그래픽, 폰트, 그래픽 디자인, 로고이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

1. **상담사 말투 파인튜닝 데이터**

| **프로젝트 주제** | 신선식품 플랫폼 고객 응대 챗봇 | | |
| --- | --- | --- | --- |
| **데이터 출처** | AI 허브 | **데이터 형태** | jsonl |
| **데이터 수집 계기** | 상담 과정에서 말투는 매우 중요한 요소이다. 하지만 일반 llm 모델의 말투는 기계적이고 딱딱하여 실제 상담사의 어투와 조화를 이루기 어렵다. 이에 따라 fine-tuning을 통해 따뜻하고 친절하면서도 전문적인 상담사 어투를 구현하고자 한다.  이 데이터 셋은 민간 민원 상담 내용을 기반으로 LLM 사전 학습 및 Instruction Tuning을 위해 가공한 자료이다. 최소 3턴 이상의 질의/답변으로 구성된 다중턴 대화 구조를 갖추고 있어 맥락 유지, 공감 표현, 완곡한 안내 등 상담사 어투의 특성을 학습하기에 적합하다고 판단한다. | | |
| **데이터 출처** | https://www.aihub.or.kr/aihubdata/data/view.do?pageIndex=1&currMenu=115&topMenu=100&srchOptnCnd=OPTNCND001&searchKeyword=%EC%83%81%EB%8B%B4&srchDetailCnd=DETAILCND001&srchOrder=ORDER001&srchPagePer=20&aihubDataSe=data&dataSetSn=71844 | | |
| **원본 데이터** |  | | |
| **데이터 전처리 시 유의할 점** | 1. 상담과 무관한 내용 제거  * 상담 목적과 직접 관련 없는 잡담, 시스템 로그, 안내 배너 문구 등은 제거한다. 광고/홍보성 문구와 중복 공지 또한 삭제한다.  1. 질의-응답 세트화  * 고객 질문과 상담사 답변을 한 세트로 정의하고, 대화는 세트 단위로 구분한다.  1. 학습 샘플링 정책  * 장문의 대화를 여러 세트로 분할하되, 세트 내부의 발화 순서는 보존한다. 학습시에는 대화 단위를 무작위로 셔플하여 특정 흐름만 과적합되는 것을 방지한다. | | |

**2. 마켓컬리 이용약관 데이터**

| **데이터 출처** | 마켓컬리 이용약관 | **데이터 형태** | 웹, 텍스트 |
| --- | --- | --- | --- |
| **데이터 수집 계기** | 플랫폼이 안정적으로 운영되기 위해서는 내부 규정이나 이용약관과 같은 제도적 장치가 반드시 필요하다. 이러한 약관은 서비스 제공자와 이용자 간의 권리와 의무를 명확히 하기 위해 마련되며, 분쟁 발생 시 기준이 되는 중요한 역할을 한다. 또한, 이용약관은 플랫폼의 신뢰도를 확보하는 데에도 크게 기여한다.  따라서 우리 플랫폼 역시 적절한 이용약관을 마련할 필요가 있다. 이를 위해 동일하거나 유사한 서비스를 운영하는 사례를 먼저 참고하는 것이 효과적이다. 특히, 우리와 성격이 유사한 플랫폼인 ‘마켓컬리’의 이용약관을 대표 사례로 살펴보고, 이를 기초 자료로 활용하기 위해 해당 데이터를 수집한다. | | |
| **데이터 출처** | https://www.kurly.com/user-terms/agreement | | |
| **원본 데이터** |  | | |
| **데이터 전처리 시 유의할 점** | 1. 조항 번호 정규화  * ‘제 1 조’와 같이 공백에 섞인 표기를 ‘제1조’로 통일한다.  1. 본문 구조화  * content를 ‘조’ 단위로 분해하여 구조화 한다. ‘항-호-목’은 ‘조’에 대한 부연설명으로 판단한다.  1. 범위 외 항목 정리 및 요약  * 우리 프로젝트와 무관한 조항은 삭제한다. 과도하게 긴 조항은 핵심 의미를 유지하는 범위에서 요약한다. | | |

**3. 마켓컬리 FAQ 데이터**

| **데이터 출처** | 마켓컬리 FAQ | **데이터 형태** | 웹, 텍스트 |
| --- | --- | --- | --- |
| **데이터 수집 계기** | 챗봇은 고객의 반복 질문을 정확하고 신속하게 답변하는 것이 핵심이므로, 고객이 자주 묻는 질문을 사전에 파악하는 작업이 필요하다. 그러나 현재는 서비스 초기 단계로 자체 데이터만으로는 고객의 니즈와 질문 패턴을 충분히 파악하기 어렵다.  이 한계를 보완하기 위해 유사 서비스의 공개 자료를 활용하고, 그 중 실제 고객 문의가 체계적으로 정리된 ‘마켓컬리 FAQ’를 기초 데이터로 수집 및 분석 한다. 이는 실제 고객의 빈번한 질문과 표준화된 답변이 체계적으로 정리 되어 있어, 챗봇이 참고하기에 적합한 데이터 소스로 판단한다. | | |
| **데이터 출처** | https://www.kurly.com/m2/service/faq.php | | |
| **원본 데이터** |  | | |
| **데이터 전처리 시 유의할 점** | 1. 플랫폼 통일  * ‘컬리’, ‘마켓컬리’ 등 원문 표기는 프로젝트의 가상 플랫폼명으로 일괄 치환한다.  1. 중복 질문 제거  * 동일하거나 유사한 질문을 통합하고 대표 질문만 유지한다.  1. 답변 보완  * 빈 답변이거나 불완전한 문장은 맥락에 맞게 보완한다. 근거가 불분명한 내용은 정정하거나 제거한다. 과도하게 긴 답변은 핵심 정보 위주로 간결화한다.  1. 채널 적합화  * 버튼/메뉴 경로 등 웹 UI 중심 표현은 챗봇 기준으로 재서술한다.  1. 범위 외 항목 삭제  * 본 프로젝트와 무관한 내용을 삭제한다. | | |