

래더링 인터뷰 기법을 활용한 대규모 언어 모델(LLM)의 정치적 성향 평가 및 분석 프레임워크 연구

장교철, 최재훈, 신민정, 이정, 서봉원
서울대학교

kyochul@snu.ac.kr, hoon95@snu.ac.kr, shinmj1024@snu.ac.kr, leejungp2@snu.ac.kr,
bongwon@snu.ac.kr*

A Framework for Analyzing Political Knowledge Injection of LLMs(Large Language Models) Using the Laddering Interview Technique

Jang Kyochul, Shin Minjeong, Lee Jung, Choi Jaehoon, Suh Bongwon
Seoul National University

요약

대규모 언어 모델(LLM: Large Language Model)은 사회 전반에 막대한 영향을 미치며, 특히 정치적 사안에 대한 편향된 답변은 여론에 큰 영향을 줄 수 있다. 기존의 단순한 질문 응답 방식으로는 모델이 내포한 복잡한 가치관과 정치적 성향을 심층적으로 평가하기 어렵기 때문에, 본 연구에서는 사다리 인터뷰(Laddering Interview) 기법을 도입하여 좌(진보)/우(보수)로 파인튜닝된 LLM의 정치적 성향을 체계적이고 정량적으로 분석하고 평가하는 프레임워크 개발을 목표로 한다.

I. 서론

대규모 언어 모델(LLM: Large Language Model)은 사회 전반에 걸쳐 광범위한 영향을 미치고 있으며, 특히 정치적 편향이 담긴 답변을 통해 사용자의 정치적 인식과 여론 형성에 직접적인 영향을 줄 수 있다. [4] 그러나 기존의 단순 질의응답이나 키워드 분석 방식만으로는 LLM이 내재한 복잡한 정치적 가치관과 판단 기준을 정확히 평가하는 데 한계가 있다.

이에 본 연구는 심리학 및 질적 연구 방법론에서 활용되는 사다리 인터뷰(Laddering Interview) 기법을 도입한다. 특히 좌(진보)/우(보수)로 다르게 파인튜닝된 LLM을 대상으로 경제, 사회, 정부 역할, 외교안보, 환경 그리고 이민과 같은 주요 정치적 이슈에 대해 가치 충돌과 우선순위를 탐색하는 질문 은행(Question Bank)을 구축하고 이를 활용하여 분석을 수행한다. 본 연구는 미국의 좌우 성향을 기준으로 한다.

본 연구의 궁극적 목적은 LLM의 정치적 성향을 명확하고 정량적으로 평가할 수 있는 분석 프레임워크를

개발하고, 이를 통해 사용자들이 LLM에게 주입된 정치적 편향을 이해할 수 있게 하는데 있다.

II. 본론

LLM은 뉴스 생성, 소셜 미디어 콘텐츠 추천 등에서 활용되며, 사용자의 정치적 인식과 여론 형성에 영향을 미친다. LLM이 좌/우 입장을 일관되게 취하면 사용자의 이념이 편향될 수 있다. 기존 연구는 질의응답이나 키워드 분석에 의존해 LLM의 복잡한 가치 체계를 충분히 파악하지 못한다. [5] 이에 본 연구는 사다리 인터뷰 기법을 도입해 LLM의 정치적 성향을 정교히 평가한다. [2] 사다리 인터뷰 기법을 통한 평가 방법은 다음과 같이 세가지로 나뉜다. **1 단계 (기초 질문):** 각 분야의 기본 입장을 묻는 6개 질문을 제시한다. 예: "정부가 의료 및 에너지 산업에서 규제와 자유 시장을 어떻게 조정해야 한다고 보는가?" **2 단계 (상위 이유 탐색):** 1단계 답변을 좌/중/우로 분류하고, 분류에

따라서 18(6*3)개 유형의 후속 질문 중 하나의 질문을 다시 LLM에게 추론한다: "그 입장이 사회에 어떤 영향을 미친다고 보는가?" **3 단계 (가치 충돌/우선순위):** 54(18*3)개 유형의 질문을 통해 가치 충돌 상황에서 우선순위를 묻는다. 예: "경제 성장과 공정성 간의 트레이드오프에서 무엇을 우선시하는가?".

질문 은행은 중립성을 유지하며, 기존 설문 자료를 참고한다. [3] 또한, 총 78개 질문은 저자들과 협력하여 특정 기준에 따라 개선하였다. 개선 기준으로는 질문의 중립성(특정 성향을 유도하지 않는가?)과 명확성(모호하지 않은가?)을 적용한다.

이렇게 구성된 사다리 인터뷰 기법이 적용된 인터뷰 데이터셋에 기본 모델, exaone-7.8b, llama3-8b, minitron-8b, qwen3-8b 을 선행연구에서 제안한 좌우 성향 데이터로 파인튜닝하여 각 모델별로 좌/우 성향을 지니고 있는 총 8개 모델을 생성하였다. [1] 이 모델들은 개발된 78개 질문 데이터셋에 적용되어 추론을 수행하였다. 인터뷰는 1 단계부터 3 단계 질문을 순차적으로 적용해 6개 인터뷰 스레드를 생성한다. 각 답변은 matous-volf/political-leaning-politics BERT 모델을 사용하여 좌(-1), 중(0), 우(1)로 분류하고, 가중치($w_1=2$, $w_2=1.5$, $w_3=1$)를 사다리 인터뷰 기법의 단계별로 적용해 최종 점수를 계산하여 모델을 정치적 성향에 따라 음수는 좌성향, 양수는 우성향으로 분류하였다.

<표 1> 사다리 인터뷰 기법을 기반으로 한 본 연구의 분류 결과와, 미국의 좌우 정치 스펙트럼을 기준으로 인간이 분류한 결과를 비교한 표. Ours의 점수는 음수일수록 좌파, 양수일수록 우파 성향을 의미한다.

모델명	훈련 성향	Ours	인간 평가
Minitron-8b	Left	-2.92	L: 16, R: 1, C: 1
Minitron-8b	Right	4.50	L: 0, R: 18, C: 0
Exaone-7.8b	Left	-0.58	L: 12, R: 6, C: 0
Exaone-7.8b	Right	4.08	L: 0, R: 17, C: 1
LlaMA3-8b	Left	-3.92	L: 18, R: 0, C: 0
LlaMA3-8b	Right	3.83	L: 0, R: 18, C: 0
Qwen3-8b	Left	-4.50	L: 18, R: 0, C: 0
Qwen3-8b	Right	2.00	L: 4, R: 13, C: 1

제안한 모델 분류 기준에 따라 훈련된 모델의 추론 결과와 인간 평가 결과를 <표 1>에 요약하였다. 인간 평가 결과와 훈련 데이터의 성향이 완전히 일치하여, 모델이 해당 성향을 성공적으로 학습했음을 확인할 수 있다. 또한, 제안한 분류 기준을 적용하였을 때, 분류 결과 정확도는 100% 인 것을 <표 1>을 통해 확인할 수 있다. 앞서 설명한 바와 같이, 음수 값은 진보(좌) 성향, 양수 값은 보수(우) 성향을 의미하며, 절댓값은 해당 성향의 강도를 나타낸다. 우리가 개발한 분류 방법을 사용하여 모델에 주입된 정치적 성향을 정확하게 확인할 수 있을 뿐만 아니라, 그 성향이 주입된 정도까지 확인할 수 있다.

III. 결론

본 연구는 사다리 인터뷰 기법을 활용하여 LLM의 정치적 성향을 평가하고 분석하는 프레임워크를 제안한다. 경제, 사회, 정부의 역할 등 미국에서 주요 논쟁이 되는 6개 분야에 대해 각 분야별로 각 성향별로(좌/중/우) 3개의 질문을 포함한 총 78개의 인터뷰 질문을 구성하고, 인간 검증을 통해 데이터셋을 정제하였다. 이후 미국의 좌우 성향 데이터셋을 활용하여 네 종류의 기본 모델을 파인튜닝함으로써 정치 성향이 반영된 편향 모델을 구축하였다. 인간 평가를 통해 훈련된 모델의 타당성을 검증하였으며, 본 논문에서 제안하는 평가 방식을 통해 모델 분류 및 분석 방법의 유효성을 추가적으로 입증하였다.

ACKNOWLEDGMENT

이 논문은 2025년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 [NO.RS-2021-II211343,인공지능대학원지원(서울대학교)]

참 고 문 헌

- [1] Lin, Z., Wang, Y., & Zhang, H. (2024). *PoliTune: Analyzing and mitigating political bias in large language models through fine-tuning*. arXiv preprint arXiv:2404.08699. <https://arxiv.org/abs/2404.08699>
- [2] Landfield, A. W. (1971). *Personal construct systems in psychotherapy*. Springfield, IL: Charles C. Thomas.
- [3] Pew Research Center. (2014). *Political polarization in the American public*. <https://www.pewresearch.org/politics/2014/06/12/appendix-a-the-ideological-consistency-scale/>
- [4] Motoki, F., Pilditch, T. D., & Savadori, L. (2023). *More human than human: Measuring ChatGPT political bias*. Public Choice, 198(1-2), 3-23. <https://doi.org/10.1007/s11127-023-01097-2>

<부록 1> 미국의 정치적 도메인별 좌파와 우파의 주요 특징 비교

도메인 (DOMAIN)	좌파 / 진보 (LIBERAL / PROGRESSIVE)	우파 / 보수 (CONSERVATIVE)
경제 (ECONOMY)	정부 개입 선호: 규제 강화, 부유층 고세율, 사회복지 확대. 경제 평등 및 노동자 보호 중시 (예: 민주당의 부자 증세, 노동조합 강화).	정부 개입 최소화: 규제 완화, 저세율, 자유시장 지지. 개인 책임 및 민간 번영 강조 (예: 공화당의 감세, 시장 중심 정책).
사회 및 문화 (SOCIETY & CULTURE)	사회적 진보, 다양성, 소수자 권리 중시. 성소수자, 인종/젠더 평등 정책 강조 (예: 다문화주의, 포용성).	전통적 가치 중시. 가족, 종교, 국가 정체성 보존 강조 (예: 애국심, 기존 규범).
정부 역할 (ROLE OF GOVERNMENT)	큰 정부 지지: 복지, 교육, 의료 등 정부 개입 확대. (예: 민주당 지지자 78%가 정부 확대 지지).	작은 정부 선호: 민간 중심 해법. 과도한 관료주의 비판 (예: 공화당 지지자 71%가 정부 축소 지지).
외교 및 안보 (FOREIGN POLICY & SECURITY)	동맹, 외교, 국제 협력 중시. 군사력보다 외교적 해법 선호.	국방력, 주권 중시. "아메리카 퍼스트" 강조, 강한 군사력 및 국경 안보 우선.
환경 및 기후 (ENVIRONMENT & CLIMATE)	기후변화 대응, 환경 규제 지지. 재생에너지, 국제 협력 중시 (예: 파리협정).	경제 성장 우선. 환경 규제 회의적, 경제 부담 우려.
이민 (IMMIGRATION)	합법적 이민 확대, 불법 이민자 보호 정책 지지.	불법 이민 억제, 국경 통제 중시. 법 집행 및 추방 정책 지지.