# 인공지능 데이터 수집 및 전처리 결과서

**2025**년 **11**월 **17**일

## **1.** 문서 개요

본 문서는 “음성에서 문서로, *AI*가 만드는 회의 자동화 시스템” 구축을 위해 인공지능 학습에 필요한 데이터 수집 및 전처리 과정을 정리하기 위해 작성되었다. 프로젝트의 목표는 음성 인식(Whisper)과 화자 분리(Pyannote)를 통해 회의 음성을 텍스트 전문으로 변환하고, 이를 기반으로 안건 추출, 태스크 추출, 요약, 최종 회의록 생성까지 자동화된 파이프라인을 구축하는 데 있다. 이를 위해 본 프로젝트에서는 실제 회의 녹취 환경을 반영한 합성 회의 데이터를 구축하였다. 특히 구어체·비문·간투사·문장부호 미사용 등 STT 전사 특성을 포함한 회의 전문과, 해당 전문에서 도출되는 주요 안건 및 태스크 정보를 하나의 JSON 구조로 통합한 데이터셋을 생성하여 파인튜닝에 활용하였다.

회의 자동화 시스템 구축을 위해 필요한 데이터 수집·정제·가공·전처리의 핵심 절차를 중심으로 정리하였으며, 파인튜닝 데이터셋의 구성 요소와 전처리 기준, 품질 관리 방향을 중심으로 개괄적으로 기술한다.

## **2.** 데이터 사용 목적

- 파인튜닝 대상 모델: *Qwen/Qwen2.5-1.5B-Instruct*
- 선정 이유: 한국어 업무 환경에서의 문맥 이해 능력이 우수하고, 도메인 적응이 용이한 경량 LLM
- 적용 목적: 복잡한 회의 내용을 도메인 특화 언어모델이 안정적으로 이해하도록 하기 위함

## **3.** 원천 데이터 확보

*Qwen/Qwen2.5-1.5B-Instruct* 모델 학습을 위해 필요한 회의 전문·안건·태스크 통합 파인튜닝 텍스트 데이터를 아래와 같이 생성·수집하였다.

본 프로젝트에서는 STT 전사 특성을 반영하여 생성된 회의 전문과, 이에 기반해 생성된 안건 목록과 태스크 정보를 하나의 JSON 구조로 통합한 합성 데이터셋을 구축하였다.
OpenAI GPT 기반 생성 방식을 통해 총 2,621개의 레코드를 확보하였으며, 태스크의 선후 관계, 담당자 변경, 기한 지연 등 실제 회의에서 발생하는 다양한 흐름을 포함하도록 설계하여, 실제 업무 회의 녹취를 입력했을 때도 안정적으로 안건 요약과 태스크 추출이 가능하도록 하는 것을 목표로 하였다.

| 구분 | 데이터 유형 | 출처 | 수집량 | 용량 | 비고 |
|---|---|---|---|---|---|
| 텍스트 | 파인튜닝용 합성 회의 데이터셋 (전문·안건·태스크 통합) | OpenAI GPT 모델 (합성 데이터) | 2,621개 | 7.4 MB | 회의 전문·안건·태스크 통합 생성 데이터 |

## **4.** 전처리

본 프로젝트에서 구축한 약 2,600여 개의 합성 회의 데이터(JSON)는 Whisper 기반 STT 후처리 환경과 유사하도록 구성되었으며, 파인튜닝에 활용하기 전에 다음 전처리 절차를 수행하였다.
전처리는 스키마 정합성 확보, 필드 단위 정규화, 중복 제거, 이상치 데이터 제거를 중심으로 진행하였다.
또한, 기존 회의록을 참고하여 바로 적용 가능한 형식의 데이터 생성을 위해 불필요한 조사, 어미, 서술형 표현 삭제 및 명사형 / 체언 중심의 데이터 생성을 위해 전처리를 진행하였다.

## 4.1 JSON 스키마 정합성 검증

합성 데이터는 아래의 구조를 따라야 하므로, 전처리 단계에서 다음 항목을 검증하였다.

```
{
  "transcript": "<string>",
  "agendas": ["<string>", ...],
  "tasks": [
    {
      "who": "<string or null>",
      "what": "<string>",
      "when": "<string or null>"
    },
    ...
  ]
}
```

1) JSON 파싱 오류 여부
2) transcript, agendas, tasks 필드 존재 여부
3) agendas가 문자열 배열 형식을 따르는지 여부
    a) 일부 데이터에서 agendas가 객체 배열로 생성되는 문제가 발견되어 수정

| 객체 배열 사례 | 문자열 배열 사례 |
|---|---|
| \"agendas\": [{<br>\"title\": \"새로운 영업 전략 및 프로모션 계획 논의\",<br>\"details\": \"시장 데이터를 바탕으로 기존 고객을 대상으로 한 새로운 프로모션 계획을 논의하며, 제품의 공급 가능 여부를 확인하는 것이 중요\"},{<br>\"title\": \"고객 서비스 개선 계획\",<br>\"details\": \"최근 증가하는 고객 불만 사항에 대한 분석과 이에 따른 예산 조정 논의\"}], | \"agendas\": [\"새로운 영업 전략 및 프로모션 계획 논의\", \"고객 서비스 개선 계획\"], |

4) tasks 배열 내부의 각 태스크가 {who, what, when} 구조를 가지는지 확인
5) 필드 타입이 사전에 정의된 스키마 규칙과 일치하는지 확인

검증 결과, 총 2,621개 중 11개 레코드에서 스키마 오류가 발견되었으며, 오류 유형은 다음과 같았다.

- agendas가 문자열이 아닌 객체 배열로 생성된 경우
- what 필드가 null로 생성된 태스크 포함
- who, when이 null로 생성된 경우를 오류로 잘못 판단하는 기존 검증 코드의 문제

이 중 what=null 태스크는 의미 정보가 없으므로 해당 태스크만 삭제, who = null 또는 when = null은 구조적 오류가 아니므로 정상값으로 처리하도록 규칙을 조정하였다.

## **4.2** 태스크 필드 **(who/when/what)** 정규화

합성 과정에서 who·when 값이 “미정”, “null”, null 등 다양한 표현으로 생성된 경우가 있어, 파인튜닝·검증 코드가 일관되게 처리하도록 다음 기준으로 정규화하였다.

- “미정”, “null” 등은 모두 null로 통합
- what = null 은 태스크 자체 제거
- who = null 또는 when = null은 태스크 유지 가능 (현실 회의에서도 미정 상태 존재)

또한 일부 데이터는

- 줄바꿈 없이 transcript가 한 문단으로 생성되어 검증 코드의 발화자 인식 실패
- 제 3자를 지칭하여 태스크를 부여하는 경우
- “ㅇㅇ팀”과 같은 팀 단위 업무 지시가 존재

등의 이유로 who 오류로 잘못 분류되는 문제가 있었기에, 검증 코드를 transcript 내 문자열 매칭 방식으로 개선하여 해결하였다.

## **4.3** 중복 데이터 제거

초기 검증에서 총 68개의 중복 샘플 쌍이 발견되었다. 중복 여부는 transcript·agendas·tasks를 모두 포함한 JSON 전체 비교로 판단하였다.

전처리 단계에서 이들 중복 레코드를 삭제하여 총 2,553개의 데이터로 정제하였다. 전처리 후 재검증 결과, 중복 샘플은 0건으로 확인되었다.

## **4.4** 명사형 **/** 체언 중심의 데이터 생성

초기 agenda 리스트 중 ‘현재 컴플라이언스 시스템 리뷰 및 데이터 정리’ 등 회의록에 적합하지 않은 동사형 어미나 ~하기, ~한다 등의 문장형 내용이 포함되어 있어 ‘컴플라이언스 시스템 리뷰 및 데이터 정리’와 같이 회의록에 사용할 수 있는 데이터로 정제하였다.

프롬프트는 아래와 같으며, open AI 활용하여 정제하였다.

```
SYSTEM_PROMPT = """
너는 회의록 데이터셋 정제 도우미다.
입력으로 주어지는 agendas(안건 문장들)를 "명사형 안건 제목"으로 다듬어야 한다.

규칙:
- 각 항목을 회의 안건 제목처럼 '명사형/체언 중심'으로 바꿔라.
- 불필요한 조사/어미/서술형 표현(~하기, ~합니다, ~하는 것, ~해야 함 등)을 제거하거나 명사구로 변환.
- 의미는 유지하고, 너무 길면 압축(권장 25자 내외).
- 출력은 반드시 JSON 배열(list) 하나만. 길이/순서는 입력과 동일해야 한다.
- 다른 텍스트(설명, 코드블록, 마크다운) 금지.
"""
```

## **4.5** 의미 기반 품질 점수 산출 및 저품질 데이터 제거

각 데이터의 품질을 정량적으로 평가하기 위해 다음 기준을 점수화하였다.

| 항목 | 감점 기준 | 설명 |
|---|---|---|
| unknown_who | -0.2 | 태스크 담당자가 transcript 내에서 확인되지 않는 경우 |
| when_missing | -0.2 | 태스크의 기한 정보가 transcript와 명백히 불일치하거나 존재하지 않는 경우 |
| semantic_low | 최대 -0.2 | 태스크의 what과 transcript 전체의 BERT cosine similarity가 낮은 경우, 태스크 개수 대비 비율만큼 감점 |

최종 점수는 아래와 같은 계산식을 따른다.

```
score = 1.0
if unknown_who:
    score -= 0.2

if when_errors:
    score -= 0.2

if semantic_low:
    total_tasks = max(len(tasks), 1)
    ratio = len(semantic_low) / total_tasks
    score -= 0.2 * ratio
score = max(score, 0.0)
```

전처리·정규화·중복 제거 후 재검증 결과는 다음과 같다.

- 평균 score: 0.8809
- 최소 score: 0.4000
- score < 0.5인 데이터는 총 49개로, 모델 파인튜닝 데이터셋에서 제외하였다.

# **5.** 데이터 품질 검증

전처리 완료 후, 정제된 2,553개 데이터에 대해 다음 항목을 중심으로 품질 검증을 수행하였다.

## **5.1** 품질 검증 요약표

| 검증 항목 | 검증 내용 | 결과 | 비고 |
|---|---|---|---|
| 스키마 유효성 | - transcript / agendas / tasks 필드 존재 여부<br>- agendas 문자열 배열 여부<br>- tasks의 who / what / when 필드 유효성 | 스키마 오류 0건 | 전처리 후 재검증 결과 기준 |
| 태스크 담당자**(who)** 유효성 | - who가 transcript 내 발화자 또는 언급 인물과 매칭되는지 확인<br>- 팀 단위 명칭 포함 여부 확인 | unknown_who 샘플 6건, 총 10개 | 문자열 매칭 방식으로 개선.<br>샘플 6건의 경우 “ㅇㅇ팀”을 인식 못 한 경우 확인. |
| 태스크 기한**(when)** 유효성 | - transcipt에 존재하지 않는 날짜·요일이 생성되었는지 확인 | when_missing 샘플 10건, 총 18개 | 일부 when은 합성 오류로 판단됨 |
| 태스크 의미 유사도 검사 | - what vs transcript BERT cosine similarity 검사<br>- 의미 불일치 태스크 비율 계산 | low_similarity 샘플 194건, 총 402개 | 품질 점수(score)에 반영 |
| 중복 여부 검사 | - transcript·agendas·tasks 전체 동일 여부 검사 | 중복 0건 | |
| 품질 점수 **(score)** | - unknown_who / when_missing / low_similarity 기반 점수 산출 | 평균 0.8809<br>최소 0.4000<br>최대 1.0 | score < 0.5 데이터는 파인튜닝에서 제외 |

## **5.2** 최종 결과

위 검증 절차를 통해 스키마 오류·중복 데이터는 모두 제거되었으며, 태스크의 담당자 및 기한 오류, 의미 불일치 등 합성 한계로 인해 발생한 품질 이슈는 점수 기반으로 관리하였다.
최종적으로 socre가 0.5 미만인 데이터는 훈련 품질 확보를 위해 파인튜닝 데이터셋에서 제외하였다. (최종 사용 데이터 약 2,500여 개)