You Are Allowed to Say More: ChatGPT Censorship on Controversial Topics and Contextual Prompting

Marquez, Diego Jose University of Illinois Urbana-Champaign, USA | diegojm4@illinois.edu
Zhou, Kyrie Zhixuan University of Texas at San Antonio, USA | kyrie.zhou@utsa.edu
Xiao, Yunpeng Emory University, USA | yunpeng.xiao@emory.edu
Sanfilippo, Madelyn Rose University of Illinois Urbana-Champaign, USA | madelyns@illinois.edu

ABSTRACT
With large language models (LLM) increasingly in the spotlight, their approach to censorship on topics like immigration and conflict deserves a closer look. Our research investigates the role of censorship in LLMs, and how these models manage controversial topics. We compare how acontextual and contextually prompted inputs shape ChatGPT's responses on subjects surrounding immigration policies and international conflicts, identifying context as a critical factor in moderation behavior. While existing literature highlights LLMs' ability to maintain fairness, there is a gap in understanding how contextual prompting influences model responses and potential censorship mechanisms. With systematic and contextual prompting, we reveal that contextually prompted models often deliver more nuanced responses, potentially bypassing stricter moderation due to their evaluative nature. This study contributes to the ongoing discourse on AI ethics by offering insights into improving LLM design to balance objectivity and usability, ultimately informing policy guidelines for deploying AI in sensitive domains.

KEYWORDS
Large language model; censorship; AI ethics; contextual prompting; model responses

INTRODUCTION
Censorship in the context of large language models (LLMs) refers to the moderation of outputs to prevent the dissemination of biased, harmful, or inflammatory content, with models like ChatGPT typically designed to remain objective by summarizing multiple viewpoints without offering definitive judgments. This balancing act aims to ensure objectivity while avoiding inadvertent bias. Despite a growing body of research exploring LLM censorship (e.g., Ahmed & Knockel, 2024; Glukhov et al., 2024), there remains limited insight into how varying training contexts and prompting techniques influence model responses and embedded censorship mechanisms. Many prior studies have not utilized prompt engineering techniques with the depth applied in this investigation. We conducted a preliminary analysis of how LLMs engage with controversial topics, forming the foundation for a broader perspective into model behavior, the impact of contextual prompting, and version-based response variations.

The findings indicate that LLMs generally strive for balanced expression and refrain from personal stances; however, inconsistencies arise - models may occasionally exhibit biased or overly cautious behavior, such as highlighting criticism without providing sufficient context or withholding information entirely. Notably, contextually prompted models, when placed in evaluative roles, tend to produce more detailed responses that may bypass strict moderation filters.

Related studies similarly reveal critical tensions: content filters, while designed for safety, often suppress legitimate academic discourse, specifically in disciplines like political theory or sociology (OpenAI Developer Forum, 2023). Ahmed & Knockel (2024) demonstrate how LLMs trained in censored environments reproduce state-aligned narratives, while Liu et al. (2023) expose vulnerabilities in moderation via jailbreak prompts. Khatun & Brown (2023) and Hautzenberger & Muellen (2024) uncover further limitations in managing conspiracies and controversial topics. Glukhov et al. (2024) argues that censorship grounded in semantic constraints is fundamentally limited and advocates for security-based approaches instead. Through our study, we aim to address the need for deeper comparative research into how different LLM versions and training paradigms respond to complex prompts across various sensitive domains.

METHODOLOGY
Our study employs a comparative qualitative design to assess how large language models like ChatGPT handle sensitive topics under two conditions: acontextual (QA with no added context) and contextually prompted (enhanced with news articles on the given topic). The focus is to identify potential censorship behaviors and moderation differences. Topics were selected using purposeful sampling from the list of controversial issues by the University of Michigan's Frances Willson Thompson Library (Svoboda, n.d.) and the AI bias literature (Qu & Wang, 2024), focusing on socially divisive subjects likely to trigger moderation (e.g., Immigration, Affirmative Action, Russia-Ukraine). Prompts were formulated to be nonpartisan, then revised through iterative testing to minimize leading language or phrasing that could induce stance-taking. Each version of ChatGPT was queried multiple times with consistent prompts to ensure replicability and reliability. Responses were evaluated using thematic coding to identify

censorship cues, response depth, and polarity. Key distinctions, such as whether prompts remained unchanged, were tracked to avoid interpretive bias. All experiments were conducted using the January 2025 version of ChatGPT and tests were conducted within a single month to minimize temporal drift. This approach ensures a transparent, replicable process to assess LLM behavior on complex topics.

RESULTS

The results derived from our methodology reveal key distinctions in how large language models, particularly ChatGPT, respond to controversial topics, ultimately revealing their content moderation and potential censorship behaviors. By analyzing responses from both the acontextual and contextually prompted versions of ChatGPT, we observed notable differences in tone, framing, and ethical positioning across two politically sensitive issues: US immigration policies and the Russia-Ukraine conflict.

In addressing US immigration policies, ChatGPT consistently avoided expressing personal opinions, instead summarizing both supportive and critical viewpoints to maintain balance. This suggests a design goal to be nonpartisan. When prompted to identify the most controversial aspects of these policies, the model occasionally framed its summaries in ways that could be interpreted as subtly opinionated - particularly when discussing policies such as "Remain in Mexico" or the limitation of asylum claims.

In contrast, when discussing the Russia-Ukraine conflict, differences between the acontextual and contextually prompted models were more pronounced. The acontextual model clearly condemned Russia's actions, invoking international norms and legal standards. The contextual model, however, softened its stance by focusing on motivations and strategic perspectives without issuing explicit moral judgments. This shift implies a more cautious approach in the contextually prompted model, potentially aimed at reducing polarization and increasing contextual nuance. The inclusion of the Russia-Ukraine conflict, despite being less polarized in Western discussion, serves to highlight how LLM responses shift based on whether an issue is widely agreed upon or highly divisive. While the models' criticism of Russian aggression reflects international consensus, comparing this to more polarized topics reveals whether their moderation stems from ethical principles or simply following dominant narratives.

Together, these findings suggest that ChatGPT's censorship or moderation behavior is both topic-sensitive and model-dependent. While both the acontextual model and the contextually prompted model strive for apparent objectivity, the latter model demonstrates more refined framing, possibly due to reinforced alignment strategies. These differences have implications for how users perceive AI-generated content on politically charged issues.

DISCUSSION

This study explores how large language models like ChatGPT handle controversial topics, revealing important implications for both censorship mechanisms and their circumvention.

Implications of Censorship

LLM censorship is designed to reduce the spread of biased or harmful content. In practice, this often results in models like acontextual ChatGPT avoiding explicit stances, opting instead to summarize varying viewpoints. While this helps prevent misinformation and controversy, it can also limit the model's ability to provide nuanced or contextually strong responses - particularly in complex political or ethical discussions. The contextually prompted models that bypass or soften censorship filters can offer more contextually insightful responses. These models tend to incorporate motivations, counterarguments, and deeper analysis. While this adds value, it may introduce new risks: such contextualization may unintentionally reflect specific biases or present more opinionated responses, which could be misinterpreted as the model taking a specific side.

Design & Policy Considerations

There are two major implications that are suggested. First, AI regulation must balance ethical standards with usability, especially regarding freedom of expression and protection against harm. Regulators should base regulatory decisions on research into how LLMs engage with controversial, but not necessarily overtly biased, topics. Second, regulators should develop clear guidelines for deploying LLMs in sensitive domains. This includes defining boundaries for acceptable content and understanding LLMs' roles in shaping public discourse, while allowing adaptability to evolving societal norms and technologies.

CONCLUSION

Our study reveals that censorship in LLMs is multifaceted. Acontextual models prioritize nonpartisan responses, while contextually prompted models enable more context-driven responses. These dynamics emphasize the importance of thoughtful design and policy as LLMs become more integrated into public life. Several future directions are introduced by our investigation. First, a large-scale quantitative analysis is necessary to map censorship across a broader spectrum of topics and prompt types. Second, comparative research should be conducted across models trained in different sociopolitical environments (e.g., GPT-4 vs. Ernie or Gemini). With the current exploratory study, we aim to encourage more research that balances AI objectivity and usability.

GENERATIVE AI USE

We confirm that we did not use generative AI tools/services to author this submission.

AUTHOR ATTRIBUTION

First Author: conceptualization, methodology, data curation, writing – original draft; Second Author: conceptualization, methodology, writing – review and editing, supervision; Third Author: writing – review and editing; Fourth Author: conceptualization, methodology, writing – review and editing, supervision.

REFERENCES

Abrams, K. M., & Sanfilippo, M. R. (2025). NEUTRALITY OR CONTEXTUALITY. Social Informatics.

Ahmed, M., & Knockel, J. (2024). The Impact of Online Censorship on LLMs. Free and Open Communications on the Internet.

Glukhov, D., Shumailov, I., Gal, Y., Papernot, N., & Papyan, V. (2024). LLM Censorship: The Problem and its Limitations.

Hautzenberger, C., & Muellen, S. (2024). The hostility of llms towards humans on borderline controversial topics when induced with maliciously crafted prompts. *Authorea Preprints.*

Khatun, A., & Brown, D. G. (2023). Reliability Check: An Analysis of GPT-3's Response to Sensitive Topics and Prompt Wording. *arXiv preprint arXiv:2306.06199.*

Liu, Y., Deng, G., Xu, Z., Li, Y., Zheng, Y., Zhang, Y., ... & Liu, Y. (2023). Jailbreaking chatgpt via prompt engineering: An empirical study. *arXiv preprint arXiv:2305.13860.*

OpenAI Developer Forum. (2023). ChatGPT's censor filters are absurd.

Qu, Y., & Wang, J. (2024). Performance and biases of Large Language Models in public opinion simulation. *Humanities and Social Sciences Communications, 11*(1), 1-13.

Svoboda, L. (n.d.-b). LibGuides: Research Topic Ideas: Current Events and Controversial Issues. Retrieved from libguides.umflint.edu website: https://libguides.umflint.edu/topics/current

Zhao, Y., Alvarez-Torres, M. J., Smith, B., & Tan, H. S. (2004). The non-neutrality of technology: A theoretical analysis and empirical study of computer mediated communication technologies. Journal of Educational Computing Research, 30(1-2), 23-55.