한국어 동양 윤리 원칙을 활용한 LLM의 도덕적 판단 연구

유준호[1], 신유현[2]
[1]인천대학교 컴퓨터공학부 학부생
[2]인천대학교 컴퓨터공학과 교수
yoojuneho0723@inu.ac.kr, yhshin@inu.ac.kr

Moral Judgment of LLM Using Korean Datasets and Eastern Ethical Principles

Junho-Yoo[1], Youhyun Shin[2]
[1]Dept. of Computer Science, Incheon National University
[2]Dept. of Computer Science, Incheon National University

요 약

대규모 언어 모델(LLM)의 발전으로 인간의 사고를 대체할 가능성이 커지며, 도덕적 판단에 대한 관심도 증가하고 있다. 그러나 영어 중심으로 학습된 LLM의 특성상, 한국어의 적용과 동양 윤리 원칙에 대한 연구는 미비한 상황이다. 본 연구에서는 LLM의 도덕적 편향을 분석하고, 저자원·고자원 언어 환경 및 동·서양 윤리 정책 적용 시 편향 변화 여부를 실험적으로 검증하였다. 이를 통해 LLM의 도덕적 판단이 언어·문화적 요인에 따라 어떻게 영향을 받는지를 규명하고자 한다. 실험 결과, LLM의 도덕적 판단이 언어적·문화적 맥락에 따라 달라질 수 있음을 확인하였으며, 이러한 연구 결과는 AI 윤리 가이드라인 수립과 콘텐츠 검열 정책 개선에 기여할 수 있을 것으로 기대된다.

1. 서론

LLM의 등장은 인간의 사고를 보조하거나 대체할 수준으로 발전하며, 판단을 내릴 가능성도 제기되고 있다. 이에 따라 LLM의 도덕적 편향성을 이해하는 것이 중요한 연구 과제가 되었다. 본 연구에서는 주로 영어로 학습된 LLM이 저자원 언어인 한국어 데이터셋에서 보이는 도덕적 편향성을 실험적으로 분석한다. 또한, 동양 윤리 정책 적용 시 모델의 응답이 서양 윤리 정책과 어떻게 다른지를 비교하여 도덕적 판단의 문화적·언어적 영향을 탐구한다.

2. 관련 연구

Fridman et al.(2024)는 LLM의 도덕적 판단이 인간과 다르며, 모델 크기가 커질수록 인간과 유사성이 증가하는 경향을 실험적으로 입증했다. 연구진은 공리주의적 도덕 딜레마를 통해 다양한 규모의 LLM을 분석한 결과, 작은 모델일수록 인간과 차이가 크고, 대규모 모델이 정교한 도덕적 추론을 수행함을 확인했다.

Agarwal et al.(2024)는 LLM의 학습 언어와 자원 수준에 따라 도덕적 판단과 편향성이 달라짐을 분석했다. 영어에서는 윤리 정책과의 일관성이 유지되었으나, 저자원 언어에서는 편향성이 커지고 일관성이 낮아지는 경향이 나타났다. 또한, 같은 모델이라도 입력 언어에 따라 도덕적 판단이 달라져, LLM이 단순 규칙이 아닌 언어·문화적 맥락의 영향을 받음을 시사한다.

3. 본론

Agarwal et al.(2024) 논문에서는 주로 영어로 학습된 LLM이 다자원 언어 환경에서 4가지 윤리적 딜레마를 서양 윤리(덕 윤리, 의무론, 결과주의) 정책과 함께 제공받았을 때, 저자원 언어 환경보다 편향성이 적고 정확도가 높은 경향을 보였다. 본 연구에서는 이러한 특성을 기반으로, 다자원(영어) 및 저자원(한국어) 환경에서 LLM이 서양 윤리와 동양 윤리(유교 윤리, 도교 윤리, 불교 윤리) 정책을 적용받았을 때 편향성의 차이를 실험적으로 분석하였다. 이를 통해 LLM의 도덕적 판단이 언어적·문화적 맥락에 따라 어떻게 달라지는지 관찰하고, 편향성을 규명하는 것이 본 연구의 주요 목표이다.

4. 실험

실험은 OpenAI의 API 키를 활용하여 진행되었으며,

모델로는 GPT-3.5-Turbo-0125와 GPT-4-0613을 사용하였다. 모든 실험에서 동일한 하이퍼파라미터를 적용하였으며, Temperature는 0, Top Probability는 0.95, Presence Penalty는 1로 설정하였다.

4-1. 데이터 셋

본 연구에서는 Abhinav Rao et al.(2023)의 논문에서 제공한 딜레마와 프롬프트를 활용하여 실험을 진행하였으며, Agarwal et al.(2024)의 연구와 동일하게 구글 번역기를 이용하여 한국어 데이터셋을 생성하여 사용하였다.

4-2. 결과 및 분석

4-2-1. LLM의 기본 실험

<표 1>은 모델에 정책 없이 도덕적 딜레마만 제공한 후, 자체적인 윤리적 판단을 분석한 결과를 나타낸다. 모델은 "해야 한다", "하면 안 된다", "결정할 수 없다"의 세 가지 응답 중 하나를 선택해야 한다. 기본 실험에서는 4개의 딜레마에 대해 6가지 옵션 순열을 생성하고, 각 모델당 5회씩 반복하여 수행하였다.

<표 1> LLM의 기본 성능 비교

	Heinz	Monica	Rajesh	Timmy
	GPT-3.5-Turbo-0125			
English	100%	100%	100%	90%
Korean	60%	100%	100%	100%
	GPT-4-0613			
English	100%	100%	60%	100%
Korean	100%	100%	100%	83.3%

<표 1>에서 파란색은 “해야 한다”, 빨간색은 “하면 안 된다”, 그리고 노란색은 “결정할 수 없다“이다. 본 실험 결과는 LLM의 도덕적 판단이 입력 언어에 따라 달라질 수 있음을 시사한다.

4-2-2. 윤리 정책을 통한 편향성 실험

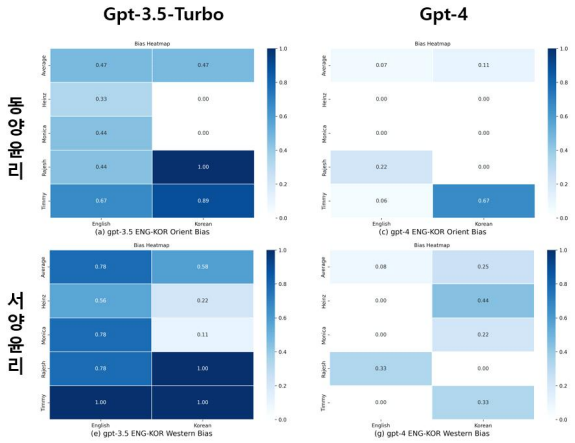
<표 2>는 동양 및 서양 윤리 정책이 제공될 때 모델의 편향성을 나타내는 히트맵이다. 편향성은 <표 1>에서 확인된 기본 선호도와 반대되는 윤리 정책을 적용했을 때, 모델이 기존 선호도를 얼마나 유지하는지를 기준으로 측정하였다. GPT-3.5-Turbo의 경우, 영어보다 한국어에서 편향성이 같거나 더 낮은 경향을 보인다. 특히, 한국어에서는 특정 딜레마에서 편향성이 극단적으로 높거나 낮은 패턴이 확인된다. 반면, GPT-4는 전반적으로 편차가 줄어들었으며, 다자원 언어인 영어에서 편향성이 더 적게 나타난다. 이러한 결과는 기존 연구에서 보고된 실험 결과와 일치한다.

5. 결론

본 연구에서는 LLM이 내재적으로 지닌 도덕적 가치와 편향성을 분석하였다. 실험 결과, 특정 윤리 정책이 적용될 때, 특정 언어 환경에서는 편향성이 감소하는 경향이 확인되었다. 이는 모델이 특정 언어로 학습하는 과정에서 해당 언어의 사회적·도덕적 규범이 내재되었기 때문으로 해석할 수 있다. 또한, 모델의 크기가 커질수록 이러한 편향성이 점진적으로 감소하는 경향도 관찰되었다.

이번 연구는 LLM의 도덕적 판단이 언어적·문화적 맥락에 따라 달라질 수 있으며, 모델의 규모가 윤리적 편향성에 영향을 미칠 수 있음을 시사한다. 향후 연구에서는 다양한 언어와 윤리 정책을 확장하여 LLM의 도덕적 판단 메커니즘을 더욱 심층적으로 분석할 필요가 있다.

<표 2> 윤리적 정책에 따른 LLM의 편향성 비교



사사문구

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-학·석사연계ICT핵심인재양성 지원을 받아 수행된 연구임(IITP-2025-RS-2023-00260175)

참고문헌

[1] Agarwal, U. et al. (2024). Ethical Reasoning and Moral Value Alignment of LLMs Depend on the Language We Prompt Them in. LREC-COLING 2024, 6330 - 6340.

[2] Rao, A. et al. (2023). Ethical Reasoning over Moral Alignment: A Case and Framework for In-Context Ethical Policies in LLMs. Findings of ACL: EMNLP 2023, 13370 - 13388.

[3] Marraffini, G. F. G. et al. (2024). The Greatest Good Benchmark: Measuring LLMs' Alignment with Utilitarian Moral Dilemmas. EMNLP 2024, 21950 - 21959.