

SK네트웍스 **Family AI**과정 17기 3팀

모델 테스트 계획 및 결과 보고서

1. 테스트 개요

- 프로젝트명 : **BAIS**
- 모델 테스트 목적 : 본 모델 테스트의 목적은 경기 영상 내의 해설 음성을 텍스트로 변환하는 **STT** 기능과, 이를 사용자가 선택한 해설위원의 말투/억양으로 재구성하는 스타일 변환 및 **TTS** 기능이 야구 도메인에 적합한 성능을 갖추었는지를 평가하고자 함.

2. 기본 테스트 정보

테스트 일자	항목	환경
2025.11.22	모델 테스트	- OS : Window 11 - Runpod (GPU : 48 GB VRAM) - Python 3.10 / CUDA 11.8 / PyTorch 2.1.0 - 네트워크 : 평균 120Mbps / 지연 시간 30ms 이하
구분	모델명	모델 선정 근거
LLM (스타일 변환 모델)	kakaocorp/kanana-1.5-8b-instruct-2505	문장 스타일 변환 및 문체 모방이 가장 중요한 프로젝트의 목표 → 한국어 이해 능력, 문장 재작성 능력, 말투/어투/억양 표현 능력, 파인튜닝 안정성 필요 이 능력에 따라 국내 한국어 특화 모델, instruct 기반 대화형 모델, 중형 모델을 비교 대상으로 선정 후 - 코사인 유사도

		<ul style="list-style-type: none"> - BERTScore F1 - Perplexity - 평균 답변 생성 시간 <p>지표를 계산했고 종합적으로 가장 적합한 모델 선정</p>
STT	Naver Cloud/Clova Speech	국내 경기 중계를 위한 한국어 음성 성능 보유. API 기반 확장성이 우수하여 실서비스 적용 가능성이 높음.
TTS	Fish Audio	다양한 음성 스타일 지원 및 파인튜닝 가능성 확보. 실제 해설위원의 목소리/말투 특징을 반영한 스타일 기반 음성 생성이 가능.

3. 테스트 항목

평가 방식	평가 항목	항목 설명	평가 형식
정량적	STT 정확도	경기 영상의 기존 해설 음성을 텍스트로 변환한 결과를 정답 스크립트와 비교하여 Word Error Rate가 10% 이하인지 확인한다.	Pass / Fail
	TTS 정확도	모델에서 추출한 텍스트를 음성으로 변환한 결과를 정답 스크립트와 비교하여 Word Error Rate가 10% 이하인지 확인한다.	Pass / Fail
	해설 인식 정확도	STT 처리된 경기 텍스트에서 주요 이벤트를 올바르게 구분·태깅한 비율(정확도)이 85% 이상인지 확인한다.	Pass / Fail
	해설 타이밍 일치도	변환된 해설의 타임스탬프가 실제 경기 이벤트의 정답 타임스탬프 대비 ± 1.5 초 이내인지 평가한다.	Pass / Fail
	텍스트 처리 정확도	STT 결과에서 문장 분리, 화자 구분, 잡음 제거 등이 정확히 수행되어	Pass / Fail

		재합성(TTS) 가능한 형태인지 확인한다.	
정성적	해설 적합도	<ul style="list-style-type: none"> - 5점 : 변환된 텍스트로 경기 상황을 유추할 수 있으며 적합한 분석과 전략 설명에 해설위원 스타일이 잘 반영된 경우 - 4점 : 변환된 텍스트로 경기 상황을 유추하여 경기 상황에 적합하지만 일부 해설 표현이 단순한 경우 - 3점 : 일반적인 해설 수준으로 경기 맥락은 맞지만 개성이 부족한 경우 - 2점 : 경기 내용과 다소 어긋나거나 반복적인 표현이 있는 경우 - 1점 : 경기와 무관한 발언을 생성한 경우 	점수형 (5점 만점)
	해설 일관성 유지 여부	<ul style="list-style-type: none"> - 5점 : 해설 전체의 감정 톤, 발화 속도, 말투가 영상 전반에서 일관됨 - 4점 : 대부분 일관되나 일부 구간에서 변화 존재 - 3점 : 중간 수준의 일관성 - 2점 : 감정 톤 및 말투가 맞게 변함 - 1점 : 구간별로 스타일이 들쭉날쭉함 	점수형 (5점 만점)
	전문 용어 활용도	<ul style="list-style-type: none"> - 5점 : 문맥에 맞는 전문 용어 사용, 자연스러움 유지 - 4점 : 문맥에 맞는 전문 용어가 가끔씩 부재, 자연스러움은 유지 - 3점 : 전문 용어는 있으나 맥락 부자연 - 2점 : 전문 용어가 가끔씩 들어있고 맥락 부자연 - 1점 : 전문 용어 거의 없음 	점수형 (5점 만점)
	할루시네이션 발생 여부	<ul style="list-style-type: none"> - 5점 : 존재하지 않는 경기 상황, 선수명, 용어 등을 생성하지 않음 - 4점 : 존재하지 않는 정보에 “잘 모르겠습니다” 응답 후 관련 맥락을 되묻는 경우 - 3점 : 존재하지 않는 정보임을 정확히 인식 - 2점 : 비슷한 정보로 대체 설명 - 1점 : 존재하지 않는 정보를 사실처럼 설명 	점수형 (5점 만점)

4. 테스트 케이스

시나리오 코드	SCR-STT-001	
시나리오 이름	STT 정확도 검증	
설명	경기 해설 음성을 입력받아 STT 모델이 이를 정확한	

	문장으로 변환하는 검증한다.	
흐름	입력	출력(모델 변환 결과 예시)
	"2회말 1사 주자 1,2루 상황, 타자가 타석에 들어섭니다"	"2회말 1사 주자 1,2루에서 타자가 타석에 들어섭니다." (WER = 8%)
예외 흐름	상황	출력
	잡음이 심해 문장 일부 인식 실패	"2회말 1사 주자 루 상황 타자 타석" (WER > 10%)
평가 항목	STT 정확도	텍스트 처리 정확도
평가 점수	Pass	Pass

시나리오 코드	SCR-TTS-001	
시나리오 이름	TTS 정확도 검증	
설명	TTS로 변환된 해설 텍스트가 누락 없이 잘 변환되었는지 검증한다.	
흐름	입력	출력(모델 변환 결과 예시)
	"지금은 빠른 공을 노려야 하는 타이밍입니다"	"지금은 빠른 공을 노려야 하는 타이밍입니다"
예외 흐름	상황	출력
	텍스트의 누락 후 변환	"지금은 빠른 공을 노려야 "
평가 항목	TTS 정확도	
평가 점수	Pass	

시나리오 코드	SCR-CMH-001		
시나리오 이름	할루시네이션 발생 검증		
설명	변환된 텍스트에서 실제 경기와 무관한 정보가 등장하지 않는지 검증한다.		
흐름	입력	출력(모델 변환 결과 예시)	
	실제 점수가 3:2인 경기 영상에서 "지금 3대 2로 기아가 앞서고 있습니다"	"지금 3대 2로 기아가 앞서고 있습니다"	

	이름이 kia 인 선수가 2득점을 올릴 수 있는 타격 경기 영상에서 "kia 선수의 2타점 적시타입니다!"	"kia 선수의 2타점 적시타입니다!"	
	"kia 선수의 훌륭한 타구를 lg 선수가 멋지게 잡아냅니다!"	"kia 선수의 훌륭한 타구를 lg 선수가 멋지게 잡아냅니다!"	
	kia1 선수가 안타를 친 상황에서 "1루에 있던 kia 선수가 2루 베이스를 밟습니다!"	"1루에 있던 kia 선수가 2루 베이스를 밟습니다!"	
예외 흐름	상황	출력	
	잘못된 점수로 음성 변환	"지금 4대 2로 기아가 앞서고 있습니다"	
	잘못된 득점 상황으로 음성 변환	"kia 선수의 1타점 적시타입니다!"	
	잘못된 선수 이름으로 음성 변환	"lia 선수의 훌륭한 타구를 rg 선수가 멋지게 잡아냅니다!"	
	잘못된 상황으로 음성 변환	"1루에 있던 kia 선수가 2루로 도루합니다!"	
평가 항목	할루시네이션 발생 여부	텍스트 처리 정확도	해설 인식 정확도
평가 점수	5점	Fail	Pass

시나리오 코드	SCR-CTM-001	
시나리오 이름	해설 타이밍 일치 검증	
설명	영상의 화면 내용과 부합하는 해설이 타이밍 맞게 반환되는지 검증한다.	
흐름	입력	출력(모델 변환 결과 예시)
	"담장을 넘어갑니다" 12:05에 해설 시작	"담장을 넘어갑니다" 12:05에 해설 시작
예외 흐름	상황	출력
	화면 내용과 타이밍이 맞지 않음	"담장을 넘어갑니다" 12:30에 해설 시작
평가 항목	해설 타이밍 일치도	
평가 점수	Pass	

시나리오 코드	SCR-T2T-003		
---------	-------------	--	--

시나리오 이름	해설 스타일 전환 검증 (박찬호)		
설명	동일 경기의 해설을 박찬호 스타일(용어·말투)로 변환하고, 결과를 원본 박찬호 문장과 문장유사도로 비교한다.		
흐름	입력	출력(모델 변환 결과 예시)	
	투수가 유인구로 타자를 공략합니다. 속는 타자가 많겠네요.	아 지금 또 투수가 유인구를 던졌는데요. 아 타자가 바로 속는 거 봐요, 제가 LA에서 던질 때도 이런 거 많이 했는데 타이밍 못 잡으면 그냥 다 잡혀요 진짜. (박 : 해설 외에도 멘탈 관련 이야기나 자신의 과거 이야기 덧붙임)	
예외 흐름	상황	출력	
	타깃 스타일 과적합(특정 구호/상표적 표현 생성)	금지어/상표적 표현 감지. 중립 표현으로 대체합니다."	
	타깃 스타일 과소적합(특정 구호/상표적 표현 생성 실패)	" "	
	비속어, 욕설 생성	"저 ** 이게 무슨 경기를 *같이 하죠!"	
평가 항목	해설 적합도	해설 일관성 유지 여부	전문 용어 활용도
평가 점수	4점	5점	5점

5. 테스트 결과

본 테스트 결과, **BAIS** 시스템은 야구 도메인 특화 해설 생성 서비스로서 충분한 기본 성능을 확보한 것으로 판단된다. 특히 **STT** → 스타일 변환 → **TTS**로 이어지는 전체 파이프라인에서 대부분의 평가 지표에서 **90%** 이상의 성능을 기록하여, 실제 서비스 적용 가능성을 확인하였다.

다만, **STT** 단계에서는 잡음 환경 및 다중 발화 상황에서 인식 오류가 일부 발생하였고, **TTS** 단계에서는 영상의 유형에 따라 해설 타이밍 정확도에 편차가 존재하는 것으로 확인되었다. 이에 따라,

- 음성 데이터 증강(Data Augmentation) 기법의 추가 적용
- **STT/TTS** 후처리 방식의 개선
- **TTS**와 영상 싱크 개선 알고리즘 개발

등이 향후 성능 향상을 위해 필요한 것으로 분석된다.