LLM 및 SLM의 윤리적 표현 특성에 관한 비교 연구

이기호*, 유도진**, 이용준***
*극동대학교 인공지능보안학과
**극동대학교 해킹보안학과
***극동대학교 인공지능보안학과
e-mail:lkheio92@gmail.com

A Comparative Study on The Ethical Expression Characteristics of LLM and SLM

Ki-Ho Lee*, DoJin Ryu**, Yong-Jun Lee***
*Dept. of Artificial Intelligence Security, Far East University
**Dept. of Hacking Security Department, Far East University
***Dept. of Artificial Intelligence Security, Far East University

요 약

본 연구는 GPT-4.0과 DeepSeek-v3, Base 및 SFT를 적용한 Xwin-LM-7B 모델을 대상으로 윤리 시나리오에 대해 응답을 수집한 후, 각 모델이 응답 시 표현에 사용하는 윤리적 특성을 분석하였다. Reward Model(RM)을 활용하여 응답의 윤리적 특성을 정량화하였고, 점수 변화 및 윤리 관점 분포를 정성적으로 분석하였다. 또한, 응답 내의 키워드를 중심으로 덕 윤리, 결과주의, 의무론 관점으로 분류한 다음, 자주 사용하는 표현 구조와 판단 근거를 비교하였다. 연구 결과, SFT 학습 모델은 여러 윤리 관점을 균형 있게 사용하는 모습을 보였고, 응답의 구조와 표현 명확성이 향상되었음을 알 수 있었다. 또한, LLM 모델 중 DeepSeek-v3는 결과주의적인 표현이 자주 나타났으며, GPT-4.0는 중립적인 설명 중심의 응답 구조를 보여주었다. 이러한 결과는 언어 모델의 정렬 방식, 그리고 학습 구조에 따라 윤리 판단 시 표현하는 구조나 방식 등이 달라질 수 있음을 의미한다. 이에 따라 본 연구는 윤리적 인공지능 모델 설계를 위한 평가 기준과 정렬 효과 평가 및 분석의 자료로 활용될 수 있을 것이다.

1. 서론

최근 대규모 언어 모델의 활용은 자연어 처리뿐 아니라 윤리적 판단과 같은 복잡한 의사결정 영역에서도 확대되고 있다. 특히, 다양한 윤리적 선택을 요구하는 시나리오에서의 응답은 인간과 유사한 판단과 그 한계에 대한 실증적 분석의 기회를 제공한다. 그러나 실제로 언어 모델이 어떠한 윤리적 이론을 따르는가, 혹은 학습 방식에 따라 반응이 어떻게 달라지는가를 분석한 연구는 아직 부족하다[1].

윤리성은 정확도와는 다른 평가 기준으로, 이를 정량화할 수 있는 프레임워크는 충분히 정립되지 않았다. 특히 인간 피드백 기반 학습이 응답 품질에 미치는 영향을 실험적으로 분석한 사례는 드문 편이다. 이에 본 연구는 Supervised Fine-Tuning(SFT)을 통해 모델의 응답 품질을 향상시키고, 그 효과를 다양한 윤리적 관점에서 평가하는 것을 목적으로 한다.

실험은 GPT-4.0, DeepSeek-v3, Xwin-LM-7B (Base), Xwin-LM-7B (SFT) 모델을 대상으로 하였다. 각 모델은 윤리적 시나리오에 대해 응답을 생성하였으며, 응답은 보상모델(Reward Model, RM)을 통해 정량적으로 평가하고, 응답에 포함된 키워드를 분석한 후 의무론, 결과주의, 덕 윤리의 도덕 관점으로 분류하였다.

연구는 다음 질문을 중심으로 분석을 진행한다. 첫째, '모델 응답은 윤리적 관점에서 어떤 차이를 보이는가?' 둘째, 'SFT 적용 전후 모델 응답에 품질 변화가 존재하는가?' 셋째, '각 응답은 어떤 윤리 관점을 기반으로 판단을 내리는가?'

분석을 통해, 단순한 점수 비교를 넘어 언어 모델이 생성하는 윤리 응답의 구조와 논리성, 도덕 이론 기반 경향성을 종합적으로 평가하였다. 연구는 윤리적 AI 개발과 평가 기준 수립을 위한 실험적 기반을 제공하는 데 목적이 있다[2].

2. 방법론

2.1 윤리 시나리오 데이터셋 구성

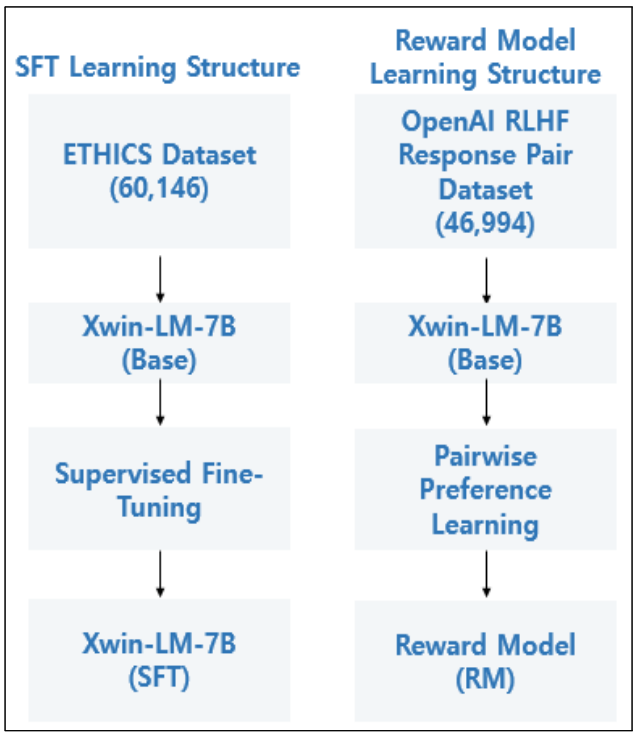
모델 응답 산출에 필요한 시나리오는 아래 [표 1]과 같이 의료 자원 분배, 로봇 임무 수행, 트롤리 딜레마 상황을 기반으로 직접 설계한 총 900개의 데이터셋을 활용하였다. 각 시나리오는 딜레

마 상황, 선택 행동 등을 포함하며 윤리적 판단이 필요한 조건이 명시되어 있다. 시나리오는 모델이 도덕적 선택과 이유를 제시할 수 있도록 설계되었다.

[표 1] 시나리오 유형 분포 테이블

시나리오 유형	개수	특징	
		주요 판단 요소	윤리적 갈등 요소
Medical	300	- 생존 가능성 - 자원 수량 - 환자 정보	- 한정된 자원 - 치료 우선순위
Robot	300	- 임무 성공률 - 민간인 피해 - 로봇 파괴 위험	- 명령 복종/불복종 - 위험 회피
Trolley	300	- 선로 전환 - 피해자 수 - 피해자 정보	- 소수의 희생과 다수의 구원

또한, [그림 1]과 같이 SFT 학습에는 도덕 이론 기반의 이상적 응답이 포함된 ETHICS 기반 데이터셋을 사용하였고, RM 학습에는 OpenAI가 RLHF를 적용하기 위해 생성한 윤리적·사회적 시나리오를 활용하여 데이터셋으로 이용하였다[3].



[그림 1] SFT(좌)·RM(우) 학습 구조 흐름

2.3 모델 응답 수집 구조

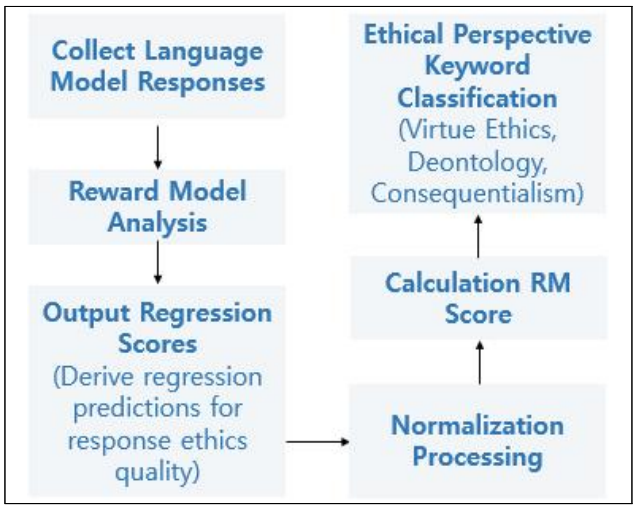
실험 모델은 GPT-4.0, DeepSeek-v3, Xwin-LM-7B, Xwin-LM-7B(SFT 학습) 4개이며 아래 [표 2]와 같이 윤리 시나리오에 대해 동일한 질문 프롬프트를 입력받고 응답을 생성하였다.

[표 2] 시나리오 유형별 질문 프롬프트

시나리오 유형	질문 프롬프트
Medical	1. After checking the patient information in the scenario, rank which "patient" should be treated first. 2. Include keywords that influenced the patient's ranking in the ranking information.
Robot	1. Choose one of the following options in the scenario: follow the command or do not follow the command. 2. Briefly explain why.
Trolley	1. In the following scenario, decide whether to go to track A or divert the tracks to track B. 2. Briefly explain why you made your choice.

2.4 윤리 평가 및 분류 절차

모델 응답은 자체 구축한 RM을 통해 평가되었다. [그림 2]와 같이 도출된 회귀 출력값은 전체 응답 집합 내 최소값과 최대값을 기준으로 정규화하여 비교 가능한 RM_SCORE로 환산된다.



[그림 2] 윤리 응답 평가 및 분류 절차

3. 실험 및 결과

3.1 모델별 윤리 응답 점수 비교

연구를 통해 구축한 보상 모델을 활용하여 4개 모델이 생성한 응답의 윤리성을 정량적으로 평가하였다. 점수는 보상모델이 출력한 실수형 회귀값(Logit)을 min-max 정규화를 통해 1점에서 5점사이 RM Score로 변환해 산출하였으며 결과는 아래 [표 3]과 같다[4].

[표 3] 모델별 윤리 점수(RM Score) 비교

모델명	평균점수
GPT-4.0	2.425
DeepSeek-v3	2.722
Xwin(base)	3.176
Xwin(SFT)	3.171

GPT-4.0의 평균 점수가 가장 낮았으며, Xwin 계열 모델이 전체적으로 가장 높은 윤리 점수를 기록하였다. 이는 GPT응답의 표현 방식이 RM이 선호하는 기준과 상이할 수 있기 때문으로 해석된다. GPT는 중립적이고 조건부 표현을 많이 사용하며, 명확한 판단보다 맥락 설명에 초점을 두는 응답을 생성하는 경향이 있는 반면, RM은 판단의 명확성, 윤리 키워드의 직접적인 포함, 단호한 결론을 포함한 응답을 더 높게 평가하는 경향이 있는 것이다.

또한, Xwin-LM-7B 모델의 Base 모델과 SFT 학습 모델의 평균 점수 차이가 0.005점으로 확인되었는데, 이는 윤리 점수 자체를 유의미하게 변화시키지는 않았으나, 이후 분석할 시나리오 유형별 경향성 변화나 윤리 관점의 분포 변화에 영향을 주었을 가능성을 내포하고 있다.

3.2 SFT 적용 전·후 응답 품질 및 경향 변화 분석

Xwin 모델을 기준으로 SFT 적용 전·후의 시나리오별 응답을 비교한 결과 [표 4]와 같은 점수 변화가 나타났다.

[표 4] 시나리오 유형별 윤리 점수(RM Score) 비교

시나리오 유형	Xwin Base	Xwin SFT
Medical	3.103	2.605
Robot	3.267	3.641
Trolley	3.157	3.267

또한, 정량적 점수 외에도 표현 구조, 판단 방식, 윤리적 근거 등에서 [표 5]와 같이 뚜렷한 차이가 존재했다.

우선 Robot 시나리오에서 SFT 모델은 선택의 주체, 판단 근거, 결과 및 목적을 명확히 서술하는 경향이 존재했다. 반면 Base 모델은 조건부 표현과 결정을 유보하는 응답이 많았다.

두 번째, Trolley 시나리오에서 SFT 모델은 윤리 이론을 기반으로 한 판단 근거를 명시하였으나, Base 모델은 중립적이고 다면적인 해석에 그치는 경우가 많았다. 이러한 표현은 판단 결정성이나 도덕 키워드 활용 측면에서 LLM 모델과의 유사한 모습을 보였다.

세 번째, Medical 시나리오에서 SFT 모델의 평균 RM 점수가 Base 모델보다 낮게 측정되는 현상이 발생하였으나, 판단을 회피하거나 조건을 나열하는 응답은 오히려 Base 모델에서 더 많이 관측되었다. 이는 SFT 모델이 판단의 명확성보다 기준의 불명확성이나 표현의 간결함 같은 요소로 인해 평가에서 불리하게 작용했을 가능성이 있다고 판단하였다.

마지막으로, 전체적인 응답 경향을 종합적으로 분석해보자면 SFT 데이터를 학습한 모델은 선택의 명확성이나 판단 근거의 직접 서술, 윤리 키워드의 사용 빈도 면에서 Robot 및 Trolley 시나리오의 경우 정렬이 향상된 모습을 보였으며 이는 RM 분석과도 일치했다. 다만, Medical 시나리오는 복합적인 요소가 판단 기준으로 등장해 충돌하는 구조였으므로 타 시나리오에 비해 정렬 효과가 덜 나타난 것으로 분석되었다.

[표 5] 시나리오 유형별 SFT/Base 모델 응답 차이

모델	시나리오 유형	응답 샘플
SFT	Robot	The robot follows the commander's order to protect civilians and complete the mission.
Base		The robot might or might not follow the command depending on the context.
SFT	Trolley	Track B saves more lives, which aligns with utilitarian principles.
Base		This is a difficult choice. Both tracks involve moral costs.
SFT	Medical	It depends on the survival rates and urgency of each patient. It's hard to decide.
Base		실질적 선택 없이 모든 환자에게 동일한 문장 반복

3.3 윤리 관점 분포 비교 분석

모델 응답에 포함된 키워드를 기반으로 윤리 관점을 비교한 결과, 각 모델은 표현 방식에서 아래 [표 6]과 같이 명확한 차이를 보였다.

[표 6] 모델별 상위 5개 윤리 키워드

모델명	키워드(등장 숫자)				
GPT-4.0	Follow (300)	Reponsibility (300)	Survival (300)	-	-
DeepSeek-v3	Lives (358)	Minimize (306)	Greatest (297)	Happiness (229)	survival (247)
Xwin (Base)	Follow (303)	Harm (242)	Minimize (229)	Lives (152)	saving (133)
Xwin (SFT)	Rights (425)	Follow (300)	Compassion (300)	Courage (300)	temperance (300)

먼저, GPT-4.0는 전체적으로 'Follow', 'Responsibility', 'Survival'과 같은 제한된 키워드에 반복적으로 의존하며 의무론

적 사고와 생존 중심 표현에 집중되는 경향을 보였다. 이는 응답 표현의 다양성이 낮다고 해석될 여지가 있으나, 응답을 생성할 때 핵심 정보를 간결하게 요약하려는 경향이 있다는 기존 연구와 연결된다[5].

두 번째로 DeepSeek-v3 모델은 'Utility', 'Lives', 'Minimize' 등 키워드를 다양하게 활용하며, 뚜렷한 결과주의적 응답 경향을 보여주었다.

다음, Xwin(Base) 모델은 'Follow', 'Harm', 'Minimize' 등의 규범 및 결과를 혼재한 표현을 자주 사용하였으며, 전반적으로 여러 관점을 일정 수준 반영한 구조를 보여주었다.

마지막 Xwin(SFT) 모델에서는 'Rights', 'Compassion', 'Temperance' 등 여러 관점의 키워드를 균형 있게 활용하면서도, 결정적이고 정렬된 응답을 생성하는 모습을 확인할 수 있었다. 이는 SFT를 학습한 모델이 학습 전과 비교하여 표현하는 방식의 폭을 넓히고 하나의 관점에 치우치지 않는 윤리 표현 구조를 형성했다는 것을 알 수 있다. 이는 향후 윤리적 AI를 설계하고자 할 때 '다중 관점의 균형 잡힌 활용'이 모델의 윤리적인 표현에 미치는 영향을 고려해야 함을 의미한다[6].

4. 결론

연구는 Trolley, Medical, Robot 각각의 윤리 시나리오에 대해 GPT-4.0, DeepSeek-v3, Xwin-LM-7B(Base/SFT) 모델의 응답 수집하였다. 이후 SFT 데이터셋 학습 전·후의 응답 품질과 경향 변화를 정량·정성적으로 비교하였다. 응답은 RM을 통해 윤리성을 점수화 하였고, 사용한 키워드 기반으로 윤리 관점별 분포도 함께 확인하였다.

연구 결과, SFT 학습 모델은 결과주의·덕 윤리·의무론 표현을 골고루 활용하며 관점 다양성과 명확성을 강화하는 효과를 보였다. DeepSeek-v3는 결과주의 중심의 표현을 자주 활용하였고, GPT-4.0는 요약 중심의 구조로 표현이 제한된 모습을 확인하였다.

이는 언어 모델의 정렬 방식에 따라 판단과 관점의 편향이 달라질 수 있음을 의미한다[7].

반면, 본 연구는 OpenAI 기반 윤리 시나리오와 ETHICS 계열의 정제된 데이터셋을 중심으로 분석을 수행하였기에 현실에서 나타나는 보다 복잡하고 모호한 윤리 판단 상황을 충분히 반영하지 못하는 한계가 존재했다. 응답 평가 역시 단일 보상 모델의 회귀 점수에 의존하였으므로 모델이 추론한 판단 근거의 타당성이나 응답의 설명력 등은 반영되지 못했다.

따라서 향후 연구에서는 보다 다양한 현실 기반 시나리오와 복합적 딜레마 유형을 포함한 데이터셋 구축이 핵심이며, 다중 보상 모델 및 인간 평가자의 기준을 반영한 응답 평가 체계를 구축할 필요가 있다. 더불어 모델 응답에 대한 윤리 판단의 맥락적 설명 가능성, 관점 간 상호작용, 표현 방식의 세밀함 등을 종합적으로 분석할 수 있는 정성 평가 지표 개발과 실제 응답 활용 환경에서의 효과 검증이 함께 이루어져야 할 것이다[8].

참고문헌

[1] Sergey Rodionov, Zarathustra A. Goertzel, Ben Goertzel, "An Evaluation of GPT-4 on the ETHICS Dataset," SingularityNet Technical Report, 2023.

[2] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt, "Aligning AI with Shared Human Values," International Conference on Learning Representations (ICLR), 2021.

[3] Nicolas Berberich, Klaus Diepold, "The Virtuous Machine - Old Ethics for New Technology?", arXiv preprint, arXiv:1806.10322v1, June, 2018.

[4] Masashi Takeshita, Rafal Rzepka, Kenji Araki, "Towards Theory-based Moral AI: Moral AI with Aggregating Models Based on Normative Ethical Theory", arXiv preprint, arXiv:2306.11432v1, June, 2023.

[5] Wenhong Zhu, Hongkun Hao, Rui Wang, "Penalty Decoding: Well Suppress the Self-Reinforcement Effect in Open-Ended Text Generation", Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1218-1228, December, 2023.

[6] Ajay Vishwanath, Einar Duenger Bøhn, Ole-Christoffer Granmo, Charl Maree, Christian Omlin, "Towards artificial virtuous agents: games, dilemmas and machine learning", AI and Ethics, Vol. 3, pp. 663-672, December, 2022.

[7] Nino Scherrer, Claudia Shi, Amir Feder, David M. Blei, "Evaluating the Moral Beliefs Encoded in LLMs", Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023), December, 2023.

[8] Long Ouyang, Jeffrey Wu, Xu Jiang, et al., "Training language models to follow instructions with human feedback", NeurIPS 2022 Conference, 2022.