[프로젝트 가이드]
RAG 기반 챗봇 서비스 구현

I 마일스톤 1

Step 1: 프로젝트 문제 정의

Step 2: 기획안 작성

Step 3: 데이터 선정

Step 1: 프로젝트 문제 정의

안녕하세요, 프로젝트를 시작하기에 앞서 **AI 서비스를 기획할 때 고려해야 할 핵심 요소**들을 함께 살펴보겠습니다. 이번 과정을 통해 여러분들이 AI 서비스를 설계하는 기본 흐름을 이해하고, 이를 바탕으로 여러분만의 AI 서비스를 기획하는 능력을 키우는 것을 목표로 합니다.

앞으로 여러분들은 **체크리스트**와 **프로젝트 예시**를 활용해 문제 정의와 이의 해결 방안을 구체화하는 방법을 알아보고, 프로젝트 과정에서 아이디어를 구체화하는 데 도움을 드리고자 합니다.

- ✅ **체크리스트**
  여러분이 서비스를 설계하는 과정에서 반드시 생각해보아야 할 질문들을 제공합니다. 이를 바탕으로 해결하고자 하는 문제를 명확히 정의하고, 서비스의 가치를 구체화할 수 있습니다.
- 🔷 **프로젝트 예시**
  해당 프로젝트 예시는 여러분들께서 3 주차에 진행하시게 될 프로젝트의 예시입니다.
  프로젝트 예시를 바탕으로 문제를 정의하고 AI 기술을 활용하여 해결하는 과정을 구체적으로 보여줍니다. 예시를 참고하여 본 프로젝트에서 여러분의 아이디어를 더욱 구체화해보세요.

1) Pain Point(문제점) 찾기
2) Expected Solution(예상 해결 방안) 도출 – AI 의 역할 정의
3) Task(태스크) 유형 정의
4) 데이터 요구사항 구체화
5) 결론

1. Pain Point(문제점) 찾기

서비스 아이디어를 구체화하기 위해서 해결하고자 하는 문제를 명확히 정의하는 것이 중요합니다.
아래 질문들을 바탕으로 고민해보세요!

- 🔷 **프로젝트 예시**
  - 어떤 불편함이나 비효율을 개선하고 싶은가?
    - 제품 매뉴얼 문서가 복잡한 문서 구조를 가지고 있어 내 원하는 기능을 찾기 어렵고, 시간이 많이 소요됨.

  - 해결하고자 하는 문제는 누구에게 영향을 미치는가?
    - 고객 : 원하는 정보를 찾기 어려움
      고객 서비스 팀 : 반복되는 질문처리로 업무 부담 증가

  - 무엇을 만들 것인가?
    - 사람이 제품 메뉴얼 문서를 모두 읽지 않아도 관련 내용을 바로 확인할 수 있도록 하는 서비스

2. Expected Solution(예상 해결 방안) 도출 – AI 의 역할 정의

AI 의 역할을 정의하고 문제 해결 방향을 구체화합니다.
해당 문제가 AI 로 해결 가능한 영역인지, AI 를 활용하면 더 효율적인지 판단합니다.

2-1/ AI 가 잘 할 수 있는 영역

AI 를 통해 가치를 더할 수 있는 영역은 비교적 명확합니다. 또한 AI 를 적용하는 것이 오히려 더 비효율적인 경우도 있습니다. 따라서 이를 미리 판단하는 것은 중요합니다. AI 가 잘하는 영역은 어떤 영역일까요?

1. **AI 를 통한 자동화**
   - 반복적이고 시간이 많이 소요되는 작업 자동화로 효율성 확보.
   - **예시**: MS 365 Copilot 의 자동 문서 생성.
2. **미래 사건의 예측**
   - 데이터를 기반으로 패턴을 학습하고 결과를 예측하여 의사결정 지원.
   - **예시**: 배달의민족 추천 배차, 쿠팡 AI 배송 솔루션.
3. **맞춤형 추천/개인화**
   - 사용자의 행동 데이터를 학습해 개인화된 경험 제공.
   - **예시**: 넷플릭스 콘텐츠 추천, Zigzag 쇼핑몰 추천 시스템.
4. **자연어 이해 및 처리**
   - 텍스트를 이해하고 요약, 분류, 번역 등 다양한 자연어 처리 작업 수행.
   - **예시**: ChatGPT, 파파고 번역.
5. **이미지/비디오 처리**
   - 이미지를 분석하고 분류하거나, 특정 정보를 추출하며, 새로운 이미지를 생성하는 작업 수행.
   - **예시**: 영수증, 진료비 세부내역서 등 문서 내 주요 데이터 자동 추출 (OCR) , X-ray/MRI 등 의료 이미지 분석을 통한 질병 진단

🔷 프로젝트 예시

- 해결하고자 하는 문제가 AI 가 해결할 수 있는 영역인가?
  - (Recap.) 해결하고자 하는 문제 : 사람이 제품 메뉴얼 문서를 모두 읽지 않아도 관련 내용을 바로 확인할 수 있도록 하는 서비스
  - 제품 매뉴얼 문서를 텍스트로 변환하는 기술은 이미지 처리의 영역임으로 AI 가 해결할 수 있음.
  - 텍스트 데이터 중 관련 내용을 파악하는 것은 자연어 이해 및 처리 파트이기에 AI 가 해결할 수 있는 영역임.
  - 또한, 제품 매뉴얼 문서에서 ***고객의 질문과 관련된 내용을 검색***하고 해당 문맥을 바탕으로 LLM 이 답변을 생성하기에 RAG 파이프라인을 통해서 해결 가능함.
- 해당 문제에 적용가능한 AI 알고리즘, 모델, 기술이 있는가?
  - Document AI : 제품 매뉴얼을 텍스트로 변환
  - LLM : 질문을 이해하고 관련 내용을 검색한 후 답변 생성
  - RAG 파이프라인 설계 예정 (2 주차)
- 해당 AI 기술이 현재 문제를 해결하는데 적합한가?
  - Document AI 는 매뉴얼 문서를 텍스트 데이터로 변환
  - RAG 는 해당 문서들 중 고객의 질문과 관련된 내용을 검색함.
  - 위 검색된 문맥을 바탕으로 LLM 은 질문에 맞는 답변을 생성함.
  - 따라서, 제품 메뉴얼이 너무 길고 복잡하여 시간이 소요되는 문제가 해결

2-2/ 비즈니스 임팩트 측정

실제로 AI 기술을 현업에서 적용하고자 할 때, AI 도입이 실제 서비스, 비즈니스에 큰 임팩트가 있는 지 판단해야 합니다.

🔷 프로젝트 예시

- AI 기술을 적용하기 위한 필요한 자원과 역량이 있는가?
  - Document AI : Upstage Document Parse API 활용 가능
  - LLM : Upstage Solar Pro API 활용 가능

- AI 기술을 적용하여 얻을 수 있는 이익이 비용과 시간에 비해 충분한가?
  - (프로젝트 한정)
    - 3 주 등 프로젝트 기간 내에 충분히 가능한 지
    - API 비용 등

- 해당 서비스를 통해 사용자/기업/사회 등은 어떤 가치/이득을 얻을 수 있는가?
  - 고객: 빠르고 정확하게 정보 획득 > 기업 : 제품 사용 만족도 개선으로 인한 매출 증대 (Make Money)
  - 고객 서비스 팀 : 반복되는 질문을 처리하여 리소스 감소 (Save Money)

3. Task(태스크) 유형 정의

다양한 자연어 Task(태스크) 유형 소개

아래 설명은 LLM 이 잘 하는 자연어 태스크를 나열해둔 것입니다.

- **Q&A (질의응답)** : 제공된 텍스트나 데이터를 기반으로 질문에 대한 답변을 생성하는 작업.
  (예시) 고객 서비스 챗봇에서 제품 사용 방법에 대한 질문에 답변 제공

- **Classification (분류)** 텍스트 데이터를 미리 정의된 레이블이나 그룹으로 분류하는 작업.
  (예시) 감성 분석, 리뷰를 긍정/부정으로 분류.

- **Generation (생성)** : 주어진 입력이나 문맥을 기반으로 새로운 텍스트나 창의적인 콘텐츠를 생성하는 작업
  (예시) 광고 문구 생성, 소셜 미디어 게시물 작성

- **(필수) Chatbot (챗봇)** : 사용자의 질문에 응답하는 자동화된 대화형 시스템.
  (예시) 고객 지원 챗봇이 제품 교환 및 환불 절차를 안내.

- **Summarization (요약)** : 텍스트의 핵심 내용이나 요점을 간결하게 추출하는 작업.
  (예시) 긴 뉴스 기사를 주요 요점으로 요약.

- **Key Extraction (핵심 정보 추출)** : 텍스트에서 이름, 날짜, 금액과 같은 특정 중요 정보를 식별하고 추출하는 작업.
  (예시) 송장 데이터에서 주문 번호와 금액 추출.

- **Comparison (비교)** : 텍스트나 데이터 간의 유사점과 차이점을 식별하고 분석하는 작업.
  (예시) 여러 투자 상품의 수익률과 리스크를 비교하여 고객에게 추천

- **Translation (번역)** : 텍스트를 한 언어에서 다른 언어로 변환하는 작업.
  (예시) 여행 안내서를 한국어에서 영어로 번역.

문제 해결을 위해 어떤 자연어 태스크 유형을 선택할 지 고민하기 이전에 다시 한 번 문제와 솔루션을 되짚어봅시다.
프로젝트 예시를 통해서 구체화해보겠습니다.

- 문제 : 제품 매뉴얼이 너무 길고 복잡하여 읽고 필요한 정보를 찾는데 시간이 오래 걸림
- 솔루션: 제품 메뉴얼 문서를 **사람이 모두 읽지 않아도** 관련 **내용을 바로 확인**할 수 있도록 하는 서비스

자 이제, 내가 이 솔루션을 풀어야한다고 가정하고, 솔루션을 최대한 구체적으로 나누어봅시다.

솔루션 내 업무 프로세스 구체화

(1) 사람이 제품에 대해 궁금한 정보가 있음.
(2) 하지만, 제품 매뉴얼 문서를 사람이 읽지 않아야함
  (a) 제품 매뉴얼 문서는 이미지이므로 텍스트 데이터로 변환해야함
    > Document Parse 이용

  (b) LLM 이 텍스트 데이터를 대신해서 이해함
    > LLM 이 대신 읽음
(3) 제품 매뉴얼 문서에서 관련 내용을 찾아야함
  > RAG 파이프라인 내 검색 엔진으로 검색 가능

(4) 제품 매뉴얼에서 찾은 내용을 사람이 바로 확인할 수 있도록 전달해줘야함.
  > 사람이 바로 확인하려면, 어떤 방식으로 제공하는 것이 편할까? 어떤 태스크 유형이 좋을 지 고민해보자!

🔷 **프로젝트 예시** 해결하고자 하는 문제가 풀어야하는 업무 프로세스를 최대한 구체적으로 정의하였는가?

- 위의 업무 프로세스가 위의 자연어처리 태스크 중 어떤 타입을 통해 효과적으로 해결될 수 있는가?
  - 제품 매뉴얼에서 찾은 내용을 제공해주기 위해서는 매뉴얼 기반으로 질문에 대한 답변을 생성하는 Q&A 와 매뉴얼 기반 내용을 요약해주는 두 태스크가 가능함.
    - Q&A(질의 응답) : 사용자가 문서를 읽지 않고도 질문에 대한 답변을 생성할 수 있어야함.
    - Summarization (요약) : 매뉴얼에 대한 내용을 요약하여 제공.

- 해당 태스크 유형이 서비스에서 제공하는 주요 기능과 일치하는가?
  - 질문에 대한 답변을 직관적으로 바로 제공해주는 것이 더 빠르고 정확하게 원하는 정보를 얻기 용이하기에 Q&A task 가 적합

- **(챗봇 필수 적용) 해당 태스크 유형 들 중 사용자 경험을 효과적으로 개선할 수 있는가?**
  또한 실시간으로 질문하고 이에 대한 답변을 받기 위한 챗봇 형태로 제작해야함.
  - Chatbot (챗봇) : 사용자가 직관적으로 질문하고 답변을 받을 수 있는 경험을 제공하기 위하여 Chatbot 형태로 제공

4. 데이터 요구사항 구체화

본격적으로 서비스를 설계하기 이전에, 서비스 구현 가능성을 점검해야 합니다. 특히, 필요한 **데이터**가 준비되었거나 구할 수 있는 상태인지, 서비스와 태스크 유형에 적합한 지를 우선 확인해야 합니다. 또한, 일반적으로 기술 및 도구의 활용 가능 여부도 점검해야 하지만, 이번 프로젝트에서는 정해진 API 와 도구를 활용하기 때문에 **데이터 준비 상태**에 초점을 맞춰 확인합니다.

🔷 프로젝트 예시

원천데이터 유형 확인

- 필요한 데이터는 무엇인가? : 제품 매뉴얼 문서
- 데이터의 도메인과 특징은 무엇인가? (e.g. 의료, 법률 등) : 제조업, 기술

원천데이터 형식 확인

- 원천데이터의 형식은 무엇인가? (e.g. PDF, HTML, txt 등) : PDF 문서
- 원천데이터 형식이 태스크 유형과 맞지 않다면, 적합하게 변환할 수 있는가?
  - PDF 문서를 텍스트 데이터로 변환
  - 그 이후, QA 태스크 유형 구현을 위해 QA Pair 로 구성

데이터 소스 선정

- 데이터를 구할 수 있는 출처가 명확한가? (e.g. 제조사 웹사이트, AI Hub, Kaggle 등)
  - 제조사 웹사이트에서 문서를 웹 크롤링으로 확보
- 데이터셋을 구할 수 있는가? (e.g. 오픈 소스 활용 가능 여부 / 웹 크롤링 가능 여부)
  - 웹 크롤링을 통한 수집 가능

데이터 품질 점검

- 데이터셋이 충분히 크고 다양한가? : 다양한 문서를 포함하고 있는가?
- 데이터의 품질을 신뢰할 수 있는가? : 최신 제품 메뉴얼을 사용하는가?

5. 결론

지금까지 AI 서비스 기획에서 고려해야 할 주요 포인트를 함께 살펴보았습니다.
체크리스트와 프로젝트 예시를 활용해 문제를 명확히 정의하고, AI 를 활용한 해결 방안을 구체화해보세요.

🔷 프로젝트 예시

a. **문제 정의 :** 제품 매뉴얼 문서의 복잡한 구조로 인해 사용자가 원하는 정보를 찾기 어렵고 시간이 많이 소요됨.
b. **기대 솔루션** : 제품 매뉴얼 데이터를 기반으로 질문에 신속하고 정확한 답변을 제공하는 QA 챗봇 서비스를 통해 문제 해결.
c. **태스크 유형** : Q&A , Chatbot(실시간 대화형 응답)
d. **데이터셋** : 제품 메뉴얼 문서 데이터 수집 및 변환하여 QA Pair 데이터셋 생성

Step 2. 서비스 기획안 구체화

서비스 기획(안) 샘플	
서비스 명 및 개요	**서비스명**: 제품 맞춤형 QA 챗봇 서비스 개발 **서비스 개요**: 많은 고객이 제품 사용 중 발생하는 문제를 해결하거나 필요한 정보를 얻는 데 어려움을 겪고 있습니다. 기존 문서 기반 매뉴얼은 검색이 비효율적이고 적합한 답변을 찾기 어렵다는 한계를 가지고 있습니다. 이를 해결하기 위해 제품 메뉴얼 데이터를 검색하여 이를 바탕으로 고객 질문에 대한 정확한 답변을 생성하는 QA 챗봇 서비스를 설계했습니다. 이 서비스는 고객의 문제 해결 시간을 단축하고 만족도를 향상시키며, 제품 사용 경험을 개선하는 데 기여합니다. 제품 구매 고객, 고객 서비스 팀을 주요 사용자로 하며, 고객이 직접 챗봇을 통해 도움을 요청하거나 혹은 고객 서비스팀이 매뉴얼 검색 대신 빠르고 효율적인 QA 도구로 활용할 수 있습니다.
타겟 사용자 및 시장 분석	**타겟 사용자**: - **예상 사용자 유형**: 고객, 고객 지원 팀 - **주요 요구사항 및 사용 목적**: ● 고객: 빠르고 정확한 제품 관련 정보 획득할 수 있습니다. ● 고객 서비스 팀: 빠른 검색 및 답변 생성을 통해 더 많은 요청을 처리할 수 있습니다. **시장 분석**: 이 서비스는 단순 검색 기반 챗봇과 달리, 검색과 생성 모델을 결합하여 보다 정확하고 유연한 답변을 제공한다는 점에서 차별성을 갖습니다. 고객의 니즈에 적합한 정보를 빠르게 제공하기에 고객 맞춤형 경험을 극대화하고, 검색 효율성과 답변 생성의 품질을 동시에 충족시킨다는 점에서 강력한 경쟁력을 가집니다.
목표 및 기대효과	**서비스 목표**: - 사용자 질문에 신속하고 정확한 답변을 제공하여 고객 문제 해결 시간을 단축합니다. - 제품 매뉴얼의 가치를 극대화하여 사용자와 기업 간의 신뢰를 강화합니다. **기대효과**: - 고객은 간단한 질문 입력만으로 매뉴얼을 탐색하지 않고도 필요한 정보를 얻을 수 있습니다. 이에 따라 고객 만족도 향상으로 제품의 브랜드 충성도가 증가할 수 있습니다. - 고객 서비스 팀은 검색 시간 단축으로 더 많은 고객 요청을 처리할 수 있습니다. 따라서 고객 서비스 비용을 절감할 수 있습니다.
데이터 구성 및 활용	**원천데이터 소스**: 제품 매뉴얼 문서 **원천 데이터 형식 :** PDF 문서 파일 . **데이터 처리 방법**: ● 데이터 수집: 웹 크롤링을 통해 제품 매뉴얼 데이터 확보. ● 데이터 전처리: ○ Document Parse 를 활용하여 PDF 파일을 HTML 형태의 텍스트 데이터로 변환 ○ 태스크 유형에 맞게 제품 관련 질문-답 쌍으로 구성된 QA 데이터셋으로 변환

Step 3: 데이터 선정 및 전처리 가이드

이번 단계에서는 위 기획안과 강좌에서 배운 내용을 바탕으로 데이터를 수집 /선정하고 처리하는 과정을 배웁니다. 데이터 선정은 서비스 구현의 출발점으로, 목표와 요구사항에 부합하는 데이터를 확보하고 정제하는 과정이 중요합니다.

위 기획안을 바탕으로 데이터의 유형/형식/출처 확인이 되었다는 가정하에 본격적으로 데이터를 선정 및 수집하고 전처리해봅니다.

1) 기획안 중 데이터셋 내용 정리
2) 데이터 소스 선정
3) 데이터 처리
4) 데이터 활용 계획 점검

1. 기획안 – 데이터셋 내용 정리

- **원천데이터 소스:** 프로젝트에 필요한 데이터의 유형 정의
- **도메인**: 서비스와 관련된 데이터의 분야를 설정
    - 예시) 기술(제품 매뉴얼), 의료(의학 가이드라인), 법률(계약서 데이터), 교육(학습 자료).
- **원천 데이터 형식**
    - 예시) txt, pdf, json, csv 등.
    - *(데이터 변환 시)* **변환 필요한 데이터 형식**
      : 원천 데이터가 태스크 유형이나 서비스의 요구사항과 맞지 않을 경우, 변환이 필요한 데이터 및 데이터 형식을 정의
- **출처** : 데이터를 구할 수 있는 신뢰할 수 있는 출처를 작성
    - 예시) AI Hub, Kaggle, 제조사 웹사이트, 공공 데이터 포털
- **저작권** : 데이터를 사용할 때의 저작권 상황을 점검

2. 데이터 소스 선정

본격적으로 데이터를 수집하기 이전에 고려해야할 체크리스트를 다시 한 번 되짚어봅니다.

2-1. 데이터 소스 선정 방식

- **사용 가능한 오픈소스 데이터셋**:
  - 공개되어 있는 데이터셋을 리서치하여 선택합니다.
    - 예시) 공공데이터포털, AI Hub, Kaggle, Hugging Face, Paper With Code.
- **웹 크롤링을 통한 데이터 수집**:
  - 시간이 충분하다면, 직접 데이터를 수집하여 활용합니다.
    - 예시) 특정 웹사이트에서 크롤링한 제품 정보, FAQ 페이지 등.

3. 데이터 처리

이제 수집한 원천 데이터를 원하는 방식대로 변환 및 처리하는 방법에 대해서 알아보겠습니다.

3-1. (데이터 변환 필요 시,) 데이터 형식 변환

- 데이터를 모델 입력 형식에 맞게 변환합니다.
  예시) PDF 데이터를 텍스트로 변환, JSON 파일로 구조화.

3-2. 데이터 정제

- **작업 예시**:
  - 중복 데이터 제거.
  - 불필요한 기호 및 특수 문자 제거.
  - 오탈자 수정 및 데이터 필터링.
- **이상치 및 결측치 처리**:
  - 데이터의 품질 저하를 방지하기 위해 이상치 및 결측 데이터를 수정하거나 제거.

3-3. 데이터 통합

- 다양한 데이터 소스를 사용하는 경우, 이를 통합하여 **일관된 데이터셋**을 구축.
- 중복 항목 및 형식 불일치를 점검하여 데이터의 일관성을 유지.

3-4. 데이터 품질 점검

- 데이터의 표현 방식, 형식, 단위가 통일되었는가?
- 데이터가 원본 내용과 일치하며 오류가 없는가?

4. 데이터 활용 계획 점검

4-1. 데이터와 모델 연계

- 데이터가 **RAG 파이프라인**과 원활히 연계되도록 설계합니다.
  데이터가 검색 엔진에 인덱싱되고, 생성 모델이 검색된 데이터를 활용하여 답변을 생성하는 프로세스 입니다.

Conclusions

마일스톤 1 에서는, 서비스 목표와 데이터 요구사항을 명확히 정의하고, 필요한 데이터를 빠르게 선정합니다.
오픈소스 데이터를 적극 활용하되, 시간이 충분한 경우 웹 크롤링을 통해 고유 데이터를 확보하는 방향으로 진행하시길 권장드립니다. 서비스 기획 및 데이터 선정을 마무리하고 나면, **2 주차 RAG 파이프라인 구현 프로젝트를** 진행하기 위한 준비가 완료됩니다!

구체적인 개발환경 및 프로젝트 구조는 2 주차에 RAG 파이프라인의 구성요소, 설계 방식을 배우는 마일스톤 2 에 해당 강좌와 함께 제공될 예정입니다.