입찰메이트 AI:
정확한 문서 검색을 위한
TTS RAG 시스템 구축

> SYSTEM_INITIALIZED: AI_6_TEAM_3

TEAM 3

박상진 김나연 서유종 안호성 장우정

01. 프로젝트 배경

문제상황

RFP 문서 검토의 어려움

입찰 및 지원사업 공고를 검토하는 기업·기관 실무자는 방대한 문서 속

- 핵심 정보(예산, 일정)를 직접 탐색
- 숙련도에 따라 해석 편차 발생 가능
- 반복적 수작업으로 업무 부담 증가

➔ 실질적인 시간 비용 발생

해결방안

지능형 RFP 분석 및 음성 브리핑 시스템

- RAG 기반 정확한 핵심 정보 검색
- TTS 기반 자동 음성 브리핑 제공
- 비정형 문서 구조를 고려한 문맥 기반 질의응답 지원

➔ 반복적인 문서 탐색 과정을 자동화하여 핵심 정보 중심의 빠른 판단을 지원

기대효과

멀티태스킹 최적화 입찰메이트 AI

- 방대한 문서에서도 핵심 정보 빠르게 파악
- 긴 문서를 읽는 피로도 해소 및 단순 반복 최소화
- 청취와 동시에 다른 업무 가능
- 실수 감소, 신속한 의사결정 지원

RFP : 제안요청서 (Request for Proposal). 발주 기관이 사업 요구사항을 정리한 공식 문서

02. 프로젝트 목표

Core Goal

분석 시간 단축

방대한 RFP 문서에서 필요한 정보를 즉각적으로 추출하여 분석 프로세스를 혁신합니다.

50% + Reduction

Tech Goal

답변 정확도 확보

하이브리드 검색 알고리즘을 통해 수치 및 기술 요건에 대한 무결한 답변을 제공합니다.

92.5% Hit Rate

User Goal

사용성 극대화

실시간 음성 합성(TTS)을 도입하여 시각적 피로도를 낮추고 정보 접근성을 강화합니다.

Near Real-time Audio

03. 서비스 데이터 처리 흐름도

문서 가공

업로드된 문서에서 텍스트와 이미지를 추출

검색 시스템

사용자의 질문 의도를 분석하여 가장 관련성 높은 내용을 찾아내고 텍스트 답변 생성

음성 생성 시스템

텍스트 답변을 실시간 음성으로 생성

사용자 인터페이스

텍스트와 음성을 화면에 출력하여, 이전 대화 문맥을 기억하여 자연스러운 질의 응답 환경 제공

유기적 데이터 흐름

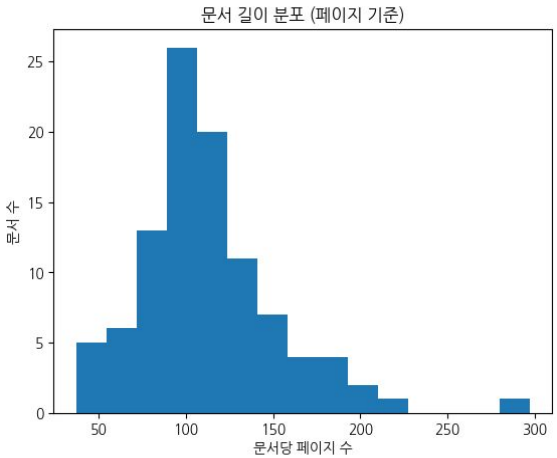
문서 및 이미지 데이터를 구조화하여 검색 엔진에 공급하고, 생성된 답변을 즉시 음성으로 변환하는 흐름을 구축했습니다.

대기 시간 최적화

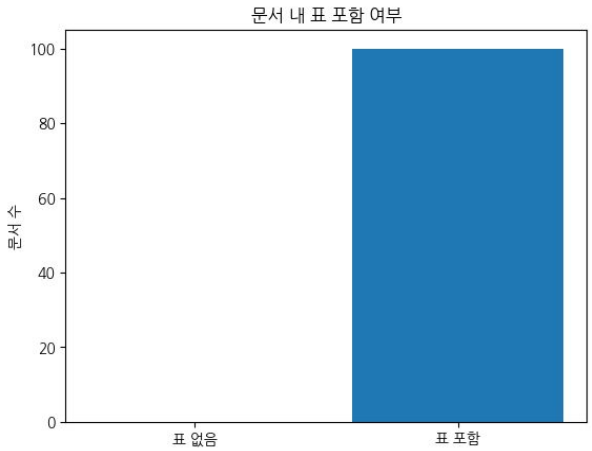
각 시스템의 독립적 작업과 경량화된 추론 모델을 통해 응답 대기 시간을 최소화하고 시스템 효율을 극대화했습니다.

04. 데이터 탐색 (EDA)

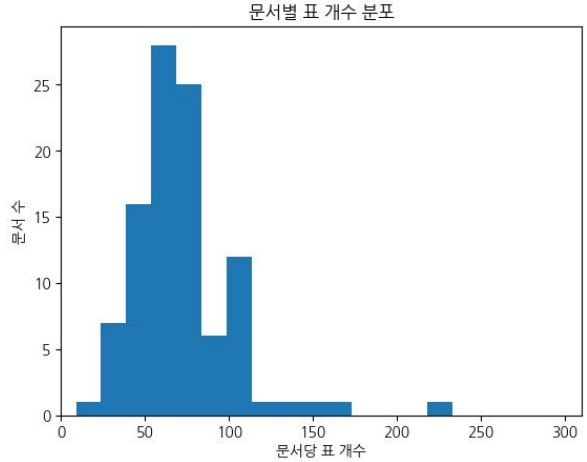
HWP 96건, PDF 4건 = 총 100건



▲ 총 100개 문서의 각 페이지 수
(평균적으로 약 110 pages를 가지고 있음)



▲ 모든 문서에 표가 1개 이상 포함되어 있음을 확인



▲ 각 문서의 표 개수
(평균적으로 약 70 개의 표를 가지고 있음)

05. EDA 결과

Result EDA

전체 문서 수: 100
평균 페이지 수 : 113.12
평균 표 개수: 71.98
최소 표 개수: 9
최대 표 개수: 233
표가 0개인 문서 수: 0

페이지 수와 표 개수 평균 비교

→ 평균 113 페이지 대비 평균 71개의 표가 포함되어 있음
→ 약 1.5페이지당 1개 표가 존재
→ 표 비중이 높은 문서 구조로 판단

페이지 수와 표를 읽을 수 있는 방법으로 문서를 읽어보는 방안을 물색

06. 데이터 파싱 파이프라인 (1)

Document Parsing Strategy

STEP 01

통합 포맷팅

HWP, PDF 등 파편화된 원본 데이터를 PDF 단일 포맷으로 통일하여 데이터 처리의 일관성을 확보합니다.

STEP 02

VLM 멀티모달 추출

단순 텍스트 추출의 한계를 극복하기 위해 VLM(Vision-Language Model)을 활용, 문서 내 표와 그림을 정확하게 해석합니다.

STEP 03

Markdown 변환

추출된 데이터를 구조적 분석이 용이한 Markdown(.md) 형식으로 변환하여 RAG 시스템의 입력 데이터로 최적화합니다.

Source Data Status

Format	Count	Processing Method	Status
HWP (Hangul Word Processor)	96 Files	HWP to PDF Conversion	**Completed**
PDF (Portable Document Format)	4 Files	Direct VLM Analysis	**Completed**

파이프라인 : 데이터가 입력되어 최종 결과물로 나오기까지의 일련의 자동화된 처리 공정
Parsing : 비정형 데이터를 분석해 의미 단위로 분해하고 구조화하는 과정

VLM : Vision-Language Model. 이미지와 텍스트를 함께 이해하여 처리하는 AI 모델
Markdown : 문서 구조(제목, 표, 목록)를 간단한 기호로 표현한 경량 마크업 언어

06. 데이터 파싱 파이프라인 (2)

Markdown Advantages

구조적 문맥 유지

텍스트뿐만 아니라 표의 시작과 끝, 페이지 구분자를 명시적으로 마킹하여 문서의 계층 구조를 보존합니다.

데이터 무결성 확보

비정형 PDF 데이터를 기계 학습 및 검색에 최적화된 정형 Markdown 포맷으로 변환하여 정보 손실을 최소화합니다.

검색 효율성 증대

구조화된 마커를 기반으로 정확한 청킹(Chunking)이 가능해지며, 이는 RAG 시스템의 검색 품질 향상으로 직결됩니다.

Before

사 업 명
통합 정보시스템 구축 사전 컨설팅
주관기관
재단법인 예술경영지원센터
2024 년 04 월
구 분
소 속
전화번호
이메일
입찰관련
경영지원팀
02-708-2212 ksj37@gokams.or.kr
사업관련
(한국 미술시장 정보시스템)
시각정보팀
02-2098-
2946
shwan@gokams.or.kr

After

```
<!-- page: 1 -->

<!-- tables: start page 1 -->

| 사업명 | 통합 정보시스템 구축 사전 컨설팅 |
|---|---|
| 주관기관 | 재단법인 예술경영지원센터 |

| 구분 | 소속 | 전화번호 | 이메일 |
|---|---|---|---|
| 입찰관련 | 경영지원팀 | 02-708-2212 |
ksj37@gokams.or.kr |
| 사업관련 (한국 미술시장 정보시스템) |
시각정보팀 | 02-2098-2946 |
shwan@gokams.or.kr |

<!-- tables: end page 1 -->
```

07. 테스트 검증 : (1) 데이터셋

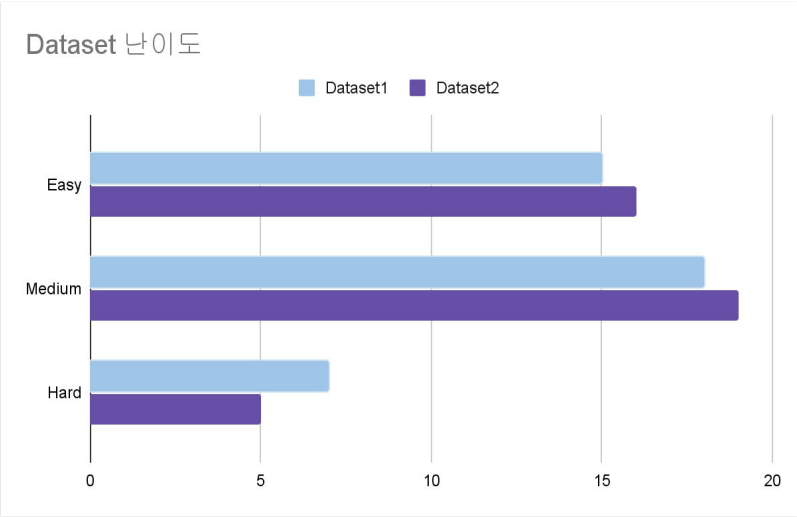
Dataset Statistics

Category	Dataset 1	Dataset 2
평가 영역	기술적 요구 사항	과업 범위
문항 수	40 문항	40 문항
주요 평가 요소	예산 및 일정 파악	문맥 및 상세 과업 파악
총 문항 갯수	총 80문항	

* Dataset 1은 예산, 일정 등 정량적 데이터 중심이며,

Dataset 2는 과업 내용 및 제안 요청 사항 등 문맥 파악 중심의 질문으로 구성되었습니다.

Difficulty Distribution



07. 테스트 검증 : (2) 난이도 별 질문

Easy

Dataset1

부산 국제 영화(BIFF) & ACFM 온라인
서비스 재개발 사업의 총 예산은 얼마인가요?

(예산 질의)

Dataset2

고려대학교 차세대 사업의 2025학년도 대가
지급 비율은 대략 몇 퍼센트인가요?

(사업 조건)

Medium

Dataset1

예술경영지원센터 통합 정보시스템 구축 컨설팅
사업의 입찰에 참여하기 위한 기업 규모 제한과
하도급 허용 여부는 어떻게 되나요?

(입찰 참가 조건)

Dataset2

인천공항운영서비스 차세대 ERP 시스템 구축
과업 범위 중 기존에 운영 중인 레거시
시스템과의 연계 항목은 무엇인가요?

(과업 범위)

Hard

Dataset1

벤처기업협회 사업의 최종 검수에서 시스템
운영 불가 판정이 날 경우 어떻게 처리되나요?

(위험 관리 규정)

Dataset2

한국수자원공사 타당성조사 및 기본계획 수립
용역에서 계약 변경이 가능한 주요 조건들은
무엇입니까?

(과업 요구사항)

08. 비교 테스트 (1) : VLM 파싱 모델

Model Performance Comparison

Chunking Method	Qwen3 (Chunks)	OpenAI (Chunks)	Variance
Paragraph	11,764	11,332	-3.7%
ContextEnriched	8,622	9,248	+7.3%
Semantic	8,622	8,246	-4.4%
Recursive	8,622	9,248	+7.3%
Hierarchical	8,622	9,437	+9.5%

OpenAI 모델 채택 근거

OpenAI 파싱 모델이 평균적으로 7~9% 더 많은 유효 청크를 생성하는 것으로 나타났습니다. 이는 텍스트 추출 품질이 더 세밀하고, 문서 내의 미세한 정보를 놓치지 않고 구조화한다는 것을 의미합니다.

Chunk Generation Variance

09. 비교 테스트 (2) : 청킹 x 임베딩 모델

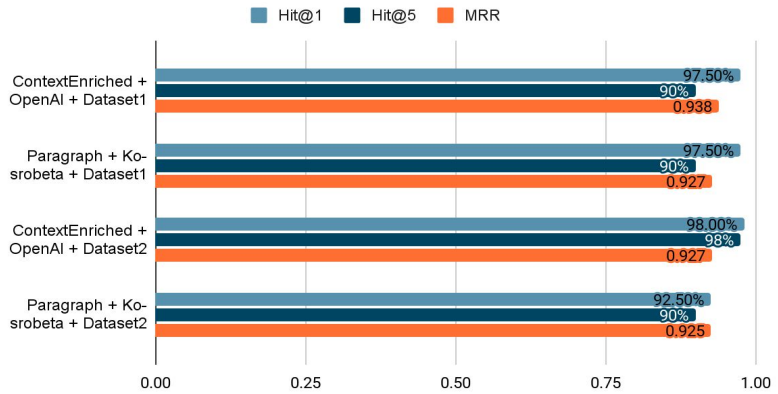
Embedding Model Comparison

Embedding Model	Key Characteristics
OpenAI	높은 일관성과 추론 능력 제공
Ko-srobeta	한국어 특화 오픈소스
MiniLM	경량 다국적 모델이나 한국어 도메인으로는 취약

Embedding : 텍스트의 의미를 고차원의 수치 벡터로 변환하여 컴퓨터가 의미적 유사도를 계산할 수 있게 하는 기술

Embedding Generation Variance

TOP 2



> FINAL_DECISION: ContextEnriched + OpenAI

OpenAI 파싱 결과가 7~9% 더 많은 유효 청크를 생성하여 텍스트 추출의 세밀함이 우수함을 확인했습니다.

문맥 보존력이 가장 뛰어난 ContextEnriched 청킹과 OpenAI 임베딩 조합을 최종 채택했습니다.

10. 파라미터 최적화 결과

Grid Search: Recall by Chunk Size & K

K Value	Chunk Size			
	400	700	900	1200
K=3	52.1%	58.3%	54.8%	49.6%
K=5	56.0%	67.0%	60.0%	60.4%
K=7	60.0%	71.7%	68.0%	66.5%
K=10	67.0%	**72.7%**	70.5%	71.4%

* Grid Search 결과, Chunk Size 700과 K=10 조합에서 가장 높은 Recall 값을 확인하였습니다.

Grid Search : 다양한 파라미터 조합을 미리 정의된 범위 내에서 전수 조사하여 최적의 성능을 내는 조합을 찾는 최적화 방식.

Final Optimized Parameters

OPTIMIZED CHUNK SIZE

700 Tokens

KEYWORD RECALL

72.7 %

RETRIEVAL HIT RATE

92.5 %

11. RAG 시스템 비교 테스트

검색과 답변 품질 정량 측정

평가 지표	설명	평가 대상
Retrieval Hit Rate 검색 성공률	정답 소스 문서가 검색 결과(Top-K)에 포함되었는가?	검색 정확도
Keyword Recall 정보 포함도	기대 키워드 중 검색된 텍스트에 포함된 비율	정보 충분성
Answer Accuracy 답변 정확도	LLM 답변에 반드시 포함되어야 할 키워드가 있는가?	생성 품질

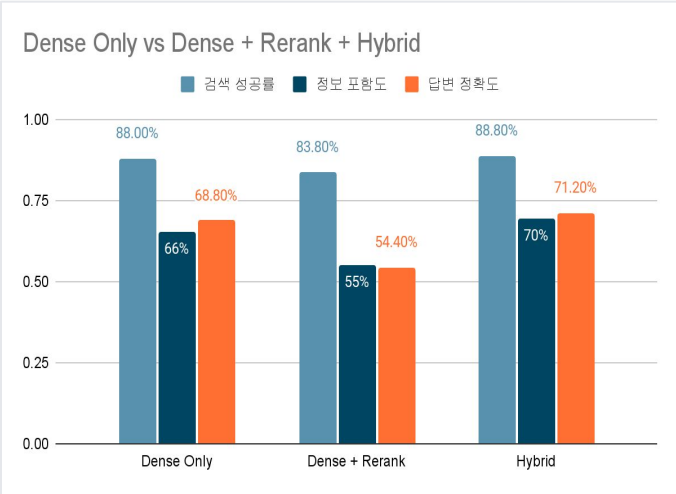
Retriever : 질문에 필요한 근거 내용만 빠르게 발췌하는 지능형 탐색기

Dense : 문장을 고차원 벡터로 변환하여, 키워드가 일치하지 않아도 의미적 유사성을 바탕으로 정보를 찾는 검색 방식

Rerank : 1차로 검색된 문서들AI 모델이 연관성을 기준으로 평가하고, 가장 정확한 문서가 최상단에 오도록 순서를 재배치 하는 과정

Hybrid : 단어의 일치 여부를 찾는 '키워드 검색' 과 문맥을 파악하는 '의미 검색' 결합하여 최적의 결과를 추출하는 검색 방식

RAG Retrieval Variance



> FINAL_DECISION: Hybrid 방식 채택

평가지표가 가장 높은 Hybrid 방식을 채택하였습니다.

12. RAG 시스템 설계 (1)

단어와 문맥을 동시에 파악하는 Hybrid Retriever

Dense Retrieval

텍스트의 의미적 유사도를 기반으로 검색합니다. 질문과 답변의 단어가 달라도 문맥이 유사하면 정보를 찾아냅니다.

Chroma DB | OpenAI Embedding | Semantic Search

BM25 Retrieval

키워드 빈도와 중요도를 기반으로 검색합니다. 사업명, 예산 수치 등 고유명사의 정확한 매칭에 탁월합니다.

Keyword Match | Term Frequency | Exact Search

Hybrid Ensemble Strategy

의미 기반 검색의 유연성과 키워드 검색의 정확성을 결합하여 RFP 분석에 최적화된 검색 품질을 확보했습니다.

0.6 : 0.4

BM25 : 문서 내 키워드의 빈도와 중요도를 계산하여 관련성을 측정하는 전통적인 키워드 기반 검색 알고리즘

12. RAG 시스템 설계 (2)

LCEL Chain Architecture

유연하고 안정적인 LCEL 시스템 구축

LangChain Expression Language(LCEL)를 사용하여 Retrieval, Prompt, LLM, Output Parser를 선언적 체인으로 조립했습니다.

Hybride 검색으로 답변 정확도 극대화

사용자의 질문에 따라 Hybrid Retriever 가 추출한 최적의 문서 조각들을 실시간으로 프롬프트에 주입합니다.

Streaming 출력으로 체감 대기시간 최소화

사용자 경험 향상을 위해 답변 생성 과정을 실시간 스트리밍으로 처리하여 대기 시간을 최소화했습니다.

프롬프트 엔지니어링

SYSTEM_PROMPT_V2

"당신은 RFP 분석 전문가입니다. 제공된 [Context]만을 바탕으로 답변하십시오.

1. 답변은 명확하고 간결하게 작성할 것.
2. 수치 데이터는 반드시 원문과 일치시킬 것.
3. 근거가 되는 문서의 페이지 번호를 명시할 것.
4. 모르는 내용은 추측하지 말고 '정보 없음'으로 답변할 것."

* 페르소나 부여 및 제약 조건 명시를 통해 환각(Hallucination) 현상을 방지하고 답변의 신뢰도를 확보했습니다.

LCEL : LangChain Expression Language. 다양한 AI 요소들을 규칙에 맞게 연결하여 자동화된 정보 처리 경로를 구축하는 언어
선언적 체인 : 질문이 들어가서 답변이 나오기까지의 과정을 하나의 파이프라인처럼 매끄럽게 연결하는 방식.
Streaming : 모델의 전체 응답이 완료될 때까지 기다리지 않고, 실시간으로 생성되는 결과를 문장(token) 단위로 전송하는 처리방식

12. TTS 시스템 설계 (1)

문제: 지연 병목

동기식 처리의 한계

음성 합성(TTS)은 고도의 연산 자원을 소모하며, 동기식으로 처리할 경우 심각한 지연 시간(Latency)이 발생합니다. 이는 사용자가 답변을 받기까지 불필요한 대기 시간을 유발하여 서비스 경험을 저해합니다.

해결: 비동기 구조

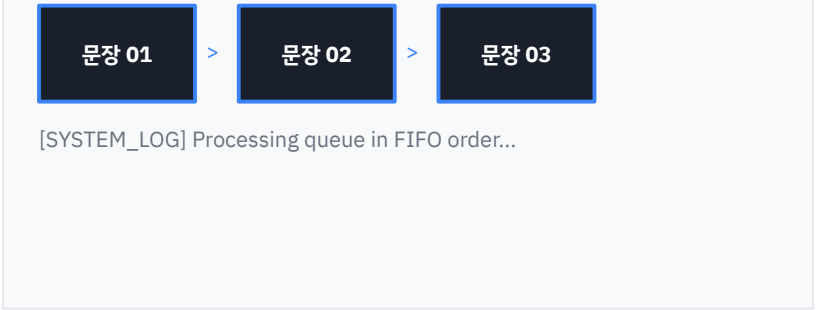
비동기 아키텍처 도입

메인 프로세스(RAG)와 분리된 비동기 워커(Async Worker)를 도입하여 음성 합성을 백그라운드에서 수행합니다. 이를 통해 텍스트 응답과 음성 출력을 병렬로 처리하여 체감 대기 시간을 획기적으로 단축합니다.

12. TTS 시스템 설계 (2)

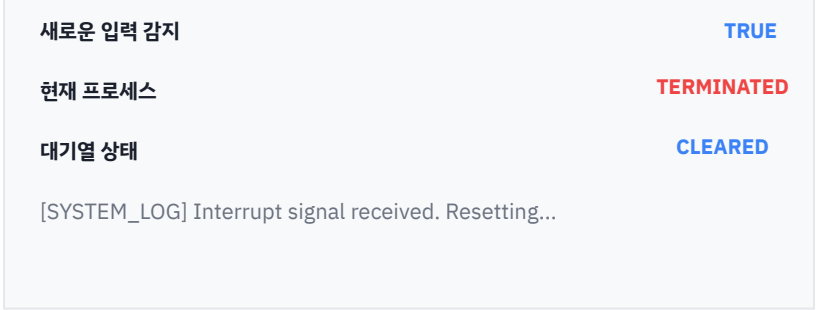
FIFO 대기열 & 제어 메커니즘

FIFO 대기열 시스템: 순서 보장



FIFO(First-In-First-Out) 대기열 시스템을 통해 문장 처리 순서를 엄격히 유지합니다. 이를 통해 오디오 겹침을 방지하고 대화의 자연스러운 흐름을 보장합니다.

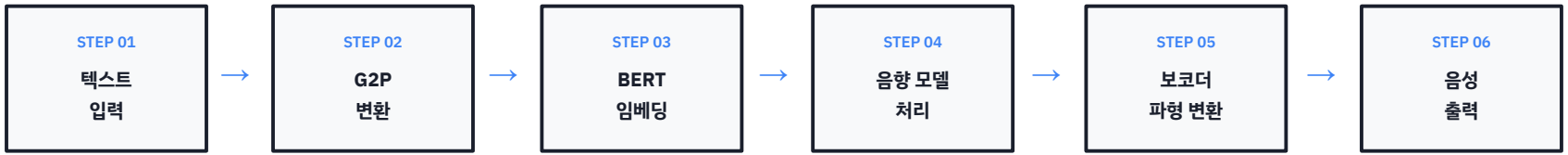
중단 제어: 즉각적 반응



사용자의 새로운 입력이 감지되면 현재 재생 중인 프로세스를 즉시 종료하고 대기열을 비웁니다. 최상의 반응성을 제공하여 대화의 단절을 최소화합니다.

13. TTS 추론 파이프라인

7 단계 추론 & ONNX 최적화



ONNX 런타임 엔진 최적화

PyTorch 모델을 ONNX 포맷*으로 변환하고 최적화된 런타임을 적용하여

CPU 환경에서도 실시간 처리가 가능한 추론 속도를 확보했습니다.

*Open Neural Network Exchange. 서로 다른 AI 프레임워크 간에 모델을 공유하고 최적화하여 실행할 수 있게 돕는 개방형 포맷.

14. 고품질 음성 합성 기술

음소 정확성: G2P

G2P (문자소-음소 변환)

텍스트(문자)를 실제 발음되는 소리(음소)로 변환하는 핵심 기술입니다. 한국어의 복잡한 음운 변화 규칙을 적용하여 '국민'을 [궁민]으로 정확하게 발음하도록 제어합니다.

Pronunciation Correction

Phonetic Transcription

문맥 반영 억양: BERT

BERT (문맥 정보 제공)

문장 전체의 문맥을 파악하여 모델이 자연스러운 억양(Prosody)을 생성하도록 문맥 정보를 제공합니다. 단어의 의미와 문장 구조를 이해함으로써, 질문이나 강조가 필요한 부분에서 인간과 유사한 음성 고저를 구현합니다.

Natural Intonation

Semantic Understanding

15. 데이터베이스 구축 및 피드백 루프

SQLite 스키마: 대화 이력

Field Name	Data Type	Description
session_id	TEXT (PK)	대화 세션 고유 식별자
user_query	TEXT	사용자 질문 원문
ai_response	TEXT	RAG 기반 생성 답변
feedback	INTEGER	만족도 (1: 좋아요, -1: 싫어요, NULL: 기본값)
timestamp	DATETIME	데이터 생성 일시

* 가벼운 SQLite를 활용하여 로컬 환경에서도 안정적인 대화 이력 관리와 빠른 조회가 가능하도록 설계했습니다.

핵심 효과

01 데이터 수집

사용자의 질문과 AI의 답변, 그리고 명시적인 피드백 데이터를 실시간으로 수집합니다.

02 오류 분석

부정적 피드백이 발생한 케이스를 분석하여 검색 실패인지 생성 실패인지 파악합니다.

03 시스템 최적화

분석 결과를 바탕으로 프롬프트를 수정하거나 모델을 강화학습하고, 리트리버의 파라미터를 재조정하여 성능을 고도화합니다.

16. 결론 및 기대효과

핵심 성과

RAG 성능 최적화 완료

하이브리드 리트리버와 파라미터 튜닝을 통해 Recall 72.7%, Hit Rate 92.5%를 달성하여 RFP 분석의 정확도를 획기적으로 높였습니다.

유사 실시간 TTS 시스템 구축

비동기 아키텍처와 ONNX 엔진 최적화를 통해 추론 속도를 향상시켰으며, 지연 없는 사용자 경험을 구현했습니다.

향후 확장성

미래 가치 및 확장 방향

01 도메인 확장: RFP뿐만 아니라 법률, 의료 등 정밀한 문서 분석이 필요한 다양한 산업군으로 솔루션 확장 가능

02 분류/ARS 기능: “메이트야~” 라고 불러서 음성 인식 시스템이 켜지면 사용자 음성을 입력받아 대화를 이어가는 서비스 구축

ex) 아이폰 “시리” 및 구글 “제미나이”

03 개인화 TTS: 사용자 피드백 루프를 활용한 맞춤형 음성 합성 및 대화형 인터페이스 고도화