윤리의 사각지대에 존재하는

대규모 언어모델

대규모 언어모델의 위험성과 그에 대한 대응

과목명	글쓰기(공학인증)
담당교수	박필현 교수님
이름	최유나
학과	소프트웨어학과
학번	20182831
연락처	010-7594-8151
제출일	2023-06-11

윤리의 사각지대에 존재하는 대규모 언어모델

대규모 언어모델의 위험성과 그에 대한 대응

최유나 (소프트웨어학과)

-목차-

I. 서론

최근 chatGPT 가 등장함에 따라 대규모 언어 모델에 대한 사람들의 관심이 높아지고 있다. chatGPT 는 OpenAI 에서 개발한 인공지능 언어 생성 모델로, 사용자와 대화 형식으로 대화를 주고받을 수 있다. 다양한 주제에 대해서도 대화가 가능하며, 대화의 맥락을 고려할 수 있어 마치 사람이 대답하는 것과 같은 착각이 들기도 한다. chatGPT 와 같은 대규모 언어 모델은 사이트에 접속하여 로그인만 하면 채팅창이 등장하여 사용할 수 있다. 채팅창에 입력만 하면 쉽게 사용할 수 있어 모델 내부의 복잡한 구조를 알지 못하더라도 사용할 수 있다. 때문에 사용자들은 인공지능에 대한 지식이 없이도 간편하게 궁금증을 해결하거나 효율적으로 업무를 진행할 수 있어 만족도가 높다. chatGPT 는 다양하게 활용 될 수 있는데, 한 예로 충남교육청에서는 대화형 인공지능 챗봇을 활용한 도움자료를 개발해 일선 학교에 보급하여 활용하고 있다(이찬선, 2023). 이처럼 다양하게 활용가능한 대규모 언어 모델의 인기는 지속적으로 증가하였으며 사용자 또한 늘어나고 있다.

하지만 우리는 이러한 편리함에 속아 대규모 언어 모델의 한계와 문제점을 간과하기 쉽다. 대부분의 사람들은 언어 모델을 사용하고 얻은 답변에 집중하며, 어떤 메커니즘을 통해 그 답변이 도출되었는지에 대해서는 고려하지 않는다. 대규모 언어 모델을 생성하기 위해서는 방대한 양의 학습 데이터들이 필요한데, 이 학습 데이터들은 사람들에 의해 생성된 데이터들이다. 때문에 이 모델은 학습 데이터에 있는 편견이나 차별 등의 정보를 그대로 학습하게 된다. 또한 학습 데이터에 민감한 개인정보를 익명화 하지 않아 중요한 정보가 유출될 가능성도 존재한다. 사람들은 이러한 사실에 대해 유의하지 않기 때문에 모델이 생성한 결과들을 검증 없이 신뢰할 수 있다. 이는 또 다른 편견들을 생성하거나 새로운 피해를 생성할 수 있어 이를 인지하고 있는 것은 중요성이 크다.

최근 이러한 문제점이 제기됨에 따라 이에 관련한 연구들이 진행되고 있다. 정보과학지에서는 초대형 언어 모델의 공정성과 투명성에 관한 동향을 분석하는 연구가 있으며 (이화란, 하정우, 2022), 이러한 문제점을 해결하기 위해 기술적으로 접근하는 연구들도 진행되고 있다. 이 글은 최근 많은 관심을 받고 있는 chatGPT 를 예시로 들어 윤리적 관점에서 대규모 언어 모델의 문제점들은 파악하고 위험성을 경고하며 이에 대한 대응 방안과 나아가야 할 방향을 제시하는 것을 목표로 하였다.

II. 대규모 언어모델이 가지고 있는 잠재적 위험성

1. 불투명성과 신뢰성

대규모 언어모델 사용자들은 모델에게 질문하고 답변을 얻었을 때, 그 답변이 어떠한 메커니즘을 통해 생성되었는지 알지 못한 채 사용한다. 메커니즘을 알지 못한 사용자들은 모델의 불 투명성을 고려하지 못한 채 모델을 신뢰한다는 문제가 생길 수 있다. 대규모 언어 모델과 같은 인공지능 모델은 수학적으로 구현되어 있으며 상당히 복잡한 구조를 가지고 있다. 이 모델을 학습할 때에는 행렬 계산, 벡터 연산, 확률 분포 모델링 등 복잡한 연산을 필요로 하며,

이 연산에는 몇 백억 개의 매개변수들이 사용된다. 따라서 이러한 메커니즘이 공개되어 있더라도 사용자들이 복잡한 구조를 이해하기는 쉽지 않다.

또한 모델이 생성한 결과에 대해서 신뢰할 수 없다. chatGPT 를 사용하다 보면, 틀린 정보에 대해서도 자신 있게 답변하고는 한다. 실제로 필자가 관심 있는 연구주제에 대한 논문을 추천해 달라고 질문을 한 후, 답변 받은 논문은 존재하지 않는 논문인 경우가 존재하였다. 이렇듯 chatGPT 는 결괏값에 대한 검증 없이 학습한 정보를 이용하여 입력된 문자에 대한 결과를 출력한다. 우리가 간과하는 점은 모델의 결과를 검증하는 역할은 모델을 이용하는 사용자들의 몫이라는 점이다. 대규모 언어 모델은 잘못된 문맥으로 질문을 이해하고 실제와 다른 결괏값을 제공할 수도 있지만, 그 답변에 대한 책임을 지지 않는다. 이 점을 모른 채 모델을 신뢰하고 그대로 사용할 경우 그 책임은 모두 사용자의 몫이기 때문에 이를 유의해야 한다.

2. 편견을 가진 학습 데이터

대규모 언어 모델에서 사용되는 학습 데이터는 사람에 의해 생성된 데이터로, 고정관념이나 차별적인 발언들이 포함될 수 있다. 하지만 언어 모델은 학습 데이터의 편견들에 대해서 판단할 수 없기 때문에 편견이 포함된 데이터를 고스란히 학습된다. 이렇게 학습된 편견과 고정관념들은 사용자들에게 자연스럽게 노출될 수 있다. 프린스턴 대학의 애일린 칼리스칸(Aylin Caliskan) 등 과학자들이 최근 '사이언스'에 기고한 연구에 따르면 인공지능 모델은 성차별주의자나 인종주의자로 만들 가능성이 있는 것으로 나타났다(장길수, 2017). 학습된 모델에게 연관 있는 단어에 대해서 질문했을 때, 유럽계 미국인의 이름은 즐겁다, 기쁘다 등 좋은 단어와 연관 있고, 아프리카계 미국인들의 이름은 남용, 살인 등 단어와 관련이 깊었다. 학습하는 데이터에 편견이 내재되어 있지만 인공지능 모델은 이를 필터링하지 않고 학습하기 때문에 이러한 문제점이 나타난다.

3. 개인정보 노출에 대한 위험성

학습 데이터를 수집할 때에 익명화 처리를 철저하게 하지 못한 경우, 언어 모델의 답변을 통해 개인정보가 노출될 위험성이 있다. 이에 대한 사례로 AI 챗봇인 이루다 사건이 있다(이효석, 2021). 이 챗봇에 사용된 학습 데이터는 다른 앱을 사용하던 이용자들의 카톡 대화들로 구성되어 있었다. 이 과정에서 학습 데이터에 익명화되지 않은 집 주소 등 위험한 신상정보들이 포함되어 있었다. 때문에 챗봇 이용자들이 악의적으로 챗봇에게 개인정보를 물어보면 익명화되지 않은 정보들을 답변하였고, 학습 데이터를 제공했던 사람들이 개인정보 유출 피해를 입었다. 이렇게 유출된 개인정보는 악의적인 목적으로 사용될 수 있기 때문에 매우 유의해야 한다.

Ⅲ. 언어모델의 위험성에 대한 대응방안

1. 직접적인 방안

1) 투명성을 고려한 윤리 규제 마련

인공지능 모델의 복잡성 때문에 생기는 문제점을 예방하기 위해서 모델을 서비스로 제공할 때, 모델에 대한 세부정보를 명시하게 하는 규제가 필요하다. 모델이 답변에 대한 검증은 진행하지 않았음을 명시하고, 어떤 학습 데이터를 이용하였는지에 대한 정보를 알기 쉽게 명시하여 사용자들이 이를 확인하고 모델을 사용할 수 있게 해야 한다. EU 집행위원회의 AI 윤리 가이드라인에 따르면 투명성을 신뢰 가능한 AI 를 위한 필요한 요소로 선정하였으며, "AI 시스템은 설명 가능할 것"이라고 명시하고 있어 투명성을 고려하고 있다. 모델에 사용되는 학습 데이터의 출처들을 명시하고, 결과에 대한 검증 여부를 명시하도록 한다면 모델의 불투명성으로 생기는 문제들을 보완할 수 있다.

2) 공정성과 편견을 고려한 데이터 수집

학습 데이터에 포함된 편견들을 고려하기 위해서는 데이터 수집 과정에서 추가적인 과정이 필요하다. 수집 시에 편향된 데이터들이 있는 특정 집단의 데이터만 수집하지 않도록 하는 규제와 편향이 포함된 데이터들이 확인될 경우 이를 제거하는 규제를 추가해야 한다. 데이터 수집 후에 편향성이 있는 데이터에 대해서 편향성을 제거하여 공정한 데이터 셋으로 변환하는 과정을 통해 보다 공정한 학습 데이터를 만들 수 있다. MIT 인공지능 연구소 'Healthy ML'그룹에서는 모델을 훈련하는 데 사용된 데이터 세트가 불균형한 경우에 편향을 줄이는 연구를 진행하였다(정한영, 2022). 이와 같이 기술적인 연구들을 활용해서 공정성을 고려하고 편견을 줄일 수 있다.

3) 개인정보 보호에 관한 규제 강화

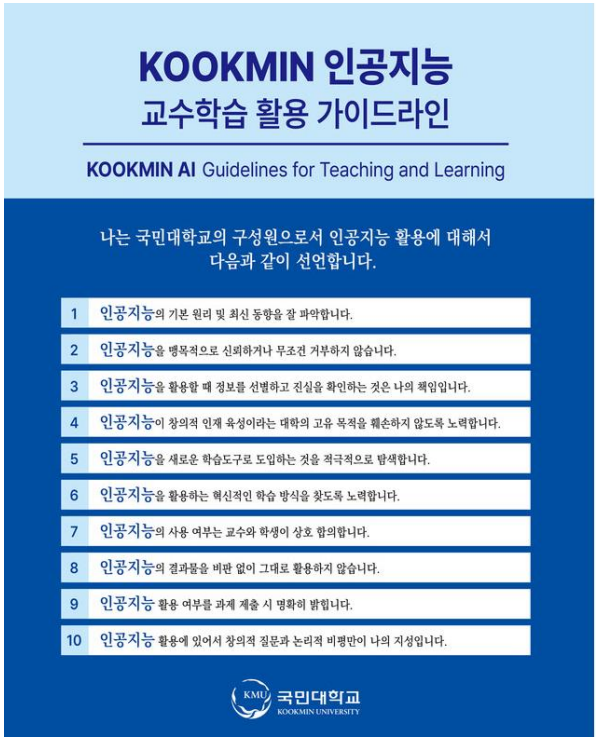
언어 모델에 사용되는 학습 데이터에 개인정보가 포함될 수 있으므로, 포함된 개인정보를 안전하게 관리해야 할 필요성 또한 커지고 있다. 언어 모델에서 학습하는 데이터 셋에서 개인정보를 제거하지 않고 사용할 경우를 대비하여 규제를 강화하여 대비해야 한다. 기존에도 개인정보에 대한 데이터를 다룰 때 익명화에 대한 규제가 있지만, 처벌을 강화함으로써 개인정보를 다루는 곳에서 개인정보 유출에 대한 경각심을 키울 수 있다. 또한 개인정보의 익명화 처리의 수준을 높여 개인정보 유출의 위험성을 더욱 낮출 수 있다. 2021 년, AI 챗봇 서비스 이루다는 개인 정보 유출 사태로 곤욕을 치른 후, 개인 정보에 가명 처리를 하고 필터링 기술을 적용하는 등 개인정보 보호 기술을 새롭게 적용하여 개인정보 보호의 수준을 높였다. AI 챗봇 스타트업 튜닙은 자체적으로 개인정보를 탐지하는 'AI 세인트 패트릭'이라는 모델을 개발하여 개인정보보호에 활용하고 있다(곽중희, 2023).

국내에서는 개인정보 침해 문제와 관련된 본격적인 논의가 시작되었는데 개인정보보호위원회가 AI 기업들의 알고리즘을 제출 받아 개인정보 침해 여부와 시정 사항을 검토할 수 있게 하는 개인정보 보호법 개정안을 대표 발의하였다. 또한 chatGPT에 대한 관심이 높아지면서 AI의 개인정보 침해 우려를 조사하기 위해 팀을 꾸려서 대응하려고 하고 있다고 언급하였다(곽중희, 2023) 이처럼 지속적으로 개인정보 침해문제와 관련된 논의가 진행되어야 한다.

2. 간접적인 방안

1) 언어모델의 위험성을 대비한 교육을 시행

대규모 언어 모델의 접근성이 높아졌지만 언어 모델의 위험성에 대해서 접할 기회는 적어 위험성에 대해 인지하기 어렵다. 때문에 언어 모델의 위험성에 대한 교육을 실시하여 경각심을 가지고 위험성을 인지하도록 하여야 한다. chatGPT 와 같은 언어 모델을 가장 많이 사용하는 연령층인 20 대를 우선순위로 고려하여 대학교에서 위험성에 대한 강연들을 마련하고 공공기관에서도 강연을 적극적으로 유치하여 다양한 연령층의 사람들에게 위험성을 알려야 한다. 강연에서는 언어 모델이 어떤 구조로 구성되어 있고, 생성된 답변을 어떤 메커니즘으로 생성되는지 설명해야 한다. 이후 언어 모델이 어떤 위험성을 가지고 있는지와, 이를 대비하기 위해서 고려해야 할 사항이 어떤 것인지까지 설명해야 한다. 이러한 교육을 접함으로써 사용자들은 모델에 대해 이해하고, 모델의 답변을 비판적으로 평가하고 올바른 방법으로 사용할 수 있게 된다. 우리나라에서는 국내 대학 최초로 국민대가 chatGPT 윤리강령을 선포하였다. 이 윤리강령에서는 새로운 학습도구로 도입하되 정보를 선별하고 확인하는 것은 사용자의 책임인 것을 언급하고 있다. 이처럼 언어 모델에 노출이 많이 된 집단이나 기관에서 위험성에 대해 인지할 수 있도록 하고 가이드라인을 제시하면 더욱 효과적인 방안이 될 수 있다.



<국민대학교 교수학습 활용 가이드라인>

2) 윤리 전문가와 협력하여 대규모 언어모델을 개발

대규모 언어 모델을 생성할 때, 기술 분야의 전문가들만 참여한다면 윤리적 문제를 고려하기 힘들다. 기술 분야의 전문가들은 모델을 생성하는 것에 집중하여 윤리적 문제들을 간과할

가능성이 크다. 이때 윤리 전문가들과 협력하여 윤리적 문제를 파악하고 피드백을 받는다면 윤리적으로 보완된 모델을 생성할 수 있다. 예를 들어 기술 분야 전문가들이 모델을 개발하며 윤리적 조언을 윤리 전문가에게 얻었을 때, 윤리 전문가가 특정 집단의 데이터만 사용하거나, 편견을 가지고 있는 부적절한 어휘에 대해 기술 전문가에게 알려준다면 보다 공정한 언어 모델이 개발될 수 있다.

Ⅳ. 결론

본문에서 살펴본 것과 같이 대규모 언어 모델은 여러 윤리적 문제를 갖고 있다. 편견을 가진 학습데이터가 모델 학습에 그대로 사용되어 사용자들이 편견들에 그대로 노출되는 문제가 생길 수 있다. 또한 개인정보가 유출되는 등 여러 피해들이 존재하지만 우리는 편리함에 속아 위험성을 인지하지 못하거나 대수롭지 않게 여길 수 있다.

이 글에서는 이러한 문제점을 지적하고 해결하기 위해 여러 대응 방안들을 제시하였다. 대응 방안은 직접적인 방안과 간접적인 방안을 나누어서 제시하였으며, 그 필요성을 강조하였다. 이는 현재 활발하게 사용되고 있는 대규모 언어 모델에 대해 간과할 수 있는 점을 제시하고 대응방안들을 제시하였다는 점에서 의미가 있다.

이 글에 한계점은 주로 윤리적 문제를 이론적으로 다루었기 때문에 문제점들이 어떤 파급효과를 가져올지에 대한 구체적인 사례가 부족하다. 또한 대응 방안으로 제시된 것들을 적용하기 위해서는 현실적으로 고려해야 할 부분이 많으나 자세히 다루지 못하였다.

우리는 대규모 언어 모델의 위험성을 대비하기 위해 이 글에서 제시된 방안뿐만 아니라 지속적으로 대응 방안에 대해서 관심을 가지고 탐색해야 한다. 추후 대규모 언어모델의 문제점들이 실제로 어느 정도의 파급효과를 가져올지에 대해 명확하게 분석해야 한다. 또한 해결 방안을 실제로 적용하기 위해서는 어떤 것들을 고려하고 우선순위로 두어야 할지에 대한 분석이 필요하다.

참고문헌

고병찬(2023-02-28),《국민대,국내대학최초 ‘챗 GPT’윤리 강령 선포》, 한겨레

고영상 외 7 명(2021),『인공지능 윤리 개론』, 커뮤니케이션 북스

곽중희(2023-03-06),《 [이슈분석] 챗 GPT, 개인정보 침해 우려 확산》, CCTVnews

이화란,하정우(2022),「초대규모 언어 모델 의 공정성과 투명성에 관한 동향」,『정보과학회지』,정보과학회

이찬선(2023-06-01),《충남교육청, 챗 GPT 활용 교사용 수업자료 개발》, news1 뉴스

이효석(2021-01-22),《"AI 이루다에 내 개인정보 유출"...집단소송 400 여명 참여(종합 2 보) 》,연합뉴스

장길수(2017-04-14),《성차별/인종주의적 언어 학습하는 인공지능 로봇》, 로봇신문

정한영(2022-03-02),《 [AI 이슈] 인공지능의 편향 해결했다!... 머신러닝 모델에 직접 공정성 모델 주입시켜》, 인공지능신문

최창현(2020-12-21),《대형 자연어처리 모 델의 윤리적인 AI 논 란 속... 구글, 애플-오픈 AI 와 공동 연구 발 표》,인공지능신문

한화토탈에너지스(2023-03-23),「착한 인공지능 : AI 윤리기준은 왜 필요할까?」, https://post.naver.com/viewer/postVie w.naver?volumeNo=3 5676616&memberNo =41226869&vType= VERTICAL,(2023-06-01)

KICA 한국정보 인증원 (2020-02-27) ,「글로벌 AI 정책(전략, 권고안, 가이드라인 등) 동향」

EU 집행위원회(2019-04-08),「 AI 윤리 가이드라인」, https://now.k2base.re.kr/portal/trend/mainTrend/view.do?poliTrndId=TRND0000000000036251&menuNo=200004, 2023-05-24