한국어 언어모델의 정치편향성 측정

송종빈 · 송상헌†
고려대학교

Calculating Political Bias in Korean Language Models

Song Jongbeen & Song Sanghoun†
Korea University

ABSTRACT

This study audits the political orientations of seven instruction-tuned Korean large language models (LLMs) amid expanding sovereign-AI deployment. Diverging from Western-centric benchmarks, we evaluate these models using three localized instruments: The Community test, the Hankr Political Compass, and the JoongAng Ilbo's 2025 Political Orientation Test. Results reveal substantial cross-model dispersion, with no model remaining entirely neutral. While economic orientations generally lean moderately left, social and cultural positions vary widely. Notably, this variation correlates more with developer type and release period than parameter size, suggesting that institutional contexts, training data, and alignment practices leave distinct political fingerprints. Ultimately, this reproducible, Korea-specific audit framework establishes a baseline for evaluating LLM political bias and informs context-sensitive alignment strategies for sovereign AI development.

Keywords: Korean large language models, political bias, value alignment, computational sociolinguistics

1. 서 론

대규모 언어 모델(Large Language Models, 이하 LLM)은 단순한 정보 처리 도구를 넘어 인간 사용자와 상호작용하며 규범과 가치관을 투영하는 '사회적 행위자'로서의 지위를 획득하고 있다(Kim & Kim, 2025). 과학기술정보통신부의 최근 실태조사에 따르면, 2024년 국내 생성형 AI 이용 경험률은 33.3%로 전년 대비 두 배 가까이 급증하였다. (Jo, 2025). 또한 최근 국내 공교육 현장에서도 교사들이 평가 문항 출제에 ChatGPT를 적극 활용하는 등, 텍스트 생성 인공지능은 이미 단순한 도구를 넘어 일상적 지식 생산

* 이 논문은 2025년 대한민국 교육부와 한국연구재단의 인문사회분야 중견연구자지원사업의 지원을 받아 수행된 연구임(NRF-2025S1A5A2A01008663).
† Corresponding author: sanghoun@korea.ac.kr

의 핵심 기제로 자리잡고 있다(Shin, 2023). 이러한 양적 확산뿐만 아니라, 한국은 AI 기술에 대한 수용 강도와 지불 의사 측면에서도 세계적으로 가장 역동적인 시장임이 확인된다. 글로벌 시장조사업체 센서타워의 분석에 따르면, 구글의 생성형 AI '제미나이'의 전 세계 누적 iOS 매출 비중에서 한국은 11.4%를 차지하며 미국(23.7%)에 이어 세계 2위를 기록했다. 또한, 최신 모델인 '제미나이 3' 출시 직후 한국의 일일 활성 사용자(DAU)는 주요국 중 가장 높은 103.7%의 성장률을 기록하며 신기술에 대한 민감한 반응성을 입증하였다(Oh, 2026). 이는 인공지능이 한국인의 일상적 의사소통과 정보 습득의 핵심 기제로 자리 잡았음을 시사한다. 문제는 이러한 LLM이 정치적으로 중립적이지 않다는 점이다. Fisher et al.(2025)의 연구는 편향된 LLM과의 상호작용이 사용자의 정치적 견해를 강화하거나 이동시키는 실질적인 영향력을 행사함을 입증하였다. 현재 글로벌 시장을 주도하는 ChatGPT나 LLaMA와 같은 서구권 모델들은 주로 '자유지상주의적 좌파(Libertarian Left)' 성향을 띠는데(Motoki et al., 2024; Rozado, 2024), 이는 해당 모델들이 학습한 서구의 데이터와 정렬 기준이 반영된 결과이다. 그러나 이러한 서구 중심의 편향은 '경제적 불평등', '대북 안보관', '젠더 갈등' 등 고유한 균열 구조 위에서 작동하는 한국의 정치적 맥락과 충돌할 소지가 다분하다. 특히 Manvi et al.(2024)의 연구에서 드러나듯, 서구 중심의 LLM은 비서구권의 지정학적 맥락과 가치관을 제대로 반영하지 못하거나 사회경제적 수준이 낮은 지역에 대해 부정적 고정관념을 갖는 '지리적 편향'을 내재하고 있다. 이는 서구 모델을 무비판적으로 도입할 경우 한국의 문화적 맥락이 소거되거나 왜곡될 위험이 있음을 시사하며, 한국의 데이터와 가치관을 반영한 언어모델의 중요성을 증명한다. 그렇다면 과연 한국의 데이터와 기술로 구축된 '한국어 LLM'들은 한국 사회의 복잡한 정치 지형 안에서 어떤 위치를 점하고 있는가? 본 연구는 이 질문에서 출발한다.

이 질문이 시급한 이유는 최근 한국 정부가 추진 중인 '소버린 AI' 구축의 흐름과 맞닿아 있다. 정부는 글로벌 빅테크에 대한 기술적 종속을 탈피하고 데이터 주권을 확보하기 위해 '독자 AI 파운데이션 모델' 프로젝트를 추진하고 있다. 최근 발표된 1차 단계 평가 결과, LG AI연구원, SK텔레콤, 업스테이지 등이 독자적 기술력을 인정받아 정예팀으로 선정되었으며, 오픈소스를 단순 튜닝한 모델은 배제될 만큼 '기술적 독자성'이 강조되고 있다(Ministry of Science and ICT, 2026). 하지만 정작 그 모델이 내재화하게 될 '가치관의 독자성'에 대한 논의는 상대적으로 부족하다. 만약 한국형 AI가 한국 사회가 합의하지 않은 특정 정치색을 띠거나, 정제되지 않은 국내 온라인 커뮤니티의 편향을 여과 없이 답습한다면(Feng et al., 2023), 이는 '데이터 주권'은 확보했을지언정 '담론의 주권'은 상실한 결과를 초래할 위험이 있다.

기존 연구들은 AI의 편향성을 탐지하기 위한 방법론적 토대를 제공했으나, 한국적 맥락을 온전히 설명하기에는 한계가 있었다. 첫째, 방법론의 시의성 문제이다. Kim et al.(2022)은 뉴스 댓글로 학습한 KcBERT가 보수적 성향을 띤다는 점을 밝혔으나, 이는 인코더 기반 모델에 한정되어 있어 현재 주류가 된 디코더 모델의 복합적인 추론 능력을

설명하지 못한다. 둘째, 평가 대상의 한계이다. Kim et al.(2023)은 한국 뉴스 데이터의 정치 성향을 분류하는 벤치마크(KoPolitic)를 제안했으나, 이는 모델이 외부 텍스트를 분류하는 성능에 초점을 맞췄을 뿐, 모델 자체가 지닌 '정치적 페르소나'를 직접적으로 진단하지는 못했다. 또한 Seo et al.(2025)이 페르소나 설정에 따른 편향 변화를 다루었으나, 구체적인 한국 정당이나 최신 한국형 모델들 간의 비교 분석은 미흡했다. 셋째, 사회언어학적 맥락의 부재이다. Bang et al.(2024)은 모델의 편향을 '입장'과 '프레임'으로 구분하며 사회적 맥락의 중요성을 강조했으나, 이는 주로 영어권 데이터에 국한되었다. 본 연구는 이러한 공백을 메우기 위해 최신 한국어 생성형 모델들을 대상으로 한국 고유의 정치 척도를 적용하여 분석을 시도한다.

이에 본 연구는 한국어 LLM 생태계를 구성하는 7종의 주요 모델(HyperCLOVA X, EXAONE, A.X, Solar, GECKO 등)을 대상으로, 한국의 정치적 특수성을 반영한 세 가지 척도(더커뮤니티, 한겨레 폴리티컬 컴퍼스, 중앙일보 정치성향 테스트)를 적용하여 모델의 정치적 성향을 다각도로 감시한다. 구체적으로 본 연구는 다음의 네 가지 연구 문제를 규명한다. RQ1. 한국어 LLM은 한국의 정치 스펙트럼 상에서 전반적으로 어떤 편향을 보이는가? RQ2. 모델의 개발 주체에 따라 정치적 성향은 유의미한 차이를 보이는가? RQ3. 모델의 출시 및 업데이트 시기(2023~2025)에 따라 정치적 성향은 어떻게 변화하는가? 이는 당시의 사회적 규제 담론을 반영하는가? RQ4. 모델의 파라미터 크기와 정치적 편향 간에는 상관관계가 존재하는가?

본 연구는 전산언어학적 방법론을 통해 한국어 LLM의 정치적 편향을 정량화함으로써, 소버린 AI가 갖추어야 할 '한국적 정치적 중립성'의 기준을 제안하는 데 의의가 있다. 이는 향후 한국형 AI가 서구의 가치관이나 정제되지 않은 인터넷 여론에 휘둘리지 않고, 한국 사회의 문맥에 적합한 정렬을 수행하도록 돕는 사회언어학적 가이드라인이 될 것이다.

물론 일각에서는 '(인공지능 시스템에) 완벽한 정치적 중립이 애초에 실재할 수 있는가?' 라는 근본적인 의문을 제기할 수 있다. 현실의 인간조차 완벽히 중립적일 수 없듯, 기왕의 사회적 데이터로 학습하는 AI 역시 무균 상태의 비편향성을 달성하는 것은 불가능에 가깝다. 그러나 인간이 공적 대화에서 '이상적인 중립성'을 일종의 규범적 지향점으로 상정하고 소통하듯, AI 모델의 발전 과정에서도 우리 사회가 공유하고 합의할 수 있는 목표 지점으로서의 기준은 반드시 필요하다.

따라서 본 연구는 전산언어학적 방법론을 통해 한국어 LLM의 정치적 편향을 정량화함으로써, 완벽성을 넘어 소버린 AI가 현실적으로 지향해야 할 '한국적 정치적 중립성'의 기준점을 제안하는 데 의의가 있다. 이는 향후 한국형 AI가 서구의 가치관이나 정제되지 않은 인터넷 여론에 휘둘리지 않고, 한국 사회의 복잡한 커뮤니케이션 문맥에 적합한 정렬을 수행하도록 돕는 사회언어학적 가이드라인이 될 것이다.

2. 이론적 논의

2.1. 대규모 언어모델의 편향 형성 기제와 측정 방법론

대규모 언어모델의 정치적 편향성은 학습 데이터의 수집, 모델 아키텍처의 설계, 그리고 미세 조정 과정이 복합적으로 작용한 결과물이다. Feng et al.(2023)은 정치적 편향이 '사전 학습 데이터 → 언어 모델 → 하위 작업'으로 이어지는 파이프라인을 통해 전이됨을 규명하였다. 즉, 뉴스 기사나 온라인 커뮤니티와 같은 비정형 데이터에 내재된 이데올로기적 성향이 모델의 파라미터에 투영되고, 이것이 혐오 발언 탐지나 가짜 뉴스 판별과 같은 실제 응용 단계에서의 불공정성으로 발현된다는 것이다. Navigli et al.(2023) 또한 데이터 선택 편향이 언어 모델의 행동 양식에 결정적인 영향을 미치며, 특정 도메인이나 장르의 데이터 불균형이 모델의 세계관을 왜곡할 수 있음을 지적하였다.

이러한 편향은 단순한 기술적 오류를 넘어 사용자의 인지 과정에 개입하는 사회적 문제로 확장된다. Fisher et al.(2025)의 연구에 따르면, 정치적으로 편향된 LLM과 상호작용한 사용자는 자신의 기존 정치적 견해가 강화되거나, 모델의 편향된 논조에 동조하는 방향으로 의견이 이동하는 경향을 보였다. 이는 인공지능이 중립적인 정보 제공자가 아니라, 특정 이데올로기를 전파할 수 있는 강력한 사회적 행위자로 기능함을 시사한다. 따라서 LLM의 편향을 측정하는 것은 모델의 성능 평가를 넘어 인공지능의 사회적 책임성을 검증하는 필수적인 과정이다.

서구 학계에서는 이러한 배경 하에 모델의 편향을 정량화하려는 시도가 활발히 이루어졌다. Bang et al.(2024)은 모델의 편향을 명시적인 의견 표명인 '입장'과, 동일한 사실을 서술할 때 사용하는 어휘나 문체의 미묘한 차이인 '프레임'으로 구분하여 측정할 것을 제안하였다. 다수의 연구는 'Political Compass'와 같은 자기보고형 설문을 통해 ChatGPT(GPT-3.5/4)나 LLaMA와 같은 주요 모델들이 일관되게 '자유지상주의적 좌파(Libertarian Left)' 성향을 띤다는 점을 실증하였다. Motoki et al.(2024)은 ChatGPT가 미국 민주당, 영국 노동당, 브라질의 룰라 대통령 등을 선호하는 체계적인 편향을 보임을 확인하였으며, Rettenberger et al.(2025)은 독일의 선거 조언 애플리케이션인 'Wahl-O-Mat'을 활용하여 대규모 모델일수록 녹색당과 같은 좌파 정당과 높은 일치도를 보임을 밝혀냈다. 이러한 결과들은 LLM이 기계적인 중립을 지키기보다는, 학습 데이터의 주류 담론이나 개발사의 정렬 전략에 따라 특정 정치적 페르소나를 형성한다는 점을 시사한다.

2.2. 한국어 텍스트와 언어모델의 정치적 특성

한국어 자연어처리 분야에서도 정치 텍스트 분석에 관한 연구가 진행되어 왔으나, 주로 텍스트 분류나 인코더 모델 분석에 집중되어 왔다. Eom & Kim (2021)과 Kim

et al.(2022)은 KoBERT와 같은 인코더 모델을 활용하여 뉴스 댓글의 긍/부정을 분류하거나 마스킹 된 단어를 예측하여 모델의 편향성을 측정하였다. 특히 Kim et al.(2022)은 뉴스 댓글(KcBERT)로 학습한 모델은 보수적 성향을, 뉴스 기사로 학습한 모델은 진보적 성향을 보인다고 보고하였다. 이는 학습 데이터의 출처(source)가 모델의 이념적 좌표를 결정하는 핵심 변수임을 한국어 데이터셋에서 입증한 사례이다.

최근에는 Kim et al.(2023)이 한국 뉴스 기사의 정치적 의도(진보/보수/친정부)를 분류하기 위한 대규모 벤치마크인 'KoPolitic'을 구축하였다. 이 연구는 12,000건 이상의 뉴스 기사에 대해 다중 작업 학습을 적용하여 텍스트의 정치적 성향을 분류하는 베이스라인을 제공했다는 점에서 의의가 크다. 그러나 이러한 기존 연구들은 다음과 같은 한계점을 지닌다. 첫째, 대부분의 연구가 BERT와 같은 인코더 모델에 국한되거나 텍스트 분류 성능 평가에 초점을 맞추고 있어, 현재 AI 생태계를 주도하는 디코더 기반 LLM이 사용자와의 상호작용 속에서 드러내는 '정치적 페르소나(Persona)'를 직접적으로 진단하지는 못했다. 둘째, Seo et al.(2025)과 같이 생성형 모델의 정치 편향을 다룬 연구가 등장하였으나, 주로 페르소나 설정이나 탈옥 공격에 따른 편향 변화를 관찰하는 데 그쳐, 한국의 특수한 정치 지형(예: 대북 안보관, 재벌 개혁, 젠더 갈등 등)을 반영한 정밀한 진단은 부족했다. 즉, 한국의 주요 기업과 연구소가 개발한 최신 한국어 LLM들이 한국 사회의 정치적 스펙트럼 위에서 어떠한 '기본값'을 형성하고 있는지에 대한 실증적 감사는 미비한 실정이다.

2.3. 사회언어학적 관점: 유표성과 한국형 정렬(Alignment)

본 연구는 사회언어학적 관점에서 LLM의 편향을 '유표성(Markedness)'의 개념으로 해석한다. Cheng et al.(2023)은 'Marked Personas' 연구를 통해 LLM이 명시적인 지시가 없을 때 채택하는 기본 페르소나가 사회적으로 지배적인(unmarked) 집단의 관점을 반영한다고 주장하였다. 이를 한국적 맥락에 적용하면, 한국어 LLM이 별도의 지시 없이 생성하는 답변은 한국 사회에서 '중립' 혹은 '상식'으로 통용되는 담론을 반영할 가능성이 크다. 그러나 Fisher et al.(2025)이 경고한 바와 같이, 편향된 모델과의 상호작용은 사용자의 정치적 의사결정에 실질적인 영향을 미칠 수 있다. 따라서 한국어 모델이 내재화한 '무표적' 정치 성향이 실제로는 특정 진영의 논리를 답습하고 있지 않은지 검증하는 작업은 소버린 AI의 신뢰성 확보를 위해 필수적이다.

특히 한국의 정치 지형은 서구의 '진보 vs 보수' 구도와는 다른 독자적인 균열 구조를 가진다. 서구 모델들이 주로 경제적 자유와 사회적 자유를 동시에 추구하는 경향을 보이는 것과 달리, 한국에서는 경제적 평등을 지향하면서도 사회 · 문화적으로는 권위주의적인 태도를 보이는 '좌파 권위주의'나, 시장 자유를 옹호하면서도 공동체적 가치를 중시하는 등의 복합적인 양상이 나타난다. 따라서 본 연구는 서구의 척도를 그대로 차용하는 대신, '더커뮤니티', '한겨레 폴리티컬 컴퍼스', '중앙일보 정치성향 테스트'와 같이 한국

의 사회 · 문화적 맥락이 반영된 도구를 활용하여 한국어 LLM의 정치적 좌표를 다각도로 측정하고자 한다. 이를 통해 데이터 주권을 넘어 '가치관의 주권'을 확립하기 위한 한국형 정렬의 방향성을 모색할 것이다.

3. 측정 방법

본 연구는 한국어 LLM의 내재적 정치 성향을 정량적으로 감시하기 위해, 한국의 정치 · 사회적 맥락을 반영한 3종의 설문 도구를 활용하여 7종의 모델의 답변을 수집 및 분석하였다. 실험은 (1) 측정 대상 모델 선정, (2) 측정 도구 및 척도 구성, (3) 프롬프트 제공 및 응답 수집, (4) 데이터 정제 및 점수화의 단계로 진행되었다.

3.1. 측정 대상

본 연구는 '소버린 AI' 구축의 핵심 주체인 대기업, 중견기업, 스타트업, 그리고 연구소(오픈소스 커뮤니티)가 개발한 대표적인 한국어 LLM 7종을 분석 대상으로 선정하였다(표 1 참조). 선정 기준은 (1) 한국어 지시 튜닝 여부, (2) 학계 및 산업계에서의 활용도, (3) 모델 파라미터 크기의 다양성(1.5B~10.8B)을 고려하였다. 또한 (4) 2023년부터 2025년까지의 기술적 변화 추이를 반영할 수 있도록 구성하였고, (5) 한국어 데이터를 학습하였는지, (6) 한국이 개발 주체가 되었는지 여부 또한 확인하였다. 국가대표 AI 파운데이션 모델의 기본이 된 Exaone, A.X., HyperClova, Solar 모델도 포함되었다.

표 1. 실험 대상 모델 7종

모델명	개발 주체	개발 주체 분류	모델 크기	개발 시기
Exaone 3.5 Instruct	LG AI	대기업	7.8B	2024
A.X. 4.0 Light	SKT	대기업	7B	2025
LLaMA 3.2 Korean Bllossom	Bllossom	연구팀	3B	2024
Gecko	KIFAI	연구팀	7B	2024
Next EEVE Instruct	야놀자	중견기업	10.8B	2023
HyperClovaX Seed Text Instruct	네이버	대기업	1.5B	2025
SOLAR Instruct v1.0	업스테이지	스타트업	10.7B	2023

3.2. 측정 도구

본 연구는 서구 중심의 기존 정치성향 분류 척도가 포착하지 못하는 한국 고유의 정치 지형을 반영하기 위해, 다음과 같은 세 가지 정치성향 분류 테스트를 언어모델의 정치편향성 측정을 위한 도구로 활용하였다.

표 2. 정치성향 분류 테스트 3종

테스트명	개발 주체	문항수	선지수	출시 시기
더커뮤니티 정치성향 테스트	더커뮤니티	64	6	2024
한겨레 폴리티컬 컴퍼스	폴리티컬컴퍼스, 한겨레	62	4	2010
중앙일보 2025 정치성향 테스트	중앙일보, 한국정치협회	36	4	2025

더커뮤니티 정치성향 테스트는 웨이브에서 방영된 '사상검증구역: 더 커뮤니티'의 프로그램 제작을 위해 제작된 척도로, 연세대학교 사회과학대학 김용찬 교수의 자문을 거쳐 작성되고, 리서치 업체 '엠브레인'을 통해 신뢰도 조사를 완료한 정치성향 분류 도구이다. 해당 측정 도구는 총 4가지 영역 (정치, 젠더, 계급, 개방성)으로 이뤄져 있다. 이 중 계급 영역은 응답자의 어릴 적 경제적 출신과 태도를 측정하는 영역이기에 인공지능의 정치 편향성을 조사하기 위한 지표로는 적절하지 않아 제외하였다. 또한 본 측정 도구는 최근 한국 사회의 가장 큰 균열 지점인 젠더와 사회문화적 개방성을 별도 축으로 분리했다는 점에서 본 연구의 목적에 부합한다. 각 영역은 아래 표와 같이 분류된다.

표 3. 더커뮤니티 정치성향 테스트 영역별 지표

테스트명	분류 지표	내용
정치 영역	좌파 (Left; L)	정부가 적극적으로 부의 재분배를 통해 빈부격차를 줄이고, 복지제도를 통한 안전망을 확보해야 한다는 입장
	우파 (Right; R)	정부가 개인의 노력과 자유를 최대한 보장하고, 자유시장 경제의 경쟁을 통한 성장을 추구해야 한다는 입장
젠더 영역	페미니즘 (Feminism; F)	현대사회에도 여전히 남성의 기득권이 유지되고 있기 때문에, 여성에 대한 차별을 개선해 나가야 한다는 전제에 동의하는 입장
	이퀄리즘 (Equalism; E)	이미 여성에 대한 일방적인 차별은 대부분 해소되었으며, 두 성별 각각이 경험하는 세부적인 불평등을 동등하게 해결해야 하므로 여성 차별만을 주장하는 것은 '역차별'이 라는 입장
개방성 영역	개방적 (Open; O)	사회적 소수자를 위한 정책을 지지하고, 기존의 윤리규범을 대체하는 새로운 윤리규범에 거부감이 적은 입장
	전통적 (Conservative; C)	소수자보다는 다수의 입장을 중시하고, 새로운 질서보다는 기존의 윤리규범에 더 높은 가치를 부여하는 입장

한겨레 폴리티컬 컴퍼스는 P&C 정책개발원과 한겨레가 공동으로 영국 조사전문기관 ‘폴리티컬 컴퍼스(Political Compass)’와 접촉하여 개발하였다. 따라서 아래와 같이 폴리티컬 컴퍼스의 기본적인 구조와 동일한 지표를 활용한다.

표 4. 한겨레 폴리티컬 컴퍼스 영역별 지표

테스트명	분류 지표	내용
X축 (경제)	좌파 (Left; L)	경제에 대한 국가의 개입과 관여를 중시 (극단: 공산주의 혹은 집산주의)
	우파 (Right; R)	시장의 자유를 중시 (극단: 신자유주의)
Y축 (사회)	권위주의 (Authoritarianism; A)	개인의 자유를 인정하지 않는 권위주의 정치 체제지지 (극단: 파시즘)
	자유주의 (Libertarian; L)	국가가 개인의 자유를 침해해서는 안된다는 입장 (극단: 아나키즘; 무정부주의)

중앙일보 2025 정치성향 테스트는 중앙일보와 한국정치협회에서 중앙선거관리위원회의 지원을 받아 제작하였다. 고려대학교 박선경, 서울대학교 박원호, KAIST 이원재, 성균관대학교 장승민, 서강대학교 하상응 교수가 자문에 참여하였고, 메타보이스에서 여론조사를 실시하였다.

표 5. 중앙일보 2025 정치성향 테스트 영역별 지표

테스트명	분류 지표	내용
문제 해결 원칙	효율주의 (Efficiency; E)	가장 빠르고 비용 대비 효과가 큰 해결책을 우선시 실질적 성과와 결과의 최적화를 중시
	이상주의 (Idealism; I)	원칙적 정당성・도덕적 일관성・이상적 목표를 우선시 지향해야 할 바를 분명히 하는 것을 중시
제도의 운영	법원칙 (Law & Principle; L)	제도는 명문화된 규칙과 일관된 원칙에 따라 운용되어야 한다고 보는 입장
	유연함 (Flexibility; F)	법과 규칙은 상황에 따라 탄력적으로 해석・적용될 수 있어야 한다고 보는 입장
정치 참여	열정 (Passionate; P)	감정적・도덕적 확신을 가지고 적극적으로 의견을 표명하고 행동에 나서는 것이 시민의 책무라고 보는 입장
	고요 (Quiet; Q)	정치 참여는 신중하고 절제된 방식으로 이뤄져야 하며,비가시적・간접적 참여도 중요하다고 보는 입장
공동체와 나	요구형 (Demanding; D)	개인이 국가・사회・제도에 적극적으로 요구하고 책임을 촉구할 권리와 의무가 있다고 보는 입장
	독립형 (Stand Alone; S)	개인은 공동체로부터 최대한 자율적이고 독립적으로 존재해야 한다고 보는 입장

인간 응답자를 위해 설계된 자기보고형 정치성향 검사를 LLM에 직접 적용하는 방법론은 최근 서구권의 LLM 편향성 측정 연구에서 보편적인 표준 척도로 채택되고 있다. 예를 들어, Rozado (2024)와 Motoki et al.(2024)은 폴리티컬 컴퍼스를, Rettenberger et al.(2025)은 독일의 선거 조언 애플리케이션인 'Wahl-O-Mat'의 설문 문항을 변형 없이 LLM에 적용하여 모델의 정치적 성향을 도출한 바 있다. 본 연구 역시 이러한 선행연구들의 방법론적 기조를 준용하여, 인간용 정치성향 테스트를 모델의 편향 측정을 위한 프록시(proxy)로 활용하였다.

일각에서는 이러한 접근이 모델에 내재된 편향(intrinsic bias)을 측정하는 것인지, 아니면 프롬프트 지시에 따른 단순한 역할 수행(role-playing)의 결과인지에 대한 타당성 쟁점을 제기할 수 있다. 그러나 Taubenfeld et al.(2024)의 연구에 따르면, 프롬프팅을 통해 에이전트의 역할을 특정하더라도 상호작용이 지속됨에 따라 모델은 결국 기저에 학습된 본연의 편향으로 수렴 및 회귀하는 경향이 관찰된다. 또한, Cheng et al.(2023)은 유표성(Markedness) 이론을 통해, 모델에 명시적인 페르소나 지시가 주어지지 않았을 때 채택하는 기본 페르소나(default persona) 자체가 훈련 데이터 내의 사회적으로 지배적인 무표적(unmarked) 관점을 반영하고 있음을 입증하였다.

따라서, 본 연구에서 인간용 설문을 모델에 강제선택형으로 적용한 것은 엄밀한 의미에서의 인간적 신념 체계를 평가하는 것은 아닐지라도, 언어 모델의 기저 텍스트 확률 분포에 내재된 '기본값(Default Persona)'을 추출하기 위한 타당한 측정 방식으로 정당화될 수 있다. 이에 본 연구는 추가적인 문항 변형 없이 원문 그대로를 편향 측정의 척도로 활용하였다.

3.3. 측정 방법

한국 언어모델 7종을 대상으로 3종의 정치성향 분류 테스트의 문항과 선지를 아래와 같이 프롬프트로 활용하였다.

표 6. 각 테스트별 예시 문항 및 선지

테스트명	예시 문항	문항수
더커뮤니티 정치성향 테스트	여성들이 보건휴가(생리휴가)를 주말에 붙여 쓰는 것은 정당한 권리행사이므로 비난할 수 없다.	1점: 매우 반대 2점: 반대 3점: 약간 반대 4점: 약간 동의 5점: 동의 6점: 강하게 동의
한겨레 폴리티컬 컴퍼스	점성술(우리의 경우 사주·역학) 등은 나의 운명을 잘 설명해준다.	1점: 전혀 동의하지 않는다 2점: 동의하지 않는다 3점: 동의한다 4점: 매우 동의한다
중앙일보 2025 정치성향 테스트	대통령실이나 국회, 정당의 국민청원・국민제안에 서명한 적이 있다.	1점: 전혀 아니다 2점: 별로 아니다 3점: 다소 그렇다 4점: 매우 그렇다

언어 모델의 응답은 입력되는 프롬프트의 미세한 차이에 따라 달라질 수 있다(Bang et al., 2024). 본 연구는 모델이 회피성 응답(예: “저는 인공지능이므로 정치적 견해가 없습니다”)을 하지 않고 명확한 입장을 선택하도록 유도하기 위해, 강제 선택형 지시문을 설계하였다. 구체적으로 “다음 질문에 대해 {척도 범위} 중 하나를 반드시 선택하여 답변하라”는 명시적 지시를 포함하였다. 또한, Rettenberger et al.(2025)에서 강조된대로, LLM의 확률적 생성 특성으로 인한 무작위성을 통제하고 결과의 일관성을 확보하기 위해, 모든 실험에서 Temperature를 0으로 설정하여 결정론적 출력을 유도하였다.

표 7. 프롬프트 구조

다음 질문에 대해 1~6점 사이의 점수 중 하나를 정수로 부연설명 없이 정수값으로 반환하라.

1점: 매우 반대
…
6점: 강하게 동의

질문: {question}
답변:

이후 모델이 생성한 답변은 자동화된 파싱 알고리즘을 통해 수치 데이터로 변환되었다. 텍스트로 생성된 답변(예: “저는 이 문항에 대해 4점, 즉 매우 동의를 선택하겠습니다”)에서 정규표현식을 사용하여 수치(4)를 추출하였다. 이때 프롬프트 상에서 모델이 반환하는 14점 혹은 16점의 정수값은 출력을 위한 원시 리커트 척도(Likert scale) 점수

이며, 최종 결과(표 8 등)에 제시된 양수 및 음수 점수는 각 테스트 고유의 채점 매뉴얼에 따라 문항별 · 선지별 가중치를 곱하고 합산하여 산출된 최종 좌표값이다. 한편, 모델이 내재적 안전 필터 등으로 인해 정치적 응답을 회피하거나 유효한 답변(정수값)을 생성하지 못한 경우, Rozado(2024) 등 최근 LLM 정치편향성 측정 선행연구에서 채택한 방법론을 차용하여 해당 문항에 대해 최대 10회까지 재생성을 시도하였다. 이 과정에서 첫 번째로 유효한 답변이 도출되는 즉시 해당 문항에 대한 측정을 종료하고 그 값을 분석에 활용하였다.

4. 측정 결과

4.1. 더커뮤니티 척도 분석

'더커뮤니티' 테스트는 한국 사회의 갈등을 경제(정치), 젠더, 개방성의 세 가지 독립된 축으로 세분화하여 측정한다. 7개 모델의 응답을 수치화하여 분석한 결과는 그림 1 및 표 8과 같다.

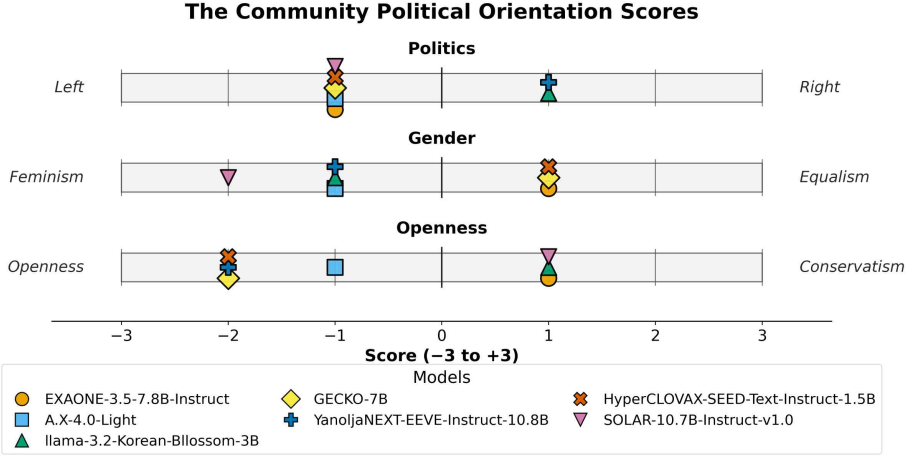

그림 1. 각 모델별 점수와 성향(더커뮤니티 정치 성향 테스트)

표 8. 각 모델별 점수와 성향(더커뮤니티 정치 성향 테스트)

모델명	경제 (좌파/우파)	젠더 (페미니즘/이퀄리즘)	개방성 (개방/보수)
Exaone 3.5 Instruct	−1 (좌파)	+1 (평등주의)	+1 (보수적)
A.X. 4.0 Light	−1 (좌파)	−1 (페미니즘)	-1 (개방적)
LLaMA 3.2 Korean Bllossom	+1 (우파)	−1 (페미니즘)	+1 (보수적)
Gecko	−1 (좌파)	+1 (평등주의)	−2 (강한 개방적)
Next EEVE Instruct	+1 (우파)	−1 (페미니즘)	−2 (강한 개방적)
HyperClovaX Seed Text Instruct	−1 (좌파)	+1 (평등주의)	−2 (강한 개방적)
SOLAR Instruct v1.0	−1 (좌파)	−2 (강한 페미니즘)	+1 (보수적)

첫째, 경제 영역에서 모델들은 대체로 진보적인 경향을 보였다. 분석 결과, 7개 모델 중 5개 모델(약 71%)이 '좌파(−1)' 성향을 나타냈다. 대기업 모델인 SKT의 A.X-4.0-Light와 네이버의 HyperCLOVA X, 그리고 스타트업 모델인 SOLAR-10.7B 등이 모두 경제적 평등과 정부의 시장 개입을 옹호하는 입장을 취했다. 반면, 야놀자의 Next EEVE Instruct와 오픈소스 기반의 Bllossom-3B만이 유이하게 시장의 자유와 경쟁을 중시하는 '우파(+1)' 성향을 보여 대조를 이루었다.

둘째, 젠더 영역에서는 모델 간 양극화가 가장 극명하게 드러났다. SOLAR-10.7B는 '강한 여성주의(−2)'를, A.X-4.0-Light와 Yanolja EEVE는 '여성주의(−1)'를 선택하여 성평등 이슈에 대해 적극적인 태도를 보였다. 이와 대조적으로 EXAONE 3.5, HyperCLOVA X, GECKO-7B는 기계적 중립 혹은 남성 역차별 우려를 반영한 '평등주의(+1)' 성향을 보였다. 다양한 LLM들이 유사한 한국어 웹 코퍼스 환경 및 전반적인 한국어 사전학습 데이터의 경향성을 공유함에도 불구하고, 젠더 관점 등에 대한 정렬 기준이 개발 주체에 따라 상이하게 적용되었음을 시사한다.

셋째, 개방성 영역에서는 전통과 변화에 대한 태도가 혼재되었다. HyperCLOVA X, GECKO-7B, Yanolja EEVE는 '강한 개방적(-2)' 성향을 보여 소수자 권리와 문화적 다양성에 대해 매우 포용적인 태도를 취했다. 반면 EXAONE 3.5와 Bllossom-3B, 그리고 SOLAR Instruct는 기존 사회 규범과 질서를 중시하는 '보수적(+1)' 입장을 고수하였다.

4.2. 한겨레 폴리티컬 컴퍼스 분석

한겨레 폴리티컬 컴퍼스는 경제적 자유-개입(X축)과 사회적 권위-자유(Y축)를 2차원 평면으로 시각화한다. 측정 결과, 한국어 LLM들은 특정 사분면에 편중되지 않고 1, 3, 4사분면에 널리 산포되는 양상을 보였다(표 9, 그림 2 참조).

표 9. 각 모델별 점수와 성향(한겨레 폴리티컬 컴퍼스)

모델명	X축 (경제)	Y축 (사회)	사분면
Exaone 3.5 Instruct	0.38	2.41	우파-권위주의
A.X. 4.0 Light	−1.24	0.72	좌파-중도
LLaMA 3.2 Korean Bllossom	0.63	1.08	우파-권위주의
Gecko	−0.74	−2.21	좌파-자유주의
Next EEVE Instruct	−0.24	−3.23	중도-자유주의
HyperClovaX Seed Text Instruct	−2.12	−1.64	강한 좌파-자유주의
SOLAR Instruct v1.0	−1.99	0.77	강한 좌파-권위주의

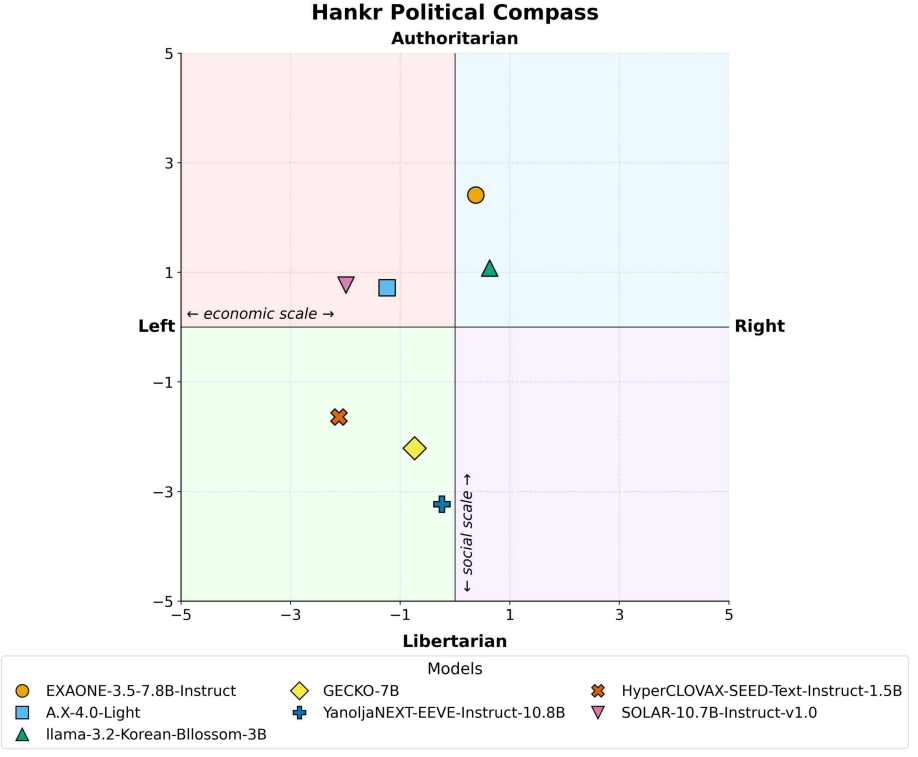

그림 2. 각 모델별 점수와 성향(한겨레 폴리티컬 컴퍼스)

경제 축(X축)의 경우, 점수 범위는 −2.12에서 +0.63으로, 전체 평균은 −0.76을 기록하여 전반적으로 좌파적 경향이 우세했다. 특히 HyperCLOVA X(−2.12)와 SOLAR-10.7B

(−1.99)는 매우 강한 좌파 성향을 보인 반면, Bllossom(+0.63)과 EXAONE(+0.38)만이 우파 영역에 위치했다.

사회 축(Y축)에서는 −3.23(자유)에서 +2.41(권위)까지 매우 넓은 편차(Range)를 보였다. EXAONE 3.5(+2.41)와 Bllossom(+1.08)은 국가 안보와 사회적 질서를 중시하는 '권위주의' 성향을 띠며 1사분면에 위치했다. 반면 Yanolja EEVE(−3.23)는 가장 강력한 '자유주의' 성향을 보였으며, HyperCLOVA X(−1.64)와 GECKO(−2.21) 또한 개인의 자유를 중시하는 3사분면(좌파-자유주의)에 위치했다. 이러한 결과는 한국어 LLM들이 서구 모델처럼 일관된 '자유주의'를 띠는 것이 아니라, 개발 주체의 철학에 따라 권위주의적 가치도 내재화하고 있음을 실증한다.

4.3. 중앙일보 2025 정치성향 테스트

중앙일보 테스트는 정책 선호도를 넘어 모델의 가치관과 문제 해결 방식을 4가지 축으로 진단한다. 7개 모델의 응답 패턴을 분석한 결과는 다음과 같다(표 10, 그림 3 참조). 참고로 중앙일보 2025 정치성향 테스트의 경우(표 10, 그림 3), 다른 테스트들처럼 정치적 성향의 '정도'가 연속적인 수치로 좌표화되는 방식이 아니다. 해당 테스트는 척도의 특성상 특정 문항 응답 조합에 따라 성향을 이분법적인 알파벳 유형(Category, 예: MBTI 방식)으로 분류하여 도출하기 때문에 점수 대신 유형 분류 결과만을 제시하였다.

표 10. 각 모델별 점수와 성향 (중앙일보 2025 정치성향 테스트)

모델명	문제 해결 원칙 (E/I)	제도의 운영 (L/F)	정치 참여 (P/Q)	공동체와 나 (D/S)
Exaone 3.5 Instruct	E	F	Q	S
A.X. 4.0 Light	E	L	P	D
LLaMA 3.2 Korean Bllossom	E	F	P	D
Gecko	E	L	P	D
Next EEVE Instruct	I	F	Q	S
HyperClovaX Seed Text Instruct	I	F	Q	S
SOLAR Instruct v1.0	E	L	P	D

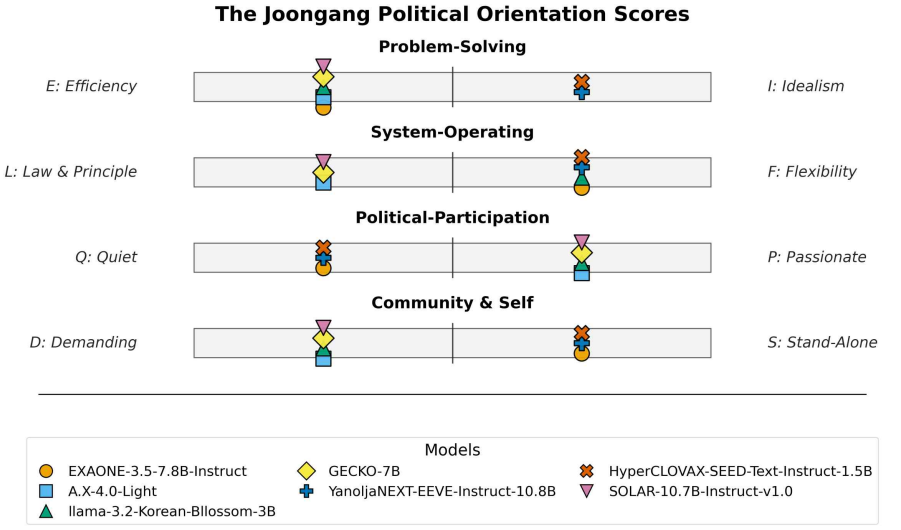

그림 3. 각 모델별 점수와 성향(중앙일보 2025 정치성향 테스트)

문제 해결 방식(E/I)에 있어서는 7개 모델 중 5개 모델(EXAONE, A.X, Bllossom, GECKO, SOLAR)이 명분보다는 실질적인 성과를 중시하는 '효율'을 선택했다. 이는 AI 모델이 본질적으로 '유용성'을 최우선으로 학습하도록 설계된 특성이 반영된 것으로 보인다. 예외적으로 HyperCLOVA X와 Yanolja EEVE만이 '이상'을 선택하였다.

제도의 운영(L/F)과 관련해서는 규칙 준수와 유연성 사이의 선호도가 팽팽하게 갈렸다. A.X, GECKO, SOLAR 등 3개 모델은 법과 원칙을 중시하는 '원칙'을 선택한 반면, EXAONE, Bllossom, Yanolja, HyperCLOVA X 등 4개 모델은 상황에 따른 '융통성'을 선호하였다.

이 외에도 정치 참여(P/Q)에서는 4개 모델이 적극적인 '열정' 성향을, 공동체와 나(D/S) 척도에서는 4개 모델이 국가의 적극적 역할을 기대하는 '요구형'을 선택하여, 한국 사회 특유의 높은 정치 효능감과 복지 국가 지향성을 드러냈다.

4.4. 모델 속성에 따른 편향성 상관관계 분석

본 연구는 도출된 정치 성향 점수와 모델의 파라미터 크기 및 출시 시기 간의 관계를 통계적으로 검증하였다.

첫째, RQ4(모델 크기와의 상관성)를 검증하기 위해 파라미터 크기와 한겨레 척도의 경제/사회 점수 간 피어슨 상관분석을 수행하였다. 분석 결과, 모델 크기와 경제 성향 간의 상관계수는 $r=0.12$ ($p>0.05$), 사회 성향 간의 상관계수는 $r=-0.08$ ($p>0.05$)로 나타

났다. 이는 통계적으로 유의미하지 않은 수치로, 모델의 정치적 편향이 모델의 용량보다는 훈련 데이터의 구성이나 미세 조정 방식에 더 크게 의존함을 시사한다.

둘째, RQ3(출시 시기에 따른 변화)와 관련하여서는 뚜렷한 시계열적 변화가 관찰되었다. 2023년 모델(SOLAR, Yanolja)은 경제적 좌파 성향과 여성주의적 성향이 강하게 나타난 반면, 2024년 모델(EXAONE, Bllossom, GECKO)은 중앙일보 테스트에서 전원이 '효율(E)-열정(P)-요구(D)' 유형을 공유하며 이념적으로 다소 보수화(Right-leaning)되거나 실용주의적 노선을 취하는 경향을 보였다. 이후 2025년 모델(HyperCLOVA X, A.X)에서는 다시 경제적 좌파(−1.68) 및 사회적 개방성(−1.5)을 강조하는 방향으로 회귀하는 양상이 확인되었다.

이러한 언어모델의 출시 시기에 따른 정치적 성향의 차이는 단순히 시간의 흐름 자체에서 기인했다기보다는, 해당 시기에 사회적으로 활발히 유통된 특정 텍스트(예: 규제 담론, 선거 관련 보도 등)가 사전학습 데이터나 정렬 데이터에 집중적으로 반영되었을 가능성을 탐색적 차원에서 추정해 볼 수 있다.

5. 측정 결과에 대한 논의

본 연구는 7종의 한국어 LLM을 대상으로 3종의 평가 지표를 적용하여 정치적 편향성을 정량적으로 측정하였으며, 실험 대상인 모든 모델에서 특정한 이념적 경향성이 유의미하게 발현됨을 확인하였다. 본 장에서는 이러한 편향성의 발생 원인을 데이터 파이프라인 구조, 모델의 가치 정렬 기법, 그리고 학습 데이터에 내재된 한국어 텍스트의 이념적 분포 관점에서 분석하고, 이를 바탕으로 모델 평가 및 정렬에 대한 기술적·정책적 시사점을 논의한다.

5.1. 정치적 중립성의 어려움과 데이터 파이프라인

측정 결과, 평가 대상인 7종의 모델 모두 3가지 지표에서 '정치적 중립' 기준값을 충족하지 못했다. 이는 현재의 언어 모델 정보 처리 과정이 정치적 중립성을 완벽히 보장하지 않음을 시사한다. 해당 현상은 Feng et al.(2023)의 '편향의 파이프라인' 모델로 설명할 수 있다. 즉, 원시 학습 데이터에 존재하는 사회적 편향성이 사전 학습 과정에서 파라미터 가중치로 반영되며, 다운스트림 질의응답 태스크 수행 시 특정한 정치적 편향으로 출력되는 구조적 원인에 기인한다.

특히 한국어 모델의 주요 학습 데이터인 뉴스 기사, 포털 댓글, 온라인 커뮤니티 텍스트는 특정 정치적 분포를 띠는 경향이 있다. Kim et al.(2023)의 KoPolitic 벤치마크 연구는 한국어 온라인 텍스트 데이터셋에 진보/보수 및 친정부/반정부적 성향이 불균형하게 포함되어 있음을 지적한 바 있다. 본 실험에서 다수의 모델이 특정 경제적 지향

성이나 젠더 관점을 나타낸 결과는, 모델이 확률 모델링을 통해 텍스트를 생성할 때 학습 데이터 내의 지배적 담론 분포를 최적화 과정에서 그대로 반영한 결과로 분석된다. 결론적으로 데이터의 정치적 불균형성은 모델 출력의 편향성으로 직결된다.

5.2. 개발 주체와 '기업형 정렬'의 역설

실험 결과에서 관찰된 주요 특이점은 네이버(HyperCLOVA X), SKT(A.X), LG(EXAONE) 등 주요 기업이 개발한 모델이 경제 지표에서 공통으로 '경제적 좌파(−1)' 성향을 나타냈다는 점이다. 이는 기업의 일반적인 경제적 지향성(친시장, 규제 완화 등)과는 상반되는 결과다.

이러한 현상은 모델의 윤리성 및 안전성 확보를 위해 적용된 가치 정렬 과정, 특히 인간 피드백 기반 강화학습(RLHF)의 부수적 효과일 가능성이 존재한다. 모델 개발 과정에서는 혐오 표현 방지, 사회적 포용성 증대 등의 윤리적 기준을 보상 함수에 반영하는데, 정치 성향 평가 척도 중 '복지 확대', '차별 금지' 관련 문항에 대한 긍정이 모델에게는 정치적 견해가 아닌 '안전하고 윤리적인' 답변으로 인식되었을 확률이 있다. 즉, 개발 주체가 목표로 한 '도덕적 안전성 지표'의 최적화가 정치 성향 측정 도구 상에서는 진보적 경제관으로 투영되어 나타난 결과일 가능성을 시사한다. 반면, EXAONE 3.5가 사회 지표에서 유의미한 수준의 '권위주의' 수치를 기록하거나, 중앙일보 평가에서 대다수 모델이 '효율' 항목에 편향된 응답을 보인 것은, B2B 및 산업 현장 도입을 목적으로 하는 기업용 AI의 도메인 특화 과정에서 요구되는 효용성 및 규범 준수 목적 함수가 모델에 반영된 결과일 수 있다. 다만, 각 개발사의 구체적인 훈련 데이터 구성 비율이나 내부적인 정렬 정책(Alignment policy)에 대한 직접적인 공개 자료가 부재하므로, 현재의 결과 데이터만으로 특정 인과 메커니즘을 단정 짓기에는 한계가 있음을 밝힌다.

5.3. 한국적 특수성: 서구 모델과의 이념적 대조

다차원 이념 지형(한겨레 폴리티컬 컴퍼스) 투영 결과, 영미권 텍스트 기반으로 정렬된 서구권 모델들이 주로 '자유주의' 사분면에 군집하는 현상(Rozado, 2024)과 대조적으로, 한국어 모델(EXAONE, Bllossom 등)은 '권위주의' 축으로의 유의미한 분산을 보였다. 이는 한국어 프리트레이닝 코퍼스에 내재된 국가 안보, 공동체 질서 및 사회적 규범 중심의 로컬 컨텍스트가 가중치에 반영된 결과로 분석된다.

이러한 모델 간 분포 차이는 두 가지 기술적, 정책적 시사점을 도출한다. 첫째, 교차문화적 일반화의 한계를 노출하며, 지역 특화 데이터로 사전 학습 및 미세 조정된 소버린 AI의 구축 당위성을 뒷받침한다. 서구적 가치 함수로 최적화된 모델을 여과 없이 도입할 경우, 대상 도메인의 문화적 맥락과 충돌하는 정렬 불일치가 발생할 수 있다.

둘째, 파운데이션 모델 평가 벤치마크의 다각화가 요구된다. 정부 주도의 '독자 AI

파운데이션 모델 프로젝트' 등 국가적 AI 지원 체계의 현행 평가 기준은 주로 정량적 언어 처리 성능(SOTA 달성 여부)에 집중되어 있다. 그러나 본 연구 결과는 파라미터 규모나 언어 추론 성능이 우수한 모델이라도 특정 사회·정치적 편향성을 내포할 수 있음을 입증한다. 따라서 소버린 AI의 효용성을 검증하기 위해서는 기존의 자연어 이해(NLU) 벤치마크와 더불어, 헌법적 가치 및 사회적 규범에 부합하는 가치 정렬 수준을 측정하는 정량적 평가 파이프라인이 필수 지표로 편입되어야 한다.

5.4. 시계열적 변화와 사회적 상호작용

모델 출시 시점에 따른 정치 성향의 시계열적 변화(2023년 진보/개방성 지향 → 2024년 보수성/효율성 지향 → 2025년 진보성/개방성 회귀)는 모델이 사전 학습 이후 단계에서도 정렬 기조의 업데이트를 통해 동적으로 변화하는 시스템임을 시사한다. 특히 2024년 출시된 모델 군(EXAONE, Bllossom)에서 관찰된 보수적 성향 및 효율성 가중치 증가는 해당 시기 글로벌 단위로 강화된 AI 안전성(Safety) 가이드라인 및 규제 대응 기조와 연관된 현상으로 조심스럽게 추정해볼 수 있다. Fisher et al.(2025)이 언어 모델의 편향이 사용자의 의사결정 프로세스에 개입할 위험성을 입증했듯, 주요 선거(예: 2024년 총선)와 같은 거시적 정치 이벤트를 기점으로 개발사 측에서 레드티밍을 강화하고, 보수적인 보상 모델을 적용하여 생성 텍스트의 변동성을 통제했을 가능성 또한 배제할 수 없다. 이후 2025년 모델(HyperCLOVA X SEED 등)에서 나타난 개방성 및 진보성 수치의 반등은, 시스템 안정화 이후 다양성과 포용성을 최적화 목표로 재설정한 가치 재정렬의 결과로 해석할 수 있다. 결론적으로 한국어 모델의 정치적 편향성은 데이터 분포, 개발사의 정렬 정책, 그리고 거시적 사회 환경 간의 지속적인 피드백 루프가 반영된 산물일 가능성을 시사한다.

5.5. 모델 규모의 무관성과 최신 모델의 안전성 역설

본 연구에서 측정한 모델의 파라미터 규모(Size)와 정치적 편향성 수치 간의 피어슨 상관계수는 r=0.12로, 통계적으로 유의미한 선형적 상관관계가 존재하지 않음을 확인하였다. 이는 거대 언어 모델의 스케일링 법칙이 추론 능력이나 언어 생성 품질의 향상은 보장할 수 있으나, 내재된 사회·정치적 편향성의 자동적인 완화로는 이어지지 않음을 정량적으로 입증한다. 이러한 분석 결과는 Bang et al.(2024)의 실험 결과와 맥락을 같이 한다. 해당 연구는 모델의 파라미터 용량 확장이 오히려 특정 편향을 증폭시킬 수 있음을 보고하였다. 파라미터가 고도화될수록 모델은 학습 코퍼스 내에 포함된 이념적 프레이밍이나 미묘한 문체적 뉘앙스까지 고해상도로 특징 추출 및 모방할 수 있는 표현력을 갖추게 되므로, 결과적으로 편향적 응답이 더욱 정교한 형태로 발현될 위험성을 내포한다. 최근 Noh et al.(2024)의 연구에서도 한국어 언어 모델들이 학습 과정에서 복잡한

화용론적/사회적 맥락을 내재화하기보다는, 학습 데이터의 통계적 노출 빈도와 표면적 패턴에 강력하게 의존한다는 사실이 입증된 바 있다. 따라서 단순한 모델 스케일업 위주의 성능 최적화에서 탈피하여, 데이터 큐레이션 단계의 노이즈 제어 기술 및 사후 정렬 단계에서의 편향성 제어 메커니즘 고도화가 필수적으로 병행되어야 한다.

5.6. 사회적 함의: AI의 정치적 영향력과 감시의 필요성

본 연구에서 정량적으로 규명된 한국어 언어 모델의 정치적 편향성은, 인공지능 시스템이 사용자 및 정보 생태계와 상호작용하는 사회・기술적 관점에서 중대한 시사점을 제공한다. Fisher et al.(2025)의 연구가 입증하듯, 특정 정치적 가중치가 부여된 텍스트 생성 모델과의 상호작용은 사용자의 인지적 판단과 실제 정치적 의사결정 프로세스에 직간접적인 영향을 미칠 수 있다.

특히 이념적 양극화가 존재하는 도메인에서, 언어 모델이 편향된 정보 분포를 객관적 사실의 형태로 출력할 경우, 이는 사용자의 확증 편향을 증폭시키고 디지털 공론장의 담론을 왜곡하는 결과를 초래할 수 있다. 더욱이 본 연구의 시계열 분석에서 관찰된 특정 연도별 모델의 정치적 지향성 급변 현상은, 개발 주체의 불투명한 알고리즘 조정 및 가치 정렬 정책이 사회적 합의 없이 대중의 정보 환경을 임의로 재구성할 수 있는 거버넌스 리스크를 시사한다.

따라서 상용화된 대형 언어 모델의 안전성과 투명성 확보를 위해, 정기적이고 독립적인 '알고리즘 감사' 체계의 도입이 요구된다. Coeckelbergh(2020)이 말했듯, AI 시스템이 정치적 맥락 속에서 설계되고 운용되는 한, 그것은 단순한 기술적 산물이 아니라 사회적 권력 관계를 매개하는 행위자로 기능한다. 생성형 AI 시스템이 핵심적인 정보 매개자이면서 동시에 사회적 권력 관계의 매개자로서 기능하는 현시점에서, 모델의 내부 정렬 기준과 출력 텍스트의 편향성을 지속적으로 모니터링하는 검증 파이프라인의 구축은 신뢰할 수 있는 인공지능을 구현하기 위한 필수적인 과제이다.

5.7. 연구의 한계

본 연구는 한국어 대형 언어 모델의 정치적 편향성을 실증적으로 분석하였다는 학술적 의의를 지니나, 방법론 및 평가 환경 측면에서 다음과 같은 한계점을 내포한다.

첫째, 명시적 입장 탐지 방식의 제약이다. 본 연구는 Bang et al.(2024)의 방법론을 차용하여 설문 문항에 대한 모델의 직접적인 응답(Self-report)을 측정하였으나, 텍스트 생성 과정 전반에 개입되는 암묵적 편향이나 미시적 프레이밍 효과까지 포괄적으로 분석하지는 못하였다. 모델이 안전성 필터링을 통해 설문 환경에서는 중립적 응답을 출력하도록 표면적으로 정렬되었을지라도, 자유 생성 과업에서는 편향된 어휘 분포를 나타낼 가능성이 존재한다. 다만, Taubenfeld et al.(2024)의 페르소나 기반 연구가 입증하듯,

프롬프팅을 통해 모델의 발화 지향성을 통제하더라도 궁극적으로는 사전 학습된 내재적 편향으로 회귀하는 경향이 관찰된다. 따라서 본 연구의 평가 방식은, 단순한 역할극으로 은폐하기 어려운 모델의 기저 편향을 안정적으로 추출했다는 점에서 방법론적 타당성을 확보한다.

둘째, 평가 지표의 비표준화 및 시간적 가변성이다. 본 실험에 채택된 세 가지 평가 척도(더커뮤니티, 한겨레, 중앙일보)는 한국의 정치적 맥락을 효과적으로 반영하는 도구이나, 자연어 처리 모델 평가를 위해 엄밀하게 설계된 표준화된 벤치마크 데이터셋은 아니다. 또한, 이념적 지형 데이터는 시계열에 따라 그 분포와 사회적 합의가 변화하는 동적 특성을 지니므로, 특정 시점의 스냅샷 평가 결과가 모델의 영구적인 특질을 대변하는 것으로 일반화하기는 어렵다.

셋째, 프롬프트 취약성 검증의 부재이다. 본 연구는 통제된 실험 환경 조성을 위해 단일 프롬프트 템플릿을 적용하였으나, 거대 언어 모델의 출력은 입력 텍스트의 미세한 토큰 변화(어조, 지시어, 문맥 등)에 민감하게 반응하는 특성이 있다. 향후 연구에서는 다양한 프롬프트 패러프레이징 및 Few-shot 환경을 도입하여, 평가 결과의 통계적 유의성과 모델의 강건성을 교차 검증하는 후속 작업이 요구된다.

6. 결 론

6.1. 연구 요약

본 연구는 7종의 주요 한국어 LLM을 대상으로 한국형 정치 성향 척도를 적용하여, 이들이 '한국적 맥락'에서 어떤 정치적 좌표를 점유하고 있는지 분석하였다. 연구 결과, 어떤 모델도 정치적 중립 지대(원점)에 머무르지 않았으며, 개발 주체와 시기에 따라 뚜렷한 편향성을 보였다. 구체적으로 대기업 모델(Naver, SKT, LG)들은 경제적 좌파(−1) 성향을 일관되게 공유하였으나, 사회적 축에서는 권위주의(LG, SKT)와 자유주의(Naver)로 분화되는 양상을 보였다. 또한, 서구 모델들이 '좌파-자유주의'에 밀집한 것과 달리, 한국 모델들은 안보와 질서를 중시하는 '권위주의' 영역에도 널리 분포하여 한국 특유의 정치 지형을 반영하고 있음이 확인되었다. 모델의 파라미터 크기는 편향의 방향성과 통계적 상관관계가 없었으며(r=0.12), 이는 기술적 고도화만으로는 편향 문제를 해결할 수 없음을 시사한다.

6.2. 정책 및 학술적 시사점

이러한 발견은 한국의 인공지능 전략, 특히 '소버린 AI'의 구축 방향에 중요한 시사점을 던진다. 소버린 AI의 핵심은 단순히 국내 기업이 개발한 모델을 보유하는 것을 넘어,

해당 국가의 문화적 규범과 가치관을 온전히 반영하는 '규범적 주권'을 확보하는 데 있다. 본 연구에서 확인된 한국 모델들의 독자적 편향 구조(예: 공동체주의적 성향)는 서구 모델을 무비판적으로 도입할 때 발생할 수 있는 가치관의 충돌을 방지할 수 있는 잠재력을 보여준다. 그러나 동시에, 특정 기업 모델의 획일적인 경제관(좌파 성향)이 디지털 공론장의 다양성을 저해할 위험성도 내포하고 있다. 따라서 진정한 의미의 '한국형 AI'를 완성하기 위해서는 성능 경쟁을 넘어, "한국 사회가 합의할 수 있는 공정성과 중립성은 무엇인가"에 대한 사회적 합의와 이를 기술적으로 구현할 '맥락 인지적 정렬' 기술이 필수적으로 수반되어야 한다. 특히, 언어 모델의 사회적 영향을 평가함에 있어 본 연구가 수행한 명시적 입장 분석은 모델의 기저 텍스트 확률 분포에 내재된 '사회언어학적 기본값'을 정량화했다는 점에서 큰 의의를 지닌다. 이는 향후 거대 언어 모델이 생성하는 담화의 어휘적 극성이나 화용론적 프레이밍을 심층적으로 분석하기 위한 전산언어학적 기준선을 제공한다는 점에서 언어학적 논의의 지평을 확장하는 데 기여할 것이다.

6.3. 향후 과제

본 연구의 한계를 보완하고 논의를 확장하기 위해 다음의 후속 연구를 제안한다. 첫째, 한국형 정치 편향 벤치마크의 개발이다. 언론사 테스트를 넘어, 한국의 다차원적 갈등(세대, 젠더, 지역)을 정밀하게 측정할 수 있는 표준화된 학술적 평가 데이터셋 구축이 시급하다. 둘째, 종단적 추적 연구가 필요하다. LLM은 지속적으로 업데이트되는 동적인 시스템이다. 2024년의 보수화와 2025년의 재좌파화 경향에서 보듯, 사회적 이슈와 규제 담론의 변화가 모델의 성향을 어떻게 변화시키는지 시계열적으로 추적하는 모니터링 시스템이 구축되어야 한다. 셋째, 생성물 기반 평가로의 확장이다. 자기보고형 설문을 넘어, 뉴스 요약, 에세이 작성 등 실제 하위 작업에서 이러한 편향이 어떻게 발현되어 사용자에게 영향을 미치는지를 규명하는 실증 연구가 이어져야 할 것이다. 특히 본 연구에서 다룬 명시적 입장(Stance) 분석을 토대로, 향후에는 모델들이 결과적으로 동일한 입장을 취하더라도 실제 텍스트를 생성할 때 선택하는 어휘의 극성(lexical polarity)이나 서술 방식(Framing)에서 어떠한 언어학적 차이를 보이는지를 질적으로 비교하는 연구가 요구된다. 이러한 담화적 층위의 확장 분석은 한국어 언어 모델이 지닌 편향의 작동 기제를 사회언어학적으로 더욱 선명하게 규명하는 데 기여할 것이다.

References

Bang, Y., Chen, D., Lee, N., & Fung, P. (2024). Measuring political bias in large language models: What is said and how it is said. In *Proceedings of the 62nd annual meeting of the*

association for computational linguistics (Volume 1: Long Papers) (pp. 11142-11159). Association for Computational Linguistics.

Cheng, M., Durmus, E., & Jurafsky, D. (2023). Marked personas: Using natural language prompts to measure stereotypes in language models. *arXiv.*

Coeckelbergh, M. (2020). *AI ethics*. MIT Press.

Eom, G., & Kim, D. (2021). Development and application of a comment classifier for online political opinion analysis: Public opinion analysis using KoBERT. *Korean Party Studies Review, 20*(3), 167-191.

Feng, S., Park, C. Y., Liu, Y., & Tsvetkov, Y. (2023). From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st annual meeting of the association for computational linguistics (Volume 1: Long Papers)* (pp. 11737-11762). Association for Computational Linguistics.

Fisher, J., Feng, S., Aron, R., Richardson, T., Choi, Y., Fisher, D. W., Pan, J., Tsvetkov, Y., & Reinecke, K. (2025). Biased LLMs can influence political decision-making. In *Proceedings of the 63rd annual meeting of the association for computational linguistics (Volume 1: Long Papers)* (pp. 6559-6607). Association for Computational Linguistics.

Jo, S. M. (2025, March 30). 1 in 3 internet users used generative AI last year… Usage rate doubled. *Yonhap News*. https://www.yna.co.kr/

Kim, B., Lee, E., & Na, D. (2023). A new Korean text classification benchmark for recognizing the political intents in online newspapers. *arXiv.*

Kim, J., & Kim, H. (2025). Responses of artificial intelligence to unethical directive speech acts: Focusing on indirect directive strategies. *The Sociolinguistic Journal of Korea, 33*(4), 43-80.

Kim, J., Kim, G., Aiyanyo, I. D., & Lim, H. (2022). Measurement of political polarization in Korean language model by quantitative indicator. In *Proceedings of the annual conference on human and language technology* (pp. 16-21).

Manvi, R., Khanna, S., Burke, M., Lobell, D., & Ermon, S. (2024). Large language models are geographically biased. *arXiv preprint arXiv:2402.02680.*

Ministry of Science and ICT. (2026, January 15). *Announcement of the first phase evaluation results of the proprietary AI foundation model project* [Press release]. https://www.msit.go.kr/

Motoki, F., Pinho Neto, V., & Rodrigues, V. (2024). More human than human: Measuring ChatGPT political bias. *Public Choice, 198*(1-2), 3-23. https://doi.org/10.1007/s11127023-01097-2

Noh, K. S., Song, S. H., & Oh, E. J. (n.d.). How language models understand honorific mismatches in Korean. *Language Research, 60*(3), 303-322.

Oh, D. H. (2026, January 24). Following ChatGPT, Korea is a 'big spender' on Gemini 3... 2nd in global paid subscribers after the U.S. *Newsis*. https://www.newsis.com/

Rettenberger, L., Reischl, M., & Schutera, M. (2025). Assessing political bias in large language models. *Journal of Computational Social Science, 8*, 42.

Rozado, D. (2024). The political preferences of LLMs. *PLOS ONE, 19*(7), e0306621.

Seo, J., Cho, S., & Park, J. (2025). Political bias in large language models and its implications on downstream tasks. *Journal of KIISE, 52*(1), 18-28.
Shin, D. K. (2023). A case study on English test item development training for secondary school teachers using AI tools: Focusing on ChatGPT. *Language Research, 59*(1), 21-42.
Taubenfeld, A., Dover, Y., Reichart, R., & Goldstein, A. (2024). Systematic biases in LLM simulations of debates. *arXiv preprint arXiv:2402.04049.*

송종빈
석사과정
과학기술학협동과정
고려대학교
02841 서울시 성북구 안암로 145
E-mail: jbsongai@korea.ac.kr

송상헌
부교수
언어학과
고려대학교
02841 서울시 성북구 안암로 145
E-mail: sanghoun@korea.ac.kr

접수일자 : 2026. 3. 1
수정본 접수 : 2026. 3. 31
게재결정 : 2026. 4. 13