

Intelligent Document Summary with Chatbot

이름 **채상준**

지도교수 **이상훈**

연구 배경

현대 사회는 방대한 양의 정보 속에서 살아가고 있으며, 매일 생성되는 문서와 보고서는 개인과 조직이 처리하기에 점점 더 큰 부담이 되고 있습니다. 특히 연구자와 전문가들은 중요한 정보를 파악하기 위해 많은 시간을 투자해야 했으며, 이는 정보 처리 도구의 필요성을 더욱 부각시켰습니다.

본 프로젝트는 이러한 문제를 해결하기 위해 사용자가 업로드한 문서를 요약하고, 문서 내용을 기반으로 질문에 답변할 수 있는 지능형 Chatbot 시스템을 개발하였습니다. 문서를 자동으로 요약하고 핵심 정보를 제공함으로써 사용자가 내용을 빠르고 효율적으로 이해할 수 있도록 지원하였습니다.

이 시스템은 다양한 형태의 문서를 처리하고, 자연스러운 대화를 통해 사용자 편의성을 극대화하도록 설계하였습니다. 이를 통해 정보 처리의 효율성을 높이고, 현대 사회의 정보 과부하 문제를 완화하는 데 기여하였습니다.

연구 진행 과정

본 프로젝트는 STT(Speech To Text) 구현, 문서 요약, 그리고 질문 응답 기능을 제공하는 지능형 시스템을 개발하기 위해 다음과 같은 과정을 거쳤습니다.

1. 문서 요약 기능 구현

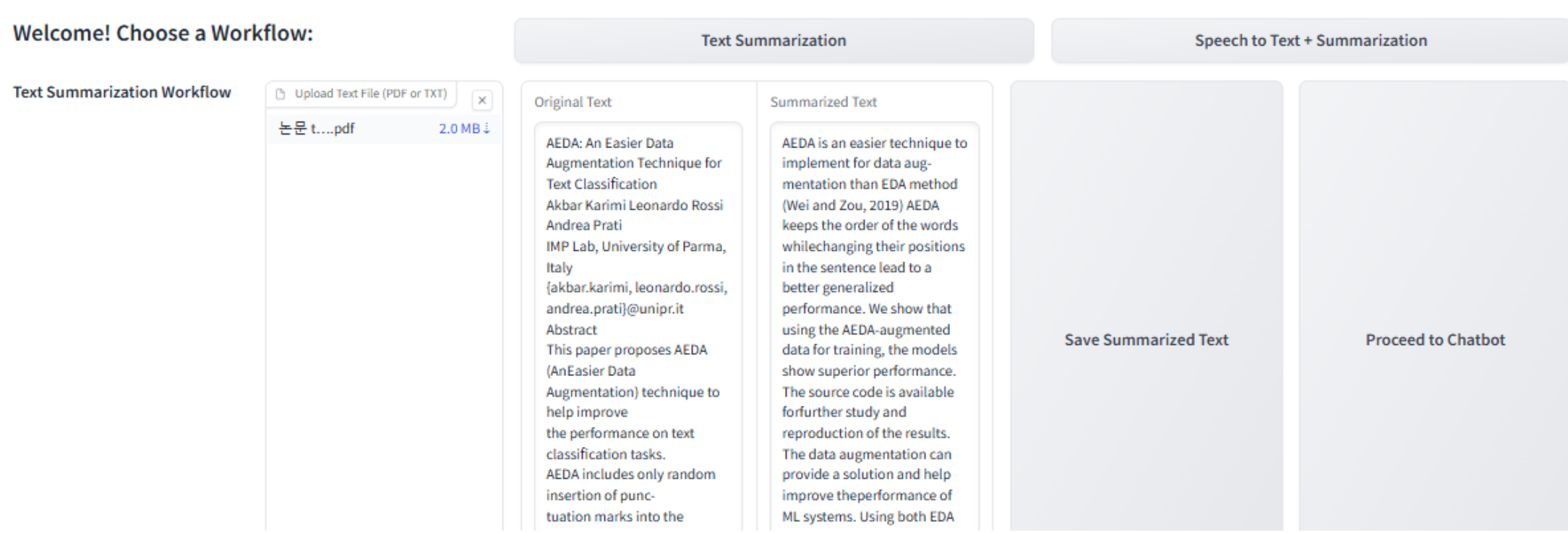
문서 요약은 텍스트 또는 PDF 파일의 내용을 간결하게 전달하기 위해 설계되었습니다.

- 문서 처리 및 전처리: PDF 파일은 PyPDF2를 사용해 텍스트를 추출하고, 불필요한 공백과 특수 문자를 제거하여 전처리하였습니다. 텍스트 파일도 동일한 과정을 거쳤습니다.

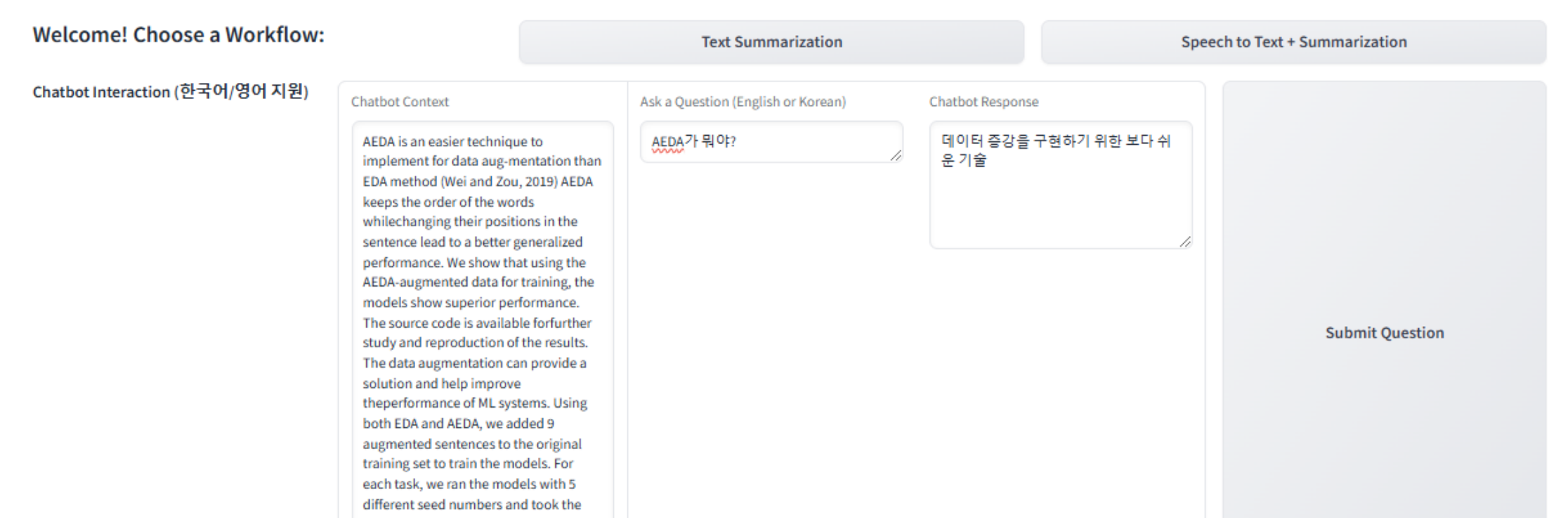
- 문서 요약 모델: Hugging Face의 facebook/bart-large-cnn 모델을 활용하여 텍스트를 요약하였습니다. 모델의 입력 길이를 초과하는 경우, 텍스트를 나누어 요약 후 병합하였습니다. 이를 통해 문서 길이로 인한 한계점을 해결하였습니다.

결과 및 분석

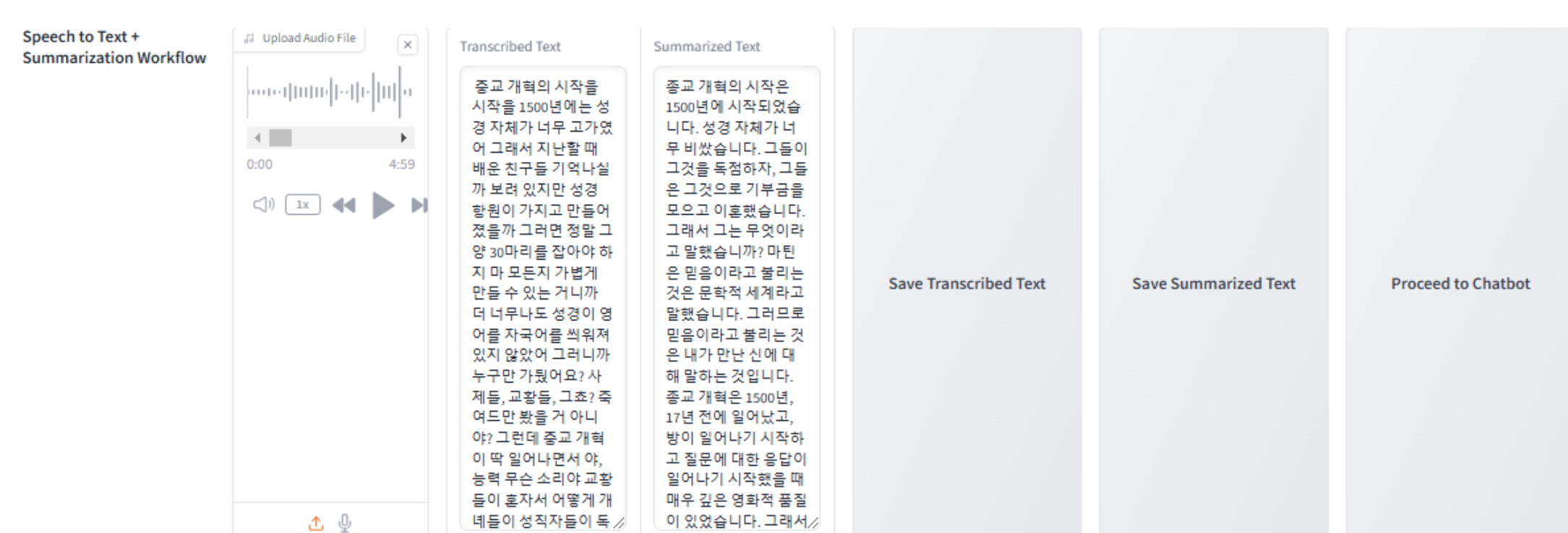
Text Summarization



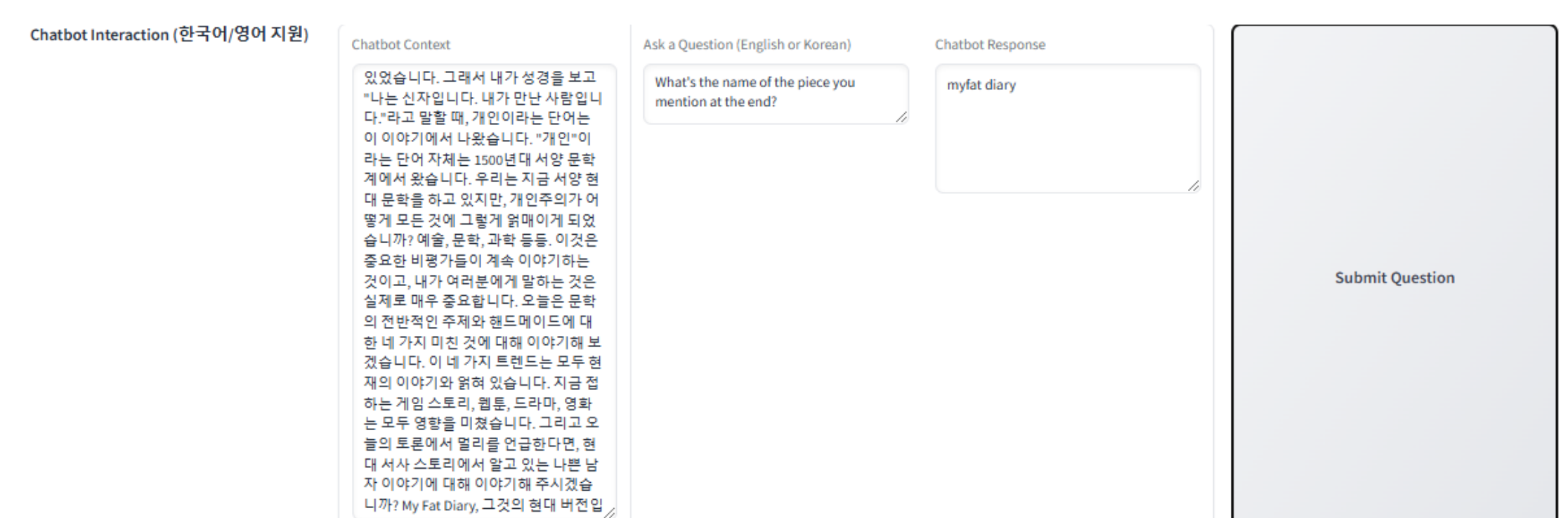
with Chatbot



STT + Summarization



with Chatbot



본 프로젝트는 텍스트 요약, STT(Speech-to-Text) 요약, 챗봇 응답 생성 기능을 성공적으로 구현하였습니다. 텍스트 요약은 PDF와 텍스트 파일의 핵심 정보를 효과적으로 추출하였고, STT 기능은 음성을 텍스트로 변환 후 요약하여 활용도를 높였습니다. 또한, 챗봇은 한국어와 영어를 지원하며 문맥에 기반한 정확한 답변을 제공하였습니다.

오픈소스 URL

https://github.com/Chae-sang-jun/2024-2_Ajou_Softcon_Self_Project.git

2. 음성 파일 처리 및 요약 기능

음성 파일을 텍스트로 변환하고 요약하는 기능을 구현하였습니다.

- 음성 인식: OpenAI의 Whisper 모델(base 버전)을 사용하여 한국어를 포함한 다양한 언어의 음성을 텍스트로 변환하였습니다.

- 다국어 지원: 변환된 텍스트가 한국어인 경우, 영어로 번역 후 요약을 진행하고, 다시 한국어로 번역하여 최종 결과를 제공하였습니다.

3. 질문 응답(Chatbot) 기능

문서를 기반으로 질문에 답변하는 챗봇을 구현하였습니다.

- 질문 응답 모델: Hugging Face의 distilbert-base-cased-distilled-squad 모델을 활용하여 원본 문서 내용에서 질문의 답변을 추출하였습니다.

- 다국어 지원: 질문의 언어를 감지한 뒤, 한국어 질문은 영어로 번역하여 처리하고, 답변을 다시 한국어로 번역하여 제공하였습니다.

4. 사용자 인터페이스 설계

Gradio를 활용하여 사용자가 직관적으로 접근할 수 있는 인터페이스를 구축하였습니다.

- 텍스트 요약: 텍스트 및 PDF 파일 업로드, 원문 및 요약문 표시, 요약문 저장 기능을 제공합니다.

- 음성 요약: 음성 파일 업로드, 변환된 텍스트와 요약문 제공, 저장 기능을 지원합니다.

- Chatbot 인터페이스: 요약 결과를 바탕으로 사용자가 질문을 입력하면 문서 내용에 기반한 응답을 제공합니다.

