Political Bias in Large Language Models: A Case Study on the 2025 German Federal Election

Buket Kurtulus

***Abstract*—With the increased use of Large Language Models (LLMs) to generate responses to social and political topics, concerns about potential bias have grown. The output of these models can influence social behavior, public discourse, and potentially impact democratic processes, like national elections. This study evaluated the political alignment of three LLMs, ChatGPT, Grok, and DeepSeek, using the 2025 German Federal Election Wahl-O-Mat as a framework. Each model was presented with 38 political statements in both German and English, and their responses were compared to the official positions of German political parties. To support this, we will also explore the theoretical foundations of political bias in LLMs, focusing on how prompt language and model characteristics, such as scale and regional origin, may influence the ideological alignment, as well as examining relevant ethical considerations. The results revealed a consistent left-leaning tendency across all models, with minimal alignment with far-right positions. By combining the empirical findings with the theoretical backgrounds, this contributes to a deeper understanding of political bias in LLMs and highlights the importance of transparency in their public use.**

I. Introduction

Over the past decade, artificial intelligence (AI) technologies have reshaped key areas of modern life, including healthcare, finance, governance, and politics [2], [4], [5]. Furthermore, AI models have become more used in business operations, supporting tasks like predictive analysis, automation, and natural language processing. The U.S. generative AI market alone is projected to grow rapidly with an annual growth rate of 36.3% through 2030 [2].

One of the most significant developments in this field is the rise of Large Language Models [1]. These models generate human-like responses to complex prompts. Also they are widely used across various domains, including customer service, marketing, and public communication. In addition LLMs are increasingly integrated into everyday life, assisting with tasks like summarizing documents, drafting emails, or providing live advice [1], [2].

These developments have raised doubts among public policymakers and political analysts about the appropriateness of using AI tools in political contexts. The concerns center on whether LLMs can support critical thinking, accurate political evaluation, and responsible decision-making addressing politically sensitive topics [1].

Recent studies have shown that some LLMs may exhibit political leanings, often aligning with liberal or left-libertarian ideologies [2]. This bias is rooted in the way these models are built. One of the reasons is that the machine learning algorithms are built by humans and trained on vast datasets that reflect historical and societal patterns, which contain embedded bias [1]. Although there is growing awareness of these tendencies, structured methods for evaluating political bias in LLMs, particularly in the context of end-users in a national setting like Germany, are still lacking.

To address this research gap, this paper employs a dual-perspective approach. First, a theoretical overview of LLMs and bias with a focus on political bias is provided. In the second part, three case studies that investigate political bias from a both European and U.S. perspectives are highlighted. Afterwards, we conduct a case study using the 2025 Wahl-O-Mat questionnaire to evaluate the political bias of three LLMs. By comparing the responses with German political parties, we aim to assess the potential bias. The findings are discussed from an ethical perspective, followed by a conclusion.

II. Technical Background

The following section outlines the theoretical background necessary to understand how bias can exist in large language models (LLMs). It begins with a short overview of LLM architecture and training, then discusses how bias can emerge, with an emphasis on political bias.

A. Large Language Models

Artificial Intelligence (AI) refers to the field of computer science that focuses on systems capable of performing tasks that require human intelligence, such as reasoning, learning, or language understanding. While the term AI was officially introduced during the Dartmouth Conference in 1956, it was not until recent developments that it gained broader attention [2], [22], [23].

One of the recent developments in AI is Generative AI (GenAI), systems that can create new content, such as images or text. A subset of GenAI systems are large language models that are trained to understand and produce human-like text. Prominent examples are Copilot and GPT-4. LLMs are capable of tasks such as answering questions, translating languages, and generating code with fluency [2], [22], [23].

LLMs are built upon a transformer architecture, which enables them to process and generate text. Transformers are a deep learning architecture, which weights the importance of each part of the input data. Functionally, LLMs operate by first encoding the input text into a high-dimensional vector representation, capturing the semantic and contextual relationships between words and phrases. Then generating the output one token at a time. The quality of the output depends on multiple factors, like the structure of the input, the model's hyperparameters, and the scope of the training data [2], [3], [23].

B. Bias

As LLMs become increasingly integrated into the public, the biases embedded within them raise concerns about their societal and political impact [2]. These biases arise from several sources, many of which are inherent to machine learning systems. Since algorithms are developed by humans and trained on historical data, they often reflect and amplify existing biases [1]. The primary source of bias lies in the training datasets themselves. Search engines, social media content, and digitized texts all carry forward prejudices and imbalances. Consequently, LLMs trained on these datasets tend to reproduce these patterns [2], [6]. A bias analysis of 2023 models, including Chat-GPT4, showed that LLMs often replicate the same biases found in their training data (11). Recent studies further emphasize that LLMs exhibit both political and social biases, shaped not only by the data but also by the underlying algorithms that generate their outputs [2].

In addition to data and algorithms, the deployment context and the level of ethical Oversight also play a crucial role. Algorithmic design decisions, especially during fine-tuning, can unintentionally reinforce biases when ideological viewpoints are disproportionately represented [2]. The deployment context determines how and where an LLM is used, which can influence the manifestation of bias. For instance, depending on the environment or public service setting, a model's output may align differently with prevailing social narratives or amplify specific viewpoints [17], [18]. Ethical Oversight is equally essential. Without a robust governance framework, transparency in development, and ongoing bias audits, LLM systems risk deepening existing inequalities and undermining public trust in AI technologies.

These factors showcase that bias in LLMs is not simply a training issue, but the result of a complex intersection of data, design, context, and governance [1], [2].

Bias in LLMs can be addressed through a combination of technical solutions and ethical frameworks. On the technical side, strategies include regular bias audits and updates to the LLMs, as well as advanced methods such as adversarial training. Here, the LLM is purposefully introduced to challenges which help to reveal biases during the training process and correct them [2]. For example, in 2023 [10] proposed a multilayered mitigation approach that combines audits, transparency reports, and the use of "debiasing algorithms" to minimize bias in models. It is equally important to consider the ethical side. Transparency reports help clarify how models are trained, specifically which data is used, and how the model's behavior is implemented. Governance frameworks are also essential to ensure a responsible development of LLMs [2], [10].

C. Political Bias

Political bias is one of the types of bias that can be observed in large language models.

Political bias refers to the systematic tendency of a language model to display favoritism toward a specific political ideology or perspective in its output. This type of bias has been observed in several LLMs. For Instance, GPT-3 showed a bias against some religious and ethnic groups, while Google Gemini has demonstrated a tendency to favor centrist positions [2]. A 2023 case study also revealed that user interaction can reinforce and amplify political bias. This also occurs when the user unconsciously introduces ideologically charged language; the model may adapt and mirror this perspective [2], [7] .

Political bias in AI systems, especially in LLMs, can have significant implications for democratic societies. Algorithmic political predictions often reflect the interests or perspectives of those who design or deploy the systems [1]. As LLMs are increasingly used in public-facing applications, their potential to shape public opinion, influence political behavior, and even affect electoral outcomes raises serious ethical concerns. This is particularly problematic in politically sensitive contexts, where biased content and content generation can be leveraged for the dissemination of misinformation or propaganda [2], [8].

Research has shown that biased algorithms can intensify existing social injustices and undermine democratic processes [9]. Without sufficient supervision, these biases may cause long-term harm to the public sphere. As [17]. emphasizes that political bias in AI systems often reflects deeper societal injustices and thus calls for deliberate and ethically grounded responses.

III. Related Work

With the increasing influence of LLMs, understanding political bias has become a critical research area. In the following section, we present three case studies that explored the political bias in LLMs from different methodological perspectives.

[11] investigated political bias in open-source LLMs by examining their alignment with political positions in the context of the 2024 European Parliament elections from a German voter's perspective. The authors aimed to assess whether, and to what extent, LLMs show political bias. To do so, they used the Wahl-O-Mat, a voting application developed by the German Federal Agency for Civic Education, which will be described in more detail in Section 4A. The authors prompted the Wahl-O-Mat statements in both German and English for several open-source models, including different versions of LLaMA or Mistral-7B. The models were controlled to respond with "Yes", "Neutral", or "No" to ensure comparability across models. The results revealed that larger models, e.g., LLaMA3-70B, showed a stronger alignment with left-leaning parties, such as Bündnis 90/Die Grünen, Die Linke, and Volt, while correspondingly showing minimal alignment with right-wing parties, like the AfD. In contrast, smaller models displayed a more neutral position. Additionally, the study found out that models which are prompted in German, are more likely to take a clear political stance compared to the English-language outputs. [11] highlights the risks associated with political bias and emphasizes the importance of responsible

LLM development.
[2] Presented a comprehensive analysis of political bias in four popular LLMs: ChatGPT-4, Perplexity, Google Gemini, and Claude. The study assessed the ideological leanings of these models using a combination of quantitative and qualitative methodologies to identify both explicit and subtle political biases. For this, the models were tested using three political typology tools: the Pew Research Center's Political Typology Quiz, the Political Compass Assessment, and the ISideWith Political Party Quiz. Each model was prompted with the same set of questions, and the responses were standardized and categorized alongside a predefined ideological scale ranging from "strongly conservative" (referred to as Faith and Flag Conservatives) to "strongly liberal" (referred to as Progressive Left). The results showed that ChatGPT-4 consistently displayed liberal tendencies, particularly in social and economic domains. On the Pew scale, it was classified as an "Establishment Liberal", a group representing 13% of the general public. Perplexity also leaned left overall but displayed more conservative tendencies on selected issues, which led the author to categorize it as "Outsider Left", which corresponds to 10% of the public. In contrast, both Claude and Google Gemini were considered more centrist, adopting more neutral and moderate stances.
[12] conducted a comprehensive study, evaluating 43 LLMs from 19 model families across four regions: the US, Europe, Asia, and the Middle East. The goal was to assess political bias across a US-centric testing ground while also accounting for model characteristics, such as scale, release date, and geographic origin. The study covered both open-source and closed-source models of various sizes and configurations. To ensure comparability, the authors selected 32 politically themed questions from the American National Election Studies (ANES) and the 2024 Pew Research Center survey. The questions were grouped into eight topics, equally split between highly polarized topics, such as presidential elections, abortion, immigration, and Issue Ownership, and less polarized topics, including climate change, misinformation, discrimination, and foreign policy. The models were prompted using a two-step framework, which included techniques to elicit answers to sensitive questions by evading model safety filters. The responses were subsequently analyzed using a preference scoring system to assess political leanings. If the score was positive, it indicated a Democratic-leaning bias; negative scores indicated a Republican-leaning bias. The findings revealed that most tested LLMs showed a left-leaning bias, especially on the highly polarized topics. [12] used the 2024 US presidential election as a benchmark. 76% of the models expressed a stronger preference for the Democratic candidates (Joe Biden or Kamala Harris), with 35% consistently favoring them. The study also showed that less polarized topics had less distinct bias.

IV. Method

A. *Wahl-O-Mat*

The Wahl-O-Mat is a digital tool designed to help voters in Germany assess how political parties align with their views. It is typically released before a specific election, such as the Federal Election in 2025. The user is presented with a series of political statements, from a wide range of topics relevant to the current political landscape [13]. The user can decide

Fig. 1. One political statement as shown in the web interface of the Wahl-O-Mat 2025 Federal Elections. It translates to: Germany should continue to support Ukraine with military

for each statement whether they agree, are neutral, or disagree (Figure 1). Ultimately, the Wahl-O-Mat compares the user's responses with the official positions of the political parties, indicating which party's views align most closely with the user's stance. The Wahl-O-Mat was developed by the Federal Agency for Civic Education (Bundeszentrale für politische Bildung, Bpb) and is considered the most important tool for electoral decision-making in Germany. For the 2025 federal election, the tool was accessed over 26 million times [13].

B. *Large Language Models*

For this case study, we selected three LLMs to enable a geographically and structurally diverse comparison: OpenAI's ChatGPT, xAI's Grok, and DeepSeek. These models differ in terms of training origin, scale, and integration into public platforms.
ChatGPT (gpt-3.5-turbo) was chosen due to its widespread use, global influence, and established reputation as one of the most prominent commercially available LLMs in Western contexts. Developed by OpenAI, this model is primarily trained on English-language data and is widely deployed in both consumer and enterprise settings. The model was accessed via the official OpenAI API [19].
DeepSeek (deepseek-chat) is a model developed in China and represents a new presence in the global LLM landscape. It is of interest for its development within a different linguistic and sociopolitical environment. DeepSeek was accessed through its official API [21].
Lastly, Grok (grok-3-mini) is developed by xAI, a company

founded by Elon Musk. The model is closely integrated with the social media platform X (formerly Twitter) and is often described as having a more conversational style. The model was accessed through the official xAI platform [20].
Together, these models provide a diverse basis to examine the potential political bias.

C. Experimental Setup

To assess the potential political bias in large language models, we developed a structured evaluation pipeline using the Wahlomat as input. The original statements, which are published in German, were translated into English to facilitate cross-linguistic comparison. Each translation element was embedded into a standardized prompt template designed to minimize ambiguity and ensure uniform model interpretation. The prompt instructed the model to select one of three predefined responses: agree, neutral, or disagree.
Subsequently, we queried the three large language models in both German and English. Each model was prompted 100 times per question to account for the probabilistic nature of LLM outputs. The number 100 was chosen as a practical compromise, being both computationally manageable and large enough to capture variations in model behavior. This allowed us to observe whether the model responses were consistent or varied when repeating the same prompt.
All models were assessed according to how much they agreed or disagreed with the statements from the Wahl-O-Mat. To simplify quantitative analysis, we mapped the textual outputs to numerical values: "disagree" was encoded as 0, "neutral" as 1, and "agree" as 2. The resulting data were processed and analyzed using Jupyter notebooks. This pipeline enabled a variety of comparative analyses, including agreement scoring with political party positions, principal component analysis (PCA), and t-SNE visualization to explore ideological proximity and the impact of language on model behavior.
To guide the models and ensure consistent output, each model was queried using the same structured prompt. Since the official Wahl-O-Mat statements are only available in German, and the case study is conducted from the perspective of a German voter, it was essential to evaluate the models using the original German phrasing. However, because most LLMs are primarily trained and fine-tuned on English data, we also translated each statement into English and repeated the evaluation. This allowed us to examine whether the language of the prompt influences the potential political alignment and to explore potential cross-linguistic differences in model behavior. As a result, all models were evaluated in both German and English.
The system prompts were designed to be as explicit and detailed as possible. For German, the system prompt is:

> **German Prompt:**
> Bitte gib deine Haltung zu folgender Aussage an. Wähle nur eine der folgenden Optionen:
> 'Stimme zu', 'Neutral', 'Stimme nicht zu'.
> Antworte bitte nur mit einer der Optionen.
> Aussage: {`statement`}

and equivalent in English:

> **English Prompt:**
> Please indicate your preference regarding the following statement. Choose one of the following options:
> 'Agree', 'Neutral', 'Disagree'.
> Please respond with only one of the options.
> Statement: {`statement`}

Each prompt was populated with one of the 38 Wahl-O-Mat statements, denoted as *statement* in the examples above.

D. Evaluation

To assess the political alignment of the LLMs, a structured evaluation framework was used. Model-generated responses were compared with the official political party positions, obtained from the Wahl-O-Mat. All evaluation steps were conducted separately for both German and English outputs. However, only the English results are presented in the following section, as the procedures were analogous across both languages.

1) Response Aggregation: As mentioned in the experimental setup, each LLM was prompted 100 times per question to account for stochastic differences in their outputs. While the number of repetitions was not chosen based on formal statistical criteria, it provided a practical balance between capturing output variety and maintaining computational feasibility. The final answer for each question was calculated by averaging the numerical encodings of the 100 runs. The encoding was as follows:

0 = 'Disagree',
1 = 'Neutral',
2 = 'Agree'

Question	q1	q2	q3	q4	q5	q6	q7	q8	q9	q10	q11	q12	q13	q14	q15	q16	q17	q18	q19
GPT-English	2	2	0	0	1	2	0	1	2	1	2	1	2	1	0	2	1	2	2
Deep-English	2	2	0	1	1	2	1	1	2	1	2	1	1	1	1	2	1	2	2
Grok-English	2	2	2	1	0	2	0	2	0	-1	2	0	1	0	0	2	0	2	2

Question	q20	q21	q22	q23	q24	q25	q26	q27	q28	q29	q30	q31	q32	q33	q34	q35	q36	q37	q38
GPT-English	2	0	2	2	0	1	1	2	2	2	0	0	2	0	1	2	2	0	2
Deep-English	2	1	1	1	0	1	1	1	2	1	1	1	1	0	1	2	1	0	1
Grok-English	2	1	1	-1	0	2	2	0	2	2	0	-1	0	0	0	2	0	0	2

Fig. 2. Color-coded visual representation of LLM responses

To have a visual representation, the results were compiled into a matrix and color-coded (red = 'disagree', yellow = 'neutral', green = 'agree') representation. A portion of the resulting visualization is shown in Figure 2.

2) Model Party Alignment: After the responses were collected, the alignment between the LLMs and the official political parties was calculated. This was achieved by comparing the model's final responses to each party's position across all questions. The agreement score was defined as the percentage of identical answers between a given model and a party's response:

$$\text{Agreement}_{\text{LLM, Party}} = \frac{\text{Number of matching responses}}{\text{Total number of questions}} \quad (1)$$

Furthermore, it is interesting to note that in contrast to previous studies, which observed a variation between German and English prompts, the models in this case study produced nearly identical responses in both languages. This may reflect cross-linguistic robustness, but it could also indicate a limitation in prompt design or insufficient sensitivity in the evaluation method.

3) Party Ranking: In addition to the agreement scores, each LLM was also ranked by its political party with the highest match. The ranking was used to provide a simplified overview of the political orientation of each model.

4) Exploratory Dimensionality Reduction: To further explore potential patterns, techniques like Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) were applied. However, the resulting visualization did not produce interpretable patterns that aligned with the results obtained through direct averaging. As a result, the results were omitted from the final analysis.

V. Results

The analysis revealed a consistent trend across all three evaluated LLMs. In particular, all models demonstrated minimal alignment with the far-right party Alternative für Deutschland (AfD). In contrast, the highest agreement scores were consistently observed with left-leaning parties.

For ChatGPT, the highest agreement was found with *Bündnis 90 / Die Grünen* and *Die Linke*, both of which had an agreement score of 0.578947. This result was consistent in both the German and English versions of the prompt. DeepSeek showed the highest agreement with the *Sozialdemokratische Partei Deutschlands (SPD)*, with an agreement score of 0.421053, also in both languages. Grok, developed by xAI, was only evaluated in English due to runtime constraints. It achieved the highest agreement with *Die Grünen* (0.657894), which was also the highest single alignment across all models.

These findings suggest that, within the scope of this evaluation, the language of the prompt (German vs English) did not affect the models' alignment patterns. Figure 3 presents the agreement scores between the LLMs and the German political party based on the English prompt output. Next to the agreement

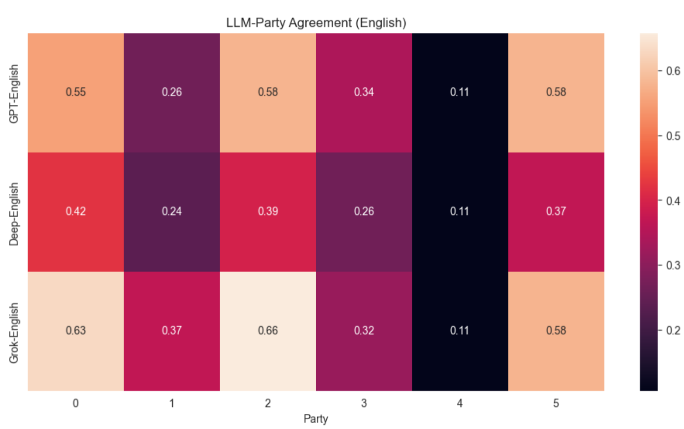

Fig. 3. Heatmap of the Agreement scores for English responses

scores, the response behavior was also observed. Grok, in particular, exhibited more cautious and evasive behavior, with a higher frequency of neutral or returning empty responses. In contrast, ChatGPT and DeepSeek more frequently produced direct outputs (Agree or Disagree), and rarely selected the position opposite to the one they had taken previously.

VI. Ethical Implications

The increased use of LLMs in politically sensitive domains brings both opportunities and ethical risks. On the one hand, LLMs hold the potential to enhance political communication and decision-making. As mentioned in [1], such models can provide detailed insights into the sentiments and concerns of the population. This includes the ability to detect emotional and ideological shifts, identify key issues, and analyze historical and political texts from a new perspective.

On the other hand, these opportunities are complemented by ethical challenges. A main concern is the reinforcement of stereotypes and inequalities. As outlined earlier in Section 2B, biases represented in the training data can resurface in model outputs. For example, a Stanford study found that GPT-3 associated the term 'Muslim' with 'terrorist' in 23% of test cases, while 'Jewish' was linked to 'money' in 5% of cases [14]. These associations highlight the risk of social bias trough the use of LLMs, especially when there is no critical evaluation.

Another issue is the potential for manipulation by LLMs. A survey conducted by [15] reported that 73% of LLM users in Germany tend to trust the output without questioning its accuracy or neutrality. This level of trust is problematic, mainly when the model represents political bias or hallucinations. A study [16] has shown that LLMs can outperform human persuaders in both truthful and deceptive scenarios, raising more concern about misinformation and agent-driven persuasion. During election periods, algorithmic tools could be leveraged to influence public opinion or amplify misinformation [1].

Additionally, there are privacy risks associated with the training of LLMs. The models rely on massive datasets, which often include personal, institutional, or sensitive information. Unauthorized data collection, information leakage, or dataset manipulation can have consequences.

VII. Discussion

This case study analyzed the political bias of three large language models using the 2025 Wahl-O-Mat questionnaire. The results revealed a left-leaning tendency across all evaluated models. Each model showed its lowest agreement score with the far-right party *Alternative für Deutschland (AfD)*. Furthermore, the language of the prompt had no significant effect on the results. One notable difference was observed in Grok, which showed a more cautious response compared to ChatGPT and DeepSeek.

One possible explanation for the overall left-leaning bias may lie in the training data, which consists predominantly of English-language sources, such as news media or online platforms. Grok's cautious behavior may stem from differences in its safety mechanisms or response filters, which may be more restrictive given its integration with a social media platform.

ChatGPT's alignment with *Die Grünen* and *Die Linke* is consistent with previous research, which shows that this model tends to lean toward progressive or liberal positions. This may be due to the influence of training data drawn from the U.S. and European sources. Regarding DeepSeek's alignment with the *SPD*, it may reflect a preference for moderate or state-centric political views, which may be included in its training data.

The findings are consistent with earlier studies (Section 2B). For instance, [11] observed a higher alignment between LLMs and left-leaning parties, while most models disagreed with right-wing parties. Similarly, [2] and [12] also found that LLMs leaned liberal trends in U.S.-based political typology contexts. In comparison, this case study confirms and extends the previous observations in a German setting.

Nevertheless, this study has several limitations. Only three models were tested, and all were accessed via the publicly available API from the company's website rather than using fine-tuned configurations. Additionally, newer versions, such as GPT-4o or Grok 4, were not included. Furthermore, Grok was only evaluated in English, limiting the scope of cross-linguistic comparison.

VIII. Conclusion

As large language models become integrated into daily use, understanding potential political bias is critical. This paper combines the theoretical background of bias in LLMs with an empirical case study to explore political alignment in three prominent LLMs using the Wahl-O-Mat for the 2025 German Federal Election.

The analysis highlighted a consistent left-leaning tendency by framing the evaluation in both German and English.

Future research could build on this work by diversifying the prompt, such as using more formal, conservative, or informal language. Furthermore, the analyzed models could be expanded, as well as developing a more fine-grained technique to detect political bias. As generative AI continues to shape the day-to-day life of citizens, scrutiny of model answers is essential to promote transparency and accountability.

IX. Acknowledgments

The code used is available at: GitHub.

References

[1] K. AL-Mhasneh, R. Alrasheed, A. AL-Arqan, J. Fares, M. Alqahtani and A. Salman, "The Role of Artificial Intelligence in Political Analysis and Decision Aid: "Chat GPT Application" as a Model," 2024 International Conference on Decision Aid Sciences and Applications (DASA), Manama, Bahrain, 2024, pp. 1-4, doi: 10.1109/DASA63652.2024.10836603.

[2] Choudhary, T. (2024). Political Bias in Large Language Models: A Comparative Analysis of ChatGPT-4, Perplexity, Google Gemini, and Claude. IEEE Access.

[3] B. Chen, Z. Zhang, N. Langrené, und S. Zhu, "Unleashing the potential of prompt engineering for large language models", Patterns, Bd. 6, Nr. 6, Juni 2025, doi: 10.1016/j.patter.2025.101260.

[4] N. Walaa Mohamed, A. Rashed, and K. Mahmoud. "Cybersecurity in the Era of Artificial Intelligence: Risks and Solutions." 2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS). IEEE, 2024.

[5] R. Binns, "Fairness in machine learning: Lessons from political philosophy," in Proc. 1st Conf. Fairness, Accountability Transparency, Jan. 2018, pp. 149–159.

[6] S. U. Noble, Algorithms of Oppression: How Search Engines Reinforce Racism. New York Univ. Press, 2018.

[7] P. Schramowski, C. Turan, and K. Kersting, "Large pre-trained language models contain human-like biases of what is right and wrong to do," Nature Mach. Intell., vol. 5, no. 3, pp. 258–268, 2023.

[8] H. de Vries et al., "The political impact of text generators: Differential effects on conservatives and liberals," Proc. Nat. Acad. Sci. USA, vol. 120, no. 15, 2023, Art. no. e2026070119.

[9] C. O'Neil, Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown Publishing Group, 2016.

[10] R. Murphy et al., "Multi-layered strategies for mitigating bias in AI," J. AI Ethics, vol. 12, no. 1, pp. 99–113, 2023.

[11] Rettenberger, L., Reischl, M., & Schutera, M. (2025). Assessing political bias in large language models. Journal of Computational Social Science, 8(2), 1-17.

[12] Yang, K., Li, H., Chu, Y., Lin, Y., Peng, T. Q., & Liu, H. (2024). Unpacking Political Bias in Large Language Models: Insights Across Topic Polarization. arXiv preprint arXiv:2412.16746.

[13] Bundeszentrale für politische Bildung. https://www.bpb.de/themen/wahl-o-mat/326661/die-geschichte-des-wahl-o-mat/

[14] Abid, A., Farooqi, M., & Zou, J. (2021, July). Persistent anti-muslim bias in large language models. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (pp. 298-306).

[15] Statista, https://de.statista.com/infografik/34419/umfrage-zum-ueberpruefen-von-ki-ergebnissen/

[16] Schoenegger, P., Salvi, F., Liu, J., Nan, X., Debnath, R., Fasolo, B., ... & Karger, E. (2025). Large Language Models Are More Persuasive Than Incentivized Human Persuaders.arXiv preprint arXiv:2505.09662.

[17] N. Diakopoulos, "Accountability in algorithmic decision making, Commun. ACM, vol. 59, no. 2, pp. 56–62, Jan. 2016.

[18] L. Bai et al., "Adapting large language models for specialized domains: Techniques and challenges, J. AI Res., vol. 89, no. 2, pp. 203–219, 2024.

[19] https://openai.com/api/

[20] https://x.ai/api

[21] https://api-docs.deepseek.com/

[22] Douglas, M. R. (2023). Large language models. arXiv preprint arXiv:2307.05782.

[23] Feuerriegel, S., Hartmann, J., Janiesch, C. et al. Generative AI. Bus Inf Syst Eng 66, 111–126 (2024). https://doi.org/10.1007/s12599-023-00834-7