SK Networks Family AI Camp 9기

**데이터 수집 결과서**

가로선

**프로젝트 명: DAIS (Divergent AI with Science)**

**담당자: 조이현**

# <목차>

1. 데이터 수집 목적

2. 데이터 수집 출처

3. 수집 데이터 요약

4. 수집 데이터 형식

# 

# 

# 1. 데이터 수집 목적

과학 AI 인플루언서 DAIS의 지식 기반 구축을 위해, 유튜브 채널에서 다음과 같은 스크립트(자막 및 설명 텍스트)를 확보하였음. 각 채널 선정 이유는 과학 관련 주제를 친근하고 이해하기 쉽게 풀어주는 특화된 콘텐츠를 제공하며, AI 모델에 다양한 설명 스타일·어투·난이도를 학습시키는 데 최적의 소스라고 판단하였음.

가. 실제 인터뷰·방송 콘텐츠 기반 Q&A 패턴 구축: 방송·인터뷰·온라인 강연 등에서 사용된 질문과 답변을 그대로 재현하여, DAIS의 대화 자연성을 높이고자 함.

나. 과학적 레퍼런스 확보: 과학 콘텐츠에서 핵심 지식을 추출하여 모델의 사실 기반 응답 품질을 강화하기 위함.

다. 다양한 질문 유형·난이도 커버리지 확장: 초보자용 기초 질문부터 심화 학습용 고난이도 해설까지 폭넓은 범위의 Q&A 샘플을 준비하여, 모든 사용자 수준을 만족하기 위함.

라. 컨텍스트 전환 및 대화 흐름 학습: 동일 주제 내 세부 질문 간 맥락 전환 패턴, 다중 턴 대화 예시를 포함하여, 자연스러운 대화 상태 전이를 모델이 익히도록 하고자 함.

마. 페르소나 몰입도 향상: 수집한 언어 표현과 어투 데이터를 통해 ‘친근하면서도 전문적인’ AI 인플루언서 캐릭터를 일관되게 학습시키기 위함.

# 

# 2. 데이터 수집 출처

| 채널명 | 핸들(Username) | 주요 콘텐츠 특징 |
| --- | --- | --- |
| 안될과학 | @Unrealscience | 실험·오류 사례 중심의 ‘실패하는 과학 실험’ 시리즈 |
| 에스오디 | @softdragon | IT·소프트웨어 원리부터 최신 기술 동향까지 폭넓은 과학 이야기 |
| 우주먼지의 현자타임즈 | @wz\_mz | 우주 물리·천문학 주제의 심화 강의 형식 콘텐츠 |
| 지식인미나니 | @iamminani | 일상 속 과학 원리를 만화·애니메이션으로 풀이 |
| 지식은 날리지 | @jisikisknowledge1620 | 퀴즈·문제 풀이를 통해 개념을 점검하는 ‘과학 퀴즈쇼’ 형식 |
| 보다 | @보다BODA | 최신 과학 뉴스·연구 성과를 쉽고 빠르게 정리하는 브리핑 |
| 과학드림 | @ScienceDream | 초중고 교과개념부터 대학 수준 심화까지 단계별 커리큘럼 제공 |
| 한눈에 보는 세상 | @kurzgesagt\_kr | 고퀄리티 애니메이션으로 과학·철학 주제를 시각적으로 설명 |
| 1분과학 | @1minscience | 1분 내외 짧고 강력한 ‘과학 핵심 요약’ 콘텐츠 |
| 과학쿠키 | @snceckie | 생활 밀착형 과학 실험·DIY 콘텐츠 |
| 긱블 | @geekble\_kr | 팟캐스트 스타일 대담과 심층 인터뷰로 과학 이슈를 다룸 |
| 이준석 | @junseoktube | 과학 정책 및 대한민국 기술 트렌드 분석 |

# 

# 3. 수집 데이터 요약

가. 채널 수: 총 12개 채널

나. 영상 수: 채널별 업로드 영상에서 수집된 전체 공개 영상 8,500여 편

다. 스크립트 샘플 수: JSONL 레코드로 환산 시 6,000여 건

라. 평균 제목 길이: 약 8 토큰 (한글 기준 약 20~30자)

마. 평균 자막 길이: 약 1,200 토큰 (한글 기준 약 1,500~2,000자)

바. 언어: 한국어 자막(자동 생성·수동 업로드 혼합)

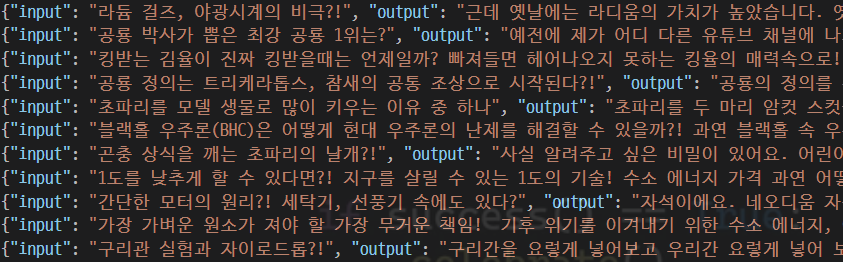
사. 수집 기간: 2025년 7월 28일 수요일 - 2025년 07월 29일 목요일

# 4. 수집 데이터 형식

가. 파일 형식: UTF-8 인코딩된 .jsonl (JSON Lines)

나. 파일 형식 개요: 영상 제목(최대 60토큰), 자막 전체 텍스트(최대 2,000토큰)

다. 파일 형식 예시:



이와 같은 형식으로 학습 데이터셋을 구성할 경우, 파인튜닝 시에는 input/output 페어로 학습 가능하며, RAG 또는 Vector DB 활용으로 용이한 장점이 있음. 이를 통해 두 가지 워크플로우를 모두 지원하는 유연한 구조를 갖추고자 함.