Using AI to Support Discursive Integration in Online Discussions

Maike Behrendt, Viviana Warnken, Dennis Friess, Marc Ziegele, Tobias Escher
Heinrich Heine University Düsseldorf
Department of Social Sciences
{maike.behrendt, viviana.warnken, dennis.friess, marc.ziegele, tobias.escher}@hhu.de

Abstract

Online discussions can be rough, especially when it comes to political issues. They are often characterized by a harsh tone which discourages many people from participating in them at all. At the same time, these discussions are very important for democracy as they promote exchange and help individuals form their own opinions. While Artificial Intelligence (AI) may be detrimental to the quality of discussions (e.g. when used in spam bots), it also offers a promising opportunity to support constructive and inclusive discussions, for example by making them more civil. To strengthen such discursive integration we have engaged in a co-creation process with non-academic stakeholders to develop a discussion assistant prototype that i) identifies likely problematic comments for a possible rephrasing and ii) offers authors help with reformulation by letting generative AI suggest improvements like more civil wording. In this paper, we describe the process of co-creative research and the current status of the discussion assistant, which is still being developed and improved.

Keywords: discussion assistant, classification, text generation, co-creation

1. Introduction

Online platforms have become a central space for political communication and public debate. Social media, comment sections, and discussion forums enable citizens to exchange arguments, encounter opposing viewpoints, and participate in democratic discourse (Stromer-Galley and Wichowski, 2011). Yet, despite their democratic potential, online political discussions are frequently characterized by incivility, hostility, and polarizing language (Beknazar-Yuzbashev et al., 2025). Harsh tones, personal attacks, and sarcastic or aggressive language not only degrade the quality of deliberation but also discourage individuals from participating altogether. As a result, the very spaces that could foster inclusive democratic engagement often reproduce exclusion and withdrawal.

Recent advances in Artificial Intelligence (AI), particularly in natural language processing (NLP) and generative models, open new possibilities to support online communication, e.g., by making them more constructive and less hostile (Shahid et al., 2025). Rather than replacing human deliberation, AI systems can be designed to assist users, e.g., when a comment risks escalating conflict due to inappropriate or uncivil phrasing (Friess et al., 2025). By detecting potentially problematic language and prompting users to reconsider or rephrase their contribution, AI-based discussion assistants may help to preserve the substance of disagreement while reducing unnecessary hostility.

In this paper, we present the development of a prototype AI-supported discussion assistant aimed at fostering discursive integration in online political discussions. We refer to *discursive integration* as the extent to which participants share certain norms of communication that are rooted in more general social norms such as transparency, civility and truthfulness. Such integration ensures that controversies do not cause participants to hate each other or abandon the discussion. In a first step, the system identifies potentially inappropriate comments. Upon identification the assistant invites users to reformulate their comment in the interest of a more constructive exchange. In a second step, the assistant offers help to the user by relying on generative AI to suggest alternative phrasings that maintain the author's intended message while improving the tone. Importantly, the prototype was developed through a co-creative process involving close collaboration with non-academic stakeholders, ensuring that normative, practical, and contextual considerations were incorporated in the development process.

We describe the participatory design process and the implementation of the discussion assistant. As the system is still under development, this paper focuses not only on technical functionality but also on the methodological and normative challenges of designing AI tools intended to shape public communication. In doing so, we aim to contribute to ongoing debates about the role of AI in supporting, rather than undermining, democratic discourse.

2. Related Work

While much research has focused on identifying and subsequently moderating problematic content in online discussion (Falk et al., 2021; Horta Ribeiro et al., 2023), our work builds upon several successful examples that focus on using AI to em-

Proceedings of The 2nd Workshop on Language-driven Deliberation Technology, pages 1–6
May 16, 2026.

power individuals to engage in discursive but not hostile discussions. Argyle et al. (2023) use GPT-3 (Brown et al., 2020) to generate three different rephrased versions of comments in one-on-one discussions on gun control. The authors found that using the AI as an assistant improved both the quality of the conversation and the willingness to listen to the opponent. This effect was particularly pronounced for the conversation partner who received the rephrased messages. However, the authors did not use any classification in advance to trigger the rephrasing of comments. We integrate the idea of letting AI rephrase comments upon request in our assistant. Shahid et al. (2025) display the effectiveness of large language models (LLMs) in the task of co-writing constructive comments. They use GPT-4 to generate constructive comments on two polarizing topics, finding that comments that have been co-written by humans and LLMs were perceived significantly more constructive than human written comments. An even greater effect has been found when comparing human vs. AI generated comments. At the same time, however, the researchers also found that when generating comments, the writers' opinions were often misinterpreted, which led to frustration. Tessler et al. (2024) could demonstrate that an LLM was able to produce statements that incorporated diverse viewpoints from a group deliberation procedure which allowed these groups to find common ground amid their conflicts. Similar to our approach, Yeo et al. (2024) developed an assistant that used generative AI to produce textual nudges that would help discussants reflect their own position before contributing to the discussion. Overall, the studies demonstrate the ability of LLMs to generate suggestions for the improvement of comments. We want to use this capability for the assistant.

3. Development of a Discussions Assistant through Co-Creation

Co-creation is a concept from the broader field of citizen science. It relies on engaging with a variety of different stakeholders from outside academia to incorporate their perspectives (Delgado et al., 2023). Importantly, it aims to engage stakeholders continuously throughout the whole development process. The co-creation process for the discussion assistant entails the following three phases.

3.1. First Phase: Idea Gathering

The goal of the first phase was to generate ideas for features to build in the discussion assistant by using co-creation to identify i) which aspects make up a integrative discussion and ii) through which technical interventions these could be supported. First, we derived criteria from different theoretical frameworks commonly associated with positive discussions, such as deliberation, agonism, and interpersonal communication. These frameworks include concepts such as politeness, empathy, and traceability. In the same way, we compiled a list of 39 (mainly technical) interventions that have so far been shown to aid such meaningful discursive exchange such as fact-checking, encouraging active participation or summaries of discussions. These interventions were primarily drawn from Kraut et al. (2012) and the Prosocial Design Network[1].

In September 2024, we conducted three workshops with the participation of eight professional moderators, content managers, and journalists (two-day workshop), approximately 150 citizens (interactive walk-in poster session for a public science event), and approximately 20 public engagement professionals from public administration. Based on feedback gathered from discussions, a focus group and prototype rankings, we identified the following aspects that an assistant should focus on in order to support discursive online discussions: i) reducing *incivility*, ii) encouraging *justification* (i.e. use of supporting arguments for statements) and iii) *empathy*, iv) provide *rephrasing suggestions* v) offer *summaries of discussions* and vi) *fact-checking* statements. The number represents the priority which we assigned based on perceived importance by the participants and technical feasibility.

3.2. Second Phase: First Prototype

We developed a prototype that focused on the first four key functionalities. It would allow to type a comment which is subsequently being analyzed for incivility, lack of justification and lack of empathy. If either of these is detected, the assistant notifies users of the potential problems and asks if they would like to reformulate. In case incivility (or a lack of empathy) is detected, it offers to provide a more polite version generated by AI. For the specific implementation details see Section 4. In a second round of co-creation, in October 2025 we invited nine experts to a two-day workshop to provide feedback. The participants had expertise in both the governance of online communities and relevant technical domains. The participants encouraged us to further pursue the application of the AI-powered assistant. At the same time, they identified a number of important areas for improvement before the prototype could be tested in a practical scenario.

Based on this we have started implementing the following changes into the next version of the assistant: i) improving detection, in particular of empathy, ii) adapting the intervention threshold to allow for

[1] https://www.prosocialdesign.org/

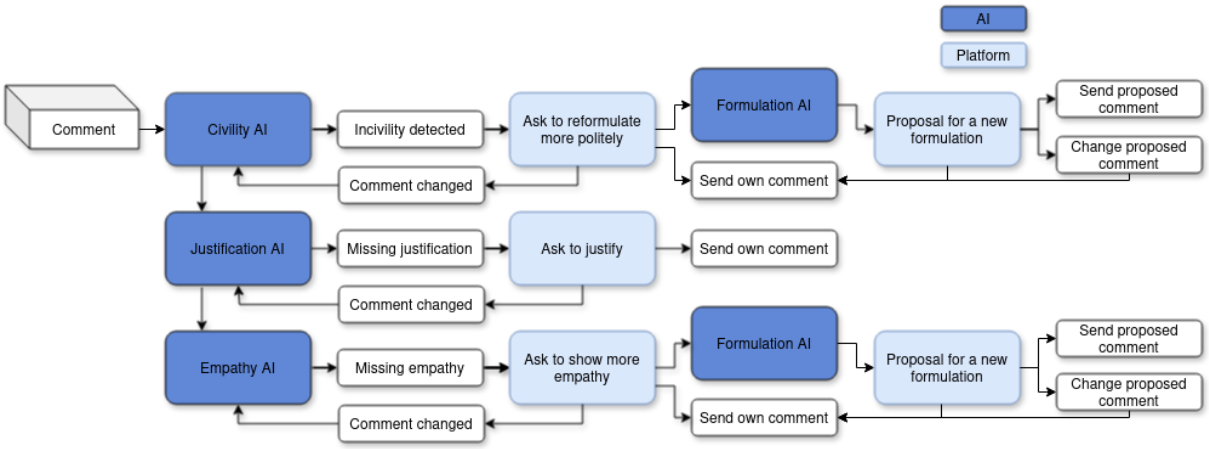

Figure 1: Flowchart for the first prototype of the assistant. A written comment is analyzed by three separate AI components that check for a lack of civility, justification, and empathy. If any of these aspects are missing, the user is asked to reconsider their formulation. If the comment is deemed incivil or lacking in empathy, the user can choose to have a generative AI suggest a reformulated comment.

more animated discussions, including strong arguments, while still interfering, when clear incivility is detected, iii) focusing on offering more help and support to users to formulate good comments instead of policing content. This includes modifications to textual elements, the interface, and functionality. While the assistant should continue to prompt users to consider rephrasing when incivility is detected, increasing empathy or adding justification is only suggested as possible improvement to a comment, iv) rephrasing only those parts of the comments deemed problematic instead the entire comment, and v) multiple improvements to the interface.

3.3. Third Phase: Evaluation

Currently, we are improving the first version of our prototype. Once these improvements are implemented we will put the system to an evaluation with real users. We plan an experimental setting in which groups of users engage in online discussions with the assistant (treatment group) while others discuss without it (control group) in order to test a) how the assistant is used (if at all) and evaluated and if this results in any measurable impacts on b) discursive integration or c) further subjective perceptions of participants. These are planned for summer 2026.

4. Implementation Details

We have developed a prototype for a discussion assistant that supports discussions below online newspaper articles and is intended to foster discursive integration. However, it is likely that it could also be used in other online discussion formats. In the following we describe the components of the assistant in detail. These include a classification, in which individual comments are being analyzed, and a generative component that is able to rephrase the user's comment, if requested.

4.1. Classification

For classification we trained adapter models (Pfeiffer et al., 2021) with a German BERT Base (Chan et al., 2020) as backbone model. The classification models predict a score for a total of six different deliberative dimensions, namely *insults*, *vulgar language*, *screaming*, *discrimination*, *justification* and *empathy* on a Likert scale from 0 to 3. We trained the adapters on the KODIE dataset (Heinbach et al., 2022), which consists of 13,587 German Facebook comments from four different news outlets. The adapter models are used to determine whether the assistant should ask the user to (i) reconsider their comment and rephrase it more politely, (ii) justify their opinion, or (iii) reconsider their comment and rephrase in a more empathetic tone. The first four mentioned adapters are responsible for identifying impolite and inappropriate comments. To evaluate the classification of incivility, we annotated 192 comments (Krippendorff's Alpha was .735.) from articles of a German online news outlet (Rheinische Post Online) for levels of incivility requiring intervention and tested every possible combination of the four variables. We achieved the best F1 score (0.5294, Recall: 0.6, Precision: 0.4737) when the thresholds is set to greater than 1 for insults, greater than 3 for discrimination, and greater than 3 for vulgar language and when omitting screaming. As the evaluation was affected by the small sample size we also tested the classification against 2,000 comments from RPCrowd (Assenmacher et al., 2021) that applied comparable crowd-based moderation and which illustrated that classification is satisfac-

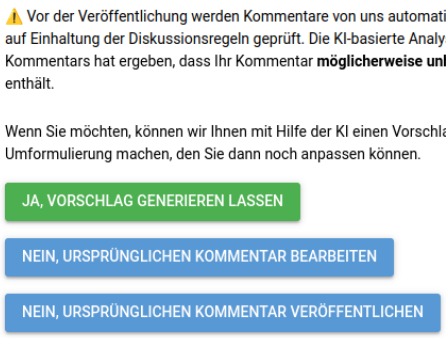

Figure 2: Screenshot of the civility detection. Translation from German: *Before publication, comments are automatically checked by us using AI to ensure they comply with the discussion rules. The AI-based analysis of your comment has revealed that your comment may contain rude content. If you wish, we can use AI to suggest a rewording, which you can then adjust as you see fit.*

You are a helpful assistant with the following task: Revise the following comment according to the rules below. Goal: Use more polite language [*empathetic wording*] while fully preserving the content and original style.
Rules: Formulate the comment politely and respectfully. Avoid offensive, derogatory, or discriminatory language - even in mild or colloquial forms. [*Formulate the post emphatically, i.e., in such a way that other perspectives or feelings are acknowledged or empathized with.*]
Preserve the content of the message in its entirety. No additions or omissions.
Revise the entire text from beginning to end.
Keep the original word count approximately the same.
Preserve the original style as much as possible - unless it conflicts with rule 1.
Precede the revision with the sentence "Here is the revision of the comment:". Comment: { comment }

Figure 3: Prompt for a revised comment (translated from German). Instructions from the prompt for a more empathetic formulation, which differ from the prompt for a more polite formulation, are marked with square brackets. The placeholder { comment } is replaced with the comment written by the user.

tory (F1: 0.71, Recall: 0.83, Precision: 0.62).

The adapter trained on *justification* and *empathy* are used to identify missing justification and empathy, respectively. As described before, users can request an AI-formulated rephrasing of their comment when incivility or a lack of empathy has been recognized. The whole process is shown as a flowchart in Figure 1.

4.2. Comment Rephrasing

To rephrase user comments we use GPT-5 mini[2] via the OpenAI API. We display the translated prompt in Figure 3. We send the user comment with prompt shown. We instruct the model to begin each response with "here is the revised comment." This allows us to cut out this part and show users only the reworded section. The user can then decide whether to accept the AI's wording, edit it, or send their original comment. Figure 2 displays the user dialog when the assistant detects a potentially inappropriate comment. The user has the option of generating a new suggestion from the AI, revising their comment, or submitting their comment.

5. Conclusion

In this work we describe the development of a prototype for a discussion assistant that should foster discursive integration. We generated and prioritized ideas for the concept and improvement of the prototype in close cooperation with non-scientific actors. The assistant should help users to formulate and communicate their opinions in online spaces such as news comments sections, while maintaining a pleasant atmosphere for discussion. Our goal is to prevent users from hostile interactions and quitting discussions early, which often occur due to rude, non-empathetic behavior of a few discussion participants. Our prototype is based on a two-staged concept that first detects comments with the potential to harm the conversation, asking users to reconsider their formulation. Afterwards it offers users the opportunity to assist them in their reformulation.

Lessons Learned Co-creation offers a promising path, especially when it comes to the development of technical advancements. Exchanging with experts and potential end users of the final product can be very beneficial. However, a co-creative approach within a research project also bears risks. The process of integrating non-scientific actors into a research project involves an enormous amount of time and organizational effort that could otherwise be used for research. The requirements for a technical development can change several times during this process. Since it often involves exchanges with small groups, it is not easy for researchers to utilize the findings scientifically, as the methodology often does not meet scientific standards. Although this allows different perspectives to be taken into account and generally leads to a better end result, research projects are often very limited in terms of time and resources. Careful planning of workshops involving non-scientists and rapid systematic evaluation of the results are therefore essential. Nevertheless, this process provides insights that would not otherwise be possible for a pure research team.

[2] https://developers.openai.com/api/docs/models/gpt-5-mini

Ethical Considerations

The development of an AI-supported discussion assistant that can potentially intervene in political communication raises a number of ethical challenges. Since the system can influence how users express their views, it directly affects core democratic values such as freedom of expression, pluralism, and autonomy.

A central concern when designing systems that detect and prompt the reformulation of comments is the potential tension with freedom of expression. Political speech enjoys particular importance in democratic societies. Any intervention that classifies language as inappropriate or suggests alternative formulations risks being perceived as censorship or viewpoint discrimination.

To address this concern, the prototype is designed as a supportive rather than restrictive tool. It does not automatically delete or suppress comments. Instead, it invites users to reconsider their wording and offers reformulation suggestions that preserve the substantive content of their message while moderating tone. The final decision to accept, modify, or ignore the suggestion remains with the user. In this way, the system aims to enhance communicative reflection rather than constrain expression.

The prototype uses the OpenAI API to process user-generated content when the reformulation feature is used. Therefore, it is very important to protect the data and privacy of users. Where possible, data should be anonymized or pseudonymized before transmission, and retention periods should be limited. Users should also be informed about how their data is processed and whether it is stored or used for further model improvement.

Acknowledgments

This publication is based on research in the project IndI, which is funded by the German Federal Ministry of Research, Technology and Space, funding no. 16SV9221. Responsibility for the content of this publication lies with the authors.

6. Bibliographical References

Lisa P Argyle, Christopher A Bail, Ethan C Busby, Joshua R Gubler, Thomas Howe, Christopher Rytting, Taylor Sorensen, and David Wingate. 2023. Leveraging ai for democratic discourse: Chat interventions can improve online political conversations at scale. *Proceedings of the National Academy of Sciences*, 120(41):e2311627120.

Dennis Assenmacher, Marco Niemann, Kilian Müller, Moritz Seiler, Dennis M Riehle, and Heike Trautmann. 2021. Rp-mod & rp-crowd: Moderator- and crowd-annotated german news comment datasets. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

George Beknazar-Yuzbashev, Rafael Jiménez-Durán, Jesse McCrosky, and Mateusz Stalinski. 2025. Toxic content and user engagement on social media: Evidence from a field experiment. CESifo Working Paper 11644, Munich.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. The participatory turn in ai design: Theoretical foundations and the current state of practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '23, New York, NY, USA. Association for Computing Machinery.

Neele Falk, Iman Jundi, Eva Maria Vecchi, and Gabriella Lapesa. 2021. Predicting moderation of deliberative arguments: Is argument quality the key? In *Proceedings of the 8th Workshop on Argument Mining*, pages 133–141, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dennis Friess, Carina Weinmann, and Mira Warné. 2025. Ai and deliberation: Normative ideals in the light of current ai research-a review. *Journal of Deliberative Democracy*, 21(1).

Dominique Heinbach, Lena Wilms, and Marc Ziegele. 2022. Effects of empowerment moderation in online discussions: a field experiment with

four news outlets. In *72nd Annual Conference of the International Communication Association (ICA)*, pages 26–30.

Manoel Horta Ribeiro, Justin Cheng, and Robert West. 2023. Automated content moderation increases adherence to community guidelines. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 2666–2676, New York, NY, USA. Association for Computing Machinery.

Robert E. Kraut, Paul Resnick, Sara Kiesler, Yuqing Ren, Yan Chen, Moira Burke, Niki Kittur, John Riedl, and Joseph Konstan. 2012. *Building Successful Online Communities: Evidence-Based Social Design*. The MIT Press.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, Online. Association for Computational Linguistics.

Farhana Shahid, Maximilian Dittgen, Mor Naaman, and Aditya Vashistha. 2025. Examining human-ai collaboration for co-writing constructive comments online. *Proc. ACM Hum.-Comput. Interact.*, 9(7).

Jennifer Stromer-Galley and Alexis Wichowski. 2011. *Political Discussion Online*, chapter 8. John Wiley & Sons, Ltd.

Michael Henry Tessler, Michiel A. Bakker, Daniel Jarrett, Hannah Sheahan, Martin J. Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tantum Collins, David C. Parkes, Matthew Botvinick, and Christopher Summerfield. 2024. Ai can help humans find common ground in democratic deliberation. *Science*, 386(6719):eadq2852.

ShunYi Yeo, Gionnieve Lim, Jie Gao, Weiyu Zhang, and Simon Tangi Perrault. 2024. Help me reflect: Leveraging self-reflection interface nudges to enhance deliberativeness on online deliberation platforms. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.