S

**SK네트웍스 Family AI 과정 12기  
데이터 수집 및 저장 데이터 조회 프로그램**



| **산출물 단계** | 데이터 수집 및 저장 |
| --- | --- |
| **평가 산출물** | 데이터 조회 프로그램 |
| **제출 일자** | 2025. 07.11. |
| **깃허브 경로** | <https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN12-FINAL-3TEAM> |
| **작성 팀원** | 이준석, 황차해 |

| **원본 데이터**  **(텍스트)** | - 데이터 유형: AMI/ICSI 회의 대화 데이터, 국회 위원회 회의록  - 원본 포맷: CSV 또는 TXT  - 구조 문제:  - 화자(speaker), 발화 시각(timestamp), 발화 내용(text)이 구분 없이 섞여 있음  - 줄마다 포맷이 상이하고, 시간/화자 누락 존재  - UTF-8 BOM 인코딩으로 필드 누락 가능성  - 전처리 필요 이유:  - 모델 학습에는 구조화된 JSONL 필드가 요구되므로, 재구성이 필수 |
| --- | --- |
| **데이터 용도** | sLLM 파인 튜닝 및 검증 |
| **데이터 전처리**  **과정** | - 입력: 원본 CSV 또는 TXT  - 출력: JSONL 형식 {"timestamp": "", "speaker": "", "text": ""}  - 전처리 도구: Python + re (정규표현식)  - 핵심 로직:  - 화자/시각/발화를 정규표현식으로 추출  - 빈 text나 speaker는 삭제 |
| **검증 코드 및**  **프롬프트 요약** | **Summarize\_conversation 메서드**  system\_message = """당신은 전문 회의록 분석가입니다. 회의록을 실무진이 활용할 수 있는 상세한 기획안 형태로 요약합니다. 한국어로 답변하세요."""  user\_prompt = f"""다음 회의록을 분석하여 기획안 형태로 요약해주세요:  파일: {filename}  참여자: {', '.join(stats['speakers'])} ({stats['unique\_speakers']}명)  총 발화 수: {stats['total\_utterances']}  회의 내용:  {conversation\_text}  다음 형식으로 상세히 요약해주세요:  ## 회의 개요  \*\*회의명:\*\* [회의 제목]  \*\*일시:\*\* [회의 일시]  \*\*참석자:\*\* [주요 참석자와 역할]  \*\*주요 안건:\*\* [논의된 핵심 안건들]  ## 주요 논의사항  ### 1. [핵심 안건 1]  - 논의 내용: [구체적 내용]  - 주요 의견: [참석자들의 의견]  - 결론: [도출된 결론이나 방향]  ### 2. [핵심 안건 2]  - 논의 내용: [구체적 내용]  - 주요 의견: [참석자들의 의견]  - 결론: [도출된 결론이나 방향]  ## 주요 결정사항  ### 확정된 사항  - [구체적 결정 내용들]  ### 보류/검토 사항  - [추후 검토가 필요한 사항들]  ## 후속 조치  - [다음 회의 일정]  - [담당자별 액션 아이템]  - [관련 부처 협조사항]  ## 핵심 키워드  - [회의에서 중요하게 다뤄진 키워드들]  모든 내용을 회의록 원문을 바탕으로 구체적이고 정확하게 작성해주세요.""" |
| **검증 결과** | **실무 제출용 기획안**으로는 **GPT-4o 버전**이 더 적절.  예산 계획, 위험 요소 등 실무 구성 요소가 있으나  Hallucination (환각 현상)이 포함  기획안의 틀은 부족하지만 회의록처럼  **내용 검토 및 회의 흐름 이해**에는 **RunPod 요약**이 더 적절 |
| **추후 개선 방안** | - 프롬프트 최적화  - Few-shot Prompting 적용 |