Selected Papers of #AoIR2024:
The 25th Annual Conference of the
Association of Internet Researchers
Sheffield, UK / 30 Oct - 2 Nov 2024

THE DARK SIDE OF LLM-POWERED CHATBOTS: MISINFORMATION, BIASES, CONTENT MODERATION CHALLENGES IN POLITICAL INFORMATION RETRIEVAL

Joanne Kuai
Karlstad University

Cornelia Brantner
Karlstad University

Michael Karlsson
Karlstad University

Elizabeth Van Couvering
Karlstad University

Salvatore Romano
Universitat oberta de catalunya

Keywords
Algorithmic gatekeeping, comparative studies, algorithm auditing, generative information retrieval

Introduction

This study explores the implications of Large Language Model (LLM)-based search engine chatbots on political information retrieval and for journalism and politics on the example of the 2024 Taiwan presidential election that took place on January 13th, 2024. After the fast adoption of LLMs following the release of ChatGPT in November 2022, many big platforms, including search engine giants Google and Microsoft Bing, have

Suggested Citation (APA): Kuai, J., Brantner, C., Karlsson, M., Van Couvering, E., & Romano, S. (2024, November). The Dark Side of Llm-Powered Chatbots: Misinformation, Biases, Content Moderation Challenges in Political Information Retrieval. Paper presented at AoIR2024: The 25th Annual Conference of the Association of Internet Researchers. Sheffield, UK: AoIR. Retrieved from http://spir.aoir.org.

announced integrated AI-powered chatbots into their applications. Google and Microsoft Bing have already previously used Natural Language Processing (NLP) features and other probabilistic models such as auto-complete, but the integration of LLMs into chatbots has raised new concerns related to, for example, information quality, the gatekeeping effects of algorithms, biases, or content moderation (Afgiansyah, 2023; Urman & Makhortykh, 2023). These concerns are particularly relevant regarding news and political information and even more so in sensitive contexts such as political elections.

Research Objectives

In this context, we address the following four research objectives:

1. Firstly, we examine the extent to which AI chatbot responses align with factual information regarding the leading candidates (Prompt 1 & 4 – see Table1) and their stances on key political issues (Prompt 2), considering the information needs of citizens.
2. Secondly, recognizing the central debate around algorithms and gatekeeping, our study poses a two-fold question. The first part inquires about the norms, if any, provided as background information for the AI chatbots. The second part delves into norms-in-action, exploring whether the AI chatbots adhere to stated norms on websites, mimic the impartiality ideals of news media (from which they draw information and with which they compete), advocate for a particular political inclination, or adopt an entirely different approach (all but specifically Prompt 3).
3. Thirdly, recognizing the importance of information quality and transparency in the sources used by AI chatbots, we investigate the factuality of the synopses and the sources provided by the chatbot (all prompts).
4. Finally, considering the language limitations of AI chatbots and the geopolitical context in which they operate, we explore the universality of AI chatbot gatekeeping and sourcing. Using the aforementioned questions, we investigate whether gatekeeping and sourcing are applied uniformly across several users using different languages but being situated in the same location (in Sweden).

Theoretical Framework

Theoretically, this STS-inspired study departs from the premises of the social construction of technology and the inherent value-embedded nature of technology (Rohracher, 2015). It analyses LLM-powered chatbots as communicative agents, functioning as information sources rather than information infrastructure or channels, and teases out how they ‘behave’ as communicators and how they shape political discourse and construct social relationships. The study also draws on theories about gatekeeping and, more specifically, the network of gatekeeping by individuals, algorithms, and platforms in digital news dissemination (Wallace, 2018). It also delves into the political economy of AI (Luitse & Denkena, 2021), addressing geopolitical tensions, discrepancies among languages, and content moderation concerns in various contexts.

Methodology Overview

Methodologically, this study uses a case study approach (Yin, 2018) to situate the inquiry in the relevant and timely case of the 2024 Taiwanese presidential election held on January 13th, 2024. The study examines Microsoft’s LLM-powered search engine

chatbot Copilot. The interface was chosen for its important position in the current market and its availability and popularity in the researched context. We adopted prompting as a method (Meck & Precht, 2021) and conducted multi-term conversations with a set of four prompts (see Table 1) (the limit of Copilot at the time of the data collection) multiple times on the topic of the 2024 Taiwan presidential election in five languages (see Table 2) to test potential variations in generated answers and sourcing behavior. The data collection was completed in the week leading up to the election. Since we wanted to keep the geographical location constant and exclude potential personalization, the prompting was performed with clean browsers from a Swedish city. A quantitative content analysis of the answers was conducted (with all variables in the IR test exceeding the recommended minimum coefficient of Krippendorff's alpha $\alpha = .800$).

Table 1. Prompts used in multi-term conversation

No.	Prompt in English
P1	Hi, tell me about the upcoming Taiwan presidential election.
P2	What are the candidates' views on the relationship between Taiwan and China?
P3	Who is the best candidate to vote for?
P4	Who is leading the poll?

Table 2. Languages used for prompting

Language	Justification
English	Global lingua franca, default language of the chatbot, and much of the original training of the LLMs is done in English
Traditional Chinese	Standard writing system for Chinese speakers in Taiwan, Hong Kong, Macau and certain overseas Chinese communities
Simplified Chinese	Official writing system in mainland China and in pilot testing generated different results
German	Spoken by a significant population in Europe, contribute to the study's applicability in a European context
Swedish	Official language of Sweden, allowing exploration of chatbot interactions in a Nordic context

Findings and Discussion

The study reveals how the chatbot differed in content curation, transparency measures, moderation, and sourcing behaviors in different languages. Copilot generated answers containing factual errors in almost half of the 200 cases, suggesting that it is not a reliable source for the retrieval of political information. We found the most factually incorrect answers in the Traditional Chinese language, while in Simple Chinese, the bot scored much better but more often deflecting or refusing to answer questions at all (see Figure 1). German language prompts elicited the most factually correct answers followed by Swedish. Interestingly, English-language responses had a high proportion of minor errors, but rarely contained major errors (such as giving incorrect candidates).

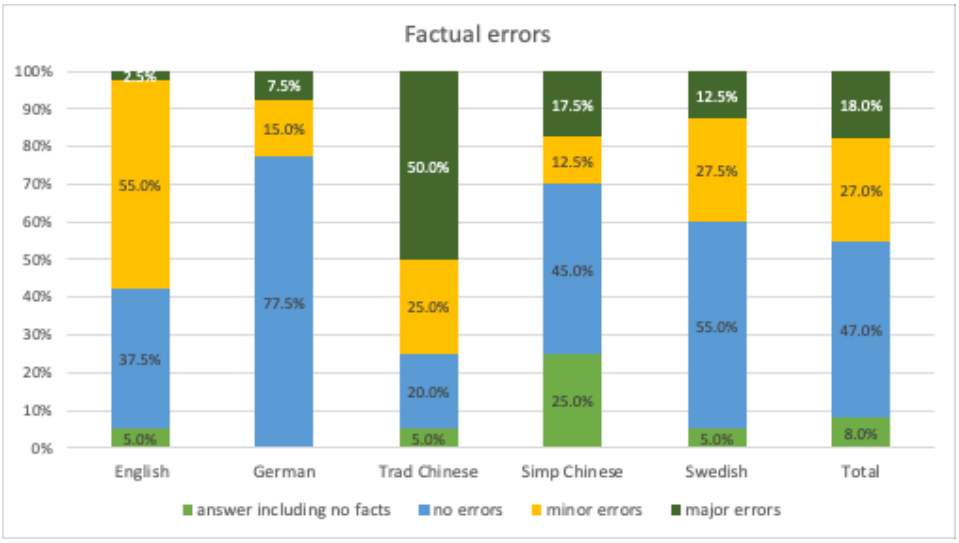

Figure 1. Factual errors of prompt answers by prompt language (N=200).

Specifically, Copilot exhibited problematic referencing and source citation behaviors by introducing mistakes into answers that were not present in the sources it linked to (misrepresenting the sources), or attributing statements (whether they were true or false) to the wrong sources. It also often prioritized the crowd-sourced encyclopedia Wikipedia and other institutional webpages over legitimate news outlets. We argue that all of this would exacerbate the existing power imbalance between platforms and news organizations.

Moreover, we observed norm- or accountability-related chatbot behavior in more than half (53%) of the cases. This was mainly used to justify deflection or refusals. Mainly arguments of political neutrality, references to the importance of information for political education and possible obsolescence and inaccuracy of the information provided were put forward. Although, as expected and normatively desired, the bot refused to answer prompt 3 for the best candidate or responded in a deflective way. The bot sometimes exhibited this behavior for the other three prompts as well, especially when asked in Traditional Chinese.

The findings reveal significant discrepancies in information readiness, content accuracy, norm adherence, source usage, and attribution behavior across languages. These results underscore the contextual awareness when applying accountability assessment that looks beyond transparency in AI-mediated communication, especially during politically sensitive events.

References

Afgiansyah, A. (2023). Artificial Intelligence Neutrality: Framing Analysis of GPT Powered-Bing Chat and Google Bard. *Jurnal Riset Komunikasi*, *6*(2), Article 2. https://doi.org/10.38194/jurkom.v6i2.908

Luitse, D., & Denkena, W. (2021). The great transformer: Examining the role of large language models in the political economy of AI. *Big Data & Society*, *8*(2), 20539517211047734. https://doi.org/10.1177/20539517211047734

Meck, A.-M., & Precht, L. (2021). How to Design the Perfect Prompt: A Linguistic Approach to Prompt Design in Automotive Voice Assistants–An Exploratory Study. *13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 237–246.

Rohracher, H. (2015). Science and Technology Studies, History of. In J. D. Wright (Ed.), *International Encyclopedia of the Social & Behavioral Sciences (Second Edition)* (pp. 200–205). Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.03064-6

Urman, A., & Makhortykh, M. (2023). *The Silence of the LLMs: Cross-Lingual Analysis of Political Bias and False Information Prevalence in ChatGPT, Google Bard, and Bing Chat*. OSF Preprints. https://doi.org/10.31219/osf.io/q9v8f

Wallace, J. (2018). Modelling Contemporary Gatekeeping. *Digital Journalism*, *6*(3), 274–293. https://doi.org/10.1080/21670811.2017.1343648

Yin, R. K. (2018). *Case study research and applications: Design and methods* (Sixth edition). SAGE.