Exploring How AI Can Interpret Discussions Through an Insider Lens to Support Future Mediation

Soobin Cho
University of Washington

Mark Zachry
University of Washington

David W. McDonald
University of Washington

Abstract

Wikipedia relies on active discussions among contributors, but not all of them go smoothly—some stall, derail, or escalate. This project focuses on article talk pages, ultimately aiming to help these discussions move forward more effectively with the support of an AI mediator. To do so, the AI must first be able to interpret discussions through a Wikipedian lens, applying insider knowledge. We seek to examine what knowledge an AI needs to interpret Wikipedia discussions, how it can acquire that knowledge, and how contributors respond when it demonstrates such understanding. We will begin by constructing Wikipedia-specific knowledge to train the initial model. Next, we will develop an interface where Wikipedians can test the AI's interpretations on actual discussions. Through iterative user interviews using the interface, we will gather feedback and refine the model to better reflect Wikipedians' perspectives. Our ultimate goal—supporting more effective discussions—benefits the entire Wikipedia community and aligns with Wikimedia's 2030 goals of knowledge equity and improved user experience. This work also offers practical insight into AI training in community contexts, informing broader research in CSCW, HCI, and human–AI interaction.

Introduction

What is the problem that you aim to solve with this proposal?

Wikipedia relies on countless discussions among community members. But not all of them go smoothly—some stall, derail, or escalate. While this can happen in many spaces, our focus is on article talk pages. These spaces are closely tied to the encyclopedia's purpose and bring together contributors with diverse perspectives and experiences.

Our ultimate goal is to help these discussions move forward more effectively—with the support of an AI mediator. But without a deep understanding of Wikipedia, an AI can never effectively mediate discussions among editors. The first step in building such a mediator is training it to use insider knowledge to interpret conversations through a Wikipedian lens.

This insider knowledge can be very broadly divided into three categories: article content, user-to-user interaction, and individual users. Article content includes knowledge about what can and cannot be included in an article, how content should be written, what editing practices are considered appropriate, etc; User-to-user interaction involves knowledge about the expected conduct during discussions, the meaning of consensus, how disputes are typically resolved, etc; Individual users refer to more dynamic, meta-level knowledge about editors themselves, implications of different

experience levels, norms related to user accounts, underlying social structures, etc.

Since covering all of these in a year is not feasible, our research will focus on the first two: article content and user-to-user interaction. These areas are extensively documented through policies, guidelines, and essays—such as those on behavior, content, deletion, editing, and style—making them a solid starting point for developing an AI capable of learning and applying insider Wikipedian knowledge.

What are the specific research questions you want to address in this proposal?

RQ1. What kinds of knowledge related to article content and user-to-user interaction are essential for an AI to understand Wikipedia discussions?

RQ2. What training strategies or representational structures can support an AI in learning such knowledge effectively?

RQ3. How do Wikipedians perceive and react to an AI and its responses when it demonstrates an understanding of Wikipedia norms?

Why is addressing the problem important for Wikimedia projects?

Helping the discussions proceed—rather than turning into harsh disputes or getting stuck—is important to administrators, editors, and article quality alike.

For administrators, a mediator can help discussions move forward before they escalate to the point of requiring arbitration or direct moderation—outcomes many admins would prefer to avoid. For editors, it can be discouraging when their active participation results only in stalled conversations or conflicts, potentially leading them to leave the community. For newcomers, discussions are more inviting when a mediator or facilitator is present. For the community as a whole, well-facilitated discussions contribute to higher-quality articles.

This also directly supports Wikimedia's 2030 Strategic Direction of Knowledge Equity, which emphasizes welcoming contributors from diverse backgrounds and removing social, political, and technical barriers to equitable participation. It also aligns with the Wikimedia 2030 Movement Strategy Recommendations, particularly Improve User Experience, which calls for inclusive platform design that supports positive experiences for all users, regardless of technical skill or experience. By proactively facilitating discussions, we aim to create a space where editors—especially those who are less experienced or bring diverse perspectives—can participate more comfortably and constructively.

Date: October 1, 2025 – October 1, 2026

Related work

Understanding Wikipedia discussions

Many disagreements occur on Wikipedia, and this is actually fundamental to how the platform operates. Some prior studies even suggest that Wikipedia is more oriented toward disagreement than collaboration itself [4, 5]. These disagreements are constructive, even when they escalate into heated conflicts [9, 10], as productive rebuttals—such as counter arguments and refutations—are linked to improved outcomes [10]. In this sense, disagreement-based discussions are important and valuable, as long as they don't escalate or derail to the point of requiring moderation. *This is why mediating and facilitating discussions is so important.*

However, understanding discussions on Wikipedia requires an insider's perspective and a strong grasp of how the community operates. Wikipedia policies, which serve as a framework for collaboration through shared language and standardized strategies [2, 11, 21], have been linked to reduced conflict in discussions [8] and are frequently referenced. Studies of AfD

discussions show that policy-based arguments are common, and strong arguments often reflect knowledge of both policies and community values [17, 18]. Similarly, research on article talk pages finds that references to Wikipedia guidelines are frequent [15, 16, 22], enough to reflect community concerns and work patterns over the long term [2]. Even when arguments don't explicitly cite policy, they gain strength when appealing to community values [13]. Our previous study on how Wikipedians read and interpret talk page discussions similarly found that they often focused on community-specific elements—nuances that were frequently missed in general summaries generated by LLMs. Taken together, these findings suggest that understanding Wikipedia discussions requires more than reading the words—it involves interpreting them through the lens of a Wikipedian.

Online discussion facilitation

Many studies have aimed to improve online discussions, but few have explored mediation and facilitation through agents—like certain advanced Wikipedia bots—that can take part in community communication and possess community-specific knowledge. The most studied method for helping users quickly and easily make sense of an online discussion is summarization [1, 14, 23]. Other approaches include organizing comments into topics [3], displaying opinions along a pro–con spectrum [12], and visualizing relationships between discussants [20]. However, these methods do not engage in active mediation. More importantly, they focus on structuring or visualizing discussion content itself, rather than interpreting it in relation to the broader community context.

In online chat environments, some studies have attempted to facilitate discussion using conversational agents. However, these agents have typically focused on structured turn-taking, encouraging participation, or managing time [6, 7], or on supporting interpersonal familiarity and intimacy [19]. In other words, they were not designed to be group- or community-aware, nor to mediate discussions based on an understanding of the community.

Taken together, previous research highlights the importance of understanding Wikipedia discussions from an insider's perspective, yet existing approaches rarely incorporate such perspectives when supporting or facilitating online discussions. To address this gap, we explore how AI might be trained to interpret Wikipedia discussions from a Wikipedian's point of view.

Methods

Phase 1. Knowledge construction and AI training

We will begin by constructing a dataset related to article content and user-to-user interaction based on three sources: 1) insights from our two previous studies (Wikimedia research project page (1), Wikimedia research project page (2)), 2) Wikipedia policies, guidelines, and essays, and 3) prior research.

The dataset will need to be structured and labeled in a way that we can feed it into an LLM to create a "fine tuned" LLM model. Fine-tuning is currently one way to improve an LLM's response performance around specialized topics, such as Wikipedia insider knowledge. One of the structuring issues for policies is that policies are often quite complex; they are multifaceted. That is, there are often several dimensions for policy. Our current thinking is that we may have to leverage both LLM and by-hand structuring and labeling to be able to convert policies into a form that can be used for fine tuning.

Once we have a fine tuned LLM model we will create a small test set of example administrator decisions. The test set will include

text of discussions and any determinations or decisions made with regard to those discussions. The test set will be used to test and validate insider knowledge or decisions from prompting or questions of the LLM.

Phase 2. Prototype development for AI testing
We currently have a web-based prototype to help with the sensemaking of Wikipedia talk page discussions. It displays a Wikipedia discussion thread on the left side of the screen and offers supporting tools on the right.

Building on this, we will develop a new prototype to test the AI. The interface will similarly show a Wikipedia discussion thread on the left, while the right side will display the AI's interpretation and explanation of the discussion. Testers will also be able to directly interact with the AI through the right-side panel.

Phase 3. Interviews for AI evaluation
Following the approach used in our previous studies, we will recruit experienced Wikipedia administrators and editors using the "Email this user" feature on user pages and by posting our research invitation on a Wikimedia research project page. We will identify potential participants by reviewing their user contributions and ensuring they have substantial experience with editing and talk page discussions.

Through semi-structured interviews, participants will be asked to use the prototype for actual discussions and evaluate how well the AI understands Wikipedia norms and whether its interpretations were appropriate or inappropriate in context. Example interview questions include: Did the AI appear to understand relevant norms? Did its explanation help you better grasp the norm? Did it help you better understand the discussion?

We will qualitatively analyze the interview data to identify cases where the AI misinterpreted or misapplied Wikipedia norms, as well as additional context that AI may need to learn.

Phase 4. Feedback-driven refinement and iteration
Based on the feedback from Phase 3, we will revise and improve the AI system. We will then repeat the testing process with additional participants, engaging in iterative refinement and evaluation until the AI reaches a sufficient level of quality.

Expected output

1) Insights to guide the training of insider AI
For AI and Wikipedia researchers, we offer practical strategies for building AI systems tailored to community-centered platforms like Wikipedia.

2) Prototype of an AI tool that understands Wikipedia talk page discussions
For Wikipedia administrators and editors, this can provide AI support in better and quickly understanding and interpreting discussions.

3) Scientific publications in peer-reviewed journals and conferences (e.g., CSCW, CHI)
For HCI, CSCW, and Human-AI interaction research communities, we contribute to research on AI for online communities.

Risks

1) Community acceptance
We address this by using an iterative approach—testing the model in small steps, closely observing community reactions, and refining the system based on feedback before wider deployment.

2) Technical limitations
To mitigate this, we will experiment with different language models and fine-tuning

strategies to identify the most reliable setup for our goals.

Community impact plan

We aim to impact the Wikipedia community by working directly with experienced administrators and editors. Through semi-structured interviews, they will interact with the AI prototype and provide feedback on how well it reflects community norms and supports discussion understanding. This process ensures that the system is shaped by the people most familiar with Wikipedia's cultural practices.

We also plan to share updates through a public project page and use Wikimedia community channels such as the Village Pump, as well as other venues like Wikimania conferences and local Wikipedia meetups, to invite broader input and increase transparency. These efforts will ensure that its development remains aligned with community values.

Evaluation

Success will be measured by:

- Whether experienced Wikipedia editors and administrators find the AI's interpretations accurate, useful, and aligned with Wikipedia norms.
- The system's improvement over iterative cycles, based on participant feedback.
- The production of generalizable insights that inform future development of AI systems for community-driven platforms.
- Engagement and interest from both academic and Wikimedia communities, as reflected through scholarly attention, feedback, and participation.

These criteria will allow us and others to assess the project's contribution to both research and practice.

Budget

https://docs.google.com/spreadsheets/d/19_hYEceFxd6967goHC_rxtMLMZVAQv4ofTwH8wbgldM/edit?usp=sharing

- Ph.D. student support: $46,298.00
- Participant compensation: $1,200.00
- Software costs: $100.00
- Total: $47,598.00

References

1. Lucas Anastasiou and Anna De Liddo. 2021. Making Sense of Online Discussions: Can Automated Reports help?. In Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (Yokohama, Japan) (CHI EA '21). Association for Computing Machinery, New York, NY, USA, Article 471, 7 pages. https://doi.org/10.1145/3411763.3451815
2. Ivan Beschastnikh, Travis Kriplean, and David McDonald. 2008. Wikipedian Self-Governance in Action: Motivating the Policy Lens. Proceedings of the International AAAI Conference on Web and Social Media 2 (09 2008), 27–35. https://doi.org/10.1609/icwsm.v2i1.18611
3. Enamul Hoque and Giuseppe Carenini. 2014. ConVis: A Visual Text Analytic System for Exploring Blog Conversations. Computer Graphics Forum 33 (06 2014). https://doi.org/10.1111/cgf.12378
4. Dariusz Jemielniak. 2014. Common Knowledge? An Ethnography of Wikipedia. Stanford University Press, Standford, California.

5. Dariusz Jemielniak and Andreea Gorbatai. 2012. Power and status on Wikipedia. (2012).
6. Soomin Kim, Jinsu Eun, Changhoon Oh, Bongwon Suh, and Joonhwan Lee. 2020. Bot in the Bunch: Facilitating Group Chat Discussion by Improving Efficiency and Participation with a Chatbot. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376785
7. Soomin Kim, Jinsu Eun, Joseph Seering, and Joonhwan Lee. 2021. Moderator Chatbot for Deliberative Discussion: Effects of Discussion Structure and Discussant Facilitation. Proc. ACM Hum.-Comput. Interact. 5, CSCW1, Article 87 (April 2021), 26 pages. https://doi.org/10.1145/3449161
8. Aniket Kittur and Robert E. Kraut. 2010. Beyond Wikipedia: coordination and conflict in online production groups. In Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work (Savannah, Georgia, USA) (CSCW '10). Association for Computing Machinery, New York, NY, USA, 215–224. https://doi.org/10.1145/1718918.1718959
9. Christine Kock and Andreas Vlachos. 2021. I Beg to Differ: A study of constructive disagreement in online conversations. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2017–2027. https://doi.org/10.18653/v1/2021.eacl-main.173
10. Christine Kock and Andreas Vlachos. 2022. How to disagree well: Investigating the dispute tactics used on Wikipedia. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 3824–3837. https://doi.org/10.18653/v1/2022.emnlp-main.252
11. Travis Kriplean, Ivan Beschastnikh, David W. McDonald, and Scott A. Golder. 2007. Community, consensus, coercion, control: cs*w or how policy mediates mass participation. In Proceedings of the 2007 ACM International Conference on Supporting Group Work (Sanibel Island, Florida, USA) (GROUP '07). Association for Computing Machinery, New York, NY, USA, 167–176. https://doi.org/10.1145/1316624.1316648
12. Travis Kriplean, Jonathan Morgan, Deen Freelon, Alan Borning, and Lance Bennett. 2012. Supporting reflective public thought with considerit. In Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (Seattle, Washington, USA) (CSCW '12). Association for Computing Machinery, New York, NY, USA, 265–274. https://doi.org/10.1145/2145204.2145249
13. Jonathan T. Morgan, Robert M. Mason, and Karine Nahon. 2011. Lifting the veil: the expression of values in online communities. In Proceedings of the 2011 IConference (Seattle, Washington, USA) (iConference '11). Association for Computing Machinery, New York, NY, USA, 8–15. https://doi.org/10.1145/1940761.1940763
14. Kevin K. Nam and Mark S. Ackerman. 2007. Arkose: reusing informal information from online discussions. In Proceedings of the 2007 ACM International Conference on Supporting

Group Work (Sanibel Island, Florida, USA) (GROUP '07). Association for Computing Machinery, New York, NY, USA, 137–146. https://doi.org/10.1145/1316624.1316644

15. Jodi Schneider, Alexandre Passant, and John Breslin. 2010. A Content Analysis: How Wikipedia Talk Pages Are Used. In Proceedings of the Web Science Conference 2010 (WebSci '10). Raleigh, North Carolina, USA.
16. Jodi Schneider, Alexandre Passant, and John G. Breslin. 2011. Understanding and improving Wikipedia article discussion spaces. In Proceedings of the 2011 ACM Symposium on Applied Computing (TaiChung, Taiwan) (SAC '11). Association for Computing Machinery, New York, NY, USA, 808–813. https://doi.org/10.1145/1982185.1982358
17. Jodi Schneider, Alexandre Passant, and Stefan Decker. 2012. Deletion discussions in Wikipedia: decision factors and outcomes. In Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration (Linz, Austria) (WikiSym '12). Association for Computing Machinery, New York, NY, USA, Article 17, 10 pages. https://doi.org/10.1145/2462932.2462955
18. Jodi Schneider, Krystian Samp, Alexandre Passant, and Stefan Decker. 2013. Arguments about deletion: how experience improves the acceptability of arguments in ad-hoc online task groups. In Proceedings of the 2013 Conference on Computer Supported Cooperative Work (San Antonio, Texas, USA) (CSCW '13). Association for Computing Machinery, New York, NY, USA, 1069–1080. https://doi.org/10.1145/2441776.2441897
19. Donghoon Shin, Soomin Kim, Ruoxi Shang, Joonhwan Lee, and Gary Hsieh. 2023. IntroBot: Exploring the Use of Chatbot-assisted Familiarization in Online Collaborative Groups. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany. ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3544548.3580930
20. Fernanda B. Viégas, Scott Golder, and Judith Donath. 2006. Visualizing email content: portraying relationships from conversational histories. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (Montréal, Québec, Canada) (CHI '06). Association for Computing Machinery, New York, NY, USA, 979–988. https://doi.org/10.1145/1124772.1124919
21. Fernanda B. Viégas, Martin Wattenberg, and Kushal Dave. 2004. Studying cooperation and conflict between authors with history flow visualizations. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (Vienna, Austria) (CHI '04). Association for Computing Machinery, New York, NY, USA, 575–582. https://doi.org/10.1145/985692.985765
22. Fernanda B. Viegas, Martin Wattenberg, Jesse Kriss, and Frank van Ham. 2007. Talk Before You Type: Coordination in Wikipedia. In 2007 40th Annual Hawaii International Conference on System Sciences (HICSS'07). 78–78. https://doi.org/10.1109/HICSS.2007.511
23. Amy X. Zhang, Lea Verou, and David Karger. 2017. Wikum: Bridging Discussion Forums and Wikis Using Recursive Summarization. In Proceedings of the 2017 ACM Conference on Computer Supported

Cooperative Work and Social Computing (Portland, Oregon, USA) (CSCW '17). Association for Computing Machinery, New York, NY, USA, 2082–2096. https://doi.org/10.1145/2998181.2998235