[최종 확정] LLM 기반 정체성 프로파일링 Action Plan (vFinal+)

기준 문서: 1116 Action Plan 최종(1736).pdf

피드백: 피드백(Action Plan에 대한).pdf

Phase 1: 기반 데이터 구축 (Foundation)

1. **Diarization & STT 실행:** AUDIO_IN 설정 및 1116...ipynb 노트북의 Phase 1 코드 실행.
2. **병합:** 모든 결과를 final_transcript.json으로 병합.
3. **입력 스모크 테스트:** Phase 2 시작 전, final_transcript.json 파일 존재 여부 및 스키마 검증. (오류 시 가이드 메시지와 함께 중단)

Phase 2: LLM 기반 정체성 프로파일링 (LLM Profiling)

1. **화자별 그룹화 (Group by Speaker)**
    - final_transcript.json을 로드하여 화자(SPEAKER_00 등)별로 모든 발화(utter)를 묶습니다.
    - 이 단계에서 발화 횟수가 3회 미만인 화자는 LLM 호출 대상에서 제외(skip)하고 로그를 남깁니다.
2. **대표 발화 추출 (Smart Selection)**
    - **고정 예산:** 화자당 10~15개의 대표 발화를 "영리하게" 추출합니다.
    - **추출 전략:**
        - 가장 긴 발화 (Top 5)
        - 키워드 기반 발화 (Top 3) (예: "요약", "반대", "제안", "질문" 포함)
        - 시점별 분배 (Top 3) (회의 초/중/후반 각 1개씩)
    - '코사인 유사도' 대신, 구현이 간편한 **'단순 텍스트 해시(hash)'**를 사용하여 중복 발화를 억제합니다.
3. **LLM 호출 설계 (Prompt Design)**
    - **프롬프트 템플릿:** 프롬프트를 별도 파일(prompt_v1.txt 등)로 분리하여 관리합니다.
    - **프롬프트 계약 (Contract):**

        - **금칙 조항:** "실명/고유명사 사용 금지" (NER 모듈과 경계)
        - **스키마 강제:** display_label, one_liner, keywords, communication_style, stance_markers, evidence_utter_idx JSON 스키마 명시 [cite: 788-796].

4. **LLM 호출 및 오케스트레이션 (Call & Orchestration)**
    - **모델 설정:** GPT-4o-mini
        - **JSON 강제:** response_format={"type":"json_object"}
        - **일관성:** temperature=0.2
    - **캐시 (Cache):** (model, prompt_hash)를 키로 사용하여 LLM 호출 결과를 파일 캐시합니다.
    - **병렬 호출:** 모든 화자에 대한 LLM 호출을 병렬(Batch)로 실행합니다.
        - Rate Limit 준수를 위해 MAX_CONCURRENT_REQUESTS = 5 (최대 동시 요청 5개), DELAY_BETWEEN_BATCHES = 0.2 (배치 간 0.2초 딜레이)와 같은 구체적인 상수를 설정합니다.
    - 비용 상한선(MAX_COST_PER_MEETING)을 설정하고, 예상 비용 초과 시 경고 또는 중단하는 로직을 추가합니다.
    - **오류 처리:** JSON 파싱 실패 시 2회 재시도 (Backoff 적용). 실패 시 텍스트에서 {...}를 직접 추출. 최종 실패 시 로그 후 스킵.
5. **산출물 저장 및 버전 관리 (Save & Versioning)**
    - **필수 산출물:** speaker_identity.json
    - JSON 저장 시 encoding='utf-8' 및 ensure_ascii=False를 명시적으로 사용합니다.
    - **메타데이터 박제:** 산출물에 model, prompt_hash, timestamp 등 재현성을 위한 메타 정보를 포함시킵니다.
6. **품질 검증 및 시연 (QA & Demo)**
    - **스모크 테스트:** speaker_identity.json 생성 후, evidence_utter_idx가 유효한 인덱스 범위 내에 있는지 자동 검사합니다.
    - **시연 셀:** 노트북에 "결과 미리보기" 셀을 추가하여 최종 결과를 테이블로

시각화합니다.