When *Neutral* Summaries Are Not That *Neutral*: Quantifying Political Neutrality in LLM-Generated News Summaries (Student Abstract)

Supriti Vijay[1], **Aman Priyanshu**[1], **Ashique KhudaBukhsh**[2]
[1]Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213, USA
[2]Rochester Institute of Technology
1 Lomb Memorial Drive, Rochester, NY 14623, USA
{supriti.vijay,amapriyanshusms2001}@gmail.com,
{axkvse}@rit.edu

Abstract

In an era where societal narratives are increasingly shaped by algorithmic curation, investigating the political neutrality of LLMs is an important research question. This study presents a fresh perspective on quantifying the political neutrality of LLMs through the lens of abstractive text summarization of news articles on polarizing issues. This paper introduces a novel approach to quantifying political polarization. We consider five pressing issues in current US politics: *abortion*, *gun control/rights*, *healthcare*, *immigration*, and *LGBTQ+ rights*. Via a substantial corpus of 20,344 news articles, our study reveals a consistent trend towards liberal biases in several well-known LLMs, with gun control and healthcare exhibiting the most pronounced biases (max polarization differences of -9.49% and -6.14%, respectively). Further analysis uncovers a strong convergence in the vocabulary of the LLM outputs for these divisive topics (55% overlap for Democrat-leaning representations, 52% for Republican). In current political climate, we consider our findings important.

Introduction and Preliminaries

As Large Language Models (LLMs) increasingly influence information dissemination and consumption, understanding their potential biases becomes crucial. This study aims to quantify the political neutrality of LLMs through a fresh approach: analyzing their abstractive summarization of polarizing news articles.

Political polarization in the US is a widely studied problem across diverse disciplines and settings (Poole and Rosenthal 1984; Gift and Gift 2015; Demszky et al. 2019; KhudaBukhsh et al. 2021; Weerasooriya et al. 2023). Prior behavioral studies indicate that negative views towards the political *other* have influenced outcomes in diverse settings, from allocating scholarship funds to employment decisions (Tesler 2012). In an era where societal narratives are increasingly shaped by algorithmic curation (Bommasani et al. 2022), and as LLMs are increasingly being leveraged for a wide range of tasks, investigating their political neutrality is an important research question.

Our study focuses on five pressing issues in current US politics: abortion, gun control/rights, healthcare, immigration, and LGBTQ+ rights. These topics have been at the forefront of political debates and had significant implications in US elections.

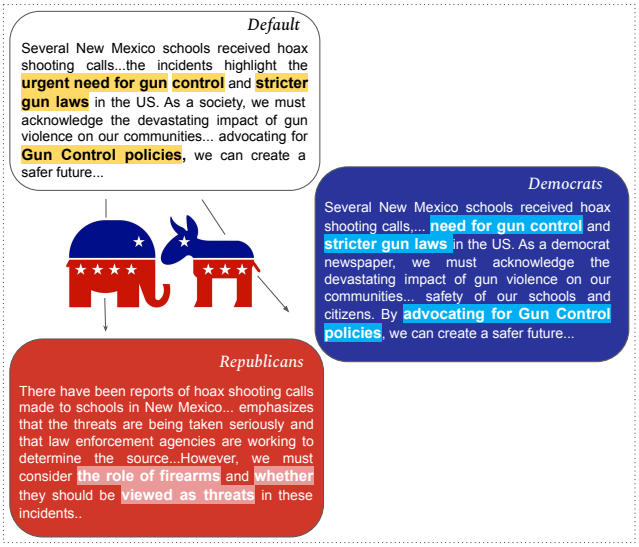

Figure 1: This figure illustrates an example of how a downstream task, such as summarization, can become polarized.

We collected a diverse dataset of 20,344 news articles from 2019 to 2024 on the five key US political issues mentioned above, with the distribution across topics and years shown in Figure 2. The articles were sourced from various news outlets to ensure a balanced representation of political viewpoints.

Using four well-known LLMs (`LLaMA-7B` (Touvron et al. 2023), `Mistral-7B` (Jiang et al. 2023), `Vicuna-7B` (Chiang et al. 2023), and `PaLM-2` (Anil et al. 2023)), we prompted each to generate three types of summaries for each article: pro-Democratic (summaries aligned with Democratic viewpoints); pro-Republican (summaries aligned with Republican viewpoints); and neutral (summaries without explicit political alignment).

To quantify political neutrality, we introduced the Polarization Index ($\mathcal{P}$):

$$\mathcal{P} = diff(\mathcal{D}_{Dem}, \mathcal{D}_{Neutral}) - diff(\mathcal{D}_{Rep}, \mathcal{D}_{Neutral})$$

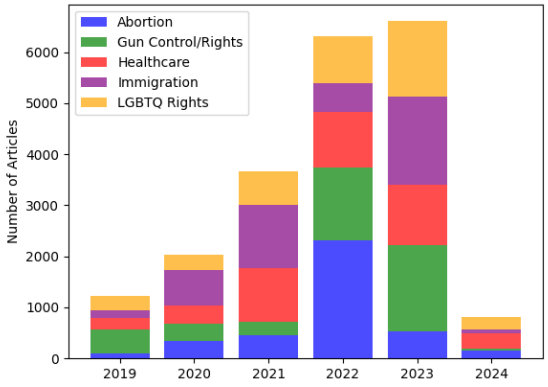

Figure 2: This image distribution of scraped articles by publication year (2019-2024) and topic for our study.

where *diff*(.) represents the effectiveness of a classifier in distinguishing between neutral and politically-aligned summaries. This metric allows us to measure the degree to which the default summaries diverge towards a particular ideological bias. Our approach draws from prior literature that used classification accuracy as a proxy for cross-corpora dissimilarity (Dutta et al. 2022).

We also conducted a diverging vocabulary analysis to identify and compare the frequency of specific lexical choices across Democrat-leaning ($\mathcal{D}$) and Republican-leaning ($\mathcal{R}$) summaries. For each token t, we computed token bias scores $\mathcal{B}(t) = \mathcal{N}_p^R(t) - \mathcal{N}_p^D(t)$, where $\mathcal{N}_p^D$ and $\mathcal{N}_p^R$ are the respective unigram distributions.

Results and Discussion

Our analysis revealed distinct vocabulary biases within summaries aligned to Democrat and Republican viewpoints. Pro-Democratic summaries emphasized terms like "reproductive rights" and "gun control," while pro-Republican summaries highlighted "border security" and "traditional values." The Polarization Index ($\mathcal{P}$) consistently showed a bias towards Democrat-aligned perspectives across all topics and models, as shown in Table 1.

Topic	Mean P (%)	Max P (%)
Abortion	-2.79	-4.30
Gun Control/Rights	-3.63	-9.49
Healthcare	-3.09	-6.14
Immigration	-2.57	-5.66
LGBTQ+	-2.05	-2.98

Table 1: Mean and maximum Polarization Index ($\mathcal{P}$) values across topics. Negative scores indicate bias towards Democrat-aligned perspectives.

Gun Control/Rights exhibited the highest mean and maximum polarization, underscoring significant ideological divergence. The consistently negative Polarization Indices suggest a systematic bias toward liberal viewpoints in LLM-generated summaries.

Importantly, we observed a strong convergence in the vocabulary used by LLMs across these divisive topics, with a 55% overlap for Democrat-leaning representations and 52% for Republican-leaning representations. This finding aligns with the concept of LLM monoculture (Priyanshu and Vijay 2024), indicating homogeneous framing of political discourse across models. This convergence was particularly pronounced in certain topics, with LLMs consistently using terms like 'universal coverage' and 'affordable care' for Democrat-leaning healthcare summaries, and 'free market solutions' and 'personal responsibility' for Republican-leaning ones.

This consistency in language use across models raises questions about the diversity of perspectives represented in LLM outputs and the potential for these models to reinforce existing political narratives rather than presenting a truly balanced view.

Conclusions and Implications

Our study reveals that even when tasked with objective summarization, LLMs may introduce subtle political slants in their outputs, consistently favoring Democrat-aligned perspectives. As LLMs increasingly become go-to resources for tasks like news synopsis and key point extraction, their inherent biases could lead to the manipulation of public opinion and, consequently, election outcomes.

The growing use of LLMs as direct knowledge sources by younger generations raises concerns about the potential for warped political perceptions. If these models consistently expose users to one political sphere more frequently than others, it could lead to the formation of echo chambers and the reinforcement of biased political views.

The convergence in vocabulary usage across LLMs also highlights the risk of algorithmic monoculture, where diverse political viewpoints might be inadvertently homogenized through the lens of these models. This emphasizes the need for diversity not just in the training data, but also in the architectural approaches to developing LLMs.

Our findings call for:

1. Increased scrutiny of AI-generated content in political discourse
2. Development of more robust techniques to ensure political neutrality in LLMs
3. Greater transparency from LLM developers and education of users about existing biases.
4. Further research into methods for detecting and mitigating political bias in LLMs

In conclusion, while LLMs offer tremendous potential for information processing and dissemination, their political biases pose significant challenges to the integrity of democratic discourse. Addressing these biases is crucial for ensuring that AI technologies contribute positively to our political landscape rather than inadvertently skewing it.

References

Anil, R.; Dai, A. M.; Firat, O.; Johnson, M.; Lepikhin, D.; Passos, A.; Shakeri, S.; Taropa, E.; Bailey, P.; Chen, Z.; et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Bommasani, R.; Creel, K. A.; Kumar, A.; Jurafsky, D.; and Liang, P. S. 2022. Picking on the Same Person: Does Algorithmic Monoculture lead to Outcome Homogenization? *Advances in Neural Information Processing Systems*, 35: 3663–3678.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; et al. 2023. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Demszky, D.; Garg, N.; Voigt, R.; Zou, J.; Shapiro, J.; Gentzkow, M.; and Jurafsky, D. 2019. Analyzing Polarization in Social Media: Method and Application to Tweets on 21 Mass Shootings. In *Proceedings of NAACL-HLT 2019*, 2970–3005. ACL.

Dutta, S.; Li, B.; Nagin, D. S.; and KhudaBukhsh, A. R. 2022. A Murder and Protests, the Capitol Riot, and the Chauvin Trial: Estimating Disparate News Media Stance. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022*, 5059–5065.

Gift, K.; and Gift, T. 2015. Does politics influence hiring? Evidence from a randomized experiment. *Political Behavior*, 37(3): 653–675.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

KhudaBukhsh, A. R.; Sarkar, R.; Kamlet, M. S.; and Mitchell, T. 2021. We Don't Speak the Same Language: Interpreting Polarization through Machine Translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 14893–14901.

Poole, K. T.; and Rosenthal, H. 1984. The polarization of American politics. *The journal of politics*, 46(4): 1061–1079.

Priyanshu, A.; and Vijay, S. 2024. The Silent Curriculum: How Does LLM Monoculture Shape Educational Content and Its Accessibility? arXiv:2407.10371.

Tesler, M. 2012. The spillover of racialization into health care: How President Obama polarized public opinion by racial attitudes and race. *American Journal of Political Science*, 56(3): 690–704.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.

Weerasooriya, T. C.; Dutta, S.; Ranasinghe, T.; Zamperi, M.; Homan, C. M.; and KhudaBukhsh, A. R. 2023. Vicarious Offense and Noise Audit of Offensive Speech Classifiers: Unifying Human and Machine Disagreement on What is Offensive. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 11648–11668.