고려대학교

LLM-기반 동화 생성 플랫폼 개발

유해 콘텐츠 필터링을 적용한 아동 맞춤형 멀티모달 AI 서비스

지도교수: 서민석 교수님
2022270661 박소정
2022270663 박예원

DeepMind

1 BACKGROUND & MOTIVATION

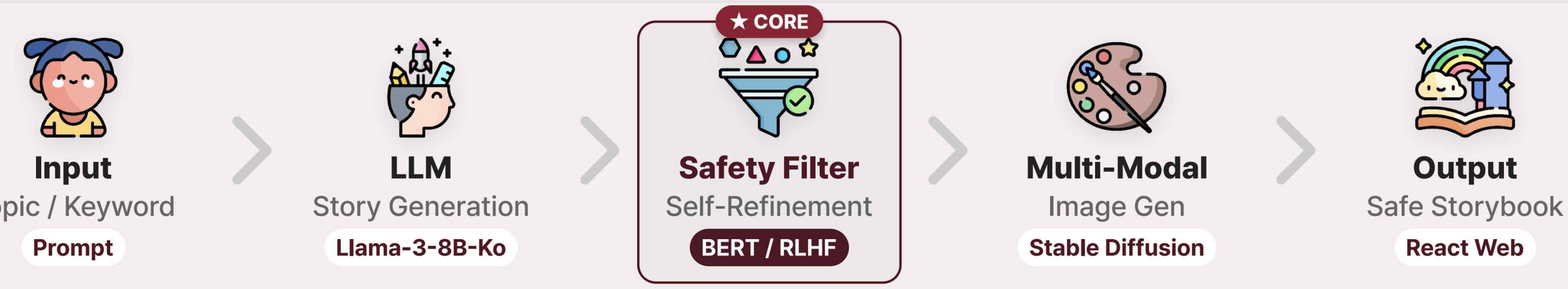
Problem Statement

- 미디어 플랫폼 알고리즘에 의한 **아동 유해 콘텐츠**(폭력·선전성) 노출 심화
- 기존 LLM의 **환각 현상(Hallucination)** 및 **데이터 편향(Bias)** 으로 인한 부적절한 생성 위험
- 기존 서비스의 **안전 장치(Safety Guardrail)** 부재

Proposed Approach

- **Safety-Aligned**를 적용한 아동 맞춤형 LLM 파인튜닝
- 부모가 신뢰할 수 있는 **윤리적 필터링 파이프라인** 구축
- 교육적 가치를 담은 **멀티모달(Text+Image)** 동화 생성
- Diffusion Model 기반의 **텍스트-이미지 일관성 유지** 기술 적용

2 SYSTEM ARCHITECTURE



3 PROJECT SPECIFICATIONS

</> Env & Lang
Python 3.9, PyTorch, React, FastAPI, Colab Pro+

Dataset
AIHub '동화 줄거리' 10만 건, 자체 유해어 데이터셋

Core Model
Text: Llama-3 (Fine-tuned)
Image: Stable Diffusion

Algorithm
RAG (검색 증강), Safety Alignment, PEFT

★ EXPECTATION

- **Safety-Aligned Ecosystem: 유해 콘텐츠 차단을 넘어선, 아동 친화적 AI 생성 모델의 안전성 기준 확립**
- **Interactive Literacy Tool: 아동의 입력에 반응하는 맞춤형 스토리텔링으로 문해력 및 창의성 증진**
- **Ethical AI Implementation: 생성형 AI의 윤리적 문제를 해결하는 실질적인 필터링 구조 구현**
- **Enhanced Immersion: 텍스트와 이미지의 맥락을 일치시키는 멀티모달로 사용자의 몰입 경험 극대화**

4 PROJECT ROADMAP