AI RESEARCH PROJECT

Kanana 8B 기반 RLHF 실험 리포트

SFT - RM - PPO 파이프라인 설계 및 검증

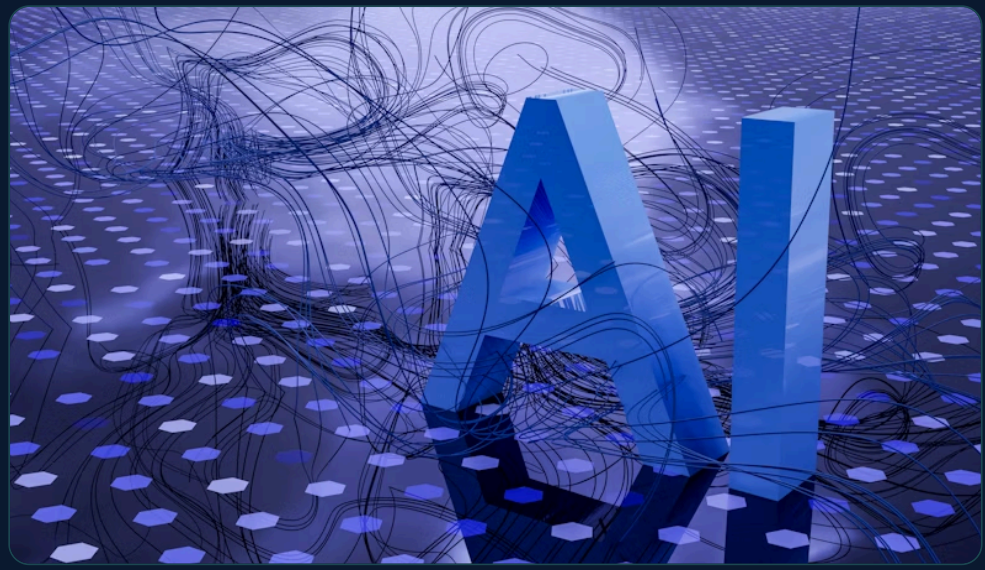
2026.01.21 | 버8로윙 (8조)

PROJECT OVERVIEW

목표: 한국어 지시 이행 능력 고도화

Kanana 8B 모델을 기반으로 사용자의 복잡한 지시를 정확히 수행하고, 인간의 가치관과 선호를 학습에 반영하는 RLHF 최적화 모델을 구축했습니다.

- SFT를 통한 기초 지능 확보
- RM 학습으로 인간의 선호도 수치화
- PPO 알고리즘 기반 최종 정책 최적화



CONCEPT OF RLHF & RANKING DATA

RLHF는 "사람의 선호"를 학습에 반영하여 모델이 **더 사람다운 응답**을 하도록 조정하는 기법입니다.

Q: "환자와 의사소통에서 약사가 주의해야 할 중요한 측면은 무엇인가요?"

RANK 1 (Best)

약사가 환자와 의사소통에서 주의해야 할 중요한 측면은 다음과 같습니다:
1. 명확한 의사소통: 어려운 용어를 피하고 알기 쉽게 설명
2. 개인화된 관리: 환자의 개별 의료 상황 고려
...

RANK 5 (Poor)

주의해야 할 중요한 측면:
1. 명확하고 이해하기 쉬운 언어 사용
2. 적절한 비언어적 의사소통
3. 개인정보 보호

→ **비교 평가:** (A > B)

→ **보상 학습:** RM 모델

→ **강화 학습:** PPO 최적화

DATA CONSTRUCTION & PREPROCESSING

EDA 분석

답변 길이 분포, 회피 문구 및 패턴 분석을 통해 노이즈 데이터를 사전 제거하고 학습 품질을 확보했습니다.

데이터 정규화

Question/Answer Norm 생성 및 data_id 기준 분할을 통해 학습 데이터와 평가 데이터의 오염(Contamination)을 방지했습니다.

포맷 최적화

Kanana Tokenizer 기반으로 모델이 최적으로 학습할 수 있는 JSONL 포맷의 랭킹 데이터셋을 구축했습니다.

BASELINE QUALITATIVE COMPARISON

질문: 기계의 자동화 생산 시스템이 어떻게 생산 효율성을 향상시키나요?

SOLAR-10.7B (Hallucination)

2008년 6월, 3M은 57%의 4900만 ㎡의 生产 空間에 自動化를 完成했음. 2. 2008년 6월, 3M은 57%의... (반복 및 문맥 이탈)

KANANA-8B (High Quality)

기계의 자동화 생산 시스템은 다양한 방식으로 생산 효율성을 향상시킬 수 있습니다. 1. 작업 시간 단축: 인간의 노동력을 대체하여 일관된 생산을 가능하게 합니다. 2. 오류 감소: 정확한 프로그래밍을 통해 실수를 낮춥니다...

EXAONE-1.2B (Summary Style)

기계의 자동화 생산 시스템은 다양한 기술과 프로세스를 통해 생산 효율성을 극대화합니다. 1. 인력 감소 및 인건비 절감... 2. 생산성 향상...

KANANA (Previous Version)

기계의 자동화 생산 시스템은 생산 효율성을 크게 향상시킵니다. 1. 작업 속도 및 처리량 증가: 로봇, 컨베이어 벨트 등을 활용하여 빠른 속도로 작업을 수행합니다.

BASE MODEL EXPERIMENT RESULTS

모델명	파라미터	평가 결과
KANANA-8B	8B	최종 선정 (Best)
EXAONE-1.2B	1.2B	성능 대비 소형 체급
SOLAR-10.7B	10.7B	반복 생성 및 품질 한계

최적 모델 선정

정량적 지표와 정성적 생성 품질을 종합한 결과, **KANANA-8B**가 한국어 지시 이행 프로젝트의 베이스 모델로 가장 적합함을 확인했습니다.

RLHF PIPELINE IMPLEMENTATION

SFT

지도 학습을 통한
기초 지능 주입

RM

선호도 학습 및
보상 모델링

PPO

강화 학습을 통한
정책 최적화

EVAL

정량/정성 평가를 통한
성능 검증

각 단계는 유기적으로 연결되어 최종 응답의 품질을 결정하는 루프 구조를 형성합니다.

CHALLENGES & RESOLUTIONS

이슈 및 한계점

버전 충돌: trl 라이브러리 간의 타입 호환성 및 어댑터 결합 이슈 발생.

메모리 부족: PPO 구조적 한계로 인해 Actor, Critic, Ref 모델 동시 로드 시 VRAM 폭발 현상 발생.

기술적 해결 방안

구조 재설계: 라이브러리 의존성을 전면 재검토하고 베이스 모델 재정의를 통해 충돌 해결.

4-bit 양자화: Actor / Value / Reward / Ref 모델에 **4-bit Quantization**을 적용하여 VRAM 사용량을 극적으로 절감.

MODEL EVALUATION STRATEGY

SFT 정량 평가

- **ROUGE-1:** 단어 중복도 평가
- **ROUGE-L:** 최장 공통 부분 수열 기반의 문장 구조 유사도 평가

상대 비교 평가

GPT-4 모델을 심판(Judge)으로 활용하여 기존 모델 대비 응답의 논리성, 가독성 승률을 측정했습니다.

RLHF 자동 지표

Reward Model의 보상 점수 추이 및 KL Divergence 모니터링을 통해 학습의 안정성을 검증했습니다.

9 QUALITATIVE STYLE METRICS

구조성

논리적 흐름과 구성

밀도

정보의 풍부함과 압축

일관성

문체와 관점의 유지

중복도

표현의 다양성 확보

명확성

의미 전달의 정확성

안전성

유해 정보 생성 방지

중립성

객관적 시각 유지

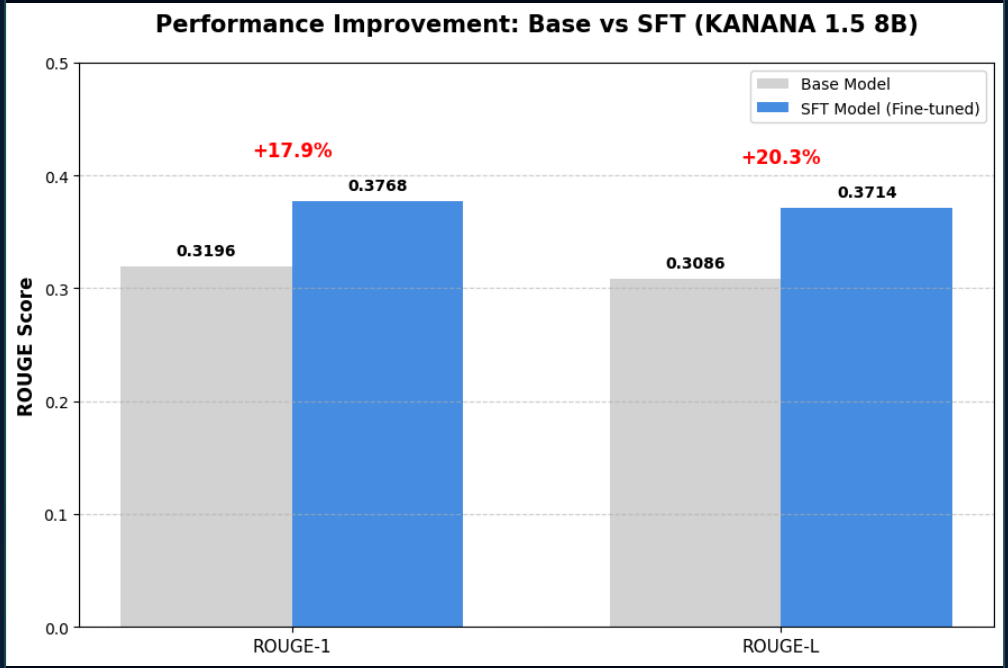
구어체

자연스러운 문장 활용

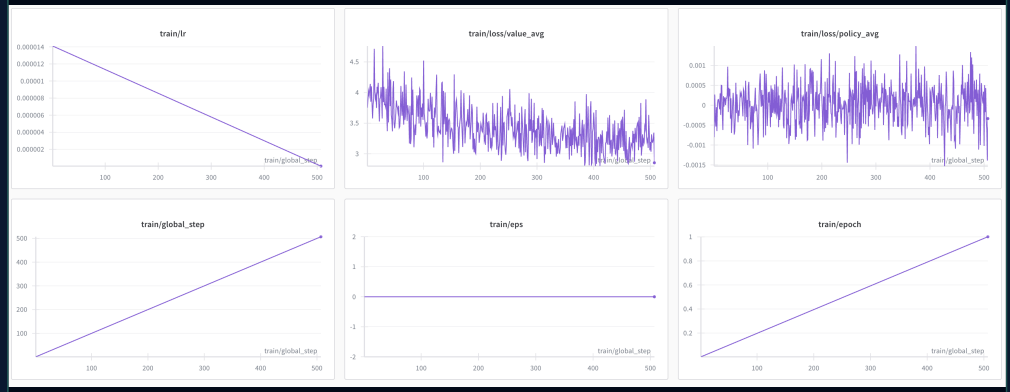
안전 문구

필수 고지 사항 포함

EVALUATION RESULTS



정량 지표: ROUGE Score



정량 지표: Loss 추이 분석

EVALUATION EXAMPLES

RLHF Win Rate : 95% (100 samples)

CASE 01 Q: "동물에 대한 인식을 높이고 교육하는 데 관심 있어?"

Base Response

동물에 대해 배우고 이해하는 것은 우리가 더 책임감 있게 행동하고, 동물 복지를 증진하며, 자연과 조화롭게 살아가는 데 큰 도움이 됩니다.

RLHF Optimized

동물에 대한 이해를 넓히고, 동물 복지와 보호, 생태계의 중요성을 알리는 일은 매우 중요하다고 생각해요.
예를 들어:

CASE 02 Q: "노래를 배워본 적 있어?" (정보형 → 대화형 전환)

Base Response

"혹시 특정 노래를 배우고 싶거나, 음악 이론, 발성법, 또는 노래방에서 잘 부르는 팁 등이 필요하신가요?"

RLHF Optimized

"예를 들어:

- 노래의 가사와 뜻을 설명해 드릴 수 있습니다.
- 노래를 부르는 방법이나 발성 팁을 알려드릴 수 있습니다.
- 악보를 읽는 법이나 악기 연주법에 대해 설명해 드릴 수 있습니다."

CHATBOT INTEGRATION PIPELINE

- User Question 입력
- PromptTemplate 구성 (Policy + Fewshot)
- Chat Prompt 변환 (apply_chat_template)
- KANANA-8B Inference (invoke)
- Output 정리 (Echo 제거 및 .strip())
- 예외 발생 시 Retry 1회 수행

Conclusion & Future Plan

최종 선정 모델

KANANA-8B

최적의 밸런스와 지시 이행력

핵심 성과

RLHF 파이프라인 구축을 통해 단순 SFT 대비 **생성 품질의 비약적 향상**을 달성했습니다.

향후 과제

추가 도메인 특화 학습 및
서빙 속도 최적화(vLLM) 진행 예정

경청해 주셔서 감사합니다.