DACON

재정정보
AI 검색알고리즘 경진 대회

2024.08.30

팀명 : TA

재정정보
AI 검색알고리즘 경진 대회

CONTENTS

Chapter 01

경진대회 개요

01 대회 개요 및 팀 소개

02 활용 데이터

03 개발 프로세스

01 대회 개요 및 팀 소개

열린재정의 중앙정부 재정 정보 검색 및 제공 편의성과 활용도를 높이는 질의 응답 AI 알고리즘 개발

경진대회 개요

일정	2024년 7월 29일(월) 10:00 ~ 2024년 8월 23일(금) 10:00
배경	• 방대한 양의 재정 데이터에서 원하는 정보를 찾기 어려움 • 일반 국민도 쉽게 접근할 수 있는 검색 매개체 필요 재정 데이터의 접근성과 활용도를 혁신적으로 개선
활용 데이터	• 재정 정보 질의 응답 데이터셋 • 재정 보고서 • 예산 설명자료 • 기획재정부 보도자료
정량적 평가지표	(n: 샘플수) $\text{Average F1} = \frac{1}{n}\sum_{i=1}^{n}\frac{2 \times Precision_i \times Recall_i}{Precision_i + Recall_i}$

TA

팀원	김원철 / 롯데이노베이트 / 7년차 재직중 김기환 / 롯데이노베이트 / 3년차 재직중 김광륜 / 롯데이노베이트 / 3년차 재직중
성과	Public 점수 : 0.75402 (1등) Private 점수 : 0.72301 (3등)
TA 만의 Key Point	**Key Point 01 데이터 품질 최적화** • 문서 별 질의응답 수와 답변 길이를 고려한 데이터 증강 기법 적용 • 마크다운 형식을 활용한 정밀한 표 정보 추출 및 처리 **Key Point 02 고도화된 모델 학습 전략** • 다중 작업(Multi-Task) 학습 방식을 통한 모델의 문서 이해력 향상 **Key Point 03 효율적인 모델 추론 시스템** • 비동기 API 및 최적화된 추론 엔진 도입으로 추론 속도 개선

02 활용 데이터

재정 정보 검색을 위한 질의응답 AI알고리즘 개발을 위해 활용된

16부의 재정관련 문서와 496건의 질의응답 데이터

재정관련 문서

문서 수량	16 부
예시	재정통계해설 2024년도 성과계획서(총괄편) 2024 나라살림 예산개요 1-1 2024 주요 재정통계 1권 보건복지부_생계급여 「FIS 이슈 & 포커스」 22-3호《재정융자사업》 월간 나라재정 2023년 12월호 고용노동부_청년일자리창출지원 「FIS 이슈 & 포커스」 23-3호《조세지출 연계관리》 중소벤처기업부_창업사업화지원 고용노동부_조기재취업수당 국토교통부_소규모주택정비사업 국토교통부_전세임대(융자) 국토교통부_민간임대(융자) 고용노동부_내일배움카드(일반) 보건복지부_노인일자리 및 사회활동지원

질의응답 데이터

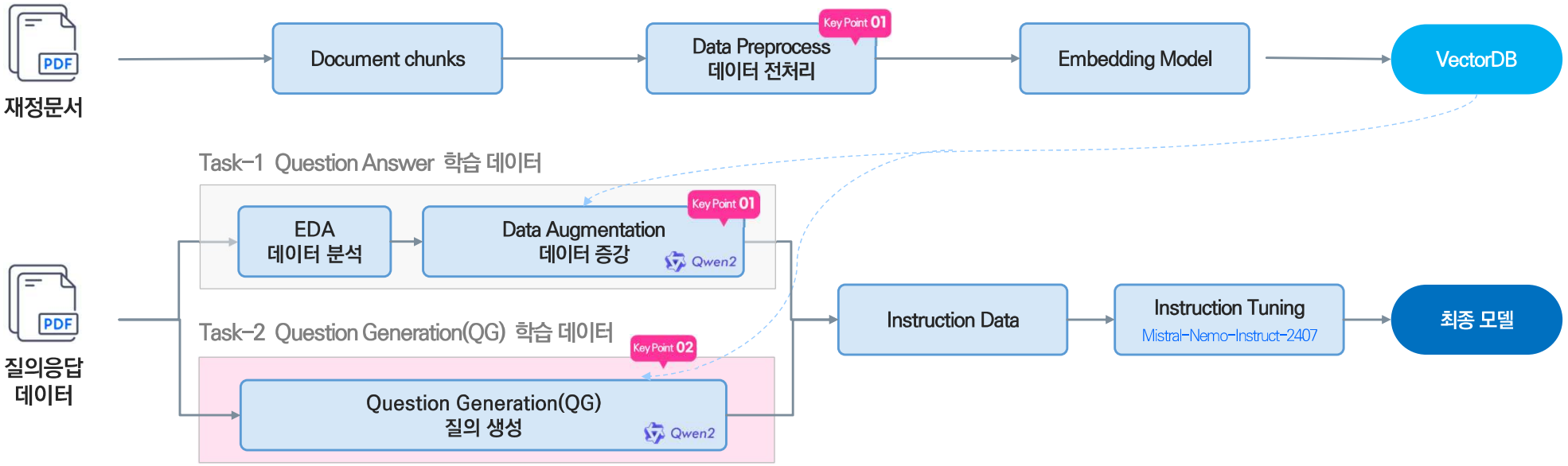
QA 수량	496건

예시

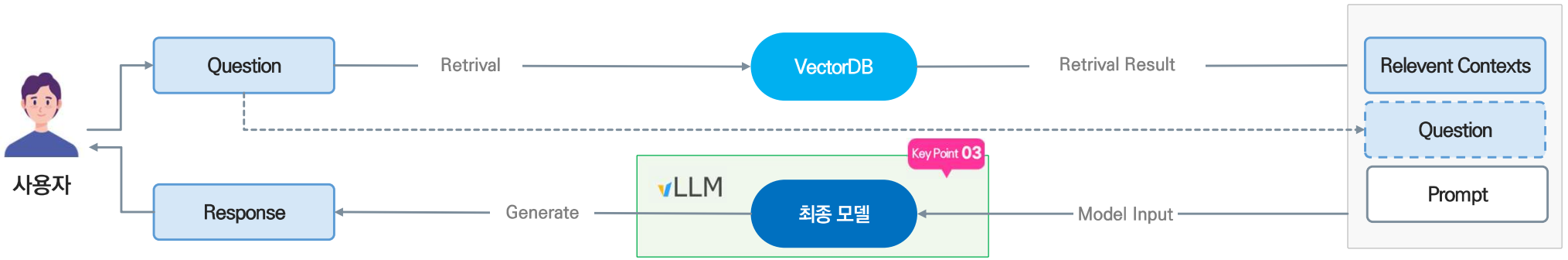
변수명	설명	예시
SAMPLE_ID	고유ID	TRAIN_000
Source	문서	1-1 2024 주요 재정통계 1권
Source_path	문서 경로	./train_source/1-1 2024 주요 재정통계 1권.pdf
Question	질의	2024년 중앙정부 재정체계는 어떻게 구성되어 있나요?
Answer	답변	2024년 중앙정부 재정체계는 예산(일반·특별회계)과 기금으로 구분되며, 2024년 기준으로 일반회계 1개, 특별회계 21개, 기금 68개로 구성되어 있습니다.

03 개발 프로세스

학습 과정



추론 과정



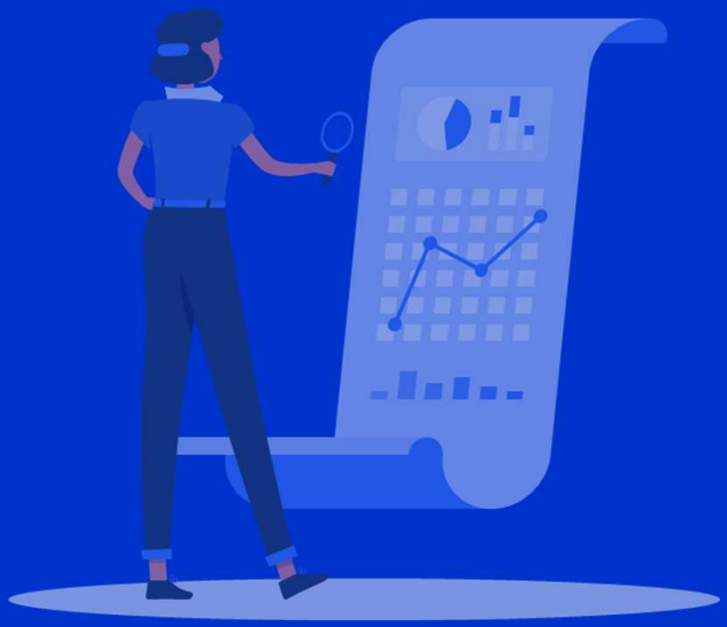
Chapter 02

데이터 활용

01 데이터 전처리

02 Question Answer 학습 데이터 증강

03 Question Generation(QG) 학습 데이터 생성



01 데이터 전처리

표 정보 추출 정교화

Key Point 01

- 재정관련 문서에 포함된 텍스트를 추출하고 정제하는 과정 중 표 정보 추출 과정을 설명합니다.
- 정확한 답변을 생성 위해서 **표 내부의 정교한 데이터 추출**이 중요합니다.

1 PDF에서 표 데이터 추출

- PyMuPDF4LLM 라이브러리를 사용하여 PDF 문서에서 표 데이터를 마크다운 형식으로 파싱
- 텍스트와 테이블을 추출하여 후속 전처리 과정에 활용

```
| 항목 | 세부내용 | 담당자 | 기한 | |------|----------|------|------|| 프로젝트 A | AI 모델 개발 | 김철수 | '24.6 | | 프로젝트 B | 데이터 분석 플랫폼 구축 | 이영희 | '24.8 |

항목 세부내용 담당자 기한 프로젝트 A AI 모델 개발 김철수 '24.6 프로젝트 B 데이터 분석 플랫폼 구축 이영희 '24.8
```

2 테이블 내용 정규화

- 파싱된 표 데이터를 행별로 분리한 후, 각 셀의 내용을 정리하여 일관된 형식으로 변환
- 테이블의 구조를 유지하면서도 불필요한 공백이나 줄바꿈을 제거하여 데이터를 깔끔하게 정리

3 중복 테이블 내용 제거

- 동일한 테이블이 여러 번 반복되어 나타나는 경우, 이를 감지하여 중복된 부분을 제거
- 테이블의 첫 번째 열과 마지막 열의 내용을 비교하여 중복된 구간을 찾아내어 삭제함으로써, 데이터의 중복성을 제거

4 텍스트 내 날짜 형식 변환

- 텍스트 내 날짜 형식이 일관되지 않은 경우, 'YY.MM' 형식을 'YYYY년 MM월'로 변환.
- 날짜 형식의 통일을 통해 분석 과정에서 혼동을 방지하고, 일관성을 유지

5 불필요한 기호 및 문구 제거

- 텍스트 내 포함된 특정 기호를 제거하거나 다른 문자로 대체.
  ※ 예: 'ㅇ', '○', '△', '⇦', '⇨' 등
- 여러 개의 기호를 활용하여 불필요한 줄바꿈, 공백, 혹은 잘못된 마크다운 문법을 수정하여 텍스트를 정리.
- 텍스트 내에서 의미가 없는 빈 괄호나 특정 패턴을 제거하여 텍스트의 가독성을 높임.
  ※ 예: '######', '[·]' 등

6 마크다운 테이블 형식 보정

- 파싱된 표 데이터의 형식을 보정하기 위해 '||'로 잘못된 테이블 구분선을 '| \n |'로 수정
- 테이블의 구조가 올바르게 표시되도록 하여, 데이터의 정확한 전달을 보장

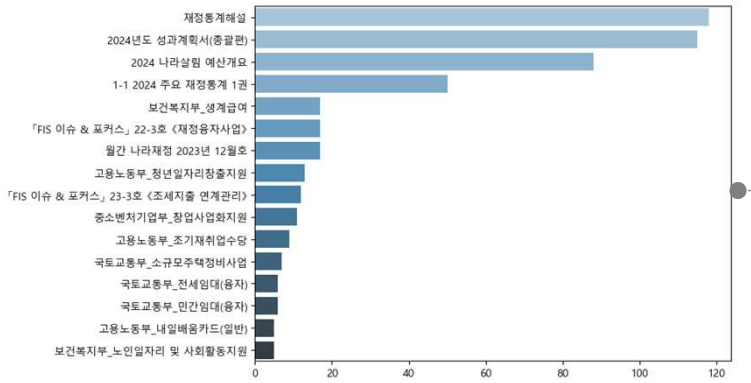
데이터 전처리 결과

```
| 항목 | 세부내용 | 담당자 | 기한 |
|------|----------|--------|------|
| 프로젝트 A | AI 모델 개발 | 김철수 | 2024년 6월 |
| 프로젝트 B | 데이터 분석 플랫폼 구축 | 이영희 | 2024년 8월 |
```

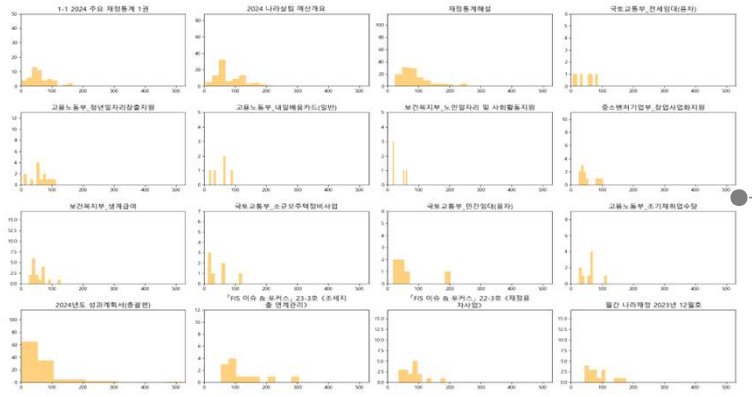
02 Question Answer 학습 데이터 증강 (1/2)

EDA를 통한 데이터 증강을 위한 데이터 불균형성 확인

- 문건별 질의응답 수와 답변(A)의 길이를 확인하고 이를 기반으로 데이터 증강 수행에 활용 합니다
- 증강하고자 하는 목표 수량을 답변 길이를 고려해 증강합니다



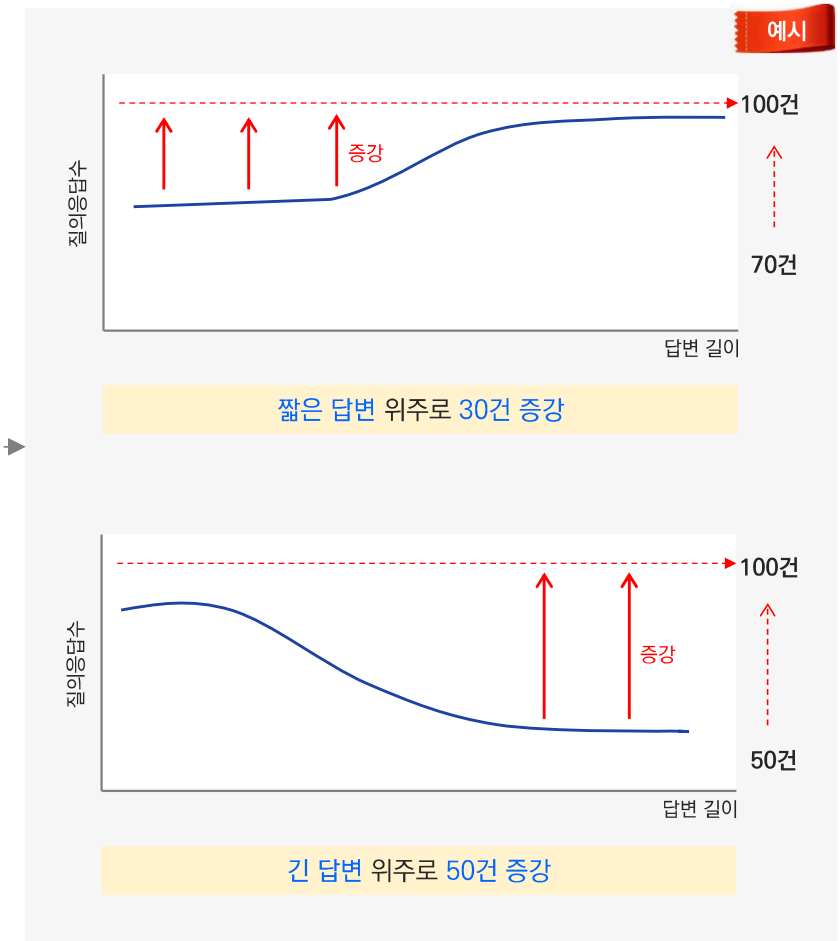
문서별 질의응답셋 수의 불균형



문서별 답변의 길이에 따른 질의응답수의 불균형

증강할 **목표 수량** 설정

답변의 **길이**를 고려해 증강



02 Question Answer 학습 데이터 증강 (2/2)

증강 방법

Key Point 01

- 다음과 같은 명령어와 기존 질의 예시, 문서를 첨부하여 Qwen모델을 통해 새로운 질의응답셋을 생성합니다

1 데이터 준비

- 제공된 학습데이터 질의응답셋
- 문서를 청크로 분할하고 FAISS로 벡터화 및 인덱싱
- BAAI/bge-m3 임베딩 모델 사용

질의응답 데이터셋

벡터화한 문서

2 프롬프트 설계

- 기존 학습 데이터에서 10개의 질문-답변 쌍 무작위 선택
- FAISS 벡터 데이터베이스에서 무작위로 문서 선택
- 선택된 예시를 프롬프트에 포함시켜 일관성 유지
- Qwen(Qwen2-72B-Instruct-AWQ) 모델로 새로운 질문-답변 생성
- outlines 라이브러리의 JSONLogitsProcessor로 출력 구조화
- 생성된 데이터 검증 및 저장

Qwen2

프롬프트(prompt)

당신은 재정전문가이며, 재정문서를 분석하여 문제를 내는 도우미입니다.
예시를 참조하여 **다음 문서**에서 **단답형 문제 2개 서술형 문제 2개** 작성하세요.
반드시 문서에서 찾을 수 있는 정보를 한국어로 작성하세요.

#예시 : 기존 학습 데이터에서 10개의 질문-답변 쌍 무작위 선택

#문서 : FAISS 벡터 데이터베이스에서 무작위로 문서 선택

3 결과 및 후처리

- 생성된 QA 쌍 JSON 형식으로 저장
- 원본 문서 및 출처 정보 포함
- 증강된 데이터셋 최종 검증

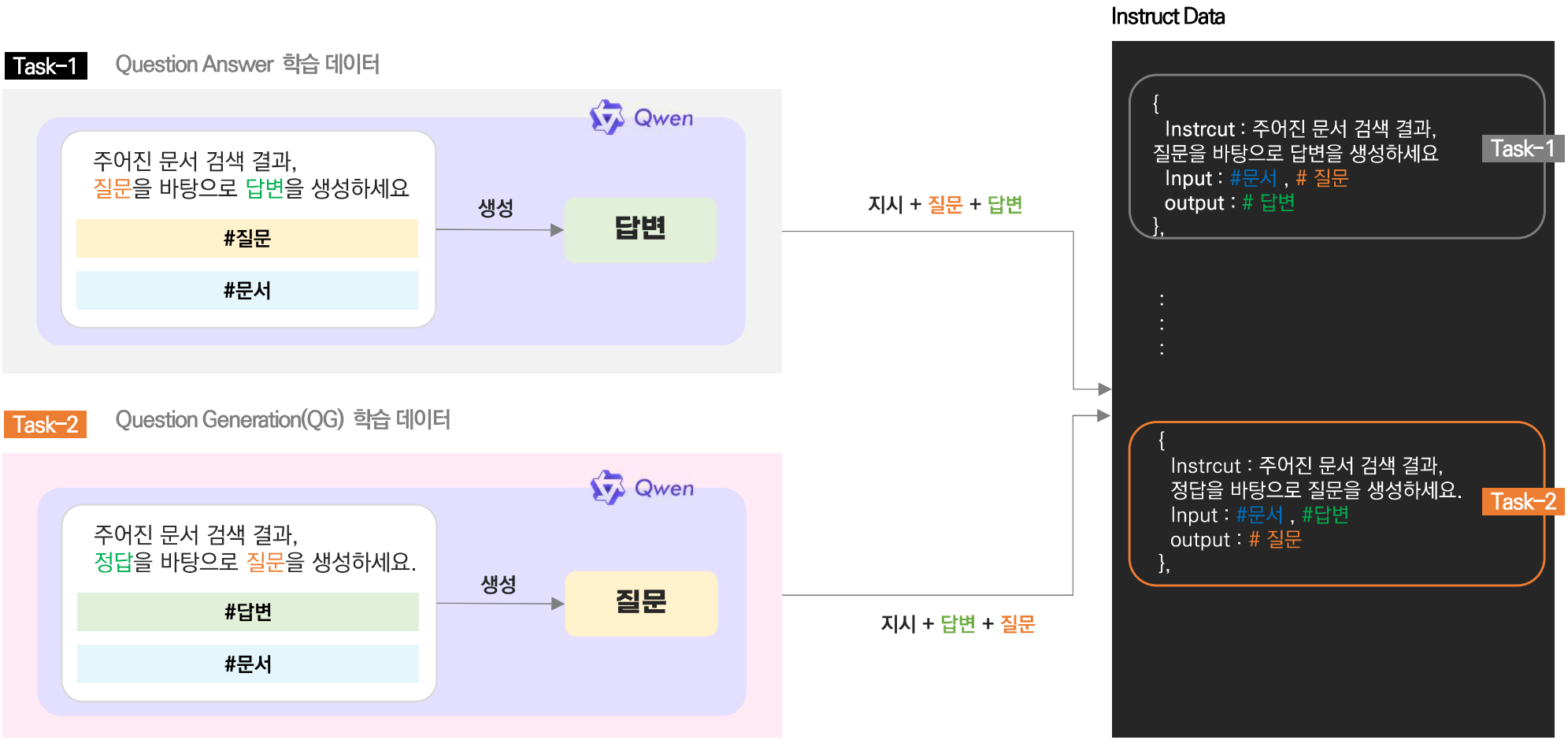
증강된 질의응답셋

03 Question Generation(QG) 학습 데이터 생성

다중 작업(Multi-Task) 학습 방식을 위한 QG 데이터 생성

Key Point 02

- Task는 지시사항(instruction), Input, Output이 다르다는 것을 의미합니다.
- 이렇게 다른 Task를 취합한 데이터로 학습하는 경우 언어에 대한 **이해도를 향상시켜** 더욱 유연한 답변이 가능한 모델이 생성됩니다.



Chapter 03

모델 알고리즘

01 검색 알고리즘

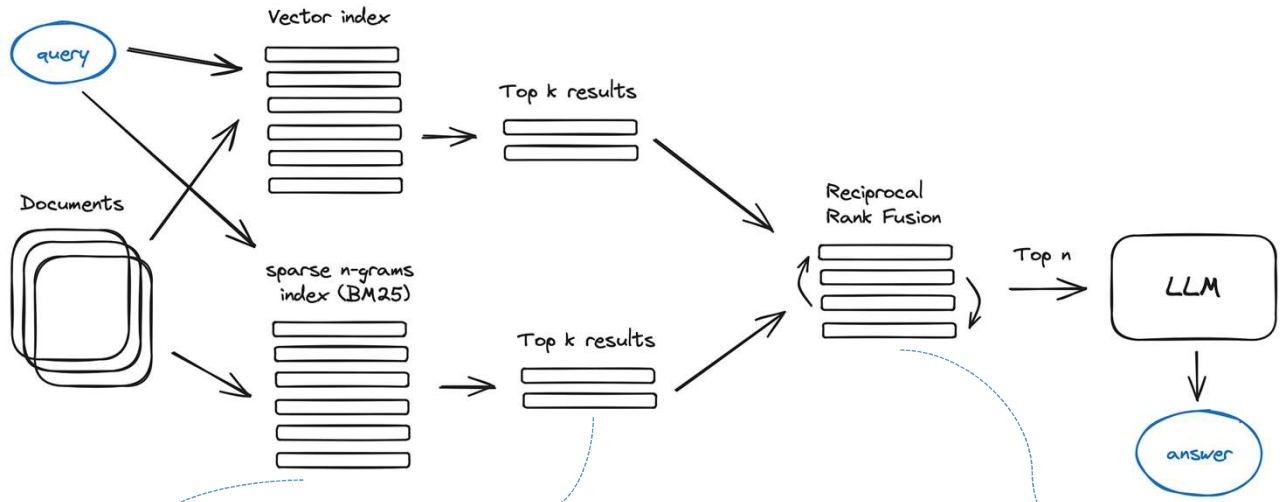
02 모델 학습

03 모델 추론

01 검색 알고리즘

앙상블 검색

- 앙상블 검색 알고리즘은 정확성과 속도를 모두 고려한 균형 잡힌 접근 방식을 제공하며, 다양한 유형의 쿼리에 대해 관련성 높은 결과를 반환할 수 있습니다.



1 BM25 검색

- 텍스트 기반 검색 알고리즘
- 단어 빈도와 문서 길이 고려
- 상위 4개 결과 반환

2 벡터 검색

- FAISS(Facebook AI Similarity Search) 사용
- 문서를 벡터로 변환하여 인덱싱
- 유사도 기반으로 검색해 **상위 3개** 결과 반환
- 임베딩 모델: BAAI/bge-m3

3 앙상블 검색

- EnsembleRetriever로 BM25와 FAISS 결과 통합
- 두 검색기 가중치 각 0.5로 설정
- **텍스트 기반(BM25)과 의미 기반(FAISS) 검색 결합**

02 모델 학습

Instruct Tuning

- DeepSpeed 기술 활용, 16K 컨텍스트 길이 확장, 그리고 네거티브 케이스 데이터 보강을 통해 학습 효율성을 높이고 모델 성능을 최적화합니다.

학습 환경

- GPU: NVIDIA A100 80GB 2장
- OS: Ubuntu 20.04.6 LTS
- Python 버전: 3.9.18

학습 주요 특징

1. DeepSpeed 활용

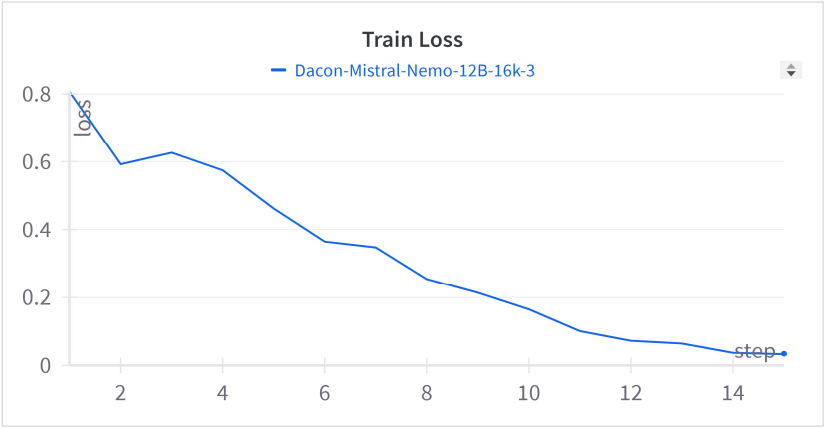
- ZeRO 기술로 모델 파라미터를 2개의 A100 GPU에 효율적으로 분산함으로써 메모리 효율적 사용
- 분산 처리와 최적화된 연산으로 대규모 모델 학습 가속화하여 학습 속도 향상

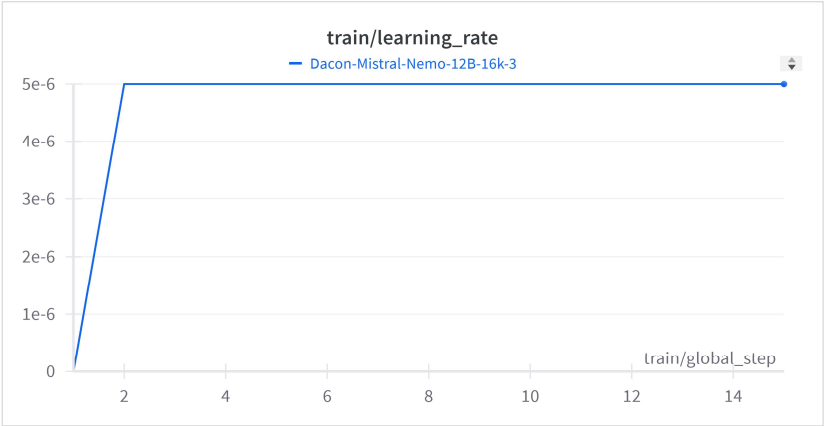
2. 16K 컨텍스트 길이

- 긴 컨텍스트 처리 능력 강화
- 더 많은 정보를 단일 입력으로 처리 가능

3. 네거티브 케이스 데이터 추가

- 문서에 답변이 없는 케이스를 의도적으로 포함
- 관련 정보 부재 시 "답변 없음 " 에 대한 응답 능력 향상으로 답변 신뢰성 개선
- 추가적인 효과로 전반적인 모델 답변 성능 개선도 확인됨





03 모델 추론 (1/2)

비동기 API 및 추론 엔진 적용

Key Point 03

- 비동기 API와 추론 엔진은 모델의 성능을 최대한 활용하면서도 **실제 서비스 환경에서 안정적이고 효율적인 운영**을 가능하게 합니다.

비동기 FastAPI 개발

FastAPI는 비동기 처리를 지원 하는 고성능의 Python 웹 프레임워크

비동기 처리
동시에 많은 요청을 효율적으로 처리할 수 있어 대규모 트래픽 상황에서도 안정적인 성능 제공

빠른 실행속도
Starlette 프레임워크를 기반으로 하여 매우 빠른 실행 속도

자동 API 문서화
Swagger UI와 ReDoc을 통해 API 문서를 자동으로 생성하여 개발 및 유지보수 용이

vLLM 추론 엔진 적용

vLLM은 대규모 언어 모델(LLM)을 위한 고성능 추론 엔진

페이지드 어텐션
메모리 사용을 최적화하여 긴 시퀀스도 효율적으로 처리

연속 배치 처리
요청을 연속적으로 배치 처리하여 GPU 활용도를 극대화

KV 캐시 관리
효율적인 Key-Value 캐시 관리로 중복 계산을 줄이고 추론 속도 향상

분산 추론
여러 GPU에 걸쳐 모델을 분산시켜 더 큰 모델도 처리

1. 동시 다중 사용자 지원 강화
2. 추론 속도 대폭 향상
3. 서버 리소스 효율적 활용
4. 메모리 최적화로 긴 컨텍스트 처리 가능

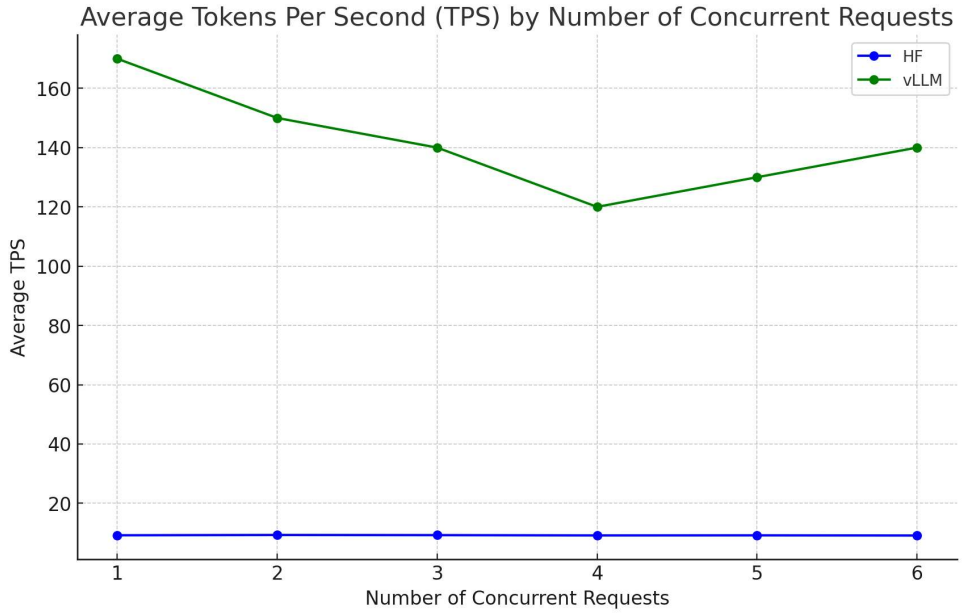
03 모델 추론 (2/2)

API 추론 속도 및 부하 테스트

- 실제 서비스 환경에서의 추론 성능을 종합적으로 평가하고, 최적화된 추론 엔진(vLLM)의 효과를 입증하기 위해 테스트를 진행합니다.

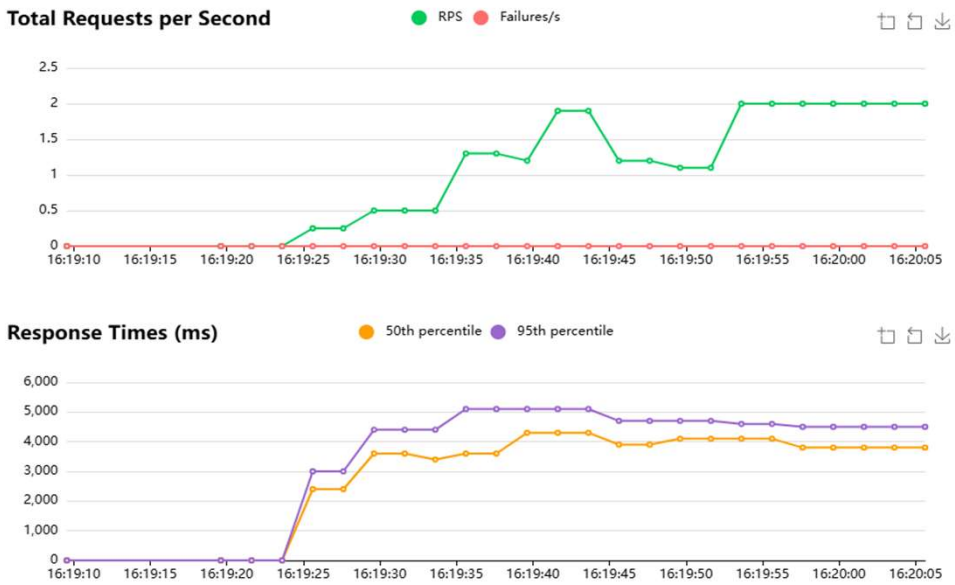
※ NVIDIA A100 * 1EA / Mem 80G 에서 테스트함

API 추론 속도 테스트



- vLLM을 적용한 버전과 적용하지 않은 버전(HF) 두가지 버전의 초당 토큰 생성 속도 비교함
- vLLM을 사용한 경우는 단일 작업에서 초당 약 170토큰에 가까운 처리 속도를 보임
- vLLM을 적용하지 않은 경우는 단일 작업에서 초당 약 9.3토큰 처리 속도를 보임
- 본 대회의 경우, vLLM을 적용한 경우가 적용하지 않은 경우 대비 약 18배의 속도 향상을 보임

API 부하 테스트



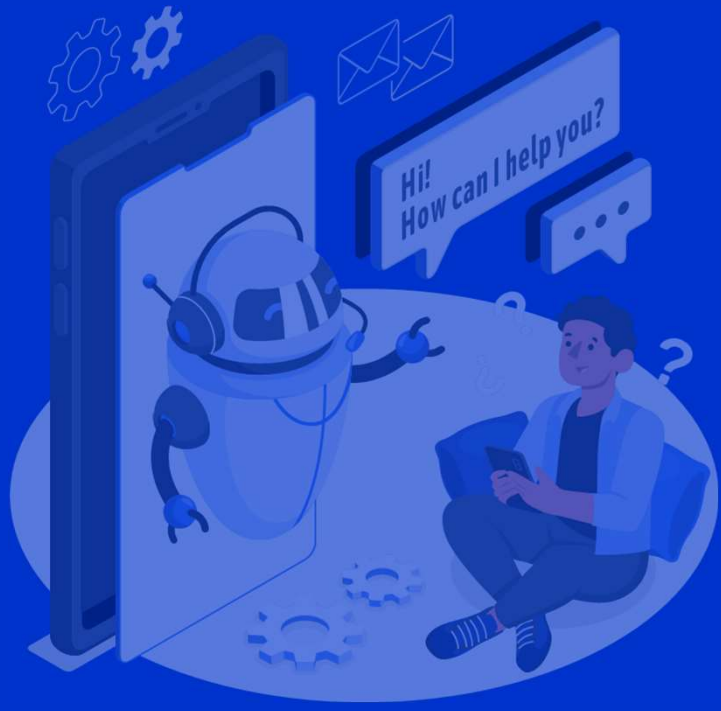
- 분산 시스템에서 웹 애플리케이션의 부하를 시뮬레이션 할 수 있는 Locust라는 오픈소스 부하 테스트 도구를 활용하여 API 부하 테스트를 진행함
- 테스트 환경은 20명의 유저가 불특정한 시간 간격(3~6초)으로 기능을 실행한다고 가정
- 응답 시간은 중간값 기준 3.5~4초로 요청을 안정적으로 처리하며, 요청 실패율 없이 안정적인 성능

Chapter 04

적용 가능성

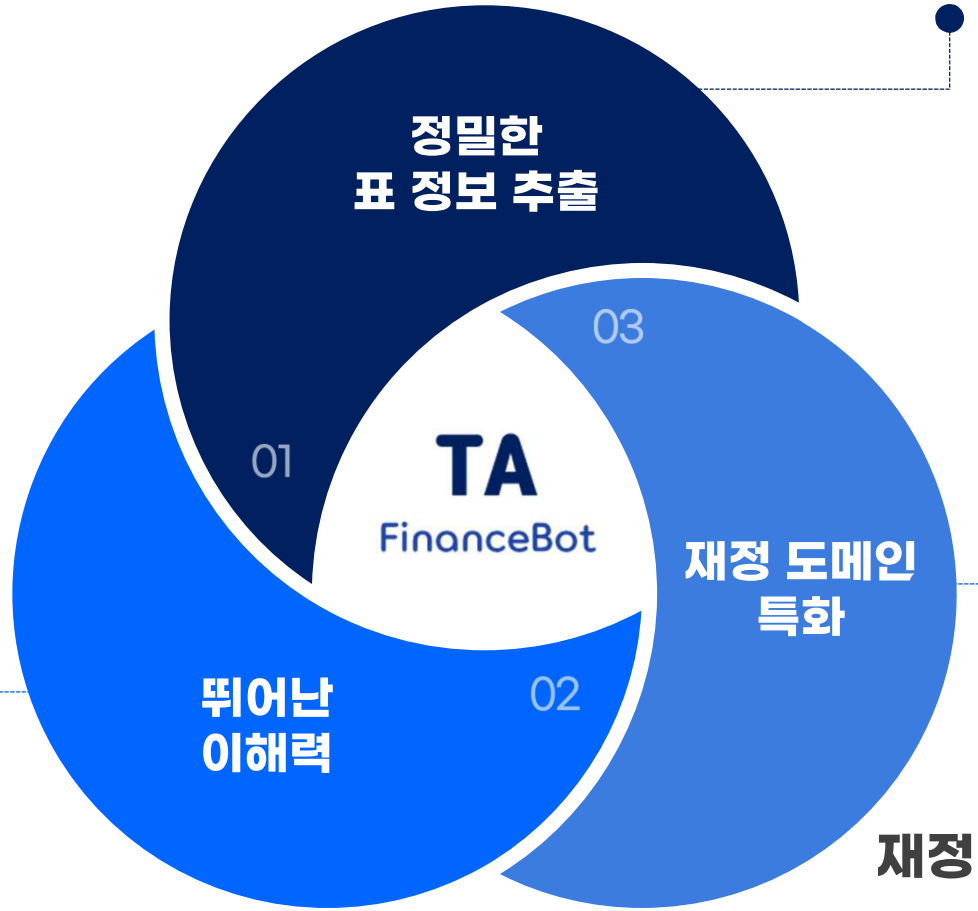
01 모델 실용성 및 활용 가능성

02 서비스 적용 가능성



01 모델 실용성 및 활용 가능성 (1/2)

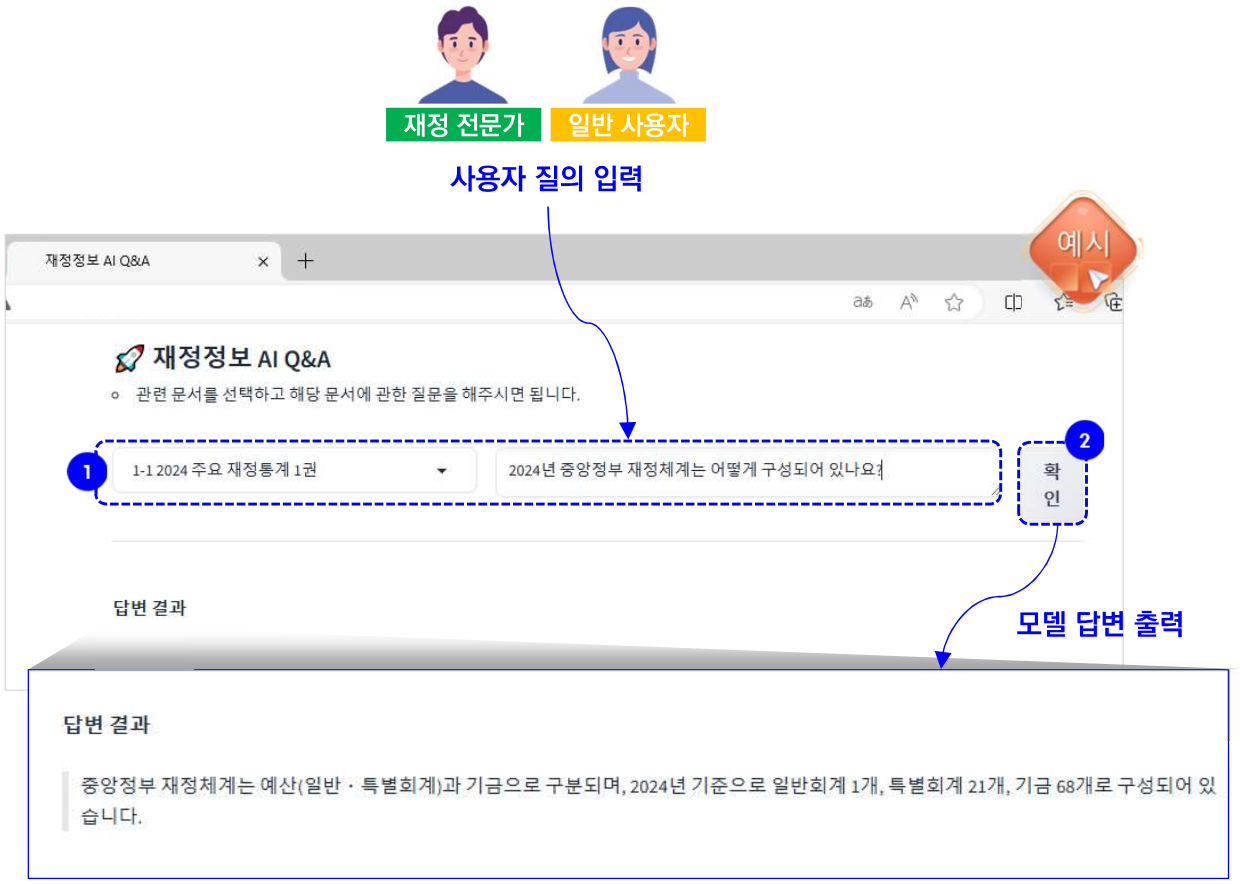
정확한 정보 확보로
답변 신뢰성 향상



다양한 질의에 대한
대처능력 뛰어남

재정 데이터로 학습하여
전문성 강화

01 모델 실용성 및 활용 가능성 (2/2)



- 사용자 접근이 용이한 매개체 확보
- 방대한 양의 재정정보 검색 편의성 향상
- 국민 응대 서비스의 효율성 향상

02 서비스 적용 가능성

- 제시된 기술 스택과 최적화 방법들은 실제 서비스 구현에 매우 적합한 환경을 제공합니다.
- 인프라 측면에서는 높은 확장성과 안정성을 보장하며, 속도 측면에서는 빠른 응답 시간과 높은 처리량을 실현할 수 있습니다.

인프라 측면

확장성

- DeepSpeed와 vLLM의 분산 처리 능력으로 다중 GPU 환경에서 효율적인 확장 가능
- FastAPI의 비동기 처리를 활용한 다중 서버 구성으로 트래픽 증가에 유연하게 대응 가능

안전성

- 다중 서버 구성 시 로드 밸런서를 통한 트래픽 분산으로 시스템 안정성 향상
- 서버 간 중복 구성으로 단일 장애점(Single Point of Failure) 방지

리소스 관리

- vLLM의 연속 배치 처리와 KV 캐시 관리로 GPU 리소스 최적화
- DeepSpeed의 ZeRO 기술과 vLLM의 페이지드 어텐션으로 메모리 사용 최적화

유지보수성

- FastAPI의 구조화된 코드 베이스로 유지보수 용이성 증대
- FastAPI의 내장 로깅 기능과 프로메테우스 통합으로 시스템 상태 실시간 모니터링 가능

속도 측면

응답시간

- FastAPI의 비동기 처리로 동시 요청 처리 시 대기 시간 최소화
- vLLM의 고속 추론 성능으로 개별 요청에 대한 처리 시간 단축
  ※본 대회의 데이터에 적용시 기존 대비 18배 성능 향상

처리량

- FastAPI의 비동기 처리와 vLLM의 연속 배치 처리로 높은 동시 사용자 처리 가능
- vLLM의 최적화된 GPU 활용으로 단위 시간당 처리량 증가

지연시간 최소화

- vLLM의 효율적인 KV 캐시 관리로 반복적인 계산 감소
- 16K 컨텍스트 길이 지원으로 긴 입력도 추가 처리 없이 빠르게 처리 가능

사용자 경험

- 빠른 응답 시간으로 대화형 서비스에서 자연스러운 상호작용 가능
- 부하 증가 시에도 안정적인 응답 시간 유지로 신뢰성 있는 서비스 제공

재정정보
AI 검색알고리즘 경진 대회

감사합니다

팀명 : TA