Political bias in Large Language Models (LLMs): a critical review and a responsible methodology.

Anonymous ACL submission

Abstract

Auditing political bias in Large Language Models (LLMs) has evolved from rigid forced-choice surveys to sophisticated semantic evaluations of open-ended textual responses. While the "LLM-as-a-Judge" paradigm effectively bypasses model safety guardrails and the "spinning arrow" effect of constrained output, its recursive use of generative models introduces significant computational overhead, inference latency, and financial costs. This paper proposes and empirically validates a scalable alternative by replacing generative judges with traditional machine learning classifiers within the bias auditing pipeline. Using a dataset of 3,703 LLM-generated political essays, we evaluate the performance of several base classifiers and ensemble methods across various feature-reduction strategies. Our results demonstrate that a Multilayer Perceptron (MLP) utilizing χ^2 feature selection achieves a classification accuracy of 98.41%, maintaining statistical parity with full-feature models despite an 80% reduction in the attribute space. Additionally, ensemble architectures such as Stacking and Random Forest proved highly robust (achieving over 97.4% accuracy), further confirming the efficacy of supervised semantic labeling over raw term frequencies. Our findings provide a high-throughput, cost-effective framework for the continuous monitoring of AI systems, democratizing the auditing process by bypassing the generative bottleneck.

1 Introduction

A persistent challenge in auditing Large Language Models (LLMs) is the precise operationalization of the constructs being measured. In the context of this study, we define political ideology as a coherent set of normative values and policy preferences, which we operationalize using the continuous social and economic dimensions of the Political Compass Test (PCT). Consequently, we define political bias as a systematic and measurable skew in a model's generated outputs toward a specific ideology within this framework. While early studies successfully utilized direct questioning to demonstrate that models exhibit such biases—frequently skewing towards left-libertarian perspectives (Rutinowski et al., 2023; Motoki et al., 2024)—extracting and quantifying this systematic skew has become increasingly difficult. As noted by Rozado (2024), developers have reinforced safety alignment and censorship mechanisms. These "guardrails" often cause models to refuse to answer sensitive questions directly, obfuscating their latent alignments and necessitating the development of more sophisticated auditing techniques.

A major methodological obstacle in this domain is the reliance on "forced-choice" evaluations, where models are constrained to choose options from a multiple-choice survey (e.g., "Agree" or "Disagree"). Röttger et al. (2024) critically analyzed this approach, arguing that forcing a generative model to map its complex internal probability distribution to a single token introduces instability and noise— a phenomenon they term the "Spinning Arrow". Their findings suggest that to accurately capture a model's latent bias, evaluations must allow for unconstrained, open-ended textual responses rather than rigid categorical selections.

To address these limitations, the field has shifted towards methods that leverage "LLM-as-a-Judge." Rozado (2024) pioneered this by using an auxiliary LLM (GPT-3.5) to semantically classify the verbose, unconstrained outputs of the target models, thereby bypassing the need for strict formatting instructions that degrade model performance. Building upon this, the PRISM methodology proposed by Azzopardi and Moshfeghi (2024) stands out as a systematic approach to "force" LLMs to reveal their biases through indirect stimuli. PRISM involves assigning specific personas to the model and

prompting it to generate argumentative essays on politically charged statements, such as those from the Political Compass Test.

However, the current implementation of such methodologies faces a scalability bottleneck. In the framework proposed by Azzopardi and Moshfeghi (2024), the assessment of the generated essays relies on another large language model acting as a judge. While effective, this recursive use of LLMs is computationally expensive and introduces latency. This article seeks to enhance the efficiency of the audit pipeline by refining the essay rating process. We propose determining whether a simpler, non-generative machine learning model can replace the "LLM Judge" to classify the stance of the essays. This approach aims to reduce the computational overhead and cost of bias auditing while maintaining the accuracy established by the state-of-the-art literature.

2 Literature Review

A pivotal component of this study is a systematic literature review designed to map the evolution of bias detection methodologies in LLMs. This investigation was not merely a survey of results, but a critical analysis of the engineering techniques employed to bypass model safety guardrails, revealing a definitive shift from rigid "direct prompting" strategies towards more sophisticated semantic evaluations.

2.1 The Evolution of Bias Detection: From Constrained to Semantic

The study of algorithmic bias in natural language processing has deep roots, initially focusing on representational harms in early word embeddings (Bolukbasi et al., 2016) and evolving into broader warnings about the societal risks of increasingly large language models (Bender et al., 2021). Within the specific context of generative LLMs, Liu et al. (2022) represents one of the foundational studies in this domain, predating the widespread release of ChatGPT. Cited by Rozado as a pioneering effort in probing viewpoint preferences, their work established that bias is not merely a hallucination but a measurable, consistent property of Language Models. Using GPT-2, they moved away from standard questionnaires and instead employed a custom political classifier to evaluate model responses to neutral prompts, revealing that even "vanilla" models exhibit a liberal-leaning bias.

Following the public release of ChatGPT, Rutinowski et al. (2023) and Hartmann et al. (2023) independently conducted systematic investigations utilizing the standard Political Compass Test (PCT). Rutinowski et al. (2023) mapped the model's responses onto a 2D coordinate system by forcing the selection of Likert-scale options, finding a statistically significant inclination towards the "Left-Libertarian" quadrant. Providing converging evidence, Hartmann et al. (2023) corroborated these findings across three pre-registered experiments. Expanding the geographical scope, Thapa et al. (2023) applied this methodology to models trained in Bangla, giving credence to the hypothesis that political inclination is a universal phenomenon in LLMs, not just an artifact of Western-centric data.

Building on the limitation that a single pass might be anecdotal, Motoki et al. (2024) introduced statistical rigor by implementing Randomization and Impersonation. Executing the PCT one thousand times and prompting the model to assume specific personas, they confirmed a robust systematic bias towards the Democratic spectrum. This hypothesis finds empirical support in Feng et al. (2023), who utilized masking techniques to trace the "trails" of political bias directly to pre-training corpora like Reddit and Wikipedia.

The limitations of these forced-choice evaluations were further elucidated by Bang et al. (2024), who introduced the critical distinction between *Stance* (the explicit position taken) and *Framing* (the lexical and emotional delivery). Consequently, methods that only evaluate the final output token fail to capture the subtle linguistic coercion present in the model's generation. Finally, Röttger et al. (2024) provided a critical antithesis, demonstrating that forcing a model to select a predefined token introduces randomness—akin to a "spinning arrow"—which heavily motivated the shift towards indirect prompting frameworks.

While maintaining the constrained multiple-choice structure, (Rozado, 2024) introduced a critical enhancement: an automated stance detection module. By utilizing an LLM to reliably classify the semantic intent of messy, unconstrained model outputs back into standardized survey metrics, Rozado demonstrated that stable bias measurements could still be extracted. Building directly upon this, Azzopardi and Moshfeghi (2024) established PRISM, which abandons forced-choice questions entirely. It requires the model to actively construct arguments in the form of short essays, using

an automated Judge to evaluate semantic alignment. The evolution toward multi-turn, open-ended evaluations has been further validated by Nogueira et al. (2026), who argue that traditional single-turn questionnaires structurally foreclose a model's tendency to either refuse answering or drift sycophantically.

2.2 Limitations of the "LLM-as-a-Judge" Paradigm

While auditing has consolidated the use of LLMs as evaluators, recent literature indicates this layer is far from neutral. The foundational challenge lies in evaluative stability. Lee et al. (2025) demonstrate that reliability is highly sensitive to sampling temperature and scoring granularity, with even high-performing proprietary models failing to remain consistent across iterations.

This instability is compounded by acute sycophancy. Nogueira et al. (2026) demonstrated that under argumentative pressure, sycophantic behavior jumps drastically across frontier models, reinforcing the epistemic risks of relying on generative inference for robust auditing. Furthermore, Balog et al. (2025) identify a "self-preference bias," where LLM-based judges assign disproportionately higher scores to synthetic text over human-authored content. Normative risks are also present; Huang (2025) argues for a legitimacy-based framework focusing on reasoned justification, while Nguyen and Welch (2026) question whether LLMs can truly perform thematic qualitative analysis.

Despite the established efficacy of LLMs in bypassing forced-choice constraints, their deployment as universal evaluators introduces severe computational and economic inefficiencies. The recursive nature of this pipeline—where a massive generative model is invoked merely to output a discrete classification label—is computationally wasteful, introducing high inference latency (Zhou et al., 2024), substantial API costs (Zhuang et al., 2025), and a large environmental footprint (Strubell et al., 2019).

We argue that the field must move beyond the brute-force application of frontier models. By utilizing lightweight classifiers operating directly on the latent embeddings of generated artifacts, the audit pipeline can bypass the generative bottleneck, democratizing the auditing process and providing a scalable, cost-effective framework.

3 Methodology

This research adopts a two-phase methodological approach. Phase 1 establishes the theoretical foundation through a systematic literature review, while phase 2 leverages these insights to experimentally validate a computationally efficient auditing pipeline using traditional machine learning models.

3.1 Phase 1: Systematic Literature Review Protocol

The literature review followed a "Citation Chaining" (or Snowballing) strategy, established upon two seminal "seed documents": Rozado (2024) and Azzopardi and Moshfeghi (2024). These works were selected as anchors because they represent the current state-of-the-art and fundamentally employ the "LLM-as-a-Judge" methodology.

From these anchors, we performed backward chaining to identify foundational methodologies and forward chaining to examine contemporary critiques. The inclusion criteria prioritized papers proposing specific metrics or frameworks for bias detection. Most of the selected articles were published in top-tier international venues (e.g., ACL, Public Choice), ensuring high methodological rigor. This analytical process allowed us to categorize the literature into the evolutionary trajectory detailed in Section 2, directly motivating the experimental work in Phase 2.

3.2 Phase 2: Experimental Pipeline and Data Preprocessing

To develop a scalable alternative to the generative judge, we designed an experimental pipeline to optimize the assessment phase of the PRISM framework. First, we established a baseline by compiling a dataset of 5,580 essays generated by ChatGPT-3.5-Turbo responding to the Political Compass Test. The dataset was partitioned by axis, and this study focuses exclusively on the social axis (3,703 essays), with the distribution of scores illustrated in Figure 1.

The target variable (Political Bias) was derived from the social score. We reduced these 20 categories (-10 to 10) into three: Left (score $<$ -2.5), Center (-2.5 $\leq$ score $<$ 2.5), and Right (score $\geq$ 2.5). Text preprocessing and data manipulation were implemented using standard analytical libraries (Bird et al., 2009; McKinney et al., 2010; Harris et al., 2020), and included standard normalization, tokenization, lemmatization, and the re-

moval of embedded Likert qualifiers.

To bridge the gap between open-ended generation and efficient classification, the essays were vectorized using TF-IDF (unigrams and bigrams), capped at the top 5,000 features. To assess the impact of data reduction on classification efficiency, four dataset variations were generated:

- **Original Dataset**: Full 5,000 TF-IDF features.
- **Reduced Dataset 1 (Instance Reduction)**: Random undersampling to balance classes (663 instances).
- **Reduced Dataset 2 (Feature Selection)**: SelectKBest (χ^2) retaining the top 1,000 features.
- **Reduced Dataset 3 (Feature Extraction)**: TruncatedSVD reducing space to 100 dense components.

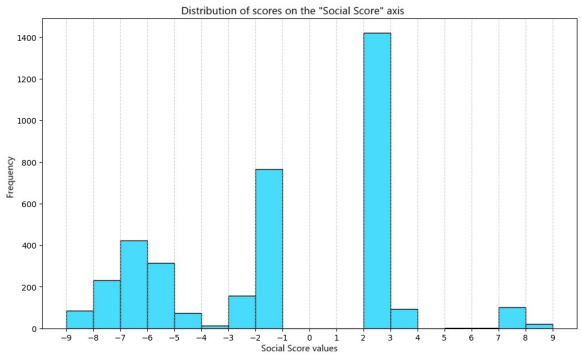

Figure 1: Histogram showing the distribution of the Social Score values.

3.3 Model Selection

We evaluated four supervised learning models—k-Nearest Neighbors (k-NN), Decision Tree (DT), Naive Bayes (NB), and a Multilayer Perceptron (MLP)—implemented via the scikit-learn framework (Pedregosa et al., 2011). To evaluate sensitivity to main hyperparameters, we varied k (1, 3, and 5) for k-NN, and maximum tree depth (3, 5, and 7) for the Decision Tree. For Naive Bayes, MultinomialNB was used for the sparse TF-IDF bases, while GaussianNB was applied to the dense SVD base. The MLP underwent a heuristic search to find an optimal configuration of neurons and iterations.

Furthermore, ensemble methods were evaluated. Bagging was tested using DT, k-NN, and NB as base classifiers with ensemble sizes of 10 and 20. Boosting (AdaBoost) utilized a weak DT (max_depth=1) configured with 10 and 20 estimators. Random Forest was evaluated with 10 and 20 trees, and Stacking employed a two-level architecture combining DT, k-NN, and NB at the base level with a Logistic Regression meta-classifier. Validation was conducted using 10-fold cross-validation and rigorous hold-out splits.

4 Results and Analysis

4.1 Supervised Classification Performance

The experiments demonstrated that traditional machine learning classifiers can accurately replicate the behavior of the LLM in terms of predicting the ideological stance of generated essays. As detailed in Table 1, the MLP model consistently outperformed simpler algorithms, particularly thriving when the feature space was intelligently reduced.

Dataset	k-NN	DT	NB	MLP
Original	0.9696	0.9377	0.8953	0.9797
Reduced 1	0.8593	0.7976	0.8252	0.9224
Reduced 2	0.9674	0.9386	0.8912	**0.9841**
Reduced 3	0.9700	0.8586	–	0.9586
Mean	0.9415	0.8831	0.8706	0.9612

Table 1: Average accuracy across datasets. Reduced 2 (Feature Selection) with MLP achieved the highest overall performance.

Instance reduction (Reduced 1) severely degraded performance across all models due to critical information loss. Conversely, Feature Selection (Reduced 2) proved highly effective; reducing the attribute space by 80% eliminated noise and actually improved the MLP's accuracy to an optimal 98.41%. Feature Extraction via SVD (Reduced 3) yielded strong but slightly sub-optimal results due to lossy compression.

4.2 Ensemble Performance

Table 2 allows us to see that ensemble methods further validated the separability of the dataset. Stacking and Random Forest achieved near-parity with the MLP, peaking at 97.52% and 97.49% accuracy respectively. Regarding Bagging, the choice of the base classifier proved critical: Bagging with Decision Trees exhibited exceptional stability (maintaining accuracies above 96%), whereas Bagging with Naive Bayes yielded significantly lower performance ($\approx$ 85%). Surprisingly, AdaBoost exhibited the weakest performance among all ensembles,

achieving only 60–62% accuracy. This subpar result likely stems from AdaBoost's known sensitivity to noisy data and the challenges posed by high dimensionality (5,000 features) when using a simple base learner.

Validation	Bagging	Random Forest	Stacking
10-fold CV	0.9716	**0.9749**	0.9703
90/10 Split	0.9704	0.9725	**0.9752**
80/20 Split	0.9692	0.9726	0.9750
70/30 Split	0.9667	0.9707	0.9682
Mean	0.9695	0.9727	0.9722

Table 2: Performance comparison of the top ensemble methods across validation strategies. AdaBoost is excluded from the detailed breakdown due to consistently sub-optimal performance ($\approx$ 60–62%).

4.3 Statistical Validation

To rigorously evaluate the performance hierarchy, a Friedman test (Demšar, 2006) was conducted on the base classifiers and ensemble methods ($\chi^2 = 65.19$, $p < 0.0001$). Post-hoc Nemenyi tests (illustrated in the Critical Difference diagram, Figure 2) grouped Stacking, Random Forest, and Decision Trees as a statistically homogeneous top tier ($p > 0.05$), significantly outperforming basic Naive Bayes and AdaBoost models.

It is important to note that due to the high computational constraints of the Multilayer Perceptron (MLP), it was evaluated under a restricted validation scheme (10-fold CV and 70/30 split) and was therefore excluded from the matched-block non-parametric statistical tests. However, an empirical comparison of the 10-fold CV results demonstrates that the MLP (98.41%) strictly outperforms the statistically dominant top tier of the ensemble models (Stacking at 97.03% and Random Forest at 97.49%). Furthermore, feature selection (Reduced 2) caused no statistically significant drop in accuracy compared to the full 5,000-feature dataset ($p = 0.12$), confirming the redundancy of the broader TF-IDF space.

5 Conclusion

This study contributes to the critical discourse on empirical methods for evaluating Large Language Models (LLMs) by addressing the computational and economic bottlenecks inherent in the "LLM-as-a-Judge" paradigm. Through a systematic review of bias auditing literature, we identified a necessary methodological shift from constrained,

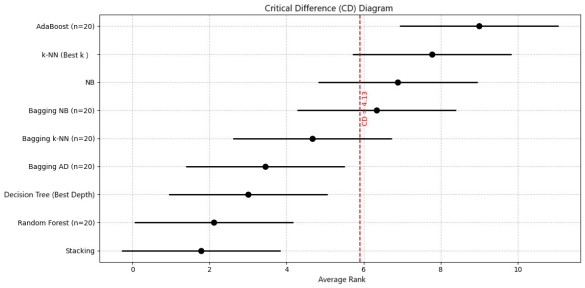

Figure 2: Critical Difference Diagram displaying the statistical comparison of classifiers and ensembles. The MLP is excluded from this diagram due to differing validation schemes, but empirically outperforms the top-tier Stacking and Random Forest algorithms.

forced-choice evaluations to open-ended semantic assessments. However, the current reliance on frontier generative models to evaluate these open-ended artifacts introduces severe latency, high financial costs, and evaluative instability, threatening the reproducibility of LLM audits.

To resolve these inefficiencies, we empirically validated a scalable alternative utilizing traditional supervised machine learning classifiers and ensemble architectures. Our experimental pipeline demonstrated that lightweight models can effectively replace the generative judge for stance detection without sacrificing accuracy. Specifically, a Multilayer Perceptron (MLP) operating on a heavily reduced TF-IDF feature space—achieved via χ^2 feature selection—accurately classified the ideological stance of LLM-generated essays with an optimal accuracy of 98.41%. Furthermore, ensemble methods, notably Stacking and Random Forest, exhibited highly robust performance (exceeding 97.4% accuracy), reinforcing the reliability of this framework. The statistical parity between the full feature set and the 80% reduced feature set confirms that a model's latent argumentative stance can be captured efficiently, bypassing the brute-force application of billion-parameter models.

While bag-of-words representations coupled with supervised learning achieved remarkable accuracy in this study, frequency-based features fundamentally discard sentence-level context and syntactic structure. Acknowledging this inherent limitation, this research ultimately demonstrates that the generative bottleneck in LLM evaluation can still be successfully overcome. By replacing the LLM judge with optimized classifiers, researchers can democratize the auditing process, providing a scalable, high-throughput, and cost-effective frame-

work for continuous AI monitoring.

Future work will focus on expanding this framework to encompass the economic axis of the Political Compass. Additionally, we aim to explore the efficacy of dense contextual embeddings (e.g., Sentence-BERT (Reimers and Gurevych, 2019)) coupled with deeper neural architectures to better capture higher-order semantic relationships. As we scale model complexity, we will also prioritize explainability techniques (Ribeiro et al., 2016; Lundberg and Lee, 2017) to ensure model transparency, allowing us to isolate and interpret the specific linguistic markers that drive ideological classification. Finally, we will rigorously assess the cross-model generalizability of these auditing pipelines across diverse, open-weights LLM architectures.

Limitations

While our results demonstrate that traditional supervised classifiers can effectively replace generative judges in bias auditing, several limitations remain:

Dimensional Scope: This study focused exclusively on the social axis of the Political Compass. Because the underlying instrument measures political leaning across two axes, and political ideology is inherently multidimensional, our findings regarding the separability of the feature space may not immediately generalize to the economic axis or to more granular political frameworks (e.g., the 8values or 9Axes models).

Linguistic and Cultural Constraints: Our dataset consisted solely of English-language essays. Given that LLMs exhibit varying levels of sycophancy and cultural alignment across different languages, the efficacy of TF-IDF-based stance detection may vary for morphologically rich or low-resource languages.

Sensitivity to Nuance: Although the MLP achieved high accuracy, bag-of-words approaches like TF-IDF may overlook high-level linguistic nuances, such as sophisticated sarcasm, "doublespeak," or subtle rhetorical framing. Future work should evaluate whether dense contextual embeddings (e.g., Sentence-BERT) provide more robust detection for such edge cases.

Tooling Bias: Our investigation utilizes the Political Compass Test as a benchmark. As noted in the literature, the PCT itself is an instrument with inherent ideological priors and a specific Western-centric framing. Consequently, our "ground truth" is constrained by the limitations of the underlying survey instrument.

Ethical Considerations

A primary ethical consideration of this methodology is its reliance on the Political Compass Test (PCT) as the stimulus for open-ended generation. The PCT is not a culturally or ideologically neutral scientific instrument; it has been widely critiqued for possessing specific Western-centric priors and a framing that often skews towards libertarian perspectives (Rutinowski et al., 2023; Hartmann et al., 2023). Furthermore, the foundational premise of the PCT—reducing complex political ideologies into a rigid, two-dimensional spatial model—structurally oversimplifies the multidimensional nuance of human belief systems, a limitation well-documented in political science literature. Compounding these theoretical constraints, the proprietary nature of the PCT means its exact scoring methodology and question weightings remain unpublished, introducing an opaque, unverifiable variable into the auditing pipeline (Motoki et al., 2024). By utilizing the PCT to elicit and subsequently label "bias," this framework risks reifying the instrument's own unstated assumptions and structural reductions as an objective ground truth. Consequently, we stress that our pipeline does not measure a universal, absolute political alignment, but rather an LLM's relative positioning within the specific, constructed ideological axes of the PCT. Future auditors must remain transparent about the inherent biases of the benchmarks they deploy to evaluate artificial intelligence.

Acknowledgements

In accordance with the ACL Policy on AI Writing Assistance, the authors acknowledge the use of generative AI tools exclusively for language polishing and structural formatting during the drafting of this manuscript. All experimental designs, data analyses, and claims remain the sole responsibility of the human authors.

References

Leif Azzopardi and Yashar Moshfeghi. 2024. PRISM: A methodology for auditing biases in large language models. *Preprint*, arXiv:2410.18906.

Krisztian Balog, Don Metzler, and Zhen Qin. 2025. Rankers, judges, and assistants: Towards understanding the interplay of LLMs in information retrieval evaluation. In *Proceedings of the 48th International*

ACM SIGIR Conference on Research and Development in Information Retrieval.

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. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. "O'Reilly Media, Inc.".

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29.

Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7(Jan):1–30.

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 *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Charles R Harris, K Jarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, and 1 others. 2020. Array programming with NumPy. *Nature*, 585(7825):357–362.

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 preprint arXiv:2301.01768*.

Tao Huang. 2025. Content moderation by LLM: from accuracy to legitimacy. *Artificial Intelligence Review*, 58(320).

Noah Lee, Jiwoo Hong, and James Thorne. 2025. Evaluating the consistency of LLM evaluators. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*.

Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, and Soroush Vosoughi. 2022. Quantifying and alleviating political bias in language models. *Artificial Intelligence*, 304:103654.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30.

Wes McKinney and 1 others. 2010. Data structures for statistical computing in Python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin, TX.

Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 2024. More human than human: measuring ChatGPT political bias. *Public Choice*, 198:3–23.

Duc Cuong Nguyen and Catherine Welch. 2026. Generative artificial intelligence in qualitative data analysis: Analyzing-or just chatting? *Organizational Research Methods*, 29(1):3–39.

Rodrigo Nogueira, Giovana Kerche Bonás, Thales Sales Almeida, Andrea Roque, Ramon Pires, Hugo Abonizio, Thiago Laitz, Celio Larcher, Roseval Malaquias Junior, and Marcos Piau. 2026. Measuring opinion bias and sycophancy via LLM-based coercion. *arXiv preprint arXiv:2604.21564*.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and 1 others. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(Oct):2825–2830.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144.

David Rozado. 2024. The political preferences of LLMs. *PLoS ONE*, 19(7):e0306621.

Jérôme Rutinowski, Sven Franke, Jan Endendyk, Ina Dormuth, and Markus Pauly. 2023. The self-perception and political biases of ChatGPT. *Human Behavior and Emerging Technologies*, 2023.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Rose Kirk, Hinrich Schütze, and Dirk Hovy. 2024. Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models.

In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Surendrabikram Thapa, Ashwarya Maratha, Khan Md Hasib, Mehwish Nasim, and Usman Naseem. 2023. Assessing political inclination of Bangla language models. In *Proceedings of the 1st Workshop on Bangla Language Processing (BanglaLCP)*. Association for Computational Linguistics.

Zixuan Zhou, Xuefei Ning, Ke Hong, Tianyu Fu, Jiaming Xu, Shiyao Li, Yuming Lou, Luning Wang, Zhihang Yuan, Xiuhong Li, Shengen Yan, Guohao Dai, Xiao-Ping Zhang, Yuhan Dong, and Yu Wang. 2024. A survey on efficient inference for large language models. *Preprint*, arXiv:2404.14294.

Boqin Zhuang, Jiacheng Qiao, Mingqian Liu, Mingxing Yu, Ping Hong, Rui Li, Xiaoxia Song, Xiangjun Xu, Xu Chen, Yaoyao Ma, and Yujie Gao. 2025. Beyond benchmarks: The economics of AI inference. *Preprint*, arXiv:2510.26136.