1. 서비스 개요

서비스명: RAG 기반 학습 보조 챗봇
목표: 사용자가 제공한 PDF 학습자료를 기반으로 질문에 대한 정확하고 신뢰성 있는 응답과 학습 지원 기능을 제공.

주요 기능:

1. **PDF 데이터 처리:** PDF 텍스트를 추출, 전처리, 임베딩 처리.
2. **질의응답:** 사용자의 질문에 대해 적합한 문서를 검색 후 응답 생성.
3. **예시 문제 생성:** 학습 주제와 관련된 실전 연습 문제 제공.

---

2. 데이터 처리 및 파이프라인 설계

2.1 데이터 전처리

- **PDF 로드 및 텍스트 추출:**
  - Python `PyPDF2`, `pdfminer` 와 `Upstage document parse` 등을 사용하여 PDF에서 텍스트를 추출.
- **텍스트 분할 (Chunking):**
  - 권장 Chunk Size: 500-1000 토큰.
  - Overlap Size: 10~20%로 설정하여 문맥 유지.

2.2 임베딩 처리

- **임베딩 모델:**
  - Upstage Solar embedding
- **벡터 데이터베이스:**
  - Pinecone 사용, 텍스트 임베딩을 벡터 형태로 저장 및 검색.

2.3 검색 및 질의응답

- **Dense Retriever:**
  - 검색 알고리즘으로 Cosine Similarity 사용.
- **Reranker:**
  - CrossEncoder Reranker로 검색된 문서의 순위 재조정.

2.4 응답 생성 및 평가

- **LLM (GPT-4o-mini):**
  - 검색된 데이터를 기반으로 질의응답.
- **평가:**
  - 답변의 신뢰도를 `Upstage Groundness Check API`로 평가.

---

3. 개발 로드맵

단계	주요 작업 내용
데이터 준비	PDF 텍스트 추출 및 전처리

단계	주요 작업 내용
임베딩 및 데이터베이스 구축	Pinecone 설정 및 데이터 임베딩 저장
검색 알고리즘 개발	Dense Retriever 및 Reranker 구현
LLM 응답 생성	질문 응답 생성 및 평가
UI/UX 설계 및 배포	React 기반 입력창, 출력창 구현 및 통합 테스트 진행

---

4. 성능 평가 및 결과

정량 평가 (RAGAS 지표)

평가지표	설명
Context Precision	검색된 문서 중 실제 관련 문서의 비율
Context Recall	실제 관련 문서 중 검색 성공 비율
Faithfulness	생성된 답변이 문서와 일치하는 신뢰도

정성 평가

1. **정확성:** 생성된 답변이 문서 내용과 얼마나 일치하는가?
2. **관련성:** 검색된 내용이 주어진 질문과 얼마나 관련 있는가?
3. **명확성:** 답변이 명확하게 이해 가능한가?

---

5. 결론 및 향후 발전 방향

결론:

RAG 기반 학습 보조 챗봇은 학습 자료를 기반으로 높은 정확도의 질문 응답을 제공하며 학습 효율을 크게 개선합니다.

향후 발전 방향:

1. **기능 확장:** 더 다양한 데이터 형식 지원 (HTML, DOCX 등).
2. **모델 최적화:** Sparse-Dense 혼합 검색 도입.
3. **자동화 강화:** 문제 유형을 자동으로 분류하고 생성하는 기능 추가.
