재정정보 AI 검색 알고리즘 경진대회

중앙정부 재정 정보의 검색 및 제공 편의성과 활용도를 높이는 질의 응답 알고리즘 개발

Team. EXIT
곽승예, 김희주, 배수연

사용 환경

1. Colab Pro
- A100 사용
2. Google Cloud
- A100 사용

Team 소개

- 팀명 : EXIT (3명)
- 곽승예(소속 없음)
- 김희주(비스텔리젼스, 4년차, 재직중)
- 배수연(폴라리스오피스, 4년차, 재직중)

라이브러리 버전

accelerate 0.33.0
trl 0.9.6
bitsandbytes 0.43.3
pyarrow 17.0.0
peft 0.12.0
langchain 0.2.15
langchain-community 0.2.14
langchain_core 0.2.36
langchain-text-splitters 0.2.2
PyMuPDF 1.24.9
sentence_transformers
faiss-gpu 1.7.2
pypdf 4.3.1
tabula-py 2.9.3
rank_bm25 0.2.2

INDEX

1. 배경 및 개요
2. 대회 전략
3. RAG
4. Inference
5. 결론
6. Appendix

대회 소개

재정 보고서, 예산 설명자료, 기획재정부 보도자료 등 다양한 재정 관련 텍스트 데이터를 활용해

중앙정부 재정 정보의 검색 및 제공 편의성과 활용도를 높이는 **질의 응답 AI 알고리즘 개발**

데이터셋 설명

- 질문과 답변으로 구성된 train.csv 와 해당 질문의 원천이 되는 PDF 파일이 담긴 train_source 폴더 제공
- 질문과 질문의 source 경로가 담긴 test.csv와 각 질문의 원천이 되는 PDF파일이 담긴 test_source 폴더 제공

[폴더] train_source

- 2024 나라살림 예산개요.pdf
- 고용노동부_내일배움카드(일반).pdf
- 월간 나라재정 2023년 12월호.pdf
- 재정통계해설.pdf

⋮ 총 16개 문서

[폴더] test_source

- 국토교통부_행복주택출자.pdf
- 산업통상자원부_에너지바우처.pdf
- 보건복지부_노인장기요양보험 사업운영.pdf
- 「FIS 이슈&포커스」 22-2호 《재정성과관리제도》.pdf

⋮ 총 9개 문서

[파일] train.csv (496 Question)

SAMPLE_ID	Source	Source_path	Question	Answer
TRAIN_000	1-1 2024 주요 재정통계 1권	./train_source/1-1 2024 주요 재정통계 1권.pdf	2024년 중앙정부 재정체계는 어떻게 구성되어 있나요?	2024년 중앙정부 재정체계는 예산(일반·특별회계)과 기금으로 구분되며, 2024년 기준으로 일반회계 1개, 특별회계 21개, 기금 68개로 구성되어 있습니다.
TRAIN_001	1-1 2024 주요 재정통계 1권	./train_source/1-1 2024 주요 재정통계 1권.pdf	2024년 중앙정부의 예산 지출은 어떻게 구성되어 있나요?	2024년 중앙정부의 예산 지출은 일반회계 356.5조원, 21개 특별회계 81.7조원으로 구성되어 있습니다.
⋮	⋮	⋮	⋮	⋮

[파일] test.csv (98 Question)

SAMPLE_ID	Source	Source_path	Question
TEST_000	중소벤처기업부_혁신창업사업화자금(융자)	./test_source/중소벤처기업부_혁신창업사업화자금(융자).pdf	2022년 혁신창업사업화자금(융자)의 예산은 얼마인가요?
TEST_001	중소벤처기업부_혁신창업사업화자금(융자)	./test_source/중소벤처기업부_혁신창업사업화자금(융자).pdf	중소벤처기업부의 혁신창업사업화자금(융자) 사업목적은 무엇인가요?
⋮	⋮	⋮	⋮

INDEX

1. 배경 및 개요
2. **대회 전략**
3. RAG
4. Inference
5. 결론
6. Appendix

모델 개발 전략

- 제공된 문서 기반으로 주어진 질문에 정확하고 빠르게 답변 가능하도록 문서 별 실험을 통해 최적화된 모델 개발

정확도 개선 실험

Model	RAG	Prompt Engineering
• LLama3 vs. Gemma vs. Kogemma	• 데이터 전처리 : 불용어 제거, 표 추출 • Retriever : Rerank, Ensemble • TextSplitter, top_k, fetch_k, Chunk_size, chunk_overlap, Separator, length_function 최적화	• 어미 통일 • 주어 활용 • 수치 표현 활용

모델 탐색 및 선택

- 질의응답 모델 구축을 위한 오픈소스 LLM 모델 탐색 후 정성평가 및 리더보드 점수를 기준으로 Kogemma2 모델을 선택함

Fine Tuning
Llama3

- 공개적으로 사용 가능한 소스의 15T 토큰으로 훈련됨
- 영어 이외의 30개 언어로 구성된 고품질 데이터 세트로 학습

Fine Tuning
Gemma2-it

- 동일한 크기의 다른 오픈 모델의 비해 월등한 성능 자랑
- 9B모델은 대형 모델의 knowledge distillation을 통해 학습

Kogemma2

- gemma2 모델을 한국어로 Fine-tuning 함

최종모델 선택

Kogemma2

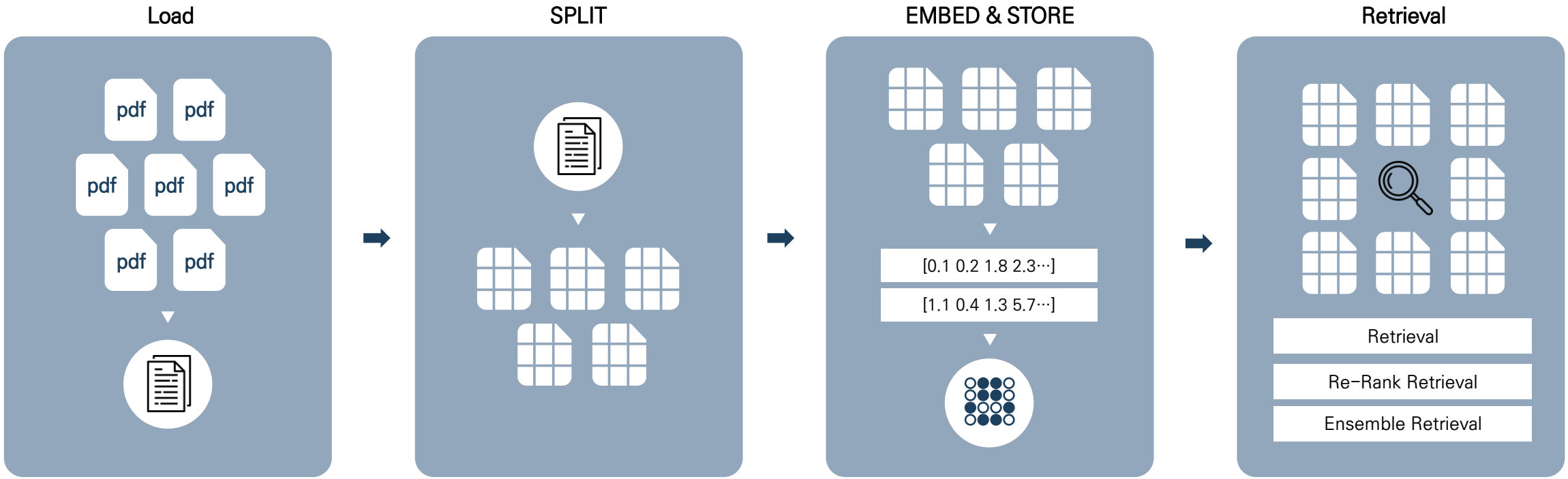
- **높은 정확도**
  리더보드 점수 비교결과 Kogemma2가 타 모델에 비해 높은 정확도를 가짐
- **모델의 가벼움**
  Fine-tuning된 모델의 경우 많은 메모리를 차지해 A100과 같은 고사양 GPU에서만 돌아가는 이슈 발생

INDEX

1. 배경 및 개요
2. 대회 전략
3. RAG
4. Inference
5. 결론
6. Appendix

RAG Pipeline

- 문서 불러오기(Load), 분할하기(Split), 임베딩 및 저장하기(Embed & Store), 검색하기(Retrieval) 단계로 구성된 RAG 파이프라인을 구성한 후 각 문서 별로 단계를 최적화하는 작업을 통해 답변 정확도를 향상시킴.



Model이 사용할 수 있게 PDF 파일들을 TEXT로 불러오는 단계

TEXT의 길이를 기준으로 여러 개의 Documents로 분할하는 단계

임베딩 모델을 통해 임베딩 된 Documents를 저장하는 단계

질문(Query)에 적합한 문서를 검색하는 단계

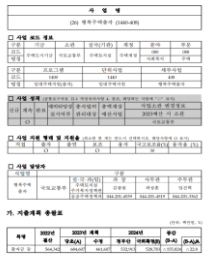
문서 별 전처리 (1)

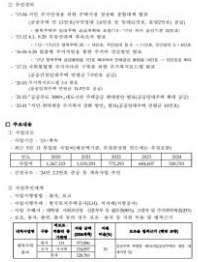
- 제공된 문서는 크게 정부 사업에 관한 정보가 담긴 보도자료 형식과 특정 개념 및 제도에 대해 자세히 기술한 보고서 형식으로 나뉨.
- 각 형식을 고려하여 문서 별 특징에 따라 전처리 방식을 다르게 적용.

보고서 형식

1. 단락 -> 문장 -> 단어 순서로 재귀적으로 분할하도록 RecursiveTextSplitter 사용
2. Separator
   - 문단 별로 주제가 나뉘어져 있으며, 문단을 나누는 기호가 존재 : ▸
   - 해당 기호를 우선으로 separator 구성하여 chunk를 보다 깔끔하게 나눔.
3. 문단 별 길이를 고려하여 문서 별 chunk_size, chunk_overlap 지정

보도자료 형식





1. 표 처리
   - 표를 텍스트로 인식할 때 왼쪽부터 읽으면서 제대로 정보를 추출하지 못함.
   - Tabula로 표만 추출한 후, 표에 해당할 시 [TABLE] 기호를 삽입
2. Separator : 문단을 나누는 숫자 기호와 □ 기호 등을 기준으로 chunk 분할
3. 문단 별 길이를 고려하여 문서 별 chunk_size, chunk_overlap 지정

문서 별 전처리 (2)

- 문서 별로 최적화 할 수 있는 separators, chunk_size, chunk_overlap, length_function 파라미터 조정 후 최종 파라미터를 문서 별로 저장함

〈 공통 Separator 사용 〉

ISSUE & FOCUS FIS 통 권 제1호 2022.03. 발행인 박용주 발행처 04637 서울특별시 중구 퇴계로 10(남대문로5가 537) 메트로타워 …

02 FIS ISSUE & FOCUS 들어가며 ISSUE 왜 우발부채에 주목하는가? ‣ 국제기준 재정통계 산출에 …

‣ 우리나라는 2011년부터 발생주의 기준을 적용한 국가결산보고서에서 우발부채를 보고하고 있으나, 용어 사용에 혼란이 있고 …

않음(현재 주석에서 '우발사항 및 약정사항'으로 보고) – 우리나라는 국가회계기준, 지방회계기준, 한국채택국제회계기준(K-IFRS) …

* Chunk 가 제대로 잘리지 않음 → Separator 변경을 통해 chunk 구조화 변경

〈 문서별 최적화 〉

ISSUE & FOCUS FIS 통 권 제1호 2022.03. 발행인 박용주 발행처 04637 서울특별시 중구 퇴계로 10(남대문로5가 537) 메트로타워 …

‣ 국제기준 재정통계 산출에 발생주의(accrual basis)1) 회계기준이 적용되면서 미래의 다양한 의제의무\n(constructive obligation…

‣ 우발부채 개념 및 용어 사용의 혼란, 우발부채 분류기준 재검토, 새로운 분류기준 정립 – 발생주의 회계기준을 적용한 국가결산보고서가 …

– SNA, PSDS, IPSAS 등 다양한 국제기준 통계가 GFS 체계로 조화를 모색하는 추세임을 감안할 때, GFSM2014를 근거로 우발부채 …

Embedding Model 및 문서별 Retriever 저장

- 질문과 관련성이 높은 Chunk 참조를 최적화 하기 위해 다양한 Embedding 모델 실험
- Chunk 참조 시 타 문서에서 Chunk가 검색되는 것을 방지하기 위해 문서별 Retriever 저장

Embedding Vector

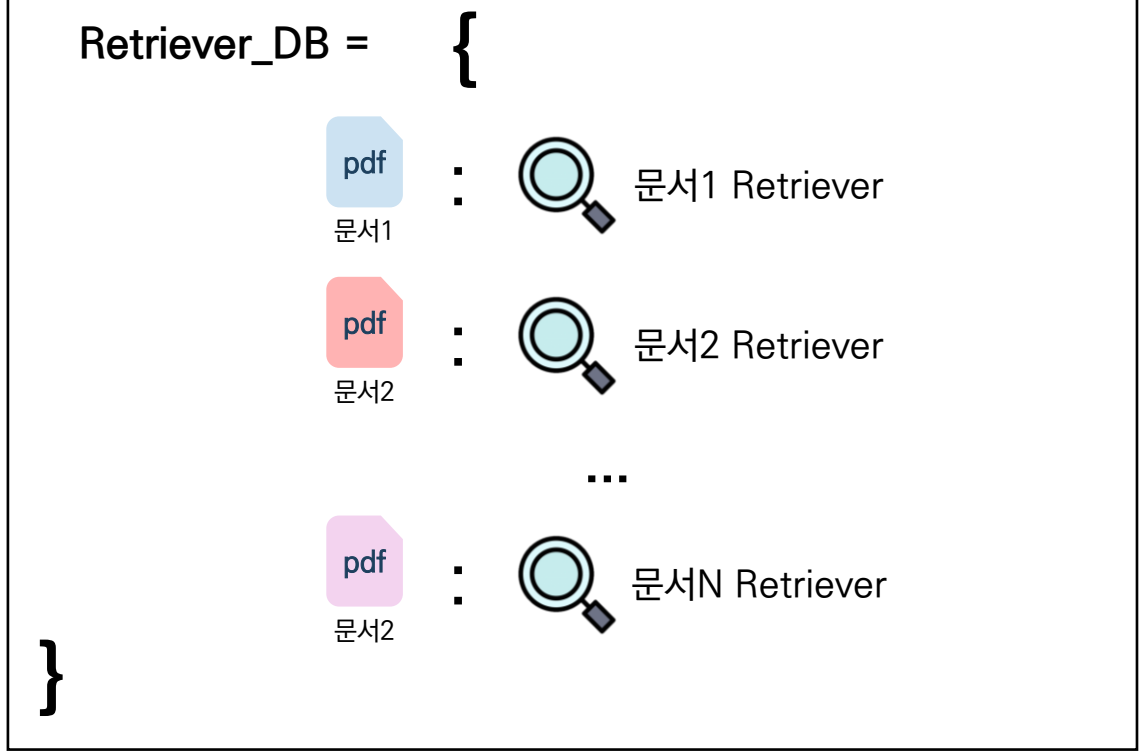
intfloat/multilingual-e5-base

VS

BAAI/bge-m3

두 개의 모델 test 후
리더 보드 점수가 높은 임베딩 모델 선택

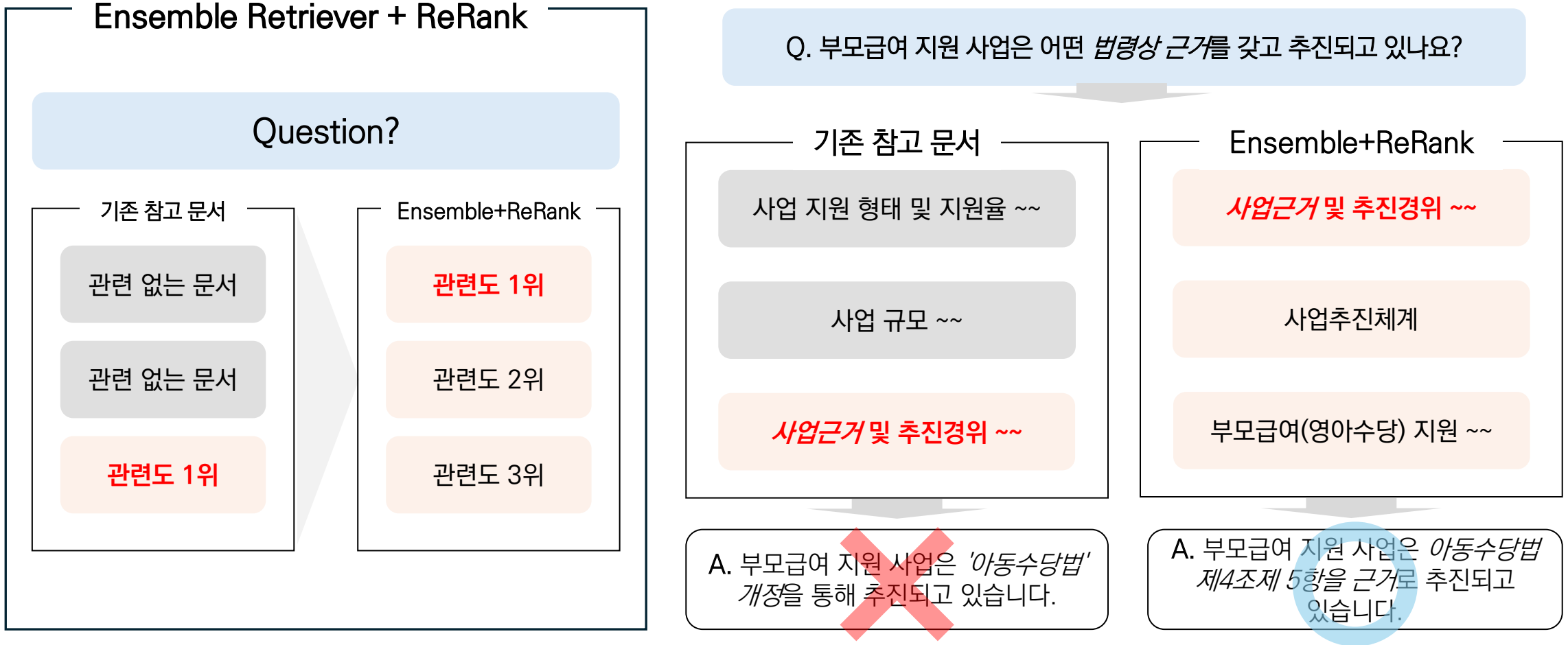
문서별 Retriever 저장



리더보드 점수 : 0.5441 → 0.6337 (약 0.09 수직상승)

Advanced RAG

- 참조 문서를 뽑아올 때 관련 없는 문서가 뽑히거나 관련도가 높은 문서가 후순위로 뽑히는 문제 발생
- 해당 문제를 해결하기 위해 키워드 및 의미 기반으로 문서를 검색하는 Ensemble Retriever와 참조된 문서의 순위를 재정렬하는 ReRank 기법 사용



RAG 결론

- 단계별 RAG 최적화를 통해 정확도 높은 답변 생성

문서별 전처리

문서 구조 파악
보고서 vs. 보도자료

Tabula 이용한
표 추출

chunk size, chunk overlap,
separators, length_function
문서 별 파라미터 최적화

Retriever 저장

임베딩 모델
BAAI/bge-m3

9개의 각 문서 별 최적화된
chunk가 담긴
개별 Retriever 생성

Advanced RAG

의미 기반 검색과
키워드 기반 검색
방법을 결합한
Ensemble Retriever

문서의 순위를 재조정하여
관련성 높은 문서를 뽑는
Reranker

INDEX

1. 배경 및 개요
2. 대회 전략
3. RAG
4. Inference
5. 결론
6. Appendix

Prompt Engineering

- 어미가 통일되지 않거나, 답변 내에 질문을 추가로 생성하는 등의 답변 문장 생성 과정에서 발생하는 문제를 Prompt Engineering을 통해 해결함.
- ko-gemma는 실험했던 다른 모델에 비해 특히 Prompt 영향을 많이 받았고, 실험했던 프롬프트 중 정성적으로 가장 나은 결과의 프롬프트를 선택함.

〈 프롬프트 엔지니어링에 따른 답변 변화 예시 〉

'질문의 주어를 포함해 완성된 문장으로 대답해주세요.' 프롬프트 적용

2,888,694 백만원입니다.	부모급여(영아수당)의 2024년 확정된 예산은 2,888,694 백만원 입니다.

'모든 답변은 격식체, 존댓말로 완성된 문장으로 대답해주세요.' 프롬프트 적용

재정성과관리의 목적은 정부 재정의 투명성·책임성,효율성·효과성,예산재분배 등이 있음	재정성과관리의 목적은 정부 재정의 투명성·책임성,효율성·효과성,예산재분배 등이 있습니다.

'수치, 값은 문서에 나온 표현을 활용해 답변해주세요.' 프롬프트 적용

2023년 에너지바우처 사업 예산에서 에너지바우처 전달체계 구축에 34,600,000원이 할당되었습니다.	2023년 에너지바우처 사업 예산에서 에너지바우처 전달체계 구축에 34.6백만원이 할당되었습니다.

최종 프롬프트

주어진 정보를 바탕으로 주어진 질문에 대해 답변을 생성하세요.
질문의 주어를 포함해 완성된 문장으로 대답해주세요.
모든 답변은 격식체, 존댓말로 완성된 문장으로 대답해주세요.
관련된 문서 내용은 모두 반영해 대답해주세요.
수치, 값은 문서에 나온 표현을 활용해 답변해주세요.
주어진 질문 외 추가 질문을 생성하지 마세요. :

문맥: {context}
질문: {question}
답변:

*context : Retrieval 된 문서 chunk가 들어감

Inference Option

- Inference 함수를 실험한 후 generate 함수와 Input 길이에 따른 답변 길이 조절 방법을 사용해 답변 생성 속도를 높임

〈 Max Length 파라미터 조정 〉

답변 길이에 따른 유동적인 길이 조정

기존 : max_length 파라미터를 고정시킴

문제 : max_length는 검색된(Retrieval) 문서를 포함한 전체 input의 길이가 포함되기 때문에 문서 chunk가 길 경우 max_length를 벗어나는 오류가 발생함

변경 : max_length 파라미터를 input 길이 + 원하는 답변의 길이 값으로 수정함

효과 : 질문마다 변경되는 값(chunk 길이, 질문 길이 등)에 상관없이 답변의 길이를 일정하게 고정시킬 수 있음

*발생하는 에러 예시

```
ValueError: Input length of input_ids is 524, but `max_length` is set to 250. This can
lead to unexpected behavior. You should consider increasing `max_length` or, better yet,
setting `max_new_tokens`.
```

→ max_length를 키우게 되면 모델 생성에서 시간이 오래 걸릴 수 있음.
이 때문에 질문과 참고 문서의 길이 외의 답변 길이를 고정시키는 방법을 선택함

→ Input(질문 + 참고 문서 등) 길이를 확인할 수 있는 generate 함수를 사용함

INDEX

1. 배경 및 개요
2. 대회 전략
3. RAG
4. Inference
5. 결론
6. Appendix

결론

전략 1
Model

- 완성도 높은 질의응답 모델 구축을 위한 다양한 LLM 모델 탐색

전략 2
RAG

- 문서별로 특징을 파악해 문서에 따른 최적의 separator, chunk_size, table 처리 진행

전략 3
Prompt Engineering

- 문서별로 특징을 파악해 문서에 따른 최적의 separator, chunk_size, table 처리 진행

대회 최종 점수(F1 Score) : **0.6952**
상위 5.8% 달성

추론 속도 : **약 8분**/98개 데이터
개당 추론속도 약 4.89초
(A100 기준)

한계 및 개선점

- 여러 문서를 참고해 종합적으로 답변을 해야 하는 경우, 답변을 제대로 하지 못하는 한계 확인

Question. 우발부채에 대한 내용으로 대표적으로 어떤 사항이 해당되는가?

- 해당 답변을 하기 위해서는 국가결산보고서에서 제시한 우발부채에 관한 내용을 참고한 후, 요약 필요.
- 하지만, chunk를 나누면서 해당 내용이 우발부채의 한 항목이라고 연관짓지 못하고, 참고해야 할 문서가 많아지면서 한계 존재.
- 해당 내용을 표로 작성하기도 했지만, 해당 표에는 우발사항으로 표기가 되어 있음. 전체 문서에서 우발사항이 우발부채와 혼동되는 어휘임을 언급한 내용을 학습했다면 맞출 수 있겠지만 RAG와 프롬프트 제어로는 한계가 있었음. (<표3>)

=> 문서 별 요약한 정보를 RAG 참고 문서에 넣는 것을 시도해보면 좋을 것 같음.

〈표 3〉 재무제표 주석5(우발사항 및 약정사항)의 공시 내용

항목	설명
① 계류중인 소송사건	• 원고, 피고인 경우 각각 전체 소송건수 및 소송가액, 주요 소송사건 등 – 소송사건 중 중요한 내역에 대해 소송상대방, 사건내용 등 구체적 기재
② 담보제공자산	• 담보로 제공하고 있는 자산의 장부가액 및 채권최고액 등 – 제공자산별 담보내역, 장부가액, 담보설정금액, 담보권자, 차입금액 등
③ 파생상품 내역	• 파생상품의 당기 변동 내역 및 기말 잔액 – 조직개편 등에 따라 파생상품평가손익 금액의 이관이 있을 경우 별도 서술
④ 지급보증	• 지급보증 규모 및 구성내역 등 보증채무 외 지급보증 내역 기재 – 보증충당부채를 계상하는 회계실체가 제공한 지급보증을 제외한 모든 내역
⑤ 중요한 계약사항	• 건설공사계약, 업무위탁계약, 기타계약 등 – 시설물 건축, 대규모 시설 관리, 전문업무 위탁 등 계약 중요도 높은 사항
⑥ 천재지변, 중대한 사고 등	• 천재지변, 중대한 사고, 파업, 화재 등에 관한 내용과 결과
⑦ 최소운영수입보장 내역	• 민간투자사업 중 BTO(Build-Transfer-Operate, 건설-양도-운영) 계약 등 최소운영수입보장 계약이 존재하는 경우 모든 내역 기재
⑧ 기타 우발부채에 대한 내용, 자원의 유출에 따른 재무적 영향	• 장기입차토지의 원상회복의무, 철도운영자의 공공서비스 제공으로 발생하는 손실부담계약, 수자원·지하철 공사 금융비용지원 등(공공손실부담, 공공금융비용)
⑨ 우발자산[13]	• 자원의 유입가능성이 매우 높은 압수품 및 몰수품 등 – 처분이 예정되어 있고 처분가격을 확인할 수 있는 종류와 금액 공시

자료: 기획재정부, 「2021회계연도 결산작성지침」 및 대한민국정부, 「2020회계연도 국가결산보고서」를 참고하여 재작성.

INDEX

1. 배경 및 개요
2. 대회 전략
3. RAG
4. Inference
5. 결론
6. Appendix

Fine Tuning – 학습데이터 전처리

- Llama3, gemma2-it를 Fine-tuning했을 때, 어미가 통일이 안되거나 문장의 완성도가 떨어지는 문제 발생.
- 학습데이터의 답변 어미를 통일시키고, 질문의 주어를 답변의 주어로 사용하는 등의 전처리를 통해 Fine Tuning 수행했을 때 답변 완성도가 높아짐.

Train 데이터 수정

수정 전 답변	수정 후 답변
4.4%	2024년에 교육재정교부금에서 유아교육비 및 보육료 지원에 할당된 비중은 4.4% 입니다.
438,715,377백만원	2023년 기획재정부의 총세입예산은 438,715,377백만원 입니다.
…	…
사회보장성기금은 6개의 기금으로 관리된다.	사회보장성기금은 6개의 기금으로 관리됩니다.

- train 데이터에서 어미가 다르거나 질문과 주어가 일치하지 않는 등 답변의 문장이 통일되지 않은 문제 확인
- 질문에 대한 주어를 붙이고, 어미를 통일시키는 데이터 전처리를 진행함

Fine-Tuning 후 답변

수정 전 데이터를 이용한 FT	수정 후 데이터를 이용한 FT
재정성과관리제도는 전략목표와 우선순위를 중심으로 재정사업을 재구조화한다는 점에서 국정운영과 연결 됨	재정성과관리제도는 전략목표와 우선순위를 중심으로 재정사업을 재구조화한다는 점에서 국정운영과 연결됩니다.
젊은층(80%)과 고령자 및 주거취약계층(20%)입니다.	행복주택출자 사업의 주요 수혜자는 대학생, 사회초년생, 신혼부부 등 젊은층(80%)과 고령자 및 주거취약계층(20%)입니다.
…	…

- 전체적인 답변의 어미가 통일되며, 문장 완성도가 높아짐