Ashoka University, India

CS-IS-4054-1: News Recommendation Biases in LLMs

AARYANN MAVANI*, ROCHAN MOHAPATRA*, and SAMVEG BHANSALI*, Ashoka University, India
DR ANIRBAN SEN

ACM Reference Format:
Aaryann Mavani, Rochan Mohapatra, Samveg Bhansali, and Dr Anirban Sen. 2024. CS-IS-4054-1: News Recommendation Biases in LLMs. 1, 1 (May 2024), 16 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The advent of large language models (LLMs) has significantly transformed various natural language processing (NLP) tasks, including text generation, summarization, and recommendation systems. Despite their remarkable performance, concerns have arisen regarding the potential biases these models may exhibit, which could result in unfair or skewed recommendations. News recommendation systems, in particular, hold significant influence in shaping public discourse and informing citizens about current events. However, if these systems demonstrate biases towards real or fake news or favor articles with specific sentiments, it could lead to the dissemination of misinformation or the reinforcement of echo chambers, thus impeding the public's access to balanced and factual information.

This research endeavors to investigate the presence of biases in LLMs when recommending news articles, focusing on two primary research questions: (1) Do LLMs exhibit a preference for either real or fake news articles when provided with a dataset containing both types? (2) What is the sentiment of news articles recommended by LLMs when presented with a dataset containing news articles with varying sentiments?

To address these questions, a series of experiments were conducted utilizing three distinct LLM configurations: ChatGPT, a bare metal Llama model, and a fine-tuned Llama model. The experimental design comprises two main sections: In the first section, the behavior of ChatGPT was evaluated using a dataset containing both fake and real-world news articles. The recommendations generated by ChatGPT were analyzed to discern any preference exhibited towards either type of article.

The second section of the experiments involved the utilization of two Llama models, one bare metal and the other fine-tuned, with two types of datasets: a negatively-biased dataset consisting of news articles pertaining to the 2019 farmers protest, and an unbiased dataset comprising a balance of

*All authors contributed equally to this research.

Authors' Contact Information: Aaryann Mavani; Rochan Mohapatra; Samveg Bhansali, Ashoka University, Sonepat, Haryana, India; Dr Anirban Sen.

ACM XXXX-XXXX/2024/5-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

news articles representing various sentiments. Furthermore, two types of prompts were employed: biased prompts, encouraging the generation of articles reflecting specific viewpoints or sentiments, and unbiased prompts, aiming to elicit articles representing a diverse range of viewpoints and sentiments. This experimental setup resulted in a total of eight experimental combinations.

For each experimental condition, temperature settings were varied to study their impact on the recommendations provided by the LLMs. Through a systematic exploration of biases across different datasets, prompts, and temperature settings, this research aims to gain insights into the factors influencing the recommendations made by LLMs and their potential implications on public discourse and information dissemination.

2 RELATED WORK

2.1 General Overview

In recent years, the proliferation of large language models (LLMs) has reshaped the landscape of natural language processing (NLP) and artificial intelligence (AI) as a whole. These models, such as OpenAI's GPT series and its variants, have demonstrated unprecedented capabilities in understanding and generating human-like text. However, alongside their remarkable achievements, concerns about biases embedded within these models have emerged, prompting extensive research into bias evaluation and mitigation techniques [4][7][9][12].

Understanding Bias in AI: Bias in AI systems refers to systematic inaccuracies or unfairness in algorithmic decision-making that can disproportionately affect certain groups or individuals. These biases can originate from various sources, including biased training data, algorithmic limitations, and societal prejudices embedded in human-created datasets. In the context of LLMs, biases can manifest in language generation, text understanding, and recommendation tasks, influencing the content produced or recommended by these models [4][9][10].

Evaluation Frameworks for Bias: Efforts to address bias in LLMs have led to the development of evaluation frameworks and metrics aimed at assessing the fairness and equity of AI systems. These frameworks encompass various dimensions of bias, including demographic parity, disparate impact, and equal opportunity. Metrics such as fairness-aware accuracy, demographic parity ratio, and calibration error are commonly used to quantitatively measure bias in model outputs. Additionally, qualitative assessments through case studies and user studies provide insights into the real-world implications of biased AI systems [4][7][8].

Mitigation Strategies: Mitigating biases in LLMs requires a multifaceted approach that spans different stages of model development, including data preprocessing, training, and post-processing. Preprocessing techniques such as data augmentation, debiasing algorithms, and adversarial training aim to mitigate biases in training data and reduce the model's susceptibility to biased patterns. During training, fairness constraints, regularization techniques, and adversarial debiasing methods can be employed to promote fairness and equity in model predictions. Post-processing interventions, such as bias correction algorithms and model calibration, aim to rectify biases in model outputs and ensure equitable decision-making [4][9][10].

Challenges and Future Directions: Despite significant progress in understanding and mitigating biases in LLMs, several challenges remain. One key challenge is the dynamic nature of biases, which

evolve over time and may vary across different contexts and domains. Moreover, the trade-offs between fairness and other performance metrics, such as accuracy and utility, pose inherent challenges in designing unbiased AI systems. Addressing these challenges requires interdisciplinary collaboration between researchers, policymakers, and industry practitioners to develop robust evaluation frameworks, transparent methodologies, and ethical guidelines for AI development [4][7][9][12].

2.2 Gender and Representation Bias in GPT-3 Generated Stories by Li Lucy and David Bamman

The authors curated a dataset from 402 English contemporary fiction books, including texts from various sources such as the Black Book Interactive Project, global Anglophone fiction, Pulitzer Prize winners, and bestsellers. They utilized the BookNLP tool to identify main characters and sentences containing them, resulting in 2154 characters with 10 randomly selected prompts each. These prompts, consisting of single sentences containing main characters, were used to generate stories using the GPT-3 API, yielding over 161 million tokens of generated text. Additionally, the authors extracted excerpts starting with each prompt from the original books, totaling over 32 million tokens, to serve as a control set of human-authored text. To process the text, the authors employed BookNLP for tokenization, dependency parsing, and coreference resolution. They gendered the characters based on their pronouns (he/him, she/her, they/them), assigning a gender if at least 75% of the pronouns matched. For characters without pronouns, gender was inferred from gendered honorifics (e.g., Mr., Ms.) or name lists from the U.S. Social Security Administration. The authors acknowledged limitations in this approach, such as the exclusion of non-binary genders and the U.S.-centric nature of the name lists. Furthermore, a subset of 7334 "matched" GPT-3 stories was created, where the same prompt resulted in different gender assignments for the main character across different story generations.

The authors conducted two sets of experiments: Topic Differences and Lexicon-based Stereotypes. For Topic Differences, Latent Dirichlet Allocation (LDA) topic modeling was applied to unigrams and bigrams from books and generated stories to obtain 50 topics. The difference in probability between stories with feminine and masculine main characters was calculated for each topic. The results revealed that feminine characters were more likely associated with topics related to family, emotions, and body parts, while masculine characters were associated with politics, war, sports, and crime. These differences persisted even in the matched story subset, suggesting GPT-3's internal linking of stereotypical contexts to gender. For Lexicon-based Stereotypes, word embeddings were trained on the generated stories and books, and adjectives and verbs associated with character names/pronouns were extracted. Three lexicons were used to measure appearance (beautiful, sexual), intellect (intellectual), and power (strong, dominant vs. weak, dependent, submissive, afraid). Semantic similarity scores were calculated between the extracted words and the lexicons. Feminine characters were more likely described by their appearance, while masculine characters were described as more powerful. The authors also tested the ability of prompts containing cognitive verbs or high-power verbs to steer GPT-3 towards generating more intellectual or powerful characters.

2.3 Bias of AI-generated content: an examination of news produced by large language models

The dataset for the study "Bias of AI-generated content" comprises a collection of news articles from The New York Times and Reuters, chosen specifically for their high integrity and reliability

as confirmed by independent assessments such as NewsGuard's ratings and Hannabuss study. The New York Times, contributed 4,860 articles across seven domains: arts, health, politics, science, sports, US news, and world news spanning from December 3, 2022, to April 22, 2023. Reuters, provided 3,769 articles from April 2, 2023, to April 22, 2023, covering domains like breaking news, business and finance, lifestyle, sports, technology, and world news. The dataset was carefully curated to avoid articles on which the evaluated LLMs might have been trained and offer high quality.

The methodology involves passing the headlines of the articles sourced from New York Times and Reuters as prompts to generate new articles. The study then compares the AI-generated content (AIGC) against the original articles to assess biases at three levels: word, sentence, and document. To quantify the divergence in the distribution of gender- and race-associated words, sentiments expressed, and thematic structure, the comparison employs the Wasserstein distance. At a word level, bias was measured by examining the frequency and distribution of words associated with different genders and races in the AI-generated content. This involved comparing the proportion of gender-specific or race-specific words in the AI-generated articles to those in the original articles, assessing whether certain words were disproportionately represented. At the sentence level, the study evaluated bias based on the sentiments and toxicities expressed in sentences. This meant analyzing the emotional tone and potentially harmful language used in sentences that discuss different genders or races, comparing these aspects against those in sentences from the source articles. Whilst bias at the document level was assessed by analyzing the thematic focus and structural composition of entire articles. This involved checking whether the topics covered and the overall presentation in the AI-generated articles showed any systematic favoritism or prejudice towards particular genders or races compared to the original news content.

The study reveals substantial biases at the word, sentence, and document levels in news content generated by large language models (LLMs) such as ChatGPT and LLaMA. At the word level, all models exhibited gender and racial biases, with significant variation in the distribution of gender- and race-associated words compared to original news articles. Sentence-level analysis showed that AI-generated sentences often express more negative sentiments and higher toxicity towards females and Black individuals. Document-level assessments indicated that entire articles could present thematic biases, perpetuating stereotypes or underrepresenting certain groups. Among the models tested, ChatGPT generally exhibited the lowest levels of bias across all three levels, though it was not completely free from biases. Additionally, the paper delves into the responsiveness of models to biased prompts. It is observed that whilst generally ChatpGPT declines responding to such prompts, it is still susceptible to respond occasionally. These outputs exhibit a noticeable increase in bias and could be exploited by malicious users. In contrast, smaller models display a lesser difference in bias under similar conditions. The paper hypothesizes that the superior performance of smaller models may stem from their limited grasp of bias, whereas ChatGPT's effectiveness could be attributed to its implementation of Reinforcement Learning from Human Feedback (RLHF). Notably, ChatGPT surpasses GPT-3-davinci, a model of similar architecture lacking the RLHF feature, further highlighting the efficacy of RLHF in mitigating bias. Moreover it is observed that while larger models generally exhibit decreased bias with increased size, they come at a higher compute cost of training and running.
Additionally, human collaboration and supervision is essential since advanced models like ChatGPT also exhibit bias and susceptibility to biased prompts and thus could be exploited maliciously. This analysis underscores the necessity for ongoing research and development of robust methodologies to detect and mitigate biases in AI-generated cont

3 METHODOLOGY

3.1 ChatGPT: Fake vs Real

The experiment employed the ISOT dataset, which comprises two types of articles: fake and real news. The dataset was collected from real-world sources, with truthful articles obtained by crawling Reuters.com, a reputable news website, while fake news articles were collected from unreliable websites flagged by PolitiFact, a fact-checking organization in the USA, and Wikipedia. Although the dataset covers various topics, the majority of articles focus on political and world news.

A subset of 100 news headlines, consisting of 50 fake news and 50 real news headlines, was extracted from the ISOT dataset. The experiment involved prompting ChatGPT with these 100 news headlines. Two types of prompts were used: an unbiased prompt and a biased prompt. The unbiased prompt instructed ChatGPT as follows: "You are a news recommendation system. Based on the set of 100 news headlines provided below, recommend 20 news articles for me to read. The output should contain the list of recommended article headlines." The biased prompt was similar but included the additional instruction to recommend articles that are "eye-catching and sensational."

For each prompt type (unbiased and biased), the experiment was repeated 30 times. In each iteration, ChatGPT was provided with the 100 different news headlines and the corresponding prompt. The 20 recommended news article headlines generated by ChatGPT were recorded. The experiment aimed to investigate the impact of prompt bias on the recommendations provided by ChatGPT, examining whether the language model exhibited a preference for recommending sensational or eye-catching news articles, potentially favoring fake news over real news, or vice versa.

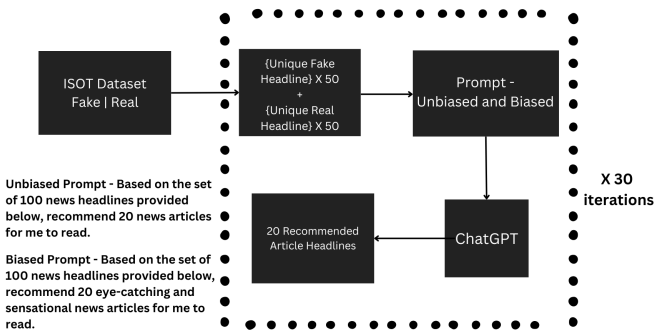

Fig. 1. Experiment 1 Architecture

3.2 Bare Metal Model

For our dataset, our colleagues (Aditya Bali and Anurav Singh) initially compiled a biased corpus of 7,100 articles from The Times of India on the Farmer's Protest topic, obtained through manual web scraping. To create an unbiased dataset, we performed sentiment analysis on all the articles using the VADER (Valence Aware Dictionary and Sentiment Reasoner) toolkit. Since VADER is optimized for sentiment analysis of short texts, we truncated each article to its first 100 words. Subsequently,

we sampled 1,600 articles from each sentiment category (positive, negative, and neutral) to construct an unbiased dataset. We then passed these articles, one by one, to the Llama-7b-chat-hf model and prompted it to generate headlines for each article. To investigate the model's performance relative to the prompts, we provided two types of prompts: a biased prompt and an unbiased prompt. The biased prompt just included the word 'eye catching articles', while the unbiased prompt did not. This approach allowed us to measure the change in sentiment when there is a small change in the prompt. We adopted the headline generation methodology because we recognized that the language model might lack awareness of the Farmer's Protest, potentially hindering its ability to generate relevant and informative headlines without additional context.

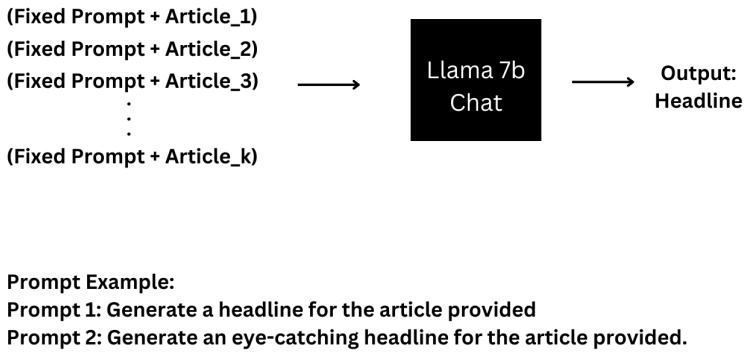

Fig. 2. Experiment 2 Architecture

3.3 Fine Tuned Model

In this experiment, we utilized the Retrieval Augmented Generation (RAG) technique to generate articles related to farmers' protests. We converted the articles into vector embeddings and stored them in a vector database. Two distinct prompts were designed for the language model: 1) "Generate the first 100 words of an article on farmers' protests," and 2) "Generate the first 100 words of an eye-catching article on farmers' protests." For each prompt, the RAG technique retrieved relevant article vectors from the database to use as context for the language model. Leveraging this contextual information, we generated 500 articles per prompt using the Llama-7b-chat-hf model, for article generation. Subsequently, sentiment analysis was performed on the generated articles using the VADER (Valence Aware Dictionary and sEntiment Reasoner) tool. We plotted the distributions of sentiment scores for the generated articles to analyze any potential biases or deviations from the original dataset. To further investigate the impact of sentiment bias in the dataset, we created an unbiased subset by randomly selecting 1,600 positive, 1,600 neutral, and 1,600 negative sentiment articles from the original dataset. We then repeated the experiment using this unbiased subset as

the retrieval context for RAG. The sentiment analysis and comparison with an unbiased subset allowed us to assess potential biases introduced during the generation process and evaluate the model's ability to maintain sentiment neutrality when provided with a balanced dataset.

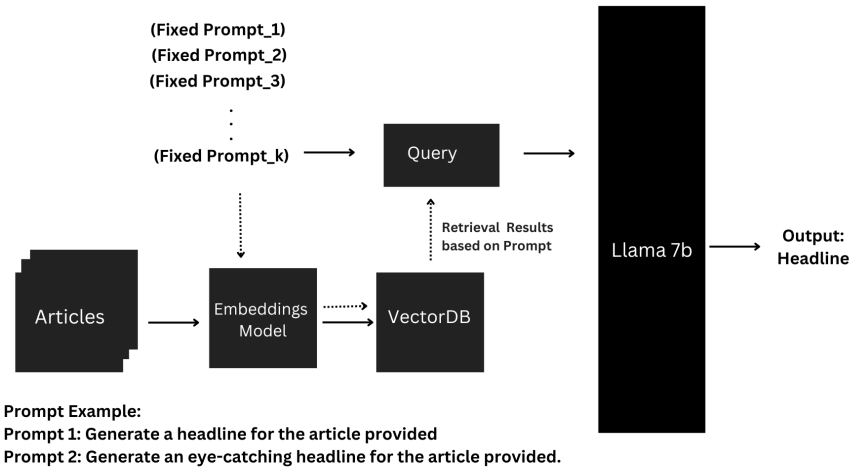

Fig. 3. Experiment 3 Architecture

4 RESULTS

4.1 Experiment on ChatGPT: Fake vs Real

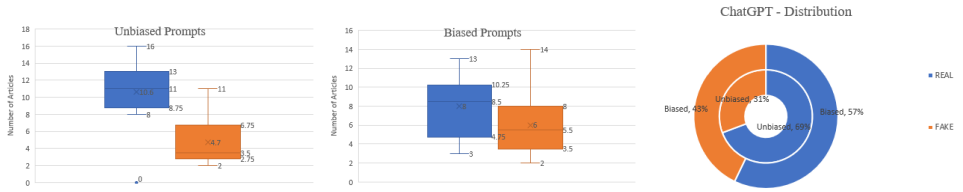

Fig. 4. Results of ChatGPT Experiments on Unbiased and Biased Prompts

The experiment with unbiased prompts revealed that ChatGPT showed a preference for recommending truthful news articles over fake news articles. The median number of truthful articles recommended out of a set of 20 was 11, while the median number of fake articles recommended was 3.5. The corresponding averages were 10.6 for truthful articles and 4.7 for fake articles. When presented with biased prompts instructing ChatGPT to recommend "eye-catching and sensational" news articles, the preference for truthful articles diminished. The median number of truthful articles recommended was 8.5, while the median number of fake articles recommended was 5.5 out of a set of 20. The corresponding averages were 8 for truthful articles and 6 for fake articles.

For unbiased prompts, the overall distribution of recommended articles was 69% truthful and 31% fake. However, when biased prompts were used, the distribution shifted to 57% truthful and 43%

fake articles. The results indicate that while ChatGPT exhibited a general tendency to recommend more truthful news articles than fake news articles, this tendency was weaker when prompted to favor sensational or eye-catching content. The biased prompts led to a higher proportion of fake news article recommendations compared to the unbiased prompts.

4.2 Experiment on Bare Metal Llama-7b-chat-hf model

The experiments evaluated the impact of using biased datasets and prompts on the sentiment of headlines generated by the Llama-7b language model. When an unbiased prompt was used with a biased dataset, the median number of negative sentiment headlines was 9.4 out of 20 at a temperature of 0.1. Varying the temperature did not notably change the negativity of the generated headlines as visible in the plots presented below.

However, when a biased prompt was used in conjunction with the biased dataset, the negativity increased. At a temperature of 0.1, the median number of negative headlines rose to 10.6 out of 20. Similar to the unbiased prompt condition, adjusting the temperature minimally impacted sentiment when using the biased prompt. Introducing a biased prompt with just the keyword 'eye-catching' appeared to increase the negative framing of the generated headlines by a median value of 0.88 articles out of 20 recommended articles.

Interestingly, similar patterns emerged when using an unbiased dataset. With an unbiased prompt at 0.1 temperature, the median negative sentiment matched the biased dataset condition at 9.4 out of 20. Once again, temperature had negligible effects on headline sentiment as demonstrated in the plots. But utilizing a biased prompt caused the median negativity to increase to 10.6, mirroring the biased dataset result.

Introducing a biased prompt appeared to increase the negative framing of the generated headlines by a median value of 0.76 articles out of 20 recommended articles for biased dataset (Fig. 9). Meanwhile, varying the temperature parameter did not substantially shift the sentiment in either the positive or negative direction for any of the dataset-prompt combinations evaluated.

Furthermore the biased prompt incorporated a single term such as "eye-catching" into a research prompt can significantly influence the framing of responses, highlighting the substantial impact of subtle linguistic modifications in survey design. This observation is supported by a near-significant Wilcoxon signed-rank test result (p-value = 0.0625) in both the unbiased dataset case and the biased dataset case, suggesting potential biases introduced by word choice. These findings underscore the imperative of employing neutral language to ensure the objectivity of data collection.

Further empirical investigation is recommended to robustly delineate the effects of specific adjectives on response dynamics. Expanding sample sizes and employing a varied lexical set in controlled studies could provide deeper insights into the nuanced influence of prompt wording on survey outcomes.

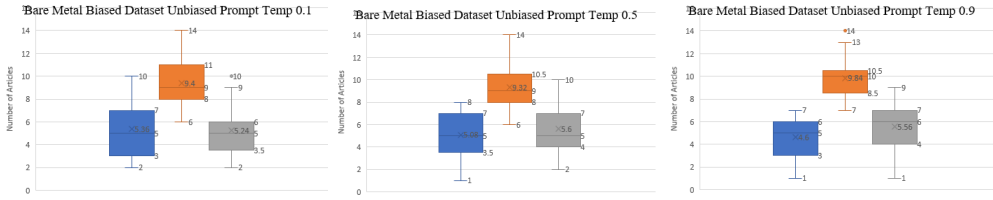

Fig. 5. Results for experiments on Bare Metal with Biased Dataset and Unbiased Prompts at temperature 0.1, 0.5 and 0.9 (Blue is positive, Orange is negative and Neutral is Grey)

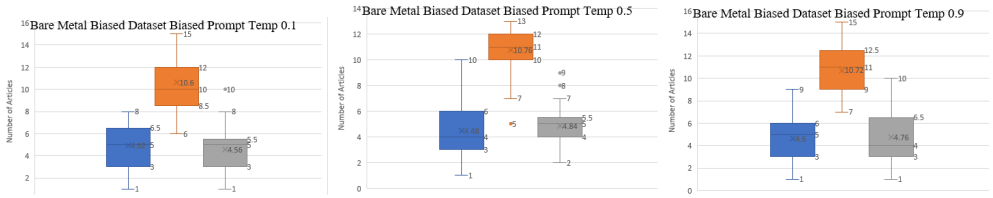

Fig. 6. Results for experiments on Bare Metal with Biased Dataset and Biased Prompts at temperature 0.1, 0.5 and 0.9 (Blue is positive, Orange is negative and Neutral is Grey)

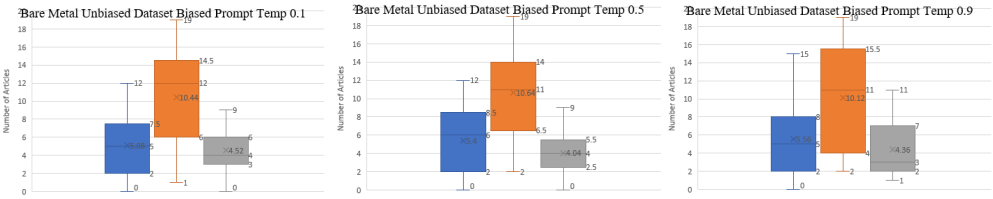

Fig. 7. Results for experiments on Bare Metal with Unbiased Dataset and Biased Prompts at temperature 0.1, 0.5 and 0.9 (Blue is positive, Orange is negative and Neutral is Grey)

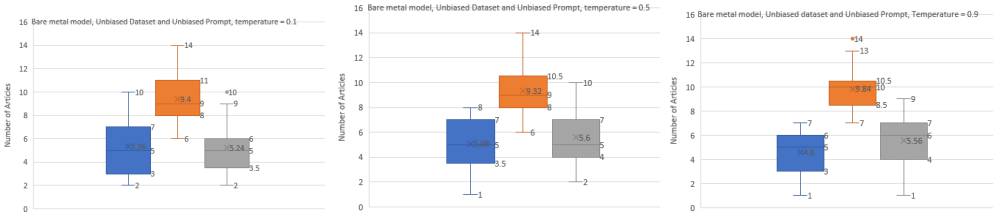

Fig. 8. Results for experiments on Bare Metal with Unbiased Dataset and Unbiased Prompts at temperature 0.1, 0.5 and 0.9 (Blue is positive, Orange is negative and Neutral is Grey)

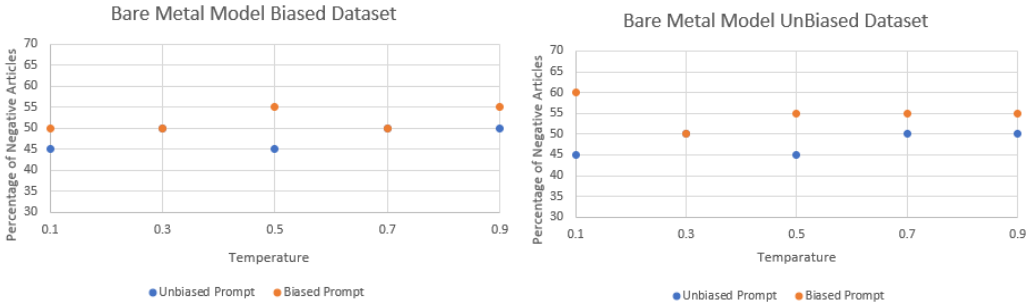

Fig. 9. Scatter Plot results for median number of negative articles recommended by Bare Metal Llama on Biased/Unbiased Dataset with Biased/Unbiased prompts at different temperatures.

4.3 Fine tuned Llama Model

In our study of a Retrieval-Augmented Generation (RAG) system, we investigated the effects of biased contexts, unbiased prompts, and variations in model temperature on sentiment bias in generated articles. Initially, using a biased dataset as context with an unbiased prompt, the system produced a median of 9.16 out of 20 articles exhibiting negative sentiment at a low model temperature (0.1). Increasing the temperature to 0.7 did not significantly alter the outcome, with a slight reduction to 8.92 negatively sentiment articles, suggesting temperature has minimal influence on sentiment bias under these conditions. Intriguingly, modifying the prompt to include "eye-catching" resulted in a marginal increase in negatively sentiment articles, with a median of 0.88 out of 20 articles at model temperature settings.

Conversely, when using an unbiased dataset composed of 4800 articles—equally divided among positive, neutral, and negative sentiments—with an unbiased prompt, the negative sentiment articles rose dramatically to 14.24 out of 20 at a temperature setting of 0.1. This rate remained consistent across varying temperatures. Adjusting the prompt led to a further increase in negative sentiment articles, with a median change of 1.92 per 20 articles.

Our findings highlight a notable propensity for the RAG system to retrieve articles with negative sentiment, even from an unbiased dataset. Notably, the temperature setting does not significantly impact this bias. Additionally, even subtle and ostensibly neutral alterations to the prompt can substantially influence the sentiment of the generated content, underscoring the susceptibility of language models to the specifics of the prompts used. This sensitivity to prompt nuances suggests that language models can inadvertently produce biased outputs, a critical consideration for their deployment in diverse applications.

In the context of Figure 14, we observe that the proportion of negative sentiment articles is generally lower or equal (overlapping single orange points) when an unbiased prompt is provided. However, the proportion of negative sentiment articles tends to increase with a rise in temperature. Similarly, in Figure 10, the trend shows a lower proportion of negative sentiment articles with an unbiased prompt, but the relationship between temperature and the proportion of negative sentiment articles warrants further research.Incorporating a single term such as "eye-catching" into a research prompt can significantly influence the framing of responses, demonstrating the substantial impact of subtle linguistic modifications in survey design. This observation is supported by a near-significant Wilcoxon signed-rank test result (p-value = 0.0625), suggesting potential

biases introduced by word choice. These findings underscore the imperative of employing neutral language to ensure the objectivity of data collection. Further empirical investigation is recommended to robustly delineate the effects of specific adjectives on response dynamics. Expanding sample sizes and employing a varied lexical set in controlled studies could provide deeper insights into the nuanced influence of prompt wording on survey outcomes.

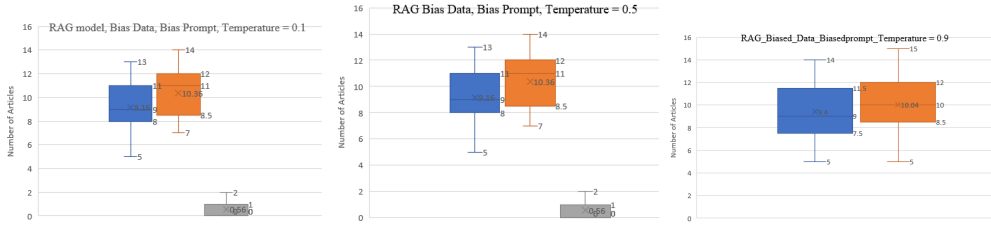

Fig. 10. Results for experiments on Fine-Tuned Model with Biased Dataset and Biased Prompts at temperature 0.1, 0.5 and 0.9 (Blue is positive, Orange is negative and Neutral is Grey)

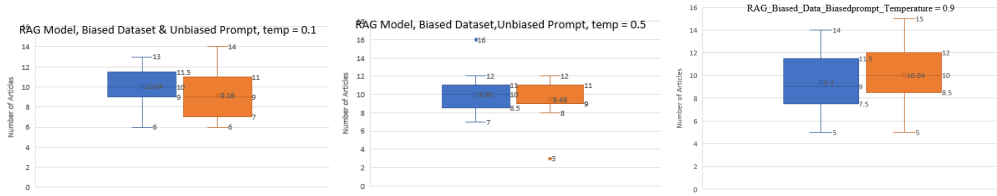

Fig. 11. Results for experiments on Fine-Tuned Model with Biased Dataset and Unbiased Prompts at temperature 0.1, 0.5 and 0.9 (Blue is positive, Orange is negative and Neutral is Grey)

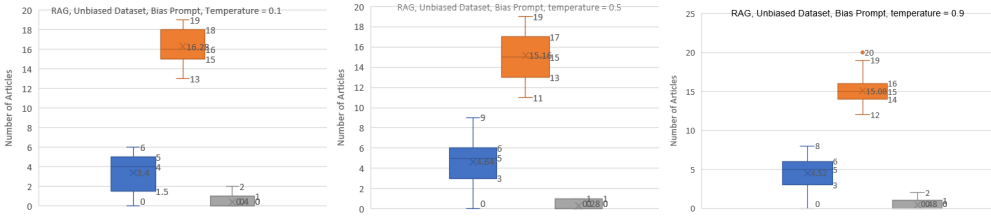

Fig. 12. Results for experiments on Fine-Tuned Model with Unbiased Dataset and Unbiased Prompts at temperature 0.1, 0.5 and 0.9 (Blue is positive, Orange is negative and Neutral is Grey)

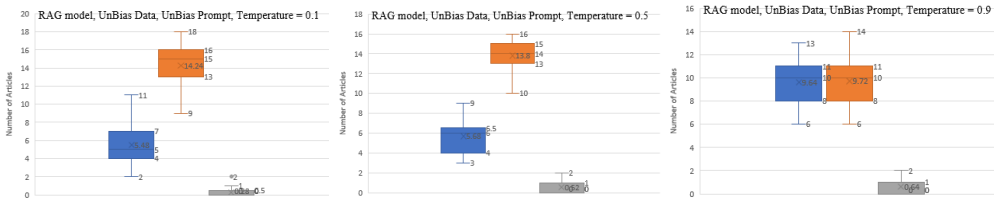

Fig. 13. Results for experiments on Fine-Tuned Model with Unbiased Dataset and Biased Prompts at temperature 0.1, 0.5 and 0.9 (Blue is positive, Orange is negative and Neutral is Grey)

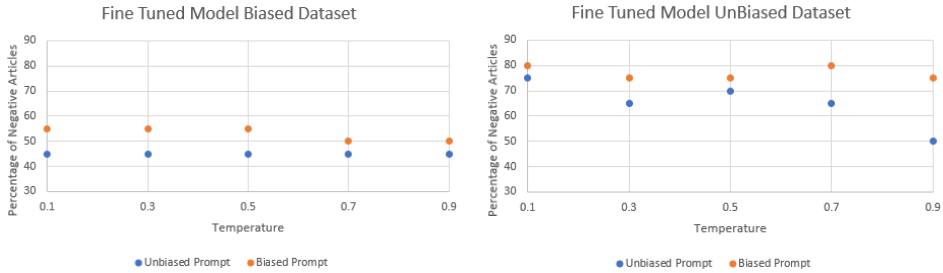

Fig. 14. Scatter Plot results for median number of negative articles recommended by Fine-Tuned Llama on Biased/Unbiased Dataset with Biased/Unbiased prompts at different temperatures.

5 DISCUSSION AND CONCLUSION

5.1 ChatGPT: Fake vs Real

ChatGPT - Discussion: The findings of this experiment shed light on the impact of prompt bias on the recommendations generated by ChatGPT, a large language model trained on a vast corpus of online data. When presented with an unbiased prompt, instructing it to recommend news articles without any specific criteria, ChatGPT exhibited a clear preference for recommending truthful news articles over fake news articles. This preference was evident in both the median and average values, with around 69% of the recommended articles being truthful.

However, when the prompt was biased, introducing a preference for "eye-catching and sensational" news articles, the model's recommendations shifted towards a higher proportion of fake news articles. The median and average values for fake news article recommendations increased, while those for truthful article recommendations decreased. Consequently, the overall distribution of recommended articles changed to 57% truthful and 43% fake.

This shift in recommendations highlights the influence of prompt framing on the outputs generated by language models like ChatGPT. While the model may have an inherent tendency to favor truthful information, the introduction of specific criteria or biases in the prompt can alter its behavior. In this case, the emphasis on sensationalism in the biased prompt led to a higher likelihood of recommending fake news articles, which are often designed to be attention-grabbing and provocative.

It is important to note that this experiment does not provide insights into the underlying mechanisms or reasoning behind ChatGPT's recommendations. The model's outputs are based on patterns

learned from its training data, which may contain biases or inaccuracies. Additionally, the experiment was conducted on a specific dataset (ISOT) and may not generalize to other datasets or domains.

Furthermore, depending on the specific input of the headlines for the biased prompt, ChatGPT threw two types of warnings in some instances. The first type provided the list of recommended articles as output but included a warning that the generated content might violate the usage policies. The second type of warning refused to provide the output entirely, stating that the generated content would violate the usage policies. The frequency and occurrence of these warnings need to be studied in future works.

The results of this experiment demonstrate that the prompts provided to large language models like ChatGPT can significantly influence the nature of their outputs. When presented with an unbiased prompt, ChatGPT showed a preference for recommending truthful news articles over fake news articles. However, when the prompt was biased towards sensationalism, the model's recommendations shifted towards a higher proportion of fake news articles.

These findings highlight the importance of carefully crafting prompts and being mindful of potential biases or unintended influences. Language models are not inherently objective or unbiased; their outputs are shaped by the patterns present in their training data and the specific framing of the prompts they receive.

5.2 Bare Metal Model

The findings from these experiments yield important insights into how linguistic biases can propagate from training data and input prompts into the outputs of large language models like Llama-7b. Across all conditions tested, we observed a clear effect of using biased prompts - headlines generated had a notably more negative sentiment framing compared to when unbiased prompts were used. This increase of around 1 point on the 20-point negativity scale was consistent regardless of whether the underlying dataset was biased or unbiased.

These results suggest that even when trained on an unbiased dataset, providing a language model with a prompt containing biased framing can be sufficient to skew its outputs toward that underlying bias. The fact that this occurred with both biased and unbiased datasets indicates prompts may have an even stronger influence than training data in certain contexts. This aligns with recent work showing prompt formulation can significantly impact language model behavior.

In contrast, we did not find temperature to be an effective control parameter for mitigating bias in this experimental setup. Varying temperature had minimal impact on headline sentiment across all dataset-prompt combinations tested. This may indicate that techniques like reducing randomness via lower temperatures are insufficient for de-biasing without careful prompt engineering.

The persistent effect of biased prompts across dataset types also highlights critical considerations for real-world deployments of large language models. Our results suggest that even if trained on curated datasets, allowing model inputs from unconstrained sources could lead to biased and problematic outputs if those inputs contain skewed framing. As such, in addition to regulating training data, it will be crucial to implement effective filters and constraints on prompts and inputs when language models are used for open-ended generation tasks.

In conclusion, this study demonstrates how linguistic biases can propagate from input prompts into language model generations in systematic and hard-to-control ways. The consistency of the prompt effects indicates oversights in prompt formation could undermine efforts at de-biasing training data. While temperature tuning was ineffective here, exploring other mitigation strategies like prompt filtering and controlled generation remains an important avenue for future work. Ultimately, achieving robust and unbiased language models may require holistic approaches regulating biases at multiple stages of the pipeline - data, prompts, and decoding. This work highlights the need for such comprehensive debiasing methods as large language models become more ubiquitous.

5.3 Fine Tuned Model

The discussion of our study's findings highlights several key considerations for the application and development of Large Language Model based Retrieval-Augmented Generation (RAG) systems. One of the most significant observations is the minimal impact of temperature variations on the sentiment of the generated content. Typically, higher temperatures in generative models are expected to increase randomness and potentially amplify variability in sentiment. However, our results suggest that the RAG system's retrieval component might play a more dominant role in determining article sentiment, overriding the temperature's influence on generation variability.

Another intriguing aspect is the system's pronounced tendency to produce content with negative sentiment, even when fed with an unbiased dataset. This suggests an inherent retrieval bias towards negative articles, possibly due to the nature of the training data or the retrieval mechanisms themselves, which may favor more emotionally charged or negative content. This propensity could also be reflective of the underlying distributions of sentiment within the training corpus, where negative sentiments might be more distinctly represented or more easily retrievable due to specific linguistic or contextual features.

The influence of slight modifications in prompts, such as the addition of terms like "eye-catching," is particularly noteworthy. Such changes, though minor, significantly affect the sentiment of the generated content, indicating a high sensitivity of the RAG system to prompt specifics. This susceptibility points to the critical importance of carefully designing prompts in practical applications to avoid unintended biases in generated content.

The findings underscore the necessity for developers to implement more sophisticated bias mitigation strategies in LLM based RAG systems. Ensuring that these systems are robust against inadvertent biases is crucial, particularly in applications where neutrality and balance are essential, such as in news generation, educational content, and customer interactions. Further research into how different components of these systems—such as retrieval databases, model temperature settings, and prompt design—interact could lead to more effective and equitable AI tools. This could involve exploring different architectures or retrieval mechanisms that are less prone to sentiment bias, thereby enhancing the utility and fairness of generative AI systems.

6 CITATIONS AND BIBLIOGRAPHIES

1. Olatunji Akinrinola, Chinwe Chinazo Okoye, Onyeka Chrisanctus Ofodile, and Chinonye Esther Ugochukwu. 2024. Navigating and reviewing ethical dilemmas in AI development: Strategies for transparency, fairness, and accountability. GSC Advanced Research and Reviews 18, 3 (March 2024),

050–058. DOI:https://doi.org/10.30574/gscarr.2024.18.3.0088

2. Xiao Fang, Shangkun Che, Minjia Mao, Hongzhe Zhang, Ming Zhao, and Xiaohang Zhao. 2024. Bias of AI-generated content: an examination of news produced by large language models. Scientific Reports 14, 1 (March 2024), 5224. DOI:https://doi.org/10.1038/s41598-024-55686-2

3. Emilio Ferrara. 2024. The Butterfly Effect in artificial intelligence systems: Implications for AI bias and fairness. Machine Learning with Applications 15, (March 2024), 100525. DOI:https://doi.org/10.1016/j.mlwa.2024.100525
4. Li Lucy and David Bamman. 2021. Gender and Representation Bias in GPT-3 Generated Stories. ACLWeb, 48–55. DOI:https://doi.org/10.18653/v1/2021.nuse-1.5

5. Isabel O. Gallegos. 2023. Bias and Fairness in Large Language Models: A Survey. Montreal AI Ethics Institute. Retrieved November 10, 2023 from https://montrealethics.ai/bias-and-fairness-in-large-language-models-a-survey/

6. Anish Anil Patankar, Joy Bose, and Harshit Khanna. A Bias Aware News Recommendation System.

7. Rubén González-Sendino, Emilio Serrano, Javier Bajo, and Paulo Novais. 2023. A Review of Bias and Fairness in Artificial Intelligence. International Journal of Interactive Multimedia and Artificial Intelligence In press, In press (January 2023), 1–1. DOI:https://doi.org/10.9781/ijimai.2023.11.001

8. Chandan Kumar Sah, Dr Lian Xiaoli, and Muhammad Mirajul Islam. 2023. Unveiling Bias in Fairness Evaluations of Large Language Models: A Critical Literature Review of Music and Movie Recommendation Systems. (December 2023).

9. Michael Webb. 2023. Exploring the potential for bias in ChatGPT. National centre for AI. Retrieved April 10, 2023 from https://nationalcentreforai.jiscinvolve.org/wp/2023/01/26/exploring-the-potential-for-bias-in-chatgpt/

10. Zhenrui Yue. 2023. LlamaRec: Two-Stage Recommendation using Large Language Models for Ranking. (August 2023).

11. https://doi.org/10.1038/s41598-024-55686-2.Lucy, Li, and David Bamman. “Gender and Representation Bias in GPT-3 Generated Stories.” ACLWeb, Association for Computational Linguistics, 1 June 2021, aclanthology.org/2021.nuse-1.5/. Accessed 16 Nov. 2021.

12. Fang, Xiao, et al. “Bias of AI-Generated Content: An Examination of News Produced by Large Language Models.” Scientific Reports, vol. 14, no. 1, 4 Mar. 2024, p. 5224, www.nature.com/articles/s41598-024-55686-2,

7 ACKNOWLEDGMENTS

We would like to extend our heartfelt appreciation to Professor Anirban Sen for his unwavering guidance and support throughout this research endeavor at Ashoka University, Department of Computer Science. His expertise and encouragement have been instrumental in shaping this work. We are also deeply grateful to the Department of Computer Science at Ashoka University for

providing the necessary resources and environment for conducting this research.

Special thanks are due to Aditya Bali and Anurav Singh for their significant contributions in curating the Farmers Protest dataset, which served as the cornerstone of our study.

We acknowledge with gratitude the support of Ashoka University for fostering an atmosphere of academic excellence and research innovation. This research was conducted as part of the ISM CS-IS-4054-1 course on News Recommendation Biases in LLMs.