터 

**SK네트웍스 Family AI 과정 12기  
 모델링 및 평가 수집된 데이터 및 전처리 문서**



| **산출물 단계** | 모델링 및 평가 |
| --- | --- |
| **평가 산출물** | 수집된 데이터 및 전처리 문서 |
| **제출 일자** | 2025.07.17 |
| **깃허브 경로** | <https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN12-FINAL-3TEAM> |
| **작성 팀원** | 이준석 |

1. **데이터 수집 개요**

데이터 수집 기간: 2025.06.30 ~ 2025.07.07

수집 방식: 수동 다운로드

1.1. 수집 목적

수집 목적: 회의 중 발생하는 음성 데이터를 기반으로, 자동 기획안 생성 및 업무 분배 기능을 수행하는 AI Agent 모델을 개발하는데 있어 파인튜닝용으로 사용하기 위함.

1.2 데이터 출처

| **데이터명** | **출처** | **수집 방식** | **형식** |
| --- | --- | --- | --- |
| 국회 상임·특별위원회 회의록 | 대한민국 국회 정보공개포털 | 수동 다운로드 + Python 전처리 스크립트 | JSONL |
| AMI 다자 대화록 (한글 번역) | The University of Edinburgh (공개데이터셋) | 수동 다운로드 + XML 파싱 및 정제 | JSONL |
| ICSI 다자 대화록 (한글 번역) | The University of Edinburgh (공개데이터셋) | 수동 다운로드 + XML 파싱 및 정제 | JSONL |

1.3 수집 대상 및 범위

* 국회 회의 발언 텍스트 (상임위원회 및 특별위원회 회의록)
* AMI/ICSI 다자 회의 발화 데이터 (다자간 자연스러운 대화 포함)
* 약 280건의 회의 세션 (국회 52건, AMI 154건, ICSI 75건),
* 총 발화 수 약 233,000건 이상

개인정보 포함 여부: 아니오 (공식 회의체 발언 및 연구용 공개 데이터셋만 포함됨)

* 민감 정보 여부 및 조치 사항: 포함되지 않음 (국회 회의록의 경우 이름, 직책 등은 공공 정보이며, 주민번호/연락처 등 민감 정보는 없음 (공개 문서 기준, 별도 익명화 불필요))

1.4 법적·윤리적 고려 사항

* 개인정보보호법 준수 여부: O
* 데이터 수집 사전 동의 유무: O (공개 회의록 및 연구 데이터로, 별도 동의 불필요)
* 자동 크롤링 시 robots.txt 확인 여부: O (국회 포털은 수동 다운로드 기반, AMI/ICSI는 수동 수집)
* 수집 데이터의 활용범위 명시 여부: O (AI 모델 학습 및 Slack App 시스템 개발에만 활용)

**2.저장 방식**

* 저장 경로:

/data/raw\_data/제22대국회 제426회(임시회) 제2차 과학기술정보방송통신위원회(정보통신방송법안심사소위원회) (2025.07.02.)\_split.jsonl

/data/raw\_data/TS3007d.jsonl 형식

* 저장 형식:
* JSONL (Line-delimited JSON) 형식으로 저장
* 각 줄마다 { "timestamp": ..., "timestamp\_order": ..., "speaker": ..., "text": ... } 구조 유지
* 일관성 확보 방법:
* 스크립트에서 검사수행 (timestamp, speaker, text 필드 존재 여부)
* null, 빈 문자열, BOM 인코딩 오류 등 필드 누락 검출 및 삭제 처리
* 텍스트에 특수기호나 기타 내용 외 부분은 전처리 필터링

2.1 중복 및 정합성 검증

* 중복 제거 기준: 동일한 "timestamp + timestamp\_order + speaker + text" 3요소가 중복되는 경우 제거
* 국회 회의록의 경우 회차가 중복된 문서 제거
* AMI/ICSI의 경우 동일 발화 반복 저장 시 제거
* 정합성 확보 방법:  
  + timestamp 필드는 HH:MM 또는 [HH:MM] 형식으로 통일
  + timestamp\_order 필드는 I-I 형태로 통일
  + speaker가 없는 발화 또는 "text": ""인 항목은 제거
  + "text" 항목이 "@", ".", "-" 등 의미 없는 문자만 포함한 경우 제거

**3. 데이터 전처리 절차**

3.1 이상치 탐지 및 처리

* 이상치 기준:
* 화자 구분 누락 혹은 오류로 인해 동일 타임스탬프에 중복 발화가 존재하는 경우
* 발화 내용이 기호(@, ., 공백 등)만 포함된 경우
* 화자의 실제 발언이 아닌 시스템 메시지(예: "회의 개회", "산회")
* 처리 방법:  
  + "text": "", "@@", "." 등 유효하지 않은 텍스트 제거
  + 시스템 메시지 등 발언 내용 외엔 전부 제거
  + 중복 발화 등 제거
* 품질 기여:
* "text": "", "@@", "." 등 유효하지 않은 텍스트들을 제거하여 데이터 품질 상승

3.2 결측치 처리

* 결측 필드:
* speaker, text, timestamp, timestamp\_order 중 하나 이상 누락된 경우
* XML -> CSV 파싱 때 단어 ID 누락 (결측치는 아니나, 구조적으로 누락되는 부분)
* 처리 방법:
* text가 비어 있거나 "text": ""인 경우 → 해당 발화 제거 후 로그 기록
* speaker 누락 시 → 제거 후 로그 기록
* XML -> CSV 파싱 때 단어 ID 누락 등 → 별 다른 처리 없이 로그에만 기록
* 품질 기여: 누락된 데이터들을 제거함으로써 데이터 품질 상승

3.3 데이터 변환 및 재현 가능성

* 텍스트 정제:  
  + 불용 기호 제거 (예: [ ], ◯, 기호-only 문장)
  + 공백·개행 문자 정리
  + 발화 단위로 줄바꿈 기준 정제
  + 빈 문자열 및 의미 없는 발화 제거 (예: text == "" 또는 "@", ".", "..." 등만 포함된 경우)
* 토큰화:  
  + 별도 형태소 분석기 없이, 전체 텍스트를 LLM 입력용으로 구성
  + Qwen3 기반 모델 입력 구조에 맞추어 speaker: text 형태 유지
* 변환 코드 예시:

import re

def clean\_text(text):

text = re.sub(r'[\s◯※…"]+', ' ', text)

text = text.strip()

if text in ["", ".", "@"]:

return None

return text

* GitHub 또는 Colab 링크:

3.4 전처리 절차 요약 (프로세스 플로우)

1. **데이터 수집:** 국회 회의록 PDF, AMI XML, ICSI CSV 수동 다운로드
2. **텍스트 변환:** PDF → TXT (국회), XML→ CSV (AMI/ICSI)
3. **기초 정제**

* BOM 오류 보정
* 발화 필드 누락 보정

1. **텍스트 클렌징**

* 특수기호 제거 (◯, ※, ., @, ...)
* 빈 문자열/비문 제거

1. **구조화 변환:** timestamp, timestamp\_order, speaker, text 필드 기준 JSONL 구조화

5-1. (AMI, ICSI의 경우) 구글 번역 API를 이용한 Text 번역

1. **불완전 발화 제거:** 길이 짧거나 기호만 포함된 발화는 필터링
2. **데이터 저장:** 파일명 기준 분리 저장
3. **최종 검토:** 화자 누락/이상 행 → 확인 후 삭제 처리

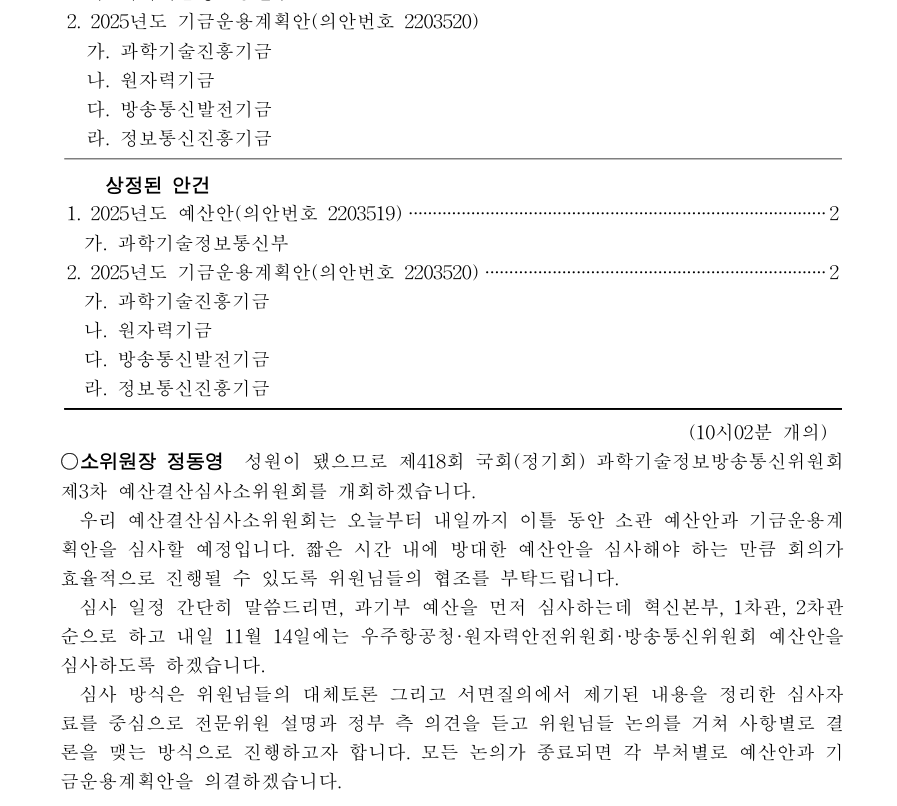
* 학습/검증/테스트 비율:

| 구분 | 건수 | 비율 |
| --- | --- | --- |
| Train | 8,000 | 64.5% |
| Validation | 2,000 | 16.1% |
| Test | 2,400 | 19.3% |

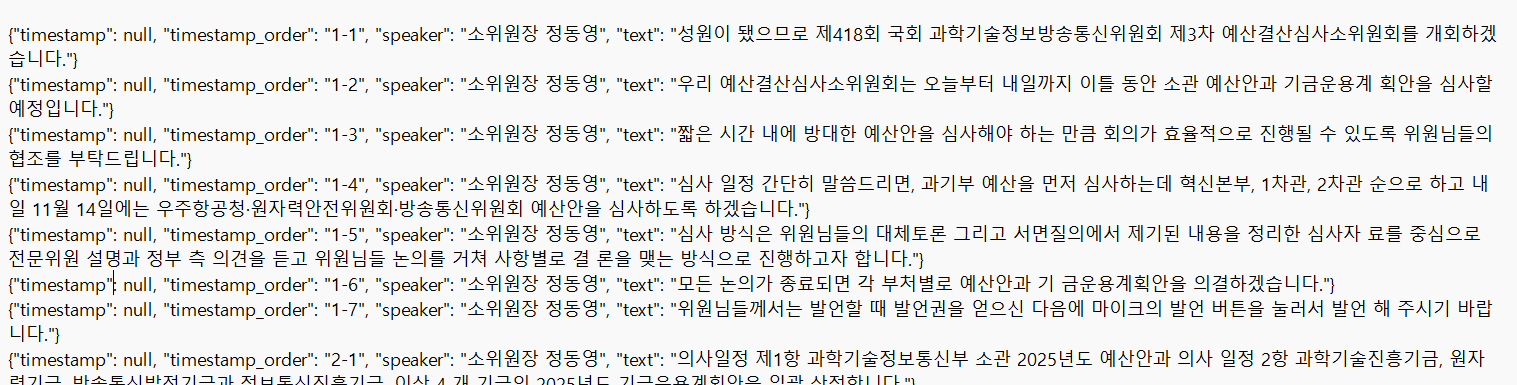
* 분리 기준: 무작위, 겹치기 없음

**4. 부록**

* 전처리 전/후 비교 예시



<원본 데이터 예시 (국회 데이터 중 일부)>



<전처리 데이터 예시 (위의 원본 전처리 후)>

* 국회 위원회 회의록: [국회 위원회 회의록](https://open.assembly.go.kr/portal/data/service/selectServicePage.do?infId=OR137O001023MZ19321)
* AMI Corpus : [AMI Corpus](https://groups.inf.ed.ac.uk/ami/corpus/)
* ICSI Corpus : [ICSI Corpus](https://groups.inf.ed.ac.uk/ami/icsi/)
* 전처리 코드 깃허브 링크: [GitHub - Keindus8404/SKN12\_Ttalkkak: skn12 final 3team ttalkkak](https://github.com/Keindus8404/SKN12_Ttalkkak)

**5. 변경 이력**

| 변경일 | 변경자 | 변경 내용 | 비고 |
| --- | --- | --- | --- |
| 2025-07-17 | 이준석 | 문서 초안 작성 |  |
| 2025-07-18 | 이준석 | 내용 수정 |  |