# Improving Wikimedia resilience against the risks of content-generating AI systems

**(WM-Research-Fund-2023 Submission 1882)**

Heather Ford
University of Technology Sydney

Michael Davis
University of Technology Sydney

Marian-Andrei Rizoiu
University of Technology Sydney

## Abstract

Large Language Models (LLMs) like ChatGPT have captured global attention, promising great advances to knowledge. Some in the Wikimedia community have identified the possibilities of LLMs: enabling editors to generate a first draft of an article, to summarise sources, to produce transcriptions of video and to more easily query Wikidata content (see Harrison, 2023; Wikimedia community, 2023). Others have highlighted the possible risks of LLMs producing  vast swathes of AI-generated content or automated comments to simulate the appearance of discussion, debate and consensus that make the job of maintaining quality, verified, consensus-driven content difficult (see Harrison, 2023; Wikimedia contributors, 2023a). The aim of this project is to explore the implications of content-generating AI systems such as ChatGPT for knowledge integrity on Wikipedia and to investigate whether Wikipedia rules and practices are robust enough to deal with the next generation of AI tools. Knowledge integrity is a foundational principle of Wikipedia practice: the verifiability of Wikipedia content is a core content policy and Wikipedians have been working to understand how to counter systemic bias on the project for almost two decades. By garnering perspectives from Wikimedia practitioners, LLM experts and academic and grey literature about its possible (and evolving) implications and by analysing current policies and practices for vetting automated tools, this project will map out the most important areas for possible policy expansion and adjustment of current practice to deal with possible risks to the Wikipedia project. This work supports the 2030 Strategic Direction in its aim to ensure that Wikimedia constitutes essential infrastructure of the ecosystem of free knowledge (Wikimedia contributors, 2023b). It will also provide insight into potential two-way information flows between Wikipedia and AI systems, with the aim of developing strategies to ensure that flow comprises comprehensive, reliable, and high-quality information.

## Introduction

Verifiability is one of Wikipedia's core content policies. For Wikipedia editors, verifiability means that "all material must be attributable to reliable, published sources." (Wikipedia:Verifiability) This principle establishes rights for readers and responsibilities for editors (Ford, 2020). Readers should have the right to be able to check whether information from Wikipedia is accurately represented by the reliable source from which it originates. Editors should ensure that all information is attributable to reliable sources and that information that is likely to be challenged is attributed using in-text citations. The latest generation of LLMs, mostly in the form of chatbots, has led to a great deal of concern within the Wikimedia community as well as the broader academic community about the potential of these tools to disrupt processes of knowledge formation, verification and dissemination.

The arrival of LLMs introduces two potential threats in relation to verifiability. The first is the threat of flooding the site with inaccurate and inaccurately cited statements that require human verification to check that they a) are reliable and b) actually support the claim summarised in the text of the article. The second is in Wikipedia being prioritised as a source by LLMs but where its content isn't being adequately preserved or cited. Research by McMahon, Johnson and Hecht (2017), for example, has demonstrated how Wikipedia is already prioritised by knowledge graph products like smart search that answer user questions, but that it is rarely cited. Aside from the inherent risk of the first threat, there is the added potential for it to exacerbate the second, with inaccurate or inaccurately cited statements in Wikipedia being used as a data source for LLMs (particularly if recently announced ChatGPT plugins can access the Wikimedia API to update their data, or in real time), in an extension of the potential problem of Wikipedia as a self-citing source (Magnus 2009).

LLMs are already notorious in their production of inaccurate, unreliable or fabricated citations and content, and experts have urged the importance of manual verification given the increased power of the latest generation of LLMs (van Dis et al., 2023). Wikipedia already suffers from citations that don't actually reflect claims in articles (Ford, 2023) and it has been estimated by Wikimedia Research that one in four articles in English Wikipedia does not have any references at all (Wikimedia Research, 2018). Moreover, while Wikipedia and other Wikimedia products have benefitted from machine-learning products in the past, the right balance needs to be struck in order to ensure the quality of Wikipedia content is maintained. The Wikimedia Research team has started to develop algorithms that determine if a statement requires a citation in order to focus the manual labour required to curate and fact-check. But adding swathes of inaccurately sourced content may tip the scales so that the manual labour of verification becomes unmanageable, particularly for less-resourced LOTE Wikipedias. Some Wikipedia language versions like Cebuano Wikipedia (the second largest Wikipedia because of the extent of automated article creation on the site) have been negatively affected by automated content generation so that the small community isn't able to adequately maintain the massive amount of content produced automatically. Improving the resilience of Wikipedia to these threats will also support the integrity of Wikipedia content accessed via the Wikimedia API, including by LLMs.

This project aims to map how Wikipedia might govern the use of AI-generated material. The research will analyse and assess current policies and practices for managing automated content and assess these against projected risks to information integrity ascertained from a series of semi-structured interviews with editors, Wikimedia Foundation employees and LLM experts as well as a review of academic and grey literature. The objective will be to issue a set of recommendations for effective interventions in policy and practice to manage LLM-derived risks.

The project will be driven by 3 research questions:

**RQ1:** Does the potential use of ChatGPT and other AI chatbots threaten knowledge integrity on Wikipedia? If so, how?

**RQ2:** To what extent are current Wikipedia policies and processes able to address any risk to information integrity posed by ChatGPT?

**RQ3:** If there are gaps in these policies and processes which could be exploited or which may otherwise present risks for information integrity, what policy and process interventions may mitigate these risks?

The project has two key objectives relating to the needs of a) the Wikimedia community and b) academic research community. First, we aim to develop a series of recommendations through community interviews and focus groups on possible next steps for governing Wikipedia's approach to LLMs. Second, we aim to develop a significant contribution to digital media policy and information systems research by presenting an example model for governing AI-generated content in public information systems.

**TImeline**: Start date July 1, 2023; end date June 30, 2024.

**Phase 1**
**July 2023**

- Literature review (including initial Wikipedia policy analysis, review of all Wikimedia conversations about ChatGPT and LLMs, relevant digital media policy and platform governance, applied epistemology and information systems literature review)
- University ethics application.

**Phase 2**
**August 2023**

- Run workshop (as focus group) and conduct interviews at Wikimania Singapore (workshop session submitted for review)

**September 2023**

- Conclude interviews online

**Phase 3**
**October - December 2023**

- Analyse interview data, complete analysis of relevant policies and practice

**January - March 2024**

- Write up results

**May - June 2024**

- Present research results at UTS and at WikiWorkshop 2024 (or equivalent);
- Submit journal article to relevant digital media policy journal, e.g. Policy and Internet.

## Related work

This interdisciplinary project is situated primarily in digital media and digital media policy studies, but overlaps with issues in information systems and applied epistemology. It responds to a gap in understanding the extent to which Wikipedia's approach to knowledge representation, information verification and free knowledge can withstand the possible risks from a new generation of AI agents delivered to the public at scale.

Our research questions extend from research on Wikipedia sources, citations and the principles of verifiability conducted by Ford over the past decade. Initial studies explored the practice of Wikipedians' working with sources and citations in documenting breaking news events on Wikipedia (see Ford, 2012) and comparing the practice of verification on Wikipedia with the same practices on other collaborative platforms (Ford, 2011). A 2013 study (Ford et al, 2013) was the first major study of Wikipedia sources where it was reported that Wikipedia relies heavily on information derived by sources other than scholarly secondary sources. In 2015 (Sen, Ford, Musicant, Graham, Keyes & Hecht) we expanded on this study to understand how geographically local sources were to the subjects of Wikipedia articles. Also in 2015, we started to investigate opposition to Google's wholesale reuse of Wikipedia content, often without credit in its knowledge panels, theorising the loss of agency experienced by Wikipedians and the greater public (Ford and Graham, 2015). This was expanded upon for a chapter for the MIT Press Wikipedia@20 book looking specifically at the verifiability principle in the context of knowledge graphs, machine learning and AI (Ford, 2020), and then again in "Writing the Revolution" (Ford, 2022). Previous research by Ford and others like McMahon, Johnson & Hecht (2017) has been conducted in the context of knowledge graph products to demonstrate the significant reliance on Wikipedia by large commercial platforms like Google.

The recent arrival of ChatGPT and other latest-generation LLMs may both exacerbate many of the problems previously identified, but likely also introduce novel or distinct problems arising from the availability, ease of use and improved output of the latest-generation tools. While the emergence of these technologies has generated substantial concern in research and practitioner communities as well as the broader public, research on the implications of LLM chatbots for knowledge integrity is naturally in its infancy. Flanagin et al. (2023) investigate the implications of ChatGPT for scientific publication in the medical field, and there have been several studies on implications for education (e.g. Perkins, 2023), while Tan et al. (2023) evaluate the performance of ChatGPT as a question answering system (QAS). This supplements an existing corpus of research on the epistemic performance or implications of previous generations of LLM tools such as GPT3 (e.g. Floridi & Chiriatti, 2020; McGuffie & Newhouse, 2020). Some very recent research briefly discusses Wikipedia in the context of broader examinations of LLM chatbots (Floridi, 2023). However, no previous academic research focused on the implications of latest-generation LLM chatbots for Wikipedia has been identified. Hence there exists a significant gap in our understanding of how resilient Wikipedia is to this new generation of AI tools made available to the public.

We are particularly interested in Wikipedia in the context of LLMs from a broader information quality perspective. Recent misinformation research is narrowly focused on social media networks and on psychological and political factors. Our broad-based approach understands information quality as contextual, i.e. dependent on features and practices of the environment in which information is developed and applied. Instead of focusing only on obvious cases of false information at the extreme, we are interested in practices of information production on Wikipedia more generally and the extent to which they enable robust, secure and ethical information production.

## Methods

As noted, the project is focused on the systematic analysis of the implications of latest-generation LLMs for Wikimedia policy and practice. Accordingly, we apply analytical methods from policy analysis and applied epistemology to:

1. analyse relevant Wikipedia content policy and practice

2. model the epistemic processes embedded in the practices of the Wikipedia community, including the application of policy
3. understand the potential interactions of LLMs in Wikipedia knowledge generation, verification and dissemination
4. assess the risks for Wikipedia information integrity presented by LLMs
5. analyse potential interventions in policy and practice to mitigate these risks.

We consider it critical to ground this analysis in the actual practices of the Wikipedia community. Policy analysis alone cannot provide adequate insight into epistemic practice, nor therefore, into any risks presented by LLMs. Our analysis will therefore be grounded in the everyday experience and actual practice of Wikipedia editors and Wikimedia Foundation experts working on information integrity issues, with particular consideration given to the context of community, academic and public conversations about LLMs and their implications for knowledge and truth.

The research will proceed in three phases:

**Phase 1 (desk-based data collection)**

In phase 1 we will conduct a comprehensive review of:

- on-wiki discussions about LLMs and their possible impact on Wikipedia
- relevant grey literature from commercial operators like OpenAI, Bard, Bing
- alternative approaches to verifiability from alternative operators like Mozilla.AI
- current moves to regulate, govern or issue moratoria on LLMs (e.g. the open letter to pause “Giant AI Experiments” organised by the Future of Life Institute)
- current Wikipedia policies (e.g. verifiability and bot policies) and practices (e.g. page patrol) most likely to be affected by the introduction of latest-generation LLMs like ChatGPT
- digital media policy and governance literature and applied epistemology literature that focuses on issues of information quality, verification, transparency and data provenance, particularly on Wikipedia.

This phase will identify a series of venues where public debate is happening around these issues. This will enable us to hone our interview list and interview questions in preparation for phase 2. We will continue to monitor these venues through to the final phase of the project. This phase will also provide a corpus of relevant Wikipedia policies and a comprehensive understanding of recent research that will provide a basis for our analysis.

**Phase 2 (interview-based data collection)**

In phase 2 we will conduct a series of semi-structured interviews that build on the data collected in phase 1. The objective of this phase is to identify risks to Wikipedia’s information integrity from LLMs (to answer RQ1) and to gather information on the application of Wikipedia policy and processes in practice (to suggest leads for RQ2). We will conduct about 15 interviews with the following groups, using snowball sampling to find those with relevant experience:

- **Wikimedians:** starting with an in-person focus group/workshop at Wikimania Singapore and interviews with individuals identified in Wikimedia-l conversations, followed by individual interviews. The goal is to understand to what extent LLMs are already having an impact on Wikipedia practice, which areas of practice might be most affected, and whether there are other risks not already identified that will be useful to consider. We will focus on community members who have direct experience working in areas most likely to be affected or related to LLMs (e.g. in new page patrol, bot policy etc).

- **Wikimedia Research Team members:** particularly those connected to the Knowledge Integrity program. The goal with this group is to understand how knowledge integrity relates in practice to questions of verifiability and provenance and to garner ideas about what is possible in terms of governing LLMs (given previous practice in relation to governing other automated processes and tools).
- **OpenAI and other LLM practitioners** (e.g. for Mozilla.AI and FOSS alternatives). The goal in interviewing this group is to understand the way that engineers are thinking about threats to information integrity from their products and what is being done or considered to mitigate against it.

**Phase 3 (analysis)**
In phase 3 we will analyse the data gathered in phases 1 and 2. This will involve the application of a range of methods from digital ethnography, applied epistemology and policy analysis:

1. Analyse interview data from phase 2 using close reading and thematic analysis techniques from similar ethnographic studies (for example, see Ford, 2023).
2. Drawing on this analysis and relevant data from phase 1, analyse and model Wikimedians' verification practices using frameworks derived from existing analyses of the epistemology of Wikipedia (e.g.Frost-Arnold, 2019; Fallis, 2008), as well as process mapping and epistemic network analysis (Reijula & Kuorikoski, 2019; Sullivan et al., 2020; Shaffer et al., 2016).
3. Identify risks in existing policy and practice presented by the latest generation of LLMs using policy analysis methodology, including scenario analysis, case-study analysis and risk analysis, against the models and data obtained from (1) and (2).
4. Identify potential interventions in Wikipedia policy and practice to mitigate risks identified in (3), drawing again on applied epistemology and policy analysis methods. This will include drawing on data obtained from phase 2 consultations with Wikimedians and Wikimedia Research as well as the results from steps 2 and 3 of the phase 3 analysis.

## Expected output

We are planning three outputs for the project:

1. A research report for the Wikimedia Community highlighting the risks and possible mitigations against those risks. The report will aim to inform Wikipedia policy and practice, supporting knowledge integrity and increasing the resilience of Wikipedia and other Wikimedia projects to threats posed by LLMs. After our initial report is first drafted, we will send it to all interviewees and ask for feedback, conduct final clarification interviews over email or video calls where necessary and then publish the final report aimed at the Wikimedia community. We also intend to present results from the report at WikiWorkshop 2024.
2. After receiving feedback on the draft report, which will be integrated into our research results, we will finalise a journal article for our research

audience. We aim to publish this in a peer-reviewed OA Q1 policy-oriented journal such as *Policy and Internet*. The intended audience is digital media and media policy scholars. As well as contributing to scholarship on knowledge integrity and verification on Wikipedia, the article is likely to make an early contribution to understanding the implications of AI-generated content for information-integrity policy and practice in the digital environment more generally.

3. We intend to hold a public-facing event at the Centre for Media Transition at UTS to communicate results to a wide audience, including researchers in digital media and media, AI and technology policy; the tech industry; and policy practitioners.

## Risks

A minor risk in the project is that we are unable to obtain adequate data from phase 2 of the project by failing to source an adequate range of expert interviewees. We will seek to mitigate this risk in phase 1 by actively monitoring Wikimedia-l discussions and using contacts and networks built through the team's prior Wikipedia research work. We will also draw on literature and expert understanding of LLMs within our team (Andrei Rizoiu from UTS Data Science team) and other CMT research associates should we fail to secure adequate interviews on the LLM side.

There is a risk that the technology or policy environment relating to LLM AI chatbots will change over the course of the project. This may make some of our preliminary data or conclusions out of date. We will mitigate this risk by engaging in a second round of consultations before publishing the final outputs. Nonetheless, in our view it is likely that the LLM-related risks for knowledge integrity on Wikipedia, and more generally, will only increase as the next generation of LLMs is developed.

## Community impact plan

Our primary output is to develop a discussion piece both based on wide consultation with Wikimedians, and aimed at Wikimedians, about a question that is front of mind for many in the movement. Our study will be grounded in community practice, drawing on interviews with volunteer editor communities, and will in turn seek to inform community practice through policy adjustment and engagement with the community at events including WikiWorkshop and online forums as well as Wikimedia Research.

The Centre for Media Transition at UTS is an interdisciplinary research centre that is actively engaged with industry and policy practitioners in media, technology and journalism. We will seek to engage these audiences through a public-facing event at UTS that will focus on encouraging informed but wide-ranging dialogue on the implications of AI for digital knowledge ecosystems. The event would benefit greatly from the participation of members of the Wikimedia community. CMT researchers are also actively engaged in media and public outreach and the research team would seek to communicate its findings through media channels such as The Conversation.

## Evaluation

Given the project's focus on policy and practice, we will evaluate the project primarily on the basis of its impact on the Wikimedia community. Success will consist in Wikimedians taking up some of our proposals for further discussion, development, or implementation. Secondarily, impact in relevant academic fields will be measured by others citing our academic research outputs to build further on our analysis. Finally, broader public impact may be gauged via attendance and feedback on our work at the event at UTS or in public engagement via media channels.

## Budget

The funds requested include budget for:

- research time from the three CIs
- employment of a research assistant
- travel for 1 researcher to Singapore to attend Wikimania 2023 to hold a workshop and / or conduct interviews
- an event at UTS Centre for Media Transition to communicate findings and engage the broader industry and policy community in Australia
- travel for 1 researcher to attend WikiWorkshop 2024 to engage with the Wikimedia community
- subscription fees for transcription software (Otter)

Budget details redacted.

## Response to reviewers and meta-reviewers

Redacted

## References

Fallis, D. (2008). Toward an epistemology of Wikipedia. Journal of the American Society for Information Science and Technology, 59(10), 1662–1674. https://doi.org/10.1002/asi.20870

Flanagin, A., Bibbins-Domingo, K., Berkwits, M., & Christiansen, S. L. (2023). Nonhuman “Authors” and Implications for the Integrity of Scientific Publication and Medical Knowledge. JAMA, 329(8), 637–639. https://doi.org/10.1001/jama.2023.1344

Floridi, L. (2023). AI as Agency Without Intelligence: On ChatGPT, Large Language Models, and Other Generative Models. Philosophy & Technology, 36(1), 15. https://doi.org/10.1007/s13347-023-00621-y

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its Nature, Scope, Limits, and Consequences. Minds and Machines, 30(4), 681–694. https://doi.org/10.1007/s11023-020-09548-1

Ford, H. (2011) Verifying information from the crowd. Ushahidi.

Ford, H. (2012). Wikipedia sources: Managing sources in rapidly evolving global news articles on the English Wikipedia. *Available at SSRN 2127204*.

Ford, H., Sen, S., Musicant, D. R., & Miller, N. (2013, August). Getting to the source: Where does Wikipedia get its information from?. In Proceedings of the 9th international symposium on open collaboration (pp. 1-10).

Ford, H., & Graham, M. (2016). Provenance, power and place: Linked data and opaque digital geographies. Environment and Planning D: Society and Space, 34(6), 957-970.

Ford, H. (2019). Rise of the underdog. Wikipedia@ 20. MIT Press.

Ford, H. (2022). Writing the Revolution: Wikipedia and the Survival of Facts in the Digital Age, MIT Press

Frost-Arnold, K. (2019). Wikipedia. In *The Routledge Handbook of Applied Epistemology* (1st ed., Vol. 1, pp. 28–40). Routledge. https://doi.org/10.4324/9781315679099-3

Harrison, S. (January 12, 2023). “Should ChatGPT Be Used to Write Wikipedia Articles?”. Slate. https://slate.com/technology/2023/01/chatgpt-wikipedia-articles.html

Magnus, P. D. (2009).“On Trusting Wikipedia.” Episteme, 6(1): 74–90

McGuffie, K., & Newhouse, A. (2020). The Radicalization Risks of GPT-3 and Advanced Neural Language Models (arXiv:2009.06807). arXiv. https://doi.org/10.48550/arXiv.2009.06807

McMahon, C., Johnson, I., & Hecht, B. (2017, May). The substantial interdependence of Wikipedia and Google: A case study on the relationship between peer production communities and information technologies. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 11, No. 1, pp. 142-151).
Perkins, M. (2023). Academic Integrity considerations of AI Large Language Models in the post-pandemic era: ChatGPT and beyond. Journal of University Teaching & Learning Practice, 20(2). https://doi.org/10.53761/1.20.02.07
Reijula, S., & Kuorikoski, J. (2021). Modeling Epistemic Communities. In M. Fricker, P. J. Graham, D. K. Henderson, & N. J. L. L. Pedersen (Eds.), The Routledge handbook of social epistemology. Routledge.
Sen, S. W., Ford, H., Musicant, D. R., Graham, M., Keyes, O. S., & Hecht, B. (2015, April). Barriers to the localness of volunteered geographic information. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 197-206).
Shaffer, D. W., Collier, W., & Ruis, A. R. (2016). A Tutorial on Epistemic Network Analysis: Analyzing the Structure of Connections in Cognitive, Social, and Interaction Data. Journal of Learning Analytics, 3(3), Article 3. https://doi.org/10.18608/jla.2016.33.3
Sullivan, E., Sondag, M., Rutter, I., Meulemans, W., Cunningham, S., Speckmann, B., & Alfano, M. (2020). Vulnerability in Social Epistemic Networks. International Journal of Philosophical Studies, 28(5), 731–753. https://doi.org/10.1080/09672559.2020.1782562
Tan, Y., Min, D., Li, Y., Li, W., Hu, N., Chen, Y., & Qi, G. (2023). Evaluation of ChatGPT as a Question Answering System for Answering Complex Questions (arXiv:2303.07992). arXiv. https://doi.org/10.48550/arXiv.2303.07992
van Dis, E. A. M., Bollen, J., Zuidema, W., van Rooij, R., & Bockting, C. L. (2023). ChatGPT: Five priorities for research. Nature, 614(7947), 224–226. https://doi.org/10.1038/d41586-023-00288-7
Wikimedia contributors (2023a). Community Call Notes. Accessed 29 March, 2023. https://meta.wikimedia.org/w/index.php?title=Wikimedia_Foundation_Annual_Plan/2023-2024/Draft/External_Trends/Community_call_notes&oldid=24785109
Wikimedia contributors (2023b). Movement Strategy. Accessed 29 March, 2023. https://meta.wikimedia.org/w/index.php?title=Movement_Strategy&oldid=24329161
Wikimedia Research (2018). Characterizing Wikipedia Citation Usage: Second Round of Analysis. https://meta.wikimedia.org/wiki/Research:Characterizing_Wikipedia_Citation_Usage/Second_Round_of_Analysis

## Appendix - Wikimania workshop proposal

(Delivered as an online focus group if not accepted for Wikimania)

The aim of this session is to gather community perspectives on the risks and opportunities of LLMs like ChatGPT across multiple language versions of Wikipedia. The session will begin with a 15 minute presentation of what LLMs are and do, what we know about their implications for knowledge integrity and information quality on Wikipedia and what we still need to know. We will then move to producing an annotated list of questions and considerations relating to chatGPT and Wikipedia by brainstorming around three key themes (either in groups or as a single cohort, depending on numbers):

1. A SWOT analysis when considering LLMs in relation to Wikipedia practice across language versions;
2. Current Wikipedia policies that are implicated by LLMs;

3. Current Wikipedia practice in relation to AI and automation that we can learn from.

Participants will leave the session with a more grounded understanding of ChatGPT and its possible implications for Wikipedia, as well as some thoughts about what we still don't know and need to do to ensure that LLMs are an opportunity rather than a threat to Wikipedia's knowledge integrity.

Wikipedians have always thought about how new technologies relate to their ultimate goals of representing the sum of all human knowledge. ChatGPT is no different. This session will garner the local knowledge of Wikipedians to think through the possible implications of ChatGPT and LLMs for their projects and to help them better understand what we currently know and need to know in order to deal with the risks that LLMs might present.