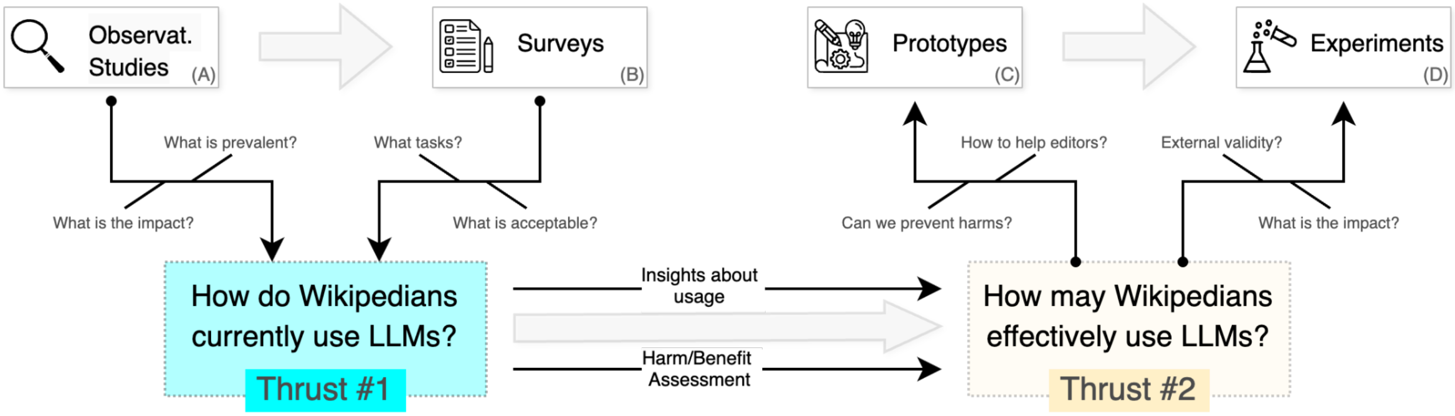
Extended: Navigating Today, Shaping Tomorrow: Studying the Role of LLMs on Wikipedia

Manoel Horta Ribeiro
Princeton University
manoel@cs.princeton.edu

Andrés Monroy-Hernandez
Princeton University
andresmh@princeton.edu

Fig. 1: We propose studying the role of LLMs on Wikipedia by measuring how these models are currently used (**Thrust #1**) and investigating how they may be effectively incorporated by Wikipedians (**Thrust #2**).



Abstract

Large language models (LLMs) threaten Wikipedia's integrity by generating unverifiable and non-neutral information, yet their use also promises improvements in knowledge quality and completeness. We propose to study LLMs' present and future role on Wikipedia through a multi-method approach combining observational studies, editor surveys, and controlled experiments. First, we will characterize how Wikipedians use LLM-based tools in their workflow through surveys and observational studies. Second, will prototype and test specialized AI assistance tools designed to mitigate risks while enhancing Wikipedia's content quality. Ultimately, our research aims to (1) comprehensively assess AI's beneficial and detrimental impacts and (2) inform policies and systems that responsibly integrate generative AI into Wikipedia.

Introduction

Large Language Models (LLMs) pose an immediate threat to Wikipedia's integrity by generating plausible-sounding but unsourced, unverifiable, or non-neutral text.[1] Yet, they may improve Wikipedia by addressing knowledge gaps and improving knowledge integrity—two key factors in achieving Wikimedia's 2030 strategic direction[2]—e.g., identifying outdated or biased information, or helping inexperienced editors write about underrepresented topics.

Recent work has provided preliminary evidence that LLM-generated text is prevalent on Wikipedia [6, 7], but whether the benefits of such AI-assisted text outweigh the harms begs the question. Ultimately, this trade-off depends on how Wikipedians use LLMs and how much LLMs or LLM-based tools can be effectively integrated into their editing workflow.

[1] https://en.wikipedia.org/wiki/Wikipedia:Large_language_models
[2] https://meta.wikimedia.org/wiki/Research:2030

We propose studying the role of LLMs on Wikipedia in two thrusts, each corresponding to a research question (See **Fig. 1**).

Thrust #1: *How do Wikipedians currently use LLMs?* Expanding on previous work examining the prevalence of LLM [3,4,6,7], we will study the nature and impact of AI use within Wikipedia. With *observational studies,* we will investigate the effect of existing AI use—does it lead to lower or higher-quality editing? Does it frequently introduce incorrect sources? Etc. We will employ freely accessible Wikipedia data and US-based web panels capturing user online activity [15]. Through *surveys with Wikipedians,* we will study Wikipedians' current attitudes towards AI. Specifically, the types of tasks AI is used for (e.g., content generation, filtering), the strategies editors employ to assess the quality of AI-generated contributions, and where Wikipedians believe AI usage to be acceptable. Importantly, we will explore these trends across various language editions.

Thrust #2: *How may Wikipedians effectively use LLMs?* Building upon **Thrust #1**, the project's second phase will explore whether AI tools designed for and with Wikipedians can amplify the benefits and mitigate the harms associated with generative AI. With as much involvement of the Wikipedia community as possible, we will develop, implement, and test a prototype tool to help editors in the most promising tasks. We will conduct experiments on the efficacy of this prototype in a sandbox environment (and would be happy to deploy it in a field experiment if possible). While the ultimate prototypes will be informed by results obtained in **Thrust #1**, as an example, we propose a direction for human AI collaboration in **Box 1**: AI agents that conduct laborious patrolling tasks, producing reports and suggestions that editors then vet.

AI Agents for Patrolling Wikipedia

Source checker patroller: Pages on Wikipedia must reference high-quality sources. Yet, unsourced pages are one of the largest clean-up categories on Wikipedia (https://w.wiki/6DS6), and there is an ongoing effort to improve reference quality (https://w.wiki/6DS3; https://w.wiki/DqDi). We propose an AI agent to access and evaluate the quality of external sources and find potential sources to be linked on the Web. (This would build on previous work using AI to source-check [42].)

New page patroller: New pages on Wikipedia must be patrolled to ensure quality. However, there is a large (and rapidly growing) backlog (https://w.wiki/7XpA). We propose an AI agent that systematically evaluates each page's adherence to relevant guidelines (https://w.wiki/DZwn), identifying potential issues such as lack of notoriety, promotional content, and vandalism. (This would build on previous work using AI for quality assessment on Wikipedia; see [43] for a survey)

Box 1. AI agents for Wikipedia. Agents' outputs would be suggestions that help streamline laborious patrolling tasks. AI agents would produce reports with recommendations for action (e.g., posted on Talk Pages). **This is an example; final prototypes will be decided based on community output (as elicited through Thrust #1).**

Outcomes: This research will inform Wikipedia of the current and future AI-related threats and equip the community with better methods and tools to reap the benefits and mitigate the harms of these new technologies. Concretely, we expect this project to produce: **(1)** data and insights into the use and impact of LLMs on Wikipedia, as well as attitudes toward LLM use; **(2)** tools for detecting LLM use in Wikipedia contexts; **(3)** prototypes that envision better human-AI collaborations; **(4)** insights on their efficiency from experiments.

Carrying out the proposed projects in the context of this call would maximize the potential impact of this research, as a key component to its success is the extent to which the Wikipedia community engages and co-designs the research proposed.

Dates: September 1st, 2025 to August 31st, 2027.

Related work

We review previous work discussing AI-generated content's impact on the Web, and specifically on Wikipedia. Then, we discuss HCI work on designing effective AI tools.

AI-generated Content and the Web

AI-generated content is increasingly prevalent on the Web. Thompson et al. (2024) find evidence that machine-translated content constitutes much of the total web content in lower resource languages [3]. Conservative estimates indicate that millions of images are generated by AI daily,[3] leading experts to worry about the future of our information ecosystem [4]. Perhaps unsurprisingly, recent work has also suggested that AI-generated content is prevalent on Wikipedia [6, 7].

At the same time, previous work suggests that users selectively engage less with existing platforms when AI provides comparable content. For example, Burtch et al. (2024) found evidence that ChatGPT reduced engagement with StackOverflow [5]. Lloyd et al. (2025) find that Reddit moderators fear LLMs will decrease their communities' utility and social values [19]. Finally, and more relevant to the proposal at hand, Liu et al. (2025) show that topics that are frequently discussed with LLMs are less viewed (and less edited) on Wikipedia [7]. The magnitude of this effect remains unclear, as other analyses find no effect when considering aggregate viewership trends [32].

Taken together, these trends imply that AI, and specifically LLMs, threaten digital public goods. LLMs act as "opaque" intermediaries between users and original knowledge sources [2], and their responses do not capture the full diversity of human thought and experiences [16, 17]. Therefore, LLMs could "pollute" the Web with well-written but low-credibility content.

Relationship to proposed work. LLMs' impact on Wikipedia is mediated by *how* the community uses these models. In **Thrust #1**, we aim to characterize such usage to make the potential threats (and solutions) more evident to Wikipedia and the research community. Assuming that some level of usage is inevitable, in **Thrust #2**, we propose imagining tools to mitigate the harms and reap the benefits of LLMs.

Human-AI Collaboration Paradigms

Successful automation and human-AI collaboration already exist on Wikipedia. For example, ORES [19] provides edit and page quality metrics widely used by various systems within Wikipedia (e.g., ClueBot, Events Dashboard). The Content Translation Tool helps port knowledge across language editions.[4] Descartes [20] "recommends potential Wikidata article descriptions for Wikipedia articles in 25 languages."[5]

More broadly, AI assistance can improve human performance in various tasks, from medical diagnosis [9] to coding [10]. However, the outputs of human-AI are often outperformed by the outputs of either humans or AIs alone [11]. Previous work looking specifically at Wikipedia suggests that providing LLMs with rules around neutrality is insufficient to ensure they behave like Wikipedians [8]. This is perhaps because decision-making on Wikipedia requires highly contextual judgements [21].

[3] https://journal.everypixel.com/ai-image-statistics

[4] https://en.wikipedia.org/wiki/Special:ContentTranslation

[5] https://w.wiki/DYXT

An extensive literature has studied how to structure and organize effective human-AI collaboration paradigms and tools [19, 22, 23, 24, 25, *inter alia]*. Yet, as LLMs advance in capabilities, new opportunities for human-AI collaboration emerge. Modern LLMs can perform entirely new sets of tasks (e.g., create text), and facilitate the creation of high-accuracy (although often imperfect) zero-shot or few-shot classifiers [26, 27]. Recent advances towards agentic AI[6] could empower these models to navigate and shape the Web very generally, without needing carefully crafted APIs [28]. These new capabilities open a design space for tools that allow natural language inputs and carry out complex, multi-step tasks [33, 12, 13, 14, 36].

Relationship to proposed work. This project aims to identify productive (and "destructive") tasks and paradigms for Human-AI collaboration in the context of Wikipedia (**Thrust #1**). Further, we propose designing tools to empower Human-AI collaboration on Wikipedia in ways that reflect community values (**Thrust #2**).

Previous Wikimedia Fund Grant

The Wikimedia community recognized the implications of AI-generated content early on,[7] resulting in a dedicated 2023–2024 Wikimedia Research Fund grant investigating generative AI's impact on Wikipedia's knowledge integrity.[8] This ongoing project aims to surface initial AI-related perceptions, concerns, and opportunities from within the community. Although the findings of this project may inform the work proposed here (pending results), its focus on perceptions (i.e., what is imagined to be the potential benefits/harms of generative AI) does not offer insight into current impacts and practices related to LLMs. Similarly, rather than eliciting perceived solutions, we propose to design, develop, and evaluate AI-assistance tools that may reap the benefits and mitigate the harms of LLMs and generative AI.

Methods

We will use a diverse methodological toolkit to study LLMs' current and potential impact on Wikipedia (**Fig. 1**). In **Thrust #1**, we will combine insights from (A) an observational study and (B) surveys. In **Thrust #2**, we will follow an HCI systems approach to (C) design prototypes and (D) experimentally assess their effectiveness. We detail these plans below.

Thrust #1A: Observational Study: LLM-Generated Content and Wikipedia

We will conduct observational analyses to understand how and where LLM-generated content appears on Wikipedia and how it impacts its overall content quality. We perform similar analyses in various language editions.

Characterizing LLM-generated text. We will develop methods to characterize LLM-generated text within articles, beyond the binary classification of entire pages, identifying which parts of articles and which type of content are LLM-generated. Following prior work [6,7,17], we will use a combination of AI detection tools, temporal editing patterns, and stylistic markers to make these inferences. To assess the reach of such content, we will link these findings to Wikipedia and third-party panel data capturing user engagement, allowing us to estimate how frequently users encounter LLM-influenced information during typical browsing sessions.

[6] https://openai.com/index/introducing-operator/
[7] E.g., from December 2022: https://w.wiki/DYXR
[8] Project page: https://w.wiki/DYXj

Assessing the impact of LLM-generated text. We will conduct analyses to evaluate the effects of LLM-generated content on article quality over time. We will compare the trajectories of articles that received edits likely to be AI-assisted with those that did not, focusing on metrics such as article completeness, neutrality, verifiability, and factual accuracy. These comparisons will use automated quality scores (e.g., ORES [19]) and expert human assessments. (We refer the reader to previous work by the Co-PI of this proposal on estimating causal effects from observational data on Wikipedia [34].)

Thrust #1B: **Survey Study:** Understanding Attitudes toward AI within the Community

We will design and administer a survey to understand attitudes toward AI within the Wikipedia editor community. The survey will focus on three core directions:

1. How do editors use AI tools in their workflow?
2. What forms of AI use are acceptable or unacceptable within Wikipedia's norms and values?
3. How do they envision AI being better used to support Wikipedia's mission in the future?

Survey development. This survey will be co-developed with Wikipedia community members to ensure the questions reflect community concerns, language, and priorities. We anticipate using participatory design approaches (e.g., discussion forums, calls for feedback on drafts; see Community Impact Plan) to ensure community input meaningfully shapes the instrument. We will also seek advice from Wikimedia affiliates and experienced editors.

Survey distribution. The survey will be distributed through established Wikipedia communication channels, including Village Pumps, community mailing lists, and interest groups (pending approval from these channels). Additionally, we will work with Wikimedia researchers to pursue the most appropriate dissemination venues (QuickSurveys, Banners).

Thrust #2C **Systems Design:** Crafting Effective Human-AI Collaboration

Building on the insights gathered in **Thrust #1**, our project's second phase will focus on designing and developing AI-powered tools that enable productive and responsible Human-AI collaboration on Wikipedia.

Prototyping. We will adopt an HCI systems approach (or what Oulasvirta and Hornbæk refer to as "*constructive problem solving*" [30]) grounded in user-centered design and iterative prototyping. This means designing tools in dialogue with the Wikipedia community, focusing on tasks where LLMs might add value—detecting knowledge gaps, improving written text, suggesting sources—while minimizing factual inaccuracy, overreliance, or content homogenization risks. As we design tools to facilitate human-AI collaboration, we will identify and support the "sweet spot" between automation and editorial control: empowering editors without replacing their judgment. Rather than proposing general-purpose AI interfaces, we will create constrained and transparent systems that offer editorial suggestions, prompt critical thinking, or assist with mechanical tasks.

Participatory design. To ensure that tools created reflect genuine community needs, we will use methods informed by participatory [38] and value-sensitive design [37]. Specifically, we will host co-design workshops with interested Wikipedia editors, conditional upon community support. We would particularly welcome collaboration with smaller language editions, recognizing that their unique challenges and insights could critically shape effective, inclusive, and context-sensitive AI solutions. Considering the prior success story of ORES [19], we foresee that adopting this design strategy will increase the chance of broad community adoption. (We refer the reader to previous work by the Co-PI of this proposal on creating tools with a participatory design in the context of gig work [39].)

Thrust #2D **Experiment**: Evaluating the Impact of Human-AI Collaboration

To understand the real-world impact of the Human-AI collaboration tools developed in **Thrust #2**, we will conduct experiments to evaluate their effectiveness, limitations, and unintended consequences. We expect these experiments to provide robust evidence about the value and risks of Human-AI collaboration in real Wikipedia editing contexts.

Controlled experiment. We will recruit experienced Wikipedians to use the prototype tools in a simulated or sandbox environment. Participants will complete tasks with and without AI assistance, allowing us to compare outcomes across various metrics—including article quality (as assessed by ORES and expert raters), editing efficiency, and user satisfaction. These experiments will help us isolate the specific contributions of the AI system and identify where it is most or least effective. We will also explore how different interface designs or guidance strategies influence relevant outcomes.

Field experiment. If the community deems it appropriate, we will also run a field experiment in a live editing context. We will deploy the tools to a randomly selected group of active editors. A comparable control group will continue editing as usual without access to the tool or incentives. By comparing the behavior and contributions of the treatment and control groups over time, we aim to assess the tools' impact on editing practices, article quality, and engagement in a naturalistic setting. (We refer the reader to a recent implementation of Post Guidance on Reddit where a new tool was also tested in a large-scale field experiment, led by the Co-PI of this proposal [31].)

Expected output

We expect this project to generate a range of outputs, which will be discussed in the following paragraphs. As the project progresses, we will document these outputs in scholarly publications and reports published within Meta-Wiki. If successful, this research may inform the academic community and the Wikipedia ecosystem and support future research and policy around AI integration in collaborative knowledge platforms.

Insights into perceptions, impact, and usage of LLMs on Wikipedia. We will analyze where and how LLM-generated content appears on Wikipedia, characterizing its reach and impact. These findings may help researchers and the Wikimedia community better understand the current use of AI on the platform. Additionally, we will obtain data about how stakeholders believe AI *should* be used on Wikipedia. Insights about these attitudes will be a starting point for imagining productive human-AI collaboration on Wikipedia.

AI detection tools for Wikipedia contexts. As part of our effort to characterize LLM-generated content, we will adapt and evaluate AI-detection techniques for Wikipedia-specific use cases. We will release code, models, and validation benchmarks to help identify likely AI-assisted text at the sentence or section level, accounting for Wikipedia's unique writing style and editing patterns.

Open-source prototypes of AI-assistance tools. We will design and release one or more open-source Human-AI collaboration tools to support Wikipedia editors in high-impact tasks. These tools will reflect best practices for ethical AI design and will accompany documentation detailing their design choices, limitations, and intended use cases. If there is interest from the community, we would be happy to help incorporate these tools (or aspects of them) into Wikipedia more formally. The final prototypes will be informed by findings obtained in Thrust #1, and co-designed with the community. Nonetheless, we provide a promising direction for AI-human collaboration in **Box 1**. Since several patrolling tasks on Wikipedia face a severe labor shortage, we propose implementing AI agents to streamline editors' workflow by producing reports and suggestions that editors would vet.

Insights from controlled and field experiments. Through experimentation, we will generate comparative evidence about the effects of AI assistance on content quality, editor behavior, and workflow efficiency. These findings will be shared through peer-reviewed publications and community forums, with particular attention to how tool use aligns (or conflicts) with Wikipedia's core values.

Risks

This project involves studying a fast-evolving technology in a complex, volunteer-driven ecosystem. As such, risks are inevitable. We believe our team is generally prepared to deal with the upcoming challenges of doing research in this environment. Our projected research team covers several languages (English, Portuguese, Spanish, French, Italian, and Korean), and one of the Co-PIs has experience working on large, multi-lingual projects in Wikipedia [40, 41]. Further, we also have expertise in the proposed methodological approaches (e.g., participatory design [39], experiments involving online communities [31], AI-detection [17, 35], and observational studies [40]).

In that context, we outline the primary risks we anticipate and the steps we will take to mitigate them.

Uncertainty in Detecting LLM-Generated Content. A core challenge of this proposal lies in reliably identifying LLM-generated content. LLM-generated content is often indistinguishable from human-written text, and existing detection tools are imperfect and error-prone. This poses a challenge for drawing precise conclusions about the prevalence or impact of LLM-generated edits.

Mitigation: We will use a multi-method approach that combines probabilistic detection tools, temporal editing patterns, and content-based features. We will also transparently communicate uncertainty and validate our methods with human-labeled samples where possible. Note that this methodology has worked in other domains such as crowd work [17] and peer reviews [35].

Community Trust and Participation. The success of our survey, tool design, and field experiments depends heavily on community engagement. Some Wikipedians may be skeptical of AI, resistant to research, or wary of interventions that could affect platform norms.

Mitigation: We will take a participatory approach, co-creating the survey and tools with community members and soliciting feedback at all stages. All interventions will be opt-in, and we will consult with relevant community bodies. Transparency, responsiveness, and respect for community values will guide our process.

Rapidly Changing AI Landscape. Advances in LLM capabilities may outpace our research timeline or alter the relevance of specific tools or questions.

Mitigation: We will remain flexible in our research focus and prioritize generalizable insights over platform-specific dependencies. We can continuously adapt our work to meet emerging needs by working closely with the Wikipedia community.

Community impact plan

This project is designed to have an impact beyond the academic community. Given the central role of the Wikipedia volunteer ecosystem in shaping the platform, editors, organizers, developers, and affiliates are essential collaborators—not just stakeholders. We aim to co-create tools, norms, and knowledge that can inform real-world practice.

Wikipedia is a *vibrant, ever-changing* community; therefore, we will have a diversified strategy for broad engagement. Our plan includes:

Directly reaching out to Wikipedians interested in related topics. We recognize the importance of directly engaging Wikipedians who have previously expressed interest in AI, editing tools, or knowledge integrity. To facilitate this, we will proactively identify and contact relevant Wikiprojects,[9] editor groups, and individuals who have been active in relevant discussions or tool development.[10] We aim to cultivate a dedicated group of informed collaborators whose expertise can help shape our research agenda and ensure its alignment with community needs and values.

Continuously seeking feedback from the community. Wikipedia already has a central hub for discussion (Village Pump), which includes a section for brainstorming ideas with the editors. [11] To maintain transparency and sustain broad community engagement, we will regularly post detailed updates and solicit feedback. We expect this will help our work evolve in a way that respects community norms and expectations.

Engaging with researchers from WMF. The proposed research requires direct collaboration with researchers from Wikimedia Foundation. This will help us better engage with the Wikipedia community (e.g., using their infrastructure) and scale and potentially deploy tools that have been developed. Additionally, some of the proposed research could greatly benefit from knowledge and infrastructure operated by WMF (e.g., running surveys on Wikipedia).

"Building in public." Finally, we commit to "building in public," adopting an open-by-default approach in all stages of our research and tool development process. All

[9] E.g., the AI Cleanup WikiProject https://w.wiki/9GCL
[10] E.g., editors here: https://w.wiki/DYZ2
[11] https://en.wikipedia.org/wiki/Wikipedia:Village_pump

code, datasets (anonymized where necessary), methodological details, and results will be made publicly accessible via dedicated pages on Meta-Wiki and GitHub repositories. We will document design decisions transparently, share openly about both successes and failures, and actively welcome external contributions. This openness aligns with Wikipedia's collaborative ethos and may foster broader community participation.

Evaluation

We propose to evaluate the project's success by considering community impact, scholarly impact, and the production of research artifacts aligned with Wikimedia's strategic goals. We propose three evaluation directions.

1. Scholarly Impact. Peer-reviewed publications in high-impact venues detailing generalizable findings regarding AI's role in Wikimedia. Impact can be evaluated here by assessing the contributions' quality and impact on subsequent research (e.g., as evidenced through citations and follow-up work).

2. Practical Relevance. Whether insights and knowledge generated through this research have informed decisions or policies created by Wikimedia or the Wikipedia community. Impact may be evaluated here by tracking mentions to research artifacts produced across discussions and policy by the Wikipedia community or the Wikimedia Foundation.

3. Dataset and Tools. Public release of well-documented datasets, codebases, and tools for human-AI collaboration. Impact here can be evaluated by the reuse or citation by independent researchers or Wikimedia and by the extent to which the Wikipedia community uses and builds upon these resources.

Budget

We ask for partial support for postdoc salaries and one of the Co-PIs' summer salaries. These resources ($143,201.94 with overhead) would be spent over two years, between September 1st, 2025, and August 31st, 2027.

Detailed budget: 🟩 Budget sheet

(https://docs.google.com/spreadsheets/d/1IXM40CVGhyRvJFrDSLGCsDBEsVuh5d-xrLDcspYR5Jc/edit?usp=sharing)

References

[1] Lyu, L., Siderius, J., Li, H., Acemoglu, D., Huttenlocher, D., & Ozdaglar, A. (2025). Wikipedia Contributions in the Wake of ChatGPT. arXiv preprint arXiv:2503.00757.

[2] Vetter, M. A., Jiang, J., & McDowell, Z. J. (2025). An endangered species: how LLMs threaten Wikipedia's sustainability. AI & SOCIETY, 1-14.

[3] Thompson, B., Dhaliwal, M., Frisch, P., Domhan, T., & Federico, M. (2024, August). A shocking amount of the web is machine-translated: Insights from multi-way parallelism. In Findings of the Association for Computational Linguistics ACL 2024 (pp. 1763-1775).

[4] Judkis, M. (2024, June 30). The deluge of bonkers AI art is literally surreal. The Washington Post. https://www.washingtonpost.com/style/of-interest/2024/06/30/ai-art-facebook-slop-artificial-intelligence/

[5] Burtch, Gordon, Dokyun Lee, and Zhichen Chen. "The consequences of generative AI for online knowledge communities." Scientific Reports 14.1 (2024): 10413.

[6] Brooks, C., Eggert, S., & Peskoff, D. (2024). The Rise of AI-Generated Content in Wikipedia. In L. Lucie-Aimée, A. Fan, T. Gwadabe, I. Johnson, F. Petroni, & D. van Strien (Eds.), Proceedings of the First Workshop on Advancing Natural Language Processing for Wikipedia (pp. 67–79). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.wikinlp-1.12

[7] Huang, Siming, et al. "Wikipedia in the Era of LLMs: Evolution and Risks." arXiv preprint arXiv:2503.02879 (2025).

[8] Ashkinaze, J., Guan, R., Kurek, L., Adar, E., Budak, C., & Gilbert, E. (2024). *Seeing Like an AI: How LLMs Apply (and Misapply) Wikipedia Neutrality Norms* (No. arXiv:2407.04183). arXiv. https://doi.org/10.48550/arXiv.2407.04183

[9] Reverberi, Carlo, et al. "Experimental evidence of effective human–AI collaboration in medical decision-making." Scientific reports 12.1 (2022): 14952.

[10] Weber, Thomas, et al. "Significant productivity gains through programming with large language models." Proceedings of the ACM on Human-Computer Interaction 8.EICS (2024): 1-29.

[11] Vaccaro, Michelle, Abdullah Almaatouq, and Thomas Malone. "When combinations of humans and AI are useful: A systematic review and meta-analysis." Nature Human Behaviour (2024): 1-11.

[12] Feng, K. J., et al. "Cocoa: Co-Planning and Co-Execution with AI Agents." arXiv preprint arXiv:2412.10999 (2024).

[13] Kim, T. S., Choi, D., Choi, Y., & Kim, J. (2022, April). Stylette: Styling the web with natural language. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (pp. 1-17).

[14] Kang, Hyeonsu B., et al. "Synergi: A mixed-initiative system for scholarly synthesis and sensemaking." Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. 2023.

[15] Feal, Alvaro, Jeffrey Gleason, Pranav Goel, Jason Radford, Kai-Cheng Yang, John Basl, Michelle Meyer, David Choffnes, Christo Wilson, and David Lazer. "Introduction to National Internet Observatory." (2024).

[16] Wang, A., Morgenstern, J., & Dickerson, J. P. (2025). Large language models that replace human participants can harmfully misportray and flatten identity groups. Nature Machine Intelligence, 1-12.

[17] Veselovsky, V., Horta Ribeiro, M., Cozzolino, P. J., Gordon, A., Rothschild, D., & West, R. (2023). Prevalence and prevention of large language model use in crowd work. Communications of the ACM.

[19] Halfaker, Aaron, and R. Stuart Geiger. "Ores: Lowering barriers with participatory machine learning in Wikipedia." Proceedings of the ACM on Human-Computer Interaction 4.CSCW2 (2020): 1-37.

[20] Sakota, Marija, Maxime Peyrard, and Robert West. "Descartes: generating short descriptions of wikipedia articles." Proceedings of the ACM Web Conference 2023. 2023.

[21] Swarts, J. (2009, October). The collaborative construction of" fact" on Wikipedia. In Proceedings of the 27th ACM International Conference on Design of Communication (pp. 281-288).

[22] Lai, V., Carton, S., Bhatnagar, R., Liao, Q. V., Zhang, Y., & Tan, C. (2022, April). Human-ai collaboration via conditional delegation: A case study of content moderation. In Proceedings of

the 2022 CHI Conference on Human Factors in Computing Systems (pp. 1-18).

[23] Mackeprang, M., Müller-Birn, C., & Stauss, M. T. (2019). Discovering the sweet spot of human-computer configurations: A case study in information extraction. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1-30.

[24] Amershi, S., Cakmak, M., Knox, W. B., & Kulesza, T. (2014). Power to the people: The role of humans in interactive machine learning. AI magazine, 35(4), 105-120.

[25] Zhang, Y., Liao, Q. V., & Bellamy, R. K. (2020, January). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In Proceedings of the 2020 conference on fairness, accountability, and transparency (pp. 295-305).

[26] Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2024). Can large language models transform computational social science?. Computational Linguistics, 50(1), 237-291.

[27] Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. Proceedings of the National Academy of Sciences, 120(30), e2305016120.

[28] Shavit, Y., Agarwal, S., Brundage, M., Adler, S., O'Keefe, C., Campbell, R., ... & Robinson, D. G. (2023). Practices for governing agentic AI systems. Research Paper, OpenAI.

[29] Hans, A., Schwarzschild, A., Cherepanova, V., Kazemi, H., Saha, A., Goldblum, M., ... & Goldstein, T. (2024). Spotting llms with binoculars: Zero-shot detection of machine-generated text. arXiv preprint arXiv:2401.12070.

[30] Oulasvirta, A., & Hornbæk, K. (2016, May). HCI research as problem-solving. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (pp. 4956-4967).

[31] Horta Ribeiro, M., West, R., Lewis, R., & Kairam, S. (2024). Post Guidance for Online Communities. *arXiv preprint arXiv:2411.16814*.

[32] Reeves, N., Yin, W., & Simperl, E. (2024). Exploring the Impact of ChatGPT on Wikipedia Engagement. arXiv preprint arXiv:2405.10205.

[33] Lee, M., Gero, K. I., Chung, J. J. Y., Shum, S. B., Raheja, V., Shen, H., ... & Siangliulue, P. (2024, May). A design space for intelligent and interactive writing assistants. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (pp. 1-35).

[34] Ruprechter, T., Ribeiro, M. H., West, R., & Helic, D. (2023). Protection from Evil and Good: The Differential Effects of Page Protection on Wikipedia Article Quality. arXiv preprint arXiv:2310.12696.

[35] Latona, G. R., Ribeiro, M. H., Davidson, T. R., Veselovsky, V., & West, R. (2024). The AI review lottery: Widespread AI-assisted peer reviews boost paper scores and acceptance rates. arXiv preprint arXiv:2405.02150.

[36] Lee, M., Liang, P., & Yang, Q. (2022, April). Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In Proceedings of the 2022 CHI conference on human factors in computing systems (pp. 1-19).

[37] Friedman, B., Kahn, P. H., Borning, A., & Huldtgren, A. (2013). Value sensitive design and information systems. Early engagement and new technologies: Opening up the laboratory, 55-95.

[38] Spinuzzi, C. (2005). The methodology of participatory design. Technical communication, 52(2), 163-174.

[39] Calacci, Dana, Varun Nagaraj Rao, Samantha Dalal, Catherine Di, Kok-Wei Pua, Andrew Schwartz, Danny Spitzberg, and Andrés

Monroy-Hernández. "FairFare: A Tool for Crowdsourcing Rideshare Data to Empower Labor Organizers." arXiv preprint arXiv:2502.11273 (2025).

[40] Ruprechter, Thorsten, et al. "Volunteer contributions to Wikipedia increased during COVID-19 mobility restrictions." Scientific reports 11.1 (2021): 21505.

[41] Ribeiro, Manoel Horta, Kristina Gligorić, Maxime Peyrard, Florian Lemmerich, Markus Strohmaier, and Robert West. "Sudden attention shifts on wikipedia during the covid-19 crisis." In Proceedings of the International AAAI Conference on Web and Social Media, vol. 15, pp. 208-219. 2021.

[42] Petroni, Fabio, Samuel Broscheit, Aleksandra Piktus, Patrick Lewis, Gautier Izacard, Lucas Hosseini, Jane Dwivedi-Yu et al. "Improving Wikipedia verifiability with AI." *Nature Machine Intelligence* 5, no. 10 (2023): 1142-1148.

[43] Moás, Pedro Miguel, and Carla Teixeira Lopes. "Automatic Quality Assessment of Wikipedia Articles—A Systematic Literature Review." ACM Computing Surveys 56, no. 4 (2023): 1-37.