGING
FACE

EDA



HWP 문서 처리

- 전체 문서의 96건이 HWP 형식
- PDF는 4건
- RAG 성능의 핵심은 텍스트 변환 안정성



매우 다양한 텍스트의 길이

- 최소 89자 ~ 최대 18,335자
- 문서에 따라 과도한 청크 or 하나로 뭉침
- 검색 성능 저하 유발



키워드 등장 빈도

- “정보화, 플랫폼, 빅데이터, 고도화, 연구개발”
- RAG 시스템: 핵심 키워드 이해 및 검색 정확도 향상
- 입찰 전략: 유망 기술·도메인에 집중
- 컨설팅: 시장 트렌드 기반 타깃 고객사 발굴

Chapter 1. 서론

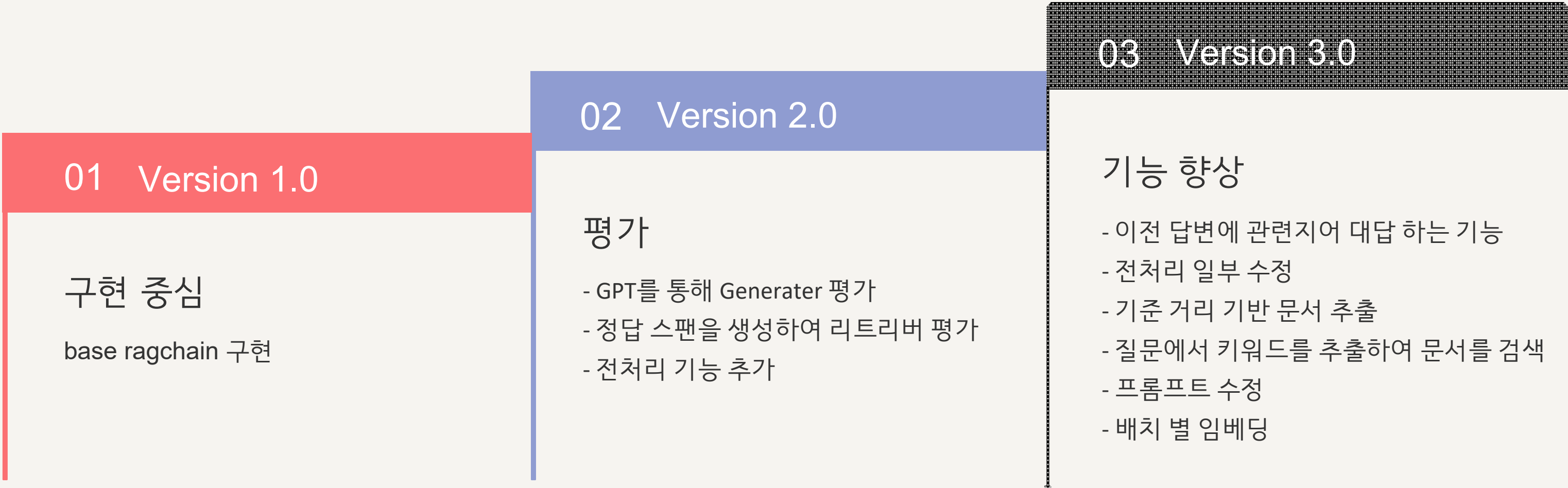
프로젝트 구조

```

codeit_mid_project/
├── data/                                # 원본 문서 및 생성된 데이터 (git: x)
│   ├── raw_data/                       # 원본 PDF, HWP 파일 등 (git: x)
│   ├── pdf_data/                       # HWP 파일을 PDF로 변환(실제로 사용할 데이터) (git: x)
│   └── processed/                      # 처리된 텍스트, chunk 등 (git: o)
│       └── span_list.json              # span_list(span_list.json) 위치
├── embeddings/                         # 벡터 임베딩 저장소 (FAISS 등)
├── experiment/                         # 실험 결과 폴더
│   ├── test_experiment1/
│   ├── test_experiment2/
│   └── test_experiment3/              # 실험 내용(결과, 사용한 LLM, 하이퍼파라미터 설정 등)
│       ...
├── rag_chain/                          # RAG 파이프라인 구성 모듈
├── scripts/                           # 임의성 스크립트 (데이터 변환, 주기화 등)
├── utils/                             # 각종 도구 스크립트
├── test.ipynb                         # 실험용 주피터 노트북
├── main.ipynb                         # 코드 통합(보고서용 주피터 노트북)
├── README.md
└── .gitignore

```

System Architecture

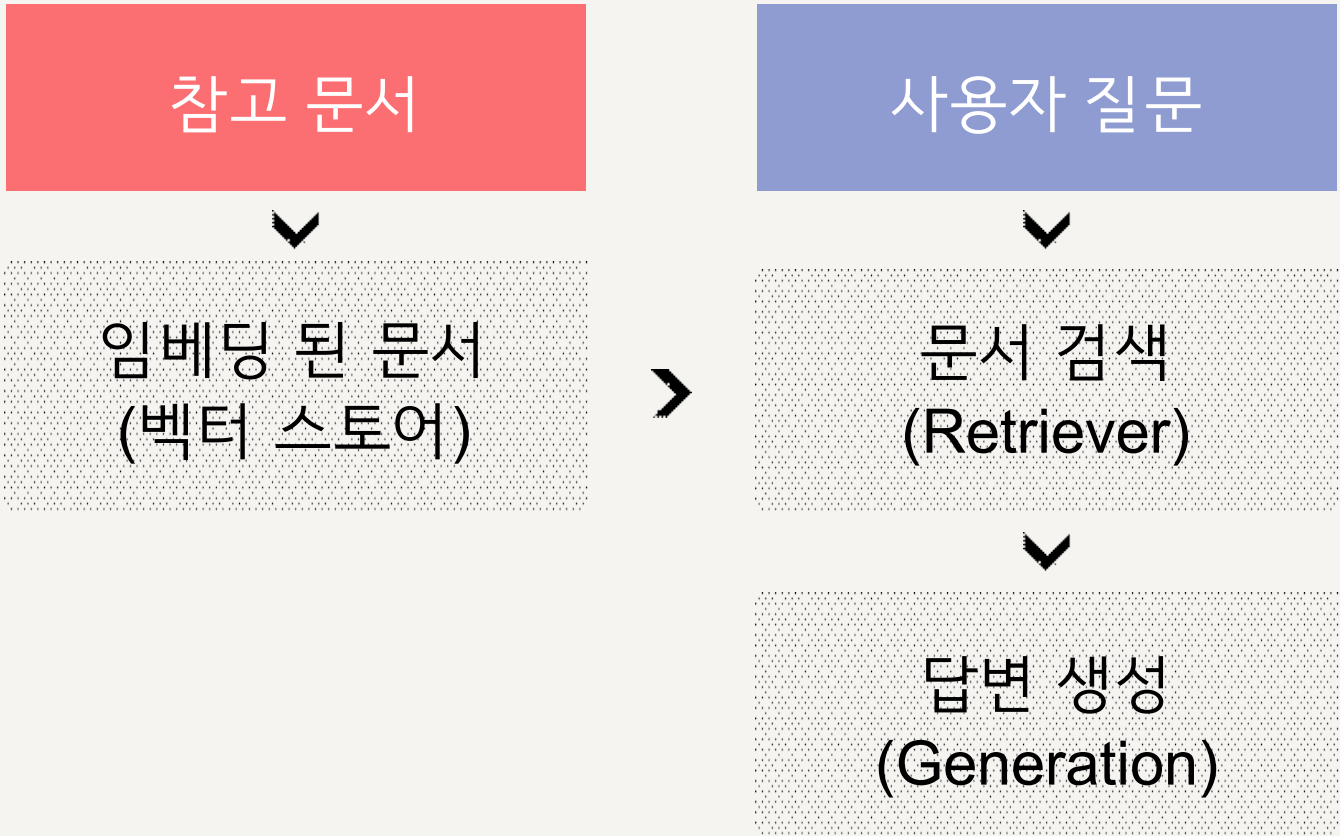


02

Chapter 2. 실험

1. Version 1.0
2. Version 2.0
3. Version 3.0

Version 1.0 - 개요



목 적

제안요청서(RFP)를 바탕으로 질문에 대한 적절한 답변을 자동으로 생성하는 RAG 시스템 기본 구조 구현.

구성 요소

1. 문서 청크 : 참고 자료를 일정길이의 청크 단위로 나누어서 저장
2. 문서 임베딩 및 리트리버 생성 : 청크 문서를 임베딩하여 벡터스토어에 저장하고 사용자의 입력을 받아서 벡터스토어에서 가장 유사한 문서를 찾아주는 리트리버를 생성
3. RAG Chain 구성 : 사용자의 질문을 받아서 유사한 문서를 찾고, 프롬프트를 만들어 자연어 처리 모델에 전달하고 생성된 답변을 돌려주는 전체 RAG Chain을 구성

Q. 한국원자력연구원이 발주한 전력망개선 사업 프로그램 사업에서 제기된 보안 위반 사항은 무엇인가요?

A. 한국원자력연구원이 발주한 전력망개선 사업 프로그램 사업에서 제기된 보안 위반 사항은 다음과 같습니다.

- 공문 소프트웨어 미설치 : 공문 소프트웨어를 설치하지 않은 경우가 보안 위반으로 규정되어 있으며, 이에 대한 위험성 및 조속감독자에 대한 증정여, 개발 방지를 위한 조치계획 제출, 특별 보안 교육 실시 등의 조치가 요구됩니다. 또한, 사업 수행 중 획득한 자료의 정보의 유출 금지, 사업 종료 시 완전 폐기 또는 반납 의무, 하도급업체에 대한 동일한 보안 준수 확인 및 보안확약서 징구, 그리고 보안사항 위반 시 법적 책임과 사업 참여 권한 또는 손해배상 갈수 등의 보안 의무 위반 사례가 포함될 수 있습니다. 요약하면, 주요 보안 위반 사항은 공문 소프트웨어 미설치와 기밀정보 유출 및 보안확약서 미준수 등이며, 이에 따른 엄격한 처벌과 관리가 이루어지고 있습니다.

Version 2.0 - 개요

v1.0 의 문제점

- 질문에 대한 답변은 확인 가능하지만, 답변의 품질 평가 불가
- 답변이 참고한 문서를 확인 결과, 동일한 어절 또는 음절의 반복과 불규칙한 띄어쓰기 확인



개선사항

- OpenAi의 'gpt-4o' 모델을 활용하여 Generator 품질 평가
- 실험 과정에서 생성된 청크 문서를 활용하여, 예상 질문·답변 쌍과 참조 문서 번호를 매칭 후, 질문 시 참조 문서와 비교해 정답 스펜(정답 범위 포함 여부) 평가를 진행

필요성

- ## 개선사항

- 11.2 알류 및 시스템 현황 복수의 결원 주식 변처 기업 요청 이 관한 특별 조치 받 시형 에 따른 복수의 결원 주식 발행 보고 기능 으론 복수의 결원 주식 발행 보고 보완 기능 코드 호너 상황 변처 기업 요청 확인 후 복수의 결원 주식 발행 사후 권리 상황 이 기업 집단 권임 등 드나타형 구현 스톡옵션 종료 변처 24 변처 기업 스톡옵션 부여 신고 기능 으론 탑재 되어 운영 되고있는 스톡옵션 DB 를 각종 자료 이관 및 검증 작업 비 상황 변처 기업 요청 확인 후 스톡옵션 부여 신고 사료은 기업 올롱한 스톡옵션 신청 운영 권리 부른 스톡옵션 신고 확인 관리자 BackOffice 1.0.0.0 발도 구성 공동 인 증서

질문: 국정과제연구소에 다중형 리눅스운영체제를 도입한 시점에서 요구하는 산출물 제출 방식은 무엇인가요?
답변: 국정과제연구소에 다중형 리눅스 운영 체제는 고도화 시점에서 요구하는 산출물 제출 방식은 다음과 같습니다.

- 표준 산출물은 USB, CD, DVD, 메모리 등 휴대용 저장 장치 사용이 불가하며, 사설 수납 장치를 정보 시스템에 연결하여 제출하여야 합니다.
- 표준 형식의 및 표준 보고서에 외부 파일 및 그림 등으로 제출하며, 서술식 정보와 각종 편의 시스템을 융합하여 병행문을 만들 및 종전 산출물을 구분하여 해당 용 이름으로하여 사설 수납 장치에 포함하여 합니다.
- 소스 코드 일체는 USB 디스크 형태로 제출합니다.
- 계약 시에는 저조지일 것을 사전 확인서를 제출해야 하며, 계약 후 10일 이내에 지정 수납 장치를 수령 또는 공공 기관 정보 시스템 구축 운영 지침에 따라 작성하여 제출해야 합니다.
- 다른 종류 및 수납 자료 일체는 USB, 외장 하드 등 저장 매체 3개를 제출하고, 경우 보고서 5부를 제출해야 합니다.
- 모든 자료를 정렬 납품하고, 자료 관리 방법을 작성하여 문서화(이력)와 전자화(파일 수명 절차) 관리 책정서에 주검 수용해야 합니다.

즉, 산출물 제출은 전차 다음과 전차로 형식으로 이루어지며, 후대용 가장 경제적 주검 사용은 거형되고, 동적 정보 시스템과 가장 대치를 통한 제출 및 출력된 모든 권리 음하를 준수하여 합니다.

중추성: 4/5 - 응답을 따르면 중추적이 단, "소스 코드 일체는 USB 디스크 형태로 제출합니다"라는 문구는 문서에서 명확히 확인되지 있었습니다. 나머지 정보는 문서에 일치합니다.

국가 예연청: 4/5 - 응답을 대부분 문서가 기반하고 있지만, 일부 세부사항은 문서에서 직접적으로 확인되지 않습니다.

중추성: 5/5 - 응답을 정확하고 이해하기 쉽게 작성되어 있습니다. 각 요구 사항에 잘 응하고 있습니다.

“““““

””””

Version 2.0 – 정답 스패를 통한 비교

필요성

GPT를 이용하여 RAG 시스템이 참고 문서로 부터 적절한 답변을 생성하였는지는 판단할 수 있지만, 적절한 참고 문서를 가져왔는지 확인 불가능

평가 방식

1. 각 질문별, '정답' 답변과 이를 근거로 하는 참고 문서(청크)의 번호를 미리 지정
2. RAG 시스템이 반환한 문서(또는 청크)와 정답 문서의 일치 여부를 비교
3. Recall@k, nDCG@k, F1 score@k 등의 검색 지표를 사용해 정량적으로 평가

장점

리트리버의 문서 검색 성능을 객관적으로 측정 가능

정답 스패 구조

```
[
  {
    "query": 질문 1,
    "answer": 답변 1,
    "chunk_id": [참고한 문서의 id]
    "source": [참고한 문서]
  },

  {
    "query": 질문 2,
    "answer": 답변 2,
    "chunk_id": [참고한 문서의 id]
    "source": [참고한 문서]
  },

  {
    "query": 질문 3,
    "answer": 답변 3,
    "chunk_id": [참고한 문서의 id]
    "source": [참고한 문서]
  },
]
```

Version 2.0 – 실험

정답 스펠 기반
질문 입력

사전에 생성한 정답 스펠(정답 문서와 답변)의 질문을 RAG 시스템에 입력하여, 시스템이 반환한 답변과
참고 문서 인덱스를 저장

Retrieval
성능 평가

RAG 시스템이 선택한 참고 문서 인덱스를 정답 스펠과 비교해 Recall, F1 score, nDCG 등 다양한
정보검색 지표로 Retrieval 성능 평가

Generato
r 성능
평가

생성된 답변을 GPT 기반 평가 기준(정확성, 근거 기반성, 명확성 등)에 따라 정량적으로 분석하여
Generator 성능 측정

실험 결과

- 전체 문서가 자연스러운 문장이 아닌 끊어진 단어를 강제로 이어 붙인 모습을 보임
- 질문과 직접적으로 관련이 없는 문서라도 내용의 유사성이 높으면 참고 문서로 선택되는 경향이 있음
- 전체 문서 중 명확하게 관련된 문서가 없더라도, 시스템은 그중 가장 유사하다고 판단한 문서를 선택하여
답변에 활용함
- 리트리버가 올바른(정확한) 문서를 찾았다는 전제 하에서, Generator의 답변 생성 성능은 전반적으로
양호하게 평가됨

Version 2.0 – 평가지표

구분

POSITIVE
NEGATIVE

POSITIVE

TP
인덱스가 찾은 문서 (O) / 정답 스펠
(O)
TN
인덱스가 찾은 문서 (X) / 정답 스펠
(X)

NEGATIVE

FP
인덱스가 찾은 문서 (O) / 정답 스펠
(X)
TN
인덱스가 찾은 문서 (X) / 정답 스펠
(O)

[이진 분류 평가 지표]

: recall@k, nDCG@k, F1score@k
- 청크 문서의 인덱스를 기준으로 정답 스펠을 만들고, 리트리버가 찾은 유사한 청크의 인덱스를 비교

```
Strict Recall@3: 0.0000 (0/33)  
nDCG@3: 0.0000
```

- 평가지표 recall@k 문제점 발생

정답 chunk_id가 33개에 비해 실제 답변의 반환한 retriever에서는 정답 chunk_id가 0으로 하나도 없었음.

Version 3.0 - 개요

v2.0 의 문제점

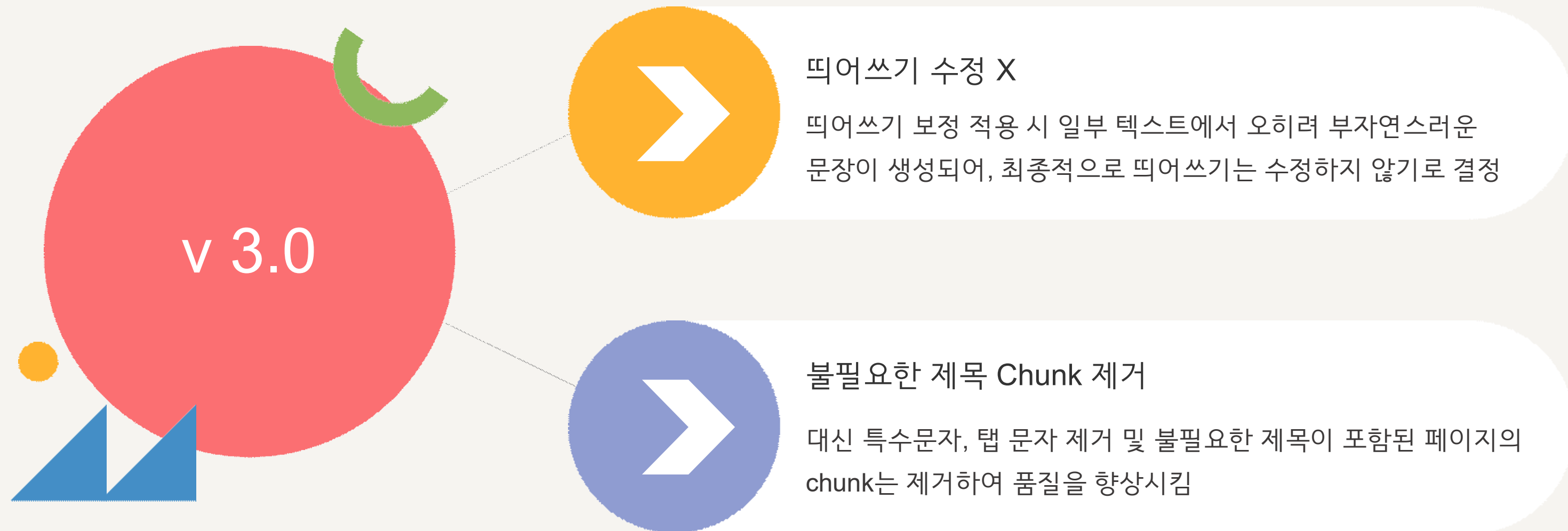
- context가 자연스러운 문장이 아니라 끊어진 단어들을 이어붙인 것처럼 만들어지는 문제
- 관련이 있는 retriever문서가 아니라 전혀 관련이 없는 문서들에서 참고하여 context를 만들고 답변하는 문제
- retriever에서 chunk_id를 기준으로 recall을 평가했는데 chunk_id가 생성되면서 정답 span에 있는 chunk_id와 간격이 달라져 매칭이 안 되는 문제



개선사항

- context가 되는 각 청크들의 text들의 전처리 방식을 수정하여 자연스러운 문장이 될 수 있게 수정
- 거리 기반으로 유사도를 보고 크게 차이 나면 해당 문서를 가지고 오지 않게 수정
- 질문에 pdf 제목에 포함된 키워드가 있는 문서들에서만 참고해서 답변을 생성하게 변경
- recall을 각 chunk_id로 비교하지 않고 조금 더 큰 틀인 page 단위로 비교

Version 3.0 – 전처리 개선



Version 3.0 – 이전 질문 기억



```
{
  "query": "국민연금공단이 발주한 이러닝시스템 관련 사업 요구사항을 정리해 줘."
  "response": "국민연금공단이 발주한 2024년 이러닝시스템 운영 용역 사업은 교육과정 운영, 콘텐츠 개발, 개인정보 보호 및 정보보안 강화 등을 주요 요구사항으로 포함하며, 계약의 투명성과 공정성을 엄격히 준수해야 합니다. 입찰 참가자는 자격 증명서류 제출, 보안서약서 작성, 기술능력 평가 및 기밀 유지 의무를 철저히 이행해야 하며, 제안서에는 실현 가능하고 정확한 계획과 보안 대책을 포함해야 합니다. 또한, 장애 발생 시 신속한 안내와 기술 이전 계획, 사용자 매뉴얼 제공 등 운영 안정성 확보를 위한 세부 조건도 명확히 규정되어 있습니다."
  "source": {
    "/data/pdf_data/국민연금공단_2024년+이러닝시스템+운영+용역.hwp.pdf": [1, 2, 38, 58, 51, 61, 41, 21],
    "/data/pdfdata/국민연금공단사업장+사회보험료+지원+고시+개정여+따른+정보시스템+보.hwp.pdf": [1, 31]}
  "query": "콘텐츠 개발 관리 요구 사항에 대해서 더 자세히 알려 줘."
  "response": "국민연금공단 콘텐츠 개발 관리 요구사항은 영상, 오디오, 이미지, 텍스트 등 다양한 미디어를 체계적으로 관리하고, 썸네일·그리드·리스트뷰 등 다양한 UI로 작업 상태와 콘텐츠 유형별 정보를 직관적으로 제공하는 것을 포함합니다. 사용자별 즐겨찾기, 태그 관리, 권한 설정과 함께 개인정보보호, 최신 법령 반영, 유지보수 및 즉각적인 수정 보완이 필수이며, 스마트러닝 연동과 학습노트 기능도 지원해야 합니다. 또한, 메타데이터 이력 관리, 통합검색, 사용자 편의성 높은 인터페이스 제공, 그리고 개발계획서 제출과 워크숍 등 체계적인 개발 및 유지관리 절차가 요구됩니다."
  "source": {
    "/data/pdfdata/KOICA+전자조달[긴급]+[지문]+[국제]+우즈베키스탄+열린+의정활동+상하원+.hwp.pdf": [39, 39, 39, 38],
    "/data/pdf_data/재단법인+한국장애인문화예술원_2024년+장애인문화예술정보시스템+이음.hwp.pdf": [43, 21],
    "/data/pdf_data/국민연금공단_2024년+이러닝시스템+운영+용역.hwp.pdf": [16, 17, 17],
    "/data/pdf_data/한국생산기술연구원_FIP3.0+고압가스+안전관리+시스템+구축+용역.hwp.pdf": [59]}
}
```

- **기능 추가 배경** : 프롬프트와 대화하듯 자연스러운 흐름을 유지하기 위해 이전 질문 기억 기능 필요성 인식
- **구현 방식** : VM 안정성을 고려하여 전체 대화 대신 이전 질문에서 5개의 핵심 키워드만 추출해 저장
- **기억 범위 제한** : 최근 대화 10개 까지만 저장되도록 제한하여 시스템 리소스 과부하 방지

Version 3.0 – 관련 문서 검색

기존 문제점	개선 방향
<p>유사도 기반 문서 선택의 한계</p> <ul style="list-style-type: none">- 질문과 유사도가 낮은 문서도 유사도 순위 상위 k개라는 이유로 선택되어 비관련 문서가 참고되는 문제 발생 <p>전체 문서 일괄 참조의 문제점</p> <ul style="list-style-type: none">- 문서 전체를 한 번에 참조하다 보니, 참고하면 안 되는 문서도 내용 유사성만으로 포함되어 부정확한 응답 가능성 존재	<p>유사도 기준 필터링 도입</p> <ul style="list-style-type: none">- 리트리버가 문서 k개를 찾더라도, 질문과의 유사도가 일정 기준 이하인 경우 해당 문서 반환하지 않음- 관련 없는 문서 참조로 인한 응답 품질 저하 방지 <p>키워드 기반 문서 탐색 방식 적용</p> <ul style="list-style-type: none">- 질문의 핵심 키워드를 활용해 관련 문서만 우선 탐색- 이후 문서 내부에서 해당 키워드가 포함된 문맥만 추출하여 정확도 향상

Version 3.0 – 배치 임베딩

기존 임베딩 방식의 문제점

- 청크 사이즈·오버랩 비율을 높이면 전체 토큰 수 증가 → 임베딩 시 토큰 초과 오류 발생
- 청크 사이즈가 너무 작으면 의미 단위가 분리되어 문맥 이해력 저하

개선 방식

- 기존: 전체 문서를 임베딩 후 Numpy Array로 변환
- 변경: 청크 단위로 임베딩 → 이후 통합하여 Numpy Array로 변환 → 메모리 분산, 효율적 처리 가능

개선 결과

- 기존에는 청크 사이즈 400 / 오버랩 10%에서도 토큰 초과 발생
- 개선 후 청크 사이즈 1000 / 오버랩 10%에서도 문제 없이 임베딩 수행 가능

기존 : 전체 청크를 한 번에 임베딩

```
texts = [doc.page_content for doc in documents]
embeddings = embedding_model.embed_documents(texts)
embeddings_np = np.array(embeddings).astype("float32")
```

개선 : 배치 단위로 임베딩

```
texts = [doc.page_content for doc in documents]
embeddings = []
max_api_batch = 100

for i in range(0, len(texts), max_api_batch):
    sub_chunk = texts[i:i+max_api_batch]
    chunk_embeddings = embedding_model.embed_documents(sub_chunk)
    embeddings.extend(chunk_embeddings)

embeddings_np = np.array(embeddings).astype("float32")
```

Version 3.0 – 프롬프트 수정

기존 프롬프트

““““
당신은 문서를 기반으로 답변하는 AI 입니다.
이전 대화 키워드: {past keyowrds}

다음의 문서를 참고하여 질문에 답하세요.

문서:
{context}

질문:
{query}
””””



변경된 프롬프트

““““
당신은 문서를 기반으로 답변하는 AI입니다.
이전 대화 키워드: {past_keywords}

문서에 있는 내용만을 사용하여 답변하세요.
문서가 충분하지 않으면 "문서에는 내용이 부족합니다."라고 답변하세요.
답변에는 반드시 관련 문서의 근거(문장 또는 항목 번호 등)를 명시하세요.
문서의 목적, 요구사항, 평가기준, 제출조건 등 RFP의 주요 항목을 우선적으로 참고하세요.
모호하거나 여러 해석이 가능하면, 가능한 해석을 모두 제시하고 각각에 대해 답변하세요.

문서를 참고하여 질문에 답하세요.

문서:
{context}

질문:
{query}
””””

03

Chapter 3. 결론

1. OPENAI Model
2. Hugging Face Model
3. 어느 모델이 더 성능이 좋을까?

OPENAI Model

Embedding Model : text-embedding-3-small



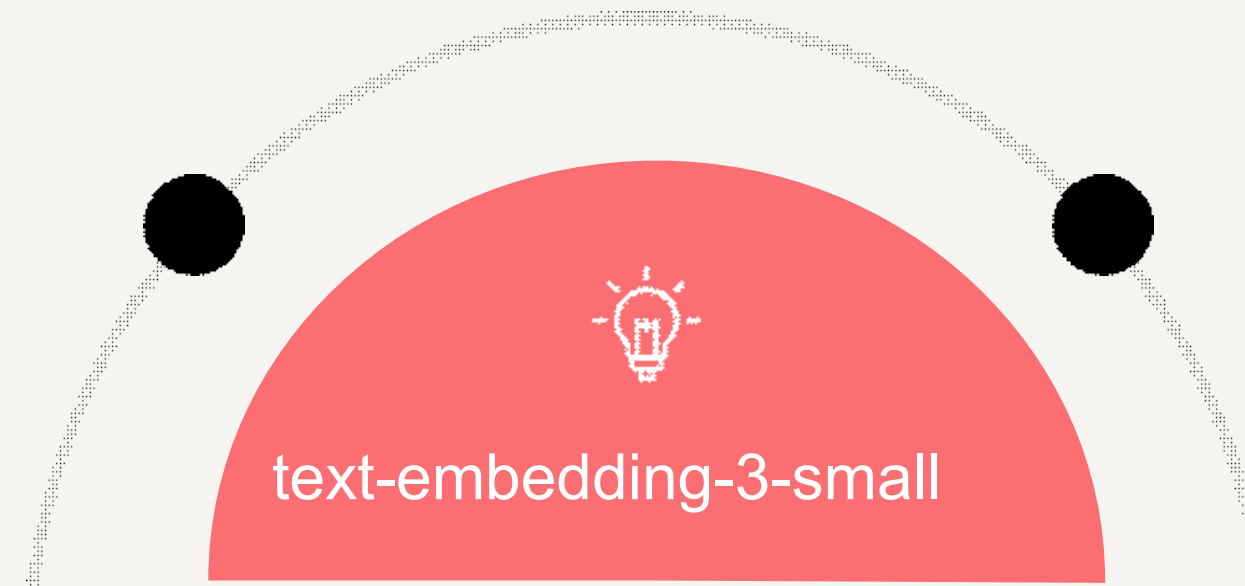
선정 이유

- OpenAI의 최신 임베딩 모델로, 의미 기반 검색 성능이 매우 우수함
- 한국어를 포함한 다양한 언어에서 안정적인 성능 보장
- 비용 대비 성능이 뛰어나고, RAG 시스템에 최적화된 구조
- LangChain 주요 프레임워크와 호환성 우수
- OpenAI API를 기반으로 안정적이고 손쉬운 사용 가능



장점

- 최신 모델로, 문장/문서 의미 표현 능력 향상
- 빠른 추론 속도와 적당한 임베딩 벡터 크기(1536차원)
- 다양한 태스크(RAG, 검색, 분류, 클러스터링 등)에 효과적
- 한국어 등 다국어 지원 → 한국어 RAG 구축에 적합
- 비교적 저렴한 가격으로 대규모 실험/운영 가능



OPENAI Model

LLM Model : gpt-4.1-mini



선정 이유

- 최신 GPT-4 계열 경량 모델
- RAG, 요약, 분류 등 다목적 대응 가능
- 높은 reasoning 성능 + 빠른 응답
- nano 대비 더 안정적이고 유연한 응답 생성
- L4 환경에서도 원활히 작동



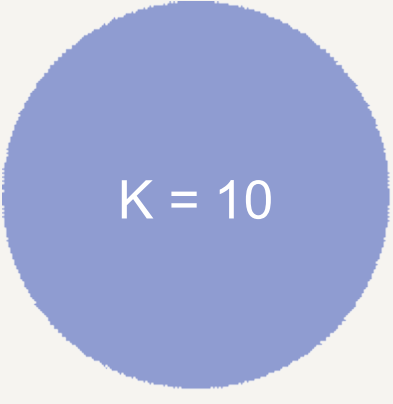
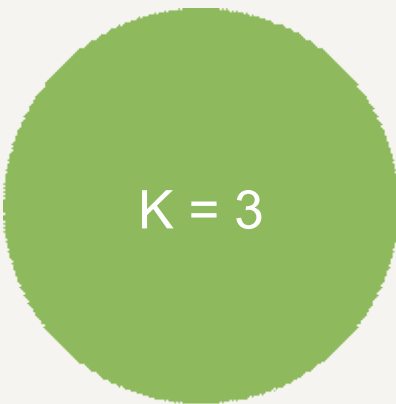
장점

- 고성능, 저비용, 빠른 응답
- GPT-4 계열 특유의 일관성과 정확성
- nano 대비 문맥 이해 및 응답 품질 우수
- 다양한 태스크에 폭넓게 적용 가능
- OpenAI API로 손쉬운 연동 및 운영 가능

OPENAI Model - 성능평가

```
{
  "query": "한국해양조사협회의 프로젝트 관리 요구사항에서 정기 및 수시보고에 대한 정의에는 어떤 내용들이 있나요?",
  "response": "한국해양조사협회의 프로젝트 관리 요구사항에서는 정기보고와 수시보고의 정의, 주기, 내용 및 형식에 대한 구체적인 지침을 명확히 규정하고 있습니다. 수시보고는 특정 상황 발생 시 이루어지며, 보고 대상과 절차가 상세히 안내되어 있어 신속한 대응이 가능하도록 되어 있습니다. 보고서 작성 시 주요 정보 포함과 제출 빈도, 제출 방법(제안서 PDF 제출 등) 및 문의처(경영관리팀)를 명확히 하여 효율적인 관리가 이루어지도록 하고 있습니다.",
  "source": ["/data/pdf_data/한국해양조사협회_2024년+항해용+간행물+품질관리+업무보조+시스템+구축.hwp.pdf": [6, 63, 38]]
}
{"query": "한국전기안전공사의 전기안전에 관련한 보안 모듈 개발할 때, 원격지 개발 보안 요구사항에 대해 설명해 주세요.",
  "response": "한국전기안전공사의 전기안전 관리시스템 보안 모듈 개발은 원격지 개발 환경에서 기밀 유지와 보안 위협 최소화가 필수적입니다. 개발자는 직무상 알게 된 정보를 엄격히 비밀로 유지하고, 보안 요구사항에 따라 수시로 보고 및 증빙서류를 작성해야 합니다. 이러한 절차와 서약서는 계약서나 확인서에 명확히 명시되어 시스템의 안전성과 신뢰성을 확보합니다.",
  "source": ["/data/pdf_data/한국전기안전공사_전기안전+관리시스템+보안+모듈+개발+용역.hwp.pdf": [58, 36, 31]]
}
{"query": "한국연구재단의 기초학문자료센터 정보시스템에 대해 설명해 주세요.",
  "response": "한국연구재단 기초학문자료센터 정보시스템은 인문학 다중화와 연구 성과 공유를 지원하며, KRM 원문출력서비스와 오픈엑세스 기능을 포함합니다. 안정적인 운영과 보안 요구사항을 충족하며, 국내외 연구자 접근성 강화와 공동기초학문자료센터 연계를 통해 종합 포털 서비스 기반을 마련합니다. 시스템은 KRM WEB과 WMS로 구성되어 IBM P728 서버에서 운영되며, 지속적인 수시 보고와 프로젝트 관리를 통해 발전하고 있습니다.",
  "source": ["/data/pdf_data/한국연구재단_2024년+기초학문자료센터+시스템+운영+및+연구성과물+DB구.hwp.pdf": [6, 1, 2, 4, 3]]
}
```

```
Strict Recall@3: 0.2143 (3/14)
nDCG@3: 0.1923
Strict Recall@5: 0.3571 (5/14)
nDCG@5: 0.2309
Strict Recall@10: 0.3571 (5/14)
nDCG@10: 0.2309
```



- k값이 클수록 recall값이 높아야 함
- 3 → 5 로 값을 바꿨을 때 유의미한 차이가 있었음

Hugging Face Model

Embedding Model : jina-embeddings-v3



선정 이유

- Jina v3는 최근 공개된 강력한 범용 임베딩 모델로, 다양한 언어와 도메인에서 좋은 성능을 보임.
- L4에서 충분히 동작 가능(large 모델 대비 경량화).
- RAG, 검색, 분류 등 다양한 태스크에 활용 가능.



장점

- 최신 모델, 다양한 태스크 지원.
- 빠른 추론 및 적당한 메모리 사용량.
- 오픈소스, 라이선스 자유로움.



Hugging Face Model

LLM Model : yanolja/EEVE-Korean-2.8B-v1.0



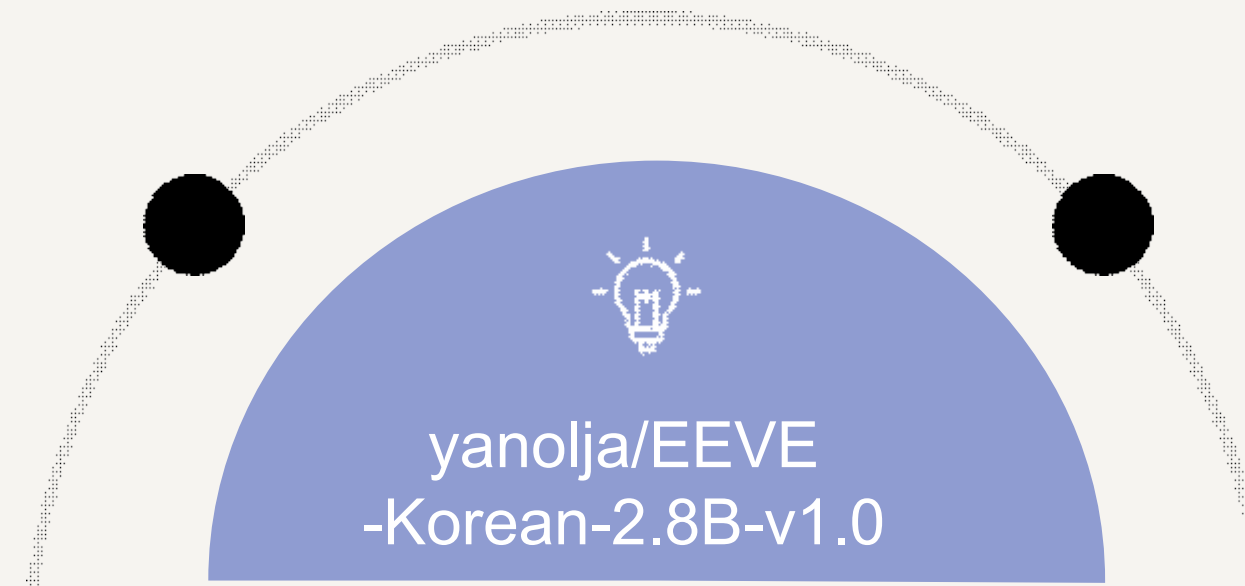
선택 이유

- 2.8B 파라미터로, L4 GPU에서 무리 없이 추론 가능(메모리 12~16GB 이상 권장)
- 한국어 RFP 문서에 최적화된 성능 기대
- 다양한 한국어 QA, 요약, 생성 태스크에서 좋은 평가



장점

- 한국어 문서에 최적화된 LLM
- L4 GPU에서 적절한 속도와 메모리 사용
- 오픈소스, 자유로운 활용



Hugging Face Model - 성능평가

```
{
  "query": "한국해양조사협회의 프로젝트 관리 요구사항에서 정기 및 수시보고에 대한 정의에는 어떤 내용들이 있나요?",
  "response": "정기 보고서는 매년 1월과 12월 두 차례씩 발행되는 것으로써 그 기간 내에 수행된 사업에 관한 결과를 정리한 것입니다. 주기적인 자료제공으로 업무 진행 상황을 파악할 뿐만 아니라 향후 계획 수립에도 도움이 됩니다.",
  "source": [
    {
      "url": "/data/pdf_data/한국철도공사+(용역)_예약발매시스템+개량+ISMP+용역.hwp.pdf",
      "score": 19
    },
    {
      "url": "/data/pdf_data/한국농수산식품유통공사_농산물가격안정기금+정부예산회계연계시스템+.hwp.pdf",
      "score": 19
    },
    {
      "url": "/data/pdf_data/그랜드코리아레저(주)_2024년도+GKI++그룹웨어+시스템+구축+용역.hwp.pdf",
      "score": 110
    },
    {
      "url": "/data/pdf_data/경기도사회서비스원_2024년+통합사회정보시스템+운영지원.hwp.pdf",
      "score": 52
    },
    {
      "url": "/data/pdf_data/한국농어촌공사_네팔+수자원관리+정보화사업-Pilot+시스템+구축용역.hwp.pdf",
      "score": 29
    }
  ]
},
{
  "query": "호계체육관 예약 시스템 구축을 위한 사업 예산이 얼마인가요?",
  "response": "사업 예산 49500만원 (총액 50억5200만원으로 약 10%)\n\n분류 : IT/IT관련정보 , 기술개발 , 로봇틱스 , 인공지능 , 머신러닝 , 데이터분석 , 기계학습 , 알고리즘 , 자동화 , 지능형제조업 , 자율주행차 , 자동차관리시스..",
  "source": [
    {
      "url": "/data/pdf_data/경기도+안양시_호계체육관+배드민턴장+및+탁구장+예약시스템+구축+용역.hwp.pdf",
      "score": 1, 27, 26, 21, 2
    }
  ]
}
```

```
➡ Strict Recall@3: 0.2143 (3/14)
nDCG@3: 0.2024
Strict Recall@5: 0.2857 (4/14)
nDCG@5: 0.2321
Strict Recall@10: 0.2857 (4/14)
nDCG@10: 0.2321
```



- k값이 클수록 recall값이 높아야 함
- 3 / 5 / 10 으로 값을 바꾸어도 큰 차이 없음

어느 모델이 더 성능이 좋을까?

1. 실험 환경 및 시간 소요

- OpenAI 실험과 동일하게 전처리 및 파라미터 조정 방식 적용
- 코랩 T4 (고용량 RAM) 기준, Experiment v2.0 전체 소요 시간 약 2시간
→ 초기 faiss.index, 청크 피클 생성 포함

2. 실험 최적화 방식

- 기존 생성된 faiss.index 및 청크 피클 재활용
- 모델 파라미터 데이터타입을 float16, bfloat16 등으로 변경하여 효율화
- 질문 1개당 평균 응답 생성 시간 4~5분으로 측정



- OpenAI는 10개 질문 기준 약 15~20분 소요
- HuggingFace 모델이 보다 더 속도가 느리고, 일부 문장에서 문장 반복 현상 발생
→ 응답 품질 및 정확도 측면에서 OpenAI 대비 성능 저하 확인

Q & A

감사합니다.

