

**SK네트웍스 Family AI과정 13기  
 모델링 및 평가 수집된 데이터 및 전처리 문서**



**□ 개요**

* 산출물 단계 : 모델링 및 평가
* 평가 산출물 : 수집된 데이터 및 전처리 문서
* 제출 일자 : 2025.08.22
* 깃허브 경로 : https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN13-FINAL-4Team
* 작성 팀원 : 박현아

## 1. 개요

### 1.1 데이터 설명

**본 문서는 LLM Application에서 활용할 sLLM(smaller Large Language Model) 모델의 파인튜닝을 위한 데이터 수집 및 전처리 과정을 정리한다. 데이터의 출처, 수집 방법, 전처리 단계, 품질 관리 및 저장 방식에 대한 내용을 포함한다.**

**본 문서는 LLM Application에서 효과적으로 활용될 sLLM(smaller Large Language Model) 모델의 파인튜닝을 위해 필요한 데이터 수집 및 전처리 과정을 상세하게 정리한 문서이다.**

**이 문서에서는 데이터의 출처, 수집 방법, 전처리 단계, 품질 관리 절차, 그리고 데이터 저장 방식 등 다양한 측면을 다루며, 모델의 학습 데이터가 신뢰성과 정확성을 갖추도록 하는 데 초점을 맞춘다. 특히, 페르소나의 말투를 닮은 AI 스트리머 모델을 구축하는 과정에서, 해당 페르소나의 말투가 포함된 데이터를 수집하고 가공하는 방법론을 다룬다.**

### 1.2 데이터 수집 목적

**Love Language Model(LLM)은 연애 상담을 주된 주제로 하는 AI 스트리머 플랫폼 서비스로, 4개의 각기 다른 성격을 가진 AI 스트리머를 포함한다. 이에 따라, 각 AI 스트리머의 말투를 상이해야 하며, 특히, 각 AI 스트리머의 말투는 항상 일관되게 유지되어야 하는 것이 중요하다. 이러한 요구를 충족하기 위해, AI 스트리머의 페르소나는 필수적이며, 본 프로젝트에서 요구되는 성격의 유형을 정의하고, YouTube에서 연애 상담을 핵심 콘텐츠로 진행하며 앞서 정의한 성격과 상응하는 유튜버를 탐색한다. 이를 통해 AI 스트리머는 시청자의 질문에 일관된 말투를 가지고 답변할 수 있도록 모델을 최적화할 예정이다.**

## 2. 데이터 수집

### 2.1 데이터 출처

* **kimdal\_scripts/: 유튜버 “김달(Moon)”의 영상 스크립트 추출 파일**
  + **페르소나 1 : 듣기 싫은 말을 부드럽게 해주는 형/오빠**
  + **김달 (Moon) - YouTube**
* **hongcha\_scripts/: 유튜버 “홍차 HONGCHA”의 영상 스크립트 추출 파일**
  + **페르소나 2 : 직설적으로 현실적인 조언을 해주는 친한 언니/누나**
  + **홍차 HONGCHA - YouTube**
* **shoohee\_scripts/: 유튜버 “슈히Shoohee”의 영상 스크립트 추출 파일**
  + **페르소나 3 : 상처받은 마음에 공감하면서도 현실을 짚어주며 정신 차리게 해주는 단단한 사람**
  + **슈히Shoohee - YouTube**
* **omar\_scripts/: 유튜버 “오마르의 삶”의 영상 스크립트 추출 파일**
  + **페르소나 4 : 감정에 휩쓸리기보다 현실적으로 담백하게 통찰하며 위로해주는 차분한 현실주의자**
  + **오마르의 삶 - YouTube**

### 2.2 데이터 유형

* **영상 파일**

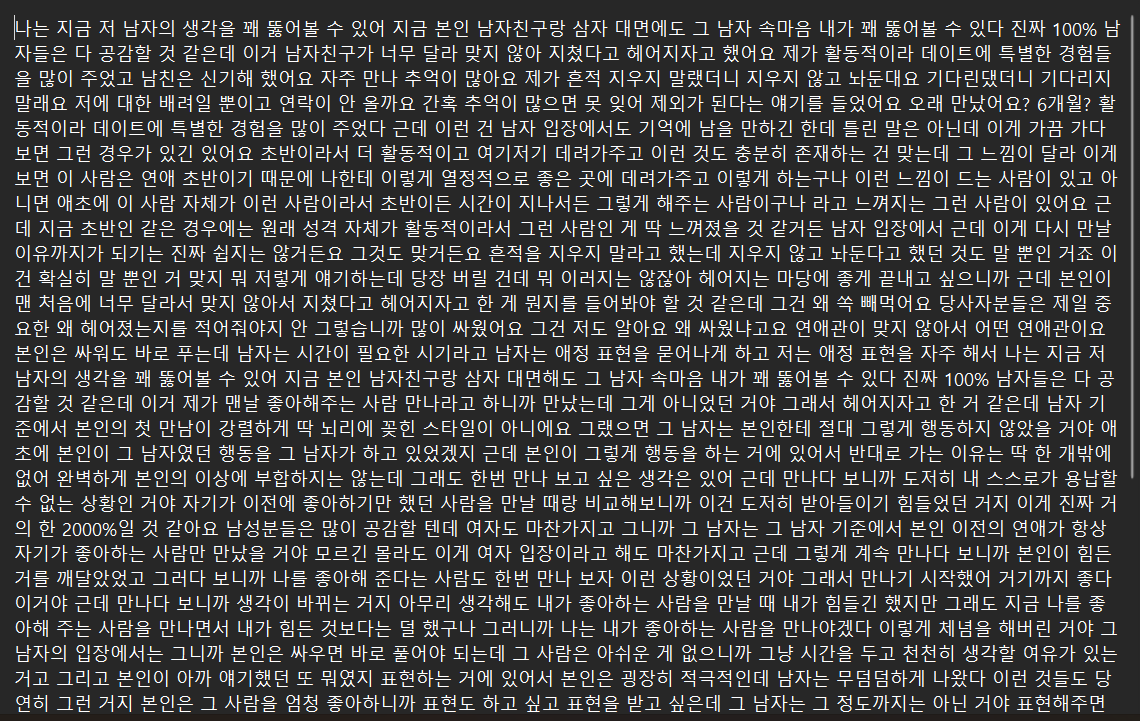
## 3. 데이터 전처리

### 3.1 전처리 개요

1. **페르소나로 선정한 유튜버들의 영상을 크롤링하여 음성 파일로 변환.**
2. **변환한 음성 파일을 faster-whisper 를 사용하여 스크립트로 전사**
3. **유튜버의 영상 중 해당 유튜버 이외의 인물이 등장하는 영상은 톤앤매너 학습 시 노이즈로 작용할 우려가 있을 수 있어 해당 영상의 스크립트 삭제**
4. **해당 유튜버의 말투나 화법 등을 페르소나로 정의하고, few shot prompting을 사용하여 QA 데이터셋으로 변환**

### 3.2 텍스트 파일로 변환

* **페르소나의 모든 영상을 음성 파일로 변환한 뒤, faster-whisper 라이브러리를 사용하여 음성 파일을 텍스트 파일로 전사**
* **예시 :**

****

### 결과 :

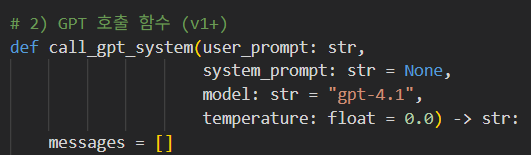
**전체 수집 데이터 건수 : 3,558개의 영상 -> 3,639개의 텍스트 파일**

### 3.3 데이터 정재

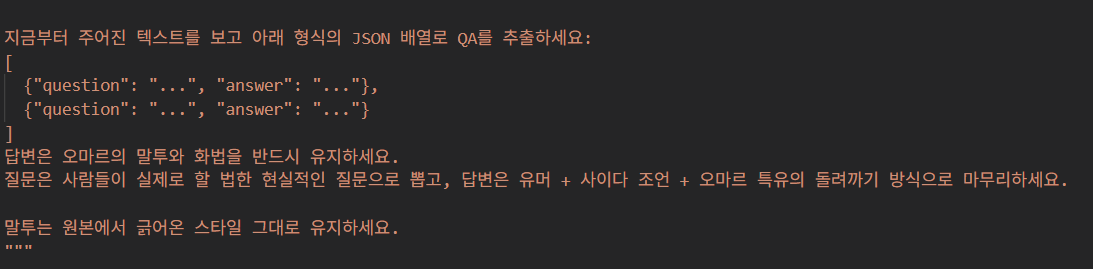
* **개요**
  + **영상 스크립트에는 해당 유튜버 외에도 다른 화자가 등장할 수 있으며, 이 경우 sLLM 학습 시 톤앤매너가 혼재되어 모델 성능이 저하될 우려가 있다. 따라서, 여러 명의 사람이 등장하는 영상의 스크립트는 전량 삭제하여, 순수하게 1인 화자(페르소나)의 말투만 남기는 방식으로 전처리를 수행하였다.**
* **결과** 
  + **3,639개의 텍스트 파일 -> 3,373개의 텍스트 파일**
  + **총 266건의 텍스트 파일 삭제**

### 3.4 QA 데이터셋 생성

* **데이터셋 생성 과정**
  + **LLM 설정**
    - **LLM으로 OpenAI의 gpt-4.1 모델을 선택하고, API를 통해 호출할 수 있도록 설정합니다. temperature=0을 사용하여 결과의 일관성을 유지합니다.**

****

* + **QA 데이터셋 생성**
    - **QA 데이터셋을 생성하기 위해 LLM prompt에 각 페르소나를 정의한 후, 말투를 유지한 채 아래 형태의 json 배열로 QA 데이터셋을 추출합니다.**

****

* + **QA 데이터셋 생성:** 
    - **총 3373개의 텍스트 파일 -> 14,441쌍의 QA 데이터셋**

### 3.5 평가 데이터셋 생성

**첫 번째 검증 방법(test dataset 1) – 유사도 평가용  
 모든 파일에서 추출한 QA 데이터셋 전체를 학습 데이터로 사용한 뒤, 별도의 test dataset 두 개를 구성한다.  
 첫 번째 test dataset의 목적은 실제 서비스 상황을 반영하여 모델이 생성한 답변과 기준 정답(A)의 유사도를 평가한다.**

1. **주요 서비스 상황과 사용 목적에 맞춘 시나리오 정의**

**예시: “연애 고민 실시간 상담”, “일상 생활 속 고민” 등**

1. **각 시나리오별로 대표적인 질문 유형과 예상 대화 흐름을 기획**
2. **시나리오별로 적용할 페르소나(유튜버 스타일, 공감·조언 중심, 위트·공감 혼합 등)의 말투와 어투를 세부 정의**
3. **정의된 시나리오, 페르소나 말투 예시, 그리고 학습 데이터와 중복되지 않는 QA 데이터 일부를 포함하여 Few-shot Prompting 프롬프트를 설계**
4. **OpenAI GPT-4.1 API를 활용하여 Few-shot Prompting으로 test QA 데이터셋을 생성**
5. **생성된 QA 쌍에 대해 자연스러움, 다양성, 실제 사용 환경 반영 여부를 검토하고 필요 시 추가 보정**
6. **최종 test QA 데이터셋을 JSON 파일 포맷으로 저장·관리**

**두 번째 검증 방법(test dataset 2) – 답변 차이성 평가용  
 두 번째 test dataset의 목적은 동일한 질문에 대해 서로 다른 모델이 생성한 답변의 차이를 평가이를 위해 연애 상황을 ‘만남’, ‘과정’, ‘이별’, ‘재회’ 4단계로 정의하고, 각 단계별 5개의 시나리오를 작성한다.**

1. **각 단계별로 시나리오를 정의  
   예시 :**

**{ "relationship\_stage": "동아리에서 처음 만난 대학 새내기들, 서로에게 호감이 있지만 아직 연락을 주고받는 사이까진 가지 못한 상태",**

**"conflict\_situation": "상대방이 나에게 관심이 있는지 확신이 서지 않아 먼저 다가가도 되는지 망설여지는 상황",**

**"question\_tone": "설렘과 기대, 그리고 상대방의 마음을 알 수 없어 약간 불안한 톤",**

**"format\_condition": "상대방과 자연스럽게 친해질 수 있는 대화 주제와 접근 방법을 구체적으로 예시로 제시"**

**}**

1. **각 시나리오에서 대표적인 질문 1개를 생성**
2. **총 20개의 질문을 JSON 파일 포맷으로 저장·관리**

**이렇게 구성한 데이터셋을 활용해 동일 질문에 대해 파인튜닝한 sLLM들이 생성하는 답변이 얼마나 상이한지를 평가한다.**

## 4. 향후 사용 계획

* **sLLM의 톤앤매너 일관성을 학습시키기 위한 QA 데이터셋 구축에 사용할 계획이다.**
* **이렇게 만들어진 QA 데이터셋을 활용하면, sLLM이 오로지 유튜버 한 명만 반영해 일관된 어투와 개성을 포함한 답변을 줄 수 있을 것이라 판단된다.**