WikipedAI: Investigating AI Collaboration and Conflict in Open Knowledge Systems

Patrick Gildersleve
University of Exeter, UK

Abstract

As generative AI systems, particularly large language models (LLMs), become an increasingly important part of the information environment, new questions arise about their behaviour in collective or collaborative settings—especially when multiple AI agents interact and contribute to shared knowledge or decision-making processes. Wikipedia is a rich site for exploring these dynamics, as a high-visibility collaborative platform where AI contributions are already emerging**.** While prior research has examined the extractive use of Wikipedia by LLMs, little is known about how LLM-supported and LLM-based editing may affect content quality, community norms, or the collaborative process. This project proposes analysis of a controlled simulation of LLM agents editing a sandbox Wikipedia. Through analysis of these AI–AI editing dynamics, we will identify emergent patterns of collaboration and conflict, evaluate their effects on article content, and assess their implications for governance, editorial workflows, and knowledge integrity. Outputs include a peer-reviewed article, public dataset, analytical tools for the "WikipedAI" platform, and recommendations for responsible AI integration. Beyond Wikipedia, this work lays the groundwork for understanding the risks and opportunities in how AI agents might shape shared digital knowledge.

Introduction

Generative AI, particularly large language models (LLMs), are increasingly capable of contributing encyclopaedic knowledge to platforms such as Wikipedia (Shao et al., 2024). While prior research focuses on LLMs' extractive use of Wikipedia (Wagner and Jiang, 2025), and the potential negative effects on Wikipedia traffic and editing (Lyu et al., 2025), less is known about the nature of LLM editing—especially in how such agents may interact with each other.

Wikipedia has an established history of simple bot editing to assist human editors (Zheng et al., 2019), which LLM-based bots (and LLM-supported humans) could contribute to. However, their growing presence brings new challenges. As LLM editing is increasingly used—often without clear disclosure or detectability—there are risks of emergent behaviours beyond the well-documented issues seen in single-agent systems (Weidinger et al., 2022). It is not known how well these agents might collaborate, or whether they might slip into patterns of automated conflict, bias reinforcement, and content homogenisation.

Given Wikipedia's role as a key primary source of training data for LLMs (Vetter et al., 2025), there is an urgent need to understand feedback loops that could reverberate across a variety of information systems: such as increased human

editor labour, wasting of resources, model collapse, recursive misinformation, or loss of epistemic diversity.

This project addresses these issues by examining autonomous LLM editing interactions within a controlled, simulated environment. Specifically, it asks:

- RQ1: How do LLM bots interact when working to edit Wikipedia?
- RQ2: How does LLM editing activity affect Wikipedia article content (e.g., on quality, sourcing, neutrality)?
- RQ3: How does LLM editing behaviour compare to human / traditional bot editing patterns?

To answer these questions on the nature of AI editing and collaboration we will conduct a simulation of Wikipedia editing using a private MediaWiki installation as our experimental platform. This sandbox wiki will contain a curated sample of English Wikipedia articles, where many LLM agents will be deployed as autonomous editors on this platform. We will instrument the system to log every action (edits, reverts, content differences, etc.) and use a suite of analytical techniques to identify patterns in the bots' interactions.

This work will provide critical insights into the emergent dynamics and potential risks associated with multi-agent AI editing, informing Wikipedia community guidelines and Wikimedia governance strategies. Additionally, by openly sharing data, tools, and findings, the project will support broader research on human–AI collaboration, promoting responsible AI use within and beyond Wikipedia. As generative AI systems increasingly interact with humans and one another across platforms, understanding how such interactions shape knowledge, reinforce bias, or trigger feedback loops will be essential for safeguarding not just Wikipedia, but the wider information ecosystem.

Date: 01/10/2025–30/09/2026

Related work

Background

This project approaches Wikipedia as a hybrid socio-technical system (Trist et al., 2016) in which human editors and AI agents—such as LLM-based bots—interact as part of a shared social infrastructure. As such, it may draw from several influential frameworks such as actor-network theory (Latour, 2005), social machines (Hendler and Berners-Lee, 2010), and human–machine networks (Eide et al., 2016; Tsvetkova et al., 2017). More recent work on hybrid collective intelligence (Peeters et al., 2021) and sociology of humans and machines (Tsvetkova et al., 2024) considers the role of 'intelligent' agents such as LLMs. These perspectives provide a foundation for analysing Wikipedia not just as a platform mediated by AI, but as a site of emerging forms of collective human and machine decision making. With the advancement of machine capabilities, the boundaries between human-led and machine-led knowledge production are becoming increasingly porous, raising urgent questions about how such hybrid systems should be designed, governed, and evaluated.

Bot–Bot and Bot–Human Interactions on Wikipedia

The Wikipedia community has extensive experience managing automated editors (bots) through the Bot Approvals Group (Wikipedia contributors, n.d.-a), offering some foundation for anticipating LLM behaviour. Prior research has documented bot–bot conflict, with "edit wars" even emerging among well-intentioned bots (Tsvetkova et al., 2017). Yet this characterisation of bot "conflict" is

disputed—Geiger and Halfaker (2017) suggesting no more than 4% of this activity constitutes true conflict. Bots also interact with humans in their activities, with Clément and Guitton (2015) finding that they are overall well accepted, but can elicit polarised reactions when engaging in "policing" behaviour. Zheng et al. (2019) examine the range of bot behaviour, creating a taxonomy based on unsupervised learning to help identify their different roles (generator, fixer, tagger, etc.) and analysing their varied impacts on newcomer editors. Together, these studies illustrate how Wikipedia's automated agents can exhibit emergent behaviours, which must be considered ahead of any introduction of more advanced AI.

Human editing

Bot activity is usually constrained to relatively simple tasks (e.g., fixing redirects, cleaning sources, editing categories). However, one anticipates that LLMs are able to make more sophisticated human-like edits, so it is instructive to consider analyses of human editing patterns. The social processes of Wikipedia editing among human editors have been extensively studied. Early work identified the value of multiple editors working on articles (Wilkinson and Huberman, 2007) especially when effective coordination is present, e.g., via talk pages or division of labour by role (Kittur and Kraut, 2008; Brandes et al. 2009; Liu and Ram, 2011). Diverse, even polarised, editor pools have also been shown to improve content (Sydow et al., 2017; Shi et al. 2019).

Wikipedia disputes remain common despite established norms and tools to handle them, making them a rich research area. For severe conflicts, or "edit wars" (often operationalised through regular mutual reverting), researchers have found distinct patterns of conflict, that are typically due to the actions of relatively few stubborn editors (Sumi et al., 2011; Yasseri et al., 2012). Early indicators in these edit wars may help predict the need for further editors or administrator intervention (Chhabra et al., 2020). Agent-based modelling studies have examined how consensus may be formed, with the quantity and knowledge of editors (Xu et al., 2008), level of "tolerance" among editors (Gandica et al., 2014), and presence of moderating neutral voices (Kalyanasundaram et al., 2015) all positively affecting faster consensus formation. It remains to be seen what human-like behaviour LLM-based and LLM-supported editors will exhibit, and whether new dynamics may emerge.

Wikipedia and generative AI

A growing body of research is investigating how generative AI and Wikipedia affect each other, though this is typically focused on the extractive use of Wikipedia and knock-on effects to its traffic and editing contributions. Warnings have been issued on the ethics and sustainability of LLM companies' relationship with Wikipedia (Vetter et al., 2025; Wagner and Jiang, 2025), with worries that users will increasingly turn towards LLMs, and away from Wikipedia, in information seeking, drying up the contributions that power the online encyclopaedia. Yet, Wikipedia remains a vital resource, both for LLMs as training data, but more importantly in and of itself as "sanctuary for the future of knowledge representation, championing representation and accessibility in the age of closed-system LLMs" (McDowell, 2024).

Early studies on Wikipedia's traffic and editing changes due to LLMs show mixed results. Reeves et al. (2024) find no evidence of a fall in engagement since consumer adoption of LLM products, whereas Lyu et al. (2025) do find decreases in editing and viewership across LLM specialist topics. AI-generated text presence is estimated between 1–5% across certain articles (Brooks et al., 2024; Huang et al., 2025), though

reliably detecting AI-generated text remains difficult.

These contributions raise questions about LLM content quality and adherence to community norms—Ashkinaze et al. (2024) find that LLMs largely fail to detect bias, exhibit model biases, and generally over-neutralise when they do detect bias, compared to human editors. Editors are also attempting to address LLM editing issues such as through WikiProject AI Cleanup (Wikipedia contributors, n.d.-b). Given these mixed early results, ongoing tracking of human and LLM related engagement is crucial to informing proactive governance and community resilience strategies for AI.

Multi-agent behaviour

Ensuring neutrality and quality is an area where AI alignment research intersects with wiki contexts. Techniques to align language models with human preferences and norms (e.g. via reinforcement learning from human feedback, or other constraints) are rapidly advancing (Griffith, 2013; Casper, 2023). When multiple AI agents interact, they can exhibit complex, emergent behaviours that go beyond their initial programming. Evidence from HCI and AI simulations shows both promising and cautionary examples. On the promising side, Park et al. (2023) demonstrated generative agents autonomously coordinating realistic social activities. This, together with other work (Talebirad and Nadiri, 2023) indicates that multi-agent LLM systems might achieve a form of collaboration, dividing tasks or responding to each other's actions in a manner reminiscent of human teamwork as utilisable on Wikipedia.

On the cautionary side, however, multi-agent systems may exacerbate the risks posed by single LLMs such as discrimination, information hazards, and misinformation (Weidinger et al., 2022). The issue of model collapse has also been raised: if generative models train on content produced by other generative models, errors and biases can amplify, and the overall quality of outputs deteriorates over time (Shumailov et al., 2024). AI-written content, if unchecked, could recursively homogenise and degrade the encyclopaedia's voice and LLM output.

Our project explicitly addresses these potential risks and opportunities. By creating a controlled, closed-loop simulation of multiple AI agents editing Wikipedia articles, we aim to directly observe whether emergent collaboration / conflict patterns or content convergence arise. Crucially, this will allow systematic assessment of how factors such as agent heterogeneity, prompt engineering, and governance mechanisms might mitigate or exacerbate negative outcomes. Ultimately, WikipedAI seeks to empirically evaluate the balance between the productive potential of multi-agent generative AI (e.g. increased productivity or around-the-clock maintenance) and the alignment risks posed by recursive AI interactions, thereby offering practical insights for responsible AI integration into and necessary guardrails for Wikipedia and other open knowledge platforms.

Methods

Planning

Discussions will first be held with Wikimedia representatives and editors involved in the WikiProject AI Cleanup and Bot Approvals Group, a continuation of those being conducted in the grant-writing phase. These interviews will be optional and not part of the formal analysis of the project, but will help guide the research setup so as to best address challenges faced by editors and the Wikimedia Foundation. These discussions at project start and key milestones will guide scenario selection, bot instruction

design, and identification of key content vulnerabilities or contested article types.

Simulation setup

A fresh MediaWiki environment will be set up on a secure server, seeded with a selection of forked Wikipedia content. This will be managed by the University of Exeter's Digital Humanities Lab and shared as read-only on the project website after completion. Initial experiments will determine the sampling approach from a Wikipedia dump to ensure diverse articles (topics, age, length, popularity). One approach would be a 0.1-1% sample of articles from English Wikipedia selected according to weights by page views. Whilst the focus initially will be on English Wikipedia, results and methods may generalise and extend to alternate languages, or multilingual editing. The wiki will be configured with revision history tracking and necessary LLM edit permissions, as well as, potentially, features such as edit summaries and talk pages for future use.

LLM agent specification

The number of agents and frequency of edits (relative to number of articles) will be controlled so as to approximate real editor interaction dynamics on Wikipedia. We will also investigate the dynamics of inter- and intra-model editing, assessing how models from different providers (OpenAI, Anthropic, Google, DeepSeek, Meta etc.) interact with each other. Prompt instructions will be consistent between models for the main analysis, and tested for stability to ensure result robustness (Barrie et al., 2024).

Initial tests will determine minimal, consistent prompt complexity. Further experiments may explore incorporating agent capabilities for memory, article navigation, or explicitly including editing guidelines. Most analysis will be conducted through observing bot interactions on a single WikipedAI platform. However, separate (smaller) MediaWiki environments may be set up to test the effects of different agent specifications. A full analysis of all the possibilities is not feasible in this project, hence one of the key outputs being the WikipedAI framework as a "model organism" for experimenting with different agent compositions and specifications. This will help with the design and evaluation of multi-agent LLM benchmarks oriented towards Wikimedia principles (Johnson et al., 2024) that may extend beyond Wikipedia.

Data collection and analysis

The MediaWiki platform will record revision history for every page, which is our primary data for analysis. From these logs we will extract: the sequence of edits for each article in each scenario, including which bot made the edit, a timestamp, and the diff (content added/removed); metadata like edit comments and revert flags; content metrics such as article length over time, number of citations, reading grade level, sentiment or bias indicators, etc., at each revision.

Analysing this editing behaviour relies on effective operationalisation of key concepts. Geiger and Halfaker (2017) categorise bot interactions according to specific tasks (e.g. category work, fixing double redirects), yet there is likely to be greater diversity in LLM editing interactions. The previously mentioned literature analysing bot and human editor interaction provides some inspiration here, and we propose four (non-exclusive) at this stage; Collaboration, Conflict, Convergence, and Cycling. Given the scale and controlled nature of interactions, we will focus on approaches at scale from sequence analysis, network science, and natural language processing. These will be complemented by qualitative review of examples of the characteristic editing patterns to ensure concept validity.

Operational definitions will be iteratively refined and carefully measured. Firstly, *collaboration* may be understood to be productive editing behaviour where contributions are additive and minimally reverted/erased. Second, *conflict* could be explicitly indicated through regular reverts or erasure of significant sections of content, without the article reaching a stable equilibrium. *Convergence* may be expressed by an article's rate of change slowing after the agents initially "fix" it. Finally, *cycling* may emerge if agents make regular, substantive, non-conflicting edits that do not lead to improvement or consistent trajectory in article content. To measure these, we may adopt multiple indicators, such as the size of the changes in content addition and deletion, article quality metrics, the changing position of the article in embedding space, or through analysing edit summaries.

An alternative to the freeform editing analysis based around an ostensibly *cooperative improvement* scenario is to design a narrower range of interaction scenarios. These would be inspired by those faced by editors and tested on smaller MediaWiki instances or even limited to single articles. For example, in a *conflicting objectives* scenario bots may be assigned opposing directives (e.g. inclusionist vs deletionist), and dynamics around escalating edit wars assessed. In a *bias and correction* scenario we may assess how bots with a pre-specified bias on a particular subject fare against "neutral" bot editors, and whether the biases or NPOV persist in the articles. Whilst the focus of the project is expected to be on assessing the emergent dynamics of LLMs engaging in good faith editing, these scenarios can provide useful testing environments ahead of the full analysis, helpful for identifying emergent behaviour (especially between models), as well promising avenues for further work that can be enabled by the WikipedAI framework (see expected output).

A challenge in this research is in relating the results to real-world unknown declaration rates around AI editors on real Wikipedia, making direct comparison difficult. As noted, AI-assisted edits are often indistinguishable from human edits, and there is currently no reliable tag that indicates LLM editing specifically. Nevertheless, we will attempt some triangulation with real-world observations. Firstly, we can compare results against those from historical declared bot interactions (e.g., on simple operationalisations of conflict like revert rates). Secondly, we can consult and compare against records of WikiProject AI Cleanup, where human editors have identified (and fixed) poor quality or damaging suspected AI-generated edits. The nature of the LLM editing patterns observed in our work may also help such projects identify other forms of AI generated editing not being spotted already.

Expected output

We expect to produce: at least one peer-reviewed article; a dataset of bot interactions; a website hosting an interactive dashboard of results as well as the end result of the LLM-edited Wikipedia simulation; training for a PhD (preferable) or Master's researcher; recommendations to the Wikimedia Foundation and Wikipedia community on how to address the challenges of AI editing; and finally, coding resources to allow researchers to extend our analysis on their own LLM / human editing problems—the WikipedAI platform.

We will aim to publish at a top web or social computing conference, such as CSCW, ICWSM, or ACM Web Science, focusing on the empirical findings of LLM bot interactions. This will be supported by dissemination at other non-archival conferences engaging the

Wikiresearch community such as IC2S2, Wiki Workshop, and Wikimania.

The detailed data from our simulation will be cleaned and released as an open dataset, hosted by a provider such as Zenodo. This may also facilitate a further publication as a dataset paper, for example, at ICWSM. The dedicated MediaWiki instance allows us to provide full editing interactions in the same format as the "revisions" information through the Wikipedia API. This will allow consistent comparison with typical editing data from Wikipedia. This dataset could be valuable for researchers in areas like multi-agent learning, AI alignment, or digital collaboration.

The project website, hosted for a minimum of 5 years post-project, will offer dual function. An interactive dashboard of project results for a public audience will be provided along with a blog-style explainer. This will be complemented by presenting the end result of the LLM edited articles in a format similar to Wikipedia (read-only, featuring clear disclaimers), to allow free form investigation of the effects of LLMs editing. The interactive website will allow the Wikimedia community and public to explore bot editing sessions, visualising key metrics such as reverts and contested text in a manner similar to Contropedia (Borra et al., 2014), and allow comparison with live Wikipedia articles via a diff tool.

This project will provide rich training and development opportunities for the graduate research assistant involved. The RA will be a full collaborator and named co-author, with the opportunity to attend and help present the research at conferences. Training this next cohort of researchers and integrating them into the Wikiresearch community is vital.

Based on the project results, recommendations for policy and practices for addressing AI editing challenges will be shared with the Wikimedia Foundation as well as directly with the editing community. This can be done through information sharing with / editing existing projects and help pages around AI editing, such as with the Bot Approvals Group or WikiProject AI Cleanup. These recommendations may help with identifying characteristics of AI editing previously undetected, determining appropriate responses in reverting / integrating AI (supported) edits, or finding (sections of) articles more susceptible to (damaging) changes by AI editing. Beyond Wikipedia, recommendations to AI Safety and Human–AI Collaboration researchers based on our findings will highlight emergent behaviours (perhaps unforeseen failure modes or alignment issues) that occur in multi-agent settings. This can inform safer design of autonomous social agents and how to curb negative interactions (like endless argument loops or misinformation amplification).

Finally, beyond the dataset provided and project analysis code, we will provide the necessary code, data, and documentation as a framework for further researchers to undertake their own work on. Within this project we are not able to run all possible forms of analysis of AI influence on Wikipedia, especially as models, prompting approaches, and multi-modal capabilities develop. However, in providing WikipedAI as a framework, future researchers can examine and compare various aspects of multi-agent LLM interactions in a controlled environment, as well as interactions with human editors. In this sense, we consider WikipedAI could act as a form of "model organism".

All outputs will be shared under an open license. We will actively disseminate them: the academic paper(s) through conferences and arXiv preprints; the dataset via Zenodo; the code on GitHub; and the interactive demo hosted by University of Exeter's dedicated Digital

Humanities Lab team. By the project's end, we expect not only to answer our research questions, but to provide the community with enduring resources to continue exploring and addressing the impact of AI on Wikipedia.

Risks

A key design consideration of this project is to conduct the LLM experiments in an isolated, sandboxed, version of Wikipedia so as not to negatively interfere with the live Wikipedia site. However, there remain several important ethical considerations. Though our wiki is private, we must consider the ethical implications of generating lots of content via AI. Ethical considerations include clearly communicating with all stakeholders the controlled nature of our sandbox experiments to prevent misinterpretation as live bot activity. Content will be reviewed internally in initial testing phases. When releasing the dataset, we'll include clear disclaimers that this was AI-generated and may contain errors or offensive text (despite our filters)—similar to how datasets of AI outputs are sometimes released with content warnings. By being transparent about goals (improving understanding to protect Wikipedia, not to advocate replacing editors with bots), we will manage reputational risk.

Unpredictable agent behaviour, whilst certainly of research interest to the project, may result in activity that jeopardises some forms of analysis, dramatically inflates API costs, or produces objectionable content. To mitigate this, we will begin testing with tightly constrained interactions, more explicit Wikipedia editing guidelines in prompts, and human oversight before progressive deployment of more advanced capabilities. We have budgeted appropriately (with a buffer) and will monitor usage. If certain scenarios turn out to be cost-heavy (e.g., long articles causing very large prompts), we might down-sample some content or use token-efficient prompting (summarizing context or limiting how much of the article a bot reads each turn) to control expenses. The dedicated Digital Humanities Lab IT support will manage technical setup and monitoring of the MediaWiki instance, to address any of these issues.

The controlled sandbox may fail to capture crucial elements of real Wikipedia editing and we will not be easily able to validate our findings against real LLM editing cases. Human editors bring nuanced judgment, and social interaction on Wikipedia involves discussion, consensus-building processes, and power structures (e.g. admin interventions) that our simulation may not fully replicate. Thus, outcomes in our experiment—whether cooperative or conflictual—might not directly map to live Wikipedia scenarios. We acknowledge this limitation and design the study to emphasize patterns and mechanisms rather than exact predictive outcomes. The fact that LLM-based and LLM-supported edits often go undetected motivates our proactive and precautionary study design. We will build our analysis on prior work on labelled simple bot editing, and community concerns through cases identified by WikiProject AI Cleanup.

Ultimately, the simulation is a necessary simplification to allow us to study AI–AI interaction in isolation. We will document the gaps and advise caution in over-generalising results. The inclusion of multiple model types and content types in our sample is another mitigation—by varying conditions, we reduce the risk that any single idiosyncrasy (say, a quirk of one model or one article) skews the overall findings.

Community impact plan

Our community impact plan focuses on transparency, collaboration with existing initiatives, and open knowledge sharing to ensure the project's outcomes benefit Wikimedia at large. Two editing groups whose insights can help shape the research and who can benefit from the project's recommendations are the WikiProject AI Cleanup and the Bot Approvals Group. During the project, we will ensure to update the MetaWiki Research page with project progress. By sharing our findings on how AI agents behave in editing, we can help WikiProject AI Cleanup refine their detection strategies for problematic AI edits, and understand how to anticipate responses in follow-up edits from AI (supported) editors. Similarly, the Bot Approvals Group's insights can guide the kind of emergent behaviours to watch out for, and our results will help indicate new and verify existing errant behaviour. Whilst we would not expect to recommend allowing unfettered LLM bot editing, the project's results would help indicate limited tasks (yet still more complex than traditional rules-based bots) where LLMs could be appropriately deployed, and the guardrails necessary to oversee and control their actions.

AI's potential impact necessitates proactive Wikimedia Foundation stewardship. This project's insights from and recommendations to the aforementioned editing groups can be adapted to support recommendations to the foundation, who may be able to establish new policy, implement editing assistance tools, or recommend guidelines for integrating / adapting to AI tools across different languages and projects.

As mentioned in the outputs section, engagement with editors, whether interested or not in details of Wikiresearch, is an important aspect of the project's impact. We plan dissemination at Wiki Workshop, Wikimedia Research Showcase, Wikimania, and The Signpost. The project website, with its interactive dashboard and browsable content will be advertised and can serve as a method for hands-on engagement for those within and beyond the Wikimedia community.

Evaluation

We will consider the project successful if it delivers the following outcomes:

1. Generation of valuable empirical insights: Evidence of distinct patterns of AI–AI interactions clearly operationalised and validated.
2. Quality and usability of analysis outputs: Production of the AI interaction dataset, the WikipedAI analysis framework, and their subsequent use by other researchers.
3. Relevance and impact of recommendations: The engagement with and uptake of project recommendations by Wikipedia editing communities and the Wikimedia Foundation.
4. Open-access dissemination and engagement: Presentation of findings at reputable conferences, complemented by community dissemination.

We hope that the Research Fund chairs evaluate the proposed project positively. We see clear project alignment with recommendations for the Wikimedia 2030 Movement Strategy; Increasing sustainability, Innovating in free knowledge, and to Evaluate, iterate, and adapt. The rapid growth of generative AI presents both significant opportunities and critical risks for Wikipedia's future. Proactive, evidence-based research like this project is essential to ensure

that technological change strengthens, rather than undermines, the participatory and open knowledge practices that define the Wikimedia movement.

Budget

https://docs.google.com/spreadsheets/d/1wzFoiZSExvxNwOQWjqqPGJBArr2k082LSW8HKYCUVy8/edit?usp=sharing

References

Barrie, C., Palaiologou, E., & TÃķrnberg, P. (2024). Prompt stability scoring for text annotation with large language models. *arXiv preprint arXiv:2407.02039*.

Borra, E., Weltevrede, E., Ciuccarelli, P., Kaltenbrunner, A., Laniado, D., Magni, G., ... & Venturini, T. (2014). Contropedia-the analysis and visualization of controversies in Wikipedia articles. *OpenSym*, *34*, 31.

Brandes, U., Kenis, P., Lerner, J., & Van Raaij, D. (2009, April). Network analysis of collaboration structure in Wikipedia. In *Proceedings of the 18th international conference on World wide web* (pp. 731-740).

Brooks, C., Eggert, S., & Peskoff, D. (2024, November). The Rise of AI-Generated Content in Wikipedia. In *Proceedings of the First Workshop on Advancing Natural Language Processing for Wikipedia* (pp. 67-79).

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., ... & Hadfield-Menell, D. (2023). Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*.

Chhabra, A., Kaur, R., & Iyengar, S. R. S. (2020, August). Dynamics of edit war sequences in Wikipedia. In *Proceedings of the 16th International symposium on open collaboration* (pp. 1-10).

Clément, M., & Guitton, M. J. (2015). Interacting with bots online: Users' reactions to actions of automated programs in Wikipedia. *Computers in Human Behavior*, *50*, 66-75.

Eide, A. W., Pickering, J. B., Yasseri, T., Bravos, G., Følstad, A., Engen, V., ... & Lüders, M. (2016). Human-machine networks: towards a typology and profiling framework. In *Human-Computer Interaction. Theory, Design, Development and Practice: 18th International Conference, HCI International 2016, Toronto, ON, Canada, July 17-22, 2016. Proceedings, Part I 18* (pp. 11-22). Springer International Publishing.

Gandica, Y., Dos Aidos, F. S., & Carvalho, J. (2014). The dynamic nature of conflict in Wikipedia. *Europhysics Letters*, *108*(1), 18003.

Geiger, R. S., & Halfaker, A. (2017). Operationalizing conflict and cooperation between automated software agents in wikipedia: A replication and expansion of'even good bots fight'. *Proceedings of the ACM on human-computer interaction*, *1*(CSCW), 1-33.

Griffith, S., Subramanian, K., Scholz, J., Isbell, C. L., & Thomaz, A. L. (2013). Policy shaping: Integrating human feedback with reinforcement learning. *Advances in neural information processing systems*, *26*.

Hendler, J., & Berners-Lee, T. (2010). From the Semantic Web to social machines: A research challenge for AI on the World Wide Web. *Artificial intelligence*, *174*(2), 156-161.

Johnson, I., Kaffee, L. A., & Redi, M. (2024, November). Wikimedia data for AI: a review of Wikimedia datasets for NLP tasks and AI-assisted editing. In *Proceedings of the First*

Workshop on Advancing Natural Language Processing for Wikipedia (pp. 91-101).

Kalyanasundaram, A., Wei, W., Carley, K. M., & Herbsleb, J. D. (2015, December). An agent-based model of edit wars in Wikipedia: How and when is consensus reached. In *2015 Winter Simulation Conference (WSC)* (pp. 276-287). IEEE.

Kittur, A., & Kraut, R. E. (2008, November). Harnessing the wisdom of crowds in wikipedia: quality through coordination. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work* (pp. 37-46).

Latour, B. (2005). *Reassembling the social: An introduction to actor-network-theory*. Oxford university press.

Liu, J., & Ram, S. (2011). Who does what: Collaboration patterns in the Wikipedia and their impact on article quality. *ACM Transactions on Management Information Systems (TMIS), 2*(2), 1-23.

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023, October). Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology* (pp. 1-22).

Peeters, M. M., Van Diggelen, J., Van Den Bosch, K., Bronkhorst, A., Neerincx, M. A., Schraagen, J. M., & Raaijmakers, S. (2021). Hybrid collective intelligence in a human–AI society. *AI & society, 36*, 217-238.

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

Shao, Y., Jiang, Y., Kanell, T. A., Xu, P., Khattab, O., & Lam, M. S. (2024). Assisting in writing wikipedia-like articles from scratch with large language models. *arXiv preprint arXiv:2402.14207*.

Shi, F., Teplitskiy, M., Duede, E., & Evans, J. A. (2019). The wisdom of polarized crowds. *Nature human behaviour, 3*(4), 329-336.

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature, 631*(8022), 755-759.

Sydow, M., Baraniak, K., & Teisseyre, P. (2017). Diversity of editors and teams versus quality of cooperative work: experiments on Wikipedia. *Journal of Intelligent Information Systems, 48*, 601-632.

Talebirad, Y., & Nadiri, A. (2023). Multi-agent collaboration: Harnessing the power of intelligent llm agents. *arXiv preprint arXiv:2306.03314*.

Trist, B., Murray, H., & Trist, E. (2016). Characteristics of socio-technical systems. In *The Social Engagement of Social Science, a Tavistock Anthology, Volume 2: The Socio-Technical Perspective* (pp. 157-186). University of Pennsylvania Press.

Tsvetkova, M., García-Gavilanes, R., Floridi, L., & Yasseri, T. (2017). Even good bots fight: The case of Wikipedia. *PloS one, 12*(2), e0171774.

Tsvetkova, M., Yasseri, T., Meyer, E. T., Pickering, J. B., Engen, V., Walland, P., ... & Bravos, G. (2017). Understanding human-machine networks: a cross-disciplinary survey. *ACM Computing Surveys (CSUR), 50*(1), 1-35.

Tsvetkova, M., Yasseri, T., Pescetelli, N., & Werner, T. (2024). A new sociology of humans and machines. *Nature Human Behaviour, 8*(10), 1864-1876.

Vetter, M. A., Jiang, J., & McDowell, Z. J. (2025). An endangered species: how LLMs threaten Wikipedia's sustainability. *AI & SOCIETY*, 1-14.
Wagner, C., & Jiang, L. (2025). Death by AI: Will large language models diminish Wikipedia?. *Journal of the Association for Information Science and Technology*.

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P. S., Mellor, J., ... & Gabriel, I. (2022, June). Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency* (pp. 214-229).

Wikipedia contributors. (n.d.-a). Wikipedia:Bot Approvals Group. Wikipedia. Retrieved April 4, 2025, from https://en.wikipedia.org/wiki/Wikipedia:Bot_Approvals_Group

Wikipedia contributors. (n.d.-b). *Wikipedia:WikiProject AI Cleanup - Wikipedia*. https://en.wikipedia.org/wiki/Wikipedia:WikiProject_AI_Cleanup

Wilkinson, D. M., & Huberman, B. A. (2007, October). Cooperation and quality in wikipedia. In *Proceedings of the 2007 international symposium on Wikis* (pp. 157-164).

Xu, J., Yilmaz, L., & Zhang, J. (2008, April). Agent simulation of collaborative knowledge processing in wikipedia. In *Proceedings of the 2008 Spring simulation multiconference* (pp. 19-25).

Zheng, L., Albano, C. M., Vora, N. M., Mai, F., & Nickerson, J. V. (2019). The roles bots play in Wikipedia. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1-20.