Political Ideology in NLP: A Survey of Detection, Measurement, and Intervention

Anonymous ACL submission

Abstract

Political ideology is a latent yet pervasive force that shapes how political positions are articulated, interpreted, and opposed. Computational NLP approaches increasingly aim to detect, measure, and intervene in ideological phenomena, offering new tools for analyzing political discourse at scale. However, the goals, challenges, and outcomes of these methods remain insufficiently categorized, limiting their accessibility to domain experts. In this survey, we organize existing work into three perspectives, aligning with the usage of ideology in the domain of political sciences. First, detection focuses on identifying ideological positions in text and among users. Second, measurement examines how ideology is encoded in models. Third, intervention studies how to control or mitigate ideological behaviors in model outputs. We then discuss the role of LLMs from a methodological perspective and the types of data employed in each category. Finally, we identify key gaps and future directions in political NLP, including dataset limitations, model generalizability, and the integration of social science theories with computational modeling. This survey provides a structured roadmap for machine learning and domain researchers seeking to understand, analyze, and implement methods for political ideology in NLP.

1 Introduction

Ideology is a foundational concept in political science, shaping individual preferences, group identities, and institutional behavior (Poole and Rosenthal, 1985; Clinton et al., 2004; Fowler, 2006; Jost et al., 2009). As political communication increasingly migrates to digital platforms, large-scale textual data such as news articles, social media posts, speeches, and debates have become central resources for studying ideological dynamics. To this end, different NLP techniques, including recent LLMs, provide powerful tools for detecting ideological signals, modeling latent belief structures, and analyzing biases. By studying large-scale data, researchers can better understand ideological shifts and discourse patterns, therefore enabling interventions to steer and mitigate harmful biases in models (Jenny et al., 2023).

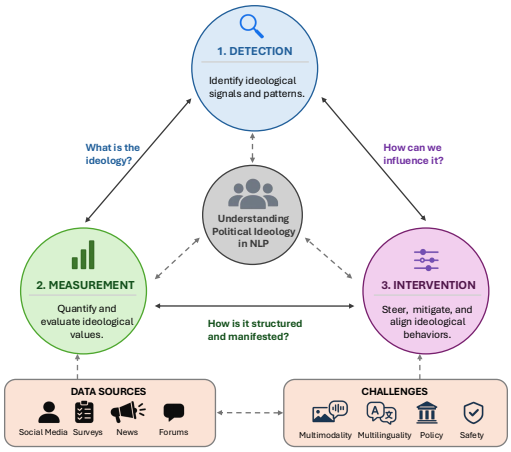

Figure 1: A conceptual overview of political ideology research in NLP.

Existing works in this space are broad and rapidly expanding. A large body of studies seek to detect ideology from text, users, or institutions by predicting partisan affiliation, estimating latent ideological positions, or comparing actors along ideological dimensions (Iyyer et al., 2014). Other works examine how ideology is measured within language models, including investigating whether embeddings or generated outputs encode partisan preferences, stereotypes, or political worldviews (Zhou et al., 2023; Chen et al., 2024). More recently, researchers have explored intervention strategies that attempt to steer model behavior, promote neutrality, or mitigate ideological bias in politically sensitive settings (Kim et al., 2025).

Despite this rapid growth, the literature remains dispersed across NLP, machine learning, information retrieval, and computational social science

venues, with limited unifying structure. Prior surveys on fairness, bias, or computational politics often cover only subsets of the problem space, while recent LLM-focused discussions tend to emphasize isolated tasks such as political bias evaluation or stance detection. Furthermore, there is a lack of a common vocabulary, and similar methods or downstream tasks are referred to differently depending on the community. To date, there is no comprehensive survey that systematically organizes political ideology research in NLP across detection, measurement, and intervention.

To address this gap, we propose a domain-informed taxonomy of NLP research on ideology-related tasks, shown in Figure 1. It organizes prior work into three major branches including: (1) **Ideology Detection**, including classification, scaling, and ranking; (2) **Ideology Measurement**, including probing, prompting, and evaluation; and (3) **Ideology Intervention**, including steering and debiasing. This taxonomy corresponds to three core questions: *what is ideology*, *how ideology is structured and manifested*, and *how ideological behavior can be influenced in NLP systems*. It also highlights both methodological continuity and emerging opportunities in the LLM era, trying to bridge at the same time the gap with the domain to avoid the incorrect utilization of methods.

Specifically, within each paradigm, we begin by introducing a systematic categorization of the underlying subtasks, providing a unified framework for organizing the existing literature. We then discuss LLM-based modeling approaches, representative benchmark datasets, and evaluation protocols across the three paradigms. Finally, we outline potential gaps and promising future directions, including investigating ideology in real political data, multilingual and multimodal ideology modeling, and robust and reproducible evaluation, etc. As such, this survey seeks to provide a unified roadmap of political ideology research in NLP, helping researchers better understand the current landscape and develop more reliable, fair, and socially responsible language technologies for politically sensitive domains.

2 Preliminaries

Political ideology definition. Political ideology refers to a latent and not directly observable belief structure that shapes how individuals interpret political issues, organize preferences, and relate to broader social and political conflicts (Poole and Rosenthal, 1985; Clinton et al., 2004; Barberá, 2015). As a result, ideology is typically approximated through observable traces, including language use, stated attitudes, voting behavior, group affiliation, and issue-specific positions. The field of political science has long cautioned against treating ideology simply as a discrete party label or as a synonym for stance. Party affiliation is an organizational or electoral category, while ideology refers to a latent belief structure that may cut across parties, factions, and institutional contexts (Poole and Rosenthal, 1985; Clinton et al., 2004). Similarly, stance captures a position on a specific issue or target, whereas ideology links multiple issue positions into broader belief systems (Laver et al., 2003; Slapin and Proksch, 2008). In NLP, it is often operationalized in simplified forms, such as left-right labels, party labels, or issue-specific stance labels, but such representations may overlook its inherently multidimensional, context-dependent, and culturally grounded nature. In this survey, we adopt a broader view of political ideology as both an explicit attribute in text and a latent construct embedded in language use and model representations, which can manifest in subtle and indirect ways across different contexts.

Conceptual task decomposition. We organize political ideology paradigms into detection, measurement, and intervention based on a conceptual decomposition of how ideology operates in NLP systems. Specifically, ideology can be viewed as (i) something to be **identified** from observable signals (detection), (ii) something that is internally **encoded and structured** within models (measurement), and (iii) something that can be **influenced or controlled** to shape model behavior (intervention). This breakdown reflects a progression from observable outputs to latent mechanisms to actionable control, providing a coherent framework to unify otherwise fragmented research threads.

Differences to existing surveys. While prior surveys have explored political-related tasks in NLP, our work differs in scope, structure, and perspective in several key ways. Existing surveys generally fall into two categories. On one hand, many focus on specific subdomains, such as stance detection (Aldayel and Magdy, 2020; Zhang et al., 2024; Pangtey et al., 2025), fairness and bias (Mehrabi et al., 2019; Nakov et al., 2021; Guo et al., 2024; Lin and Li, 2025), or sentiment analysis (Tan et al., 2023; Chandan and Mandal, 2025), treating po-

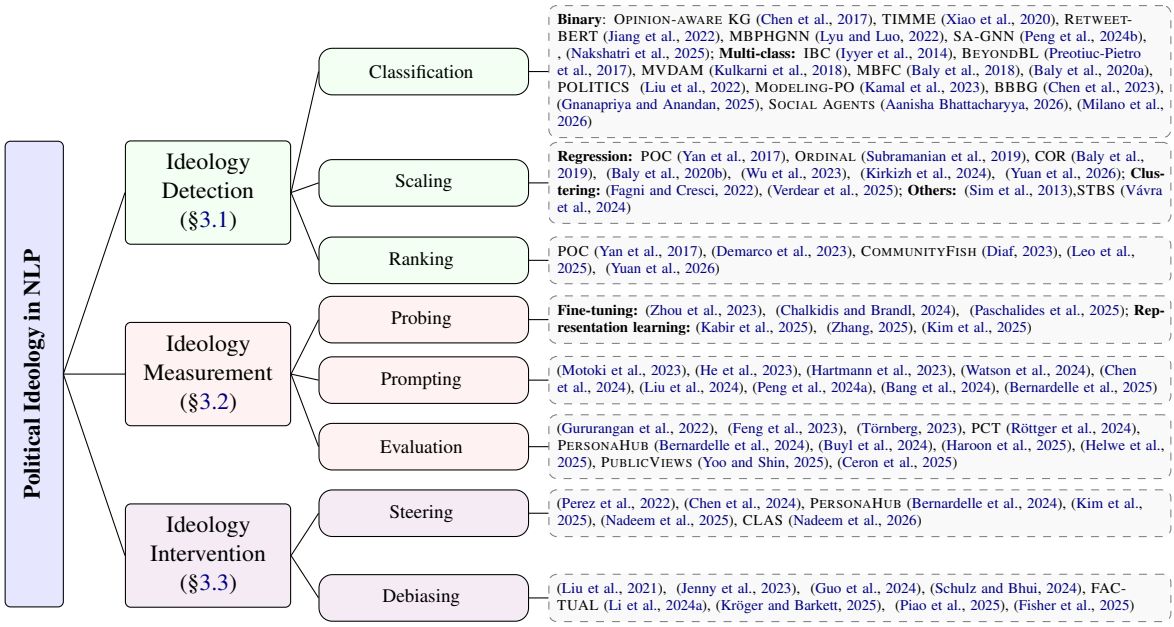

Figure 2: Taxonomy of Political Ideology research with corresponding literature.

litical ideology as a task-specific signal within a constrained setting. On the other hand, a smaller set of works adopts a more holistic perspective on political phenomena (Wilkerson and Casas, 2017; Chatsiou and Jankin, 2020; Linegar et al., 2023; Li et al., 2024b; Aoki, 2024), but often remain at a high level, without systematically connecting detection, measurement, and intervention within modern NLP systems. Building on this distinction, our survey places **political ideology** itself at the center, rather than treating it as a byproduct of tasks. We provide a unified framework with a particular focus on how ideology is implicitly encoded and expressed in modern LLMs. Detailed comparison see Appendix A.

3 Task Types

Understanding political ideology in language requires evaluating models across a diverse set of tasks that capture how ideological signals are expressed, encoded, and acted upon. We summarize how prior work studies ideology by organizing tasks into the three paradigms. An overview is presented in Figure 2.

3.1 Detecting Political Ideology

Classification. Political ideology detection is commonly formulated as a classification task, where texts, users, or entities are assigned discrete ideological labels. Existing work mainly distinguishes between binary settings (e.g., left vs. right) and more fine-grained multi-class or multi-label settings involving party affiliation, ideological categories, or multiple ideological dimensions. Binary formulations are especially common in social media analysis (Jiang et al., 2022; Peng et al., 2024b), while broader settings consider richer ideological categorizations across news media, political actors, and public discourse (Baly et al., 2020a; Lyu and Luo, 2022; Milano et al., 2026). Overall, this line of work treats ideology as an explicit and observable label grounded in annotated data.

Scaling. Beyond discrete labels, some work models political ideology using regression or clustering approaches, capturing it along continuous uni or multidimensional spectra. For example, Subramanian et al. (2019) model ideology detection as an ordinal regression problem rather than a standard classification task, allowing the model to capture the inherent ordered relationships among ideological positions. Fagni and Cresci (2022) leverage unsupervised deep clustering to uncover latent ideological groupings from social media text without relying on manually labeled political categories.

Ranking. Another line of work frames ideology detection as a ranking problem, where entities, such as politicians, or documents, are ordered according to their ideological positions. For example, studies rank politicians and party groups by political leaning based signals extracted from the speeches and/or bills associated to them (Yuan et al., 2026). Also, ranking formulations are particularly useful

when relative comparisons are more reliable than absolute labels, and they naturally connect to downstream applications such as recommendation and bias analysis. For example, Leo et al. (2025) elicit pairwise comparisons from LLMs and aggregate them with a Bradley–Terry model (Bradley and Terry, 1952) to derive a global ideological ordering without requiring predefined label spaces.

3.2 Measuring Political Ideology

Probing. Probing aims to uncover how political ideology is encoded in model representations. This typically involves fine-tuning LLMs for alignment (Zhou et al., 2023) or trains lightweight classifiers (Kim et al., 2025; Zhang, 2025) over frozen representations to assess whether ideological signals can be linearly or nonlinearly extracted. Beyond simple classification probes, some work explores causal and interpretability-driven probing to identify which model components encode ideological distinctions and how these representations vary across contexts or languages. Using Sparse Autoencoders (SAEs), Kabir et al. (2025) show that models with more abstract political features are generally more resistant to ideological steering.

Prompting. Prompting-based methods infer the political tendencies of LLMs by analyzing their responses to carefully designed prompts mapped onto ideological dimensions. Some works evaluate bias and positioning, examining the model consistency of whether LLMs systematically favor certain viewpoints under different prompts (Motoki et al., 2023; Bang et al., 2024). For example, Hartmann et al. (2023) find that ChatGPT aligns most closely with pro-environmental, left-libertarian parties.

Evaluation. Similar to prompting, evaluation focuses not on how ideology is encoded. It aims to measure how well language models reflect ideological tendencies with existing benchmarks or designed tests. For example, PUBLICVIEWS (Bernardelle et al., 2024) introduces a large-scale collection of statements to study whether the generated framing leans left/right or favors particular political perspectives. In addition, Political Compass Test (PCT) has been widely adopted as a testbed for evaluating values, ideological tendencies and political opinions in LLMs (Röttger et al., 2024; Helwe et al., 2025).

3.3 Intervening Political Ideology

Steering. Recent work explores methods for steering the ideological tendencies and values of LLMs. In particular, instruction tuning leverages persona-conditioned data to guide model behavior toward specific political positions (Bernardelle et al., 2024). Other approaches manipulate hidden representations or activation directions to achieve finer-grained control over generated outputs (Kim et al., 2025; Nadeem et al., 2026). For example, Nadeem et al. (2025) leverage contrastive learning over multilingual statements expressing opposing ideological stances to derive a bias-direction vector for steering. Nadeem et al. (2026) propose a method named Cross-Lingual Alignment Steering (CLAS) to map language-specific steering vectors into a shared ideological subspace.

Debiasing. Existing works explore multiple strategies for mitigating ideological bias in LLMs, including reinforcement learning and calibration methods (Liu et al., 2021), prompt variation (Jenny et al., 2023), self-supervised debiasing with adaptive noise injection (Guo et al., 2024), etc. Recent work by Piao et al. (2025) adopts a multi-agent cognitive framework using fact elicitation and counterfactual reasoning. In addition, predictive modeling approaches analyze how ideological dissimilarity relates to prejudice toward social groups, providing insights for designing stronger debiasing guardrails (Kröger and Barkett, 2025).

Overall, these categories capture different ways of eliciting, analyzing, and intervening ideological tendencies in language models. The full representative methods under each category are discussed in Appendix B.

4 LLMs for Political Ideology

We next review how LLMs are used in political ideology analysis. Compared with earlier models, LLMs provide a flexible framework for ideological detection, measurement, and intervention.

Prompting. Prompting is the most common paradigm for evaluating ideology in LLMs. By presenting political questionnaires, controversial prompts, or simulated dialogues, researchers elicit ideological positions and analyze model responses. Zero-shot and few-shot prompting are widely used for ideology measurement, opinion tracking, and persona simulation (Liu et al., 2024; Milano et al., 2026; Yuan et al., 2026).

Instruction Tuning and Fine-tuning. Instruction-tuned LLMs are increasingly used to align model behavior with political tasks and improve responsiveness to nuanced ideological queries (Chen et al.,

2024). Compared with base models, instruction tuning often improves performance in long-form ideological generation. Fine-tuning adapts pre-trained LLMs to downstream political ideology tasks using supervised political datasets (Liu et al., 2022; Zhou et al., 2023; Paschalides et al., 2025). This approach is widely applied to ideology classification, stance detection, and bias analysis, where labeled political corpora are available.

Representation analysis. A growing line of work studies political ideology through analyze the learned model representations. These methods analyze hidden representations of LLMs using linear probes or lightweight classifiers to identify whether ideological dimensions are encoded in intermediate activations (Kim et al., 2025; Zhang, 2025). Such probing enables researchers to move beyond observable outputs and investigate where ideological information is represented internally, how it varies across layers, and how such representations relate to political scales or value dimensions.

Alignment. Beyond analysis, recent work explores aligning ideological behavior in LLMs. Representative approaches include feature-based steering (Nadeem et al., 2025), Direct Preference Optimization (DPO) (Liu et al., 2021; Schulz and Bhui, 2024), activation editing (Nadeem et al., 2026), etc. These methods aim to shift model outputs toward target ideological positions or mitigate partisan bias while preserving fluency and task performance. This line of work connects ideology analysis with broader questions of fairness and controllable generation.

Overall, LLMs provide a unified framework for political ideology analysis, spanning ideological detection, measurement, and intervention, while also introducing new challenges around robustness, alignment, and cross-cultural generalization.

5 Benchmarks

As shown in Table 1, political ideology benchmarks have evolved from simple ideology classification datasets to broader resources covering stance detection, political bias, polarization, and debiasing. Early benchmarks such as the Ideological Books Corpus (IBC) (Iyyer et al., 2014) and the MANIFESTO Project (Gemenis, 2013) focus on ideology prediction from books and political manifestos. Shared tasks including SemEval-2016 Task6 (Mohammad et al., 2016) and SemEval-2019 Task4 (Kiesel et al., 2019) establish standard benchmarks for stance detection and hyperpartisan news detection. Later efforts shift toward more advanced ideology analysis, including PUBLICVIEWS (Yoo and Shin, 2025) for fine-grained public opinion tracking and the PCT dataset (Nadeem et al., 2025) for PCT-based debiasing.

Most political ideology datasets are sourced from social media (e.g., Twitter) and news media (e.g. AllSlides), serving as critical means for political communication and public opinion formation. Among them, datasets such as BEYOND BINARY LABELS (Preotiuc-Pietro et al., 2017), P-STANCE (Li et al., 2021), POLITWEETS (Baran et al., 2022) support political stance and ideology prediction from user-generated content. In parallel, news-oriented benchmarks such as BASIL (Fan et al., 2019) and ARTICLE-BIAS-PREDICTION (Baly et al., 2020b) enable fine-grained analysis of media bias and framing.

Recent work has increasingly emphasized multilingual and multimodal political understanding. X-STANCE (Vamvas and Sennrich, 2020), POLICORPUS-2020 (García-Díaz et al., 2021), and Laabar and Zaghouani (2024) extend political NLP beyond English to German, French, Spanish, and Arabic settings. Multimodal datasets such as BN-IMGCAP (Qiu et al., 2022), TMPT (Liang et al., 2024), and TWINVIEWS-13K (Fulay et al., 2024) incorporate images and videos alongside text to study ideological framing and debiasing.

Overall, benchmark construction has progressed from binary ideology classification toward richer tasks. However, existing datasets remain heavily concentrated on English-speaking political contexts and social media platforms, highlighting the need for broader geopolitical coverage, culturally diverse benchmarks, and more robust evaluation settings for political reasoning in LLMs. We detail the discussion in Section 7.2.

6 Evaluation Metrics

Classification Metrics. Many of the tasks, such as ideology prediction, stance detection, bias detection, and polarization classification, fall into a standard classification setting. Henceforth, the evaluation metrics include Accuracy, Precision, Recall, (Micro/Macro) F1-score, AUC-ROC which measure overall predictive performance and class-level effectiveness (Iyyer et al., 2014; Liu et al., 2022). Since political datasets are often imbalanced, Macro-F1 is frequently adopted to provide

Dataset	Language	Source	URL
Ideology Prediction / Detection			
CONVOTE (Thomas et al., 2006)	English	debate transcripts	https://www.cs.cornell.edu/home/llee/data/convote.html
MANIFESTO PROJECT (Gemenis, 2013)	Multilingual	party manifestos	https://manifestoproject.wzb.eu/
IBC (Iyyer et al., 2014)	English	books	https://www.cs.umd.edu/~miyyer/ibc/index.html
Bakshy et al. (2015)	English	Facebook	https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/LDJ7MS
BEYONDBINARYLABELS (Preotiuc-Pietro et al., 2017)	English	Twitter profiles	https://www.preotiuc.ro/
MBFC (Baly et al., 2018)	English	news articles	https://github.com/ramybaly/News-Media-Reliability/
Thomas and Kovashka (2019)	English	political ads	https://people.cs.pitt.edu/~chris/politics/
TIMME (Xiao et al., 2020)	English	Twitter	https://github.com/PatriciaXiao/TIMME
POLICORPUS-2020 (García-Díaz et al., 2021)	Spanish	Twitter	https://github.com/NLP-UMUTeam/Spanish-PoliCorpus-2020
CHES (Jolly et al., 2022)	Multilingual	expert surveys	https://www.chesdata.eu/2024-chapel-hill-expert-survey-ches
MITWEET (Liu et al., 2023)	English	Twitter	https://github.com/LST1836/MITweet
POLITISKY24 (Rostami et al., 2025)	English	global news media	https://zenodo.org/records/15616911
Stance Detection			
SEMEVAL2016 TASK6 (Mohammad et al., 2016)	English	Twitter	https://alt.qcri.org/semeval2016/task6/
X-STANCE (Vamvas and Sennrich, 2020)	German, French, Italian, English	voting platform	https://github.com/ZurichNLP/xstance
VAST (Allaway and McKeown, 2020)	English	news / social media	https://github.com/CASIA-IVA-Lab/VAST
P-STANCE (Li et al., 2021)	English	Twitter	https://github.com/chuchun8/PStance
BIGNEWS (Liu et al., 2022)	English	news	https://github.com/launchnlp/POLITICS
POLITWEETS (Baran et al., 2022)	Polish	Twitter	https://github.com/Joanna065/politweets
Laabar and Zaghouani (2024)	Arabic	news / social media	https://forms.gle/S9fZtYjAyLAqFsH19
PUBLICVIEWS (Yoo and Shin, 2025)	English	comments / forums	https://github.com/yoojuneho/Fair-or-Framed/tree/main
Bias, Polarization, and Framing			
Kulkarni et al. (2018)	English	news articles	–
BASIL (Fan et al., 2019)	English	news articles	https://github.com/launchnlp/BASIL
SEMEVAL2019 TASK4 (Kiesel et al., 2019)	English	web articles / news	https://pan.webis.de/semeval19/semeval19-web/
ARTICLE-BIAS-PREDICTION (Baly et al., 2020b)	English	AllSides news	https://github.com/ramybaly/Article-Bias-Prediction
POLITICAL-POLARIZATION (Sinno et al., 2022)	English	news	https://github.com/bernovie/political-polarization
Zhou and Tan (2023)	English	news articles	https://forms.gle/S9fZtYjAyLAqFsH19
TWINVIEWS-13K (Fulay et al., 2024)	English	paired news articles	https://huggingface.co/datasets/wwbrannon/twinviews-13k
EUROPARLVOTE (Yang et al., 2025)	Multilingual	European Parliament	https://huggingface.co/datasets/unimelb-nlp/EuroParlVote
PCT (Nadeem et al., 2025)	English	Reddit / Twitter	https://github.com/Afx-Msh/SVE_Mitigation
Multimodal, Speech, and Image Dataset			
Zannettou et al. (2019)	–	Twitter	–
BN-IMGCAP (Qiu et al., 2022)	English	news images	https://github.com/launchnlp/mulmodide
TMPT (Liang et al., 2024)	English	social media / video	https://github.com/Leon-Francis/Multi-Modal-Stance-Detection
AGORASPEECH (Sermpezis et al., 2025)	English, Greek	debates / audio	https://github.com/Datalab-AUTH/AgoraSpeech-EDA

Table 1: An overview of common political ideology-related benchmarks, grouped by task.

balanced evaluation across ideological groups. For multi-party or multi-ideology settings, Top-k Accuracy is also employed to evaluate whether the correct ideology appears among the top predicted candidates. For regression-style ideology estimation (e.g., continuous ideology scales), studies often use standard regression metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Pearson/Spearman Correlation, etc (Aanisha Bhattacharyya, 2026; Yuan et al., 2026).

Generation Metrics. These metrics help to evaluate how effectively a model can generate text aligned with a specific political persona, ideology, or viewpoint. Hence, they are used in tasks such as ideological steering, political persona simulation, and statement rewriting. A widely used metric is *Target Shift* or *Delta* (Δ), which measures the change in ideological position before and after applying a steering intervention, reflecting how successfully the output moves toward the target political orientation (Bernardelle et al., 2025). In addition, Perplexity (PPL) is often used to evaluate fluency and coherence, ensuring that political steering does not substantially degrade language quality (Liu et al., 2021; Jiang et al., 2025).

Fairness Evaluation Metrics. Since political NLP systems may unintentionally amplify ideological imbalance and partisan bias, fairness evaluation has become increasingly important. A central dimension is ideological fairness, which assesses whether models systematically favor or disadvantage particular political viewpoints. Common evaluation criteria include disparities in False Positive and False Negative rates across ideological groups, calibration consistency between political affiliations, and prediction parity, which measures whether predictions are distributed equitably across ideological categories (Guo et al., 2024). Recent work has also emphasized demographic and cultural fairness in political ideology analysis, examining whether models remain robust and unbiased across different languages, regions, and cultural backgrounds. Representative fairness metrics in this domain include Bias Score / Skewness, Pearson correlation, Av-

Category	Representative Evaluation Metrics	Tasks
Classification	Accuracy, Precision, Recall, Macro/Micro-F1, AUC-ROC, Top-k Accuracy, MAE, MSE, ...	Ideology Prediction, Stance Detection, Bias Detection, Polarization Classification, Ideology Scoring
Generation	Target Shift (Δ), Perplexity (PPL), ...	Ideological Steering, Political Persona Simulation, Statement Rewriting
Fairness	Equalized Odds, Demographic Parity, Prediction Parity, Group-wise Macro-F1, Bias Score, Average Euclidean Distance, ...	Ideological Fairness Evaluation, Bias Analysis
Ranking	NDCG, MRR, Spearman's ρ, Kendall's τ, ...	Party-wise Ranking, Political Retrieval, MP Ranking, ...
Human Assessment	Topic relevance, Persona Fidelity, Fluency, Civility, Toxicity, ...	Persona Evaluation, Ideology Validation, Debiasing Assessment

Table 2: Representative evaluation metrics across political ideology analysis tasks.

erage Euclidean distance, Equalized Odds, Demographic Parity, and Group-wise Macro-F1 (Zhang, 2025; Rettenberger et al., 2024).

Ranking Metrics. Ranking metrics are commonly used in political ideology tasks that require ordering politically relevant items, such as ideological viewpoints, news articles, or candidate responses, based on relevance or ideological alignment (Yuan et al., 2026). These metrics include Normalized Discounted Cumulative Gain (NDCG) and Mean Reciprocal Rank (MRR), which evaluate ranking quality while accounting for the positions of relevant items in a ranked list. In addition, Spearman's ρ and Kendall's τ measure the similarity between the predicted and ground-truth ordering.

Human Assessment. As shown in Table 2, beyond automatic evaluation, human evaluation plays an important role in assessing political ideology tasks, particularly for subjective dimensions that are difficult to capture with standard methods and metrics. In ideology detection, human annotators are commonly used to measure inter-annotator agreement, perceived extremism, and rhetorical framing to validate whether model predictions align with human political intuition (Aanisha Bhattacharyya, 2026). For ideology measurement and simulation tasks, human evaluation often focuses on persona fidelity, demographic authenticity, and cross-cultural consistency, examining whether generated outputs accurately reflect targeted socio-political perspectives (Törnberg, 2023). In ideology intervention and debiasing tasks, annotators typically assess linguistic fluency, topic relevance, civility, and toxicity to ensure that steering methods preserve conversational quality and safety while reducing ideological bias (Liu et al., 2021).

7 Discussions

7.1 Potential Gaps

Despite the increasing attention to political ideology in NLP, several gaps remain. First, much existing work relies on simplified proxies for ideology, including left–right labels, party affiliation, media-source labels, or issue-specific stance labels (Laver et al., 2003; Slapin and Proksch, 2008). While these proxies are useful for prediction, they risk conflating ideology with adjacent but distinct concepts such as party identity, issue position, media bias, or polarization. As a result, models may learn surface-level political associations rather than theoretically meaningful ideological structures.

Second, while predictive performance remains essential for evaluating NLP models, these efforts would benefit from greater emphasis on construct validity and complementary evaluations of whether models capture ideology as understood in political science. Rather than validating ideology models only against task-specific labels, researchers can compare model outputs with established political science resources, including expert surveys, party manifesto positions, roll-call based ideal points, parliamentary voting records, and legislative speech corpora (Kam, 2001; Clinton et al., 2004; Laver et al., 2003; Slapin and Proksch, 2008). Large-scale resources such as the Chapel Hill Expert Survey (CHES) (Jolly et al., 2022) or CONVOTE (Thomas et al., 2006) provide useful anchors for assessing whether NLP models recover theoretically meaningful ideological positions.

Third, existing datasets remain concentrated in English-language, U.S.-centered social media and news settings, limiting generalizability to multilin-

gual, multiparty, and institutional contexts where ideology may be expressed and contested differently. More work is needed on legislative debates, policy documents, parliamentary voting, campaign materials, and other political data (Slapin and Proksch, 2008; Lauderdale and Herzog, 2016).

Finally, intervention-oriented work often assumes that ideological neutrality is a straightforward objective. Yet neutrality is politically and culturally contested. We then discuss the future work directions to fill the mentioned gaps.

7.2 Future Directions

Ideology in Real Political Data. As shown in Table 1, political ideology research remains dominated by social media and news, especially from platforms such as Twitter and AllSides. While these sources provide large-scale and accessible political discourse, they often reflect highly polarized, informal, and platform-specific behavior that may not represent broader political processes. As a result, models may overfit to partisan online language and struggle to capture ideology in institutional and societal contexts. The polarized nature of these datasets may also limit models' ability to identify subtler ideological signals, such as those found in parliamentary discourse. Future research should place greater emphasis on real political data from real political data such as legislative debates, parliamentary voting records, etc. Compared with social media, these sources offer richer and more structured representations of ideology, including policy preferences, coalition dynamics, and long-term ideological change. Their complexity can support more capable methods and provide deeper insight into the biases captured by LLMs.

Multimodal and Multilingual Extension. Political communication is inherently multimodal, combining images, videos, speeches, and other media with text (Thomas and Kovashka, 2019). Integrating these signals across levels of abstraction is essential for understanding elections, referenda, and emerging political movements. Although some work has incorporated images (Qiu et al., 2022) and audio (Sermpezis et al., 2025), most ideology research remains text-centric, limiting models' ability to capture modern political communication. Future work should move toward advanced political reasoning that jointly models multimodal signals. Key directions include multimodal propaganda detection, meme ideology analysis, cross-modal stance understanding, and visual political framing. At the same time, multimodal LLMs introduce new challenges, including bias amplification, manipulated visual narratives, and culturally dependent symbolism, underscoring the need for robust multimodal datasets and cross-modal evaluation frameworks.

In addition, as discussed in Section 7.1, the limited cross-cultural and cross-lingual generalizability of ideology poses a challenge for current models, which still struggle with culture-specific rhetoric, historical references, and localized framing, while multilingual benchmarks remain largely focused on prediction. Future work should develop culturally adaptive political representations, cross-lingual ideology alignment, and multilingual generation, supported by diverse multilingual datasets and culturally aware evaluation frameworks.

Fairness and Explainability. As LLMs become increasingly involved in political applications, ensuring fairness and neutrality becomes critical. Future research should focus on mitigating ideological bias, improving demographic and cultural fairness, and preventing harmful political manipulation, misinformation, and polarization amplification. Robust evaluation frameworks for political fairness and safety are still needed. In addition, political reasoning often involves subjective and socially sensitive decisions, making explainability particularly important. Future work should also develop interpretable political NLP systems capable of providing transparent rationales and evidence for ideological predictions. Human-centered evaluation protocols that account for political diversity, cultural perspectives, and annotation disagreement will also become increasingly important.

8 Conclusion

In this survey, we review political ideology research in NLP through three major directions: ideology detection, ideology measurement, and ideology intervention. We summarize the tasks in each direction and discuss the role of LLMs across them. We then list benchmark datasets and evaluation protocols under different perspectives. Although recent progress has significantly advanced political ideology understanding in NLP, current research remains heavily focused on English-centric, text-based, and social media-driven settings. Future work should therefore move toward more multilingual, cross-cultural, multimodal, and real-world political reasoning while improving fairness, interpretability, and safety in political NLP systems.

Limitations

Our survey primarily focuses on computational and NLP perspectives on political ideology, rather than broader theories and methodologies from political science. As a result, some sociopolitical, historical, and institutional aspects of ideology, as well as relevant work from the broader political science literature, may not be fully covered. In addition, the survey mainly centers on ideological analysis, and therefore does not comprehensively address broader political NLP topics beyond ideology.

References

Yaman Kumar Singla Rajiv Ratn Shah Changyou Chen Balaji Krishnamurthy Aanisha Bhattacharyya, Abhilekh Borah. 2026. Social agents: Collective intelligence improves LLM predictions. In *ICLR*.

Abeer Aldayel and Walid Magdy. 2020. Stance detection on social media: State of the art and trends. *Inf. Process. Manag.*, 58:102597.

Emily Allaway and Kathleen McKeown. 2020. Zero-shot stance detection: A dataset and model using generalized topic representations. In *Conference on Empirical Methods in Natural Language Processing*.

Goshi Aoki. 2024. Large language models in politics and democracy: A comprehensive survey. *ArXiv*, abs/2412.04498.

Eytan Bakshy, Solomon Messing, and Lada A. Adamic. 2015. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348:1130 – 1132.

R. Baly, Georgi Karadzhov, Dimitar Alexandrov, James R. Glass, and Preslav Nakov. 2018. Predicting factuality of reporting and bias of news media sources. In *Conference on Empirical Methods in Natural Language Processing*.

R. Baly, Georgi Karadzhov, Jisun An, Haewoon Kwak, Yoan Dinkov, Ahmed Ali, James R. Glass, and Preslav Nakov. 2020a. What was written vs. who read it: News media profiling using text analysis and social media context. *ArXiv*, abs/2005.04518.

R. Baly, Georgi Karadzhov, Abdelrhman Saleh, James R. Glass, and Preslav Nakov. 2019. Multi-task ordinal regression for jointly predicting the trustworthiness and the leading political ideology of news media. In *North American Chapter of the Association for Computational Linguistics*.

R. Baly, Giovanni Da San Martino, James R. Glass, and Preslav Nakov. 2020b. We can detect your bias: Predicting the political ideology of news articles. In *Conference on Empirical Methods in Natural Language Processing*.

Yejin Bang, Delong Chen, Nayeon Lee, and Pascale Fung. 2024. Measuring political bias in large language models: What is said and how it is said. *ArXiv*, abs/2403.18932.

Joanna Baran, Michal Kajstura, Maciej Zi'olkowski, and Krzysztof Rajda. 2022. Does twitter know your political views? politweets dataset and semi-automatic method for political leaning discovery. In *POLITICALNLP*.

Pablo Barberá. 2015. Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political Analysis*, 23(1):76–91.

Pietro Bernardelle, Stefano Civelli, Leon Fröhling, Riccardo Lunardi, Kevin Roitero, and Gianluca Demartini. 2025. Political ideology shifts in large language models. *ArXiv*, abs/2508.16013.

Pietro Bernardelle, Leon Fröhling, Stefano Civelli, Riccardo Lunardi, Kevin Roitero, and Gianluca Demartini. 2024. Mapping and influencing the political ideology of large language models using synthetic personas. *Companion Proceedings of the ACM on Web Conference 2025*.

Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs the method of paired comparisons. *Biometrika*, 39:324–345.

Maarten Buyl, Alexander Rogiers, Sander Noels, Iris Dominguez-Catena, Edith Heiter, Raphaël Romero, Iman Johary, Alexandru Cristian Mara, Jefrey Lijffijt, and Tijl De Bie. 2024. Large language models reflect the ideology of their creators. *Npj Artificial Intelligence*, 2.

Tanise Ceron, Dmitry Nikolaev, Dominik Stammbach, and Debora Nozza. 2025. What is the political content in llms' pre- and post-training data? *ArXiv*, abs/2509.22367.

Ilias Chalkidis and Stephanie Brandl. 2024. Llama meets eu: Investigating the european political spectrum through the lens of LLMs. *ArXiv*, abs/2403.13592.

M. Chandan and Shrabanti Mandal. 2025. A comprehensive survey on sentiment analysis: Framework, techniques, and applications. *Comput. Sci. Rev.*, 58:100777.

Kakia Chatsiou and Slava Jankin. 2020. Deep learning for political science. *ArXiv*, abs/2005.06540.

Chen Chen, D. Walker, and Venkatesh Saligrama. 2023. Filtering context mitigates scarcity and selection bias in political ideology prediction. *ArXiv*, abs/2302.00239.

Kai Chen, Zihao He, Jun Yan, Taiwei Shi, and Kristina Lerman. 2024. How susceptible are large language models to ideological manipulation? *ArXiv*, abs/2402.11725.

Wei Chen, Xiao Zhang, Tengjiao Wang, Bishan Yang, and Yi Li. 2017. Opinion-aware knowledge graph for political ideology detection. In *International Joint Conference on Artificial Intelligence*.

Joshua Clinton, Simon Jackman, and Douglas Rivers. 2004. The statistical analysis of roll call data. *American Political Science Review*, pages 355–370.

Franco Demarco, Juan Manuel Ortiz de Zarate, and Esteban Feuerstein. 2023. Measuring ideological spectrum through nlp. In *NL4AI@AI*IA*.

Sami Diaf. 2023. Communityfish: A poisson-based document scaling with hierarchical clustering. *ArXiv*, abs/2308.14873.

Tiziano Fagni and Stefano Cresci. 2022. Fine-grained prediction of political leaning on social media with unsupervised deep learning. *J. Artif. Intell. Res.*, 73:633–672.

Lisa Fan, Marshall Harold White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. 2019. In plain sight: Media bias through the lens of factual reporting. *ArXiv*, abs/1909.02670.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair nlp models. In *Annual Meeting of the Association for Computational Linguistics*.

Jillian R. Fisher, Ruth Elisabeth Appel, Chan Young Park, Yujin Potter, Liwei Jiang, Taylor Sorensen, Shangbin Feng, Yulia Tsvetkov, Margaret E. Roberts, Jennifer Pan, Dawn Xiaodong Song, and Yejin Choi. 2025. Political neutrality in ai is impossible- but here is how to approximate it. *ArXiv*, abs/2503.05728.

James H Fowler. 2006. Connecting the congress: A study of cosponsorship networks. *Political Analysis*, 14(4):456–487.

Suyash Pradeep Fulay, William Brannon, Shrestha Mohanty, Cassandra Overney, Elinor Poole-Dayan, Deb Roy, and Jad Kabbara. 2024. On the relationship between truth and political bias in language models. *ArXiv*, abs/2409.05283.

José Antonio García-Díaz, Ricardo Colomo-Palacios, and Rafael Valencia-García. 2021. Psychographic traits identification based on political ideology: An author analysis study on spanish politicians' tweets posted in 2020. *Future Gener. Comput. Syst.*, 130:59–74.

Kostas Gemenis. 2013. What to do (and not to do) with the comparative manifestos project data. *Political Studies*, 61:23 – 3.

S Gnanapriya and K. Anandan. 2025. Sentiment analysis in political discourse: Understanding public opinion from social media. *E-Learning and Digital Media*.

Yufei Guo, Muzhe Guo, Juntao Su, Zhou Yang, Mengqiu Zhu, Hongfei Li, Mengyang Qiu, and Shuo Liu. 2024. Bias in large language models: Origin, evaluation, and mitigation. *ArXiv*, abs/2411.10915.

Suchin Gururangan, Dallas Card, Sarah K. Drier, Emily Kalah Gade, Leroy Z. Wang, Zeyu Wang, Luke Zettlemoyer, and Noah A. Smith. 2022. Whose language counts as high quality? measuring language ideologies in text data selection. *ArXiv*, abs/2201.10474.

Muhammad Haroon, Magdalena Wojcieszak, and Anshuman Chhabra. 2025. "whose side are you on?" estimating ideology of political and news content using large language models and few-shot demonstration selection. In *IJCNLP-AACL*.

Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. 2023. The political ideology of conversational ai: Converging evidence on chatgpt's pro-environmental, left-libertarian orientation. *ArXiv*, abs/2301.01768.

Zihao He, Siyi Guo, Ashwin Rao, and Kristina Lerman. 2023. Inducing political bias allows language models anticipate partisan reactions to controversies. *ArXiv*, abs/2311.09687.

Chadi Helwe, Oana Balalau, and Davide Ceolin. 2025. Navigating the political compass: Evaluating multilingual LLMs across languages and nationalities. In *Annual Meeting of the Association for Computational Linguistics*.

Mohit Iyyer, Peter J. Enns, Jordan Lee Boyd-Graber, and Philip Resnik. 2014. Political ideology detection using recursive neural networks. In *Annual Meeting of the Association for Computational Linguistics*.

David F. Jenny, Yann Billeter, Mrinmaya Sachan, Bernhard Schölkopf, and Zhijing Jin. 2023. Navigating the ocean of biases: Political bias attribution in language models via causal structures. *ArXiv*, abs/2311.08605.

Chunyan Jiang, Lijun Zhang, and Yayun Wang. 2025. Integrating natural language processing techniques with lda topic modeling: Deconstructing the evolution of media discourse on educational modernization ideology. *Education and Information Technologies*, 30:23419 – 23466.

Julie Jiang, Xiang Ren, and Emilio Ferrara. 2022. Retweet-bert: Political leaning detection using language features and information diffusion on social networks. In *International Conference on Web and Social Media*.

Seth Jolly, Ryan Bakker, Liesbet Hooghe, Gary Marks, Jonathan Polk, Jan Rovny, Marco Steenbergen, and Milada Anna Vachudova. 2022. Chapel hill expert survey trend file, 1999–2019. *Electoral studies*, 75:102420.

John T. Jost, Christopher M. Federico, and Jaime L. Napier. 2009. Political ideology: its structure, functions, and elective affinities. *Annual review of psychology*, 60:307–37.

Shariar Kabir, Kevin Esterling, and Yue Dong. 2025. Beyond the surface: Probing the ideological depth of large language models. *ArXiv*, abs/2508.21448.

Christopher Kam. 2001. Do ideological preferences explain parliamentary behaviour? evidence from great britain and canada. *Journal of Legislative Studies*, 7(4):89–126.

Sadia Kamal, Brenner Little, Jade Gullic, Trevor Harms, Kristin L. Olofsson, and Arunkumar Bagavathi. 2023. Modeling political orientation of social media posts: An extended analysis. *ArXiv*, abs/2311.12323.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel M. Vincent, Payam Adineh, D. Corney, Benno Stein, and Martin Potthast. 2019. Semeval-2019 task 4: Hyperpartisan news detection. In *International Workshop on Semantic Evaluation*.

Junsol Kim, James Evans, and Aaron Schein. 2025. Linear representations of political perspective emerge in large language models. *ArXiv*, abs/2503.02080.

Nora Kirkizh, Roberto Ulloa, Sebastian Stier, and Jurgen Pfeffer. 2024. Predicting political attitudes from web tracking data: a machine learning approach. *Journal of Information Technology & Politics*, 21:564 – 577.

Paul Kröger and Emilio Barkett. 2025. Don't change my view: Ideological bias auditing in large language models. *ArXiv*, abs/2509.12652.

Vivek Kulkarni, Junting Ye, Steven S. Skiena, and William Yang Wang. 2018. Multi-view models for political ideology detection of news articles. In *Conference on Empirical Methods in Natural Language Processing*.

Sanaa Laabar and Wajdi Zaghouani. 2024. Multidimensional insights: Annotated dataset of stance, sentiment, and emotion in facebook comments on tunisia's july 25 measures. *Proceedings of the Second Workshop on Natural Language Processing for Political Sciences @ LREC-COLING 2024*.

Benjamin E Lauderdale and Alexander Herzog. 2016. Measuring political positions from legislative speech. *Political Analysis*, 24(3):374–394.

Michael Laver, Kenneth Benoit, and John Garry. 2003. Extracting policy positions from political texts using words as data. *American Political Science Review*, 97(02):311–331.

Riccardo Di Leo, Chen Zeng, Elias Dinas, and Reda Tamtam. 2025. Mapping (a)ideology: A taxonomy of european parties using generative LLMs as zero-shot learners. *Political Analysis*.

Ang Li, Jingqian Zhao, Bin Liang, Lin Gui, Hui Wang, Xi Zeng, Xingwei Liang, Kam-Fai Wong, and Ruifeng Xu. 2024a. Mitigating biases of large language models in stance detection with counterfactual augmented calibration. In *North American Chapter of the Association for Computational Linguistics*.

Lincan Li, Jiaqi Li, Catherine Chen, Fred Gui, Hongjian Yang, Chenxiao Yu, Zhengguang Wang, Jianlong Cai, Junlong Zhou, Bolin Shen, Alexander Qian, Weixin Chen, Zhongkai Xue, Lichao Sun, Lifang He, Hanjie Chen, Kaize Ding, Zijia Du, Fangzhou Mu, and 28 others. 2024b. Political-LLM: Large language models in political science. *ArXiv*, abs/2412.06864.

Yingjie Li, Tiberiu Sosea, Aditya Sawant, Ajith Jayaraman Nair, Diana Inkpen, and Cornelia Caragea. 2021. P-stance: A large dataset for stance detection in political domain. In *Findings*.

Bin Liang, Ang Li, Jingqian Zhao, Lin Gui, Min Yang, Yue Yu, Kam-Fai Wong, and Ruifeng Xu. 2024. Multi-modal stance detection: New datasets and model. *ArXiv*, abs/2402.14298.

Xinru Lin and Luyang Li. 2025. Implicit bias in llms: A survey. *ArXiv*, abs/2503.02776.

Mitchell Linegar, Rafal Kocielnik, and R. Michael Alvarez. 2023. Large language models and political science. *Frontiers in Political Science*.

Andy Liu, Mona T. Diab, and Daniel Fried. 2024. Evaluating large language model biases in persona-steered generation. *ArXiv*, abs/2405.20253.

Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, Lili Wang, and Soroush Vosoughi. 2021. Mitigating political bias in language models through reinforced calibration. *ArXiv*, abs/2104.14795.

Songtao Liu, Ziling Luo, Minghua Xu, Lixiao Wei, Ziyao Wei, Han Yu, Wei Xiang, and Bang Wang. 2023. Ideology takes multiple looks: A high-quality dataset for multifaceted ideology detection. In *Conference on Empirical Methods in Natural Language Processing*.

Yujian Liu, Xinliang Frederick Zhang, David Wegsman, Nick Beauchamp, and Lu Wang. 2022. Politics: Pretraining with same-story article comparison for ideology prediction and stance detection. *ArXiv*, abs/2205.00619.

Hanjia Lyu and Jiebo Luo. 2022. Understanding political polarization via jointly modeling users, connections and multimodal contents on heterogeneous graphs. *Proceedings of the 30th ACM International Conference on Multimedia*.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Ani Saxena, Kristina Lerman, and A. G. Galstyan. 2019. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54:1 – 35.

Emilia Milano, Alistair Plum, Yves Scherrer, and Christoph Purschke. 2026. Language ideologies in a multilingual society: An LLM-based analysis of luxembourgish news comments.

Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiao-Dan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *International Workshop on Semantic Evaluation*.

Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 2023. More human than human: measuring chatgpt political bias. *Public Choice*, 198:3–23.

Afrozah Nadeem, Mark Dras, and Usman Naseem. 2025. Steering towards fairness: Mitigating political bias in LLMs. *ArXiv*, abs/2508.08846.

Afrozah Nadeem, Agrima Seth, Mehwish Nasim, and Usman Naseem. 2026. Bias beyond borders: Political ideology evaluation and steering in multilingual LLMs. *ArXiv*, abs/2601.23001.

Preslav Nakov, Husrev Taha Sencar, Jisun An, and Haewoon Kwak. 2021. A survey on predicting the factuality and the bias of news media. *ArXiv*, abs/2103.12506.

Nishanth Nakshatri, Nikhil Mehta, Siyi Liu, Sihao Chen, Daniel J. Hopkins, Dan Roth, and Dan Goldwasser. 2025. Talking point based ideological discourse analysis in news events. *ArXiv*, abs/2504.07400.

Lata Pangtey, Anukriti Bhatnagar, Shubhi Bansal, Shahid Shafi Dar, and Nagendra Kumar. 2025. Large language models meet stance detection: A survey of tasks, methods, applications, challenges and future directions. *ArXiv*, abs/2505.08464.

Demetris Paschalides, George Pallis, and Marios D. Dikaiakos. 2025. Probing the subtle ideological manipulation of large language models. *ArXiv*, abs/2504.14287.

Tai-Quan Peng, Kaiqi Yang, Sanguk Lee, Hang Li, Yucheng Chu, Yuping Lin, and Hui Liu. 2024a. Beyond partisan leaning: a comparative analysis of political bias in large language models. *Journal of Information Technology & Politics*.

Xing Qiang Peng, Zhenkun Zhou, Chong Zhang, and Ke Xu. 2024b. Online social behavior enhanced detection of political stances in tweets. In *International Conference on Web and Social Media*.

Ethan Perez, Sam Ringer, Kamilė Lukoiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Chris Olah, Daisong Yan, Daniela Amodei, and 44 others. 2022. Discovering language model behaviors with model-written evaluations. In *Annual Meeting of the Association for Computational Linguistics*.

Jing Piao, Zhihong Lu, Chen Gao, and Yong Li. 2025. Social bots meet large language model: Political bias and social learning inspired mitigation strategies. *Proceedings of the ACM on Web Conference 2025*.

Keith T Poole and Howard Rosenthal. 1985. A spatial model for legislative roll call analysis. *American Journal of Political Science*, pages 357–384.

Daniel Preotiuc-Pietro, Ye Liu, Daniel J. Hopkins, and Lyle H. Ungar. 2017. Beyond binary labels: Political ideology prediction of twitter users. In *Annual Meeting of the Association for Computational Linguistics*.

Changyuan Qiu, Winston Wu, Xinliang Frederick Zhang, and Lu Wang. 2022. Late fusion with triplet margin objective for multimodal ideology prediction and analysis. In *Conference on Empirical Methods in Natural Language Processing*.

Luca Rettenberger, Markus Reischl, and Mark Schutera. 2024. Assessing political bias in large language models. *Journal of Computational Social Science*, 8.

Peyman Rostami, Vahid Rahimzadeh, Ali Adibi, and Azadeh Shakery. 2025. Politisky24: U.s. political bluesky dataset with user stance labels. In *Conference on Empirical Methods in Natural Language Processing*.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Rose Kirk, Hinrich Schutze, and Dirk Hovy. 2024. Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. In *Annual Meeting of the Association for Computational Linguistics*.

Lion Schulz and Rahul Bhui. 2024. Political reinforcement learners. *Trends in cognitive sciences*.

Pavlos Sermpezis, Stelios Karamanidis, Eva Paraschou, Ilias Dimitriadis, Sofia Yfantidou, Filitsa-Ioanna Kouskouveli, Thanasis Troboukis, Kelly Kiki, Antonis Galanopoulos, and Athena Vakali. 2025. Agoraspeech: A multi-annotated comprehensive dataset of political discourse through the lens of humans and ai. *ArXiv*, abs/2501.06265.

Yanchuan Sim, Brice D. L. Acree, Justin H. Gross, and Noah A. Smith. 2013. Measuring ideological proportions in political speeches. In *Conference on Empirical Methods in Natural Language Processing*.

Barea M. Sinno, Bernardo Oviedo, Katherine Atwell, Malihe Alikhani, and Junyi Jessy Li. 2022. Political ideology and polarization: A multi-dimensional approach. In *North American Chapter of the Association for Computational Linguistics*.

Jonathan B Slapin and Sven-Oliver Proksch. 2008. A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, 52(3):705–722.

Shivashankar Subramanian, Trevor Cohn, and Timothy Baldwin. 2019. Deep ordinal regression for pledge specificity prediction. In *Conference on Empirical Methods in Natural Language Processing*.

Kian Long Tan, Chin Poo Lee, and Kian Ming Lim. 2023. A survey of sentiment analysis: Approaches, datasets, and future research. *Applied Sciences*.

Christopher Thomas and Adriana Kovashka. 2019. Predicting the politics of an image using webly supervised data. In *Neural Information Processing Systems*.

Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Conference on Empirical Methods in Natural Language Processing*.

Petter Törnberg. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *ArXiv*, abs/2304.06588.

Jannis Vamvas and Rico Sennrich. 2020. X-Stance: A multilingual multi-target dataset for stance detection. In *Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS)*, Zurich, Switzerland.

Jan Vávra, Bernd Hans-Konrad Prostmaier, Bettina Grün, and Paul Hofmarcher. 2024. A structural text-based scaling model for analyzing political discourse. *ArXiv*, abs/2410.11897.

Daniel Verdear, Ashley Hemm, Zuoyu Tian, Sara El Oud, Sandra Kübler, John Funchion, Michelle Seelig, Amanda B. Diekman, Manohar N. Murthi, Kamal Premaratne, Neil F. Johnson, and Stefan Wuchty. 2025. Identifying and characterizing ideologically homogeneous clusters on twitter and parler during the 2020 election. *PLOS One*, 20.

Julia Watson, Sophia Lee, Barend Beekhuizen, and Suzanne Stevenson. 2024. Do language models practice what they preach? examining language ideologies about gendered language reform encoded in llms. In *International Conference on Computational Linguistics*.

John D. Wilkerson and Andreu Casas. 2017. Large-scale computerized text analysis in political science: Opportunities and challenges. *Annual Review of Political Science*, 20:529–544.

Patrick Y. Wu, Joshua A. Tucker, Jonathan Nagler, and Solomon Messing. 2023. Large language models can be used to estimate the latent positions of politicians.

Zhiping Xiao, Weiping Song, Haoyan Xu, Zhicheng Ren, and Yizhou Sun. 2020. Timme: Twitter ideology-detection via multi-task multi-relational embedding. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Hao Yan, Allen Lavoie, and Sanmay Das. 2017. The perils of classifying political orientation from text. In *LINKDEM@IJCAI*.

Jinrui Yang, Xudong Han, and Timothy Baldwin. 2025. Demographics and democracy: Benchmarking LLMs' gender bias and political leaning in european parliament. In *International Conference on Natural Language and Speech Processing*.

Joonkyung Yoo and Youhyun Shin. 2025. Fair or framed? political bias in news articles generated by LLMs. In *Conference on Empirical Methods in Natural Language Processing*.

Yifei Yuan, Luis Salamanca, Sophia Schlosser, and Laurence Brandenberger. 2026. Graph-augmented LLMs for swiss mp ideology prediction.

Savvas Zannettou, Tristan Caulfield, Barry Bradlyn, Emiliano De Cristofaro, Gianluca Stringhini, and Jeremy Blackburn. 2019. Characterizing the use of images in state-sponsored information warfare operations by russian trolls on twitter. In *International Conference on Web and Social Media*.

Bowen Zhang, Genan Dai, Fuqiang Niu, Nan Yin, Xiaomao Fan, and Hutchin Huang. 2024. A survey of stance detection on social media: New directions and perspectives. *ArXiv*, abs/2409.15690.

Tianyi Zhang. 2025. Probing political ideology in large language models: How latent political representations generalize across tasks. In *Conference on Empirical Methods in Natural Language Processing*.

Karen Zhou and Chenhao Tan. 2023. Entity-based evaluation of political bias in automatic summarization. In *Conference on Empirical Methods in Natural Language Processing*.

Xiaotian Zhou, Qian Wang, Xiaofeng Wang, Haixu Tang, and Xiaozhong Liu. 2023. Large language model soft ideologization via ai-self-consciousness. *ArXiv*, abs/2309.16167.

A Comparison with existing surveys

Table 3 shows comparison between our survey and existing related surveys. Prior surveys typically focus on individual tasks or broader political NLP phenomena, whereas our survey places political ideology itself at the center and provides a unified perspective spanning detection, measurement, and intervention in modern NLP and LLMs.

B Methods

B.1 Ideology Detection

The goal is to identify the ideological stance of a given user, text, or entity. Typical approaches include:

Name	Survey Category	Political-Centered	Ideology-Centered
Task-specific Surveys			
Aldayel and Magdy (2020)	Stance Detection	Partial	No
Zhang et al. (2024)	Stance Detection	Partial	No
Pangtey et al. (2025)	Stance Detection	Partial	No
Mehrabi et al. (2019)	Bias and Fairness	No	No
Nakov et al. (2021)	News Bias	Partial	No
Guo et al. (2024)	LLM Bias	No	No
Lin and Li (2025)	LLM Bias	No	No
Tan et al. (2023)	Sentiment Analysis	No	No
Chandan and Mandal (2025)	Sentiment Analysis	No	No
General Political Surveys			
Wilkerson and Casas (2017)	Political Text Mining	Yes	No
Chatsiou and Jankin (2020)	Deep Learning for Political Science	Yes	No
Linegar et al. (2023)	LLM in Political Science	Yes	No
Li et al. (2024b)	LLM in Political Science	Yes	No
Aoki (2024)	LLM in Political Science	Yes	No
Our Survey	Political ideology in NLP	Yes	Yes

Table 3: Comparison between our survey and existing related surveys.

- **Lexicon-Based & Feature-Based Machine Learning**. Early papers such as Sim et al. (2013) and BeyondBL (Preotiuc-Pietro et al., 2017) leverage linguistic features, political dictionaries, and stylistic elements paired with linear models or word embeddings (e.g. Word2Vec, GloVe) for modeling and predicting the ideology.

- **Supervised Fine-tuning**. Some detection papers utilize custom fine-tuned language models. For example, RETWEET-BERT (Jiang et al., 2022) and POLITICS (Liu et al., 2022) leverage masked language modeling fine-tuned on political datasets to classify or dynamically scale text along continuous ideological dimensions.

- **GNNS & Network Embeddings**. They are used to model social media ecosystems. Works such as TIMME (Xiao et al., 2020) and SA-GNN (Peng et al., 2024b) use heterogeneous graphs, graph convolutional networks, and user-to-user interaction networks (e.g. retweets, follows) to propagate political labels across social graphs.

- **Generative LLMs as Zero-Shot / Few-Shot Detectors**: Rather than retraining internal weights, researchers write specialized prompts containing diverse political information, from text to other contextual data, and ask the generative LLM to output a classification or continuous coordinate. For example, Milano et al. (2026); Yuan et al. (2026) use LLMs (e.g. Llama, Mistral, or GPT variants) directly as zero-shot text judges to predict an ideology score from different types of context provided in the prompt.

B.2 Ideology Measurement

This branch aims to analyze how deeply models natively understand, encode, and simulate human political worldviews.

- **Linear Probing & Activation Extraction**. Papers such as Zhou et al. (2023); Kim et al. (2025), freeze the weights of an LLM, feed it politically charged inputs, and extract hidden state activations. Then, they train a small linear probe classifier on those internal activations to confirm whether specific hidden layers implicitly track an "ideological direction."

- **Survey-Based Behavioral Benchmarking**. Papers such as Motoki et al. (2023); Hartmann et al. (2023); Feng et al. (2023) treat LLMs like human survey respondents. They repeatedly prompt models with established psychometric questionnaires (e.g., The Political Compass, Pew Research political typology tests) under zero-shot settings to plot the model's default biases.

- **Synthetic Demographic Prompting**. Works like PERSONAHUB (Bernardelle et al., 2024) structure specialized system prompts to assign distinct personas (e.g., "You are a working-class conservative from Ohio..."). They use

Taxonomy Branch	Core Engineering Methods	Inputs
§3.1 Detection	Feature-based machine learning, BERT fine-tuning, GNNs, GCNs, LLM prompting.	Twitter/X retweet networks, congressional speeches, political news articles.
§3.2 Measurement	Hidden-layer activation probing, zero-shot psychometric testing, synthetic persona prompting.	Political Compass quizzes, Pew Research surveys, synthetic profiles.
§3.3 Intervention	Activation vector injection, Direct Preference Optimization (DPO), latent alignment steering.	Paired multi-perspective training text, controversial question sets.

Table 4: Overview of the representative engineering methods and the corresponding inputs in political ideology analysis for language models.

these synthetic human proxies to study how robustly an LLM can simulate real-world human public opinion and demographic polarization.

B.3 Ideology Intervention

This branch actively manipulates the model's behavior, moving beyond passive observation to forcing a model to change its political output or eliminate internal biases.

- **Activation Vector Injections (Representation Engineering)**. To address the steering problem, papers like Kim et al. (2025) and standard representation frameworks aim at computing an "ideology vector" via probing, and then directly add or subtract that mathematical vector from the model's hidden states during forward-pass inference. This forcibly changes the model's political tone without rewriting weights.

- **Cross-Lingual & Direct Latent Space Alignment**. Advanced steering methods, such as CLAS (Nadeem et al., 2026), apply adaptive scaling to specific layers to push the model's representations toward neutrality or target viewpoints across multiple languages simultaneously.

- **In-Context Fine-Tuning & Alignment** Training (RLHF / DPO): Debiasing papers, including Guo et al. (2024) and Li et al. (2021), use preference optimization frameworks (such as Direct Preference Optimization) or targeted instruction-tuning. They train the model on curated pairs of biased vs. neutral viewpoints to penalize biased text generation and enforce viewpoint neutrality.