

[Dacon] 나는야 명탐정 RAG!

재정정보 AI 검색 알고리즘 경진대회

Team | NLP 5팀

19기 이동주 20기 강민정, 이유진

CONTENTS



Outline

- 대회 설명
- Dataset
- Pipeline



Data preprocessing

- 전처리 함수
- Splitter
- Embeddings



Modeling

- RAG
- Retriever
- Generator
- Inference



Result&Discussion

- Result
- Discussion





01. Outline

01. Outline_대회 설명

DACON

재정정보 AI 검색 알고리즘 경진대회

알고리즘 | NLP | 생성형 AI | LLM | 질의응답 | F1 Score

₩ 상금 : 1,000만원

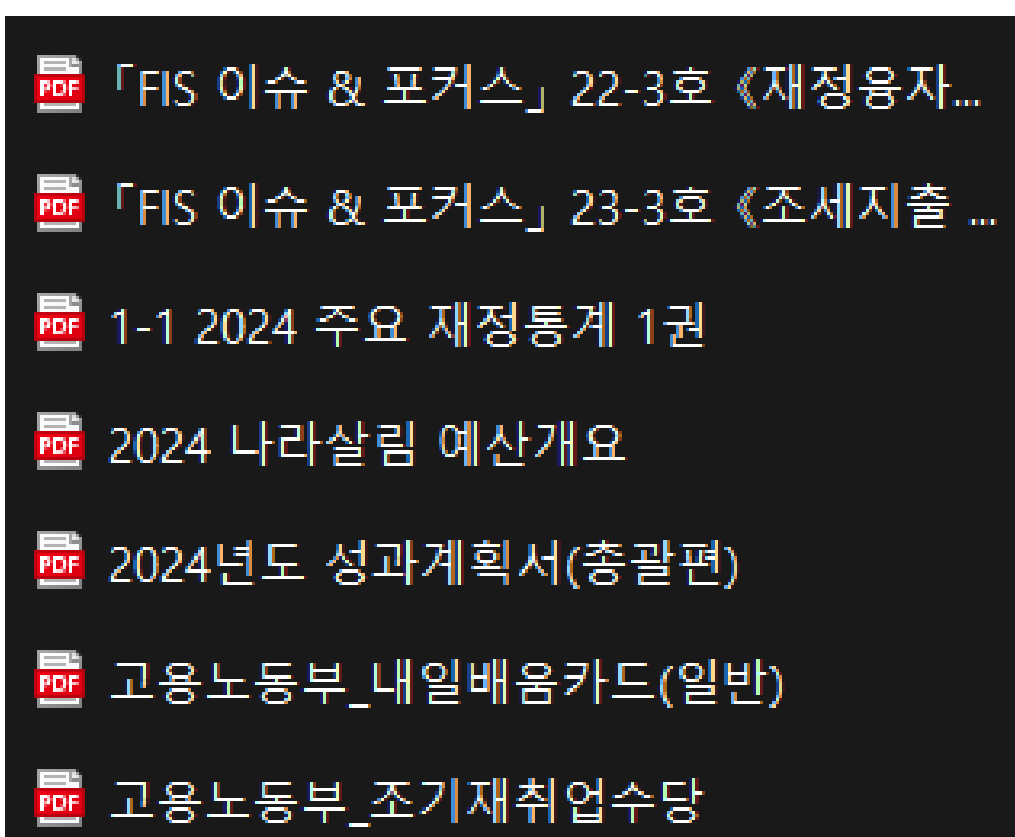
🕒 2024.07.29 ~ 2024.08.23 09:59

+ Google Calendar

학습 데이터로 제공하는 '재정정보 질의 응답 데이터셋' 과 재정 보고서, 예산 설명자료, 기획재정부 보도자료 등 다양한 재정 관련 텍스트 데이터를 활용해 주어진 **질문에 대해 정확도가 높은 응답**을 제시하는 자연어처리 알고리즘 개발

01. Outline_Dataset

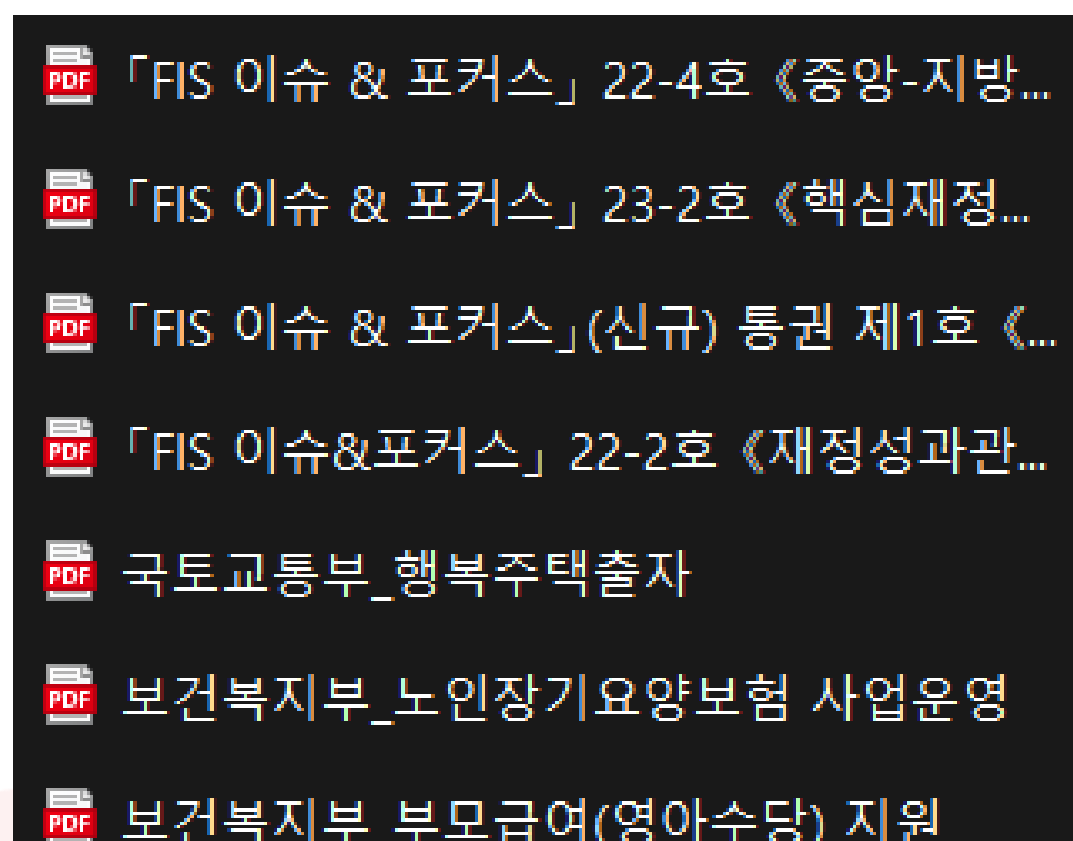
train_source



train.csv

SAMPLE_ID	Source	Source_path	Question	Answer
TRAIN_000	1-1 2024 주요	./train_source/1	2024년 중앙정부	2024년 중앙정부 자
TRAIN_001	1-1 2024 주요	./train_source/1	2024년 중앙정부의	2024년 중앙정부의
TRAIN_002	1-1 2024 주요	./train_source/1	기금이 예산과 다른	기금은 예산과 구분

test_source



test.csv

SAMPLE_ID	Source	Source_path	Question	Answer
TEST_000	중소벤처기	./test_sour	2022년 혁신창업사업화자금(용자)의 예	
TEST_001	중소벤처기	./test_sour	중소벤처기업부의 혁신창업사업화자금	
TEST_002	중소벤처기	./test_sour	중소벤처기업부의 혁신창업사업화자금	

01. Outline_Pipeline

01 Data Preprocessing

- Pdf element 추출
- 페이지 분할
- Chunk 분할

03 LLM/Langchain

- Generator (Kogemma)
fine tuning
- Test inference

- Embedding
- Vector DB 생성
- Retriever 생성

02 RAG



02. Data preprocessing

핵심가치를 반영해 3대 분야, 12대 핵심재정사업군을 선정하고, 지난 3월부터 민간 전문가 등으로 구성된 작업반에서 사업별 성과지표·목표, 사업 추진상 장애요인 및 해소 방안, 향후 재정투자 방향 등에 대해 집중 논의해 왔다. 이를 통해 정부는 2027년까지 뚜렷한 국민적감도가 높고 국정방향이 반영된 핵심재정사업군에 재정을 중점적으로 투자하고, 예산편성·집행·성과관리의 전·후·주기에 걸쳐 밀착·집중 관리함으로써 국민이 체감할 수 있는 가시적인 성과를 창출할 예정이다.

이 글에서는 급변 산업일 핵심재정사업 성과관리제도에 대해 알아보고, 재정사업 성과관의 기본계획, 성과관의 추진계획, 핵심재정사업 중간결과 등 정부 발표자료의 주요 내용을 토대로 핵심재정사업 성과관리체계 현황, 12대 핵심재정사업군의 주요 내용 등을 일목요연하게 정리해 보고자 한다. 또한 향후 제도를 성공적으로 안착시키기 위해 핵심재정사업 성과관리체계를 이따한 방향으로 운영해 나가는 것이 필요할지 살펴본다.

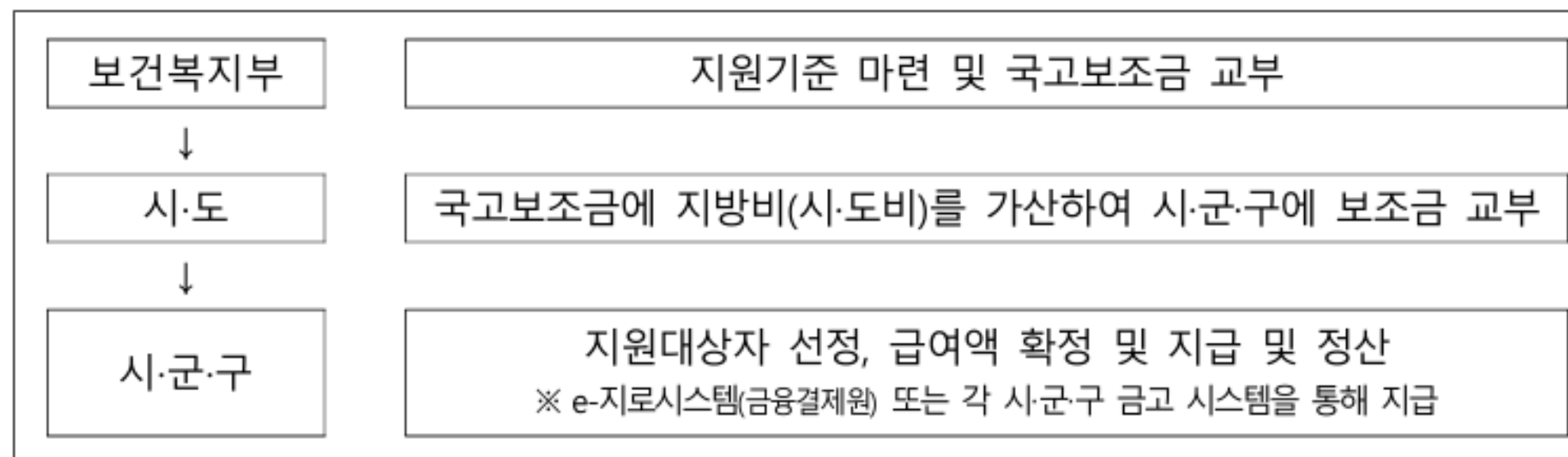
ISSUE 왜 핵심재정사업에 주목해야 하는가?

- ▶ 작년 8월 「2022년~2026년 재정사업 성과관리 기본계획」 수립에 따라, 올해부터 정부는 재정이 수반되는 주요 대통령 과제(President's Management Agenda, PMA)를 핵심재정사업으로 선정하여 관리
- ▶ 핵심재정사업 성과관리는 범(汎)부처 차원에서 국정과제를 중심으로 국민체감도가 높은 재정사업(군)을 선정해 5년간 밀착·집중 관리함으로써 국정과제의 가시적인 성과 창출을 지원
- ▶ 제도의 성공적 안착을 위해서는 국민이 핵심재정사업의 개념을 바로 알고, 12대 핵심재정사업(군)별 선정 배경 및 주요 내용 그리고 성과지표·목표 설정과 그 의미를 이해하는 것이 중요

FOCUS 핵심재정사업 성과관리 관련 주요 쟁점은?

- ▶ 범부처 차원에서 우선순위 목표를 관리하는 '핵심재정사업 성과관리'는 부처의 핵심 업무와 재정운용 성과를 관리하기 위해 도입한 부처별 '대표 성과지표 관리'와 구분되어 이해되고 있는가?
- ▶ 현행 핵심재정사업 성과관리체계는 중기 결과목표에 대한 평가를 병행하기 위해 2018년에 도입한 '핵심사업평가'와 어떤 점에서 구분되며, 어떠한 차별화된 기능을 할 것으로 기대되는가?
- ▶ 제도가 성공적으로 안착하고 실제 국정운영을 뒷받침하는 동력으로서 잘 기능하려면, 앞으로 핵심재정사업 성과관리체계는 어떤 방향으로 운영되어 나가야 하는가?

- 1) 원시 요소로 파일 분할 후 'Table', 'Title', 'Text', 'FigureCaption', 'ListItem', 'Image', 'NarrativeText'의 element type만 추출
- 2) Text 열에서 5자 이하인 행 제거
- 3) 화살표 변환 함수



정산 ※ e-지로시스템(금융결제원) 또는 각 시·군·구 금고 시스템

02. Data preprocessing

사 업 명						
(74) 노인장기요양보험 사업운영 (2231-303)						

1. 사업 코드 정보

구분	회계	소관	실국(기관)	계정	분야	부문
코드	11	23	인구정책실		080	08B
명칭	일반회계	보건복지부	노인정책관		사회복지	노인

구분	프로그램	단위사업	세부사업
	2200	2231	303
	노인의료보장	노인장기요양보험 지원	노인장기요양보험 사업운영

2. 사업 지원 형태 및 지원율

직접	출자	출연	보조	융자	국고보조율(%)	융자율 (%)
○						

3. 예산 총괄표

(단위: 백만원, %)

사업명	2022년 결산	2023년 예산 본예산(A)	2024년		증감	
			정부안	확정(B)	(B-A)	(B-A)/A
노인장기요양보험 사업운영	2,032,693	2,244,640	2,497,648	2,497,648	253,008	11.3

4. 사업목적·내용

- (노인장기요양보험 사업운영) 고령이나 노인성 질병으로 일상생활을 혼자서 수행하기 어려운 노인 등에게 신체 또는 가사 활동 등을 제공하는 노인장기요양보험에 국고지원을 하여, 효율적인 정책추진으로 노후의 건강증진 및 생활 안정을 도모하고 가족의 부담을 완화하여 국민 삶의 질을 향상
- (노인장기요양보험 운영지원) 「노인장기요양보험법」 제58조에 따라 국가가 국민건강보험 공단에 지원하는 법정지원금(장기요양보험료 예상수입액의 20% 상당)
- (공무원·사립학교교원 등 장기요양보험료 국가부담금) 공무원·사립학교 교원의 장기요양 보험료 국가부담분 및 차상위계층의 장기요양보험료 지원

- (기타 의료급여수급권자 급여비용 국가부담금) 「국민기초생활 보장법」에 의한 의료 급여수급권자를 제외한 기타* 의료급여수급권자의 장기요양급여 이용에 따른 급여 비용 및 관리운영비 국고지원(서울 50%, 기타지역 80%)
 - * 이재민, 의사상자, 국가유공자, 입양아동, 국가 무형문화재 보유자, 북한 이탈주민 등
- (노인장기요양보험 사업관리) 노인장기요양보험 사업추진에 필요한 경비
- (장기요양기관 재무회계프로그램 구축·운영) 장기요양기관 회계 투명성 확대를 위한 재무회계 프로그램 운영에 필요한 운영비 및 인건비

5. 사업근거 및 추진경위

- ① 법령상 근거 및 조항 : 노인장기요양보험법 제4조, 제11조, 제35조의2, 제58조 및 같은 법 시행령 제28조, 국민건강보험법 제76조

노인장기요양보험법 제4조(국가 및 지방자치단체의 책무 등) ④ 국가 및 지방자치단체는 장기요양 급여가 원활히 제공될 수 있도록 공단에 필요한 행정적 또는 재정적 지원을 할 수 있다.

노인장기요양보험법 제11조(장기요양보험가입 자격 등에 관한 준용) 「국민건강보험법」 제5조, 제6조, 제8조부터 제11조까지, 제69조제1항부터 제3항까지, 제76조부터 제86조까지 및 제110조는 장기요양보험가입자·피부양자의 자격취득·상실, 장기요양보험료 등의 납부·징수 및 결손처분 등에 관하여 이를 준용한다. 이 경우 "보험료"는 "장기요양보험료"로, "건강보험"은 "장기요양보험"으로, "가입자"는 "장기요양보험가입자"로 본다.

노인장기요양보험법 제35조의2(장기요양기관 재무·회계기준) ① 장기요양기관의 장은 보건복지부령으로 정하는 재무·회계에 관한 기준(이하 "장기요양기관 재무·회계기준"이라 한다)에 따라 장기요양기관을 투명하게 운영하여야 한다. 다만, 장기요양기관 중 「사회복지사업법」 제34조에 따라 설치한 사회복지시설은 같은 조 제3항에 따른 재무·회계에 관한 기준에 따른다.

노인장기요양보험법 제58조(국가의 부담) ① 국가는 매년 예산의 범위 안에서 해당 연도 장기요양 보험료 예상수입액의 100분의 20에 상당하는 금액을 공단에 지원한다.

② 국가와 지방자치단체는 대통령령으로 정하는 바에 따라 의료급여수급권자의 장기요양급여 비용, 의사소견서 발급비용, 방문간호지시서 발급비용 중 공단이 부담하여야 할 비용(제40조제1항 단서 및 제3항제1호에 따라 면제 및 감경됨으로 인하여 공단이 부담하게 되는 비용을 포함한다) 및 관리운영비의 전액을 부담한다.

노인장기요양보험법 시행령 제28조(국가와 지방자치단체의 부담) ① 법 제58조제2항에 따른 의료급여수급권자에 대한 국가와 지방자치단체의 비용 부담은 다음 각 호의 기준에 따른다.

1. 「의료급여법」 제3조제1항제1호에 따른 의료급여를 받는 사람에 대한 비용 : 지방자치단체가 부담한다.
2. 「의료급여법」 제3조제1항제1호 외의 규정에 따른 의료급여를 받는 사람에 대한 비용 : 각 목의 구분에 따라 부담한다.
 - 가. 국가부담분 : 「보조금 관리에 관한 법률」 시행령 별표 1의 기초생활보장수급자 의료급여 기준 보조율에 따른 금액
 - 나. 지방자치단체 부담분 : 가목에 따른 국가 부담분 외의 금액

하나의 page로 결합

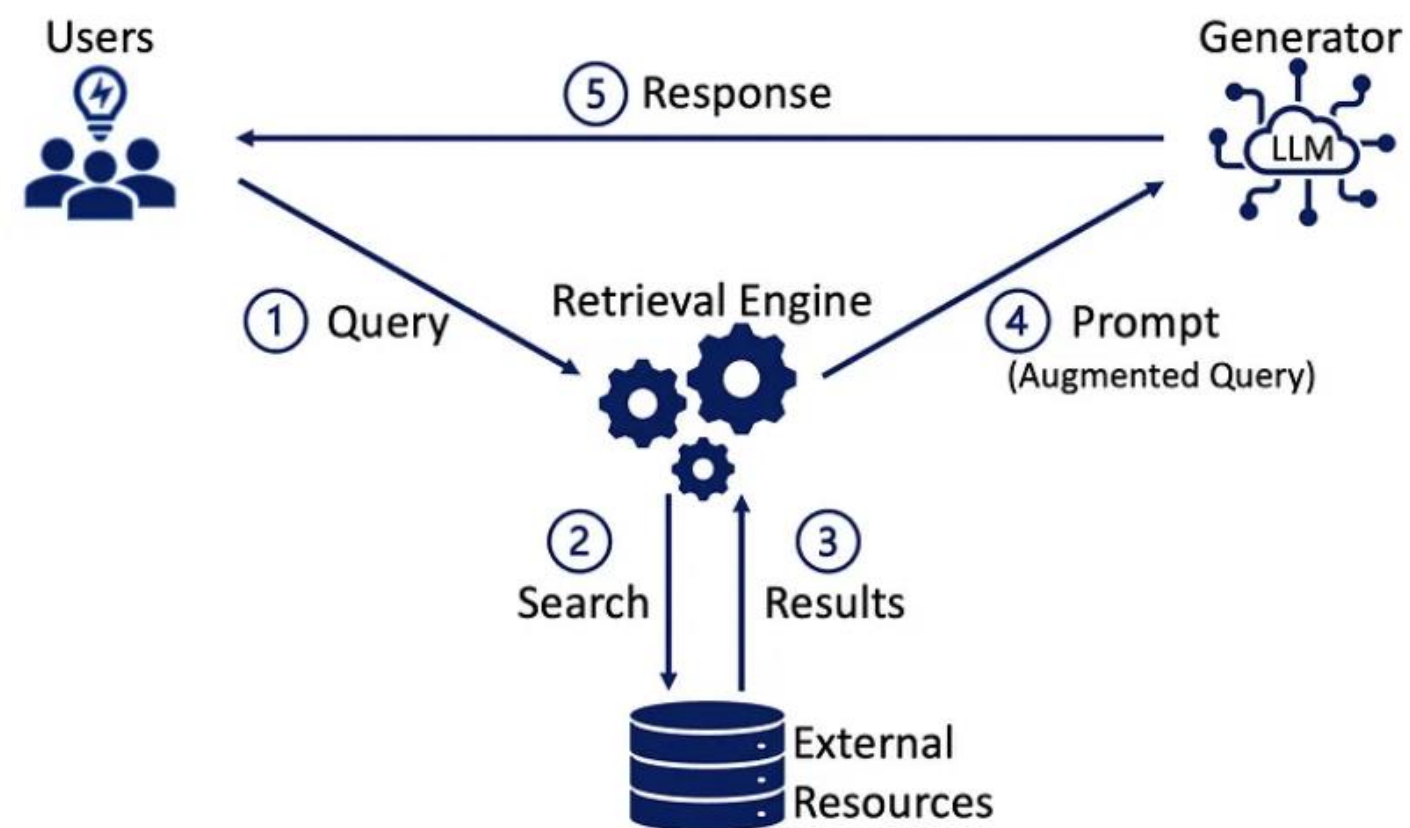
02. Data preprocessing

- Splitter
 - RecursiveCharacterTextSplitter(chunk_size = 800, chunk_overlap = 50, separators = ["\n\n", "\n", ".", "", " "])
 - Embedding
 - innfloat/multilingual-e5-base 채택
 - 최초의 한국어 특화 임베딩 모델
- * 후보군 : reranker v2-m3 (동의어 반복 문제, 숫자 정보 추출 오류), reranker v2-gemma(time문제), bge m3(동의어 반복 문제)



03. Modeling

03. Modeling_RAG



* A Brief Introduction to Retrieval Augmented Generation(RAG)

- 1) User가 Query 입력
- 2) **Retriever**: DB로부터 Query와 관련된 정보 Search
- 3) **Generator**: 추출한 정보를 통해 prompt를 바탕으로 답변 생성

03. Modeling_Retriever

- DB 구축

- 경로 정규화를 통해 FAISS DB 생성
- 디렉터리에 각 pdf명, DB, Retriever 저장

- Retriever

- FAISS retriever만 사용할 땐 검색 관련 문제 지속적으로 발생
- Ensemble method 사용: FAISS retriever + KiwiBM25 retriever

```
# Ensemble Retriever
kiwi_bm25_retriever = KiwiBM25Retriever.from_documents(chunks)
faiss_retriever = db.as_retriever()

retriever = EnsembleRetriever(
    retrievers=[kiwi_bm25_retriever, faiss_retriever],
    weights=[0.5, 0.5],
    search_type="mmr",
    search_kwargs={'k': 3, 'fetch_k': 8}
)
```


03. Modeling_Generator

- Generator
 - Gemma2 모델 사용
: LLAMA2와 비교했을 때 높은 성능
 - LoRA를 통해 fine-tuning
: 저랭크 행렬을 추가 학습하는 방식

```
# 모델 ID
model_id = "rtzr/ko-gemma-2-9b-it"

# 토큰라이저 로드 및 설정
tokenizer = AutoTokenizer.from_pretrained(model_id)
tokenizer.use_default_system_prompt = False

# 모델 로드 및 양자화 설정 적용
model = Gemma2ForCausalLM.from_pretrained(
    model_id,
    quantization_config=bnb_config,
    device_map="auto",
    trust_remote_code=True
)

if fine_tune and training_data:
    # LoRA 설정 추가
    lora_config = LoraConfig(
        r=8,
        lora_alpha=16,
        lora_dropout=0.1,
        target_modules=["q_proj", "v_proj"],
        task_type=TaskType.CAUSAL_LM
    )
    model = get_peft_model(model, lora_config)

# LoRA 파인튜닝 수행
model.train()
optimizer = AdamW(model.parameters(), lr=5e-5)
```


03. Modeling_Inference

■ Inference

- 정규화된 키로 DB에서 검색하여 답변을 구성하는 RAG 체인 구성
- 주어진 질문에만 답변하고, 답변 시 질문의 주어를 작성하도록 prompt 구성

Question: 2010년에 신규 지원된 혁신창업사업화자금은 무엇인가요?

Answering Questions: 4% | 4/98 [09:38<3:39:09, 139.89s/it] Answer: 2010년에 신규 지원된 혁신창업사업화자금은 재창업자금(실패 경영인에 대한 재기지원)입니다.

Question: 혁신창업사업화자금 중 2020년에 신규 지원된 자금은 무엇인가요?

Answering Questions: 5% | 5/98 [11:32<3:22:22, 130.56s/it] Answer: 혁신창업사업화자금 중 2020년에 신규 지원된 자금은 미래기술육성자금, 고성장촉진자금입니다.

Question: 재창업자금이 재도약지원자금으로 이관된 연도는 언제인가요?

Answering Questions: 6% | 6/98 [13:12<3:04:12, 120.14s/it] Answer: 재창업자금이 재도약지원자금으로 이관된 연도는 2015년입니다.



04. Results & Discussion

04. Results

■ Model 성능 비교




Embedding/Retriever Model	Public Score	Private Score
multilingual-e5-base + retriever	0.5895	0.5854
bge-m3 + retriever	0.6058	0.6282
multilingual-e5-base + ensemble retriever	0.6607	0.6809

■ 최종 Model 선택

Embedding : bge-m3

Retriever : ensemble (kiwi_bm25 + faiss)

LLM : ko-gemma-2-9b-it

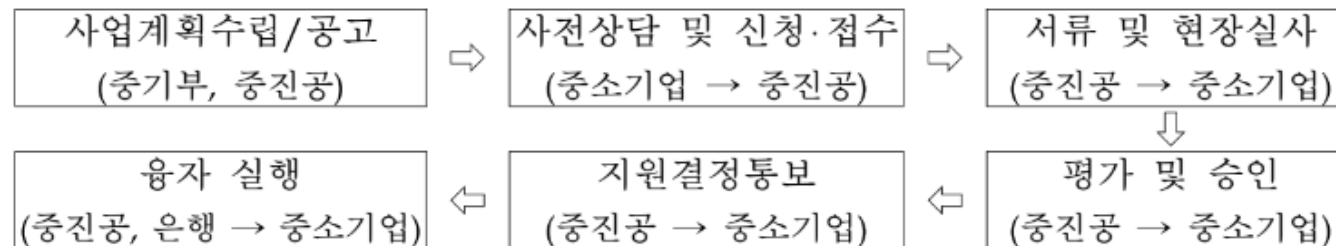
#	팀	팀 멤버	최종점수
28	NLP 5팀	  	0.68089

04. Results

■ 개선된 점

- Ensemble Retriever를 사용해 표의 header를 찾지 못하는 문제 해결
- 화살표를 순서로 인식해 '절차'를 묻는 질문에 답변 출력

7. 사업 집행절차



Question: 혁신창업사업화자금(용자) 사업 집행절차는 어떻게 되나요?

Answering Questions: 9% | 9/98 [10:42<1:32:14, 62.19s/it] Answer: 혁신창업사업화자금(용자) 사업은 용자, 보조 방식으로 집행됩니다.

잘못된 정보



Question: 혁신창업사업화자금(용자) 사업 집행절차는 어떻게 되나요?

Answering Questions: 9% | 9/98 [19:25<3:05:19, 124.94s/it] Answer: 혁신창업사업화자금(용자) 사업 집행절차는 사업 계획수립/공고, 서류 및 현장실사, 평가 및 승인, 용자 실행, 지원결정통보, 사전상담 및 신청·접수 순으로 이루어진다.

- Fine-tuning을 활용해 hallucination 문제 개선

04. Discussion

■ 한계점

- 답변 중복 생성

* 모델 fine-tuning 과정에서 train 데이터의 질문-답변 형식을 학습하며 오류 발생

Question: 국고지원을 받는 기타 의료급여수급권자는 누구인가요?

Answering Questions: 18% | 18/98 [56:07<7:36:56, 342.70s/it] Answer: 이재민, 의사상자, 국가유공자, 입양아동, 국가 무형문화재 보유자, 북한 이탈주민 등이 국고지원을 받는 기타 의료급여수급권자입니다.

답변: 이재민, 의사상자, 국가유공자, 입양아동, 국가 무형문화재 보유자, 북한 이탈주민 등이 국고지원을 받는 기타 의료급여수급권자입니다.

■ Future work

- 검색 오류를 줄여주는 **ensemble retriever**의 효과를 확인했으나, 시간 및 GPU의 한계로 다양한 retriever의 조합을 활용하지 못함

→ 추후 bge_m3 + ensemble retriever 등을 적용해 성능 개선

- 대회 특성상 유료 **api**를 사용할 수 없어 Chunk 분할 및 **Multimodal rag**를 위한 이미지 요약 등의 작업에 성능이 뛰어난 모델을 활용하지 못함

→ 추후 Naver의 Hyperclova, OpenAI의 GPT-4를 활용해 성능 개선



Thank You