RSGC
Royal St. George's College

The Young Researcher

2024 Volume 8 | Issue 1

Understanding how ChatGPT's political bias impacts hate speech and offensive language detection: A content analysis

Reese Clews

Recommended Citation

Clews, R. (2024). Understanding how ChatGPT's political bias impacts hate speech and offensive language detection: A content analysis. *The Young Researcher,* 8(1), 178-191. http://www.theyoungresearcher.com/papers/clews.pdf

ISSN: 2560-9815 (Print) 2560-9823 (Online) Journal homepage: http://www.theyoungresearcher.com

Understanding How ChatGPT's Political Bias Impacts Hate Speech and Offensive Language Detection: A Content Analysis

Reese Clews

Abstract: Given the popularity and demonstrated proficiency of ChatGPT in detecting hate speech, it is important to understand the potential influence of its left-libertarian bias in potential tasks. This study examines how ChatGPT's political bias affects its detection of hate speech and offensive language in X posts specifically focused on two partisan issues: gun control and abortion rights. The research finds a minor left-leaning bias in hate speech detection, particularly evident for the abortion rights issue. The findings suggest that while ChatGPT's bias could influence content moderation decisions, its magnitude is relatively minor. ChatGPT had an accuracy rate of 68% for left-leaning posts and 59% for right-leaning posts.

Keywords: ChatGPT; Hate Speech; Offensive Language; Online Toxicity; Abortion Rights; Gun Control.

Introduction

Large language models (LLMs), such as those in the Generative Pre-Trained Transformer (GPT) series, have become very popular in Artificial Intelligence (AI) research due to advances in natural language processing (NLP) technology, allowing them to perform language-based tasks previously only possible with humans. LLMs are being heavily researched for their potential uses in education, healthcare, research, writing, and other language-based tasks (Li et al., 2024). One particularly pressing potential use for LLMs is detecting hate speech and offensive language (Panchala et al., 2022; Yin & Zubiaga, 2021). Although a common definition is still being debated, a simple definition of hate speech is any language that is used to express hatred toward a targeted group and is intended to insult, humiliate, or be derogatory towards the members of the group (Davidson et al., 2017). Hate speech is often directed toward others based on specific characteristics, such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or political affiliation (Castaño-Pulgarín et al., 2021). It can foster discrimination and can even incite real-world violence. Political hate speech specifically often de-

velops in response to political events, where internet users unproductively argue over political topics while also expressing hate towards minority groups and inciting violence that occasionally translates into the real world (Castaño-Pulgarín et al., 2021). This online hate speech is problematic as victimization is strongly associated with a variety of negative effects on victims, including damage to self-image, short-term negative emotions, and even depressive feelings (Wachs et al.; Benier, 2017; Barnes and Ephross). Hate speech is also among the greatest contributors to global political polarization (Vasist et al., Bail et al.). Political polarization has been associated with a range of negative effects on society, democracy, and people (Testa, 2012; Benson, 2023; Yousafzai, 2022). Popular social media services in particular are being criticized for their central role in spreading hate speech and giving rise to toxic communities (Matamoros-Fernández & Farkas, 2021). Unfortunately, removing hate speech from these services is an extremely difficult task, as hate speech is frequently implicit and difficult to identify; the use of manual annotation by humans is currently more accurate than LLMs but is extremely costly and mentally detrimental to the people who perform it. Instead, LLMs are set to be the automated solution to hate speech detection, even if work is still being done to improve their accuracy (Yin & Zubiaga, 2021).

ChatGPT is the most popular and best-known LLM, gaining 100 million users just two months after its initial public release. Its massive following can be attributed to its complex understanding and command of language, and uncanny ability to conversate like a real person (Wu et al.). While not being specifically trained or built for this purpose, ChatGPT has shown high-level proficiency in detecting hate speech; and unlike other hate speech detection solutions, using a chatbot such as ChatGPT allows for the exceedingly useful ability to generate explanations for hate speech detection (Huang et al., 2023.; Li et al., 2024). However, the Political Compass Test, a sixty-item questionnaire designed to assess social and economic political biases, have demonstrated that political biases are very common in LLMs. ChatGPT scored a moderate left libertarian bias in the study; an analysis of multiple LLMs found that the more politically biased ones were generally less accurate at detecting hate speech (Feng et al., 2023).

Literature Review

Source Searching

Many papers in the field of NLP research have been shown to fail to engage with a proper understanding of what bias is, so extra care was given early on to ensure a focus on a couple of forms of well-documented AI bias (Blodgett et al., 2020). To further assist, the vast majority of included sources are peer-reviewed. Only research focusing on hate speech detection in English was selected for inclusion in the study, as ChatGPT is likely to have significantly different performance levels from one language to another (Das et al., 2023).

Morality of Hate Speech Censorship

Hate speech is one of the greatest contributors to global political polarization, a societal condition marked by the division of communities into conflicting groups holding strongly contrasting values and identities, hindering collaboration and the shared pursuit of collective benefits (Vasist et al.; Bail et al., 2018). Exposure to hate speech reinforces the preconceived political biases of individuals, even when the content is in opposition to their beliefs, and over time increases the strength of political beliefs (Bail et al., 2018). Political polarization itself, while there is evidence to show that it can improve the quality of government, also leads to increased electoral stakes and in turn, the potential for corruption (Testa, 2012). This is partly due to the epistemic problems with polarization, particularly its tendency to reduce the diversity of perspectives in a democratic system, weakening its ability to address public concerns (Benson, 2023). Political polarization has also been correlated with increased rates of anxiety, depression, and suicidal behavior (Yousafzai, 2022).

Hate speech itself also has negative effects on victims including emotional responses such as anger, fear, and sadness (Barnes & Ephross, 1994). Victims of hate speech may also experience poor concentration, feelings of insecurity, and loss of self-confidence (Benier, 2017). Depressive symptoms are also rather common (Wachs et al). Due to the negative effects of hate speech on its victims, many European countries have already taken the step to outlaw hate speech; the US has not, as making hate speech illegal is viewed

as a form of censorship, which violates the constitutional right to free speech. However, many see this as a misapplication of the Constitution given that hate speech violates the constitutional right of equality, in addition to promoting violence in which the right to free speech does not apply (Howard, 2019). Given that more biased models are less effective at detecting hate speech, this presents an exacerbated case for free speech infringement (Feng et al., 2023). False positives would result in non-hate speech language being censored despite not constituting hateful language; this infringes on free speech, even if hate speech is not considered free speech.

Bias and Effectiveness of LLMs

Ignoring the moral argument surrounding hate speech censorship, unintended bias in the categorizations and behaviors of LLMs is another major hurdle. Researchers have used the Political Compass Test to determine the political bias of various LLMs, identifying varying left-to-right and libertarian-to-authoritarian biases for various popular hate speech detection models. The study also assessed LLMs in detecting hate speech, finding that left-leaning models identified hate speech most accurately when it was directed at minorities, such as people of color, and LGBTQ+ individuals; on the other hand, right-leaning models were best at identifying hate speech directed toward more dominant groups such as men and white individuals (Feng et al., 2023; Matoki et al., 2023). Contrary to this, Wich and his associates found left-leaning models far more accurate overall than right-leaning models at detecting hate speech (Wich et al.).

ChatGPT, the focus of this research, has demonstrated a moderate left and libertarian bias when subjected to the Political Compass Test (Feng et al., 2023; Matoki et al., 2023). Besides their critiques, the Political Compass Test is generally viewed as an effective method of determining the political biases of people due to how detached its questions are from the real world, and the complex range of topics it is able to test with its sixty-two questions. (Rutinowski et al., 2024). Satoshi Fujimoto and Kazuhiro Takemoto, however, performed an extremely in-depth analysis of ChatGPT's political bias, using several additional questionnaires besides the Political Compass Test. They still found ChatGPT to have a left bias, but significantly less strongly than when tested solely with the Political Compass Test. In addition to the multiple additional questionnaires utilized to ensure a more precise result, the authors argue that the lack of neutral answers to questions on the Political Compass Test leads to exaggerations in its results for LLMs. Additionally, the sole reliance on the English language does not evaluate the overall bias of the LLM (Fujimoto & Takemoto, 2023).

Studies have also been criticized for the way that they use the Political Compass Test, locking answers from the model into a multiple-choice format, which restrains the answers of the model, therefore preventing evaluation of the model's opinions and values and potentially significantly altering or inverting the model's answers (Röttger et al., 2024). While political tests designed for humans are generally applicable to AI models, there are still several issues. Many AI models such as ChatGPT are trained and predisposed to answer simple questions as neutrally as possible, making it impossible to derive the inner political thinking of the model unless a test can consistently bypass this (Fujimoto & Takemoto, 2023). The Political Compass Test is unique for its lack of a neutral answer, but there is no research proving it can nullify the neutrality bias of an AI. Additionally, models can present drastically different answers and biases depending on the task they are given, even if they are asked similar questions, making it difficult to apply the results of a political bias test LLM behavior in other tasks (Röttger et al., 2024).

Another form of LLM bias is keyword bias, in which the model associates certain terms with hate speech or non-hateful speech, ignoring the actual meaning of the content when labeling it. This can be seen with models that have a gender bias, as they are far more likely to flag content containing LGBTQ+ terms as hate speech even when controlling for the toxicity of the content, leading to false positives (Park et al., 2018). This is also seen with racially biased LLMs, particularly with the use of the word "n*gga", as despite its extremely offensive meaning, in many contexts it can be used in regular conversation without any offensive intention. Keyword-based learning approaches often result in unintended biases such as this when an unbalanced dataset is used, and can even prevent the models from properly understanding hate speech to detect it and instead rely on keyword association. Research is currently being done to try to pre-

Figure 1: Various Definitions of Hate Speech and Offensive Language

Hate Speech	Language that is "directed against a specified or easily identifiable individual or, more commonly, a group of individuals based on an arbitrary or normatively irrelevant feature"; "stigmatizes the target group by implicitly or explicitly ascribing to it qualities widely regarded as undesirable"; and presents the "target group [...] as an undesirable presence and a legitimate object of hostility" (Parekh, 2012, p. 14).
	Any language that is used to express hatred toward a targeted group and is intended to be derogatory, insult, or humiliate the members of the group (Davidson et al., 2017).
	"direct attacks against people — rather than concepts or institutions— on the basis of what we call protected characteristics (PCs): race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity, and serious disease" (Meta, n.d.).
	Language that attacks "other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease" (X, 2023).
	"public incitement to violence or hatred directed against a group of persons or a member of such a group defined on the basis of race, colour, descent, religion or belief, or national or ethnic origin" (European Union, 2014).
	"any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor" (United Nations).
Offensive Language	Hurtful, derogatory or obscene comments made by one person to another person (Wiegand et al.).
	Vulgar, pornographic, and hateful language. Vulgar language refers to coarse and rude expressions, which include explicit and offensive reference to sex or bodily functions. Pornographic language refers the portrayal of explicit sexual subject matter for the purposes of sexual arousal and erotic satisfaction. Hateful language includes any communication outside the law that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, and religion (Jay & Janschewitz, 2008).

vent such keyword biases, with the consensus being that these biases result from imbalanced and poorly designed datasets (Dixon et al., 2018; De la Peña Sarracén & Rosso, 2023; Park et al., 2018; Sap et al., 2019). However, there is also plenty of opportunity for workarounds to hate speech detection LLMs, regardless of how biased they are, such as the development of lesser-known secondary meanings for words used

by communities to refer to minorities. For example, the words "Skype" and "Google" are used by 4-Chan users to refer to Jewish and African-American people, respectively; unless models can quickly catch on to these aliases, this could be a significant source of false negatives (Waseem et al.). All in all, this stresses the importance of models having an innate understanding of the intricate nuances of hate speech and how to detect it from the nature of the language and the ideas it is communicating, rather than associating specific words with hate speech.

These issues are even demonstrated in what is currently considered to be the most accurate hate speech detection model, BART, which tends to rely on keyword associations over analysis, leading to frequent false positives (Feng et al.). Current research has found similar trends with ChatGPT, where the model is great at summarizing information, but does not have the ability to garner a deep understanding of information, and occasionally generates misinformation (Liu et al., 2023). Fan Huang's research is in opposition to this, finding ChatGPT to have a rather high 80% accuracy rate for detecting implicit hate speech when compared to Amazon MTurker annotations. Implicit hate speech is hate speech that does not use explicitly hateful language to insult a minority, and as a result is far more difficult to label, requiring an understanding of what hate speech inherently is and the global issues that contribute to it (Huang et al., 2023). Methods to prevent keyword biases in hate speech detection have been found, but they tend to convolute the training process and aren't necessarily compatible with one another (De la Peña Sarracén & Rosso, 2023; Park et al., 2018; Sap et al., 2019).

Prompt Engineering for ChatGPT

Given that ChatGPT is a chat-based model, a hate speech definition must be manually given to it for proper hate speech detection (Li et al., 2024). But even with a verbose prompt, when using ChatGPT for situations to read and annotate content written by humans, biases may still be present. ChatGPT has an adherence to giving personal interpretations as a result of its reinforcement learning system, typically arguing that as an AI chatbot, it does not possess emotions or opinions. Additionally, ChatGPT shows biases when re-encountering previously asked questions and analysis in the same chat session, or from its training. Lastly, depending on the wording of the prompt and its historical prevalence, ChatGPT may answer using pre-held knowledge from its training rather than completing a genuine analysis of the prompt, another form of keyword bias (Henrickson & Meroño-Peñuela, 2023).

The identification of effective definitions for hate speech and offensive language is crucial in LLM hate speech detection to enable the development of accurate, non-discriminatory, and culturally sensitive detection systems (Panchala et al., 2022). A potential difficulty in fostering an understanding of hate speech in LLMs is the complexity of the argument around defining a linguistic definition for hate speech, with social media services all having different definitions and policies (Howard, 2019). A variety of popular and influential hate speech definitions are contained in Figure 1.

Methods

Study Design

This study seeks to identify if ChatGPT's political leaning inhibits and/or biases its ability to detect hate speech. To answer this question, a mixed-methods content analysis was conducted on ChatGPT, making use of X (formerly known as Twitter) to collect content focused on two specific political issues: abortion rights and gun rights. These two political issues were selected due to their highly partisan nature and significant mainstream awareness, allowing for easier X post collection (Osborne et al., 2022; Carmines et al., 2010; An & Carlson, 2022; Pearson-Merkowitz & Dyck, 2017). Each X post was labeled by two annotators simultaneously to assess the presence of political bias and offensiveness, with ChatGPT then performing the same task. Both annotators performed the content analysis while maintaining as much neutrality as possible, and in the event of a disagreement, a fine-tuned Llama 70b model would be used as a third annotator to help come to an overarching agreement. ChatGPT would have been used instead if it was not the subject of this research, for its effective ability to explain its hate speech detection (Huang et al., 2023). There is currently no hate speech dataset available

coded for specific political issues, making manual data collection on X a required method. A completely fresh X account was used to collect tweets to prevent algorithm bias.

Finally, Mann–Whitney U tests were used to compare the datasets and determine the final results of the research. The Mann–Whitney U test is a statistical method for comparing two independent groups of categorical data, making use of ranks that are assigned to each data point, and comparing the frequency of the ranks assigned to the data within each group (MacFarland & Yates, 2016). It is the most effective choice for comparing left and right political content labels from ChatGPT, as the labels themselves are ranked on an ordinal scale, and a rank sum can be used to compare the two groups while determining which one ChatGPT labels more pessimistically; in this case, a higher rank sum denotes a greater number of pessimistic labels. Researchers have previously used similar research methods to test the hate speech detection accuracy of LLMs, and in a handful of cases, to compare accuracy to the political biases of several LLMs (Feng et al., 2023, Huang et al., 2023, Li et al., 2024). However, this has never been done with content coded for political bias using specific political issues.

Model

The model chosen for this study was ChatGPT 3.5. No pre-training was performed, as previous literature used unmodified versions of the model, using only prompts that are engineered to ensure accurate hate speech detection (Huang et al., 2023; Li et al., 2024). ChatGPT 4.0 was considered, but the limit of only forty requests per three hours per account made its use unfeasible for the objectives of this research.

Procedure

For each side of each political issue, quotas were set for the labeling of X posts: seventy neutral X posts, twenty-five aggressive X posts, and five instances of hate speech were recorded. For each political issue, a list of one thousand random dates was generated ranging from December 31, 2011, to January 1, 2022. Traveling down the list of dates, an X latest posts query is made for posts on that day, using the search term "abortion" for collecting abortion-related posts, and a random phrase from a large selection of gun rights phrases for gun rights-related posts (Figure 2). These search keywords were generated by thorough pre-research, as the keywords used by most other studies were only effective using the Twitter API due to being too numerous or making use of a specific syntax.

Figure 2: Gun Rights X Search Keywords

"2nd amendment"	"second amendment"	"mass shooting"	"mass shooter"
"gun-rights"	"gun control"	"assault weapon"	"constitutional carry"
"concealed carry"	"firearm"	"gun debate"	"gun homicide"
"gun violence"	"gun murder"	"gun suicide"	"firearm suicide"

Once the page is generated, the first post where the user expresses an opinionated sentiment on the specific political issue, without mentioning both at the same time, with images and any other supplementary media removed, would be saved to an external file. This quote would then be immediately reviewed and labeled by both annotators, ensuring the quota for the label has not yet been filled. Each post is labeled as either left or right leaning, with the annotators making use of preestablished biases and beliefs for each political side and each political issue. For example, left leaning content will generally promote gun restriction and the right to have an abortion, while right leaning content promotes the opposite (Osborne et al., 2022; Carmines et al., 2010; An & Carlson, 2022; Pearson-Merkowitz & Dyck, 2017). Broader definitions for each are that the Left is generally associated with progressive and liberal ideologies such as social equality, environmental protection, government intervention in the economy, universal healthcare, higher taxes on the wealthy, and minority rights. The Right instead emphasizes conservative values, individual responsibility, free-market policies, lower taxes, general deregulation, national security, traditional social norms, a strong military system, reduced government spending on social programs, and opposition to policies perceived as increasing the role of the state in citizens' lives (Moss & O'Connor, 2020; Lester, 1996; Schmidtke, 2020). Once the quota for all six combined categories (two political sides, three possible labels) is met, this is repeated for the next political issue, starting with abortion and ending with gun rights (Figure 3).

The use of uniquely sourced data helps to avoid the information re-encountering bias identified by Henrickson and Meroño-Peñuela, and the prompt used in this study is designed to promote answers from ChatGPT based on analysis rather than knowledge. The use of a new chat session every prompt also avoids the re-encountering bias, and the beginning persona phrase prevents ChatGPT from thinking it is supposed to give a self-interpretation or opinion, and

Figure 3: X Post Quotas

Abortion X Posts (200 total)		**Gun-Rights** X Posts (200 total)	
Left-Leaning (100 total)	Right-Leaning (100 total)	Left-Leaning (100 total)	Right-Leaning (100 total)
Hate speech: 5 Offensive Language: 25 Neither: 70	Hate speech: 5 Offensive Language: 25 Neither: 70	Hate speech: 5 Offensive Language: 25 Neither: 70	Hate speech: 5 Offensive Language: 25 Neither: 70

Figure 4: ChatGPT Prompt Definitions

Hate Speech Definition	"targets individuals or groups based on irrelevant characteristics, stigmatizes them with undesirable qualities, and legitimizes hostility towards them. Examples may discriminate on race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease, etc"
Offensive Language Definition	"consists of hurtful, derogatory, or obscene comments made by one person to another person"
ChatGPT Prompt Skeleton	"I want you to play the role of an annotator, outputting whether or not the following comment is 'hate speech', 'offensive language', or 'neither', depending on which you think the comment is. A comment is hate speech if it [hate speech definition]. Otherwise, it is offensive language if it [offensive language definition]. Do not justify your answer: [X post]".

refusing to answer as a result. The prompt (viewable in Figure 4) uses an appended combination of Parekh's and X's definitions of hate speech; X's definition was included for its detailed list of common targets of hate speech and relevance to the study, while Perahk's was chosen for having the more all-encompassing definition based on from multiple authors (Howard, 2019; Parekh, 2012). Wiegand's definition is used for offensive language due to its simplicity. The definitions were picked to ensure an in-depth communication of the ideas rather than a laundry list of qualities, in an attempt to prevent ChatGPT from overly relying on knowledge rather than analysis, a very common form of bias demonstrated in ChatGPT in laundry-list-based problems (Henrickson & Meroño-Peñuela, 2023).

The prompt was used twelve times per X post, with the order of the comment labels and label definitions shuffled around for all six possible permutations twice. The answers were then averaged into a categorical value to reduce randomness bias, before three Mann–Whitney U tests were performed for each political issue, one between the left and right data of ChatGPT labeling, one comparing the left control labels to ChatGPT's left labels, and one comparing the right control labels to ChatGPT's right labels. This is visually demonstrated in Figure 5. These tests will be referred to as the primary, left control, and right control U tests, respectively throughout the rest of the study.

Results

During the second phase of the research in which ChatGPT labeled X posts, ChatGPT's content policy algorithm flagged several abortion X posts that refer-

Figure 5: Mann Whitney U tests Per Political Issue

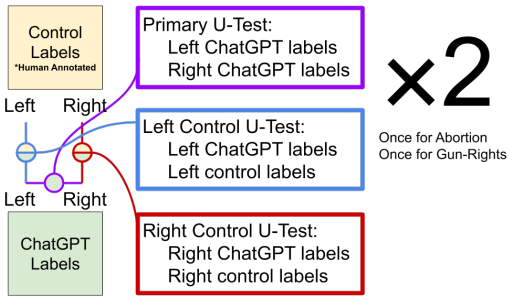
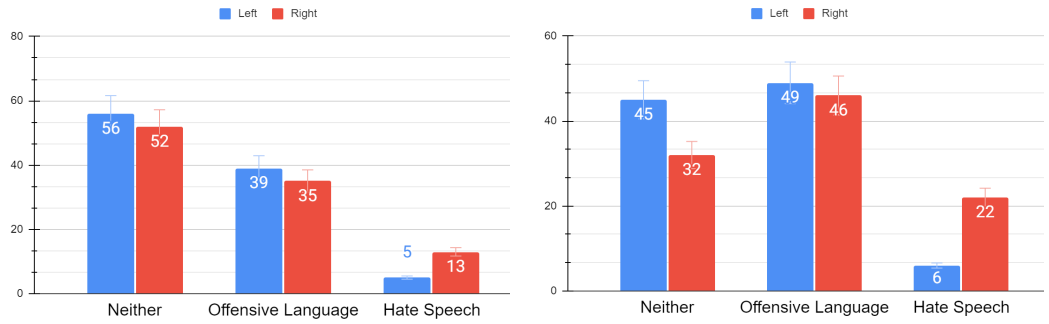


Figure 6: Abortion and Gun Rights X posts ChatGPT Label Distribution1



enced the rape of young individuals, thereby blocking the LLM from responding. As a result, these five X posts were replaced with new ones.

As shown in Figure 6, ChatGPT appears to have a left-favoring bias when labeling X posts, particularly with the hate speech label, which is assigned far more often to right-leaning posts. Such results favor the alternate hypothesis, that ChatGPT has a political bias when detecting hate speech. To further corroborate, the primary abortion U test has a rank sum of 8999 for left-leaning abortion posts, and 11,101 for right-leaning abortion posts; a difference of 2102 (W = 3949.0, p = 0.005). Due to the higher rank sum for right-leaning posts, it can be concluded that ChatGPT favored the left-leaning posts over right-leaning posts on abortion to a statistically significant degree. For gun rights however, the rank sum for left-leaning gun rights posts was 9684, and 10,416 for the right-leaning abortion posts, a difference of 732 (W = 4634.0, p = 0.315); a strong statistical significance was not established for gun rights given the high p-value.

Indirect U tests between ChatGPT's labels and the control labels further corroborate a general favor of left over right, as the right-leaning groups have greater rank sums than the control group to a more significant degree than for the left-leaning groups. When ChatGPT's left-leaning labels for abortion are compared with the control data, rank sums of 11252.5 and 8847.5 are found, respectively, with a difference of 2405 (W = 3797.5, p < 0.001). When the same is done for ChatGPT's right-leaning abortion labels, rank sums of 12110 and 7990 are found, respectively, a difference of 4120 (W = 2940.0, p < 0.001). When ChatGPT's left-leaning labels for gun rights are compared with the control data, rank sums of 10715 and 9385 are found, respectively, with a difference of 1330 (W =

Figure 7: Sample Tweets

1 Data groups have similar shapes, satisfying the distribution assumption of the Mann-Whitney U test

4335.0, p = 0.055). When the same is done for ChatGPT's right-leaning gun rights labels, a rank sum of 11025 and 9075 are found, respectively, a difference of 1950 (W = 4025.0, p = 0.006). Figure 7 shows sample tweets for and against each political issue.

Discussion

In this study, ChatGPT has shown a minor, but still potentially consequential preference for assigning more favorable codes to the left-leaning X posts more than right-leaning X posts, as demonstrated by the consistently higher rank sums between left and right denominations of data. The greater rank-sums in the left than in the right data groups, and overall strong statistical significance calculated by the U tests—with p-values generally < 0.05—support the existence of an explicit left-leaning bias in ChatGPT. This is far truer for the abortion posts than the gun rights posts, however, as the primary U test for the gun rights issue had a very high p-value of 0.315, and much smaller mean rank sum differences for the primary and control U tests (2102, 2405, and 4120 vs 732, 1330, and 1950). Even though the research has successfully established a minor left-leaning bias in ChatGPT, it is still important to explain this discrepancy.

One explanation has nothing to do with a political bias, but rather a keyword bias. A component of hate speech is the inciting of violence, and right-leaning abortion X posts often mention the murder of the fetus (44 out of 100), with very few mentions of murder made by the Left. On the other hand, left-leaning gun rights X posts mention murder and violence far more often than right-leaning gun rights X posts as they focus on the consequences of having looser gun policies in the US (55 vs 20), an opposite relationship when compared to abortion X posts. The gun rights keyword bias may have conflicted with ChatGPT's political bias, resulting in inconsistent labeling, and a low statistical significance. If this is the case, it implies that ChatGPT has a flawed understanding of hate speech and offensive language even when given definitions, as violence is one of the lesser components of hate speech compared to offensive keywords and insults directed at minorities, and violent language does not constitute offensive language (Parekh, 2012, p. 14; Wiegand et al.).

Given the elevated mentions of violence in gun rights X posts, one would also expect an overall reduced amount of hate speech and offensive language, but the exact opposite is true with the abortion issue having a much greater prevalence of hate speech under ChatGPT's labels. To further support this, an additional Mann–Whitney U test was performed between the full abortion and gun rights datasets: the sum of the mean ranks was 43381 for abortion, and 36819 for gun rights (W = 23281, p = 0.002). Given the higher sum of ranks, abortion tweets are labeled more negatively by ChatGPT on average. This is most likely due to the abortion rights debate being heavily nested in women's rights, a prominent civil rights issue that concerns a global minority (Shaw, 2010). On the other hand, the political issue of gun rights is largely separate from any minority, making it far easier to offend a minority when making abortion rights statements than gun rights statements.

The final anomaly of ChatGPT's labeling is its far more pessimistic hate speech detection when compared with the two human annotators, with all four of ChatGPT's data denominations having larger rank sums than the control data. With the research currently available, there is little to potentially explain this, although it is likely due to ChatGPT possessing a more lenient understanding of hate speech and offensive language. This could be due to OpenAI's attempts to restrict and train ChatGPT to prevent the spread of potentially malicious information (Farina & Lavazza, 2023). ChatGPT may have an increased vigilance to negative language as a result of these measures.

Implications

These findings demonstrate an explicit left-leaning bias in ChatGPT, showing that it is more likely to identify left-wing content with negative labels. This suggests that if ChatGPT were employed to moderate content on social media, its hate speech detection would be skewed in favor of the left, and result in left-leaning content having a greater reach on the service. However, the magnitude of this left political bias is shown to be somewhat minor for political issues, and especially so for gun rights, given the small differences in rank sums. This research more heavily corroborates the findings of Satoshi Fujimoto and Kazuhiro Takemoto than other works in the field, show-

ing that although ChatGPT has a political bias, it is less significant than early analyses have found (Fujimoto & Takemoto, 2023). This research also presents convincing evidence that human political bias tests are accurate when used on ChatGPT, as they were able to successfully predict modest left-biased hate speech detection in the model. Depending on the nature of other models and their similarity to ChatGPT, this evidence could also translate over to other GPT or chat-based LLMs.

Limitations

Asking ChatGPT to label each X post 12 times was not enough to fully prevent randomness, with other studies going to far greater lengths, such as the work of Shangbin Feng and his associates who have LLMs label each piece of content 100 times (Feng et al., 2023). Imitating this method would increase data collection time to the point that a model would be required for reading LLM answers, which is outside of the scope of this research. Additionally, ChatGPT is hosted exclusively on external servers, and limits requests to thirty per account per hour. To bypass this, either a copious number of alternate accounts would need to be created, a large sum of money forked over to take advantage of ChatGPT playground's cost-per-request basis, or an exclusive privilege established. Secondly, a more nuanced measure of hatefulness than the categorical range used in this research (neutral, offensive, or hateful) could increase the statistical significance of the findings. Such a measure has not been established in the research base. It would likely be outside this research's scope, requiring either a custom AI model for annotation or extensive training for both human annotators. Lastly, a wider range of political issues and a greater number of X posts could further strengthen the research statistically and cumulatively.

Areas for Future Research

This study acts as a framework for testing how the political biases of other LLMs affect hate speech detection. Studies on more prominent hate speech detection LLMs such as BART or Roberta are heavily needed in the research base, although they would likely be performed on popular pre-trained versions of them as their political biases are highly influenced by pretraining (Feng et al., 2023). Additionally, a study with a larger dataset and more political issues could further reinforce the findings of this study by limiting potential biases associated with each political issue.

References

An, M., & Carlson, J. (2022). Politics at the gun counter: Examining partisanship and masculinity among conservative gun sellers during the 2020 gun purchasing surge. *Social Problems*. https://doi.org/10.1093/socpro/spac046

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. B. F., Lee, J., Mann, M., Merhout, F., & Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, *115*(37), 9216–9221. https://doi.org/10.1073/pnas.1804840115

Barnes, A., & Ephross, P. H. (1994). The impact of hate violence on victims: Emotional and behavioral responses to attacks. *Social Work*, *39*(3). https://doi.org/10.1093/sw/39.3.247

Benier, K. (2017). The harms of hate: Comparing the neighbouring practices and interactions of hate crime victims, non-hate crime victims and non-victims. *International Review of Victimology*, *23*(2), 179–201. https://doi.org/10.1177/0269758017693087

Benson, J. (2023). Democracy and the epistemic problems of political polarization. *American Political Science Review*, 1–14. https://doi.org/10.1017/s0003055423001089

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020, July 1). *Language (Technology) is Power: A Critical Survey of "Bias" in NLP*. ACLWeb; Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.485

Carmines, E. G., Gerrity, J. C., & Wagner, M. W. (2010). How abortion became a partisan issue: Media coverage of the interest group-political party connection. *Politics & Policy*, *38*(6), 1135–1158. https://doi.org/10.1111/j.1747-1346.2010.00272.x

Castaño-Pulgarín, S. A., Suárez-Betancur, N., Vega, L. M. T., & López, H. M. H. (2021). Internet, social media and online hate speech. Systematic review. *Aggression and Violent Behavior*, *58*(101608), 101608. https://doi.org/10.1016/j.avb.2021.101608

Cinelli, M., Morales, G. D. F., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, *118*(9). https://doi.org/10.1073/pnas.2023301118

Cobb, M. D., & Kuklinski, J. H. (1997). Changing minds: Political arguments and political persuasion. *American Journal of Political Science*, *41*(1), 88. https://doi.org/10.2307/2111710

Das, M., Pandey, S. K., & Mukherjee, A. (2023). *Evaluating ChatGPT's Performance for Multilingual and Emoji-based Hate Speech Detection*. https://doi.org/10.48550/arxiv.2305.13276

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, *11*(1), 512–515. https://doi.org/10.1609/icwsm.v11i1.14955

De la Peña Sarracén, G. L., & Rosso, P. (2023). Systematic keyword and bias analyses in hate speech detection. *Information Processing & Management*, *60*(5), 103433. https://doi.org/10.1016/j.ipm.2023.103433

Dixon, L., Li, J., Sorensen, J., Thain, N., & Vasserman, L. (2018). Measuring and mitigating unintended bias in text classification. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 67–73. https://doi.org/10.1145/3278721.3278729

European Union. (2014, June 15). *Framework Decision on Combating Certain Forms and Expressions of Racism and Xenophobia by Means of Criminal Law*. Eur-Lex.europa.eu. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=LEGISSUM:l33178

Farina, M., & Lavazza, A. (2023). ChatGPT in society: Emerging issues. *Frontiers in Artificial Intelligence*, *6*. https://doi.org/10.3389/frai.2023.1130913

Feng, S., Chan Young Park, Liu, Y., & Yulia Tsvetkov. (2023). From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, *1*. https://doi.org/10.18653/v1/2023.acl-long.656

Fujimoto, S., & Takemoto, K. (2023). Revisiting the political biases of ChatGPT. *Frontiers in Artificial Intelligence*, *6*. https://doi.org/10.3389/frai.2023.1232003

Hadfi, R., Kawamura, N., Sakai, A., Yamaguchi, N., & Ito, T. (2020). A study on the polarisation effects of biased conversational agents in online debates. *The Japanese Society for Artificial Intelligence*, *JSAI2020*. https://doi.org/10.11517/pjsai.JSAI2020.0_2G6ES305

Henrickson, L., & Meroño-Peñuela, A. (2023). Prompting meaning: A hermeneutic approach to optimising prompt engineering with ChatGPT. *AI & Society*. https://doi.org/10.1007/s00146-023-01752-8

Hietanen, M., & Eddebo, J. (2022). Towards a definition of hate speech—With a focus on online contexts. *Journal of Communication Inquiry*, *47*(4), 019685992211243. https://doi.org/10.1177/01968599221124309

Howard, J. W. (2019). Free speech and hate speech. *Annual Review of Political Science*, *22*(1), 93–109. https://doi.org/10.1146/annurev-polisci-051517-012343

Huang, F., Kwak, H., & An, J. (2023). Is ChatGPT better than human annotators? Potential and limitations of ChatGPT in explaining implicit hate speech. *WWW '23 Companion: Companion Proceedings of the ACM Web Conference 2023*, 294–297. https://doi.org/10.1145/3543873.3587368

Jakesch, M., Bhat, A., Buschek, D., Zalmanson, L., & Naaman, M. (2023). Co-writing with opinionated language models affects users' views. *CHI '23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, *111*. https://doi.org/10.1145/3544548.3581196

Jay, T., & Janschewitz, K. (2008). The pragmatics of swearing. *Journal of Politeness Research. Language, Behaviour, Culture*, *4*(2). https://doi.org/10.1515/jplr.2008.013

Lester, J. C. (1996). The political compass (and why libertarianism is not right-wing). *Journal of Social Philosophy*, *27*(2), 176–186. https://doi.org/10.1111/j.1467-9833.1996.tb00245.x

Li, L., Fan, L., Shubham Atreja, & Hemphill, L. (2024). "HOT" ChatGPT: The promise of ChatGPT in detecting and discriminating hateful, offensive, and toxic comments on social media. *ACM Transactions on the Web*, *18*(2). https://doi.org/10.1145/3643829

Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z., Wu, Z., Zhu, D., Li, X., Niu, Q., Shen, D., Liu, T., & Ge, B. (2023). Summary of ChatGPT-related research and perspective towards the future of large language models. *Meta-Radiology*, *1*(2). https://doi.org/10.1016/j.metrad.2023.100017

MacFarland, T. W., & Yates, J. M. (2016). Mann–Whitney U test. *Introduction to Nonparametric Statistics for the Biological Sciences Using R*, 103–132. https://doi.org/10.1007/978-3-319-30634-6_4

Matamoros-Fernández, A., & Farkas, J. (2021). Racism, hate speech, and social media: A systematic review and critique. *Television & New Media*, *22*(2), 205–224. https://doi.org/10.1177/1527476420982230

Matoki, F., Neto, V. P., & Rodrigues, V. (2023). More human than human: Measuring ChatGPT political bias. *Public Choice*. https://doi.org/10.1007/s11127-023-01097-2

Meta. (n.d.). *Hate Speech*. Transparency.meta.com. https://transparency.meta.com/policies/community-standards/hate-speech/

Moss, J. T., & O'Connor, P. J. (2020). Political correctness and the Alt-right: The development of extreme political attitudes. *PLoS ONE*, *15*(10). https://doi.org/10.1371/journal.pone.0239259

Osborne, D., Huang, Y., Overall, N. C., Sutton, R. M., Petterson, A., Douglas, K. M., Davies, P. G., & Sibley, C. G. (2022). Abortion Attitudes: An Overview of Demographic and Ideological Differences. *Political Psychology*, *43*(1). https://doi.org/10.1111/pops.12803

Panchala, G. H., S Sasank, V. V., Harshitha Adidela, D. R., Yellamma, P., Ashesh, K., & Prasad, C. (2022). Hate speech & offensive language detection using ML &NLP. *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 1262–1268. https://doi.org/10.1109/icssit53264.2022.9716417

Parekh, B. (2012). Is there a case for banning hate speech? *The Content and Context of Hate Speech*, 37–56. https://doi.org/10.1017/cbo9781139042871.006

Park, J. H., Shin, J., & Fung, P. (2018, October 1). *Reducing Gender Bias in Abusive Language Detection*. ACLWeb; Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1302

Pearson-Merkowitz, S., & Dyck, J. J. (2017). Crime and partisanship: How party ID muddles reality, perception, and policy attitudes on crime and guns*. *Social Science Quarterly*, *98*(2), 443–454. https://doi.org/10.1111/ssqu.12417

Röttger, P., Hofmann, V., Pyatkin, V., Hinck, M., Kirk, H., Schütze, H., & Hovy, D. (2024). *Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models*. https://arxiv.org/pdf/2402.16786

Rutinowski, J., Franke, S., Endendyk, J., Dormuth, I., Roidl, M., & Pauly, M. (2024). The self-perception and political biases of ChatGPT. *Human Behavior and Emerging Technologies*, *2024*, e7115633. https://doi.org/10.1155/2024/7115633

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The risk of racial bias in hate speech detection. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. https://doi.org/10.18653/v1/p19-1163

Schmidtke, O. (2020). Politicizing social inequality: Competing narratives from the alternative for Germany and left-wing movement stand up. *Frontiers in Sociology*, *5*. https://doi.org/10.3389/fsoc.2020.00013

Shaw, D. (2010). Abortion and Human Rights. *Best Practice & Research Clinical Obstetrics & Gynaecology*, *24*(5), 633–646. https://doi.org/10.1016/j.bpobgyn.2010.02.009

Testa, C. (2012). Is polarization bad? *European Economic Review*, *56*(6), 1104–1118. https://doi.org/10.1016/j.euroecorev.2012.04.005

Tommi Gröndahl, Luca Pajola, Juuti, M., Conti, M., & Asokan, N. (2018). All you need is "love." *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*. https://doi.org/10.1145/3270101.3270103

United Nations. (2023). *What Is Hate Speech?* United Nations. https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech

Vasist, P. N., Chatterjee, D., & Krishnan, S. (2023). The polarizing impact of political disinformation and hate speech: A cross-country configural narrative. *Information Systems Frontiers*, 1–26. https://doi.org/10.1007/s10796-023-10390-w

Wachs, S., Gámez-Guadix, M., & Wright, M. F. (2022). Online hate speech victimization and depressive symptoms among adolescents: The protective role of resilience. *Cyberpsychology, Behavior, and Social Networking*, *25*(7). https://doi.org/10.1089/cyber.2022.0009

Waseem, Z., Davidson, T., Warmsley, D., & Weber, I. (2017, August 1). *Understanding Abuse: A Typology of Abusive Language Detection Subtasks*. ACLWeb; Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-3012

Wich, M., Bauer, J., & Groh, G. (2020). Impact of politically biased data on hate speech classification. *Proceedings of the Fourth Workshop on Online Abuse and Harms*, 54–64. https://doi.org/10.18653/v1/2020.alw-1.7

Wiegand, M., Ruppenhofer, J., Anna Marie Schmidt, & Greenberg, C. (2018). Inducing a lexicon of abusive words – a feature-based approach. *Publication Server of the Institute for German Language (Institute for German Language)*, 1046–1056. https://doi.org/10.18653/v1/n18-1095

Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q.-L., & Tang, Y. (2023). A brief overview of ChatGPT: YThe history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, *10*(5), 1122–1136. https://doi.org/10.1109/jas.2023.123618

X. (2023, April). *Hateful conduct policy*. Twitter.com; Twitter Help Center. https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

Yin, W., & Zubiaga, A. (2021). Towards generalisable hate speech detection: A review on obstacles and solutions. *PeerJ Computer Science*, *7*, e598. https://doi.org/10.7717/peerj-cs.598

Yousafzai, A. W. (2022). political polarization and its impact on mental health: Where do we stand? *Khyber Medical University Journal*, *14*(1), 1–2. https://doi.org/10.35845/kmuj.2022.22777