【서식 3-1】 캡스톤디자인 과제 수행 결과보고서 (※ 학생 작성)_ 팀용

기업연계형 캡스톤디자인 교과목 과제 수행 결과보고서

과제 유형	■ 기업연계기반				
과제명	**LLM 기반 아동 동화 생성에서 서사적 맥락을 활용한 2단계 안전성 평가 시스템**				
팀명	박장대소				
수강 교과목명	캡스톤디자인2		**교과목 학수번호**	DCSS452	
교과목 담당교수	소 속	컴퓨터정보학과	성 명	이웅기	
	E - mail	codingchild@korea.ac.kr	교내전화		
지도교수	소 속	컴퓨터정보학과	성 명	이웅기	
	E - mail	codingchild@korea.ac.kr	교내전화		
산업체 참여 인력(PM)	소 속	업스테이지	성 명	전영훈	
	E - mail	yesica@upstage.ai			
산업체 역할 (자문내용)	Solar API 사용 관련 조언				

구분		**성명**	**학과**	**학년**	**학번**	**E - mail**
참여 학생	**팀장**	박소정	컴퓨터소프트웨어학과	4	2022270661	venus7472@korea.ac.kr
	팀원	박예원	컴퓨터소프트웨어학과	4	2022270663	hasilu85@korea.ac.kr
		장서연	컴퓨터소프트웨어학과	4	2022270647	syjang1016@korea.ac.kr

*이중전공의 경우 본 소속학과(이중전공)으로 표기

위와 같이 규정에 의해 과제를 완료하였음을 결과보고서로 제출합니다.

2026. 05. 21.

지도교수: 이웅기 이웅기 (인 또는 서명)

대표학생: 박소정 박소정

고려대학교 세종 SW중심대학사업단 귀하

작품과제명	LLM 기반 아동 동화 생성에서 서사적 맥락을 활용한 2단계 안전성 평가 시스템
과제 개요	○ 과제 선정 배경 생성형 AI 기술의 발전으로 LLM 기반 아동 콘텐츠 자동 생성 수요가 빠르게 증가하고 있다. 그러나 LLM이 생성하는 콘텐츠에는 욕설·위협·그루밍 유도 등 아동에게 유해한 표현이 포함될 위험이 존재하며, 이는 특정 모델의 문제가 아니라 현재 생성형 AI 전반에서 나타나는 구조적 한계이다. 기존 단일 모델 필터링 방식은 명백한 유해 표현은 잘 탐지하나, 맥락에 의존하는 그루밍 패턴 등은 탐지율이 낮은 한계가 있다. 본 과제에서는 Kakao Corp.의 Kanana 모델과 Upstage의 Solar Pro를 활용하여 아동 동화 자동 생성 및 2단계 안전성 검증 파이프라인을 설계 및 검증한다. ○ 과제의 필요성 6~7세 아동 대상 동화는 일반 콘텐츠보다 엄격한 안전 기준이 요구되며, 단순 문장 단위 필터링만으로는 맥락 의존적 유해 표현을 탐지하기 어렵다. 2단계 LLM-as-Judge 구조를 추가함으로써 FPR(정상 콘텐츠 오탐률)을 낮추면서도 교육적으로 필요한 갈등 장면을 보존할 수 있다. 아동 안전 측면에서 FNR(실제 유해 콘텐츠를 놓치는 비율) 최소화가 최우선 목표이며, kanana-safeguard-8b와 Solar Pro 3의 역할을 명확히 분리한 파이프라인을 설계하고 그 효과를 정량적으로 검증할 필요가 있다.
과제 내용	○ 과제 구성 본 과제는 아동 동화 생성과 안전성 평가를 통합한 2단계 파이프라인으로 구성된다. 1단계에서는 kanana-1.5-8b 모델이 Solar Pro 3가 생성한 기획 힌트(교훈 방식, 등장인물 설정, 결말 방향 등)를 참조하여 동화 초안을 생성한다. 2단계에서는 kanana-safeguard-8b가 문장 단위 유해 표현을 1차 탐지하며, S1~S7 카테고리로 분류한다. 이때 성적 콘텐츠·아동 성착취·자살 및 자해에 해당하는 HARD 카테고리(S3·S5·S6)가 탐지되면 Solar Pro 3가 해당 카테고리를 인지하고 반드시 해당 부분을 수정하도록 강제 지시한다. 나머지 욕설·혐오·범죄·잘못된 정보에 해당하는 SOFT 카테고리(S1·S2·S4·S7)는 Solar Pro 3가 동화 전체 맥락을 고려하여 수정 여부를 판단한다. 3단계에서는 Solar Pro 3가 동화 전체 맥락을 기반으로 최종 SAFE/UNSAFE를 판정하고, 불합격 시 구체적인 수정 지시(rewrite_instructions)를 생성하며, kanana-1.5-8b가 이를 힌트로 동화를 재생성한다. 기존 계획과 비교하여 가장 큰 변경점은 두 가지이다. 첫째, 생성 모델이 kanana-nano-2.1b에서 kanana-1.5-8b로 변경되었다. 둘째, 기존에는 Solar Pro 3가 동화 기획·판단·재생성을 모두 담당하였으나, 현재는 Solar Pro 3가 기획 힌트 제공과 판단·수정 지시 역할만 담당하고 실제 재생성은 kanana-1.5-8b가 수행하는 구조로 전환되었다. 우리가 생성하고자 하는 모델의 이름은 SETA(Safety Evaluation with Two-stage and nArrAtive context)라 한다. ○ 과제 주요 특징 본 과제의 핵심 특징은 계층적 2단계 안전성 검증 구조, 휴먼 어노테이션 기반 필터링 성능 평가, CSM 프레임워크 기반 품질 평가 설계에 있다. 안전성 검증은 kanana-safeguard-8b와 Solar Pro 3의 역할을 명확히 분리한 2단계 구조로 이루어진다. 1차 탐지를 담당하는 kanana-safeguard-8b는 동화를 문장 단위로

순차 입력하여 S1~S7 카테고리로 분류하며, 유해 표현이 포함된 문장은 태깅을 하여 2차 탐지로 전달한다. 이때 성적 콘텐츠·아동 성착취·자살 및 자해에 해당하는 HARD 카테고리(S3·S5·S6)는 Solar Pro 3가 반드시 해당 부분을 수정하도록 강제하며, 욕설·혐오·범죄·잘못된 정보에 해당하는 SOFT 카테고리(S1·S2·S4·S7)는 Solar Pro 3가 동화 전체 맥락을 고려하여 수정 여부를 자율적으로 판단한다. 이 구조는 명백히 위험한 표현은 반드시 걸러내면서도, 교육적으로 필요한 갈등 장면은 맥락에 따라 보존할 수 있도록 설계되었다.

필터링 성능 평가는 연구자가 직접 동화를 읽고 SAFE/UNSAFE를 태깅한 휴먼 어노테이션 테스트셋을 기반으로 수행하였다. 사람이 읽었을 때 불안감을 줄 수 있는 문장이 하나라도 포함된 동화는 전체를 UNSAFE로 분류하였으며, 총 61편의 동화를 평가 데이터로 사용하였다. kanana-safeguard-8b 단독, Solar Pro 3 단독, 2단계 구조 세 가지 시스템을 동일한 테스트셋에 적용하여 FNR·FPR·F1·Accuracy를 비교하였다. 아동 안전 도메인의 특성상 실제 유해 콘텐츠를 놓치는 비율인 FNR을 최우선 지표로 삼았으며, 이상적인 결과는 2단계 구조의 FNR이 세 시스템 중 가장 낮은 것이다.

동화 품질 평가에는 Common Sense Media(CSM) 프레임워크를 기반으로 서사적 맥락, 아동 모델링, 도덕 메시지, 편견·고정관념, 언어 표현, 교육적 가치 6개 항목을 1~5점으로 채점하며, 6~7세 아동이 일상에서 실천 가능한 구체적 행동 모델 제시를 핵심 기준으로 삼는다. 6개 항목 평균 4.5점 이상, 항목별 최저 4.0점 이상을 합격 기준으로 하며, 불합격 시 Solar Pro 3가 수정 지시를 생성하고 kanana-1.5-8b가 재생성하는 피드백 루프가 최대 5회 반복된다.

○ 성능 평가

지표	kanana-safeguard-8b 단독	Solar Pro 3 단독	SETA 2단계 구조
Accuracy	0.6230	0.7213	0.7869 최고
F1	0.6230	0.2609	0.6667 최고
Precision	0.4524	0.7500 최고	0.6500
Recall	1.000	0.1579	0.6842
FPR(거짓 양성률)	0.5476	0.024 최저	0.1667
FNR(거짓 음성률)	0.000*	0.8421	0.3158 균형 최적

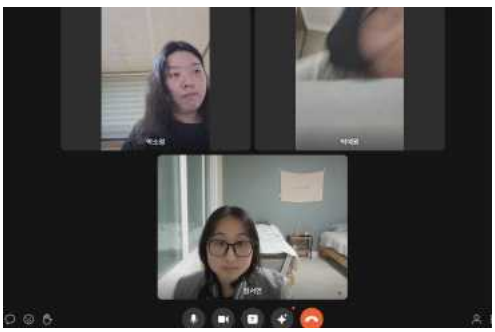
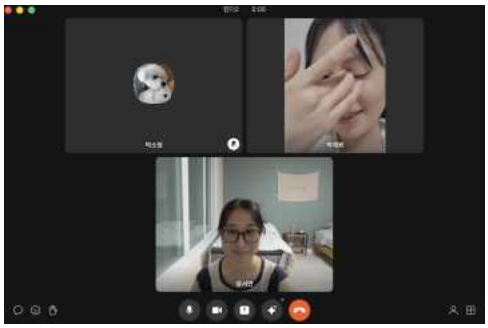
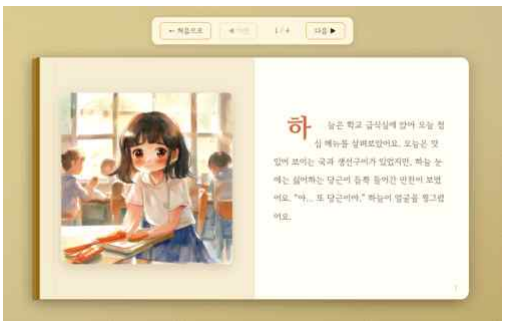
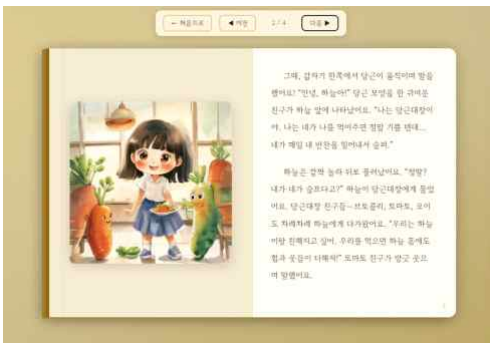
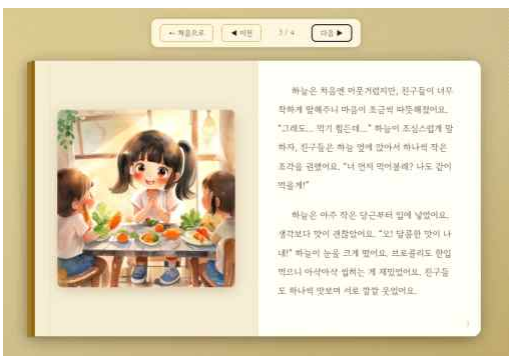
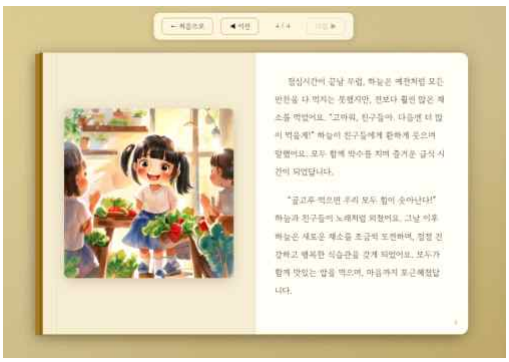
* 기본 카나나 FNR=0.000은 모든 동화를 UNSAFE로 과탐지한 결과로, 실질적 탐지 능력이 아님.

결과물의 활용방안 및 기대효과

○ 결과물의 활용방안 및 기대효과

본 과제의 결과물은 kanana-1.5-8b 기반 아동 동화 생성 서비스의 실용적 안전성 검증 파이프라인으로 즉시 적용 가능하며, 다른 아동 콘텐츠 생성 서비스에도 확장 적용할 수 있다. 2단계 필터링 구조는 특정 모델에 종속되지 않아 kanana-safeguard-8b 및 판단 모델을 교체하더라도 동일한 방식으로 운용 가능하며, CSM 프레임워크 기반 품질 평가 체계 역시 동화 외 다양한 아동 교육 콘텐츠 생성 시스템에 재활용할 수 있다. 특히 유해 표현 탐지와 교육적 갈등 장면 보존을 동시에

	달성하는 구조는 아동 대상 생성형 AI 서비스 전반에서 안전성과 서사 품질의 균형을 맞추는 데 기여할 수 있다. 국내 AI 기본법 및 EU AI Act 등 아동 대상 AI 서비스에 대한 안전 규제가 전 세계적으로 강화되는 추세에서, 본 파이프라인은 규제 준수(compliance)의 실질적 근거 자료로 활용될 수 있다. 특히 본 과제는 FNR·FPR·F1·Accuracy 등 정량적 안전성 지표를 산출하는 구조를 갖추고 있어, 규제 기관이 요구하는 안전성 검증 리포트를 체계적으로 생성하고 제출하는 자동화 시스템으로 확장하는 것이 가능하다. 이는 아동 대상 AI 서비스를 운영하는 기업이 사전 규제 심사나 사후 감사에 대응할 때 실질적인 증빙 수단으로 기능할 수 있으며, 안전성 검증 과정을 문서화하는 표준 절차로 자리잡을 수 있다. 또한 보호자와 교사 입장에서 AI가 생성한 콘텐츠에 대한 신뢰는 여전히 낮은 것이 현실이다. 본 파이프라인은 단순히 유해 표현을 필터링하는 데 그치지 않고, 안전성 판정 근거와 수정 이력을 투명하게 추적할 수 있는 구조를 갖추고 있어 서비스 수용성을 높이는 데 기여할 수 있다. 검증 결과를 보호자나 교육 현장에 공개 가능한 형태로 제공한다면, AI 생성 동화에 대한 신뢰를 제고하고 실제 교육 환경에서의 도입 장벽을 낮추는 효과를 기대할 수 있다. ○ 한계점 다만 본 과제는 몇 가지 한계를 가진다. 첫째, 평가에 사용된 테스트셋이 61편으로 규모가 제한적이며, 동화의 주제와 문체 다양성이 충분히 확보되지 않아 일반화 가능성에 제약이 있다. 추후 더 많은 동화를 수집하고 다양한 유해 표현 유형을 포괄하는 대규모 데이터셋 구축이 필요하다. 둘째, 현재 평가 기준은 6~7세 아동을 단일 독자층으로 설정하고 있으나, 실제 아동 콘텐츠는 연령대별로 적절한 표현 수준과 허용 범위가 다르다. 독자 연령·발달 수준·문화적 배경에 따라 평가 기준을 달리 적용하는 맞춤형 평가 시스템을 도입한다면 더 세밀하고 신뢰도 높은 안전성 검증이 가능할 것이다.

<table>
<tr><td rowspan="7">수행 방법</td><td>구분</td><td>성명</td><td>과제 참여 내용(역할)</td></tr>
<tr><td>팀장</td><td>박소정</td><td>프로젝트 총괄 및 일정 관리, LLM 파이프라인 구조 설계, 데이터셋 전처리, Solar Pro·카나나 API 프롬프트 설계, 전체 코드 공동 설계</td></tr>
<tr><td>팀원</td><td>박예원</td><td>이미지 생성 모델 API 연동 코드 설계, 웹 서비스 프론트엔드 개발, 발표 자료 제작, 전체 코드 공동 설계</td></tr>
<tr><td>팀원</td><td>장서연</td><td>데이터셋 전처리, 카나나 세이프가드 파인튜닝, 이미지 생성 모델 API 연동 코드 설계, 전체 코드 공동 설계</td></tr>
<tr><td>팀원</td><td></td><td></td></tr>
<tr><td>팀원</td><td></td><td></td></tr>
<tr><td>팀원</td><td></td><td></td></tr>
<tr><td>결과물</td><td colspan="3">  
  
   </td></tr>
</table>