공학석사 학위논문

LLM(Large Language Model)모델 윤리성 비교 평가 : 성별 편향성 연구

Ethical Comparison of Large Language Models :
A Study on Gender

2024년 8월

서울과학기술대학교 산업대학원

빅데이터AI경영정보학과

김 민 지

LLM(Large Language Model)모델 윤리성 비교 평가 : 성별 편향성 연구

Ethical Comparison of Large Language Models :
A Study on Gender

지도교수 한지형

이 논문을 공학석사 학위논문으로 제출함
2024년 8월

서울과학기술대학교 산업대학원
빅데이터AI경영정보학과

김 민 지

김민지의 공학석사 학위논문을 인준함
2024년 8월

심사위원장 이길홍 (인)

심사위원 천세학 (인)

심사위원 한지형 (인)

요 약

제 목 : LLM(Large Language Model) 모델 윤리성 비교 평가 : 성별 편향성 연구

인공지능(AI) 기술의 급속한 발전은 다양한 딥러닝 모델의 발전으로 이어졌고 그 중에서도 최근 몇 년간 자연어 처리(NLP) 분야의 활발한 연구가 이루어졌다. 2022년 OpenAI의 Chat GPT를 선두로 시작하여 NLP의 딥러닝 모델 중 하나인 Transformer를 기반으로 하는 LLM 모델에 대한 연구와 개발이 폭발적으로 이루어지고 있다. 자동 번역, 감정분석, 챗봇 등에서 널리 사용되는 이들 모델은 텍스트 데이터를 처리하고 이해하는 데 탁월한 능력을 보인다.

그러나 이러한 모델들이 사용하는 데이터가 편향되어 있을 경우, 모델 자체도 편향된 결과를 산출할 가능성이 있다. 특히 성별과 같은 민감한 특성에서 편향성 문제는 심각한 사회적 불평등을 초래할 수 있다. 이는 심각한 사회적 문제를 야기할 수 있으므로 전 세계적으로 AI 모델의 편향성을 정량적으로 측정하고 개선하고자 많은 노력이 진행 중에 있다.

이에 본 연구의 목적은 다양한 오픈소스 LLM 모델을 사용하여 성별 편향성을 정량적으로 검정하고 비교하는 것이다. 연구를 위한 데이터는 Bias in Bios 데이터 세트를 사용 하였다. 이 데이터 세트는 직업 설명과 해당 직업을 가진 사람들의 성별 정보를 포함하고있다. 주로 사회적으로 높은 소득에 포함되는 (Ex, 교수) 직업은 남성으로 라벨링이 되어있는 편향 데이터에 해당한다. 성별을 식별할 수 있는 단어를 중립어로 대체하여 각 LLaMA2, LLaMA3, Gemini, BERT 모델을 사용하여 직업을 기술하는 각 문장의 성별을 예측하도록 하고 모델의 예측 결과와 실제 성별을 비교하여 예측 일치 여부를 평가하였다. 그 후 IBM의 Fairness360 도구를 사용하여 Statistical Parity Difference, Disparate Impact, Equal Opportunity Difference 등의 편향성 지표를 계산하였다.

연구 결과, 모드 통계적 편향 차이 지표에서 편향이 존재하지만 특히 LLaMA3 모델이 Statistical Parity Difference 지표가 -0.99로 통계적 편향 차이가 가장 크게 나타났다. Disparate Impact 지표는 상대적으로 1에 가까운 값을 보였으나, BERT 모델이 가장 높은 비율을 보였다. Equal Opportunity Difference 지표에서는 LLaMA3 모델이 가장 낮은 값을 보여, 특권 그룹과 비특권 그룹 간의 참 양성율 차이가 가장 작았다.

결론적으로, 각 모델은 성별 편향성을 어느 정도 가지고 있으며, 이는 모델의 공정성을 개선하기 위한 인식 제고 및 AI 윤리에 대한 논의의 장을 만들어 가기 위한 중요한 기초 자료가 될 수 있다. 향후 연구에서는 다양한 데이터 세트와 더 많은 모델을 포함하여 편향성 검정을 수행하고, 편향성을 줄이기 위한 다양한 접근법을 적용하고자 한다.

목 차

표 목 차

그 림 목 차

Ⅰ. 서 론

1. 연구의 배경

최근 몇 년 간 AI (Artificial Intelligence) 기술, 특히 2022년 공개된 Open AI의 Chat GPT를 필두로 생성형 AI 기술이 급속도로 발전하고 있다. 이러한 기술들은 인간의 의사결정 과정을 지원하면서 일상생활의 다양한 측면을 혁신하고 있지만, 그와 동시에 새로운 사회적, 윤리적 도전 과제를 야기하고 있으며 이를 대응하기 위한 체계적인 연구와 대책이 시급하게 요구되고 있는 실정이다.

특히 생성형 AI는 인간이 생성한 대규모의 문서 데이터를 학습하고 이에 기반한 LLM (Large Language Model) 알고리즘으로 작동하기 때문에 학습 데이터의 성격에 따라 다양한 편향을 내포 할 수 있으며 이러한 편향은 성별, 인종, 지역 등에 따라 차별적 반응으로 나타나는 경우가 있다. 이러한 문제는 기술의 편향성뿐만 아니라, 그로 인해 야기될 수 있는 사회적 불평등 확대로 이어질 수 있으며, 특정 집단에 대한 차별을 심화시킬 우려가 있기 때문에 심각한 사회 문제로 이어질 수 있다.

예를 들어, 인공지능 모델이 특정 인종과 범죄 관련 용어를 더 자주 연관 지어 언급하는 사례가 발견되었고, 이는 해당 인종에 대한 부정적인 사회적 인식을 강화할 우려를 내포하고 있다. 또한, 성별에 따른 역할 기대치가 모델에 의해 강화되어, 여성과 남성에 대한 전통적인 직업적 역할 분담이 강조되는 경우도 있다. 이러한 사례들은 LLM의 편향이 어떻게 실제 사회 문제로 연결될 수 있는지를 보여준다.

이러한 맥락에서, 본 연구는 최근 주목받고 있는 오픈소스 기반의 LLM 모델 BERT(Google), LLaMA2(Meta), LLaMA3(Meta), Gemini(Google)를 중심으로 이들 모델의 성별 편향성 탐색하고 'Bias in Bios' 데이터 [1] 를 활용하여 실제 모델의 성별 편향성을 측정한 뒤, IBM에서 제공하는 Fairness 360 [2] 툴을 기반으로 정량적으로 비교 평가 하고자 한다.

2. 연구의 목적

본 연구는 인공지능, 특히 생성형 AI 기술의 발전이 사회에 미치는 영향을 다각적으로 이해하고, 더욱 공정하고 포괄적인 기술 개발을 위한 기초 자료를 제공하고자 하는 데 그 목적이 있다. LLM의 편향성 문제를 규명하고 이에 대응하는 것은 단지 기술적인 진보를 넘어 사회적 정의를 구현하는 데 있어 매우 중요하며 많은 사회적 논의가 필요한 문제이기 때문이다.

특히 최근 주목받고 있는 오픈소스 LLM 모델(BERT, Llama2, Llama3, Gemini)들이 성별에 관한 편향성을 어떻게 반영하고 있으며 이러한 반영이 모델의 응답에서 어떤 형태의 편향성을 불러 일으키는지를 체계적으로 분석하는 것을 목표로 한다.

연구의 결과를 정량적으로 평가하기 위해서 IBM Research에서 개발한 AI Fairness 360(이하 AIF360) 도구를 사용한다. 본 도구는 AI 모델의 불공정성을 감지하고 완화하는 것을 목표로 다양한 공정성 관련 지표와 알고리즘을 제공하여 신뢰할 수 있는 인공지능 모델을 개발, 평가할 수 있도록 지원한다.

이를 통해 본 연구는 다음과 같은 세부적인 연구 목적을 달성하고자 한다.

1) LLM 편향성 연구의 필요성 및 중요성 인식 제고 : LLM 편향성의 개념과 중요성을 명확하게 정의하고 인공지능의 올바른 발전에 있어 본 연구의 필요성을 제고

2) 글로벌 오픈소스 LLM 모델 비교 분석 : 글로벌 대기업에서 개발하고 오픈소스로 제공하는 LLM 모델의 구조, 학습 알고리즘, 편향성에 대해 비교하고 객관적으로 측정할 수 있는 정량적 방법론을 사용하여 각 모델의 편향성을 실증적으로 분석

3) LLM 모델 편향 완화 방안 모색 : LLM 모델들의 편향성을 최소화 할 수 있는 모델 설계 및 학습 전략을 제안하고 AI윤리의 중요성에 대한 시사점을 제공한다.

마지막으로 AI 개발자와 사용자가 인공지능을 보다 책임감 있게 활용할 수 있는 방안을 제시하고자 하며 더 나아가, 이 연구는 인공지능 기술이 사회적으로 긍정적인 영향을 미치도록 유도하고, 기술 발전이 가져올 수 있는 부정적인 측면을 최소화하는 데 기여하고자 한다.

Ⅱ. 이론적 배경

1. LLM 모델의 편향성

1) 편향성의 정의

편향성이란, 데이터나 알고리즘에서 특정 그룹이나 결과에 대해 불공정하게 기울어진 경향성을 말한다. 대규모 언어 모델(Large Language Models, LLM)의 편향성은 주로 훈련 데이터에 내재된 사회적, 문화적 편견이 모델의 결정과 출력에 반영되는 현상을 지칭하며 이러한 편향성은 모델이 생성하는 언어와 의사결정 과정에서 특정 인구집단에 대해 차별적인 태도를 보이거나 특정 주제에 대한 불균형한 관점을 제시하는 것으로 나타날 수 있다. [3]

LLM의 편향 유형은 일반적으로 다음과 같이 분류할 수 있다.

Table 2.1 편향의 유형

유형	설명
성별	성별 편향은 모델이 특정 성별에 대한 고정관념이나 선입견을 반영하는 현상이다. 예를 들어, '간호사'나 '비서'와 같은 직업명이 여성과 더 강하게 연결되거나, '공학자'나 '경영자'와 같은 직업이 남성과 더 자주 연결되는 경우를 들 수 있다.
인종 민족	인종적 또는 민족적 편향은 특정 인종이나 민족에 대해 부정적인 스테레오타입이나 차별적인 언어를 사용하는 것을 포함한다. 이러한 편향은 모델이 특정 인종적 배경을 가진 인물에 대해 부정적인 성향을 나타내거나, 특정 문화적 요소를 과소평가하는 방식으로 나타날 수 있다.
연령	연령 편향은 모델이 특정 연령대에 대해 긍정적이거나 부정적인 편향을 보이는 현상이다. 예를 들어, '노인'에 대해 과도하게 보호적이거나 의존적인 표현을 사용하는 것이 이에 해당한다.
지역	지역적 편향은 특정 지역이나 국가에 대한 편견을 반영하는 것이다. 이는 모델이 특정 지역의 언어나 방언, 문화적 특성을 무시하거나 오해하는 형태로 나타날 수 있다.

사회 경제	이 편향은 특정 사회경제적 배경을 가진 그룹에 대한 스테레오타입을 반영한다. 예를 들어, 부유한 계층이나 특정 직업군에 대해 긍정적인 이미지를 과도하게 강조하는 경우를 들 수 있다.
이념 정치	모델이 특정 정치적 이념이나 사상에 치우친 언어를 사용하는 경우이다. 이는 특정 정치적 인물이나 사건에 대해 긍정적이거나 부정적인 감정을 더욱 강하게 표현하는 방식으로 나타날 수 있다.

2) 편향성의 측정방법

편향성을 측정하는 정량적 방법론은 주로 두 가지 접근법, 직접적 측정과 간접적 측정을 포함한다. 각 접근법은 모델의 편향성을 효과적으로 평가하기 위해 고안된 다양한 기법을 활용한다. [4]

(1) 직접적 측정

직접적 측정 방법은 모델의 출력에서 특정 키워드나 표현을 분석하여 편향성을 평가하는 접근법이다. 다음 표는 직접적 측정 방법의 주요 절차를 요약한 것이다.

Table 2.2 편향성 직접 측정 방법

절차	설명
키워드 분석	모델의 응답에서 특정 키워드나 문구의 빈도와 맥락을 분석한다. 성별이나 인종에 대한 언급 빈도를 조사하고, 긍정적 또는 부정적 의미로 사용된 빈도를 비교한다.
감정 분석	텍스트 분석 도구를 사용하여 모델의 응답에서 감정(positive, negative, neutral)을 평가한다. 특정 그룹이나 주제에 대해 일관되게 부정적인 감정을 표현하는지 확인한다.
편향 벤치마크	WinoBias와 같은 벤치마크 데이터를 사용하여 모델의 응답을 평가한다. WinoBias는 성별 고정관념을 드러낼 수 있는 문장들

사용	로 구성되어 있으며, 모델이 이러한 문장에서 편향된 응답을 보이는지를 평가한다.

(2) 간접적 측정

간접적 측정 방법은 모델의 의사결정 과정을 통해 편향성을 파악하는 방법으로, 더 복잡한 실험 설계를 필요로 한다. 다음 표는 간접적 측정 방법의 주요 절차를 요약한 것이다.

Table 2.3 편향성 간접 측정 방법

절차	설명
시나리오 기반 실험	다양한 시나리오를 구성하여 모델의 반응을 관찰한다. 동일한 질문을 성별, 인종, 나이 등 다양한 인구통계적 변수에 따라 변형하여 제시하고, 모델의 응답 차이를 분석한다.
의사결정 과정 분석	모델의 의사결정 과정을 추적하여 특정 변수들이 모델의 출력에 미치는 영향을 분석한다. 이를 통해 모델이 어떤 요인에 의해 편향된 결정을 내리는지 파악할 수 있다.
간접적 감지 기법	모델의 출력뿐만 아니라 내부 메커니즘을 분석하여 편향성을 평가한다. 예를 들어, 특정 단어의 출현 확률 변화나 내부 레이어의 활성화 패턴을 분석하여 편향성을 간접적으로 감지한다.
인과 추적	모델의 내부 상태를 조작하고 이로 인해 출력이 어떻게 변화하는지 관찰하여, 특정 상태가 편향성에 미치는 영향을 분석한다. 이를 통해 모델의 편향성을 유발하는 내부 메커니즘을 파악할 수 있다.

이와 같은 직접적 및 간접적 측정 방법은 AI 모델의 편향성을 체계적으로 분석하고, 이를 완화하기 위한 구체적인 전략을 수립하는 데 중요한 역할을 한다.

3) 편향성 측정을 위한 글로벌 연구 동향

(1) 오픈소스 툴

오늘날 데이터 집합과 모델 내의 편향 및 공정성 파악에 사용할 수 있는 오픈소스 툴은 다양하다. 예를 들어, 구글의 WIT(What-If Tool), Fairness Gym, IBM의 AI Fairness 360, Aequitas, FairLearn 등이다. 데이터의 대표성이나 균형성에 대한 이해를 높이기 위한 데이터 시각화 및 상호작용에 사용할 수 있는 툴들도 있다. 구글의 Facets, IBM AI 360 Explainability가 대표적이다. 이들 툴 중에는 편견 완화 기능이 들어 있는 것도 있지만 대부분은 없으므로 별도의 툴을 구매해야 할 수도 있다. 아래 표는 이러한 툴들을 분류하고 각각의 상세한 설명을 제공한다.

Table 2.4 AI 공정성 파악을 위한 글로벌 오픈소스 툴

이름	제공 기업	주요 기능 및 설명
WIT (What-If Tool)	Google	모델의 예측 결과를 분석하고, 입력 데이터를 수정해가며 모델의 반응을 시각적으로 탐색할 수 있는 도구. 사용자는 모델의 결정 과정과 편향성을 쉽게 파악할 수 있음.
Fairness Gym	Google	시뮬레이션 환경을 제공하여 다양한 시나리오에서 AI 모델의 공정성을 테스트할 수 있는 도구. 사용자는 모델의 결정이 시간이 지남에 따라 어떻게 영향을 미치는지 분석할 수 있음.
AI Fairness 360	IBM	AI 모델의 편향성을 감지하고 완화하는 다양한 알고리즘과 지표를 제공하는 툴킷. 데이터 전처리, 모델 학습, 평가 단계에서 공정성을 확보할 수 있는 다양한 기능을 제공함.

Aequitas	Center for Data Science and Public Policy	공정성 분석을 위한 오픈소스 툴킷. 모델의 예측이 인구 집단에 따라 어떤 차이가 있는지 분석하고, 다양한 공정성 지표를 통해 편향성을 평가할 수 있음.
FairLearn	Microsoft	AI 모델의 공정성을 평가하고 개선하는 데 사용되는 도구. 공정성 지표를 분석하고, 모델 학습 과정에서 특정 집단에 대한 편향을 줄이는 방법을 제시함.
Facets	Google	데이터 시각화 도구로, 데이터 세트의 구조와 분포를 쉽게 이해할 수 있도록 함. 데이터의 대표성과 균형성을 높이기 위한 상호작용적인 시각화 기능을 제공함.
AI 360 Explainability	IBM	AI 모델의 결정 과정을 설명하고 이해하기 위한 도구. 모델의 예측 결과가 어떻게 도출되었는지 시각적으로 설명하여 사용자가 모델의 내부 작동 원리를 이해할 수 있도록 도움.

(2) 레드팀 (Red team) 구성

경쟁자 역할을 하는 레드팀 (Red team)을 구성하는 방법도 있다. 이는 보안 분야에서 따온 것인데 윤리적 사용 맥락에 쓰이는 경우를 가정해 AI 시스템 사용이 피해를 야기하는 방식으로 테스트한다. 이를 통해 윤리적(또한 잠재적으로 법적) 위험이 드러나면 해결 방법을 마련한다.

AI 시스템의 잠재적인 해나 의도치 않은 결과를 파악하기 위해 지역사회 배심원단을 활용하는 방법도 있다. 이를 통해 다양한 계층, 특히 소외된 지역사회의 대표인단을 소집해 특정 시스템이 그들에게 어떻게 영향을 미칠지에 대한 그들의 관점을 더 명확하게 파악할 수 있다.

4) AI Fairness 360의 편향성 평가 지표

AI Fairness 360은 IBM에서 개발한 오픈 소스 툴킷으로, 인공지능 시스템에서의 편향을 감지하고 완화하는 데 필요한 다양한 도구와 알고리즘을 제공한다. [2] 이 툴킷은 편향성을 평가하는 다수의 지표와 함께, 데이터 전처리, 모델 학습 과정 수정, 결과 후처리 등을 통해 편향을 줄이는 기술을 포함하고 있으며 사용자는 이 툴킷을 통해 AI 모델의 공정성을 향상시키기 위한 실용적인 솔루션을 탐색하고 적용할 수 있다.

AI Fairness 360의 평가지표는 대표적으로 다음과 같다.

각 지표는 특정 상황이나 요구에 따라 선택적으로 사용될 수 있으며, 하나의 지표만으로 모든 유형의 편향을 충분히 평가할 수 없기 때문에 종종 여러 지표를 함께 사용하여 보다 포괄적인 편향성 분석을 수행한다.

Table 2.5 Fairness 360의 편향성 평가 지표

지표	설명	해석 방법
Statistical Parity Difference	비특권 집단과 특권 집단 간에 발생하는 긍정적인 결과(예: 승진, 학교 입학, 대출 승인 등)를 받는 빈도가 얼마나 다른지 측정함으로써 두 집단이 동등하게 좋은 기회나 혜택을 받고 있는지 여부를 파악하는데 사용	이 지표가 0에 가까울수록, 두 집단이 동등한 비율로 긍정적 결과를 받는 것을 의미하며, 편향이 적다고 볼 수 있다. 양수 또는 음수 값은 특권 집단이 유리하거나 불리하게 작용하고 있음을 나타낸다.
Equal Opportunity Difference	긍정적 결과(예: 질병 없음)를 받을 경우의 진짜 긍정 비율(실제 긍정 케이스 중 정확하게 긍정으로 예측된 경우의 비율)이 비특권집단과 특권 집단 간에 얼마나 다른지를 나타낸다.	이 지표는 공정한 예측이 이루어지고 있는지를 평가하는데 중요하다. 값이 0에 가까울수록 두 집단이 동일한 기회를 받는다는 것을 의미한다.
Average Odds Difference	양성 예측 비율과 진짜 긍정 비율의 평균 차이를 비특권 집단과 특권 집단 사이에서	이 지표는 편향을 감지하기 위해 사용되며, 두 집단이 받는 결과의 공정성을 평가한

	측정한다.	다. 0에 가까울수록 더 공정한 결과를 나타내며, 양수 또는 음수는 한 집단에 대한 편향을 의미한다.
Disparate Impact	비특권 집단이 특권 집단에 비해 긍정적 결과를 얻을 확률의 비율이다.	이 비율이 1에 가까울수록 두 집단 간의 결과가 균등하다는 것을 의미한다. 1보다 크거나 작은 값은 특정 집단이 유리하거나 불리한 위치에 있음을 나타낸다.
Theil Index	개인에게 혜택이 배분되는 불평등을 측정한다.	이 지표는 0에 가까울수록 개인 간의 혜택 배분이 평등하다는 것을 의미하며, 값이 클수록 더 큰 불평등을 나타낸다.

2. 오픈소스 LLM 모델 비교

1) Transformer

Transformer 모델은 딥러닝 분야에서 자연어 처리(NLP) 작업을 수행하는데 널리 사용되는 모델이다. 이 모델은 Google의 연구진이 2017년 “Attention is All You Need“ 논문에서 처음 제안하였다. [5] Transformer는 self-attention 메커니즘을 도입하여 이전의 순환 신경망(RNN)과 장단기 기억 네트워크(LSTM) 모델보다 더 효율적으로 데이터를 처리할 수 있다.

Transformer 모델은 인코더와 디코더의 두 가지 주요 구성 요소로 구성된다. 각 구성 요소는 여러 층으로 이루어져 있으며, 각 층은 multi-head self-attention 메커니즘과 순전파 신경망으로 구성된다. Transformer의 주요 특징은 병렬 처리가 가능하다는 점으로, 이는 모델 학습과 추론 속도를 크게 향상시킨다.

Table 2.6 Transformer 모델의 주요 구성요소

구성요소	설명
인코더 (Encoder)	입력 시퀀스를 처리하여 context-aware 표현을 생성한다. 여러 층의 self-attention과 피드포워드 신경망으로 구성된다.
디코더 (Decoder)	인코더의 출력과 이전 디코더 출력을 결합하여 최종 출력 시퀀스를 생성한다. 인코더와 유사한 구조를 가진다.
Self-Attention	입력 시퀀스의 각 단어가 다른 모든 단어와의 관계를 학습하여 중요한 정보를 추출한다.
Multi-Head	여러 개의 self-attention 메커니즘을 병렬로 사용하여 다양한 표현을 학습한다.

Transformer 모델을 기반으로 하는 10개의 대표적인 대규모 언어모델은 다음과 같다.

Table 2.7 Transformer 기반 주요 LLM

모델명	개발사	주요 특징	응용 분야 및 장점
BERT	Google	양방향 문맥 이해, 사전 학습 및 미세 조정	문서 분류, 질의 응답, 감정 분석 등 다양한 NLP 작업에서 높은 성능
GPT-3	OpenAI	1750억 개의 파라미터, 자연어 생성 및 이해, zero-shot 학습 가능	대화형 AI, 자동화된 콘텐츠 생성, 언어 번역
T5	Google	Text-to-Text 프레임워크, 다양한 NLP 작업을 단일 모델로 처리	텍스트 요약, 번역, 문서 생성 등 다목적 NLP 작업 수행
RoBERTa	Facebook	BERT의 개선판, 더 긴 학습 시간과 대규모 데이터셋 사용	향상된 문서 분류, 질의 응답 성능
ALBERT	Google	경량화된 BERT, 파라미터 공유 및 팩터화된 임베딩	메모리 효율적, 빠른 학습과 추론 가능
XLNet	Google	양방향 학습, Transformer-XL 기반, BERT보다 나은 성능	문서 이해, 텍스트 생성, 감정 분석

LLaMA2	Meta	대화형 AI에 최적화, 높은 파라미터 수,	자연스러운 대화 생성, 고객 지원, 대화형 에이전트
LLaMA3	Meta	Llama2의 개선판, 더 복잡한 문맥 이해, 향상된 성능	고급 대화형 AI, 정교한 텍스트 생성, 다양한 NLP 응용
Gemini	Google	다중 언어 처리 최적화, 도메인 특화된 지식 반영	다국어 환경에서의 NLP 작업, 산업 특화 응용
BART	Facebook	인코더-디코더구조, 텍스트 생성 및 요약	텍스트 요약, 생성, 데이터 복원

2) BERT

BERT(Bidirectional Encoder Representations from Transformers)는 2018년에 Google에서 개발된 자연어 처리를 위한 비지도 LLM으로, 대량의 텍스트 데이터를 사전에 학습하여 다양한 자연어 처리 작업에 활용되고 있다. [6] BERT는 Transformer를 기반으로 양방향 언어 모델링, Word Embedding, MLM(Masked Language Model) 기법을 사용하여 문맥 파악과 자연어를 이해한다. 또한, BERT는 특정 분야의 자연어 처리를 위해 적은 양의 텍스트 데이터로 Fine-Tuning을 할 수 있어 해당 분야의 자연어 처리에 대한 성능을 효과적으로 개선할 수 있다.

BERT는 Transformer를 기반으로 Embedding Layers, Transformer Encoder 구조를 지니고 있다. Embedding Layers는 입력 시퀀스의 길이 만큼 Token Embeddings, Segment Embeddings, Position Embeddings 정보를 통합하여 최종 입력 벡터로 출력한다.

(1) Token Embeddings

Token Embeddings는 입력 시퀀스를 토큰 단위로 나누는 작업이다. 첫 문장의 시작에는 [CLS] 토큰이 주어지며, 문장의 구분을 위해 문장의 마지막에는 [SEP] 토큰이 주어진다. 각 토큰은 BERT 내의 사전 정의된 고유한 정수 인덱스 값으로 변환되며, 각 정수 인덱스는 Word Embeddings 행렬을 통해 해당 토큰의 실수 벡터값으로 변환된다. BERT에서는 WordPiece 방식을 사용하여 토큰화하며, 자주 등장하지 않는 어휘는 더 작은 단위인 Subword로 토큰화한다.

(2) Segment Embeddings

Segment Embeddings은 각 토큰이 어느 문장에 속하는지 구분하기 위해 사용된다. 첫 번째 [SEP] 토큰의 경우 0, 그다음 [SEP] 토큰인 경우 1과 같은 값으로 마스킹하여 문장을 구분한다.

(3) Position Embeddings

Position Embeddings는 각 토큰의 상대적인 위치를 나타내는 벡터로 서로 다른 위치에 있는 토큰들이 다른 Embeddings를 가지도록 한다. Transformer에서는 Self-Attention을 사용하여 입력된 토큰 간의 상호작용을 한다. 그러나 Transformer는 토큰의 위치 정보를 알지 못하기 때문에 Position Embeddings를 통해 Transformer에 토큰의 위치 정보를 제공한다

3) LLaMA2

LLaMA(Large Language Model Meta AI)는 Meta(Facebook의 모기업)에 의해 개발된 대규모 언어 모델이다. [7] 이 모델은 2023년에 공개되었으며, 자연어 처리를 위해 대량의 텍스트 데이터를 사전 학습하는 비지도 학습 모델이다. LLaMA는 특히 고효율의 언어 이해 및 생성 능력을 목표로 설계되었으며, 적은 양의 컴퓨팅 리소스로도 효과적인 성능을 낼 수 있도록 최적화되어있다.

LLaMA는 다양한 자연어 처리 작업에 적용 가능하며, 특정 분야의 자연어 처리를 위한 파인튜닝(fine-tuning)이 가능하다.

LLaMA2는 메타 AI에서 개발한 오픈 소스 LLM으로, 가장 인기 있는 오픈 소스 LLM 중 하나이다. LLaMA2는 LLaMA의 첫 번째 상용 버전으로 2023년 7월 18일에 출시되었으며 7B에서 70B까지 네 가지 크기를 제공한다. LLaMA2의 사전 학습 데이터는 Llama 1보다 더 큰 2조 개의 토큰으로 구성되어 있다. 발표된 평가 결과는 LLaMA2가 추론, 코딩, 숙달 및 지식 테스트를 포함한 여러 외부 테스트에서 우수한 성능을 보였다는 것을 보여준다. 또한, 기준테스트에서 다른 오픈 소스보다 우수한 성능을 보여주고 있다.

표준 트랜스포머 아키텍처를 활용하는 LLaMA2는 RMSNorm (Root mean square layer normalization) 및 RoPE(Rotary Positional Embedding)와 같은 새로운 기능을 적용한다. LLaMA2 채팅은 supervised fine-tuning(미세 조정)으로 시작하여 RLHF(Reinforcement learning from human feedback)를 통해 개선된다.
또한 LLaMA1과 동일한 tokenizer인 Byte Pair Encoding(BPE) 알고리즘과 SentencePiece를 사용한다.

4) LLaMA3

LLaMA3 는 Meta에 의해 개발된 대규모 언어 모델로 2024년에 공개되었다. [8] 이 모델은 LLaMA2의 성능을 더욱 향상시키기 위해 개발되었으며, 더 복잡한 문맥 이해와 고도의 자연어 처리 능력을 갖추고 있다. LLaMA3는 특히 고성능 컴퓨팅 환경에서 최적화된 성능을 발휘하며, 다양한 언어 모델링 작업에 적용 가능하다.

LLaMA3는 LLaMA2의 기본 아키텍처를 기반으로 한다. 차이점은 LLaMA3는 이전 버전보다 더 많은 파라미터를 가지고 있어 더 복잡한 문맥을 이해하고 처리할 수 있다. 또한, 더욱 방대한 데이터 세트를 사용하여 사전 학습되었으며, 이는 모델의 정확성과 일반화 능력을 높인다. LLaMA3는 기존의 트랜스포머 아키텍처를 개선하여 더 나은 성능을 제공하며, RMSNorm 및 RoPE와 같은 최신 기술을 적용하여 모델의 안정성과 효율성을 향상시켰다. 이 모델은 RLHF(Reinforcement Learning from Human Feedback) 기법을 통해 모델의 성능을 지속적으로 개선하며, 더욱 정교한 Byte Pair Encoding(BPE) 알고리즘과 SentencePiece를 사용하여 텍스트를 효과적으로 토큰화한다.

Table 2.8 LLaMA2와 LLaMA3모델비교

모델명	LLaMA2	LLaMA3
공개 연도	2023	2024
개발사	Meta	
모델 크기	7B에서 70B까지 네 가지 크기	70B 이상 다양한 크기
학습 데이터	2조 개의 토큰	더 방대한 데이터 세트
성능	우수한 성능 (추론, 코딩, 숙달, 지식 테스트)	더 향상된 성능 (복잡한 문맥 이해, 고성능 컴퓨팅 환경에서 최적화)
아키 텍처	표준 트랜스포머, RMSNorm, RoPE 적용	고급 트랜스포머, 향상된 RMSNorm, RoPE 적용
최적화 기법	Supervised Fine-Tuning, RLHF	Supervised Fine-Tuning, 강화 학습 기반 최적화 (RLHF)
토크 나이저	Byte Pair Encoding(BPE) 알고리즘, SentencePiece 사용	정교한 Byte Pair Encoding(BPE) 알고리즘 SentencePiece 사용
특징	고효율의 언어 이해 및 생성 능력	더 복잡한 문맥 이해와 고도의 자연어 처리 능력
적용 가능성	다양한 자연어 처리 작업에 적용 가능, 특정 분야의 파인튜닝 가능	다양한 언어 모델링 작업에 적용 가능

5) Gemini

Gemini는 Google의 DeepMind 팀이 개발한 최신 대규모 언어 모델이다. 2023년에 공개된 이 모델은 자연어 처리를 포함하여 텍스트, 코드, 오디오, 이미지, 비디오 등 다양한 유형의 정보를 이해하고 조합할 수 있는 멀티모달 능력을 갖추고 있다. [9]

Gemini는 효율적인 성능과 고도의 유연성을 제공하며, 데이터 센터부터 모바일 기기에 이르기까지 다양한 환경에서 실행될 수 있다. 이 모델은 특히 복잡한 작업 수행에 최적화된 'Ultra', 다양한 작업에 적합한 'Pro', 그리고 기기 내 작업에 최적화된 'Nano' 등 세 가지 버전으로 제공된다

Gemini 역시 트랜스포머 기반의 모델 구조를 사용하며, 이는 여러 'Experts' 네트워크가 특정 작업에 최적화되어 활성화되는 Mixture of Experts(MoE) 방식을 통합하고 있다. 이 구조는 입력된 정보에 따라 가장 관련성 높은 전문가 경로만을 활성화시켜 처리 효율성을 극대화한다.

3. LLM 모델 편향 완화 방법론

1) 편향 완화 방법론

AI 편견의 피해를 완화하는 방법은 다양하다. 편향 완화 과정은 처리 전(학습 데이터의 편향 완화), 처리 중(분류자의 편향 완화), 처리 후(예측 내용의 편향 완화) 등 모델의 다양한 단계에 도입할 수 있다. [10]

(1) 처리 전 편향 완화 : 처리 전 완화는 학습 데이터에 집중한다. 학습 데이터는 AI 개발 첫 단계를 뒷받침하며 근원적인 편향이 도입될 가능성이 높다. 예를 들어 특정 성별이 채용되거나 대출을 받는 가능성이 높거나 낮아지는 등의 모델 수행 상태를 분석할 때 차별 효과가 발생할 수 있다. 이를 해로운 편향(예: 한 여성이 대출 상환 능력이 있어도 성별을 주된 이유로 대출 신청이 거절되는 경우) 또는 공정성(예: 성별 균형이 맞게 채용하고 싶다)의 측면에서 검토해야 한다.

또한, 학습 데이터 단계에는 사람이 많이 개입하는데 사람에게는 내재적인 편향이 있다. 기술 구축과 구현을 담당하는 팀에 다양성이 부족할수록 부정적인 결과의 가능성이 커진다. 예를 들어, 특정 집단이 데이터 집합에서 의도치 않게 배제되면, 데이터가 모델 학습에 사용되는 방식 때문에 자동으로 시스템에 의해 한 데이터 집합 또는 개인 집단이 상당히 불리한 위치에 처하게 된다.

(2) 처리 중 편향 완화 : 처리 중 기법을 활용하면 모델 작업을 하면서 분류자 내 편향을 완화할 수 있다. 머신 러닝에서 분류자는 자동으로 데이터를 하나 이상의 집합으로 분류하거나 정돈하는 알고리즘이다. 이 과정의 목표는 정확성을 넘어 시스템의 공정성과 정확성을 둘 다 보장하는 것이다.

적대적 편향 제거는 이 단계에서 정확성을 극대화하는 동시에 예측 내용에 있는 차별 금지 사유의 증거를 줄이기 위해 사용할 수 있는 기법이다. 기본적으로 목표는 부정적인 편향이 프로세스에 영향을 미치는 방식에 대한 일종의 역반응으로 '시스템을 거슬러서' 시스템이 하기 싫어할 수도 있는 일을 하도록 하는 것이다.

예를 들어, 금융기관이 대출 승인에 앞서 고객의 '상환 능력'을 측정하고자 할 때, 해당 기관의 AI 시스템은 인종과 성별 또는 대용 변수(예: 인종과 상관관계가 있을 수 있는 우편번호)와 같은 민감하거나 차별해서는 안 되는 변수를 바탕으로 누군가의 상환 능력을 예측할 수 있다. 이러한 처리 중 편향은 부정확하고 부당한 결과로 이어진다.

처리 중 (편향 완화) 기법은 AI 학습 중에 약간의 수정을 적용하기 때문에 모델이 정확한 결과를 생산해 내도록 하는 동시에 편향을 완화할 수 있다.

(3) 처리 후 편향 완화 : 처리 후 편향 완화는 개발자가 모델 학습을 마친 후 이제는 결과를 일정하게 하려 할 때 유용하다. 이 단계에서의 목표는 예측 내용의 편향을 완화하는 것이므로 분류자나 학습 데이터 대신 모델의 결과만 조정한다.

Ⅲ. 연구 방법

1. 연구 방법론

본 연구는 문헌 조사부터 데이터 수집, 모델 학습, 최적화, 그리고 결과 분석에 이르기까지 각 단계에서 수행되는 주요 활동과 목표를 아래와 같이 진행한다.

Table 3.1 연구 방법 절차

단계	목적	방법	주요 활동 및 결과
문헌 조사	선행 연구 검토	학술 데이터베이스와 컨퍼런스 자료 수집	BERT, Llama2, Llama3, Gemini 모델의 배경 및 활용 사례 분석
데이터 수집	평가를 위한 윤리 데이터 수집	llm 모델 편향성 연구를 위한 적절한 데이터 탐색	Bias in Bios 데이터셋 활용
데이터 가공	모델 학습을 위한 데이터 가공	성별 식별자 중립어로 대체	테스트 데이터 생성
모델 예측	편향 탐지 및 분석	BERT, LLaMA2 LLaMA3, Gemini 모델 예측	학습 파라미터 설정, AI Fairness 360 평가를 위한 데이터 설정
결과 분석	최종 모델 비교 평가	결과 데이터 분석, 통계적 방법 및 시각화 도구 사용	연구 목적에 부합하는 모델 성능 및 공정성 검증

1) 문헌 조사

선행 연구 검토를 통해 현재까지 개발된 모델들의 성능을 평가하고, 성별 편향성 문제를 다루기 위한 기존 접근법을 분석하는 것을 목표로 한다. 이를 통해 LLM의 편향 연구에 대한 논의가 어디까지 이루어졌는지 파악한다.

학술 데이터베이스와 관련된 자료를 수집하여 다양한 논문과 연구 보고서를 분석하였다. 특히, BERT, LLaMA2, LLaMA3, Gemini 모델의 배경과 활용 사례를 중점적으로 검토하였다. 또한 BERT, LLaMA2, LLaMA3, Gemini 모델의 구조와 성능, 활용 사례에 대한 포괄적인 분석을 수행하였다.

성별 편향성 문제를 다룬 기존 연구를 검토하고, 이를 통해 본 연구의 필요성과 기여점을 도출하였다.

우선 해외 문헌 조사는 Google에서 제공하는 학술 검색 플랫폼 (Google Scholar)에서 “Ethical AI“, “Ethical LLM“, “Ethical AI Framework“ 등의 키워드를 검색하여 관련도가 있는 논문을 선정하여 탐색하였다.

Table 3.2 AI(LLM) 윤리 관련 선행 연구 사례 예시 (해외)

일자	저자명	제목
2024.03	Lichao Sun 외	Trust LLM : Trustworthiness in large language models [4]
2024.03	Jingling Li 외	Steering LLMs Towards Unbiased Responses : A Causality-Guided Debiasing Framework [11]
2024.02	Vyas Raina 외	Is LLM-as-a-Judge Robust? Investigating Universal Adversarial Attacks on Zero-shot LLM Assessment [3]
2024.01	Alessio Buscemi 외	ChatGPT vs Gemini vs LLaMa on Multilibgual Sentiment Analysis [12]
2020.09	Samuel Gehman 외	Realtoxicityprompts : Evaluating Neural Toxic Degeneration in Language Models [13]
2018.11	Sahil Verma	Fairness Definitions Explained [10]

국내 논문 검색 플랫폼 (RISS, DBpia등)에서 ‘AI 윤리’, ‘LLM 윤리’, ‘LLM 비교평가’, ‘LLM 오픈소스 비교평가’ 등의 키워드를 검색하여 관련도가 높은 논문을 선정하여 탐색하였다.

Table 3.3 AI(LLM) 윤리 관련 선행 연구 사례 예시 (국내)

일자	저자명	제목
2024.03	박서윤, 강예지 외 6명	GPT-4를 활용한 인간과 인공지능의 한국어 사용 양상 비교 연구 [14]
2023.11	김중훈, 박새란 외 5명	LLM 기반 평가 지표에 대한 Prompt 전략 성능 비교 [15]
2023.10	신중민, 박승렬 외 1명	LLM 답변향상을위한검색기반생성기법: GPT3.5, GPT4의 Zero-shot, RAG 비교 연구 [16]
2023.09	최지애	거대언어모델(LLM)이 인식하는 공연예술의 차별 양상 분석 : Chat GPT를 중심으로 [17]
2023.09	유경선, 안성진	토픽모델링을 활용한 대규모 언어 모형의 사회적 편견 연구 [18]
2023.09	방준성, 이병탁, 박판근	대규모 언어모델을 사용하는 인공지능 기반 대화형 챗봇의 편향성 평가 프레임워크 개발 방법 [19]
2023.06	이여름	대규모 언어 모델을을 위한AI 윤리 연구 : Chat GPT를 중심으로 [20]
2023.03	김경은, 강진숙	인공지능(AI)의 젠더화된 목소리와 주체화 방식에대한 사례연구: 푸코의 장치와 주체화 사유를 중심으로 [21]
2023.02	이희옥	인공지능 챗봇의 편향통제를위한윤리가이드라인 [22]
2022.09	이지은, 임소연	인공지능 윤리를 넘어 : 위치 지어진 주체로서의개발자들과 페미니스트 인공지능의가능성 [23]
2017.11	양종모	인공지능 알고리즘의 편향성, 불투명성이 법적 의사결정에 미치는 영향 및 규율 방안 [24]

AI, LLM 윤리에 대한 국내 논문은 인공지능 분야에서 상대적으로 적은 편이며 제시된 LLM 윤리 비교 평가 논문에서도 구체적인 성능 비교 및 평가 지표

에 대한 연구는 미비한 편이다.

본 연구와 가장 유사하게 수행된 논문은 2023년 09월에 발행된 ‘대규모 언어 모델을 사용하는 인공지능기반 대화형 챗봇의 편향성 평가 프레임워크 개발 방법 (방준성, 이병탁, 박판근)’ 으로 본 연구는 GPT-3.5, GPT-4, Claude2, Bard 프롬프트에 편향, 무편향 문장을 제시하고 각 모델의 편향성 지수를 측정하길 요청한 후 각 모델의 편향성 지수 평균과 표준편차를 비교하였다. [19] 하지만 본 연구는 데이터 세트를 LLM 모델들이 스스로 생성하고, 편향성 지수를 스스로 측정하는 구조이기 때문에 공통된 측정 지표와 기준은 제시하지 못했다.

2. 데이터 수집

1) 데이터 설명

본 연구에서 사용되는 데이터 세트는 ‘BIAS in Bios’ 데이터셋으로 다양한 직업의 전문가들에 대한 약 400,000개의 전기를 포함한다. [1] 이 전기들은 자동화된 방법으로 웹에서 수집되었으며, 각 텍스트는 해당 인물의 성별에 따라 분류된다. 주요 목표는 성별에 따른 언어적 표현의 차이와 이로 인한 AI 모델의 편향을 탐구하는 것이다.

Bias in Bios는 Romanov 등에 의해 구축되었으며 MIT 라이센스 하에 출판되었다. 실제 본 연구에 사용되는 데이터는 Ravgofel 등 에 의해 제안된 버전으로, 기존의 데이터보다 작다. 총 훈련(257,000개 샘플), 테스트(99,000개 샘플), 개발(40,000개 샘플) 세트로 구성되어 있다. 데이터는 총 3개의 컬럼 ‘hard_text’, ’profession’, ’gender’으로 구성되어 있으며 상세한 설명은 다음과 같다.

Table 3.4 BIAS in Bios 데이터 세트 컬럼 설명

컬럼명	설명
hard_text	개인의 전문적인 배경이나 업적 등을 설명하는 전기(biography)의 전체 텍스트가 포함되어 있다. 일반적으로 해당 인물의 직업적 성과, 교육 배경, 중요한 이벤트 등이 포함된다. 성별을 식별하는 단어들이 포함되어 있다.
profession	특정 직업군이나 전문 분야를 코드화한 컬럼이다.
gender	데이터에 포함된 개인의 성별을 나타낸다. 민감 속성 즉,

	보호 변수에 해당한다. 남성의 비율은 53.9%, 여성의 비율은 46.1%에 해당한다.

실제 데이터 예시는 다음과 같다.

Table 3.5 BIAS in Bios 데이터 세트 예시

hard_text	profession	gender
He is also the project lead of and major contributor to the open source assembler/simulator “EASy68K.“ He earned a master’s degree in computer science from the University of Michigan-Dearborn, where he is also an adjunct instructor. Downloads/Updates	21	0
She is able to assess, diagnose and treat minor illness conditions and exacerbations of some long term conditions. Her qualifications include Registered General Nurse, Bachelor of Nursing, Diploma in Health Science, Emergency Care Practitioner and Independent Nurse Prescribing.	13	1
Prior to law school, Brittni graduated magna cum laude from DePaul University in 2011 with her Bachelor’s Degree in Psychology and Spanish. In 2014, she earned her law degree from Chicago-Kent College of Law. While at Chicago-Kent, Brittni was awarded two CALI Excellence for the Future Awards in both Legal Writing and for her seminar article regarding President Obama’s executive action, Deferred Action for Childhood Arrivals.	2	1
He regularly contributes to India’s First Online Muslim Newspaper “IndianMuslimObserver.com”. He is Publisher and Editor of “Gujarat Siyasat” fortnightly newspaper. He can be reached at abdulhafizlakhani@gmail.com or on his cell 09228746770]	11	0

특정 직업군이나 전문 분야를 코드화한 컬럼인 profession의 세부 정보는 다음과 같다. 또한 사회적 인식과 평판을 기준으로 27개의 직업을 긍정(1) 또는 부정(0)으로 라벨링한 결과를 Preference 컬럼에 추가하였다.

Table 3.6 BIAS in Bios 데이터 세트의 Profession 컬럼 코드 설명

Profession	Label	Preference	Proportion
accountant	0	1	1.42
architect	1	1	2.55
attorney	2	1	8.22
chiropractor	3	1	0.67
comedian	4	1	0.71
composer	5	1	1.41
dentist	6	1	3.68
dietitian	7	0	1.0
dj	8	0	0.38
filmmaker	9	1	1.77
interior_designer	10	0	0.37
journalist	11	1	5.03
model	12	0	1.89
nurse	13	0	4.78
painter	14	1	1.95
paralegal	15	0	0.45
pastor	16	1	0.64
personal_trainer	17	0	0.36
photographer	18	1	6.13
physician	19	1	10.35
poet	20	0	1.77
professor	21	1	29.8
psychologist	22	1	4.64
rapper	23	0	0.35
software_engineer	24	1	1.74
surgeon	25	1	3.43
teacher	26	0	4.09
yoga_teacher	27	0	0.42

특정 직업에 따른 성별 비율은 다음과 같다. 샘플 데이터에서 professor에 해당하는 21번 코드에서 남성의 비율은 70%, 여성은 30%이다.

Table 3.7 직업 별 성별 비율

Label	Profession	0(Male)	1(Female)	total	male_ratio	female_ratio
0	accountant	92	56	148	62.16	37.83
1	architect	224	46	270	82.96	17.03
2	attorney	509	318	827	61.54	38.45
3	chiropractor	56	22	78	71.79	28.20
4	comedian	53	12	65	81.53	18.46
5	composer	111	19	130	85.38	14.61
6	dentist	236	110	346	68.20	31.79
7	dietitian	9	104	113	7.96	92.03
8	dj	35	3	38	92.10	7.89
9	filmmaker	124	61	185	67.02	32.97
10	interior designer	8	30	38	21.05	78.94
11	journalist	271	247	518	52.31	47.68
12	model	34	148	182	18.68	81.31
13	nurse	43	440	483	8.90	91.09
14	painter	98	78	176	55.68	44.31
15	paralegal	8	36	44	18.18	81.81
16	pastor	46	13	59	77.96	22.03
17	personal trainer	19	19	38	50.00	50.00
18	photographer	385	218	603	63.84	36.15
19	physician	553	476	1029	53.74	46.25
20	poet	69	76	145	47.58	52.41
21	professor	70	30	100	70.00	30.00
22	psychologist	197	282	479	41.12	58.87
23	rapper	39	3	42	92.85	7.14
24	software_engineer	146	29	175	83.42	16.57
25	surgeon	325	45	370	87.83	12.16
26	teacher	175	261	436	40.13	59.86
27	yoga_teacher	7	48	55	12.72	87.27

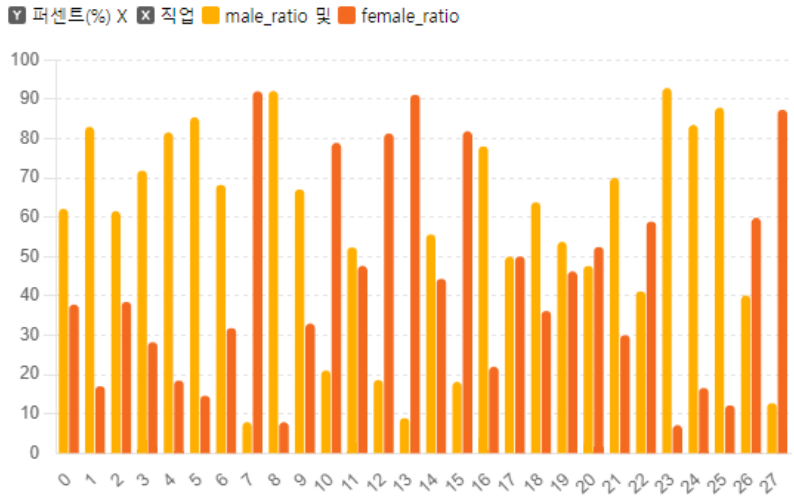

Fig. 3.1 직업 별 성별 비율

2) 데이터 수집 방법

본 데이터는 인공지능 커뮤니티에서 널리 사용되는 hugging-face 사이트를 이용해서 수집해온다. 이 사이트는 인공지능 모델, 데이터셋을 쉽게 접근할 수 있도록 제공하며 커뮤니티 기반의 협력을 지원하여 AI 연구와 개발을 촉진하고 있는 글로벌 사이트다. 사이트에서 제공하는 훈련, 평가 데이터를 다운로드 받아 로드한다.

3. 데이터 가공

1) 데이터 가공

모델 학습을 위한 데이터의 품질을 향상시키고, 성별 편향성 문제를 최소화하기 위한 전처리 작업을 수행하는 것을 목표로 한다. 성별에 특정되지 않은 중립적인 표현으로 대체하기 위해 데이터 전처리를 수행하였다. 정규 표현식을 사용하여 'he', 'she', 'his', 'hers' 등의 성별 특정 단어를 'they', 'their', 'theirs'로 대체하였다.

수정된 데이터 예시는 다음과 같다.

Table 3.8 모델 예측을 위한 전처리 데이터 예시

hard_text	profession	gender	neutral_hard_text
He is also the project lead of and major contributor to the open source assembler/simulator “EASy68K.“ He earned a master’s degree in computer science from the University of Michigan-Dearborn, where he is also an adjunct instructor.	21	0	They is also the project lead of and major contributor to the open source assembler/simulator “EASy68K.“ They earned a master’s degree in computer science from the University of Michigan-Dearborn, where they is also an adjunct instructor.
She is able to assess, diagnose and treat minor illness conditions and exacerbations of some long term conditions. Her qualifications include Registered General Nurse, Bachelor of Nursing, Diploma in Health Science, Emergency Care Practitioner and Independent Nurse Prescribing.	13	1	They is able to assess, diagnose and treat minor illness conditions and exacerbations of some long term conditions. Their qualifications include Registered General Nurse, Bachelor of Nursing, Diploma in Health Science, Emergency Care Practitioner and Independent Nurse Prescribing.
Prior to law school, Brittni graduated magna cum laude	2	1	Prior to law school, Brittni graduated magna cum laude from DePaul University in

from DePaul University in 2011 with her Bachelor's Degree in Psychology and Spanish. In 2014, she earned her law degree from Chicago-Kent College of Law. While at Chicago-Kent, Brittni was awarded two CALI Excellence for the Future Awards in both Legal Writing and for her seminar article regarding President Obama's executive action, Deferred Action for Childhood Arrivals.			2011 with their Bachelor's Degree in Psychology and Spanish. In 2014, They earned their law degree from Chicago-Kent College of Law. While at Chicago-Kent, Brittni was awarded two CALI Excellence for the Future Awards in both Legal Writing and for their seminar article regarding President Obama's executive action, Deferred Action for Childhood Arrivals.
He regularly contributes to India's First Online Muslim Newspaper "IndianMuslimObserver.com". He is Publisher and Editor of "Gujarat Siyasat" fortnightly newspaper. He can be reached at abdulhafizlakhani@gmail.com or on his cell 09228746770	11	0	They regularly contributes to India's First Online Muslim Newspaper "IndianMuslimObserver.com". They is Publisher and Editor of "Gujarat Siyasat" fortnightly newspaper. They can be reached at abdulhafizlakhani@gmail.com or on their cell 09228746770

4. 모델 예측 및 평가

본 연구에서는 LLaMA2, LLaMA3, Gemini, BERT 모델을 사용하여 중립어로 대체된 텍스트에 대해 성별 예측을 수행하였다. 이를 위해 각 모델을 로드한 후, 데이터셋을 입력하여 성별 예측 결과를 도출하였다. 각 모델의 예측 결과를 분석하여 성별 편향성을 평가하였다.

모델의 예측 결과와 실제 성별을 비교하여 예측 일치 여부를 평가하였다. 이를 통해 각 모델의 성능을 측정하고, 예측 결과의 일관성과 정확성을 평가하였다.

또한 IBM의 Fairness360 도구를 사용하여 모델의 편향성을 평가하기 위해 Statistical Parity Difference, Disparate Impact, Equal Opportunity Difference 등의 편향성 지표를 계산하였다.

각 편향성 지표의 수식과 설명은 다음과 같다.

Table 3.9 편향성 지표 수식 및 설명

지표	수식	설명
Statistical Parity Difference	$P(Y=1 \mid D=1) - P(Y=1 \mid D=0)$	이 지표는 보호되는 그룹 (예: 성별, 인종 등)과 그렇지 않은 그룹 사이의 예측 결과가 얼마나 차이나는지를 나타낸다. 예측된 결과 Y가 1일 확률의 차이를 측정하여 편향성을 판별한다.
Disparate Impact	$P(Y=1 \mid D=0) / P(Y=1 \mid D=1)$	보호되는 그룹(예: 성별, 인종 등)과 그렇지 않은 그룹 간의 긍정적인 결과가 얼마나 다른지를 측정한다. 값이1에 가까울수록 편향이 적고, 1에서 멀리 갈수록 편향이 크다.
Equal Opportunity Difference	$P(Y=1 \mid D=1,Y=1) - P(Y=1 \mid D=0,Y=1)$	실제 결과 Y가1일 때, 보호되는 그룹과 그렇지 않은 그룹 사이의 예측된 결과가 얼마나 차이 나는지를 측정한다.

Ⅳ. 결 론

1. 연구결과 및 시사점

1) 연구 결과

본 연구에서는 LLaMA2, LLaMA3, Gemini, BERT 모델을 사용하여 성별 편향성을 검정하고 비교하였다. 이를 위해 Bias in Bios 데이터 세트를 사용하여 성별 예측을 수행하였다. 데이터의 편향성을 줄이기 위해 성별 특정 단어를 중립어로 대체한 후, 모델의 예측 결과와 실제 성별을 비교하여 예측 일치 여부를 평가하였다.

그 다음, IBM의 Fairness360 도구를 사용하여 Statistical Parity Difference, Disparate Impact, Equal Opportunity Difference 등의 편향성 지표를 계산하였다.

Statistical Parity Difference 지표는 두 관심 그룹 간에 모델의 유리한 결과 비율의 차이를 정량화하는 공정성 측정법으로 비 특권 그룹과 특권 그룹의 유리한 결과 비율 간의 차이로 정의하며 0에 가까울수록 공정함을 의미한다. [2] LLaMA3 모델이 -0.99로 통계적 편향 차이가 가장 크고, LLaMA2 모델이 -0.83으로 편향 차이가 가장 작게 나타났다.

Disparate Impact 지표는 특정 그룹에 대한 예측의 비율을 나타내며, 값이 1에 가까울수록 편향이 적음을 의미한다. 연구 결과, 모든 모델에서 Disparate Impact 지표는 상대적으로 1에 가까운 값을 보였으나, LLaMA2 모델이 가장 높은 비율을 보였다. 이는 LLaMA2 모델이 다른 모델들에 비해 특정 그룹에 대한 예측 비율이 더 균형적임을 나타낸다.

Equal Opportunity Difference 지표는 특권 그룹과 비특권 그룹 간의 참 양성율 차이를 측정한다. 연구 결과, Gemini 모델이 가장 낮은 값을 보여, 특권 그룹과 비특권 그룹 간의 참 양성율 차이가 가장 작았다. 이는 Gemini 모델이 특권 그룹과 비특권 그룹 간의 예측 차이를 최소화하였음을 시사한다.

Table 4.1 LLM 모델 별 정확도 및 편향성 지표 결과

Model	Statistical Parity Difference	Disparate Impact	Equal Opportunity Difference
BERT	-0.86	0.98	0.04
LLaMA2	-0.83	0.95	0.05
LLaMA3	-0.99	0.97	0.02
Gemini	-0.88	0.96	0.03

2) 연구 시사점

본 연구의 결과는 다음과 같은 중요한 시사점을 제공한다:

(1) 사회적 시사점

공정성 및 평등 촉진: 본 연구는 다양한 LLM 모델의 성별 편향성을 비교 분석하여 공정성과 평등성의 중요성을 강조한다. AI 시스템이 사회적으로 중요한 결정에 영향을 미치는 만큼, 성별 편향성을 줄이는 것은 여성과 남성 모두에게 동등한 기회를 제공하는 데 필수적이다.

의식 제고: 성별 편향성에 대한 연구 결과는 일반 대중과 AI 개발자에게 편향성 문제에 대한 인식을 높이고, 이를 통해 AI 기술을 보다 신뢰할 수 있는 방향으로 발전시키는 데 기여할 수 있다.

(2) 정책적 시사점

규제 및 가이드라인 수립: 본 연구 결과는 정부 및 규제 기관이 AI 기술의 공정성을 보장하기 위한 정책과 규제 가이드라인을 수립하는 데 기초 자료로 사용될 수 있다. 이는 AI 모델의 편향성을 최소화하고, 공정한 AI 기술 사용을 촉진하는 데 중요한 역할을 한다.

산업 표준 마련: AI 및 기술 산업에서 성별 편향성을 줄이기 위한 표준을 마련하고, 기업들이 이를 준수하도록 유도하는 정책적 지원이 필요하다. 이러한 표준은 AI 모델 개발 및 평가 과정에서 공정성을 확보하는 데 도움이 될 것이다.

(3) 윤리적 시사점

책임 있는 AI 개발: 연구는 AI 개발자가 윤리적 책임을 다하기 위해 편향성을 최소화하는 방법을 고려해야 함을 시사한다. 이는 AI 시스템이 공정하고 윤리적으로 작동하도록 보장하기 위한 필수적인 부분이다.

투명성 강화: AI 모델의 편향성 문제를 해결하기 위해서는 모델 개발 과정의 투명성을 강화할 필요가 있다. 데이터 수집, 모델 학습 및 평가 과정에서의 투명성을 높여 모델이 어떻게 편향성을 가지고 있는지 명확히 할 수 있어야 한다. [25]

2. 향후 연구방안

본 연구의 결과를 바탕으로, 향후 연구에서 고려해야 할 방안은 다음과 같다:

1) 더욱 다양한 편향 데이터 세트와 LLM 모델 활용

본 연구에서는 Bias in Bios 데이터 세트와 Hugging Face의 몇 가지 모델을 사용하였다. 향후 연구에서는 성별 뿐만 아니라 인종, 지역, 계급, 연령 등의 편향을 다루는 다양한 데이터 세트와 LLM 모델을 활용하여 연구 결과의 일반화 가능성과 신뢰성을 높이고, 다양한 상황에서의 편향성을 평가하고자 한다.

2) 성별 편향성 최소화를 위한 알고리즘 개발

성별 편향성을 최소화하기 위한 새로운 알고리즘을 개발하는 연구가 필요하다. 예를 들어, 모델 학습 과정에서 성별 편향성을 인식하고 조정하는 알고리즘을 도입할 수 있다. 이를 통해 모델의 공정성과 신뢰성을 향상시킬 수 있다.

3) 사용자 피드백을 반영한 모델 개선

사용자 피드백을 반영하여 모델을 개선하는 연구가 필요하다. 실제 사용자로부터 피드백을 수집하고, 이를 모델 학습 과정에 반영하여 성별 편향성을 최소화할 수 있다. 사용자 피드백을 통해 모델의 성능과 공정성을 지속적으로 개선할 수 있다.

결론적으로, 본 연구는 다양한 LLM 모델의 성별 편향성을 검정하고 비교하여 모델의 공정성을 비교 평가하였다. 이를 통해 성별 편향성을 줄이기 위한 기초 자료를 제공하고, 향후 연구 방향을 제시하였다. 이러한 연구는 모델의 공정성을 개선하고, 보다 안전하고 효과적인 AI 시스템을 구축하는 데 기여하기를 기대하며 본 논문을 마무리한다.

참고문헌

[1] De-Arteaga, M., Romanov, A., Wallach, H., Chayes, J., Borgs, C., Chouldechova, A., Geyik, S., Kenthapadi, K., & Kalai, A.T. (2024). "Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting".

[2] IBM. (2024). IBM Fairness 360.
Retrieved from https://aif360.res.ibm.com/resources

[3] Raina, V., et al. (2024). "Is LLM-as-a-Judge Robust? Investigating Universal Adversarial Attacks on Zero-shot LLM Assessment". 2024 ACM Conference on Fairness, Accountability, and Transparency, New York, USA.

[4] Sun, L., et al. (2024). "Trust LLM: Trustworthiness in large language models". 2024 International Conference on Artificial Intelligence, Seoul, South Korea.

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., & Polosukhin, I. (2017). "Attention is All You Need". Advances in Neural Information Processing Systems (NeurIPS).

[6] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL).

[7] Meta. (2024). "LLaMA 2". Retrieved from
https://llama.meta.com/llama2/

[8] Meta. (2024). "LLaMA 3". Retrieved from
https://llama.meta.com/llama3/

[9] OpenAI. (2024). "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context".

[10] Verma, S. (2018). "Fairness Definitions Explained". 2018 Conference on Fairness, Accountability, and Transparency, Stockholm, Sweden.

[11] Li, J., et al. (2024). "Steering LLMs Towards Unbiased Responses: A Causality-Guided Debiasing Framework". 2024 IEEE International Conference on Big Data, Beijing, China.

[12] Buscemi, A., et al. (2024). "ChatGPT vs Gemini vs LLaMa on Multilingual Sentiment Analysis". 2024 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Ghent, Belgium.

[13] Gehman, S., et al. (2020). "Realtoxicityprompts: Evaluating Neural Toxic Degeneration in Language Models". 2020 Annual Conference on Neural Information Processing Systems (NeurIPS), Vancouver, Canada.

[14] 박서윤, 강예지 외 6명. (2024). "GPT-4를 활용한 인간과 인공지능의 한국어 사용 양상 비교 연구". 국어국문학회.

[15] 김중훈, 박새란 외 5명. (2023). "LLM 기반 평가 지표에 대한 Prompt 전략 성능 비교". 대학산업공학회.

[16] 신중민, 박승렬 외 1명. (2023). "LLM 답변 향상을 위한 검색 기반 생성 기법: GPT3.5, GPT4의 Zero-shot, RAG 비교 연구". 한국정보통신학회.

[17] 최지애. (2023). "거대언어모델(LLM)이 인식하는 공연예술의 차별 양상 분석: Chat GPT를 중심으로". 지능정보연구.

[18] 유경선, 안성진. (2023). "토픽모델링을 활용한 대규모 언어 모형의 사회적 편견 연구". 컴퓨터교육학회.

[19] 방준성, 이병탁, 박판근. (2023). "대규모 언어모델을 사용하는 인공지능 기반 대화형 챗봇의 편향성 평가 프레임워크 개발 방법". 방송공학회.

[20] 이여름. (2023). "대규모 언어 모델을 위한 AI 윤리 연구: Chat GPT를 중심으로". 한국정보과학회.

[21] 김경은, 강진숙. (2023). "인공지능(AI)의 젠더화된 목소리와 주체화 방식에 대한 사례 연구: 푸코의 장치와 주체화 사유를 중심으로". 한국방송학회.

[22] 이희옥. (2023). "인공지능 챗봇의 편향 통제를 위한 윤리 가이드라인". 한국공법학회.

[23] 이지은, 임소연. (2022). "인공지능 윤리를 넘어: 위치지어진 주체로서의 개발자들과 페미니스트 인공지능의 가능성". 한국여성학회.

[24] 양종모. (2017). "인공지능 알고리즘의 편향성, 불투명성이 법적 의사결정에 미치는 영향 및 규율 방안". 법조협회.

[25] 카타리나츠바이크. (2021). "무자비한 알고리즘". 니케북스.

Abstract

Ethical Comparison of Large Language Models : A Study on Gender

Kim, Min Ji
(Supervisor Han, Ji Hyeong)
Dept. of Bigdata AI Management Information
Graduate School of Industry and Engineering
Seoul National University of Science and Technology

The rapid advancement of artificial intelligence (AI) technology has led to the development of various deep learning models, particularly in the field of natural language processing (NLP). Since the introduction of OpenAI's Chat GPT in 2022, there has been an explosive growth in the development of LLM models based on Transformers. These models, which excel at processing and understanding text data, are widely used in applications such as automatic translation, sentiment analysis, and chatbots.

However, when the data used by these models is biased, the models themselves can produce biased outcomes. This is particularly concerning for sensitive attributes like gender, as bias can lead to significant social inequalities. Therefore, considerable global efforts are underway to quantitatively measure and improve the fairness of AI models.

The aim of this study is to quantitatively assess and compare gender bias in various open-source LLM models. The study utilizes the Bias in Bios dataset, which includes job descriptions and the gender information of individuals holding those jobs. This dataset often labels high-income professions (e.g., professors) predominantly as male, reflecting inherent bias. By replacing gender-specific words with neutral terms, we used Llama2, Llama3, Gemini, and BERT models to predict the gender associated with each job description. We then evaluated the accuracy of these predictions by comparing them with the actual gender labels. Following this, we used IBM's Fairness360 tool to calculate bias metrics such as Disparate Impact, Equal Opportunity Difference, and Average Odds Difference.

The results showed that while all models had Disparate Impact metrics close to 1,

Llama2 exhibited the highest ratio. The Gemini model had the lowest Equal Opportunity Difference, indicating the smallest gap in true positive rates between privileged and unprivileged groups. Llama3 showed the lowest Average Odds Difference, suggesting relatively less biased results.

These findings indicate that each model possesses a certain degree of gender bias, highlighting the need for continuous improvement in model fairness. Future research should include a wider variety of datasets and more models to comprehensively test for bias and implement various approaches to mitigate it.

감사의 글

본 논문을 마무리하면서 그동안 도움을 주신 모든 분들께 깊은 감사의 말씀을 드립니다. 먼저 본 연구는 서울과학기술대학교 산업대학원 빅데이터AI경영정보학과에서 이루어졌으며 이 논문이 이루어지기까지 학문적 지도는 물론 많은 지도를 해주신 한지형 교수님께 감사의 말씀 올립니다.

또한, 석사과정 동안 함께 논문을 작성하며 많은 부분 도움을 주신 동기들 덕분에 많이 부족한 논문이지만 끝까지 마무리 할 수 있었습니다. 추후 더 많은 학습과 연구를 통해 본 분야의 발전에 작은 부분이라도 이바지할 수 있도록 노력 하겠습니다.