Uncounted Dataset

기술 사양서 v1.0 · 2026-02-11

> “해석과 후속(다운스트림) 처리는 구매사의 책임입니다. 우리는 내러티브가 아니라 원신호(raw signal)를 제공합니다.”

01. 데이터셋 유형

Voice Session

다양한 엔터프라이즈 환경에서 수집된 고품질 원음(원시 신호) 세션.

형식 MP3 / WAV (무손실)	샘플레이트 16kHz / 44.1kHz
단위 세션(평균 5분+)	채널 Mono / Stereo (옵션)

App Usage Session

사용자 행태 패턴을 밀리초 단위로 기록한 상호작용 이벤트 스트림.

형식 JSONL	정밀도 밀리초 타임스탬프
구조 이벤트 스트림	단위 세션 / 이벤트

02. 설계 원칙

세션 기반(Session-Based)
데이터는 의미 있는 사용자 세션 단위로 묶여 맥락을 보존합니다.

의미 라벨 없음(No Semantic Labels)
편향을 줄이기 위해 사전 라벨을 제공하지 않습니다. 분류는 구매사 영역입니다.

원시 충실도(Raw Fidelity)
원 신호 특성과 노이즈를 유지하기 위해 전처리를 최소화합니다.

프라이버시 우선(Privacy-First)
수집 단계에서 PII를 제거하며, 익명화는 되돌릴 수 없도록 설계합니다.

03. 메타데이터 레이어

모든 납품 데이터에는 품질·희귀도 평가를 포함한 기술 메타데이터가 함께 제공됩니다.

Q. 품질 점수(Quality Scoring)

각 파일/세션의 신호대잡음비(SNR) 등 자동 품질 지표를 산정합니다.

- 오디오: SNR, 배경잡음 레벨(dB), 클리핑 탐지
- 앱 로그: 세션 완결성, 이벤트 밀도

R. 희귀도 지수(Rarity Index)

오픈 데이터셋에서 해당 데이터 포인트가 등장할 확률을 추정합니다.

- Common Crawl / LibriSpeech 등과의 벡터 유사도 기반
- 지수가 높을수록 엣지 케이스 학습 가치가 높음

04. 지원 포맷

JSONL - 표준 텍스트 기반 로그
Apache Parquet - 컬럼형 저장/압축
Apache Avro - 스키마 진화 지원

05. 수집(인제스트) 워크플로우

1	2	3	4
데이터 캡처 SDK 연동	익명화 PII 제거	포맷팅 스키마 검증	납품 S3 버킷 / API

06. 범위 및 책임

Uncounted가 제공합니다

- 원시(미가공) 데이터 신호
- 포괄적인 기술 메타데이터
- 일관된 스키마 검증
- 포맷 변환(Parquet/Avro)

Uncounted가 제공하지 않습니다

- 사람 검증 의미 라벨링
- 피처 엔지니어링/벡터화
- BI/인사이트 분석
- 사전학습 모델 가중치