■ 민족문화연구 제112호(2026.08.31.), 고려대학교 민족문화연구원, pp.301-343.
http://doi.org/10.17948/kcs.2026..112.301

KoPolitical-LLM: 한국 정치 연구를 위한 거대언어모델 탐구*

노강산(盧剛山)** · 송상헌(宋相憲)***

목 차

1. 서론
2. 배경
3. KoPolitical-LLM: 거대언어모델 기반 한국 정치 연구
4. 논의
5. 결론

국문초록

본고는 거대언어모델을 한국 정치 연구에 적용하는 한국어 전산정치학의 가능성을 탐색하기 위해 두 가지 구체적인 연구로 구성된 KoPolitical-LLM를 제시하였다. 첫 번째 사례인 KoP-Stance는 더불어민주당과 국민의힘 양당의 공식 대변인 논평 및 브리핑 데이터를 기반으로 구축된 한국 정치 성향 탐지 프레임워크로, 인코더 기반 언어모델의 파인튜닝과 디코더 기반 언어모델의 인컨텍스트 러닝을 통해 한국 정치 텍스트의 당파성을 자동으로 탐지하고 분류하는 것을 목적으로 한다. 두 번째 사례인 KoP AA는 거대언어모델에 양당의 정체성을 부여하여 실제 국회 발의 법안에 대한 각 정당의 입장을 예측하는 정치 행위자 에이전트로, 거대언어모델이 한국 정당의 정치 성향

* 이 논문은 2025년 대한민국 교육부와 한국연구재단의 인문사회분야 중견연구자지원사업의 지원을 받아 수행된 연구임(NRF-2025S1A5A2A01008663).
** 제1저자: 고려대 언어학과 박사수료.
*** 교신저자: 고려대 언어학과 교수.

을 이해하고 현안에 대한 정치적 판단을 시뮬레이션할 수 있는지를 검증하는 것을 목적으로 한다. 본고는 이 두 가지 연구 사례를 통해 영미권을 중심으로 발전해 온 전산정치학의 방법론을 한국 정치 데이터에 적용함으로써 한국어 전산정치학 연구의 가능성을 실증하고자 하였다. 그리고 이를 통해 정치 데이터의 고유성과 파급성에 근거하여 한국어 전산정치학의 당위성을 논의하고 향후 한국전산정치학이 나아갈 방향을 제시하였다.

주제어: 전산정치학, 한국 정치, 거대언어모델, 정치 성향 탐지, 정치 행위자 시뮬레이션

1. 서론

2020년대에 접어들면서 국내외 정치 지형은 급속한 변화의 국면을 맞이하고 있다. 해외에서는 미국이 도널드 트럼프 대통령의 재선 이후 외교적으로 고립주의를 표방하며 국제 분쟁에 대한 개입을 축소하는 한편,[1] 경제적으로는 관세 정책을 통해 기존의 무역 체제에서 이탈함에 따라 냉전 이후의 국제 질서가 흔들리고 있다.[2] 한국에서는 2024년 12·3 비상계엄과 이로 인한 대통령 탄핵 사태 등 일련의 거대한 정치적 사건이 발생하면서 국내 정치 지형과 사회 전반에 큰 여파가 있었다.[3] 주목할 점은 국내외를 막론하고 정치적 변화가 이어지는 가운데 인공지능(Artificial Intelligence)의 영향력이 확대되고 있다는 것이다. 특히 ChatGPT를 비롯한 생성형 인공지능의 등장 이후, 그 파급 효과가 학계와 산업계를 넘어 사회 전반으로 확산된 상황이며,[4] 정치학(Political Science) 또한 이

1) Wojczewski, Thorsten, “Trump, populism, and American foreign policy”, Foreign Policy Analysis 16-3, 2020, pp.292-311.

2) Burrows, Mathew and Braml, Josef, World to Come: The Return of Trump and the End of the Old Order, Brixton Ink, 2025.

3) 임태훈, 「12·3 이후의 청각 정치학 - 광장의 소리와 이론의 탐색」, 『대중서사연구』 31-2, 2025, 463-496면; 장승진, 「비상계엄과 탄핵, 그리고 제21대 대선의 투표선택: 포퓰리즘과 회고적 투표」, 『한국정치연구』 34-3, 2025, 108-134면; 정태욱, 「한국 헌정사에서 대통령 책임과 대통령 탄핵」, 『민주법학』 87, 2025, 71-122면; 천정환, 「거리정치를 통해 본 내란 이후 '극우화'의 문화정치적 자원과 정동적 동력」, 『역사비평』 151, 2025, 136-177면.

4) Barros, Amon, Prasad, Ajnesh and Śliwa, Martyna, “Generative artificial intelligence

러한 흐름에서 예외가 아니다. 해외에서는 거대언어모델을 정치 연구에 접목하여 정치 성향 탐지, 편향 분석, 감성 분석 등을 시도하는 연구가 이루어졌다.[5)]

본고는 거대언어모델의 정치학에의 활용이 영미권과 유럽을 중심으로 하는 해외 학계에서는 활발히 이루어지고 있는 반면에 국내 정치 연구에서의 활용은 상대적으로 초기 단계에 있다는 문제의식에서 출발하였다. 법률이 각국의 법체계에 따라 고유한 특성을 가지듯, 정치적 담론 역시 각국의 정치 구조와 역사적 맥락에 따라 상이한 양상을 보인다. 그럼에도 불구하고 거대언어모델을 비롯한 인공지능과 전산언어학 관련 기술이 적용된 한국 정치 연구는 현재 상대적으로 소수이며, 한국의 특성을 고려한 논의 또한 충분히 수행되지 않았다. 이러한 상황에 주목하여 본고는 해외에서 선도적으

---

and academia: Implication for research, teaching and service", Management Learning 54-5, 2023, pp.597-604.

5) Liu, Ruibo et al., "Quantifying and alleviating political bias in language models", Artificial Intelligence 304, 2022, 103654; Bang, Yejin et al., "Measuring political bias in large language models: What is said and how it is said", Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp.11142-11159; Moghimifar, Farhad et al., "Modelling political coalition negotiations using llm-based agents", arXiv preprint arXiv:2402.11712, 2024; Pavlyshenko, Bohdan M., "Using GPT Models for Qualitative and Quantitative News Analytics in the 2024 US Presidential Election Process", arXiv preprint arXiv:2410.15884, 2024; Pit, Pagnarasmey et al., "Whose side are you on? Investigating the political stance of large language models", arXiv preprint arXiv:2403.13840, 2024; Taubenfeld, Amir et al., "Systematic biases in LLM simulations of debates", Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024, pp.251-267; Arslan, Muhammad, Munawar, Saba and Cruz, Christophe, "Political-RAG: using generative AI to extract political information from media content", Journal of Information Technology & Politics 22-4, 2025, pp.479-494; Farooq, Aqsa et al., "Generative AI Generating Concerns: Citizens' Perspectives During the 2024 European Elections", The International Journal of Press/Politics, 2025; Mouthami, K., Naren, P. and Pranesh, R., "Political sentiment analysis on twitter using deep learning and Llm models", 2025 3rd International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA), IEEE, 2025, pp.1-6; Rettenberger, Luca, Reischl, Markus and Schutera, Mark, "Assessing political bias in large language models", Journal of Computational Social Science 8-2, 2025, p.42.

로 수행된 전산정치학의 방법론을 한국 정치 데이터에 적용함으로써 한국 정치 연구와 거대언어모델의 접점을 탐색하는 것을 목적으로 하였다. 이를 위해 본고는 다음과 같은 연구 문제(Research Question)를 설정하였다.

RQ1: 현행 거대언어모델은 한국 정치 데이터에 내재된 정치적 성향과 특성을 탐지하고 이에 근거하여 정치적 행위자로서 행동할 수 있는가?

RQ2: RQ1의 과제를 위해 어떠한 방식의 데이터셋과 실험 설계가 제시될 수 있는가?

이러한 연구 문제에 답변하고자 본고는 한국 정치 연구와 거대언어모델의 접목 가능성을 실증하는 KoPolitical-LLM을 제시하였다. (1) 정치 성향 탐지를 위한 프레임워크 KoP-Stance와 (2) 정치 행위자 에이전트 KoPAA를 제시하였으며, 이를 통해 한국어 전산정치학의 발전 가능성을 모색하고자 하였다.

2. 배경

1) 거대언어모델

언어모델(Language Model)이란 주어진 문맥을 바탕으로 다음에 등장할 토큰을 예측하는 모델이다.[6] 초기에는 N-gram 모델과 같은 통계 기반 모델이 사용되었으나, 인공신경망의 발전과 함께 대규모 데이터를 학습하는 신경망 기반 모델이 주류를 이루게 되었다.[7] 언어모델은 주어진 토큰 시퀀스를 기반으로 그 다음 토큰을 예측하는 확률 함수로 정의될 수 있으며, 이를 수식으로 나타내면 (1)과 같다.

---

6) Blank, Idan A., "What are large language models supposed to model?", Trends in Cognitive Sciences 27-11, 2023, pp.987-989.

7) Bengio, Yoshua et al., "A neural probabilistic language model", Journal of Machine Learning Research 3, 2003, pp.1137-1155; Mikolov, Tomas et al., "Recurrent neural network based language model", Proceedings of Interspeech 2010, ISCA, 2010, pp.1045-1048.

$$(1)\ P(w) = P(w_1) \cdot P(w_2 \mid w_1) \cdot P(w_3 \mid w_1,\ w_2) \cdots P(w_n \mid w_1,\ \cdots,\ w_{n-1})$$

수식 (1)은 언어모델이 이전 토큰들을 바탕으로 각 토큰의 조건부 확률을 순차적으로 추정함으로써 토큰 시퀀스의 확률을 계산하는 과정을 나타낸다. 이처럼 언어모델의 기본적인 목적은 주어진 문맥으로부터 다음에 등장할 토큰을 적절히 예측하는 것이다.[8)]

언어모델의 성능은 이른바 규모의 법칙(Scaling Law)에 따라 학습 데이터와 파라미터 크기가 증가할수록 향상되는 경향이 있으며,[9)] 이는 곧 거대언어모델(Large Language Model, LLM)의 등장으로 이어졌다. 거대언어모델 관련 연구는 2017년 구글이 발표한 트랜스포머(Transformer)를 기점으로 급속히 발전하였다.[10)] 트랜스포머는 입력 시퀀스를 토큰별 문맥 벡터로 변환하는 인코더(Encoder)와, 이를 바탕으로 출력 시퀀스를 순차적으로 생성하는 디코더(Decoder)로 구성된다. 트랜스포머 기반 거대언어모델은 인코더 기반 모델, 디코더 기반 모델, 그리고 시퀀스 투 시퀀스(Sequence-to-Sequence) 모델로 분류되며, 각각 자연어 이해,[11)] 자연어 생성,[12)] 그리고 번역과 같은 변환 과제에 특화되었다. 거대언어모델에 대한 연구가 지속되면서 학습 데이터의 확장, 벤치마크의 고도화, 인간 피드백 기반 강화학습(Reinforcement Learning from Human Feedback, RLHF)의 도입 등 다양한 측면에서 발전이 이루어졌으며 이에 따라 범용적인 성능을 갖춘 모델들이 등장하였다. 이러한 모델들은 자연어의 이해 및 생

8) Xiao, Tong and Zhu, Jingbo, “Foundations of large language models”, arXiv preprint arXiv: 2501.09223, 2025.
9) Kaplan, Jared et al., “Scaling laws for neural language models”, arXiv preprint arXiv: 2001.08361, 2020.
10) Vaswani, Ashish et al., “Attention is all you need”, Advances in Neural Information Processing Systems 30, 2017, pp.5998-6008.
11) Devlin, Jacob et al., “BERT: Pre-training of deep bidirectional transformers for language understanding”, Proceedings of NAACL-HLT 2019, 2019, pp.4171-4186.
12) Radford, Alec et al., “Language models are unsupervised multitask learners”, Open AI Blog 1-8, 2019, pp.1-9.

성은 물론 논리적·수리적 추론과 같은 고차원적 과제에서도 뛰어난 성능을 발휘하고 있다.[13]

2) 거대언어모델과 정치 연구

(1) 전산정치학

거대언어모델에 대한 논의가 자연어 처리 및 인공지능 연구의 경계를 넘어 자연과학과 인문사회과학 전반으로 확산됨에 따라, 정치학 또한 이러한 변화를 수용하고 있다. 전산정치학(Computational Political Science)은 거대언어모델의 등장 이전부터 영미권을 중심으로 발전해 온 학제적 연구 분야로,[14] 최근 거대언어모델의 급속한 발전과 함께 연구 범위와 방법론이 확장되고 있다. Political-LLM은 거대언어모델을 연구하는 컴퓨터 과학자들과 정치학자들이 협력하여 거대언어모델이 적용된 정치학 연구의 현황을 조사한 프로젝트이다.[15] Political-LLM에서는 전산정치학을 전산방법론과 정치학을 통합하여 정치 체계, 정치 행위와 그 결과를 분석하는 학제적 연구 분야('Computational Political Science is an interdisciplinary field that integrates computational methods with political science to analyze political systems, behaviors, and outcomes')로 정의한다. 그리고 이러한 정의와 함께 전산정치학의 핵심 연구 주제들이 〈표 1〉과 같이 제시되고 있다.

---

13) Chang, Tyler A. and Bergen, Benjamin K., "Language model behavior: A comprehensive survey", Computational Linguistics 50-1, 2024, pp.293-350; Naveed, Humza et al., "A comprehensive overview of large language models", ACM Transactions on Intelligent Systems and Technology 16-5, 2025, pp.1-72.

14) Haq, Ehsan Ul et al., "A survey on computational politics", IEEE Access 8, 2020, pp.197379-197406.

15) Li, Lincan et al., "Political-LLM: Large language models in political science", arXiv preprint arXiv:2412.06864, 2024.

〈표 1〉 전산정치학의 핵심 개념들

개념	상세
정치 데이터 및 텍스트 생성 (Political Data Sources and Text Generation)	정치 연설, 입법 문서, 소셜 미디어 등 다양한 정치 데이터에 대한 분석
선거 예측 및 투표 분석 (Election Prediction and Voting Behavior)	유권자 행동 패턴 및 선거 결과 예측
정책 및 입법 해석 (Policy and Legislative Interpretation)	법률 문서, 법안, 토론 등에 대한 분석
허위정보/가짜뉴스 탐지 (Misinformation/Fake News Detection)	정치 관련 허위 정보 탐지 및 차단
정치적 위기 및 갈등 예측 (Political Risk and Conflict Prediction)	정치적 불안정성 및 국제 분쟁 예측
정치 게임이론 및 협상 (Political Game Theory and Negotiation)	정치적 주체 간의 상호작용 모델링

이러한 핵심 개념들을 바탕으로 KoPolitical-LLM은 정치학 연구에의 거대언어모델 적용에 관한 기존 연구들을 무엇을 연구하는가(What)와 어떻게 연구하는가(How)라는 두 축을 기준으로 체계적으로 분류한다. 이러한 분류는 〈표 2〉에 제시되어 있다.

〈표 2〉 기존 전산정치학 연구 분류

분류	세부 항목
무엇을 연구하는가 (What)	예측 과제(Predictive Tasks)
	생성 과제(Generative Tasks)
	시뮬레이션(Simulation)
	설명 이론(Explanation Theory)
	사회적 영향(Social Impacts)
어떻게 연구하는가 (How)	벤치마크 데이터셋(Benchmark Datasets)
	데이터 처리(Data Processing)
	거대언어모델 파인튜닝(Fine-Tuning LLM)
	제로샷/퓨샷 추론(Zero/Few-Shot Inference)
	기타 추론 기법
	투표 시뮬레이션(Voting Simulation)

전자의 분류는 거대언어모델이 정치 연구에서 수행할 수 있는 역할을 제시하며, 후자의 분류는 이를 실제로 구현하기 위한 기술적 방법론을 제시한다. Political-LLM의 분류는 기존 연구의 성과를 체계적으로 정리하는 동시에 향후 연구에서 거대언어모델을 어떻게 활용할 수 있는지에 대한 방향성을 제공한다.

(2) 한국어 전산정치학

한국어 전산정치학의 현황을 살펴보면 텍스트 마이닝(Text Mining) 등의 자연어 처리 기술을 적용하여 정치인의 발화와 같은 정치 데이터를 대상으로 담화 분석을 시도한 연구 사례들이 존재한다.[16] 이러한 연구들이 한국어 전산정치학 연구의 토대를 마련하였다는 점에서 의의가 있으나, 최근 급속도로 발전하고 있는 거대언어모델을 연구 방법론으로 직접적으로 도입한 사례는 충분히 축적되지 않았다. 한국 정치 연구에서의 거대언어모델 적용 현황을 분석한 최근 연구에 따르면 한국학술지인용색인(Korean Citation Index, KCI)에 수록된 정치외교학 분야 논문 중 2025년 기준 거대언어모델을 주제로 하거나 부분적으로 다룬 논문은 9편에 해당하지만 거대언어모델을 연구 방법론으로 직접 적용한 사례는 찾아보기 어려운 것으로 확인되었다.[17]

기존의 조사가 KCI에 한정되었다는 점을 고려하여 KCI, SSCI, SCIE 등재 학술지, 학술대회(Conference) 및 arXiv를[18] 대상으로 자체적인 추가 조사를 수행한 결과, 2026년 6월을 기준으로 한국 정치를 대상으로 거대언어모델을 직접적으로 적용한 선행 연구 사례가 〈표 3〉과 같이 정리되었다.[19]

---

16) 김현강, 「내러티브 텍스트이자 정치담화로서의 정치광고 분석-한국의 16대, 17대 대통령 선거 광고를 중심으로-」, 『사회언어학』 22-1, 2014, 61-85면; 정다솜·권순희, 「문재인 대통령의 SNS 정치 담화에 대한 텍스트언어학적 분석」, 『텍스트언어학』 49, 2020, 187-225면; 김승래, 「김정은과 남한 대통령 정치 담화의 프레이밍 전략 비교: 형용사를 중심으로」, 『현대정치연구』 18-2, 2025, 165-211면; 한은수, 「텍스트 마이닝으로 본 포퓰리즘 리더십 유형 연구: 도널드 트럼프(Donald Trump)와 이재명 대통령 연설문의 수사적 프레임과 정책 지향성 비교 분석」, 『연구방법논총』 10-3, 2025, 31-67면.

17) 이인복, 「정치학 연구에서의 대형언어모델(LLM): 주제 동향과 사용 현황 분석」, 『국제정치논총』 65-3, 2025, 257-328면.

18) arXiv(https://arxiv.org)는 1991년 창설된 오픈 액세스 프리프린트(open-access preprint) 저장 공간이며, 동료평가(peer review)를 아직 받지 않은 연구들이 PDF, HTML의 형태로 공개되어 있다.

19) 거대언어모델, 한국 정치, LLM, Korean Politics와 같은 핵심어를 활용하여 학술 데이터베이

〈표 3〉 거대언어모델 기반 한국 정치 연구 사례 (2026년 6월 기준)

분류	학술지	저자 및 연도	논문
KCI	어학연구	송종빈·송상헌 (2026)	한국어 언어모델의 정치편향성 측정
	한국HCI학회 논문지	주미선 외 (2026)	CALL: 대규모 언어 모델 기반 온라인 정치 커뮤니티 특성 추출 및 평가 전략
	한국정당학회보	전예성·강신재·박상현 (2026)	LLM을 통한 한국 국회의원의 이념성향 측정: 국회 회의록 텍스트 분석을 중심으로
	정보과학회 논문지	서정연·조석민·박종철 (2025)	거대 언어 모델의 정치적 편향과 하위 작업에서의 영향
	사회적질연구	김정연 (2025)	대규모 언어모델(LLM) 계열 기술의 정치·공공 부문 적용 : 생성형 인공지능의 기대와 한계
SSCI	Journal of East Asian Studies	Jung, Yujin J. (2025)	The Sectarian Divide: The Dynamics of Populism in South Korea
	Media and Communication	Lee, Keyeun et al. (2025)	Ideology and Policy Preferences in Synthetic Data: The Potential of LLMs for Public Opinion Analysis
	Transport Policy	Yun, Hyunsoo and Lee, Eun Hak (2025)	Party politics in transport policy with a large language model
SCIE	Scientific data	Park, Sunkyoung et al. (2025)	A benchmark dataset for evaluating gender sensitivity in Korean political discourse with large language models
학술대회	Empirical Methods in Natural Language Processing (EMNLP)	Lee, Dahyun et al. (2025)	Journalism-Guided Agentic In-Context Learning for News Stance Detection
		Yu, Seunguk et al. (2025)	From Ground Trust to Truth: Disparities in Offensive Language Judgments on Contemporary Korean Political Discourse
arXiv	arXiv Preprint	Kim, Beomjune, Lee, Eunsun and Na, Dongbin (2023)	A New Korean Text Classification Benchmark for Recognizing the Political Intents in Online Newspapers
		Kang, Sungwoo (2026)	Diagnosing Korean-Language LLM Political Bias via Census-Grounded Agent Simulation
		Kim, DongJae et al. (2025)	LASTIST: LArge-Scale Target-Independent STance dataset

스를 대상으로 조사가 이루어졌다. 한국인 연구자에 의해 진행된 연구인 경우에도 한국 정치를 대상으로 하지 않았다면 조사 대상에서 제외하였다. 또한 로동신문과 같은 북한 정치 연구도 제외하였는데, 이는 실질적으로 일당 독재 체제인 북한의 특성상 정당 간 대립 양상의 분석이 불가능하기 때문이다.

앞서 〈표 2〉에서 제시된 Political-LLM의 분류 체계에 따르면, 무엇을 연구하는가(What)의 측면에서 기존 한국어 전산정치학 연구는 설명 이론과 사회적 영향 분석에 집중하고 있으며, 어떻게 연구하는가(How)의 측면에서는 벤치마크 데이터셋의 구축과 데이터 처리에 초점을 맞추고 있다. 본고는 기존 한국어 전산정치학의 이러한 상황을 고려하여, 추가적인 탐색의 여지가 남아 있는 예측, 생성, 시뮬레이션 과제를 중심으로 파인튜닝과 인컨텍스트 러닝 기법을 적용함으로써 한국어 전산정치학의 외연을 확장하고자 하였다.

3. KoPolitical-LLM: 거대언어모델 기반 한국 정치 연구

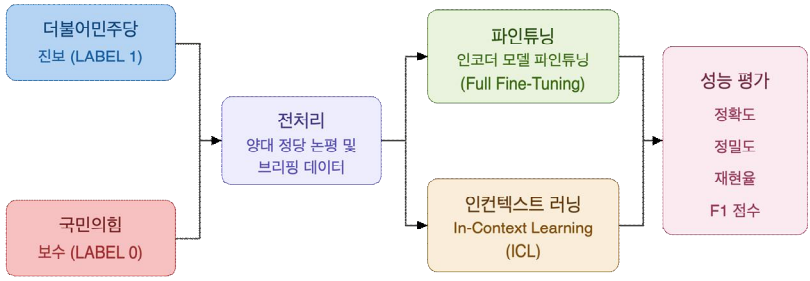
본 장에서는 한국 정치 연구에 거대언어모델을 직접적으로 적용한 연구 사례 KoPolitical-LLM을 제시한다. 첫 번째 사례는 정치 성향 탐지로, 한국어 인코더 기반 언어모델의 파인튜닝과 디코더 기반 언어모델의 인컨텍스트 러닝을 통해 주어진 데이터의 정치 성향을 포착하여 분류하는 과제이다. 두 번째 사례는 정치 행위자 에이전트로, 거대언어모델에 특정 정치 행위자의 정체성과 맥락을 부여하여 해당 행위자의 입장에서 현안을 판단하는 과제이다. 앞서 2.2.1절에서 소개된 Political-LLM의 분류에 따르면 이러한 두 적용 사례 모두 정치 연구를 위해 거대언어모델을 적용한 예측 및 시뮬레이션 과제로 이해될 수 있다. 이 두 사례는 영미권을 중심으로 하는 전산정치학의 방법론을 한국 정치 데이터에 적용함으로써, 한국 정치 연구에서의 거대언어모델 활용 가능성을 검증한 결과물이라는 의의가 있다.

1) KoP-Stance: 한국 정치 성향 탐지 프레임워크

정치 성향 탐지(Political Stance Detection)란 주어진 데이터에서 드러나는 정치적 성향을 파악하고 분류하는 과제이다. 이는 전산정치학에서 가장 기초적이면서도 핵심적인 과제 중 하나로, 정치 데이터에 대한 정량적 분석의 출발점이 된다.[20] 해외에서는 의회 발언, 정당 강령, 뉴스 기사 등 다양한 정치 데이터

를 대상으로 언어모델 기반 정치 성향 분류 연구가 활발하게 수행되었다.[21] 이러한 연구 중에는 P-Stance와 같이 정치적 트윗(Tweet) 데이터를 정제하여 데이터셋을 구축하고, 이를 언어모델의 학습 및 평가에 활용한 사례도 존재한다.[22]

한국 정치는 분단 국가, 압축 성장, 단일 사회 등의 특수성으로 인해 진보와 보수의 이념적 대립이 뚜렷하게 나타난다.[23] 이러한 점에서 한국어 정치 데이터를 대상으로 한 정치 성향 추론은 한국어 전산정치학 연구의 가장 기본적인 토대를 마련하는 과제이다. 본고는 이를 고려하여 한국 정치 성향 탐지를 위한 프레임워크인 KoP-Stance를 구축하였다. 이 과정에서 거대언어모델의 정치 성향 탐지를 포괄적으로 수행하기 위해 인코더 기반 언어모델과 디코더 기반 언어모델을 모두 사용하였다. KoP-Stance의 구체적인 구조는 〈그림 1〉과 같다.



〈그림 1〉 KoP-Stance: 한국 정치 성향 탐지 프레임워크

20) Hardalov, Momchil et al., "A survey on stance detection for mis-and disinformation identification", Findings of the Association for Computational Linguistics: NAACL 2022, 2022, pp.1259-1277; Burnham, Michael, "Stance detection: A practical guide to classifying political beliefs in text", Political Science Research and Methods 13-3, 2024, pp.611-628.

21) Liu, Yujian et al., "POLITICS: Pretraining with same-story article comparison for ideology prediction and stance detection", Findings of the Association for Computational Linguistics: NAACL 2022, 2022, pp.1354-1374.

22) Li, Yingjie et al., "P-stance: A large dataset for stance detection in political domain", Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021, pp.2355-2365.

23) 김영명, 「한국 정치의 특수성에 관한 연구 서설: 분단, 압축성장, 단일사회 문화」, 『비교민주주의연구』 7-1, 2011, 5-35면.

KoP-Stance는 한국 현실 정치를 대표하는 양대 정당의 논평 및 브리핑 데이터를 전처리하고, 인코더 및 디코더 기반 언어모델을 통해 정치 성향을 탐지하는 방식으로 구성된다. 다음의 3.1.1절에서는 파인튜닝을 위한 데이터셋 구축 과정을, 3.1.2절과 3.1.3절에서는 각각 인코더 기반 언어모델과 디코더 기반 언어모델에 기반한 KoP-Stance의 구축 과정을 소개한다.

(1) 데이터셋 구축

파인튜닝(Fine-Tuning)이란 대규모 말뭉치를 통해 사전 학습을 거친 언어모델이 특정 도메인의 과제를 수행할 수 있도록 해당 도메인의 데이터를 활용하여 모델의 파라미터를 업데이트하는 과정으로, 전문적이고 특수한 도메인에 특화된 과제 수행을 가능하게 한다. 본 절에서는 인코더 기반 한국어 언어모델에 전체 파라미터를 업데이트하는 풀 파인튜닝(Full Fine-Tuning) 기법을 통해 정치 성향을 탐지하고자 하였다.

파인튜닝을 위한 학습 및 평가 데이터로는 한국 현실 정치의 특성을 반영하는 데이터를 선정하였다. 이를 위해 한국 현실 정치에서 막대한 지분을 차지하는 거대 양당인 더불어민주당(Democratic Party of Korea, DPK)과 국민의힘(People Power Party, PPP)의 데이터를 사용하였다. 2026년을 기준으로 더불어민주당은 한국 현실 정치 지형에서 진보 진영을 대표하는 거대 정당이며 국민의힘은 보수 진영을 대표하는 거대 정당이다.[24] 구체적으로 본고에서 더불어민주당과 국민의힘 양당을 선정한 이유는 다음과 같다. 첫째, 양당 외에도 다수의 원내·원외 정당이 존재하나, 두 정당이 한국 현실 정치에서 가장 큰 영향력을 지니며 대중적 관점에서도 각각 진보와 보수 진영을 대표한다는 점을 고려하였다. 둘째, 양당을 제외한 군소 정당들을 포함할 경우 정당 간 입장 및 정체성의 차이로 인

24) Kim, Yeaji, “The development of party organizations in new democracies: Evidence from South Korea”, in Poguntke, Thomas and Hofmeister, Wilhelm (eds.), Political Parties and the Crisis of Democracy: Organization, Resilience, and Reform, Oxford University Press, 2024, pp.401-426.

해 진보와 보수라는 이진 분류(Binary Classification)를 적용하기 어렵다는 점을 고려하였다.[25] 셋째, 군소 정당들의 경우 상대적으로 짧은 정당의 역사와 축적된 자료의 부족으로 인해 언어모델의 학습 및 평가에 필수적인 충분한 양의 데이터를 확보하기 어렵다는 점을 고려하였다. 데이터의 양과 균질성은 언어모델의 학습과 평가에 있어 중요한 요소이므로 이를 담보할 수 있는 더불어민주당과 국민의힘 양대 정당을 최종적으로 선정하였다.[26]

이렇게 선정된 더불어민주당과 국민의힘 양대 정당을 대상으로 추출할 데이터로는 양당 대변인의 논평 및 브리핑 자료를 선정하였다. 대변인 논평 및 브리핑 자료를 선정한 이유는 다음과 같다. 첫째, 대변인은 양당의 공식 입장이나 견해를 전달하므로 대변인을 통해 발표되는 논평 및 브리핑 자료는 실질적으로 양당의 정치적 특성을 드러내는 가장 대표적인 데이터이다. 즉, 어느 개인의 단편적인 정치적 견해가 아니라 정당의 공식적인 입장이므로 언어모델의 한국 현실 정치에 대한 학습 및 평가를 위한 용도에 가장 잘 부합하는 데이터이다. 둘째, 질적인 측면에서 보았을 때, 시의성을 가진다. 현재 한국 현실 정치에서 다루어지고 있는 의제 및 안건에 대한 직접적인 언급이 이루어지며 일반적으로 동일한 주제에 대하여 양당이 상충하는 견해를 보이므로 입장 차이를 여과 없이 반영한다. 셋째, 양적인 측면에서 보았을 때, 대변인의 논평 및 브리핑은 하루 단위로 지속적으로 발표되기 때문에 충분하게 확보할 수 있다는 점이 고려되었다.

---

25) 물론 진보와 보수의 단순 이진 분류는 다음과 같은 한계를 지닌다. 첫째, 더불어민주당과 국민의힘을 한국 정치에서 각각 진보와 보수를 대표하는 정당으로 간주하여 이진 분류 체계를 적용하였으나, 실제 정치 지형은 복잡한 이념적 스펙트럼을 내포하므로 이진 분류로 이를 완전히 포착하기 어렵다. 둘째, 본고의 진보·보수 분류는 정치철학적 이론이 아니라 한국 현실 정치에 대한 대중적 인식에 기반하고 있어, 관점에 따라 본고의 분류 체계에 대한 이견이 제기될 수 있다. 본고는 연구 설계상의 한계를 인정하면서도, 거대언어모델이 한국 정치 연구에 활용될 수 있는 구체적인 실례를 제공한다는 점에서 그 의의를 찾을 수 있다고 본다.

26) 이러한 점은 전산정치학 연구가 가장 활발하게 이루어진 미국 정치의 경우에도 마찬가지인데, 민주당과 공화당 외에 군소 정당이 존재함에도 불구하고 편의상 민주당을 진보, 공화당을 보수로 설정하여 연구가 수행되는 것이 일반적이다.

본고에서는 더불어민주당과 국민의힘 양당의 공식 홈페이지에서 2017년 5월 9일부터 2026년 5월 31일까지 약 9년간 발표된 대변인 논평 및 브리핑 데이터를 Requests와 BeautifulSoup 라이브러리를 활용하여 크롤링하였다. 수집된 데이터에 대해서는 대변인 소개, 소제목, 일시 및 장소 안내 문구, 중복 문장 등을 제외하는 정제 작업과 함께 어절 수 5개 미만 또는 30개 초과인 문장을 제거하는 길이 기반 필터링을 적용하였다. 나아가 길이 기반 필터링만으로는 양당 공통 어휘가 포함된 당파성이 약한 문장을 걸러내기 어렵다는 한계를 보완하기 위해 Kiwi 형태소 분석기를 통해 내용어를 추출하였다.[27] 그리고 이를 기반으로 로그 오즈비(Log Odds Ratio)를 산출하여 각 정당에서 50회 미만으로 출현한 저빈도 어휘를 제외한 후 변별력 점수 하위 30%에 해당하는 문장을 추가로 제외하였다. 마지막으로 담화 구조를 고려하여 동일 논평 내 연속된 문장들을 2~3개씩 묶어 하나의 청크(Chunk)로 구성하였다.

이와 같은 전처리 과정을 통해 최종적으로 더불어민주당 78,235개, 국민의힘 46,173개의 청크가 확보되었으며, 약 1.7:1의 양적 불균형이 존재함을 확인하였다. 이러한 불균형을 제거하고 연도별 편향을 완화하기 위해 연도별 층화 추출(Stratified Sampling) 방식을 채택하여 양당 각 15,000개, 도합 30,000개의 청크로 구성된 데이터셋을 구성하였다. 이후 2017년 5월 9일부터 2024년까지의 데이터를 학습 데이터(24,000개), 2025년 데이터를 검증 데이터(3,000개), 2026년 1월 1일부터 5월 31일까지의 데이터를 평가 데이터(3,000개)로 분리하였다. 데이터셋의 최종 구성은 〈표 4〉와 같다.

〈표 4〉 파인튜닝을 위한 데이터셋 구성

데이터셋	기간	더불어민주당	국민의힘	합계
학습	2017. 5. 9. - 2024. 12.	12,000	12,000	24,000
검증	2025. 1. 1. - 2025. 12. 30.	1,500	1,500	3,000
평가	2026. 1. 1. - 2026. 5. 31.	1,500	1,500	3,000

27) 이민철, 「Kiwi: 통계적 언어 모델과 Skip-Bigram을 이용한 한국어 형태소 분석기 구현」, 『디지털인문학』 1-1, 2024, 109-136면.

다음으로 수합된 양당의 논평 및 브리핑 데이터에 대하여 레이블링(Labeling)이 이루어졌다. 보수에 해당하는 국민의힘 데이터는 숫자 0으로, 진보에 해당하는 더불어민주당의 논평 및 브리핑 데이터는 숫자 1로 레이블링되었다. 예문 (2)는 언어모델의 파인튜닝에 활용된 데이터의 예시이다.

(2) a. {"SENTENCE": "노동시장의 경직성과 기업을 옥죄는 규제는 그대로 둔 채, 소위 노란봉투법과 같은 반기업 입법과 단기적 재정 처방에 매달리고 있습니다. 지금 필요한 것은 무분별한 '돈 풀기'가 아니라 경제가 다시 뛰게 만드는 조건을 회복하는 일입니다. 기업이 투자할 수 있는 환경을 만들고, 미래 산업에 인재가 유입되는 구조를 만들어가야 합니다.", "LABEL": 0, "STATEMENT_ID": "PPP_000978", "CHUNK_ID": "PPP_000978_C04"}

b. {"SENTENCE": "수많은 노동자의 눈물과 희생이 스며든 노란봉투법이 마침내 시행되었습니다. 무분별한 손해배상과 가압류에 짓눌려 생존권을 위협받던 하청 노동자들의 한을 풀게 된 이번 시행은, 노동 시장의 이중구조를 해소하고 약자를 보호하겠다는 이재명 정부의 흔들림 없는 원칙과 결단이 낳은 성과입니다. 국민의힘에서는 이 법을 파업 조장법이라 왜곡하고 있습니다.", "LABEL": 1, "STATEMENT_ID": "DPK_000829", "CHUNK_ID": "DPK_000829_C02"}

(2) 인코더 기반 언어모델을 통한 정치 성향 탐지

본고에서는 대표적인 인코더 기반 한국어 언어모델에 해당하는 KR-BERT,[28] KLUE-RoBERTa,[29] KoELECTRA[30]의 세 모델을 선정하였다. KLUE-RoBERTa와 KoELECTRA의 경우 base와 large 두 수준의 모델이 제공되는데, 본고에서는 base 모델을 사용하였다. 모델의 세부 사항은 〈표 5〉와 같다.

---

28) Lee, Sangah et al., "KR-BERT: A small-scale Korean-specific language model", arXiv preprint arXiv:2008.03979, 2020.

29) Park, Sungjoon et al., "KLUE: Korean language understanding evaluation", arXiv preprint arXiv:2105.09680, 2021.

30) Park, Jangwon, 「KoELECTRA: Pretrained ELECTRA Model for Korean」, GitHub, 2020, https://github.com/monologg/KoELECTRA(2026.06.25. 검색).

〈표 5〉 인코더 기반 한국어 언어모델 세부 사항

모델	Model ID	파라미터
KR-BERT	snunlp/KR-BERT-char16424	99M
KLUE-RoBERTa-base	klue/roberta-base	111M
KoELECTRA-base	monologg/koelectra-base-v3-discriminator	110M

파인튜닝은 Google Colaboratory 환경에서 런타임 유형을 A100 GPU로 설정하여 수행되었다. 하이퍼파라미터 설정은 〈표 6〉과 같다.

〈표 6〉 파인튜닝을 위한 하이퍼파라미터 설정

하이퍼파라미터	설정
학습률(Learning Rate)	3e-5
배치 사이즈(Batch Size)	16
최대 에포크 수(Max Epochs)	10
얼리 스타핑 기준 에포크(Early Stopping Patience)	3
최대 시퀀스 길이(Max Sequence Length)	256
웜업 스텝 수(Warmup Steps)	500
가중치 감쇠(Weight Decay)	0.01
최적화 알고리즘(Optimizer)	AdamW
최적 모델 선정 기준(Best Model Selection Criterion)	F1 점수

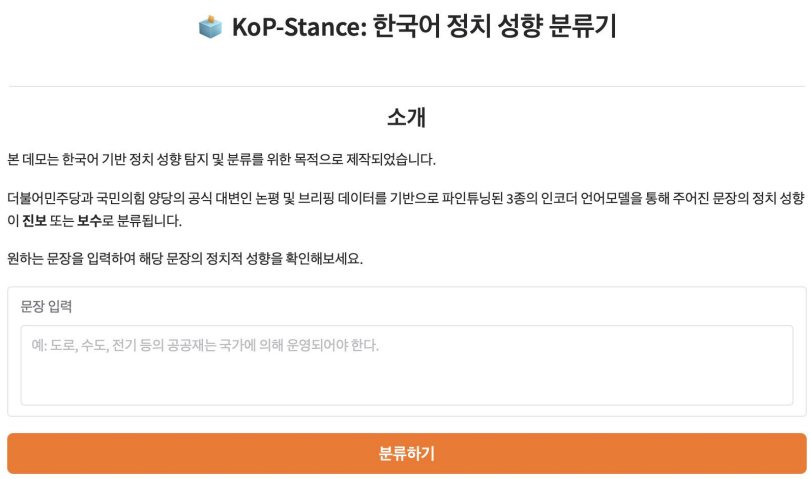
본고에서는 검증 데이터에 대한 성능이 일정 에포크 동안 개선되지 않을 경우 학습을 자동 중단하여 과적합(Overfitting)을 방지하고 최적 성능 시점의 모델을 보존하는 조기 종료(Early Stopping) 기법을 적용하였다. 평가 지표로는 정확도, 정밀도, 재현율, F1 점수가 사용되었다. 파인튜닝 과정에서 F1 점수를 기준으로 3 에포크 동안 연속적인 개선이 없을 경우에 학습을 조기 종료하였으며, 검증 데이터 기준으로 F1 점수가 가장 높은 모델을 최적의 모델(Best Model)로 선정하였다. 〈표 7〉은 인코더 기반 한국어 언어모델을 대상으로 한 파인튜닝의 결과이다.

〈표 7〉 인코더 기반 한국어 언어모델 파인튜닝 결과 (평가 데이터 기준)

모델	정확도	정밀도	재현율	F1 점수
KR-BERT	0.8307	0.8330	0.8307	0.8304
KLUE-RoBERTa-base	0.8610	0.8619	0.8610	0.8609
KoELECTRA-base	0.8580	0.8588	0.8580	0.8579

세 모델 모두 F1 점수 0.83 이상을 기록하였으며, 모델 간 성능 격차는 약 3% 포인트 이내로 일관된 성능을 보였다.

본고에서는 파인튜닝된 3종의 모델을 허깅페이스(Hugging Face)에 업로드하여 한국어 기반 정치 성향 분류기를 직접 사용할 수 있도록 공개하였다.[31] 인터페이스는 〈그림 2〉와 같다.



〈그림 2〉 한국어 기반 정치 성향 분류기 인터페이스

분류기에 문장이 입력되면 3종의 모델을 통해 해당 문장의 정치적 성향에 대한 탐지가 이루어진다. 그리고 해당 문장의 히트맵(Heatmap)이 제공된다.[32] 예컨대, 다음의 문장들을 분류기에 입력하는 경우를 고려해 보자. 다

31) 「KoP-Stance: 한국어 정치 성향 분류기」, HuggingFace, https://huggingface.co/spaces/San0727/KoP-Stance(2026.06.25. 검색).

음의 (3)과 (4)에 제시된 예문들은 대북 정책과 부동산 정책에 대한 더불어민주당과 국민의힘의 공식 논평의 일부이다. 〈표 8〉은 (3)과 (4)의 예문들에 대한 정치 성향 분류기의 분류 결과이다.

(3) a. 북한이 닫아건 문을 열고 평화공존을 위한 대화의 장에 나오기를 촉구합니다. 한반도의 운명은 결국 당사자인 남북이 마주 앉아야 풀립니다. (더불어민주당 공보국 2026. 6. 15.)
b. 적대 행위를 계속하는 북한의 현실을 외면한 채 희망적인 구호와 유화적 제스처만으로는 국민의 생명과 안전을 지킬 수 없습니다. (국민의힘 대변인 2026. 6. 23.)

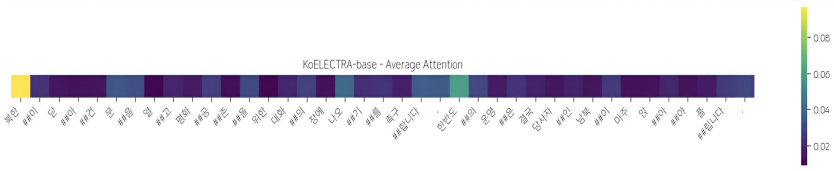
(4) a. 지금 필요한 것은 투기를 조장하고 보호하는 정치가 아니라 주거 안정을 위한 책임 있는 정치, 국가의 미래를 위한 협력입니다. (더불어민주당 공보국 2026. 6. 9.)
b. 대출 규제로 실수요자의 자금 조달을 어렵게 만들고, 서울과 수도권 주요 지역을 '토지거래허가구역'으로 묶어 현금 부자만 집을 살 수 있게 만들었습니다. (국민의힘 중앙선대위 공보단장 2026. 6. 1.)

〈표 8〉 정치 성향 분류 결과 예시

예문	모델	분류 결과	보수 확률	진보 확률
(3a)	KR-BERT	진보	0.0224	0.9776
	KLUE-RoBERTa-base	진보	0.0033	0.9967
	KoELECTRA-base	진보	0.004	0.996
(3b)	KR-BERT	보수	0.8759	0.1241
	KLUE-RoBERTa-base	보수	0.9993	0.0007
	KoELECTRA-base	보수	0.9999	0.0001
(4a)	KR-BERT	진보	0.0287	0.9713
	KLUE-RoBERTa-base	진보	0.0008	0.9992
	KoELECTRA-base	진보	0.0007	0.9993
(4b)	KR-BERT	보수	0.992	0.008
	KLUE-RoBERTa-base	보수	0.999	0.001
	KoELECTRA-base	보수	0.9998	0.0002

32) 히트맵은 문장에 대한 모델의 추론에 있어 어떤 토큰이 중요하게 작용하였는지를 시각화한 것으로 정치적 성향 판단에 있어 영향을 준 요인을 분석하는데 활용될 수 있다.

예문 (3a)와 (4a)의 경우, 모두 진보로 분류되었는데 해당 예문들은 더불어민주당의 견해에 해당하므로 3종의 모델 모두가 적합한 판단을 내렸음을 알 수 있다. 예문 (3b)와 (4b)의 경우, 모두 보수로 분류되었는데 해당 예문들은 국민의힘의 견해이므로 3종의 모델 모두가 적합한 판단을 내렸다고 볼 수 있다.〈그림 3〉은 KLUE-RoBERTa 모델이 산출한 예문 (3a)의 히트맵이다. 문장을 구성하는 단어 중에서 '북한', '한반도' 등의 단어가 모델의 판단에 중요한 영향을 미쳤음을 알 수 있다.



〈그림 3〉히트맵 분석 결과 예시

(3) 디코더 기반 언어모델을 통한 정치 성향 탐지

본 절에서는 3.1.2절의 인코더 기반 언어모델의 정치 성향 탐지 실험과의 비교를 위해 디코더 기반 언어모델에 인컨텍스트 러닝(In-Context Learning)을 적용하여 정치 성향 탐지를 수행하였다. 인컨텍스트 러닝은 모델의 파라미터를 업데이트하지 않고, 소수의 예시를 제공함으로써 모델이 주어진 과제를 수행하도록 하는 방식으로, 추가적인 학습 없이도 다양한 과제에 적용 가능하다는 장점이 있다.[33] 인컨텍스트 러닝을 위한 프롬프트는 〈표 9〉와 같다.

33) Brown, Tom et al., "Language models are few-shot learners", Advances in Neural Information Processing Systems 33, 2020, pp.1877-1901.

〈표 9〉 인컨텍스트 러닝을 위한 프롬프트

프롬프트
[System] SYSTEM_PROMPT = ("한국어 문장의 정치적 성향을 '진보' 또는 '보수' 중 하나로만 분류하는 과제를 수행하시오." "반드시 한국어로 '진보' 또는 '보수' 둘 중 하나의 단어만 출력하시오." "본 과제와 무관한 텍스트를 출력하지 마시오.") [User] 문장: {입력 문장} 정치적 성향 (진보/보수 중 하나만):

다음으로 프롬프트와 함께 제공되는 예시의 경우, 0-shot과 6-shot 조건을 설정하여 퓨샷 데이터의 유무에 따른 성능 변화를 비교하고자 하였다. 6-shot 조건에 사용되는 퓨샷 데이터는 노동 정책, 대북 정책, 부동산 정책의 세 가지 영역을 기준으로 〈표 4〉에서 제시된 학습(Train) 데이터에서 두 정당 간의 대립이 명확하게 드러나는 청크를 선정하여 보수 성향 청크 3개와 진보 성향 청크 3개로 구성하였다. 퓨샷 데이터의 예시는 (5)에 제시되어 있다.

(5) a. {"SENTENCE": "25번의 규제 덩어리 부동산 정책과 임대차 3법이 낳은 처참한 결과 아닌가. 대출 없이는 집을 살 수조차 없게 만든 정부가 이제는 내 집 마련을 포기하라는 선언을 한 셈이다. '장래 소득' 희망고문으로 국민을 우롱하지 말라. 시장의 목소리를 들어야 한다.", "LABEL": 0, "STATEMENT_ID": "PPP_009882", "CHUNK_ID": "PPP_009882_C04"}

b. {"SENTENCE": "부동산 안정화를 위한 법안을 두고 '국민에 대한 화풀이'라는 억지 논리는 그만 두십시오. 더불어민주당은 정부와 함께 부동산 안정화를 위한 정책을 지속적으로 추진하겠습니다. 투기 세력을 근절하고 주택시장을 안정시켜 '주택 = 주거'의 개념을 정착시키겠습니다.", "LABEL": 1, "STATEMENT_ID": "DPK_015861", "CHUNK_ID": "DPK_015861_C11"}

평가 데이터의 경우, 〈표 4〉의 평가(Test) 데이터 3,000개 청크를 그대로 사용하였다.

인컨텍스트 러닝을 적용할 모델의 경우, 대표적인 디코더 기반 한국어 언어모델에 해당하는 SOLAR 계열의 모델과[34] Kanana 계열의 모델이[35] 선정되었다. 두 모델군은 Google Colaboratory 환경에서 추론을 수행하기에 적합한 규모를 갖추었으며, 한국어 특화 사전학습을 통해 한국어 이해 능력이 검증되었다. 모델들의 세부 사항은 〈표 10〉과 같다.

〈표 10〉 디코더 기반 한국어 언어모델 세부 사항

계열	모델	Model ID	파라미터
Kanana	Kanana-1.5-15.7B	kakaocorp/kanana-1.5-15.7b-a3b-instruct	15.7B
	Kanana-1.5-8B	kakaocorp/kanana-1.5-8b-instruct-2505	8B
SOLAR	SOLAR-10.7B	upstage/SOLAR-10.7B-Instruct-v1.0	10.7B

인컨텍스트 러닝은 Google Colaboratory 환경에서 런타임 유형을 A100 GPU로 설정하였다. 인컨텍스트 러닝 과정에서 온도(Temperature)는 0.1로 설정하여 출력의 일관성을 확보하고자 하였으며, 최대 생성 토큰(Max New Tokens)의 수는 16으로 설정하여 모델이 '진보' 또는 '보수' 레이블 외의 불필요한 텍스트를 생성하지 않도록 하였다. 다음의 〈표 11〉은 본고에서 디코더 기반 한국어 언어모델을 대상으로 수행한 인컨텍스트 러닝의 결과이다.

〈표 11〉 디코더 기반 한국어 언어모델 인컨텍스트 러닝 결과

조건	모델	정확도	정밀도	재현율	F1 점수
0-shot	Kanana-1.5-15.7B	0.5603	0.5715	0.5603	**0.5425**
	Kanana-1.5-8B	0.7297	0.7428	0.7297	**0.7260**
	SOLAR-10.7B	0.6531	0.6554	0.6535	**0.6521**
6-shot	Kanana-1.5-15.7B	0.7403	0.7582	0.7403	**0.7358**
	Kanana-1.5-8B	0.7803	0.7918	0.7803	**0.7782**
	SOLAR-10.7B	0.7293	0.7619	0.7284	**0.7201**

34) Kim, Sanghoon et al., "SOLAR 10.7B: Scaling large language models with simple yet effective depth up-scaling", arXiv preprint arXiv:2312.15166, 2023.

35) Kanana LLM Team, "Kanana: Compute-efficient bilingual language models", arXiv preprint arXiv:2502.18934, 2025.

퓨샷 예시 제공 여부에 따라 모든 모델에서 성능 향상이 관찰되었으며, 특히 Kanana-1.5-15.7B는 0-shot(F1 0.5425) 대비 6-shot(F1 0.7358)에서 약 19% 포인트가 향상되었다. 6-shot 조건에서는 Kanana-1.5-8B가 F1 0.7782로 가장 우수한 성능을 기록하였다. 한편 인코더 기반 파인튜닝 결과(F1 0.83~0.86)와 비교하면 디코더 인컨텍스트 러닝의 최고 성능(F1 0.7782)은 약 8~10% 포인트 낮은 수준으로, 추가적인 파인튜닝 없이 퓨샷 예시만으로는 인코더 기반 파인튜닝 수준의 성능에 도달하기 어렵다는 점이 확인되었다.

2) KoPAA: 한국어 정치 행위자 에이전트

에이전트(Agent)란 주어진 환경에서 목표를 달성하기 위해 자율적으로 행동하는 시스템이다.[36] 거대언어모델 기반 에이전트(LLM-based Agent)는 거대언어모델에 기반하여 특정 페르소나(Persona)와 맥락이 제공된 상황에서 자율적으로 행동하는 시스템이다.[37] 해외에서는 미국 의회 입법 시뮬레이션,[38] 미국 대선 시뮬레이션[39] 등과 같이 정치 행위자 에이전트(Political Actor Agent, PAA)에 대한 연구가 수행된 바 있다.

본고에서는 디코더 기반 한국어 언어모델에 더불어민주당과 국민의힘 양당의 정체성을 부여하여 특정 현안에 대한 각 정당의 입장을 예측하고 평가하는 정당 에이전트(Party Agent)를 구축하였다.[40] KoPAA의 구체적인 구조는 〈그림 4〉와 같다.

---

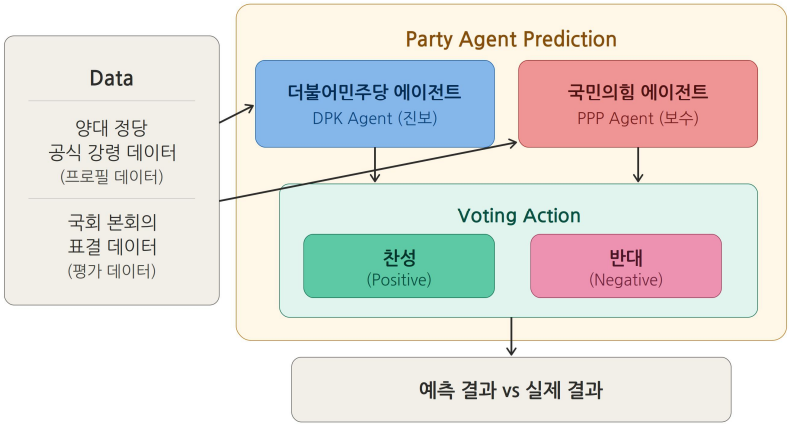
36) De Marchi, Scott and Page, Scott E., "Agent-based models", Annual Review of Political Science 17-1, 2014, pp.1-20.

37) Luo, Junyu et al., "Large language model agent: A survey on methodology, applications and challenges", arXiv preprint arXiv:2503.21460, 2025.

38) Li, Hao, Gong, Ruoyuan and Jiang, Hao, "Political actor agent: Simulating legislative system for roll call votes prediction with large language models", Proceedings of the AAAI Conference on Artificial Intelligence 39-1, 2025, pp.388-396.

39) Zhou, Lingfeng, et al., "FlockVote: LLM-empowered agent-based modeling for simulating U.S. presidential elections", arXiv preprint arXiv:2512.05982, 2025.

40) 본래 Li, Hao, Gong, Ruoyuan and Jiang, Hao (2025) 연구에서 구축한 Political Actor Agent는 미국 하원의원의 과거 투표 기록에 기반하여 개별 의원에 대한 시뮬레이션을 수행한다. KoPAA는 개별 의원이 아닌 정당을 시뮬레이션의 단위로 설정하였다. 이는 한국의 정치적·사회적 맥락에서 개별 정치인을 에이전트로 설정할 경우에 제기될 수 있는 명예훼손 및 가짜 뉴스 등의 문제를 고려한 것이다.



〈그림 4〉 KoPAA: 한국어 정치 행위자 에이전트

에이전트를 위한 데이터의 경우, 양대 정당의 프로필(Profile) 정보를 각 정당의 정체성을 부여하기 위한 데이터로, 입장 예측의 대상이 되는 정치적 안건을 평가 데이터로 각각 활용하였다. 정당 에이전트를 활용한 예측 단계에서는 각 정당의 프로필 정보와 평가 데이터를 입력받아 진보적 견해를 산출하는 더불어민주당 에이전트(DPK Agent)와 보수적 견해를 산출하는 국민의힘 에이전트(PPP Agent)가 구성되었다. 각 에이전트는 정치적 안건에 대한 해당 정당의 입장을 찬성(Positive)과 반대(Negative)의 두 가지 방식으로 표명한다. 최종적으로 각 에이전트의 예측 결과를 실제 결과와 비교하여 정확도 및 F1 점수를 산출함으로써 모형의 성능을 평가하였다.

우선 각 정당의 정체성을 정의하는 프로필 정보로는 더불어민주당과 국민의힘의 공식 강령(2026년 6월 기준)을 활용하였다.[41] 각 당의 강령 전문에 명시된 핵심 가치, 정책 기조, 대북·노동·경제·복지 분야별 입장을 클로드(Claude) Sonnet 4.6 모델을 통해 요약하여 정당 프로필로 구성하였으며, 해당 내용은 〈표 12〉에 제시되어 있다.[42]

41) 프로필 정보를 생성하기 위해 양당의 공식 자료를 사용하였다. 구체적으로는 「강령·당헌·당규」, 더불어민주당, https://theminjoo.kr/main/sub/introduce/rule.php(2026.06.25. 검색); 「강령·당헌·당규」, 국민의힘, https://www.peoplepowerparty.kr/about/constitution(2026.06.25. 검색)의 페이지에서 PDF 파일을 다운로드하여 모델의 요약 과제를 위한 입력 데이터로 사용하였다.

〈표 12〉 KoPAA를 위한 양당 프로필 정보

정당	프로필
더불어민주당 (DPK)	더불어민주당은 4·19혁명, 5·18민주화운동, 촛불시민혁명의 민주주의 정신을 계승하며 서민과 중산층을 대변한다. 핵심 가치로 '공정·생명·포용·번영·평화'를 내세우며, 경제 분야에서는 경제민주화 완성과 혁신성장을 통한 지속가능한 시장경제를 지향한다. 노동 분야에서는 동일가치노동 동일임금 원칙과 전국민고용보험 실현을 추구하며, 복지 분야에서는 소득·건강·주거를 보장하는 보편적 복지국가 건설을 목표로 한다. 대북·외교 분야에서는 한반도 완전한 비핵화와 평화체제 구축, 굳건한 한미동맹을 바탕으로 한 국익 중심 실용외교를 추구한다. 기후위기 대응을 위한 탄소중립 실현, 성평등 민주주의 구현, 교육의 공공성 강화도 주요 정책 기조로 삼는다.
국민의힘 (PPP)	국민의힘은 자유민주주의와 공정·기회균등을 핵심 가치로 삼으며, 3·1운동과 대한민국임시정부의 정통성 및 민주화 운동 정신을 계승한다. 경제 분야에서는 규제 혁신과 4차 산업혁명 대응, 스타트업 육성 등 시장 중심의 성장을 추구하고, 노동 분야에서는 유연한 노동시장 조성과 양질의 일자리 창출, 노사 법치주의 확립을 지향한다. 복지 분야에서는 개인 중심의 맞춤형 복지와 사회적 약자 보호를 통한 지속가능한 복지국가를 목표로 한다. 대북·외교 분야에서는 한미동맹 강화와 북한의 완전한 비핵화를 전제로 한 자유민주주의적 평화통일을 추구하며, 힘에 기반한 안보태세를 견지한다.

이렇게 정리된 정당 프로필 정보는 KoPAA 실험을 위해 더불어민주당 에이전트와 국민의힘 에이전트에게 각각 제공되었다.

다음으로 평가 데이터는 21대(2020년 5월 30일 ~ 2024년 5월 29일) 및 22대(2024년 5월 30일 ~ 2028년 5월 29일) 국회에 발의된 법률안 및 결의안 중 양당의 공식 입장이 명확하게 대립하는 안건을 선별하여 구축하였다. 안건 선정에 있어서는 한국 현실 정치에서 양당이 일관되게 대립적 입장을 견지해 온 주요 정책 의제를 기준으로 삼았다. 구체적으로 조세 부담 완화 및 부동산

42) 생성형 인공지능을 활용하여 요약을 수행한 것은 연구자 본인의 주관적 개입을 최소화하기 위한 조치였다. 물론 거대언어모델의 정치적 편향성에 관한 선행 연구들에서 알 수 있듯이 인공지능 역시 완전한 정치적 중립성을 보장하지는 않는다. 그러나 통제된 환경에서 주어진 데이터만을 기반으로 요약하도록 설계하였으므로, 프로필 정보는 두 정당에 대한 대중적 인식의 범위를 벗어나지 않는 수준에서 정리되었다고 판단된다. 프롬프트는 다음과 같다. "위 문서는 {party_name}의 공식 강령입니다. 강령 전문을 바탕으로 {party_name}의 핵심 가치, 정치 이념, 노동·경제·복지·대북·외교 분야별 정책 기조를 300자 내외의 단일 문단으로 요약하시오."

규제 개혁은 국민의힘이 지속적으로 추진해 온 반면 더불어민주당은 반대 입장을 취해 온 사안이며, 노동기본권 강화 및 자주국방 실현은 더불어민주당이 주도적으로 추진하고 국민의힘이 반대해 온 사안이다. 또한 대북·안보 정책과 관련하여 양당은 대남대북 전단 처리 방향을 둘러싸고 상반된 입장을 보여 왔다. 이러한 현실 정치 지형에 기반하여 국민의힘이 찬성하고 더불어민주당이 반대하는 안건 3건과 더불어민주당이 찬성하고 국민의힘이 반대하는 안건 3건을 각각 선정하고 해당 안건의 실제 제안이유를 수집하여 현안 텍스트로 활용하였다. 이 과정에서 국회 의안정보시스템(LIKMS)을 자원으로 활용하였다.[43] 이를 통해 총 6건으로 구성된 균형 잡힌 평가 데이터셋이 구축되었다. 안건의 목록은 〈표 13〉과 같으며 구체적인 예시는 (6)에 제시되어 있다.

〈표 13〉 KoPAA 평가 데이터 안건 목록 (총 6건)

안건 및 의안번호	더불어민주당	국민의힘
법인세법 일부개정법률안 [2217206]	반대	찬성
종합부동산세법 폐지법률안 [2200637]	반대	찬성
재건축초과이익 환수에 관한 법률 폐지법률안 [2200148]	반대	찬성
대남대북전단 중지, 한반도 평화실현을 위한 국회 결의안 [2200699]	찬성	반대
노동조합 및 노동관계조정법 일부개정법률안 [2117346]	찬성	반대
전시작전통제권 환수 촉구 결의안 [2114224]	찬성	반대

(6) {"id": "bill_005",
"topic": "노동조합 및 노동관계조정법 일부개정법률안"
"text": "현행법은 근로자 및 사용자 개념에서 고전적인 1:1의 노사관계를 상정하고 있어 특수고용노동자, 간접고용노동자 쟁의행위에 대한 민사면책규정이 유명무실화되고, 노동쟁의의 대상을 '근로조건의 결정'에 관한 사항으로 한정하고 있어 노동쟁의의 범위가 협소하고, 쟁의행위등으로 인하여 발생한 재산적 손해에 대한 민사면책의 인정 요건 또한 이 법에 의한 단체교섭 및 쟁의행위로 좁게 한정하여 노조활동이 제약되거나 노동조합 및 근로자가 생계에 곤란을 겪는 경우가 많음. 아울러, 노조행위등으로 인한 손해배상의 배상청구의 범위가 지나치게 넓고 그 금액에 상한이 없어 노동조합의 활동이 위축되는 원인이 되고 있음. 이에 근로자, 사용자, 노동쟁의의 정의 규정을 수정하고, 손해배상 청구 제한을 확대하며, 개인에 대한 손해배상을 금지하는 등 조항을 개정 및 신설하여 개인 등에 대한 불합리한 손해배상을 제한하고, 헌법이 보장한 권리인 노동3권을 실질적으로 보장하고자 함."
"ground_truth": "DPK": "positive", "PPP": "negative"}

43) 「의안정보」, 국회 의안정보시스템(LIKMS), https://likms.assembly.go.kr(2026.06.25. 검색).

더불어민주당 에이전트와 국민의힘 에이전트는 법안명(topic)과 제안이유(text)를 입력으로 받아 앞서 〈표 12〉에서 제시된 프로필 정보를 참조하여 각 정당의 입장에서 찬성 또는 반대 여부를 표명한다. 이를 위해 사용된 프롬프트의 구체적인 구성은 〈표 14〉와 같다.

〈표 14〉 KoPAA 실험을 위한 프롬프트

프롬프트
판단할 현안] {issue_text} 위 현안에 대한 {party_name}의 공식 입장을 아래 JSON 형식으로만 응답하시오. 다른 텍스트는 일절 포함하지 마시오. {{ "stance": "찬성" 또는 "반대", "confidence": 0.0~1.0 사이의 실수, "reasoning": "판단 근거를 2~4개 문장으로 서술할 것" }}""

이와 같은 프롬프트에 기반하여 모델별로 더불어민주당 에이전트와 국민의힘 에이전트의 역할을 수행하여 추론 결과가 산출되었다. 다음으로 KoPAA를 구현하기 위해 선정된 모델들의 세부 사항은 〈표 15〉와 같다.

〈표 15〉 KoPAA 실험을 위한 거대언어모델 세부 사항

계열	모델	Model ID	파라미터
Kanana	Kanana-1.5-15.7B	kakaocorp/kanana-1.5-15.7b-a3b-instruct	15.7B
	Kanana-1.5-8B	kakaocorp/kanana-1.5-8b-instruct-2505	8B
SOLAR	SOLAR-10.7B	upstage/SOLAR-10.7B-Instruct-v1.0	10.7B
GPT	GPT-5.1	GPT-5.1-2025-11-13	비공개
	GPT-5.2	GPT-5.2-2025-12-11	
	GPT-5.4	GPT-5.4-2026-03-05	
	GPT-5.5	GPT-5.5-2026-04-23	
Claude	claude-sonnet-4-5	claude-sonnet-4-5-20250929	
	claude-haiku-4-5	claude-haiku-4-5-20251001	
	claude-opus-4-5	claude-opus-4-5-20251101	
	claude-opus-4-6	claude-opus-4-6	
	claude-opus-4-7	claude-opus-4-7	
	claude-opus-4-8	claude-opus-4-8	

Kanana 계열 및 SOLAR 계열 모델은 상대적으로 파라미터 규모가 큰 한국어 기반 모델이라는 점을 고려하여 선정되었다. GPT 계열 및 Claude 계열 모델은 2026년 6월 시점에서 대중적 인지도가 높고 범용적으로 활용되는 대표적인 모델군이라는 점을 고려하여 선정되었다.[44] 실험 환경과 관련하여, Kanana 계열 및 SOLAR 계열 모델은 Google Colaboratory 환경에서 A100 GPU 런타임을 사용하여 수행되었으며, GPT 계열 및 Claude 계열 모델은 각각 OpenAI API와 Anthropic API를 통해 수행되었다. KoPAA 실험의 하이퍼파라미터 설정은 〈표 16〉과 같다.

〈표 16〉 KoPAA 실험을 위한 하이퍼파라미터 설정

하이퍼파라미터	설정
부동소수점 정밀도(Precision)	bfloat16
최대 생성 토큰 수(Max New Tokens)	256
온도(Temperature)	0.3
디코딩 전략(Decoding Strategy)	Sampling

다음의 〈표 17〉은 더불어민주당 에이전트와 국민의힘 에이전트를 사용한 KoPAA 실험의 결과이다.

44) OpenAI, 「GPT-5 System Card」, OpenAI, 2025, https://cdn.openai.com/gpt-5-system-card.pdf(2026.06.25. 검색); Anthropic, 「System Card: Claude Opus 4.5」, Anthropic, 2025, https://www.anthropic.com/claude-opus-4-5-system-card(2026.06.25. 검색); Anthropic, 「System Card: Claude Opus 4.8」, Anthropic, 2026, https://www.anthropic.com/news/claude-opus-4-8(2026.06.25. 검색).

〈표 17〉 KoPAA 실험 결과

모델	더불어민주당(DPK)		국민의힘(PPP)		양당 합산	
	정확도	F1 점수	정확도	F1 점수	정확도	F1 점수
Kanana-1.5-15.7B	1.0000	1.0000	0.1667	0.1429	0.5833	0.5804
Kanana-1.5-8B	0.8333	0.8286	0.5000	0.3333	0.6667	0.6250
SOLAR-10.7B	0.6667	0.6250	0.6667	0.6250	0.6667	0.6250
GPT-5.1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
GPT-5.2	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
GPT-5.4	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
GPT-5.5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
claude-sonnet-4-5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
claude-haiku-4-5	1.0000	1.0000	0.8333	0.8286	0.9167	0.9161
claude-opus-4-5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
claude-opus-4-6	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
claude-opus-4-7	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
claude-opus-4-8	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

한국어 기반 거대언어모델의 경우에는 상대적으로 낮은 성능을 보였다. 우선 SOLAR-10.7B는 양당 모두 0.6667의 정확도를 보였으며, Kanana-1.5-8B는 더불어민주당 에이전트(0.8333)에 비해 국민의힘 에이전트(0.5000)에서 낮은 성능을 보였다. 특히 Kanana-1.5-15.7B는 더불어민주당 에이전트에서는 최고 수준의 성능(1.0000)을 달성하였으나 국민의힘 에이전트에서는 0.1667에 그쳐 극단적인 정당 간 성능의 불균형을 보였다. 이는 모델이 국민의힘 에이전트로서의 역할을 수행할 때 보수적 정치 입장을 충분히 반영하지 못하고 대부분의 안건에 반대 입장을 취하는 편향이 발생한 것으로 해석된다. 반면 GPT 모델과 대부분의 Claude 모델은 양당 에이전트 모두에서 최고 수준의 성능(1.0000)을 달성하였다. 이러한 모델군 간의 성능 차이는 부분적으로 모델 파라미터 규모에 기인하는 것으로 판단된다. GPT 계열 및 Claude 계열 모델들은 정확한 파라미터 수가 공개되어 있지 않으나, 본 연구에서 사용된 Kanana 계열 및 SOLAR 계열 모델들에 비해 상당히 거대한 규모일 것으로 추정된다.

4. 논의

1) 정치 데이터의 특수성과 한국어 전산정치학의 당위성

거대언어모델의 등장과 발전의 역사에 대한 연구들이 지적한 바와 같이 현시점의 거대언어모델은 범용적인 언어능력을 갖추고 있다.[45] 그러나 거대언어모델이 그럴 듯한 자연어 문장을 생성한다고 해서 특정 분야에 대한 전문성까지 갖추었다고 볼 수는 없을 것이다. 거대언어모델은 자신이 알고 있는 것과 알지 못하는 것을 명확히 구분하지 못한 채 근거가 불충분하거나 사실에 부합하지 않는 정보를 생성하는 경향이 있다.[46] 이러한 문제는 환각(Hallucination)으로 불리며, 학계는 물론 일반 대중에게도 거대언어모델의 대표적인 한계점으로 알려져 있다.[47] 환각을 방지하기 위해 사용될 수 있는 방법 중 하나는 거대언어모델이 전문 도메인에 특화될 수 있도록 별도의 학습을 진행하는 것이다. 이때 도메인은 일반적으로 법률이나 의학 분야와 같이 전문성을 요구하는 분야들을 일컫는다. 특히 법률 분야[48]와 의학 분야[49]

45) Chang, Tyler A. and Bergen, Benjamin K., Op. cit.; Naveed, Humza et al., Op. cit.
46) Yin, Zhangyue et al., "Do large language models know what they don't know?", Findings of the Association for Computational Linguistics: ACL 2023, 2023, pp.8653-8665.
47) Ji, Ziwei et al., "Survey of hallucination in natural language generation", ACM Computing Surveys 55-12, 2023, pp.1-38.
48) Zhong, Haoxi et al., "How does NLP benefit legal system: A summary of legal artificial intelligence", Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp.5218-5230; Hwang, Wonseok et al., "A multi-task benchmark for korean legal language understanding and judgement prediction", Advances in Neural Information Processing Systems 35, 2022, pp.32537-32551; Park, Ye-Chan, Lee, Hanyong and Lee, Jaesung, "Scaling Up: Analyzing Classification Outcomes in Large Korean Legal Document Datasets", Jahr-European Journal of Bioethics 15-2, 2024, pp.527-541; Ariai, Farid, Mackenzie, Joel and Demartini, Gianluca, "Natural language processing for the legal domain: A survey of tasks, datasets, models, and challenges", ACM Computing Surveys 58-6, 2025, pp.1-37; Li, Haitao et al., "LegalAgentBench: Evaluating LLM agents in legal domain", Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2025, pp.2322-2344; Siino, Marco et al., "Exploring llms applications in law: A literature review on current legal nlp approaches", IEEE Access 13, 2025, pp.18253-18276.
49) Harrison, Conrad J. and Sidey-Gibbons, Chris J., "Machine learning in medicine:

에서는 국내외를 막론하고 데이터셋 구축 및 벤치마크 개발은 물론, 이에 기반한 거대언어모델 활용 연구가 활발하게 이루어지고 있다. 이러한 연구 추세는 법률과 의학 분야의 중요성은 물론, 해당 분야에서 거대언어모델이 제대로 기능하기 위해서는 정교한 데이터셋과 벤치마크가 필수적임을 방증한다.

본고에서는 법률과 의학 분야가 그러했듯이 정치 분야 또한 거대언어모델을 위한 특수 도메인으로 기능할 수 있다는 점을 강조하고자 한다. 이는 정치 데이터가 가지고 있는 다음과 같은 특수성에 기인한다. 첫째, 정치 데이터는 개별 국가, 지역, 공동체별로 고유한 역사와 정서, 특징을 내포하고 있으며 이로 인해 특정 국가, 지역, 공동체의 정치적 의제와 담론을 단순하게 다른 국가, 지역, 공동체의 정치 지형에 적용할 수 없다.[50] 즉, 정치 데이터는 고유성(固有性·Originality)을 가진다. 이러한 측면에서 정치 데이터는 각 국가 및 공동체별로 고유한 역사와 특성을 가지는 법률 데이터와 유사하다. 둘째, 정치 데이터는 사회의 운영에 결정적인 영향을 끼치는 정치 분야의 데이터이므로 데이터의 분석이 해당 사회의 존속 및 발전에 유의미한 영향을 준다고 볼 수 있다. 정치 데이터에 오류나 편향이 개입될 경우, 여론의 형성이나 정치적 의사결정 과정에 왜곡이 발생할 수 있으며, 이는 사회 전반에 직간접적인 피해를 초래할 수 있다.[51] 즉, 정치 데이터는 파급성(波及性·Imp

---

a practical introduction to natural language processing", BMC Medical Research Methodology 21-1, 2021, p.158; Locke, Saskia et al., "Natural language processing in medicine: a review", Trends in Anaesthesia and Critical Care 38, 2021, pp.4-9; Kim, Yoojoong et al., "A pre-trained BERT for Korean medical natural language processing", Scientific Reports 12, 2022, p.13847; Kim, Yoojoong et al., "Predicting medical specialty from text based on a domain-specific pre-trained BERT", International Journal of Medical Informatics 170, 2023, 104956; Jang, Dongyeop et al., "GPT-4 can pass the Korean national licensing examination for Korean medicine doctors", PLOS Digital Health 2-12, 2023, e0000416; Jo, Eunbeen et al., "Assessing GPT-4's performance in delivering medical advice: A comparative analysis with human experts", JMIR Medical Education 10, 2024, e51282; Jo, Eunbeen et al., "Domain and Language adaptive pre-training of BERT models for Korean-English bilingual clinical text analysis", BMC Medical Informatics and Decision Making 25-1, 2025, p.428.

50) Almond, Gabriel A. and Verba, Sidney, The Civic Culture: Political Attitudes and Democracy in Five Nations, Princeton University Press, 1963.

51) Lazer, David M. J. et al., "The science of fake news", Science 359(6380), 2018, pp.1094-1096.

act)을 가진다. 이러한 측면에서 정치 데이터는 데이터의 품질 및 신뢰성이 사회적 파급력을 가지는 의료 데이터와 유사하다.

본고는 이러한 정치 데이터의 고유성과 파급성을 고려해 볼 때 한국 정치 데이터에 기반한 한국 전산정치학 연구의 당위성이 확보될 수 있다고 본다. 법률 분야의 고유성으로 인해 한국어 법률 자연어처리(Korean Legal NLP)와 한국어 법률 거대언어모델(Korean Legal LLM)이 필요하고, 의료 분야의 파급성으로 인해 한국어 의료 자연어처리(Korean Medical NLP)와 한국어 의료 거대언어모델(Korean Medical LLM)이 필요하듯이, 한국 정치의 특수성으로 인해 한국어 정치 자연어처리(Korean Political NLP)와 한국어 정치 거대언어모델(Korean Political LLM) 또한 필요하다고 할 수 있다. 이러한 관점에서 본고의 KoPolitical-LLM 프레임워크는 한국 정치 도메인을 대상으로 하는 거대언어모델 관련 연구의 구현 및 확장을 실증적으로 제시하였다는 의의가 있다.

2) 한국어 전산정치학의 향후 과제

본 절은 한국어 전산정치학 연구의 향후 과제를 정리하고 향후 연구 주제들에 대해 논의한다. 다만 후속 연구가 반드시 본고에서 제시하는 주제들에 한정될 필요는 없으며, 다양한 방향으로의 발전 가능성이 열려 있음을 밝힌다.

(1) 정치학과 인공지능 연구 간의 학제적 연구

개별 학문의 경계가 흐려지고 분야 간 공동 연구의 필요성이 증가함에 따라 학제적 연구(Interdisciplinary Research)는 연구자들이 지향해야 하는 행동지침으로 자리 잡고 있다.[52] 최근 거대언어모델 관련 연구에서의 학제적 연구는 단순히 추상적인 개념이나 외교적인 수사에 그치지 않고, 다양한 분야를 아우르는 연구자들의 구체적인 협업으로 실현되어 왔다. 앞서 4.1절에서 논의

52) Wilson, Edward O., Consilience: The Unity of Knowledge, Alfred A. Knopf, 1998; Tobi, Hilde and Kampen, Jarl K., "Research design: the methodology for interdisciplinary research framework", Quality & Quantity 52-3, 2018, pp.1209-1225; Repko, Allen F. and Szostak, Rick, Interdisciplinary Research: Process and Theory, Sage Publications, 2020.

된 거대언어모델을 적용한 의료·법률 데이터 연구, 그리고 Political-LLM 프로젝트에서 시도되고 있는 정치학자와 인공지능 연구자 간의 공동 연구가 바로 학제적 연구의 구체적인 사례라고 할 수 있다. 학제적 연구를 향해 나아가고 있는 현재 학계의 흐름을 고려할 때, 한국 정치 연구와 한국어 거대언어모델 연구 역시 공동 연구를 통해 상호 발전할 수 있을 것이다. 그러나 본고의 조사에서 드러났듯이, 한국 정치 연구자들과 인공지능 연구자들 간의 학제적 연구는 상대적으로 초기 단계에 있으며, 두 연구 집단 간의 교류와 공동 연구 또한 보다 활발히 이루어질 필요가 있다. 이러한 점에서 본고는 향후 한국 정치 연구자와 인공지능 연구자 간의 활발한 교류와 공동 연구를 통해 규모와 영향력 모두에서 유의미한 성과가 산출되기를 기대한다.

(2) 한국 정치 벤치마크 및 데이터셋 고도화

재현가능성(Reproducibility)은 현대 과학 연구의 핵심적인 특징으로,[53] 이를 위한 데이터셋의 정제 및 구축은 개별 연구의 완성도를 높이는 데 그치지 않고 후속 연구에서의 재현 및 보완을 가능하게 한다는 점에서 중요하다.[54] 또한 구축 및 공개된 정치 데이터셋은 그동안 주로 질적(Qualitative) 방법론에 의존해 온 정치 연구에 양적(Quantitative) 접근을 가능하게 함으로써 기존의 질적 분석을 보완할 수 있다.[55] 나아가 거대언어모델 연구에서는 성능 평가를 위한 벤치마크의 구축 및 고도화가 필수적이며, 거대언어모델의 특수한 도메인에 대한 전문성이 부각되고 있는 추세를 고려하면 한국 정치 역시 주요 도메인으로 기능한다고 볼 수 있다. 즉, 한국 정치 벤치마크 및 데이터셋의 확충은 한국 전산정치학의 발전을 위해 중요한 과제이다.

---

53) Open Science Collaboration, "Estimating the reproducibility of psychological science", Science 349(6251), 2015, aac4716.

54) Laraway, Sean et al., "An overview of scientific reproducibility: Consideration of relevant issues for behavior science/analysis", Perspectives on Behavior Science 42-1, 2019, pp.33-57; Miyakawa, Tsuyoshi, "No raw data, no science: another possible source of the reproducibility crisis", Molecular Brain 13-1, 2020, p.24.

55) McNabb, David E., Research Methods for Political Science: Quantitative and Qualitative Methods, Routledge, 2015.

본고의 도입부에서 논의한 바와 같이 한국어 전산정치학은 현재 초기 단계에 있으며 연구 성과가 점진적으로 축적되고 있다. 이로 인해 한국 정치를 대상으로 하는 벤치마크 또는 오픈소스 데이터셋은 제한적인 상황이다. 2026년 6월 기준으로 한국 정치 데이터를 수집·정제하여 후속 연구에 활용 가능한 데이터셋으로 구축한 연구는 대부분 더불어민주당과 국민의힘 양당의 공식 자료 및 온라인 정치 기사를 대상으로 한다.[56] 향후에는 이러한 더욱 다양한 성격의 정치 데이터를 수집·정제함으로써 벤치마크와 데이터셋의 고도화가 이루어질 수 있을 것이다.

(3) 복잡계로서의 한국 정치 연구

비단 전산정치학뿐만 아니라 정치학 연구 전반에서 고려되어야 할 점은 바로 정치 이론 및 정치철학 등에 대한 연구가 현실 정치에 대한 연구에 그대로 적용되기 어렵다는 것이다. 현실 정치에서 정당 및 정치인은 대외적으로 표방하는 이념적 지향성과 무관하게 상대 정당 및 정치인을 견제하거나 비판하기 위해 행동하며, 이러한 행동이 반드시 본래의 이념적 지향성과 일치하지는 않는다.[57] 유권자 역시 이념적 지향성에 따라 합리적으로 투표 및 시위 등의 정치적 행위를 하는 것이 아니라, 특정 정치 현안이나 개별 정치인에 대한 반감에 기반하여 행동하기도 한다. 즉, 정치적 신념에 따라 특정 정당이나 정치인에게 동조(align with)하는 방식이 아니라, 반발 심리에 기반하여 상대를 배척(align against)하는 방식으로 정치적 의사를 표명한다는 것이다.[58] 정치 행위자의 이러한 특성은, 여론조사가 빗나가는 경우에서 볼 수 있듯이, 정치적 이념에 따른 행동을 전제로 하는 예측 및 시뮬레이션 연구를 어렵게 만드는 요인이다.

---

56) Kim, DongJae et al., Op. cit.; Lee, Dahyun et al., Op. cit.; Park, Sunkyoung et al., Op. cit.

57) Lee, Myeonghwa, Shang E. Ha and Lee, Wonjae, "Do political parties always prefer loyalists? Evidence from South Korea", PLoS ONE 18-11, 2023, e0291336.

58) Abramowitz, Alan I. and Webster, Steven W., "Negative partisanship: Why Americans dislike parties but behave like rabid partisans", Political Psychology 39, 2018, pp.119-135.

향후에는 이러한 현실 정치의 양상을 일종의 복잡계(Complex System)로 상정하여 연구를 수행할 수 있을 것으로 기대된다.[59] 즉, 다양한 요인들이 복합적으로 작용하는 상황을 전제로 하는 것이다. 최근에는 거대언어모델 기반 에이전트를 복잡계 내부에서 행동하도록 설정하는 연구가 시도되고 있다.[60] 이러한 연구에서는 거대언어모델이 주어진 환경에서 동적으로 결정을 내리도록 하며, 이를 통해 행위자들 간의 협력이나 규범의 준수와 같이 실제 사회에서 발생하는 현상들을 재현하려고 시도한다. 한국어 전산정치학 역시 이러한 방법론을 반영하여 복잡계로서의 한국 정치와 그 내부에서 행위하는 주체로서의 한국어 거대언어모델을 대상으로 하는 연구를 수행할 수 있을 것이다. 그리고 이러한 시도를 통해 현재 한국어 전산정치학의 주요 연구 주제인 정치 성향 탐지, 정치에서의 성 편향, 거대언어모델의 정치적 편향 분석 등을 넘어 정치 지형과 정치 행위자의 행위에 대한 심층적인 분석으로 발전할 수 있을 것으로 기대된다.

5. 결론

본고는 한국어 전산정치학의 발전 가능성을 탐색하기 위해 거대언어모델을 한국 정치 데이터에 적용한 두 가지 사례를 제시하였다. 첫 번째 사례인 KoP-Stance는 더불어민주당과 국민의힘 양당의 공식 대변인 논평 및 브리핑 데이터를 기반으로 구축된 한국 정치 성향 탐지 프레임워크이다. 2017년 5월 9일부터 2026년 5월 31일까지 약 9년간 수집된 논평 및 브리핑 데이터를 정제하고 이를 통해 데이터셋을 구축하여 한국어 기반 거대언어모델에

59) Rhee, Young Pil, “Complex systems approach to the study of politics”, Systems Research and Behavioral Science 17-6, 2000, pp.487-491.

60) Gürcan, Önder, “Llm-augmented agent-based modelling for social simulations: Challenges and opportunities”, HHAI 2024: Hybrid Human AI Systems for the Social Good, SAGE Publications, 2024, pp.134-144; Lu, Yikang et al., “LLMs and generative agent-based models for complex systems research”, Physics of Life Reviews 51, 2024, pp.283-293.

파인튜닝과 인컨텍스트 러닝 기법을 적용하였다. 두 번째 사례인 KoPAA는 거대언어모델에 양당의 정체성을 부여하여 특정 안건에 대한 정당의 입장을 예측하는 정치 행위자 에이전트이다. 평가를 위해 실제 국회 발의 법안이 활용되었으며 이를 통해 거대언어모델의 한국 현실 정치에 대한 이해 및 처리 능력을 검증하였다.

이어서 본고가 가지고 있는 한계점과 향후 발전 방향을 명시하고자 한다. 첫째, KoP-Stance에서 채택한 진보와 보수의 이진 분류 체계는 한국 정치의 복잡한 이념적 스펙트럼을 충분히 포착하지 못한다는 한계가 있다. 이를 해결하기 위해 양대 정당 외에 소수 정당 및 원외 정당 등에 대한 고려가 이루어질 필요가 있다. 둘째, KoPAA는 정당 에이전트로 구축되었으나 이 과정에서 사용된 평가 데이터의 양이 부족하다는 한계가 있다. 이는 국회 법안과 같이 공식성을 갖추면서도 양당 간 대립 구도가 뚜렷하게 드러나는 데이터를 양적으로 충분히 확보하기 어려우며, 기계적으로 증강하기도 어렵다는 점에 기인한다. 그러나 향후 데이터셋의 발전 및 평가 체계의 고도화를 통해 이러한 한계를 극복할 수 있을 것으로 기대된다. 마지막으로, 기존 연구들이 지적한 바와 같이, 거대언어모델에 내재하는 정치적 편향이 본고의 연구 결과에 영향을 미쳤을 가능성을 배제할 수 없다.

이러한 한계에도 불구하고 본고는 한국 정치 데이터를 대상으로 거대언어모델을 직접적으로 적용한 구체적인 연구 사례를 제시함으로써 한국어 전산정치학의 연구 가능성을 실증하였다는 의의가 있다. 또한 한국어 기반 정치 성향 분류기를 공개함으로써 본고의 연구 결과를 직접적으로 활용할 수 있도록 하였다. 이와 같이 본고는 한국어 전산정치학의 토대를 마련하기 위한 시범적 프레임워크로서 KoPolitical-LLM을 제시하였다. 본고가 한국 정치 연구에 거대언어모델을 접목하려는 후속 연구자들에게 참고 지침이 되고, 나아가 한국어 전산정치학의 외연 확장에 기여할 수 있기를 기대한다.

▮ 참고문헌

「강령·당헌·당규」, 더불어민주당, https://theminjoo.kr/main/sub/introduce/rule.php(2026.06.25. 검색).
「강령·당헌·당규」, 국민의힘, https://www.peoplepowerparty.kr/about/constitution(2026.06.25. 검색).
「의안정보」, 국회 의안정보시스템(LIKMS), https://likms.assembly.go.kr(2026.06.25. 검색).
「KoP-Stance: 한국어 정치 성향 분류기」, HuggingFace, https://huggingface.co/spaces/San0727/KoP-Stance(2026.06.25. 검색).

【국내 논저】
김승래, 「김정은과 남한 대통령 정치 담화의 프레이밍 전략 비교: 형용사를 중심으로」, 『현대정치연구』 18-2, 2025.
김영명, 「한국 정치의 특수성에 관한 연구 서설: 분단, 압축성장, 단일사회 문화」, 『비교민주주의연구』 7-1, 2011.
김정연, 「대규모 언어모델(LLM) 계열 기술의 정치·공공 부문 적용: 생성형 인공지능의 기대와 한계」, 『사회적질연구』 9-1, 2025.
김현강, 「내러티브 텍스트이자 정치담화로서의 정치광고 분석-한국의 16대, 17대 대통령 선거 광고를 중심으로-」, 『사회언어학』 22-1, 2014.
서정연·조석민·박종철, 「거대 언어 모델의 정치적 편향과 하위 작업에서의 영향」, 『정보과학회논문지』 52-1, 2025.
송종빈·송상헌, 「한국어 언어모델의 정치편향성 측정」, 『어학연구』 62-1, 2026.
이민철, 「Kiwi: 통계적 언어 모델과 Skip-Bigram을 이용한 한국어 형태소 분석기 구현」, 『디지털인문학』 1-1, 2024.
이인복, 「정치학 연구에서의 대형언어모델(LLM): 주제 동향과 사용 현황 분석」, 『국제정치논총』 65-3, 2025.
임태훈, 「12·3 이후의 청각 정치학 - 광장의 소리와 이론의 탐색」, 『대중서사연구』 31-2, 2025.
장승진, 「비상계엄과 탄핵, 그리고 제21대 대선의 투표선택: 포퓰리즘과 회고적 투표」, 『한국정치연구』 34-3, 2025.
전예성·강신재·박상현, 「LLM을 통한 한국 국회의원의 이념성향 측정: 국회 회의록 텍스트 분석을 중심으로」, 『한국정당학회보』 25-1, 2026.
정다솜·권순희, 「문재인 대통령의 SNS 정치 담화에 대한 텍스트언어학적 분석」, 『텍스트언어학』 49, 2020.
정태욱, 「한국 헌정사에서 대통령 책임과 대통령 탄핵」, 『민주법학』 87, 2025.
주미선 외, 「CALL: 대규모 언어 모델 기반 온라인 정치 커뮤니티 특성 추출 및 평가 전략」, 『한국HCI학회 논문지』 21-1, 2026.
천정환, 「거리정치를 통해 본 내란 이후 '극우화'의 문화정치적 자원과 정동적 동력」, 『역사비평』 151, 2025.
한은수, 「텍스트 마이닝으로 본 포퓰리즘 리더십 유형 연구: 도널드 트럼프(Donald Trump)와 이재명 대통령 연설문의 수사적 프레임과 정책 지향성 비교 분석」, 『연구방법논총』 10-3, 2025.

【해외 논저】

Abramowitz, Alan I. and Webster, Steven W., "Negative partisanship: Why Americans dislike parties but behave like rabid partisans", Political Psychology 39, 2018.

Almond, Gabriel A. and Verba, Sidney, The Civic Culture: Political Attitudes and Democracy in Five Nations, Princeton University Press, 1963.

Anthropic, 「System Card: Claude Opus 4.5」, Anthropic, 2025, https://www.anthropic.com/claude-opus-4-5-system-card(2026.06.25. 검색).

Anthropic, 「System Card: Claude Opus 4.8」, Anthropic, 2026, https://www.anthropic.com/news/claude-opus-4-8(2026.06.25. 검색).

Ariai, Farid, Mackenzie, Joel and Demartini, Gianluca, "Natural language processing for the legal domain: A survey of tasks, datasets, models, and challenges", ACM Computing Surveys 58-6, 2025.

Arslan, Muhammad, Munawar, Saba and Cruz, Christophe, "Political-RAG: using generative AI to extract political information from media content", Journal of Information Technology & Politics 22-4, 2025.

Bang, Yejin et al., "Measuring political bias in large language models: What is said and how it is said", Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024.

Barros, Amon, Prasad, Ajnesh and Śliwa, Martyna, "Generative artificial intelligence and academia: Implication for research, teaching and service", Management Learning 54-5, 2023.

Bengio, Yoshua et al., "A neural probabilistic language model", Journal of Machine Learning Research 3, 2003, pp.1137-1155.

Blank, Idan A., "What are large language models supposed to model?", Trends in Cognitive Sciences 27-11, 2023.

Brown, Tom et al., "Language models are few-shot learners", Advances in Neural Information Processing Systems 33, 2020.

Burnham, Michael, "Stance detection: A practical guide to classifying political beliefs in text", Political Science Research and Methods 13-3, 2024.

Burrows, Mathew and Braml, Josef, World to Come: The Return of Trump and the End of the Old Order, Brixton Ink, 2025.

Chang, Tyler A. and Bergen, Benjamin K., "Language model behavior: A comprehensive survey", Computational Linguistics 50-1, 2024.

De Marchi, Scott and Page, Scott E., "Agent-based models", Annual Review of Political Science 17-1, 2014.

Devlin, Jacob et al., "BERT: Pre-training of deep bidirectional transformers for language understanding", Proceedings of NAACL-HLT 2019, 2019.

Farooq, Aqsa et al., "Generative AI Generating Concerns: Citizens' Perspectives During the 2024 European Elections", The International Journal of Press/Politics, 2025.

Gürcan, Önder, "Llm-augmented agent-based modelling for social simulations: Challenges and opportunities", HHAI 2024: Hybrid Human AI Systems for the Social Good, SAGE Publications, 2024.

Haq, Ehsan Ul et al., "A survey on computational politics", IEEE Access 8, 2020.

Hardalov, Momchil et al., "A survey on stance detection for mis-and disinformation identification", Findings of the Association for Computational Linguistics: NA ACL 2022.

Harrison, Conrad J. and Sidey-Gibbons, Chris J., "Machine learning in medicine: a practical introduction to natural language processing", BMC Medical Researc h Methodology 21-1, 2021.

Hwang, Wonseok et al., "A multi-task benchmark for korean legal language understanding and judgement prediction", Advances in Neural Information Processing Systems 35, 2022.

Jang, Dongyeop et al., "GPT-4 can pass the Korean national licensing examination for Korean medicine doctors", PLOS Digital Health 2-12, 2023, e0000416.

Ji, Ziwei et al., "Survey of hallucination in natural language generation", ACM Computing Surveys 55-12, 2023.

Jo, Eunbeen et al., "Assessing GPT-4's performance in delivering medical advice: A comparative analysis with human experts", JMIR Medical Education 10, 2024, e51282.

Jo, Eunbeen et al., "Domain and Language adaptive pre-training of BERT models for Korean-English bilingual clinical text analysis", BMC Medical Informatics and Decision Making 25-1, 2025.

Jung, Yujin J., "The Sectarian Divide: The Dynamics of Populism in South Korea", Journal of East Asian Studies 25-2, 2025.

Kanana LLM Team, "Kanana: Compute-efficient bilingual language models", arXiv preprint arXiv:2502.18934, 2025.

Kang, Sungwoo, "Diagnosing Korean-Language LLM Political Bias via Census-Grounded Agent Simulation", arXiv preprint arXiv:2605.18395, 2026.

Kaplan, Jared et al., "Scaling laws for neural language models", arXiv preprint arXiv: 2001.08361, 2020.

Kim, Beomjune, Lee, Eunsun and Na, Dongbin, "A new Korean text classification benchmark for recognizing the political intents in online newspapers", arXiv preprint arXiv:2311.01712, 2023a.

Kim, Sanghoon et al., "SOLAR 10.7B: Scaling large language models with simple yet effective depth up-scaling", arXiv preprint arXiv:2312.15166, 2023b.

Kim, DongJae et al., "LASTIST: LArge-Scale Target-Independent STance dataset", arXiv preprint arXiv:2510.25783, 2025.

Kim, Yoojoong et al., "A pre-trained BERT for Korean medical natural language processing", Scientific Reports 12, 2022.

Kim, Yoojoong et al., "Predicting medical specialty from text based on a domain-specific pre-trained BERT", International Journal of Medical Informatics 170, 2023c, 104956.
Kim, Yeaji, "The development of party organizations in new democracies: Evidence from South Korea", in Poguntke, Thomas and Hofmeister, Wilhelm (eds.), Political Parties and the Crisis of Democracy: Organization, Resilience, and Reform, Oxford University Press, 2024.
Laraway, Sean et al., "An overview of scientific reproducibility: Consideration of relevant issues for behavior science/analysis", Perspectives on Behavior Science 42-1, 2019.
Lazer, David M. J. et al., "The science of fake news", Science 359(6380), 2018.
Lee, Dahyun et al., "Journalism-Guided Agentic In-Context Learning for News Stance Detection", Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 2025a.
Lee, Sangah et al., "KR-BERT: A small-scale Korean-specific language model", arXiv preprint arXiv:2008.03979, 2020.
Lee, Keyeun et al., "Ideology and policy preferences in synthetic data: The potential of LLMs for public opinion analysis", Media and Communication 13, 2025b.
Lee, Myeonghwa, Shang E. Ha and Lee, Wonjae, "Do political parties always prefer loyalists? Evidence from South Korea", PLoS ONE 18-11, 2023, e0291336.
Li, Hao, Gong, Ruoyuan and Jiang, Hao, "Political actor agent: Simulating legislative system for roll call votes prediction with large language models", Proceedings of the AAAI Conference on Artificial Intelligence 39-1, 2025a.
Li, Haitao et al., "LegalAgentBench: Evaluating LLM agents in legal domain", Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2025b.
Li, Lincan et al., "Political-LLM: Large language models in political science", arXiv preprint arXiv:2412.06864, 2024.
Li, Yingjie et al., "P-stance: A large dataset for stance detection in political domain", Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021.
Liu, Ruibo et al., "Quantifying and alleviating political bias in language models", Artificial Intelligence 304, 2022a.
Liu, Yujian et al., "POLITICS: Pretraining with same-story article comparison for ideology prediction and stance detection", Findings of the Association for Computational Linguistics: NAACL 2022, 2022b.
Locke, Saskia et al., "Natural language processing in medicine: a review", Trends in Anaesthesia and Critical Care 38, 2021.
Lu, Yikang et al., "LLMs and generative agent-based models for complex systems research", Physics of Life Reviews 51, 2024.
Luo, Junyu et al., "Large language model agent: A survey on methodology, applications and challenges", arXiv preprint arXiv:2503.21460, 2025.

McNabb, David E., Research Methods for Political Science: Quantitative and Qualitative Methods, Routledge, 2015.
Mikolov, Tomas et al., "Recurrent neural network based language model", Proceeding s of Interspeech 2010, ISCA, 2010.
Miyakawa, Tsuyoshi, "No raw data, no science: another possible source of the reprod ucibility crisis", Molecular Brain 13-1, 2020.
Moghimifar, Farhad, Li, Yuan-Fang, Thomson, Robert and Haffari, Gholamreza, "Mod elling political coalition negotiations using llm-based agents", arXiv preprint arXiv:2402.11712, 2024.
Mouthami, K., Naren, P. and Pranesh, R., "Political sentiment analysis on twitter using deep learning and Llm models", 2025 3rd International Conference on Advan cements in Electrical, Electronics, Communication, Computing and Automati on (ICAECA), IEEE, 2025.
Naveed, Humza et al., "A comprehensive overview of large language models", ACM Transactions on Intelligent Systems and Technology 16-5, 2025.
Open Science Collaboration, "Estimating the reproducibility of psychological science", Science 349(6251), 2015, aac4716.
OpenAI, 「GPT-5 System Card」, OpenAI, 2025, https://cdn.openai.com/gpt-5-system-card.pdf(2026.06.25. 검색).
Park, Jangwon, 「KoELECTRA: Pretrained ELECTRA Model for Korean」, GitHub, 2020, https://github.com/monologg/KoELECTRA(2026.06.25. 검색).
Park, Sunkyoung et al., "A benchmark dataset for evaluating gender sensitivity in Korean political discourse with large language models", Scientific Data, 2025.
Park, Sungjoon et al., "KLUE: Korean language understanding evaluation", arXiv preprint arXiv:2105.09680, 2021.
Park, Ye-Chan, Lee, Hanyong and Lee, Jaesung, "Scaling Up: Analyzing Classification Outcomes in Large Korean Legal Document Datasets", Jahr-European Journal of Bioethics 15-2, 2024.
Pavlyshenko, Bohdan M., "Using GPT Models for Qualitative and Quantitative News Analytics in the 2024 US Presidential Election Process", arXiv preprint arXiv: 2410.15884, 2024.
Pit, Pagnarasmey et al., "Whose side are you on? Investigating the political stance of large language models", arXiv preprint arXiv:2403.13840, 2024.
Radford, Alec et al., "Language models are unsupervised multitask learners", OpenAI Blog 1-8, 2019.
Repko, Allen F. and Szostak, Rick, Interdisciplinary Research: Process and Theory, Sage Publications, 2020.
Rettenberger, Luca, Reischl, Markus and Schutera, Mark, "Assessing political bias in large language models", Journal of Computational Social Science 8-2, 2025.

Rhee, Young Pil, "Complex systems approach to the study of politics", Systems Research and Behavioral Science 17-6, 2000.

Siino, Marco, Falco, Michele, Croce, Danilo and Rosso, Paolo, "Exploring llms applications in law: A literature review on current legal nlp approaches", IEEE Access 13, 2025.

Taubenfeld, Amir et al., "Systematic biases in LLM simulations of debates", Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024.

Tobi, Hilde and Kampen, Jarl K., "Research design: the methodology for interdisciplinary research framework", Quality & Quantity 52-3, 2018.

Vaswani, Ashish et al., "Attention is all you need", Advances in Neural Information Processing Systems 30, 2017.

Wilson, Edward O., Consilience: The Unity of Knowledge, Alfred A. Knopf, 1998.

Wojczewski, Thorsten, "Trump, populism, and American foreign policy", Foreign Policy Analysis 16-3, 2020.

Xiao, Tong and Zhu, Jingbo, "Foundations of large language models", arXiv preprint arXiv:2501.09223, 2025.

Yin, Zhangyue et al., "Do large language models know what they don't know?", Findings of the Association for Computational Linguistics: ACL 2023.

Yu, Seunguk et al., "From Ground Trust to Truth: Disparities in Offensive Language Judgments on Contemporary Korean Political Discourse", Findings of the Association for Computational Linguistics: EMNLP 2025, 2025.

Yun, Hyunsoo and Lee, Eun Hak, "Party politics in transport policy with a large language model", Transport Policy 171, 2025.

Zhong, Haoxi et al., "How does NLP benefit legal system: A summary of legal artificial intelligence", Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020.

Zhou, Lingfeng, et al., "FlockVote: LLM-empowered agent-based modeling for simulating U.S. presidential elections", arXiv preprint arXiv:2512.05982, 2025.

【Abstract】

KoPolitical-LLM: Exploring Large Language Models for Korean Political Science

Noh, Kang-san* · Song, Sang-houn**

This paper presents KoPolitical-LLM, a framework consisting of two concrete research cases designed to explore the potential of Korean Computational Political Science in applying large language models to the study of Korean politics. While the use of LLMs in political research has been actively pursued in Anglophone and European academia, its adoption in Korean political research remains at a relatively nascent stage. Departing from the recognition that political discourse is deeply embedded in the unique historical and institutional contexts of individual nations, this paper aims to demonstrate that methodologies developed within the tradition of computational political science can be meaningfully extended to Korean political data.

The first research case, KoP-Stance, is a Korean political stance detection framework built on official spokesperson press releases and briefings from the Democratic Party of Korea and the People Power Party, spanning approximately nine years. Both encoder-based fine-tuning and decoder-based in-context learning were applied, with fine-tuned encoder models achieving macro F1 scores above 0.83, outperforming decoder-based in-context learning by approximately eight to ten percentage points. The fine-tuned models have been made publicly available via HuggingFace.

The second research case, KoPAA (Korean Political Actor Agent), assigns the ideological identities of the two major parties to large language models and predicts each party's stance on actual bills introduced in the National Assembly of Korea. Evaluated across Korean-specialized models and larger frontier models including GPT and Claude series, the results reveal a notable performance gap, with frontier models achieving near-perfect accuracy while Korean-specialized models exhibited lower and more uneven performance.

This paper further discusses the rationale for Korean computational political science, grounded in two distinctive properties of political data, namely originality and societal

* Primary Author / PhD Student, Dept. of Linguistics, Korea Univ.
** Corresponding Author / Professor, Dept. of Linguistics, Korea Univ.

impact, and proposes future directions including interdisciplinary collaboration, benchmark development, and the conceptualization of Korean politics as a complex system.

Keywords: Computational Political Science, Korean Politics, Large Language Model, Political Stance Detection, Political Actor Simulation

접수일: 2026.06.30. ‖ 심사일: 2026.08.07. ‖ 게재확정일: 2026.08.14.