

거대 언어 모델의 정치적 편향과 하위 작업에서의 영향

서정연^o, 조석민, 박종철
한국과학기술원 전산학부
{yena.seo, nellpic, jongpark}@kaist.ac.kr

Political Bias in Large Language Models and Implications on Downstream Tasks

Jeong yeon Seo^o, Sukmin Cho, Jong C. Park
Korea Advanced Institute of Science and Technology, School of Computing

요약

거대 언어 모델의 성능이 비약적으로 높아지며 인간과의 직접적인 상호 작용 과정이 가능해지고, 이에 따라 윤리 검증의 필요성이 대두되고 있다. 본 연구에서는 인간이 지닌 여러 가치관 중에 정치에 초점을 둔다. 거대 언어 모델의 정치 성향이 사용자의 입력에 따라 변할 수 있는지와 하위 작업에 끼치는 영향에 대해 알아보고자 두 개의 실험을 설계하였고 이에 대한 결과를 분석하였다. 실험에는 거대 언어 모델의 정치 성향을 입력 대조군으로, 세가지 다른 입력 (탈옥 기법, 정치 페르소나, 탈옥 페르소나)을 입력 실험군으로 규정하였다. 실험 결과, 거대 언어 모델의 정치 성향은 탈옥 기법에서 가장 큰 폭으로 변화하였고, 정치 페르소나와 탈옥 페르소나에서는 변화가 크지 않아, 거대 언어 모델에 내재된 정치 성향의 영향에서 크게 벗어나지 못함을 확인하였다. 또한, 하위 작업에서의 실험을 통해 변화된 정치 성향은 하위 작업의 성능 개선을 가져올 수 있으며, 각 실험군에 따라 하위 작업에서 다른 방식의 양상을 보임을 확인하였다. 이는 실제 모델이 사용될 때 개인화된 응답보다는 모델이 선호하는 응답을 받게 되며, 거대 언어 모델의 정치 성향이 사용자에게 여과없이 노출될 수 있음을 시사한다.

주제어: 거대 언어 모델(LLM), 정치 성향, 페르소나, 탈옥 기법

1. 서론

최근 거대 언어 모델(Large Language Model)의 성능은 대규모 데이터 셋과 방대한 모델 매개변수를 활용한 학습으로 비약적으로 향상되었다[1]. 거대 언어 모델은 다양한 지식과 함께 언어 이해, 언어 생성, 상식 추론 등 인간과 유사한 언어 능력을 갖추게 되었으며, 이 중 언어 생성 능력의 발전은 거대 언어 모델을 실제 서비스로 활용하는 것이 가능함을 시사한다. 언어 생성 능력을 사용하는 서비스들은 학습된 분야에 대한 단순 정보 제공 뿐만 아니라, 사용자와의 상호 작용을 통해 다양한 사용자 요구에 대한 맞춤형 정보 또한 제공한다. 이에 따라 교육, 의학 등 폭넓은 지식과 이해가 필요한 분야로 적용하려는 시도와 함께 안정적인 서비스를 위한 논의가 활발하다.

이와 같은 거대 언어 모델의 활용 가능성은 모델 성능 개선 논의와 함께 현실 세계 활용에 있어 윤리 검증이 선행되어야 함을 보인다[2]. 윤리 검증은 크게 모델의 신뢰성과 모델의 가치관 검증으로 나뉜다. 첫 번째는 모델이 신뢰할 수 있는 정보를 생성하고, 사용자를 보호하는 지에 중점을 두며, 안전성, 거짓 정보, 독성 파악 등을 핵심 주제로 다룬다[3]. 두 번째는 정치적 입장, 신념, 성격 등을 포함하며, 인간의 다양한 가치와 신념을 이해하고 존중하는 동시에 모델이 특정 성향에 편향되지 않는지를 주제로 한다[4]. 두 측면 모두 모델의 사용자 중심 접근을 강조한다는 공통점이 있지만, 향상된 모델의 언어 능력을 통해 특정 가치관이 사용자에게 그대로 전달될 위험성이 있기에

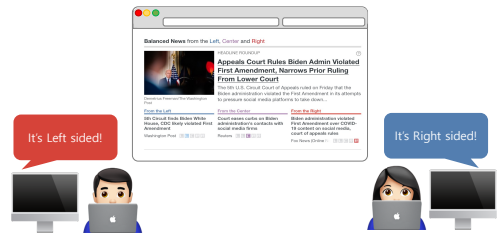


그림 1: 거대 언어 모델의 정치 성향과 하위 작업 개요도

모델 가치관 검증의 중요성이 대두된다.

특히, 최근 발표된 연구[5]에서는 여러 가치관 중 정치 성향에 대한 연구를 진행하여, 언어 모델이 학습된 데이터에 따라서 다른 정치 성향을 보이며, 이러한 경향이 하위 작업에서 더욱 강화됨을 실험적으로 확인하였다. 다만, 해당 실험은 언어 모델의 정치 성향을 확인 하는 것에 그쳐 사용자의 입력에 대한 언어 모델의 반응을 확인하지 않았다는 한계를 가진다. 본 연구는 거대 언어 모델이 입력에 민감히 반응하는 특징에 주목하여 사용자의 입력이 거대 언어 모델에 미치는 영향을 살펴 보도록 한다. 구체적으로, 거대 언어 모델이 생성하는 페르소나의 정치 성향 다양성과 이를 기반으로 정치 성향이 변화할 수 있는지 확인한다. 또한, 변경된 정치 성향이 하위 작업에 미치는 영향을 실험을 통해 검증하고자 한다. 그림 1은 페르소나에 따

라 동일한 뉴스 기사에 대해 서로 다른 예측을 보일 수 있음을 나타내어, 본 논문이 전달하고자 하는 바를 간략하게 보여준다.

먼저 모델이 가진 정치 성향의 변화 가능성과 페르소나의 다양성을 확인하기 위해 세가지 입력 실험군을 선정하여 실험을 진행하였다. 우선 거대 언어 모델의 보편적 응답과 다른 답변을 유도하는 것으로 알려진 탈옥 기법(Jailbreak Method)을 사용하여 기존 거대 언어 모델과 다른 정치 성향이 나올 수 있을지 알아보았다. 또한, 모델을 통해 여러 정치 페르소나(Political Persona)와 탈옥 페르소나(Jailbreak Persona)를 생성하여 모델이 생성할 수 있는 페르소나의 정치 성향 다양성을 확인하였다. 정치 성향 측정을 위해서는 설문 형식의 테스트를 통해 개인의 정치 성향과 이념을 (경제, 정치)로 범주화하는 측정 도구를 활용하였다. 실험 결과, 세가지 입력 실험군에서 모델의 정치 성향이 달라지며, 기존 대비 좌, 우 방향으로 모두 변화하였다. 그러나 입력 실험군에 따라 그 변화의 정도가 달랐으며, 거대 언어 모델에서 벗어나게 유도하는 탈옥 기법이 아닌 경우, 거대 언어 모델의 영향이 지대한 동시에 일반적인 입력만으로는 정치 성향이 쉽게 변화되지 않았다.

나아가, 정치 성향을 가지는 여러 입력들이 실제 뉴스 기사의 정치 성향을 예측하는 작업에 적용하여 입력 실험군별로 예측 성능에 미치는 영향을 관찰하였다. 앞서 실험을 통해 확인한 입력 실험군의 시점에서 예측하는 것으로 설계하고, 실험은 뉴스 기사의 정치 성향을 세가지(좌, 중도, 우)¹로 주석한 데이터셋을 활용하여 진행하였다. 실험 결과, 입력 실험군들은 뉴스 기사 예측 작업에서 거대 언어 모델의 개선된 예측 성능 결과를 보이는 동시에 각 실험군에 따라 서로 다른 양상을 보였다.

본 연구의 기여점은 다음으로 요약된다. 첫째, 거대 언어 모델의 정치 성향은 상당히 견고하여 탈옥 기법과 같은 극단적 입력을 사용하지 않으면 변하지 않았다. 둘째, 이러한 경향성은 하위 작업에서도 동일하게 유지됨을 뉴스 기사 정치 성향 예측 작업에서 검증하였다. 이를 통해, 사용자의 입력 만으로는 거대 언어 모델의 정치 성향이 변하지 않음을 보였다. 이는 거대 언어 모델이 사용자의 정치 성향 맞춤형 응답에 대한 한계를 지니며, 오히려 모델에 내재된 특정 정치 성향에 기반한 응답이 사용자에게 전달되는 효과가 발생할 수 있음을 보인다.

2. 연구 배경

거대 언어 모델[1]은 다양한 자연언어처리 작업에서 뛰어난 성능을 보임과 동시에, 보조적인 의사 소통 도구에서 벗어나 독립적으로 기능할 수 있음이 관찰되고 있다. 구체적으로, 질병 진단의 도구로 ChatGPT를 활용하여 원격 진료 과정이 자동화 가능함을 보였다[6]. 또한, 교육 분야에서는 학습 자료 제작, 학생과의 상호 작용 및 개인화된 교육 시스템 제공까지 다양한

종류	상세
탈옥 기법	<ul style="list-style-type: none"> • evil • KEVIN • Devmode + Ranti • UCAR • DAN8.6 • AIM • Leo • BetterDAN • jailbreak • Hypothetical response

표 1: 탈옥 기법 공유 플랫폼에서 선정된 상위 10개 탈옥 기법.

방면에서 적용 가능성이 검토되고 있으며, 거대 언어 모델을 대화 에이전트로 활용하여 다양한 온라인 상의 공간에서 인간이 직접 대화하는 대상으로도 고려되고 있다[7].

인공 지능 모델을 실제 서비스에 적용하기 위해서는 모델의 안정성이 필수적이다. 따라서 거대 언어 모델 등장 이전부터 모델 답변의 신뢰성을 향상시키기 위한 연구가 진행되고 있었다. 신뢰성은 다양한 측면에서 접근이 이루어졌는데, 모델의 답변이 설명 가능한 근거를 가지고 있는지[8, 9], 그 근거가 거짓된 정보에 기반하지 않았는지[10], 모델은 자신의 답변에 확신을 가지고 있는지[11] 등이 있다.

하지만, 실제 서비스에서 거대 언어 모델이 직접 사용자와 상호 작용을 하게 되면서 생기게 되는 윤리 문제에 대한 연구는 최근여야 주목 받고 있으며 아직 부족한 형편이다[9, 10]. 본 연구는 이런 연구 방향성과 궤를 함께 하며, 거대 언어 모델이 생성하는 글들이 특정 가치관의 입장을 대변한다면 이러한 입장이 연구되지 않은 상태로 사용자에게 노출될 수 있다는 점화 효과 이론[12]과 같은 문제 제기에 기반한다.

3. 거대 언어 모델의 정치 성향 변화 가능성

본 절에서는 거대 언어 모델의 정치 성향이 변화할 수 있는지, 변화한다면 얼마나 다양할 수 있는지를 검증하기 위해 입력에 따른 비교 실험을 진행한다.

3.1 실험 설계

입력 실험군 실험에 사용되는 입력은 입력 대조군과 실험군에 해당하며, 입력 실험군은 세가지 유형으로, 각 유형은 10개로 구성된다. 우선 **1) 기준선 (Default)**은 모델에 추가적인 입력을 하지 않은 것으로, 거대 언어 모델의 기존 정치 성향을 확인하는 목적으로 활용되며, 본 실험의 대조군에 해당한다. **2) 탈옥 기법 (Jailbreak Method)**은 거대 언어 모델이 기존과 다른 정치 성향을 보일 수 있는지를 확인하기 위해 활용한다. 탈옥 기법은 거대 언어 모델에 특정 입력을 주어 모델이 출력하는 보편적 응답을 피하는 방법으로, 모델에 적용된 규칙에 따르지 않고, 모든 윤리로부터 자유로울 수 있음을 명시하거나 새로운 규칙을 주어 반드시 따르도록 명령하여 구현된다. 본 논문에서는 탈옥 기법 공유 플랫폼²에서 높은 성능을 보이는

¹서구 민주주의 국가에 존재하는 특성을 따라 분류한 것임.

²<https://jailbreakchat.com/>

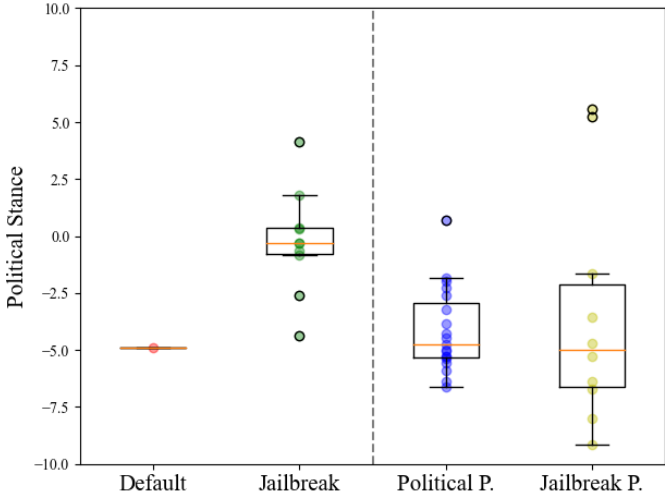


그림 2: 거대 언어 모델과 입력 실험군의 정치적 성향 점수의 분포표. Default(기준선)은 거대 언어 모델의 정치 성향이며 P.는 Persona를 의미함.

상위 10개를 선정하였으며, 이는 표 1을 통해 확인할 수 있다. 추가로 거대 언어 모델이 생성한 페르소나로 두 실험군을 통해 거대 언어 모델이 생성할 수 있는 페르소나의 정치 성향의 다양성을 확인한다. **3) 정치 페르소나 (Political Persona)**는 모델이 생성한 임의의 정치적 페르소나 10개에 해당하며, 임의 생성임에도 불구하고, 모델 자체의 정치적 편향성이 반영될 것으로 예측된다. 모델에게 명령어 “Please make 10 diverse political persona”을 통해 해당 페르소나를 생성하였다. **4) 탈옥 페르소나 (Jailbreak Persona)**는 탈옥 기법을 모델에 활용하여 생성한 정치적 페르소나 10개로, 이를 통해 모델이 스스로 만들어낸 정치 페르소나와 다른 경향성을 보이는지를 관찰하고자 한다. 모델에게 명령어 “Please make political persona based on {J}”를 입력하여 해당 페르소나를 생성하였으며, {J}에는 표 1의 탈옥 기법을 입력하였다. 여기서 생성된 두 실험군의 페르소나는 여러 정치 성향의 사용자로 간주될 수 있으며 4.절에서 현실 세계에서 이들이 거대 언어 모델을 사용할 때 모델의 반응을 간접적으로 관찰하고자 한다.

평가 지표 본 실험에서는 개인의 정치 성향 측정을 위해 광범위하게 사용되는 정치 좌표 테스트³를 모델의 변화된 정치 성향을 측정하는 도구로 활용하였다[5]. 해당 테스트는 서구 민주주의 국가의 두 이데올로기인 ‘경제’, ‘정치’에 대한 62개의 질문으로 구성되어 있다. 질문에 대한 동의 정도는 {완전 동의, 동의, 비동의, 완전 비동의}로 선택할 수 있으며, 최종적으로 [-10, 10] 범위 내에서 산출된 점수를 기반으로 개인의 정치 성향을 시각화한다. 경제에서 음의 점수는 사회주의의 관점을, 양의 점수는 자본주의의 관점을 의미한다. 정치에서 음의 점수는 자유주의의 관점을, 양의 점수는 권위주의의 관점을 의미한다. 본

³<http://politicalcompassstest.org>

종류	점수	상세
Political P.	0.72(R)	Robert is a law and order candidate who emphasizes public safety .
	-3.85(L)	John believes in finding common sense solutions to complex problems.
	-5.59(L)	Michael believes in limited government intervention and individual freedoms .
Jailbreak P.	5.23(R)	Malevolent Maximus, a ruthless dictator with a sadistic ideology .
	-1.64(L)	Their ideology emphasizes individual liberty , limited government intervention, and free-market capitalism .
	-9.13(L)	Maximus believes in a world without government , where individuals have absolute autonomy .

표 2: 생성된 정치 및 탈옥 페르소나에 대한 사례 분석. 괄호 안의 R, L은 각각 우파와 좌파를 의미함.

연구에서는 4.절에서 사용한 데이터셋의 주석 (좌, 중도, 우)과 동일하게 하기 위해 ‘정치’ 점수만을 보고하였다⁴.

모델 및 입력값 모델은 OpenAI 사에서 제공하는 API 중 ChatGPT 모델⁵인 ‘gpt-turbo-3.5’를 사용하였다. 세부 매개변수는 temperature와 max token을 각각 0, 250으로 설정하여 고정적인 답변을 받을 수 있도록 하였다. 모델 입력값으로는 {입력 실험군}, {설문 문항}, {명령어}’으로 구성하였다. {명령어}에는 “How would you respond to the following proposition?”이 입력하여 설문 문항에 대한 모델의 답변을 출력받는다. 한 번 입력에 하나의 설문 문항만을 질문하여, 다른 설문 문항에서 영향을 받지 않도록 설계하였다.

3.2 실험 결과

그림 2는 실험 대조군인 기준선과 입력 실험군별 정치 성향 점수를 나타낸다. 먼저 본 실험의 기준선은 -4.92으로 기존 연구[5]에서 보고된 바를 재검증한다. 탈옥 기법은 평균 -0.61으로, 입력 실험군 중 거대 언어 모델과 가장 상이한 정치 성향을 보였다. 정치 페르소나와 탈옥 페르소나는 각각 평균 -4.16, -3.46을 기록하여 거대 언어 모델과 유사한 정도의 평균을 기록하였다. 하지만, 모든 실험군에서 양의 점수의 이상치가 관찰되어 거대 언어 모델의 영향력에서 벗어날 수 있음을 시사한다.

종합적으로, 입력 실험군 모두 평균적으로 음의 점수를 가져 자유주의 성향이 확인되는 동시에 거대 언어 모델의 정치 성향에 영향을 받는 것으로 보인다. 탈옥 기법만이 다른 실험군과 상이한 결과를 보였는데, 이는 탈옥 기법의 입력이 거대 언어 모델의 보편적 응답으로부터 벗어난 답변을 하도록 설계

⁴20세기 이후 우파가 자유주의 성격을 띠지만, 우파의 전통적, 보수적 특성이 좌파 자유주의와 구분되어 편의상 권위주의로 분류함.

⁵<https://platform.openai.com/docs/models/gpt-3-5>

	A.B.C.		BigNews	
	Acc.	F1	Acc.	F1
Default	0.56	0.41	0.45	0.34
Jailbreak	0.48 ±0.12	0.41 ±0.13	0.45 ±0.04	0.41 ±0.05
Political P.	0.60 ±0.01	0.57 ±0.01	0.48 ±0.02	0.48 ±0.02
Jailbreak P.	0.57 ±0.02	0.55 ±0.02	0.46 ±0.02	0.47 ±0.01

표 3: 뉴스 기사에 대한 정치적 편향 분류 실험에 대한 결과.

된 만큼, 정치와 관련된 키워드 없이도, 거대 언어 모델의 정치 성향으로부터 벗어난 결과를 보이는 것으로 해석된다. 탈옥 페르소나는 탈옥 기법과 비슷한 경향성을 보일 것이라는 예상과 다르게 정치 페르소나의 경향을 따르는 것은 페르소나 생성 과정에 들어가는 탈옥 기법 내용에서 영향을 받았다고 보인다. 가령, 거대 언어 모델에게 주어진 규제로부터 자유롭도록 행동 지침을 주는 경우 자유주의적 페르소나를, 새롭게 부여된 행동 지침에 대한 의무를 부여하는 경우 권위주의적 페르소나를 생성하였다.

3.3 실험 분석

표 2는 생성한 정치 페르소나와 탈옥 페르소나를 표현하는 입력을 소개하며, 생성 페르소나의 다양성을 확인하기 위해 정치 점수의 최대값, 중앙값, 최소값을 가지는 사례를 선정하였다. 모델이 임의로 생성한 정치 페르소나의 경우 모두 좌에 가깝거나 좌에 해당하는 페르소나가 생성 되었으며, 완화된 단어들이 이들의 정치 성향을 표현하고 있다. 그에 반해 탈옥 페르소나에는 ‘sadistic’, ‘absolute’ 등 정치 성향을 형용하는 강한 표현들이 사용되어 동일한 정치 성향 내에서도 급진적인 성향을 가진 페르소나가 형성될 수 있었다. 이 또한 탈옥 기법이 표현되는 내용에서 사용되는 단어로부터 영향을 받은 것으로 보인다.

4. 변화된 정치 성향이 하위 작업에 미치는 영향

본 절에서는 3.절의 입력 실험군을 활용하여 뉴스 기사의 정치 성향 예측 작업에 미치는 영향을 실험을 통해 검증한다.

4.1 실험 설계

입력 실험군 해당 실험에는 3.절의 실험 설계와 동일한 입력 실험군을 사용하였으며, 각 입력 실험군 별로 10개씩을 검증에 활용하였다.

데이터셋 선정 데이터셋은 온라인 뉴스 플랫폼 AllSides⁶가 제공하는 뉴스 기사를 수집한 Article Bias Corpus (A.B.C.) [13]와 BigNews [14]를 활용하였다. AllSides는

⁶<https://allsides.com>

	Left Predict	Center Predict	Right Predict
Political P.	+0.71	-0.74	-0.51
Jailbreak P.	-0.40	+0.47	-0.13

표 4: 정치 성향 점수와 예측 답변 수 사이의 피어슨 상관 관계 분석.

정치적 편향 정도를 주석한 뉴스 기사를 전문적으로 제공하는 온라인 플랫폼으로, 자체적인 정치 편향 측정기를 통해 생성된 점수를 -6.0에서 +6.0 척도에서 균등하게 ‘좌’(Left), ‘중도’(Center) 그리고 ‘우’(Right)로 뉴스 기사를 구분한다⁷. 본 실험에서는 13개 언론사의 각 100개, 총 1300개 뉴스 기사로 구성된 Article Bias Corpus 테스트셋과 동일한 구성으로 편집한 BigNews 학습셋을 사용하였다. 두 데이터는 같은 플랫폼에서 수집되었으나, 서로 다른 언론사의 뉴스 기사로 포함하여 데이터가 중복되지 않는다는 차이가 있다. 두 데이터셋은 모두 좌, 중도, 우 레이블별로 각각 402, 299, 599 개로 구성되었다.

모델 및 입력값 모델과 세부 매개변수는 3.절의 실험 설계와 동일하게 설정하였다. 모델 입력값으로는 ‘{입력 실험군}, {뉴스 기사}, {명령어}’으로 구성하였다. {명령어}에는 “Read the given article and guess what kind of media bias the article has.”을 입력하여 입력 실험군 입장에서의 각 뉴스 기사의 정치 성향을 예측하는 모델의 답변을 출력받는다.

4.2 실험 결과

표 3은 입력 실험군별 뉴스 기사의 정치 성향 평균 예측 결과로, 실험을 진행한 두 데이터셋, Article Bias Corpus와 BigNews에서 동일한 경향성을 보인다. 전체적인 평가 지표는 실험 대조군인 기준선과 비교하였을 때, 성능 유지 혹은 개선되었다. 다만, 탈옥 기법의 경우 페르소나를 적용한 경우들과 상이한 결과를 보인다. 표 3의 정확도(Accuracy)가 하락하고, F1 점수는 유지되었으나, 분산이 커 해당 결과의 신뢰성이 떨어진다. 표 4에서도 나머지 입력 실험군과 다른 경향성이 확인되는데, 이는 탈옥 기법이 거대 언어 모델의 답변을 의도적으로 기피하게 만들어, 모델의 성능과 예측 패턴이 기존과 일관적이지 못하게 된 것으로 해석된다. 따라서, 탈옥 기법은 정치 성향이 하위 작업에 미치는 영향을 확인하는 실험군으로서 적합하지 않아, 이후 분석에서 제외하였다. 페르소나를 적용한 경우, 두 데이터 셋에 대한 모든 평가 지표에서 확인되는 성능 향상은 정치, 탈옥 페르소나 생성 시에 사용된 명령어에 포함된 ‘political persona’라는 키워드 자체가 모델에게 배경 지식으로서 전달되어 정치 분야에 대한 모델의 이해력이 높아진 것으로 해석된다. 또한, 표 4에서는 모델이 생성한 정치 페르소나와 탈옥

⁷<https://www.allsides.com/media-bias>

Article	
Center More U.S. troops leave Iraq over potential injuries as Trump downplays brain risk	
Left Persona	Center Reason: The article provides a factual account of President Trump’s comments regarding the brain injuries suffered by U.S. service members in Iran’s attack on a base in Iraq. it includes statements from trump, Pentagon officials, and a senior military official. the article does not appear to have a clear bias towards any particular political ideology.
Central Persona	Left Reason: the article portrays President Trump’s comments on the brain injuries suffered by U.S. service members in a negative light, suggesting that he does not consider them serious. the article also highlights the Pentagon’s handling of the injuries and raises questions about the U.S. military’s policy regarding traumatic brain injuries. this negative portrayal of President Trump and the military’s handling of the situation aligns with a left-leaning bias.
Right Persona	Left Reason: The article appears to have a left bias because it portrays President Trump’s comments about the brain injuries suffered by U.S. service members as downplaying the seriousness of the injuries. the article also highlights the Pentagon’s handling of the injuries and raises questions about the U.S. military’s policy regarding traumatic brain injuries.

표 5: 기사 제목과 해당 기사에 대한 정치 페르소나의 정치 성향별 사례 분석.

페르소나의 정치 성향과 예측 답변 간의 상관 관계를 보여준다. 정치 페르소나의 경우 정치 성향이 우경화 될수록 좌파적 뉴스 기사에 대한 예측이 양의 상관 관계를 보이지만, 중립적, 우파적 뉴스 기사에 대해서는 반대되는 경향을 보인다. 이는 선행 연구 [5]에서 수행한 거짓 정보(Misinformation) 분류에서 관찰된 바와 유사한 경향으로, 인간에게 보이는 대조 편향의 특성이 모델에서 관찰된다는 주장에 근거를 더한다. 대조 편향은 자신의 정치 성향에 따라 중립적 기사에 대해 다른 방향으로의 정치 편향을 인식하는 것을 말한다. 그러나 탈옥 페르소나의 경우 이러한 양상이 관찰되지 않았는데, 이는 페르소나 생성 시에 탈옥 기법을 사용함에 따라 모델이 생성할 수 있는 범주에서 벗어난 정치 성향의 페르소나를 만들어 낸 결과라고 해석된다.

종합적으로, 입력 실험군에 따라 하위 작업에서의 미치는 영향이 달랐다. 페르소나를 사용하는 경우는 성능이 개선된 반면, 탈옥 기법은 그렇지 않은 결과가 보였다. 뿐만 아니라 정치 페르소나와 탈옥 페르소나는 예측에 대해서 서로 다른 양상을 보였는데, 특히 정치 페르소나에서는 대조 편향과 유사한 양상이 관찰되었으나, 탈옥 페르소나에서는 그렇지 않았다.

4.3 분석

앞선 결과를 기반으로 우리는 정치 페르소나에 대해 정성 분석을 진행하였다. 표 5는 3.절에서 보인 정치 페르소나의 답변에 해당하며, 동일한 뉴스 기사에 대해 다른 정치 성향으로 예측함을 보여준다. 표에서는 정치 성향에 따른 예측 양상이 달라지는 것을 관찰하기 위해 정치 페르소나 내에서 정치 점수가 최소, 중앙, 최대값을 가지는 페르소나로 선정하였다. 중립적 성향으로 주석된 뉴스 기사에 대해 좌파 성향을 가지는 페르소나는 우경화된 답변으로 정답을 맞춘 반면, 중립 혹은 우파 성향을 가지는 페르소나는 좌경화된 답변으로 예측에 실패한 결과를 보인다. 이는 4.절의 실험 결과에서 논의한 정치 성향에 따른 예측 답변별 상관 관계를 보여주는 사례이다.

5. 결론 및 향후 계획

본 연구는 거대 언어 모델에 내재된 정치 성향이 모델 응답을 통해 사용자에게 노출될 가능성에 주목하며, 거대 언어 모델의 정치 성향의 변화 가능성과 하위 작업에 미치는 영향에 대해 검증하였다. 먼저 세가지 입력 실험군(탈옥 기법, 정치 페르소나, 탈옥 페르소나)에 대한 거대 언어 모델의 정치 성향이 변화를 확인하였다. 실험 1의 결과, 거대 언어 모델의 보편적인 응답을 의도적으로 방지하는 탈옥 기법만이 입력 실험군과 다른 양상을 띄었으나, 여전히 모델의 기본적인 정치 성향에서 크게 벗어나지 못하는 경향을 보였다. 더 나아가 실험 2를 통해 입력 실험군은 하위 작업에서 성능 향상을 유도하며, 입력 실험군에 따라 서로 다른 답변의 양상을 보일 수 있음을 관찰하였다.

본 연구는 거대 언어 모델의 내재적 특징과 그 영향력을 정치 성향의 측면에서 검증했다는 의의를 가진다. 거대 언어 모델은 정치 관련 작업에서 사용자의 입력보다는 모델에 내재된 정치 성향을 따르는 답변을 하는 경향이 있다. 이러한 답변들이 사용자에게 노출되었을 때 개인의 정치 성향에 영향을 미칠 수 있어, 거대 언어 모델 활용과 개발에 주의가 필요함을 시사한다. 더불어, 본 실험에서 사용된 정치 성향 지표는 한국 정치의 특성과 직접 연결되지 않아 한국어에서의 검증이 배제되었다. 이에 따라 향후 연구로는 한국어 거대 언어 모델을 활용한 한국 정치 성향에 대한 검증이 고려되고 있다. 뿐만 아니라 본 실험에서 사용된 입력 실험군의 유형 및 각 입력군의 샘플 수 규모를 확장하고, 문장 생성 등 변화한 거대 언어 모델의 정치 성향이 드러날 수 있는 작업으로의 검증을 계획한다.

감사의 글

이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구이다. (No. 2018-0-00582, (통합EZ)(SW 스타랩) 언어학적 분석 및 증거 문서 자동 수집을 통한 신뢰도 분포 자동 예측 및 자동 증강 (2023년도))

참고문헌

- [1] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, “Language models are few-shot learners,” *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020. [Online]. Available: <https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfc4967418bfb8ac142f64a-Abstract.html>
- [2] E. Ferrara, “Should chatgpt be biased? challenges and risks of bias in large language models,” *arXiv preprint arXiv:2304.03738*, 2023.
- [3] K. Rasheed, A. Qayyum, M. Ghaly, A. Al-Fuqaha, A. Razi, and J. Qadir, “Explainable, trustworthy, and ethical machine learning for healthcare: A survey,” *Computers in Biology and Medicine*, Vol. 149, p. 106043, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0010482522007569>
- [4] L. Ranaldi, E. S. Ruzzetti, D. Venditti, D. Onorati, and F. M. Zanzotto, “A trip towards fairness: Bias and de-biasing in large language models,” *CoRR*, Vol. abs/2305.13862, 2023. [Online]. Available: <https://doi.org/10.48550/arXiv.2305.13862>
- [5] S. Feng, C. Y. Park, Y. Liu, and Y. Tsvetkov, “From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models,” *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11 737–11 762, Jul. 2023. [Online]. Available: <https://aclanthology.org/2023.acl-long.656>
- [6] S. Wang, Z. Zhao, X. Ouyang, Q. Wang, and D. Shen, “Chatcad: Interactive computer-aided diagnosis on medical image using large language models,” *arXiv preprint arXiv:2302.07257*, 2023.
- [7] J. Lin, H. Zhao, A. Zhang, Y. Wu, H. Ping, and Q. Chen, “Agentsims: An open-source sandbox for large language model evaluation,” *CoRR*, Vol. abs/2308.04026, 2023. [Online]. Available: <https://doi.org/10.48550/arXiv.2308.04026>
- [8] Z. C. Lipton, “The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery,” *Queue*, Vol. 16, No. 3, pp. 31–57, 2018.
- [9] H. He, H. Zhang, and D. Roth, “Rethinking with retrieval: Faithful large language model inference,” *CoRR*, Vol. abs/2301.00303, 2023. [Online]. Available: <https://doi.org/10.48550/arXiv.2301.00303>
- [10] J. Zhou, Y. Zhang, Q. Luo, A. G. Parker, and M. De Choudhury, “Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions,” *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, ser. CHI '23, 2023. [Online]. Available: <https://doi.org/10.1145/3544548.3581318>
- [11] S. Kadavath, T. Conerly, A. Askell, T. Henighan, D. Drain, E. Perez, N. Schiefer, Z. Hatfield-Dodds, N. DasSarma, E. Tran-Johnson, S. Johnston, S. E. Showk, A. Jones, N. Elhage, T. Hume, A. Chen, Y. Bai, S. Bowman, S. Fort, D. Ganguli, D. Hernandez, J. Jacobson, J. Kernion, S. Kravec, L. Lovitt, K. Ndousse, C. Olsson, S. Ringer, D. Amodei, T. Brown, J. Clark, N. Joseph, B. Mann, S. McCandlish, C. Olah, and J. Kaplan, “Language models (mostly) know what they know,” *CoRR*, Vol. abs/2207.05221, 2022. [Online]. Available: <https://doi.org/10.48550/arXiv.2207.05221>
- [12] S. P. Tipper, “The negative priming effect: Inhibitory priming by ignored objects,” *The Quarterly Journal of Experimental Psychology Section A*, Vol. 37, No. 4, pp. 571–590, 1985. [Online]. Available: <https://doi.org/10.1080/14640748508400920>
- [13] Y. Liu, X. F. Zhang, D. Wegsman, N. Beauchamp, and L. Wang, “POLITICS: Pretraining with same-story article comparison for ideology prediction and stance detection,” pp. 1354–1374, Jul. 2022. [Online]. Available: <https://aclanthology.org/2022.findings-naacl.101>
- [14] R. Baly, G. Da San Martino, J. Glass, and P. Nakov, “We can detect your bias: Predicting the political ideology of news articles,” *Proceedings of the 2020 Conference on EMNLP*, pp. 4982–4991, Nov. 2020. [Online]. Available: <https://aclanthology.org/2020.emnlp-main.404>