SK네트웍스 **Family AI**과정 **5**기

모델링 및 평가 테스트 계획 및 결과 보고서

☐ 개요

- 산출물 단계 : 모델링 및 평가
- 평가 산출물 : 테스트 계획 및 결과 보고서
- 제출 일자 : 2025-02-13
- 깃허브 경로 : https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN05-FINAL-4TEAM.git
- 작성 팀원 : 김지연, 배윤관, 박보람

개요	본 테스트는 AI기반 인터뷰 시스템의 질문 생성 모델과 답변 평가 모델, 음성 분석 및 비언어적 평가 모델, 사용자 답변 요약 모델의 성능을 검증하기 위해 수행됩니다. 테스트의 주요 목표는 다음과 같습니다. <질문 생성 모델(GPT-3.5-turbo) 테스트> - 지원자의 이력서 및 직무 공고 데이터를 기반으로 적절한 면접 질문을 생성하는지 확인 - 평가 기준을 반영하여 다양한 유형의 질문이 생성되는지 검증 <답변 평가 모델(GPT-4) 테스트> - 주어진 면접 질문과 응답을 평가 기준에 따라 일관되게 채점하는지 확인 <whisper 기반 음성-텍스트 변환 모델 텍스트> - 입력된 음성 파일을 whisper 모델이 정확하게 텍스트로 변환하는지 검증 <발음, 말 빠르기, 말더듬 분석 모델 테스트> - whisper의 변환 결과를 활용하여 발음, 말 빠르기, 말더듬 횟수를 분석하고 평가하는 모델의 성능 검증 <AI기반 비언어적 평가 모델 테스트> - 음성 분석 데이터를 기반으로 GPT-4가 평가 점수를 일관성 있게 부여하는지 확인 <사용자 답변 요약 모델 테스트> - 사용자의 긴 답변을 논리적인 핵심 내용 중심으로 요약하는 모델의 성능을 검증 - 면접자의 답변을 요약할 때 중요한 정보가 유지되는지 확인

질문 생성 모델 테스트 과정	테스트 방법 1. 입력 데이터 준비 - 예제 이력서 텍스트, 담당 업무, 지원 자격 데이터 생성 - 평가 기준(evaluation_metrics) 사전 정의 2. 질문 생성 함수 호출 - 질문 생성 함수를 실행하여 면접 질문을 생성 - 변환된 json 데이터에서 질문의 개수, 내용, 다양성을 확인 3. 테스트 케이스 정의 4. 출력 확인 - 질문이 평가 기준을 반영하는지 확인
답변 평가 모델 테스트 과정	테스트 방법 1. 입력 데이터 준비 - 예제 질문 및 지원자의 응답을 정의 - 직무 관련 이력서와 공고 내용을 준비 2. 답변 평가 함수 호출 - 답변 평가 함수를 실행하여 응답을 평가 - 평가 결과가 정상적으로 반환되는지 확인 3. 테스트 케이스 정의 4. 출력 확인 - 각 평가 기준별 점수 부여가 논리적인지 확인
stt 모델 테스트 과정	테스트 방법 1. 입력 데이터 준비 - 명확한 발음, 속도가 빠른 발음 음성 파일 준비 2. 음성 변환 함수 호출 - 음성 변환 함수 호출 - 반환된 텍스트가 원본 음성과 얼마나 일치하는지 확인 3. 테스트 케이스 정의 4. 출력 확인

발음, 속도, 말더듬 분석 모델 테스트	테스트 방법 1. 입력 데이터 준비 - 다양한 발음 정확도 및 속도를 가진 음성 파일 준비 - 말더듬이 포함된 음성 파일 추가 2. 음성 분석 함수 호출 - 음성 분석 함수 실행 - 발음, 속도, 말더듬에 대해 개별 테스트 수행 3. 테스트 케이스 정의 4. 출력 확인
AI 기반 비언어적 평가 모델 테스트	테스트 방법 1. 입력 데이터 준비 - whisper 변환된 텍스트 및 음성 분석 데이터 활용 - 평가 기준(발음, 속도, 말더듬이 포함된 테스트 데이터 생성) 2. 평가 모델 호출 - 평가 모델 함수 실행 - GPT-4의 점수 부여 및 개선사항 일관성 검토 3. 테스트 케이스 정의 4. 출력 확인
답변 요약 모델 테스트	테스트 방법 1. 입력 데이터 준비 - 예제 질문 및 사용자의 장문 답변을 준비 - 다양한 길이와 복잡도를 가진 답변을 포함 2. 요약 모델 호출 - 답변 보정 모델 호출 - 답변 요약 모델 호출 3. 테스트 케이스 정의 4. 출력 확인 - 보정이 잘 됐는지, 누락이 없는지 확인 - 문맥이 자연스러운지 확인

결과	질문 생성 모델 ● 질문이 지원자의 이력서 및 채용 공고와 얼마나 연관성이 있는지 확인 -> SBERT 문장 임베딩 유사도 계산 (0.82) ● 질문의 문법적 오류와 가독성이 얼마나 뛰어난지 확인 -> LanguageTool을 활용한 오류 탐지 및 점수화 (0.91) ● 질문이 직무 관련 키워드와 얼마나 밀접하게 연결되는지 확인 -> SBERT활용한 키워드 유사도 점수화 (0.78) 결론: 질문 품질을 전반적으로 우수하나, 특정 직무의 전문성 반영이 필요 답변 평가 모델 ● 동일한 답변을 여러 번 평가하여 점수 변동을 확인했을 때 1점 차이 ● AI피드백이 반복되지 않고 다양한 표현을 사용하는지 확인 -> 텍스트 엔트로피 값 측정 (피드백1: 3.91, 피드백2: 3.58) ● 피드백이 쉽게 이해될 수 있는지 가독성 점수 측정 (98.25) 결론: 답변 평가 모델은 정확하고 일관된 평가 결과를 제공하며, 피드백 가독성이 높고, 논리적 전개가 명확함. 답변 요약 모델 ● 답변 보정 모델 ○ 프롬프트 최적화 모델이 **더 간결하면서도 명확하게 핵심 내용을 전달**하고 있습니다. 특히 불필요한 단어를 줄여 가독성이 향상되었으며, 문장 간 연결이 매끄러워 일관성이 뛰어납니다. 또한, 결과를 강조하는 방식도 더 효과적이라 전달력이 높아졌습니다 ● 답변 요약 모델 ○ 프롬프트 최적화 모델이 **압축적인 전달이 필요한 경우** 더 효과적이에요