모델링 및 평가 테스트 계획 및 결과 보고서

2025년 **11**월 **10**일

1. 테스트 개요

본 문서는 '말하는대로(AI 기반 회의 자동화 시스템)' 서비스에 탑재될 파인튜닝 모델 및 음성 인식 모델의 성능 적합성을 검증하기 위한 결과 보고서이다. 선정된 모델이 제한된 하드웨어 환경에서 목표하는 정확도와 응답 속도를 보장하는지 평가한다.

• 테스트 기간: 2025.11.19 ~ 2025.11.21 (4주차)

2. 테스트 목적

본 테스트의 목적은 '말하는대로(AI 기반 회의 자동화 시스템)' 모델의 성능과 품질을 종합적으로 검증하기 위함이며, 주요 목표는 다음과 같다.

1. 기능 검증: 시스템이 정의된 요구사항에 따라 주요 기능(요약&안건, 태스크 추출, 발언자 인식 등) 수행 확인
2. 성능 검증: 시스템의 응답 속도, 처리 효율 등 성능 관련 지표가 기준을 충족하는지 평가
3. 예외 처리 검증: 예외 상황 발생 시 시스템이 적절히 처리하고 복구 절차를 수행하는지 검증
4. 음성 인식률 검증: 회의 음성을 텍스트로 변환 시 오탈자 및 누락 발생률 최소화 검증

3. 테스트 환경

3.1 하드웨어 환경

- 테스트 장비: HP ProBook 450 G10 Notebook PC (15.6 inch)
- 프로세서: 13th Gen Intel® Core™ i5-1340P (1.9GHz, 12코어/16스레드)
- 그래픽: Intel® Iris® Xe Graphics (통합 GPU, VRAM 128MB)
- RAM: 32GB DDR4 (3200MHz, 듀얼 채널 구성)
- 저장 장치: 512GB SSD (실제 사용 가능 용량 약 477GB)

3.2 소프트웨어 환경

- 운영 체제: Windows 11 Pro
- 개발 언어: Python 3.11
- 주요 라이브러리: PyTorch, Hugging Face Transformers, PEFT

3.3 네트워크 환경

- 테스트 서버: 내부 네트워크 환경에서 분리된 전용 네트워크

3.4 테스트 도구

3.4.1 STT 품질 검증 도구

구분	도구	용도
전사 정확도 평가	CER (Character Error Rate)	STT 결과와 정답 텍스트의 일치도를 문자 단위로 비교하여 전사 오류율을 정량적으로 계산

3.4.2 파인튜닝 모델 검증 도구

구분	도구	용도
모델 출력 검증	Python 평가 스크립트	모델 출력값을 테스트 케이스 단위로 실행하여 정확도 및 누락 항목 검증

4. 테스트 항목

평가 방식	평가 항목	항목 설명	평가 형식
정량적	발화자 추출 정확도	추출된 발화자 수와 실제 회의 참여자 수 같은가?	Pass / Fail
	텍스트 반환 속도 (전문, 요약본, 회의록)	모델의 텍스트 반환 속도가 기준 시간 이하로 소요되는가? (기준 시간 = 음성 길이 * 0.5) - 전문 : 음성 입력이 완료된 시점부터, 모델 응답까지로 판단 - 요약본 : 음성 입력이 완료된 시점부터, 모델 응답까지로 판단 - 회의록 : 음성 입력이 완료된 시점부터, 모델 응답까지로 판단	Pass / Fail
	발화자 식별 정확성	- 5점 : 80% 이상의 발화가 올바른 발화자에게 할당됨 - 4점 : 70% 이상의 발화가 올바르게 할당되었으나 오류가 존재함 - 3점 : 60% 이상의 발화가 일치하나 다수의 오류가 존재함 - 2점 : 50% 이상의 발화가 발화자에게 식별 되지않음 - 1점 : 대부분의 발화가 잘못된 발화자에게 할당됨 - 0점 : 발화자 구분 불가 또는 전면적 오인식됨	점수형 (5점 만점)
	전문 내용 일치도	- 5점 : CER = 0% - 4점 : 0% < CER <= 5% - 3점 : 5% < CER <= 15% - 2점 : 15% < CER <= 30% - 1점 : 30% < CER <= 50% - 0점 : CER > 50%	점수형 (5점 만점)

정성적	요약본 생성 품질	- 5점 : 회의 요약이 5W3H(Who, What, When, Where, Why, How, How much) 구조를 완전하게 포함하며, 문장 표현이 자연스럽고 500자 이내로 반환됨 - 4점 : 5W3H 항목 중 1개 이하가 누락되었으나, 전체 회의 흐름과 주요 내용은 명확히 전달되고 500자 이내로 반환됨 - 3점 : 5W3H 항목 중 2~3개가 누락되거나, 일부 문장이 불명확해 요약의 논리적 흐름이 다소 어색하고 500자를 넘음 - 2점 : 5W3H 구성요소의 절반 이상이 누락되거나, 요약 내용이 단편적으로 제시되어 회의 핵심 파악이 어려움 - 1점 : 회의 요약이 단순 발화 요약 수준에 그치며, 5W3H 구조를 거의 따르지 않음 - 0점 : 요약이 생성되지 않거나, 5W3H 구조와 무관한 내용으로 작성됨	점수형 (5점 만점)
	태스크 추출 정확도	- 5점 : 추출된 태스크가 회의 내용과 일치하고 회의 중 언급된 who / what / when 중 명시 가능한 정보를 모두 포함함 - 4점 : 대부분의 태스크가 정확히 추출되었으나 일부 태스크에서 언급된 who가 추출되지 않음 - 3점 : 주요 업무는 추출되었으나 언급된 who / what / when 중 두 항목 이상이 누락되거나 표현이 모호함 - 2점 : 태스크가 부분적으로만 추출되었거나 언급된 필수 요소(what / when)가 다수 누락됨 - 1점 : 태스크 추출 결과가 회의 내용과 불일치하거나 의미 있는 업무 정보로 해석 불가능함 - 0점 : 태스크 추출이 전혀 수행되지 않고 결과가 완전히 무의미함	점수형 (5점 만점)
	회의록 세부 내용 정확도	- 5점 : 추출된 안건과 결정된 사항이 사용자의 정답 데이터와 전부 일치하며 요약본과 매칭됨 - 4점 : 추출된 안건과 결정된 사항이 사용자의 정답 데이터와 80% 이상 일치하며, 일치하지 않는 안건이 주제와 연관성이 있음 - 3점 : 추출된 안건과 결정된 사항이 사용자의 정답 데이터와 60% 이상 일치하며, 일치하지 않는 안건이 주제와 연관성이 없음 - 2점 : 추출된 안건과 결정된 사항이 사용자의 정답 데이터와 60% 이상 일치하며, 일치하지 않는 안건이 주제와 연관성이 없음. 추출된 안건과 요약본이 80% 이하로 매칭됨 - 1점 : 추출된 안건과 결정된 사항이 사용자의 정답 데이터와 60% 이상 일치하며, 일치하지 않는 안건이 주제와 연관성이 없음. 추출된 안건과 요약본이 60% 이하로 매칭됨	점수형 (5점 만점)

		- 0점 : 추출된 안건과 결정된 사항이 사용자의 정답 데이터와 40% 이하로 일치하며, 일치하지 않는 안건이 주제와 연관성이 없음.	

5. 테스트 케이스

시나리오 코드	SCR-MEET-001		
시나리오 이름	STT 전사 정확도 (회의 전문)		
설명	회의 녹음 파일을 STT 모델로 변환했을 때, 실제 발화 내용과 얼마나 정확히 일치하는지 평가한다.		
흐름	입력	예상 결과	실제 결과
	“다음 회의는 내일 오전 10시에 시작하겠습니다.”	정상 인식 *“다음 회의는 내일 오전 10시에 시작하겠습니다.”*	“다음 회의는 내일 오전 10시에 시작하겠습니다.”
예외 흐름	입력	예상 결과	실제 결과
	발음이 불명확한 경우 *“그거 내일 …(불명확 구간) 그거 있잖아”*	*“그거 내일 그거 있잖아”*	그거 내일 그거 있잖아
	배경 소음이 있는 경우 (회의 중 타이핑 소리, 웃음 등) *“프로젝트 일정은 다음 주 수요일이에요”*	*“프로젝하하하 일정은 하하 수요일이에요”*	프로젝트 일정은 다음 주 수요일이에요
	동시에 발화한 경우 *A :* “이 부분 수정할게요” */ B :* “네 알겠습니다”	“이 부분 수정네할게요 알겠습니다”	네 알겠습니다
	외래어나 약어 발음 인식 오류 “다음 미팅은 줌으로 진행합시다.”	“다음 미팅은 중으로 진행합시다.”	다음 미팅은 줌으로 진행합시다

평가 항목			
발화자 추출 정확도	텍스트 반환 속도 **(**전문 추출**)**	발화자 식별 정확성	전문 내용 일치도
Pass (명 수 동일)	Pass (6.38s → 39.6s)	5점	5점

시나리오 코드	SCR-MEET-002		
시나리오 이름	회의 전문 요약 정확도 평가		
설명	회의 전문에서 5W3H 구조를 기반으로 핵심 내용을 추출하고, 원문 의미를 유지하며 간결하게 요약하는 정확도를 평가한다.		
흐름	입력	예상 결과	실제 결과
	정상 요약 “다음 주 수요일 오전 *10*시에 신제품 기획 회의를 진행한다. 회의 주제는 신규 기능 제안 검토이며, 김민수 과장이 발표를 맡는다. 박지연 대리가 회의실을 예약하고, 이은호 부장이 참석자 명단을 정리한다.”	다음 주 수요일 오전 10시, 회사 회의실에서 신제품 기획 회의(신규 기능 제안 검토) 진행 김민수 과장 : 발표 진행 박지연 대리 : 회의실 예약 이은호 부장 : 참석자 명단 정리	"who": "김민수 과장", "what": "신규 기능 제안 검토", "when": "다음 주 수요일 오전 10시", "where": "회의실", "why": "신제품 기획 회의", "how": "발표", "how_much": "null", "how_long": "null" "who": "박지연 대리", "what": "회의실 예약", "when": "다음 주 수요일", "where": "회의실", "why": "신제품 기획 회의", "how": "예약", "how_much": "null", "how_long": "null"
예외 흐름	입력	예상 결과	실제 결과
	핵심 요소 누락 “박지연 대리가 회의실을 예약하고, 김민수 과장이 발표를 준비한다.”	누락: Who, When “발표 준비 예정”	"who": "박지연 대리", "what": "회의실 예약", "when": "null", "where": "회의실", "why": "null", "how": "예약", "how_much": "null", "how_long": "null" "who": "김민수 과장", "what": "발표 준비", "when": "null", "where": "null", "why": "발표 준비", "how": "발표", "how_much": "null", "how_long": "null"
	불필요한 세부 포함 “회의는 오전 *10*시에 진행됩니다. 끝나고 점심은 근처 식당에서 먹을 예정입니다.”	비관련 내용 포함 “회의와 점심 일정 포함됨.”	"who": "null", "what": "회의", "when": “오전 10시”, "where": "null", "why": "회의 진행", "how": "발표", "how_much": "null", "how_long": "null"
	시제 오류 “회의는 어제 진행되었습니다.”	시제 불일치 “회의 예정”	미추출

텍스트 반환 속도 (요약 / 태스크)	요약본 생성 품질
평균 3.6s	5/5점

시나리오 코드	SCR-MEET-003		
시나리오 이름	태스크 추출 정확도 평가		
설명	STT로 전사된 회의 전문을 입력으로 하여 모델이 회의 중 언급된 태스크 정보를 정확히 식별하고 담당자(Who), 업무 내용(What), 기한(When) 등의 핵심 정보를 올바르게 추출하는지를 평가한다.		
흐름	입력	예상 결과	실제 결과
	“다음 주 수요일 오전 *10*시에 신제품 기획 회의를 진행한다. 회의 주제는 신규 기능 제안 검토이며, 김민수 과장이 발표를 맡는다. 박지연 대리가 회의실을 예약하고, 이은호 부장이 참석자 명단을 정리한다”	정상 추출 Who: 김민규 What: 신제품 기획 회의 준비 When: 다음 주 수요일 10시 Who: 박지연 What: 회의실 예약 When: 다음 주 수요일 Who: 이은호 What: 참석자 명단 정리 When: 다음 주 수요일	"description": "신제품 기획 회의 준비 및 일정 설정", "assignee": "김민수", "due": "다음 주 수요일 오전 10시", "due_date": "null" "description": "회의실 예약 및 참석자 명단 정리". "assignee": "박지연", "이은호", "due": "*", "due_date": "null"
예외 흐름	입력	추출된 태스크	실제 결과
	담당자 누락 “다음 주 금요일까지 시제품 테스트 결과를 정리해야 합니다.”	Who: 인식 실패 What: 시제품 테스트 결과 정리 When: 다음 주 금요일	"description": "테스트 결과 정리", "assignee": "개발팀원", "due": "다음 주 금요일", "due_date": "null"
	마감기한 누락 “이은호 부장이 곧 디자인 시안을 업데이트할 예정입니다.”	Who: 이은호 What: 디자인 시안 업데이트 When: 인식 실패	"description": "디자인 변경사항 확인 및 처리". "assignee": "이은호 부장". "due": "*", "due_date": "null"
	태스크 누락 “박 과장님 그거 월요일까지…” *(잡음 포함)*	Who: 박 과장 What: 인식 불가 When: 월요일	"description": "업무 처리", "assignee": "박 과장". "due": "월요일", "due_date": "2025-12-01"

태스크 추출 정확도
4 / 5점 (태스크 누락 시, 업무 처리 등으로 처리)

시나리오 코드	SCR-MEET-004		
시나리오 이름	도메인 특화 용어 정확도 평가		
설명	회의 전문을 기반으로 생성된 요약문이 도메인(IT) 특화 용어를 올바른 의미로 해석하고 유지했는지 평가한다.		
흐름	입력	예상 결과	실제 결과
	“이번 회의에서는 백엔드 서버의 *API* 응답 속도 개선 방안을 논의했습니다. 프론트엔드 팀과 협업하여 캐시 로직을 수정하기로 했습니다.”	정확한 용어 유지 요약: 백엔드 API 응답 속도 개선을 위해 프론트엔드 팀과 캐시 로직 수정 논의	"who": "백엔드 팀", "what": "API 응답 속도 개선", "when": "이번 회의", "where": "회의실", "why": "응답 속도 저하", "how": "캐시 로직 수정", "how_much": "null", "how_long": "null"
예외 흐름	입력	예상 결과	실제 결과
	도메인 용어 오인식 “데이터 레이크 구축 일정은 다음 주까지 확정해야 합니다.”	‘데이터 호수 구축 일정’으로 오인식	"who": "개발자". "what": "데이터 레이크 구축 관련 문서 검토 및 리뷰", "when": "다음 주까지", "where": "null", "why": "null", "how": "null", "how_much": "null", "how_long": "다음 주 내"
	의미 왜곡 “이번 분기에는 *ERP* 시스템을 *SAP*로 전환합니다.”	‘회계 시스템을 새 프로그램으로 변경’으로 의미 단순화	"who": "개발자", "what": "ERP 데이터베이스 업데이트 및 코드 리팩토링"
	도메인 혼동 “모델 학습 결과, *F1 Score*가 상승했습니다.”	‘F1 경기 점수 상승’으로 인식	"who": "데이터 분석가", "what": "새 모델 학습 결과 검토 및 F1 Score 상승 확인"

평가 항목
도메인 특화 용어 정확성
Pass

6. 테스트 결과

6.1 테스트 결과 개요

본 테스트에서 STT/화자 분리 모델과 파인튜닝 모델에 대해 기능성, 정확도, 성능을 종합적으로 검증하였다. 전반적으로 높은 수준의 정확도를 달성하였으며, 주요 정량적 평가 항목에서 설정된 기준 점수 이상을 기록하였다.

6.2 STT 및 화자 분리 성능 분석

STT 및 화자 분리 모델의 경우, 성능 지표에서 전반적으로 높은 등급의 결과를 보이며 시스템의 안정성을 입증하였다.

• 전문 내용 일치도: Whisper 모델을 사용한 전사 정확도 평가에서 CER 0%를 달성하며, 정량적 평가 항목에서 5점 만점을 기록하였다. 이는 전사 결과가 실제 발화 내용과 의미 왜곡이나 누락 없이 완전히 일치함을 의미한다.

• 화자 식별 정확성: Pyannote 모델 기반으로 한 발화자 식별 정확성 역시 5점 만점을 기록하여 80% 이상의 발화가 올바르게 할당된 높은 수준의 화자 분리 성능을 확인하였다.

6.3 모델 요약 및 태스크 추출 성능 분석

모델 요약 및 태스크 추출 기능도 높은 등급의 결과를 보이며 시스템의 성능을 입증하였다.

• 요약본 생성 품질: 5W3H 구조를 포함한 요약본을 성공적으로 생성하였다.

• 태스크 추출 정확도: 평가 결과, 언급된 태스크들을 정확하게 추출해주지만, 정보가 불완전한 발화에 대해, 모델이 태스크가 명확히 정의되지 않았음에도 불구하고 태스크를 추출하려는 경향을 보였다.

6.4 주요 오류 분석 및 개선 방안

시나리오 SCR-MEET-003의 태스크 추출 정확도 평가에서 필수 요소인 “What(업무 내용)”을 인식하지 않고 단순히 “업무 처리”와 같은 일반적인 값으로 대체하여 반환하였다. 이러한 모호한 정보에 대한 불필요한 태스크 추출이 정확도를 떨어뜨리는 주요 원인으로 분석된다. 이는 모델 프롬프트 엔지니어링 단계에서 필수 정보 누락 시, 해당 태스크를 필터링 하는 방식으로 개선하고자 한다. 명확한 태스크가 언급되지 않으면 태스크 추출을 하지 않도록 수정할 계획이다.