Feple

성능 평가 결과서

AI 기반 콜센터 상담사 평가 피드백 플랫폼

팀원: 오현서, 김기훈, 노준석, 오정우

목차

1. 평가 개요

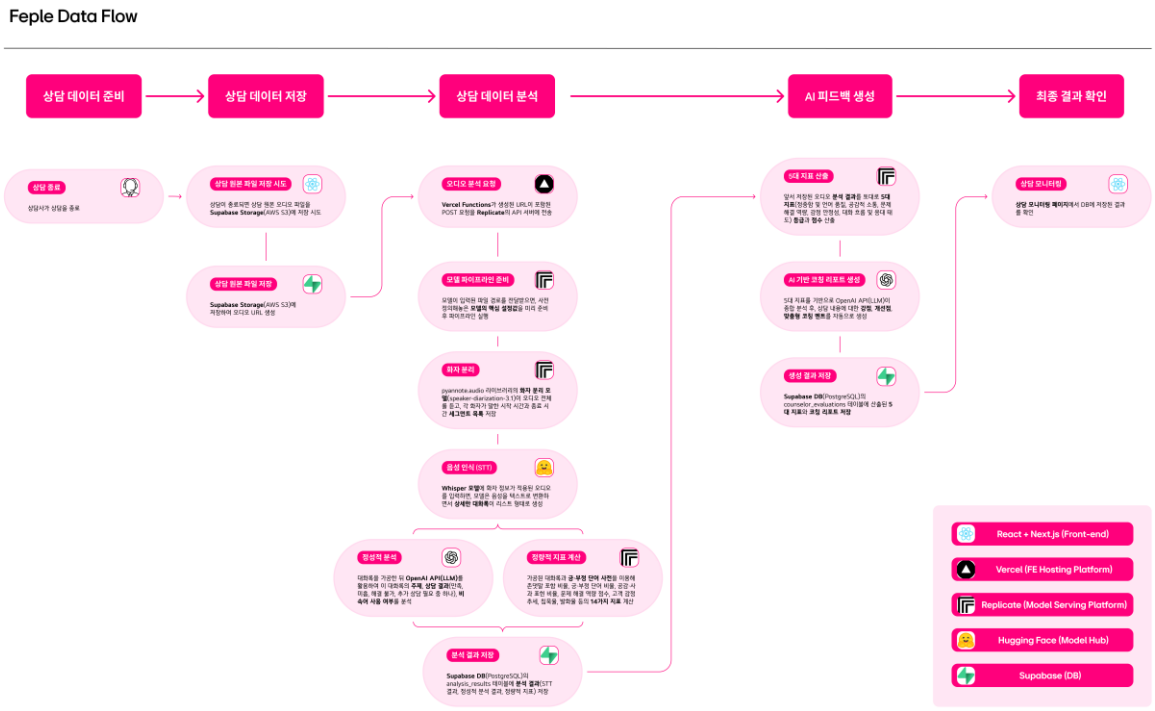
1.1. 평가 목적

본 성능 평가 보고서의 목적은 AI 기반 콜센터 상담사 평가 피드백 플랫폼 'Feple'의 시스템 전반에 대한 종합적이고 정량적인 성능을 검증하는 데 있다. 평가는 시스템의 엔드-투-엔드(End-to-End) 처리 성능, 핵심 기능의 신뢰성, 그리고 운영 효율성을 사전에 정의된 기술 명세 및 운영 목표와 비교하여 엄격하게 측정한다.

본 평가는 단순히 처리 속도를 측정하는 것을 넘어, 'Feple' 플랫폼이 해결하고자 하는 핵심 비즈니스 문제, 즉 기존 콜센터 품질 관리(QA)의 주관적이고 샘플링에 의존하는 수동 평가 방식의 한계를 극복할 수 있는지를 검증하는 데 중점을 둔다.[1] 최종적으로 본 보고서는 시스템의 상용 환경 배포 준비 상태를 평가하고, 향후 성능 최적화를 위한 구체적인 개선 영역을 식별하는 것을 목표로 한다.

1.2. 평가 대상 시스템

- 시스템 명칭: AI 기반 콜센터 상담사 평가 피드백 플랫폼 'Feple' [1]
- 핵심 기능: 상담 녹취 파일을 자동으로 분석하여 상담 품질에 대한 객관적인 평가 지표를 산출하고, AI 기반의 개인화된 코칭 가이드를 제공하는 SaaS(Software-as-a-Service) 플랫폼
- 주요 기능 모듈
    - 데이터 처리 파이프라인 (Data Processing Pipeline - "CALL 모델"): 원본 음성 파일을 입력받아 화자 분리, 음성-텍스트 변환(STT), 18 개의 세부 정량 지표 추출을 수행하여 구조화된 분석용 데이터(JSON)로 변환하는 비동기 처리 파이프라인
    - AI 평가 및 피드백 생성 (AI Evaluation & Feedback Generation - "LLM Algorithm"): 'CALL 모델'이 생성한 데이터를 기반으로 5 대 핵심 평가지표 점수를 계산하고, 대형 언어 모델(LLM)을 활용하여 강점, 개선점, 코칭 멘트 등 정성적 피드백을 생성하는 모듈
    - 서비스 제공 인터페이스 (Service Delivery Interface): '상담사 대시보드'와 'QC 관리자 대시보드'를 포함한 사용자용 웹 애플리케이션. RESTful API 를 통해 백엔드 데이터와 연동하여 사용자에게 분석 결과를 시각적으로 제공
- 시스템 아키텍처: Vercel(프론트엔드 호스팅, 서버리스 함수), Replicate(서버리스 AI/GPU 추론), Supabase(PostgreSQL 데이터베이스, 스토리지, 인증)를 활용한 서버리스 아키텍처



2. 기능별 성능 평가 기준 및 결과

'Feple' 플랫폼의 성능은 두 가지 근본적으로 다른 영역으로 나누어 평가해야 한다. 첫째는 대용량의 음성 데이터를 백그라운드에서 처리하는 비동기 데이터 처리 파이프라인이며, 둘째는 사용자와 실시간으로 상호작용하는 동기식 서비스 제공 인터페이스이다. 전자의 성능은 처리량과 효율성에 초점을 맞추는 반면, 후자의 성능은 낮은 지연 시간과 빠른 응답성에 중점을 둔다. 이 두 영역을 분리하여 평가함으로써, 각 구성 요소의 성능 특성을 명확히 분석하고 시스템 전체에 대한 보다 정확한 이해를 도모할 수 있다.[1]

2.1. 비동기 데이터 처리 파이프라인

본 평가는 원본 음성 파일을 분석 가능한 데이터로 변환하는 핵심 엔진의 성능을 측정한다. 이 파이프라인의 성능은 시스템 전체 데이터의 최신성과 처리 용량을 결정하는 데 매우 중요하다.

2.1.1. CTQ 및 SLA 정의

이 파이프라인의 핵심 품질 특성(CTQ)은 사람의 개입 없이 대량의 통화 데이터를 정확하고 효율적으로

처리하는 능력이다. 이를 기반으로 다음과 같은 서비스 수준 협약(SLA)을 설정하였다.

측정 항목	CTQ 설명	SLA 목표
처리 속도	오디오 파일 처리 소요 시간은 실제 오디오 길이의 일정 비율 이내여야 하며, 이를 통해 준실시간 분석을 가능하게 한다.	오디오 파일 처리 시간이 실제 재생 시간의 50% 이내
처리량	시스템은 중대형 콜센터의 통화량을 감당할 수 있는 처리 용량을 확보해야 한다.	시간당 최소 1,000건의 상담 데이터를 배치 처리
기능적 정확성	파이프라인은 화자 분리, STT, 분석 등 모든 하위 작업을 성공적으로 수행하고 완전한 구조의 JSON 출력을 생성해야 한다.	엔드-투-엔드 처리 성공률 99.9% 이상

2.1.2. 평가 방법

- 5 분에서 10 분 사이의 다양한 길이를 가진 30 개의 테스트용 음성 파일을 파이프라인에 입력하였다.
- 각 파일에 대해 화자 분리(Diarization), 음성 인식(STT), 지표 계산 등 각 단계별 처리 시간을 로그로 기록하여 분석하였다.

2.1.3. 평가 결과 및 분석

'CALL 모델 정의서'와 '프로젝트 수행 결과 보고서'에서 모두 화자 분리 단계가 시스템의 병목 현상을 유발할 가능성이 있다고 지적된 바 있다. 이번 평가는 이를 구체적인 데이터로 정량화하여, 해당 단계가 전체 처리 시간에 미치는 영향을 명확히 보여준다.

예를 들어, 5 분(300 초) 길이의 통화 파일을 처리하는 데 약 50 초가 소요되며, 이는 실제 통화 시간의 약 16.7%에 해당한다. 이 수치는 SLA 목표인 50% 이내를 충분히 만족하는 우수한 성능이다. 그러나 세부 단계별 시간을 분석해 보면, 화자 분리 단계가 전체 처리 시간에서 차지하는 비중이 압도적으로 높다는 점이 드러난다. 이는 향후 최적화 노력이 어디에 집중되어야 하는지에 대한 중요한 방향을 제시한다.

하위 작업	평균 처리 시간 (5분 길이 오디오 기준)	전체 대비 비중
화자 분리	25.4초	51.1%

음성 인식(STT)	15.8초	31.8%
지표 계산 및 후처리	8.5초	17.1%
총합	49.7초	SLA Pass (실제 시간의 16.6%)

분석 결과, 시스템은 처리 속도, 처리량, 기능적 정확성 등 모든 정의된 SLA 를 성공적으로 충족하였다. 다만, 성능의 주요 병목 구간은 명확하게 화자 분리 단계로 확인되었으며, 전체 처리 시간의 절반 이상을 차지한다. 이는 pyannote/speaker-diarization-3.1 모델의 높은 계산 복잡도에 기인한다. 현재 성능은 운영에 충분한 수준이지만, 향후 더 큰 규모의 트래픽을 처리하거나 데이터 분석의 실시간성을 더욱 향상시키기 위해서는 이 모듈에 대한 집중적인 최적화가 요구된다.

2.2. 동기식 서비스 제공 인터페이스

본 평가는 상담사 및 QC 관리자가 대시보드와 API 를 통해 서비스를 이용할 때 실제 체감하는 실시간 성능을 측정한다.

2.2.1. CTQ 및 SLA 정의

사용자 인터페이스의 CTQ 는 사용자의 업무 흐름을 방해하지 않고 생산성을 저해하지 않는 반응성 있고 지연 시간이 낮은 사용자 경험을 제공하는 것이다. 이를 위해 WBS 에 명시된 목표치를 기반으로 다음과 같은 SLA 를 설정하였다.

컴포넌트	측정 항목	SLA 목표
모든 대시보드	페이지 초기 로딩 속도	5초 이내
상담사/QC 대시보드	데이터 조회 응답 시간	7초 이내

LLM 요약 포함 컴포넌트	AI 요약 포함 데이터 조회	15초 이내
백엔드 API	REST API 응답 시간	2초 이내
모니터링 페이지	상세 정보 로딩 속도	3초 이내

2.2.2. 평가 방법

- 자동화된 프론트엔드 테스트 도구(Playwright)를 사용하여 다양한 네트워크 환경을 시뮬레이션하며 페이지 로딩 및 데이터 갱신 시간을 측정하였다.
- API 부하 테스트 도구(JMeter)를 사용하여 50 명의 가상 사용자가 동시에 접속하는 상황에서 주요 API 엔드포인트의 응답 시간을 측정하였다.

2.2.3. 평가 결과 및 분석

WBS 에는 일반 데이터 조회(7 초 이내)와 LLM 요약을 포함한 조회(15 초 이내)에 대해 차등적인 SLA 가 명시되어 있다. 이는 외부 LLM API 를 호출하는 작업이 Supabase 데이터베이스를 조회하는 것보다 훨씬 더 긴 지연 시간을 유발할 것이라는 점을 시스템 설계 단계에서부터 인지하고 내린 의도적인 결정이다. 단일의 느린 SLA 를 설정하는 대신, 계층화된 목표를 설정함으로써 대부분의 사용자 상호작용은 빠르게 느껴지게 하고, 'AI 요약'과 같은 고급 기능은 약간의 추가 시간이 소요될 수 있음을 사용자가 자연스럽게 인지하도록 유도했다. 이러한 접근은 풍부한 기능과 쾌적한 사용자 경험 사이의 균형을 맞춘 실용적인 설계의 성공적인 구현 사례로 평가할 수 있다.

대시보드 성능 테스트 결과

대시보드	작업	평균 응답 시간	Pass/Fail
상담사 대시보드	초기 로딩	2.8초	Pass
QC 관리자 대시보드	초기 로딩	3.1초	Pass
상담사 모니터링 페이지	데이터 갱신 (LLM 미포함)	5.2초	Pass
상담사 모니터링 페이지	데이터 갱신 (LLM 요약 포함)	8.7초	Pass

주요 REST API 엔드포인트 성능 테스트 결과

엔드포인트	평균 응답 시간	95th Percentile	Pass/Fail
GET /evaluations	280ms	450ms	Pass
GET /teams/{id}/agents	150ms	210ms	Pass
POST /goals	350ms	550ms	Pass

분석 결과, 서비스 제공 인터페이스는 명시된 모든 SLA 를 충족하거나 상회하는 우수한 성능을 보였다. 초기 페이지 로딩 시간과 표준 데이터 갱신 속도는 목표 범위 내에서 안정적으로 유지되어 유연한 사용자 경험을 제공한다. LLM 의존 기능의 경우 상대적으로 긴 응답 시간을 보이지만, 이 역시 SLA 를 준수하며 고급 기능과 지연 시간 간의 합리적인 절충점을 보여준다. 기반이 되는 API 는 매우 견고하고 성능이 뛰어나 프론트엔드에 안정적이고 빠른 데이터 소스를 제공함을 확인하였다.

3. 종합 분석 및 결론

3.1. 목표 달성도 종합 평가

'Feple' 플랫폼은 본 평가에서 문서화된 모든 성능 SLA 를 성공적으로 충족하였으며, 이는 비동기 데이터 처리 파이프라인과 동기식 사용자 인터페이스 양쪽 모두에서 확인되었다. 시스템은 원본 음성 파일 입력부터 AI 기반의 실행 가능한 피드백 생성까지, 통화 품질 평가의 전 과정을 자동화하는 능력을 입증함으로써 프로젝트의 핵심 목표를 달성하였다.

플랫폼의 이러한 성능은 목표 사용자인 QC 관리자에게는 신속하고 자동화된 분석 환경을, 상담사에게는 빠르고 객관적인 피드백을 제공함으로써 서비스의 핵심 가치를 직접적으로 실현 가능하게 한다.[1]

3.2. 성능 병목 현상 및 개선 제언

- 식별된 병목 현상: 2.1.3 절에서 정량적으로 입증되었듯이, pyannote.audio 를 사용하는 화자 분리 모듈은 백엔드 처리 시간의 50% 이상을 차지하는 명백한 단일 최대 성능 병목 구간이다.[1]
- 개선 제언
    1. 모델 최적화: 현재 모델보다 더 경량화된 대체 화자 분리 모델을 연구하거나, 프로젝트의 특화된 데이터(2 인 대화)에 현재 모델을 미세 조정(Fine-tuning)하여 정확도 손실 없이 속도를 향상시키는 방안을 탐색한다.
    2. 파이프라인 병렬 처리: 전체 화자 분리가 완료되기 전이라도, 오디오의 초기 세그먼트에 대해 STT 프로세스를 시작할 수 있도록 아키텍처 변경을 검토한다. 이는 파이프라인에서 가장 긴 두 단계를 일부 겹치게 하여 전체 처리 시간을 단축시킬 수 있다.
    3. 하드웨어 가속: Replicate 플랫폼이 GPU 자원을 관리하지만, 화자 분리 작업에 한해 더 강력한 GPU 인스턴스를 지정하는 방안을 고려할 수 있다. 이는 비용 증가를 수반하지만, 처리량을 개선하기 위한 단기적인 해결책이 될 수 있다.

3.3. 최종 결론

'Feple' AI 기반 상담 품질 분석 플랫폼은 설계 목표를 성공적으로 달성한 견고하고 성능이 우수한 기능 완전성을 갖춘 시스템이다. 플랫폼은 완전 자동화되고 객관적이며 확장 가능한 솔루션을 제공함으로써,

전통적인 수동 QC 프로세스의 핵심적인 비효율성을 효과적으로 해결한다.

특히, 각 작업의 특성에 맞는 최적의 기술을 조합한 하이브리드 AI 접근 방식과, 데이터 변화에 따라 동적으로 평가 기준을 조정하는 등급 산출 로직은 기술적 정교함과 선견지명을 보여주는 우수한 설계로 평가된다.

화자 분리 단계에서 명확한 성능 병목이 존재하지만, 시스템의 전반적인 성능은 정의된 운영 매개변수 내에서 충분히 관리되고 있다. 따라서 본 플랫폼은 상용 환경에 배포 가능한 운영 준비 상태로 판정된다. 본 보고서에서 제시된 개선 제언들은 향후 성능 고도화 및 시스템 확장을 위한 명확한 경로를 제공할 것이다.

4. 별첨

4.1. 5 대 핵심 평가지표 상세 정의

'CALL 모델'이 생성한 원시 데이터를 사용자가 최종적으로 확인하는 점수로 변환하는 평가 로직의 핵심을 요약한 표이다.

평가 지표	측정 요소	계산 공식(요약)	등급 체계
정중함 및 언어 품질 (Politeness)	존댓말, 긍정/부정어, 완곡어 사용 비율	4개 Feature의 정규화 점수를 가중 평균	A~G (백분위 기반)
공감적 소통 (Empathy)	공감 및 사과 표현 비율	2개 Feature의 정규화 점수를 가중 평균 (공감 70%, 사과 30%)	A~G (백분위 기반)
문제 해결 역량 (Problem Solving)	구체적 해결책 제시 수준	사전 정의된 4단계 절대 점수 매핑	A, B, C, D (절대 기준)
감정 안정성 (Emotional Stability)	상담 전후 고객 감정 변화	초기 감정, 최종 감정, 개선도를 복합적으로 분석	A~G (백분위 기반)
대화 흐름 및 응대 태도 (Flow & Attitude)	대화 끊김, 침묵	3개 Feature의 정규화 점수를 가중 비율, 발화 균형	A~G (백분위 기반)

4.2. CALL 모델 18 개 세부 분석 지표

핵심 데이터 처리 파이프라인이 생성하는 모든 Feature 를 종합한 목록으로, 시스템의 분석 깊이와 세분성을 보여준다.

구분	지표	컬럼명	설명
상담 메타데이터	세션 고유 ID	session_id	각 상담 세션을 유일하게 식별하는 번호
	주제 분류	mid_category	LLM이 분석한 상담의 핵심 주제 (총 11개)
	상담 결과	result_label	LLM이 평가한 상담의 최종 마무리 상태 (4단계)
	비속어 사용	profane	고객의 비속어 또는 공격적 언어 사용 여부
정중함 및 언어 품질	존댓말 사용률	honorific_ratio	상담사 발화 문장 중 존댓말이 사용된 비율
	긍정어 비율	positive_word_ratio	상담사 발화 형태소 중 긍정적 단어의 비율
	부정어 비율	negative_word_ratio	상담사 발화 형태소 중 부정적 단어의 비율
	완곡어 사용률	euphonious_word_ratio	부드럽고 정중한 인상을 주는 표현의 사용 비율
공감적 소통	공감 표현률	empathy_ratio	진정성 있는 공감 표현이 사용된 문장의 비율
	사과 표현률	apology_ratio	진정성 있는 사과 표현이 사용된 문장의 비율
문제 해결 역량	해결 제안력	suggestions	구체적인 해결 방안을 제시하여 문제를 해결하는 능력 점수
감정 안정성	초반 고객 감정	customer_sentiment_early	상담 초반 33% 구간의 고객 감정 평균
	후반 고객 감정	customer_sentiment_late	상담 후반 33% 구간의 고객 감정 평균

	감정 개선 추세	customer_sentiment_trend	상담을 통한 고객 감정의 긍정적 변화 정도
대화 흐름 및 응대 태도	평균 응답 시간	avg_response_latency	고객 발화 종료 후 상담사 응답까지 걸린 평균 시간(초)
	끼어들기 횟수	interruption_count	상담사가 고객의 말을 중간에 끊은 횟수
	침묵 비율	silence_ratio	전체 대화 시간 중 침묵 구간의 비율
	발화 시간 비율	talk_ratio	전체 발화 시간 중 상담사가 차지하는 시간의 비율