

Can AI Gauge Deliberative Quality? Evaluating Deliberation on Social Media With Large Language Models

Patrick Frey ¹ , Smita Bhattacharya ² , Gregor Wiedemann ² ,
and Cornelius Puschmann ¹ 

¹ Centre for Media, Communication and Information Research (ZeMKI), University of Bremen, Germany

² Media Research Methods Lab, Leibniz Institute for Media Research | Hans-Bredow-Institute, Germany

Correspondence: Cornelius Puschmann (puschmann@uni-bremen.de)

Submitted: 31 March 2026 **Accepted:** 18 August 2026 **Published:** 22 October 2026

Issue: This article is part of the issue “The Role of AI for Counter Speech: Detection, Intervention, and Risks” edited by Lena Frischlich (University of Southern Denmark – Odense), Cathy Buerger (Dangerous Speech Project), and Magdalena Obermaier (LMU Munich), fully open access at <https://doi.org/10.17645/mac.i500>

Abstract

Although social media has become a central arena for deliberative discourse, debates are frequently characterized by incivility. Due to the volume of online discourse, it is challenging to assess when and where interventions, such as community-based counterspeech, are required. Accordingly, automated measurement of deliberative quality is a powerful tool for facilitating constructive online communication. Since deliberation is multidimensional, attempts to monitor online discourse must consider alternative forms of communication to foster inclusive deliberation. Whereas Type I deliberative dimensions, such as rationality and civility, are grounded in Habermasian ideals, Type II deliberation extends this perspective by incorporating alternative communicative forms, including emotionality and storytelling, into a more inclusive assessment of deliberative quality. Addressing the limitations of prior studies that primarily focused on classic Type I deliberative dimensions and tested only a small number of large language models, this study evaluates multiple open-weight models to conduct a nuanced assessment of deliberative quality. We applied the large language model that achieved the best overall balance between predictive performance, model size, and computational efficiency to a dataset of 64,223 comments from German-language climate change discourse on X. Our results demonstrate that large language models can reliably capture complex Type II deliberative dimensions despite class imbalances and subtle linguistic cues that generally hinder automated text classification. Class imbalance often impairs classification performance due to highly skewed label distributions, and Type II dimensions are particularly difficult to detect as they are often signaled by implicit, context-dependent markers. The empirical analysis also reveals distinct communicative patterns: While public speakers predominantly employ civil, rational, constructive, and common-good-oriented communication, other debating users exhibit higher levels of reciprocal engagement, humor, and negative emotions, alongside more frequent incivility. Our study demonstrates that AI-driven measurement can serve as a foundational tool for identifying norm violations, directing counterspeech toward specific debates,

promoting inclusive digital public spheres, and observing broader long-term effects of moderation and interventions on the quality of debate.

Keywords

argumentation; automatic content analysis; counterspeech; debate quality; deliberation; discourse; political discourse; social media

1. Introduction

While social media can provide an accessible arena for political communication, it may also catalyze the fragmentation of the public sphere and entrench political polarization (Jungherr & Schroeder, 2023). Online, the dissemination of political information is not based on traditional ideals such as authoritative backing, but is instead influenced by platform-specific affordances (Theocharis et al., 2020) that often serve to enable incivility (Duncan et al., 2026). A recent example is X (formerly Twitter), which has long been a hotspot for political discussions. Following the acquisition by Elon Musk, new platform rules were implemented, including fewer restrictions on platform-driven content moderation, which sparked a significant increase in incivility (Hickey et al., 2025). These dynamics hinder the unfolding of constructive political deliberation and can trigger a so-called “spiral of incivility” as uncivil comments spark further incivility in subsequent discussions (Friess et al., 2021).

An alternative, bottom-up, and community-based moderation strategy is “counterspeech” (Lasser et al., 2025; Mun et al., 2024). Counterspeech encompasses various forms of communicative user responses to hateful comments, ranging from providing facts on a discussion topic to supporting victims of different forms of incivility (e.g., Friess et al., 2021). Counterspeech is less dependent on platform-based, top-down moderation structures and regulations than other interventions, while adhering to norms of free speech (Lasser et al., 2025). Yet, as political discussions on social media platforms are complex to monitor, it is challenging to assess whether they require support from counterspeakers to stay on track. This creates a crucial barrier (Garland et al., 2022), as the decision to respond is affected by the amount of time counterspeakers need to identify posts and comment sections that are promising for impactful counterspeech (Mun et al., 2024). Due to rapid advances in artificial intelligence (AI)—particularly large language models (LLMs), which show promising results for measuring deliberative quality online (e.g., Stolwijk et al., 2025)—AI appears to be well-suited to support counterspeech. However, the few existing studies utilizing LLMs (Bosley, 2025; Stolwijk et al., 2025) compared only a small number of models and focused only on dimensions such as rationality and incivility—so-called Type I deliberation—while largely omitting alternative forms of communication, such as empathy and storytelling—so-called Type II deliberation—that are particularly prevalent among marginalized groups and necessary to foster inclusive discourse (Esau & Friess, 2022).

Addressing this gap, we evaluate multiple open-weight models based on our operationalization of Type I and Type II deliberative dimensions for asynchronous and dialogical social media communication. Applied to a large-N dataset of German-language climate discourse on X under posts by representatives of major German political parties and other relevant public speakers, such as journalists, covering the period from March 2021 to March 2023, we systematically evaluate LLM performance on this task and gain insights into the distribution of deliberative communicative dimensions in real-world political discourse. Beyond providing a

validated measurement approach, our analysis enables the systematic identification of user groups that demonstrate higher rates of comments that violate deliberative norms in polarized climate change discussions. Consequently, rather than automating counterspeech, we focus on laying the groundwork for tools that can support counterspeakers by detecting user comments that violate deliberative norms (e.g., incivility) and by distinguishing between X conversations (i.e., an initial post by a public speaker and all comments posted underneath it, including reply chains in which users respond both to the original post and to one another) in which deliberative quality has already eroded and those where deliberative comments create conditions that make bystander activation more likely (e.g., Friess et al., 2021). Therefore, we examine the following research questions:

RQ1: How reliably and validly can Type I and Type II deliberation be measured with the help of LLMs?

RQ2: How are deliberative norm violations and norm adherence distributed across different types of users in German-language climate discourse on X?

2. Counterspeech and Online Deliberation

2.1. Definition and Process of Counterspeech

Counterspeech is generally understood as “any direct response to hateful or harmful speech that seeks to undermine it” (Obermaier et al., 2026, p. 2). In comparison with formal content moderation, for instance the deletion of comments by algorithmic detection, this approach exerts informal social control through ordinary users who have no access to platform-specific tools (Friess et al., 2021). Counterspeech can occur in the form of individual interventions or coordinated action organized around specific hashtags (e.g., #iamhere) or communities (Buerger, 2021a; Wu et al., 2025). As Buerger (2021b) explains, counterspeakers themselves often describe that, in addition to supporting users who are targeted by hateful speech, they try to reach the “silent” audience with the goal of convincing them to adopt a certain position or motivating them to engage in counterspeech themselves. Accordingly, ongoing counterspeech tries to motivate bystanders to contribute to a more respectful online discourse (Buerger, 2021b), rather than altering the behaviors or views of users expressing hateful speech (Wu et al., 2025). This can trigger a “contagion effect” (Buerger, 2021b) or a “spiral of deliberativeness” that may motivate bystanders to engage in more respectful political discourse (Friess et al., 2021).

Counterspeech encompasses a variety of actions and objectives, as Garland et al. (2022, p. 1) state, “users of online platforms themselves generate responses to hateful content in order to stop it, reduce its consequences, discourage it, as well as to support the victim and fellow counterspeakers and ultimately increase civility and deliberation quality.” According to Mun et al. (2024), counterspeech consists of three phases: (a) deciding to respond, starting with the detection of an utterance that is perceived as hateful and the decision to respond to it based on considerations such as the trade-off between effort and anticipated effectiveness; (b) formulating a response, by drafting a comment aligned with individual or collective objectives; and (c) posting and aftermath, which includes dealing with responses from other users. Different barriers exist in this complex process that potentially hinder the completion of a counterspeech act (Garland et al., 2022). For instance, the decision to respond is affected by the amount of time counterspeakers need to identify posts that are promising for impactful counterspeech (Mun et al., 2024).

2.2. Definition and Types of Deliberation

An important condition for effective counterspeech that tackles hate and enhances discourse quality is the alignment with deliberative norms such as civility (Ziegele et al., 2018). For instance, Friess et al. (2021) conclude, based on a quantitative content analysis of the German action group #ichbinhier (#iamhere), that collective counterspeech “can potentially increase the quality of online discussions in comment sections, as long as this engagement lives up to the norms of deliberation itself” (p. 638).

According to communication-centric deliberative democratic theory, which became prominent through the work of Habermas (1996), communication has to follow specific norms and provide egalitarian inclusion to enable democratic reciprocity, referring to “a willingness to grant political opponents the same right to express and advocate their views in the public sphere that we hope they will grant us” (Argyle et al., 2023, p. 1). Thus, deliberative processes are an essential component of representative democracies, for instance, to enable domination-free opinion formation (Duncan et al., 2026). However, social media comments often do not adhere to deliberative norms (Lasser et al., 2025; Naab et al., 2025).

A differentiation between so-called Type I and Type II deliberation is widely established (e.g., Bächtiger et al., 2010), with both referring to the question of how deliberation should be carried out (Friess & Eilders, 2015). Type I deliberation mostly originates from Habermasian logic (Bächtiger et al., 2010), which theorizes deliberation as a “communication process in which equal participants justify their positions in a respectful manner, always willing to accept the forceless force of the better argument” (Esau & Friess, 2022, p. 2). This, for instance, includes respect toward other discourse participants, that is, civility (e.g., Steenbergen et al., 2003). As critical scholars argue, deliberation is multidimensional; hence, a sole focus on Type I deliberation would represent only the communicative practices of more privileged actors (Esau & Friess, 2022) and fail to take into account the nature of non-institutionalized discussions in the digital public sphere (Bodrunova, 2024). In contrast, Type II deliberation “involve[s] a shift away from the idea of purely rational discourse toward a conception of deliberation that incorporates alternative forms of communication” (Bächtiger et al., 2010, p. 3).

However, despite shared denominators such as civility and reciprocity, there is no consensus on the relevant dimensions for measuring asynchronous online deliberation (Esau & Friess, 2022; Ziegele & Eilders, 2022), nor is there a comprehensive framework integrating classical and more inclusive forms of deliberation (Esau, 2022). In light of this debate, we focus on a theoretically informed set of Type I and Type II dimensions to derive indicators for the LLM-based monitoring of social media discourse. On this basis, we identified the following key dimensions for Type I deliberation, which are largely shared in the literature (e.g., Behrendt et al., 2024; Esau et al., 2023) and summarized in the literature review by Friess and Eilders (2015): (a) Rationality (reasoned argumentation); (b) Reciprocity (the interactive mode of deliberation as a foundation for mutual understanding); (c) Civility (respectful tone between users); (d) Constructiveness (attempts to find solutions or a common ground to reach a consensus); and (e) Orientation Toward the Common Good (the connection of contributions to the interests of others, which is particularly relevant in climate change discourse). The only dimension we excluded, despite its relevance for deliberation, was Equality, which means that all users should have the same chance to contribute to an online discussion (Friess et al., 2025). The key reason for this decision was that its coherent operationalization is difficult for human coders to evaluate, including determining the gender distribution in an X discussion based solely on user names and comment content.

In contrast, there is no clear consensus on relevant Type II dimensions. We included salient dimensions presumed to influence the quality of political online discourse: (a) Emotionality (expression of emotions); (b) Storytelling (use of stories to justify a stance); (c) Empathy (perspective-taking or empathizing with others); (d) Reflexivity (reconsidering one's own stance in light of others' contributions); and (e) Humor (humorous contributions, including irony and sarcasm; see Esau et al., 2023). Prior studies indicate that the expression of emotions exerts an ambivalent influence on deliberation depending on their valence (negative or positive) and target (characteristics of another user or discussion topic; e.g., Esau et al., 2023). Storytelling is widely seen as an inclusive alternative to rational argumentation, although findings on its deliberative effects are mixed (e.g., Bächtiger et al., 2010; Esau et al., 2023). Empathy and reflexivity are both considered necessary for mutual understanding between discourse participants (e.g., Esau et al., 2023; Graham, 2010). Finally, humor can enhance deliberative quality, particularly along classical deliberative dimensions (e.g., Esau, 2022; Esau & Friess, 2022). However, humor can also be used to disguise uncivil communication (e.g., Schmid & Greipl, 2025).

2.3. Deliberative Norms and Counterspeech Strategies

This distinction is useful when looking at counterspeech strategies, which, following Obermaier et al. (2026), can be grouped into five different approaches that largely align with different deliberative communicative norms. Obermaier et al. (2026) differentiate reason-based, empathy-based, norm-based, sanction-based, and support-based forms of counterspeech. With regard to “reason-based” forms, such as providing factual knowledge, which overlap with Type I deliberative dimensions, Friess et al. (2021) show that constructive, rational, and polite comments by members of the German counterspeech group #ichbinhier (#iamhere) are associated with more deliberative follow-up discussions, while Lasser et al. (2025) find that especially in highly polarized situations on German X, the most effective counterspeech strategy is to express opinions that are delivered without uncivil language, even when these opinions are not supported by factual information. “Norm-based” strategies focus on reminders and calls for Type I-related communication norms, such as politeness (Obermaier et al., 2026). For example, metacommunicative comments can prompt subsequent users to model metacommunicative behavior, as users were more likely to intervene and point out uncivil comments after being informed of the uncivil tone of a discussion (Masullo et al., 2022).

“Empathy-based” strategies are attempts at “demonstrations of and appeals to the perpetrator’s cognitive empathy” (Obermaier et al., 2026, p. 3) and thus account for Type II deliberation. For instance, as Bilewicz et al. (2021) demonstrate, bot-generated empathy-based interventions in social media discourse can reduce subsequent verbal aggression. “Support-based” approaches focus on victims by encouraging them or expressing solidarity (Obermaier et al., 2026) and thus also include empathic communicative elements. In contrast, “sanction-based” strategies target the authors of destructive communication. These can encompass humorous contributions (Type II deliberation), but can also violate deliberative norms by insulting users in an attempt to explicitly sanction their behavior (Obermaier et al., 2026). Lasser et al. (2025) show that simple expressions of opinion and sarcasm can be more effective than evidence-based or constructive responses, whereas insults and other non-deliberative counterspeech reduce discourse quality. Focusing on journalistic counterspeech, Ziegele et al. (2018) demonstrate that only sociable forms of interactive moderation (e.g., harmless jokes) reduce subsequent incivility.

Existing research highlights that many counterspeech strategies employ deliberative communication and benefit from adhering to its norms. However, effective counterspeech does not necessarily need to conform to classical deliberative norms such as civility; it can also benefit from alternative deliberative styles such as empathy. However, counterspeech and deliberative communication are not interchangeable. While counterspeech refers to intentional user-responses to hateful speech, not all forms of counterspeech are deliberative (e.g., sanction-based strategies). Moreover, users' adherence to deliberative norms can contribute to similar objectives, but does not necessarily constitute a response to incivility or other norm violations.

2.4. Conditions for Effective Counterspeech

Aside from the type of counterspeech strategy, other factors influence counterspeech effectiveness, especially in activating bystanders. Bystanders' willingness to intervene is, for instance, influenced by the target of hateful speech. For example, the likelihood is lower when hate is directed at politicians compared to marginalized groups; thus, perceived vulnerability influences the perception of responsibility to enact counterspeech (Obermaier & Schmid, 2026).

Moreover, the normative context of a discussion is important. Obermaier and Schmid (2026, p. 16) show that the perceived presence of prosocial communication norms (e.g., other users providing emotional support) strengthens the bystanders' sense of personal responsibility and willingness to engage in counterspeech, as they are "more motivated to act when they believed others approved of intervention." In this respect, identifying the presence of deliberative comments, such as empathic contributions, may be useful for locating X conversations in which counterspeakers are more likely to receive bystander support. Conversely, a low presence of deliberative comments (e.g., rational contributions) may indicate social media environments in which counterspeech is less likely to be effective in tackling hateful speech or where discussions are at risk of escalating into less deliberative discourse, warranting closer monitoring ("spiral of incivility"; see Friess et al., 2021).

Consequently, rather than conceptualizing deliberation monitoring as a direct trigger for counterspeech, we view it as a tool that (a) flags user comments that contain uncivil content or violate other deliberative norms and (b) helps to differentiate between conversations with higher or lower levels of deliberative communication, supporting targeted counterspeech coordination. Additionally, tracking changes in deliberative quality over time can help counterspeech groups monitor how discussions evolve after interventions and adjust their strategies accordingly.

To support counterspeakers in this matter, LLMs could enable the automated assessment of deliberative quality to detect instances where counterspeech might be needed and prove fruitful. LLMs may also assist in generating counterspeech responses themselves, along dimensions of deliberation such as reason-based strategies (R. Gupta et al., 2023). This raises the question of how LLMs can be integrated into deliberative counterspeech practices, ranging from discourse monitoring to automated response generation, which is addressed in the following section.

3. Automated Monitoring and Counterspeech

Cypris et al. (2022) stipulate that automated counterspeech can, in principle, serve the same mechanisms and purposes as human counterspeech, for example, by acting as a “behavioral prompt” for human bystanders that positively alters the tone and mode of online discussions. In contrast, Mun et al. (2024, p. 10) highlight different fears of counterspeakers concerning the use of LLMs to support or substitute human counterspeech, which include: “short-term concerns of credibility, long-term negative impact on authenticity leading to disengagement in counterspeech, loss of agency over and dehumanization of counterspeech, and functionality doubts about AI tools.” These concerns include poor performance in detecting uncivil language that targets minority groups (Kathirgamalingam et al., 2025; Stoll et al., 2025).

Moreover, robust empirical evidence on the effectiveness of LLM-generated counterspeech is still missing (Wu et al., 2025). As Bär et al. (2024) show based on a large-scale field experiment on X, certain types of LLM-generated counterspeech are not only ineffective but can also lead to higher levels of incivility in subsequent comments. Similarly, Jungherr and Rauchfleisch (2025) identify, based on a survey experiment, an “AI-penalty” in deliberative formats for both willingness to participate and evaluations of discourse quality. The aforementioned user survey of Mun et al. (2024) highlights that counterspeakers could, in principle, imagine involving LLM tools in the different phases of counterspeech if they are useful, particularly with regard to the ease of adoption and functionality of the tools, as many counterspeakers seek support to make counterspeech more efficient and less time-consuming. Still, as “counterspeech is deeply rooted in human empathy, cultural understanding, and personal motivation” (Ping et al., 2025, p. 2), it seems necessary to avoid generating counterspeech without human involvement (Mun et al., 2024).

A focus on easy-to-use LLM-based monitoring of deliberative quality with human oversight (e.g., an LLM-powered browser add-on) could be a first step toward integrating LLMs into counterspeech. This would enable counterspeech communities to coordinate collective action more precisely and assess the effects of their actions at scale while preserving human agency that is central to counterspeech. Therefore, automated measurement of deliberative quality could be a precondition for community-based facilitation in online mass communication (Korre et al., 2025).

However, the few studies that have tested LLM-based analyses of deliberation in political discourse have not accounted for alternative Type II deliberation dimensions and have compared only a small number of LLMs (Bosley, 2025; Stolwijk et al., 2025). Moreover, in the case of Bosley (2025), only parliamentary speeches were analyzed, which do not capture the complexity involved in analyzing short text passages on social media. Prior to the growing relevance of LLMs, automated approaches often relied on fine-tuned transformer-based classifiers such as BERT or RoBERTa, for instance, to predict deliberative quality scores or argument quality (see, e.g., Behrendt et al., 2026; for overviews, see Wachsmuth et al., 2024). Research projects that have developed successful automated classifiers also exhibit relevant gaps. These include failing to account for “the nuance of human speech that could indicate quality deliberative discourse, features that manual coders could observe” (Duncan et al., 2026, p. 13) and including only selected Type II dimensions (e.g., Behrendt et al., 2024). Nevertheless, the results of Stolwijk et al. (2025) highlight that LLMs show promise for measuring deliberation and outperform previous computational approaches.

4. Methods and Data

4.1. Sampling and Datasets

The dataset used in this study comprises public dialogical conversations on X from German political discourse on climate change. Climate change represents an ideal issue for our study, as it is a defining topic in public debates and the media (Schäfer & Painter, 2021). Climate change is underpinned by factual information while at the same time being emotionally salient, provoking heated debates and triggering high levels of online engagement (e.g., Meyer et al., 2023). Against this backdrop, debates about climate-related topics are politicized and often polarized (e.g., Esteve-del-Valle, 2025). These dynamics are particularly relevant on social media platforms, which not only function as communication channels for political elites but also create spaces for more informal political participation by younger and marginalized groups (Boulianne & Theocharis, 2020).

As a basis for the data collection, we utilized the Database of Public Speakers (Schmidt et al., 2023)—a set of curated account lists of German politicians, political parties, and other relevant public actors and organizations, along with their profiles on various social media platforms. Drawing on the Database of Public Speakers, this classification distinguishes public speakers from other X users in our dataset based on their contribution to societal public discourse in Germany, either by producing public reporting (e.g., as journalists or media outlets) or by being generally in the public eye and subject to public reporting (e.g., as politicians or political organizations). Using the (now closed) Twitter Academic API v2, all posts by public speakers and their metadata were collected over the period from March 2021 to March 2023. To update interaction metrics (e.g., likes and retweets), the dataset was subsequently rehydrated in April 2023 using the twarc library (Summers et al., 2023). In addition, entire user conversations were retrieved via the API for all collected root comments.

From this large dataset (approximately 12 million comments), we sampled two topic-specific subsets related to climate change for our study. First, thematic subsampling was performed using a regular expression that combined 29 key terms from the climate debate into a single comprehensive search pattern (e.g., “Klima” [climate], “Erderwärmung” [global warming]; see Supplementary File, Appendix D, Table D.1 for the complete pattern). This filtering step produced a subset of comments centered on climate-related topics and linked to their conversation context via a specific conversation ID. Using this ID, we selected all conversations that: included at least one climate-related comment; contained between five and 360 comments, corresponding to the second and third quartile of all conversation lengths; and included at least one public speaker comment (e.g., by a German politician or journalist).

This final analysis dataset totals 64,223 comments, of which 3,447 were authored by public speakers and another 60,776 by other debating users in 1,251 conversations, with a median length of 17 turns (see Appendix D, Table D.2). This conversation-level filtering naturally results in an imbalanced distribution of comments across speaker types, as conversations typically contain substantially more comments by debating users, while still ensuring that both short and long conversations are included.

For the validation of the manual content analysis, we drew a stratified subsample of 266 comments. This subsample was created with the aim of obtaining a more balanced distribution of categories, after an initial

coding round on a random subsample of 320 comments from the main dataset revealed substantial class imbalance. For the validation of the automated content analysis, we combined both subsamples ($n = 586$), using human coding decisions as the ground truth for evaluating the LLM-based classification (see Appendix B, Tables B.2 and B.4 for the distributional statistics).

4.2. Human Validation of Codebook

To ensure consistency and methodological rigor, a detailed codebook was developed for the manual content analysis (see Appendix A, Table A.2 for the operationalization), following current recommendations for the validation of LLM-based content analysis against a human benchmark (e.g., Törnberg, 2023). For Type I deliberation, the codebook includes rationality, reciprocity, civility, constructiveness, and common good orientation. For Type II deliberation, we selected emotionality, empathy, storytelling, reflexivity, and humor. Some dimensions in our codebook, which was iteratively refined through several coding sessions, include different subdimensions to comprehensively capture associated sub-phenomena (see Appendix A, Table A.2). For instance, rationality encompasses various forms of justification, such as external justification through evidence from external sources and internal justification through one's own reasoning. Apart from the ordinal category external justification (high quality/low quality/absent) and the nominal category emotions (no emotion/negative emotion/positive emotion/unclassifiable emotion), all categories are dichotomous (present/absent).

The validation proceeded in two steps: First, human intercoder reliability was conventionally tested, and second, LLMs were evaluated against ground-truth labels derived from human coding. In the initial validation sample, ground-truth labels were based on majority decisions by three trained human coders. Cases without a majority decision, as well as N/A cases, were excluded. In a targeted follow-up validation based on a stratified sample, two trained human coders established the ground truth through consensus, i.e., both coders agreed on the same label.

Intercoder reliability was tested with extensively trained human coders. As deliberation is a relational phenomenon, the preceding comment in a conversation provides essential context for annotating most categories (e.g., substantial dialogic exchange, which assesses whether users reference prior arguments). Therefore, each annotated unit consisted of a user comment together with the specific comment to which it replied within the same conversation. In the initial random validation sample ($n = 320$), some categories showed low Krippendorff's α values ($\alpha < .667$) despite very high percentage agreement between three trained human coders and a highly skewed distribution, with the majority of cases in many categories coded as "absent," particularly for Type II deliberation dimensions (e.g., storytelling). This pattern suggested that the low α values were likely driven by severe class imbalance rather than fundamental problems in the codebook or human coding. In line with other recent research facing similar challenges (e.g., Stolwijk et al., 2025), we computed alternative reliability coefficients that are less sensitive to skewed distributions (κ ; Brennan & Prediger, 1981), which indicated strong agreement ($\kappa > .80$) for almost all affected categories (see Appendix B, Table B.1).

We then drew a second, stratified sample from the climate-related subset, using preliminary LLM-based labels to oversample likely positive instances of low-prevalence categories, particularly for Type II deliberation (e.g., emotions, reflexivity). All items in this sample ($n = 266$) were then independently coded by

two human coders, and these labels were used to compute the final reliability results. Krippendorff's α exceeded .80 for 11 of 13 categories and remained above .75 for civility and storytelling ($\emptyset = .89$; see Appendix B, Table B.3). Our results clearly surpass the commonly used threshold of $\alpha = .667$ for tentative conclusions and meet the benchmark of $\alpha = .80$ for reliable agreement for almost all categories (Krippendorff, 2004). This indicates that all 13 deliberative dimensions can be considered sufficiently reliable as a human benchmark for validating the LLM-based coding and analyzing the larger dataset. Nonetheless, the comparatively lower alpha values for civility and storytelling warrant slightly more cautious interpretation.

5. Results

5.1. Prompting Strategy and Classification Pipeline

The automatic classification of the 13 deliberative dimensions was performed in an LLM-based one-shot approach, i.e., presenting one positive and one negative example per category in addition to the LLM coding instructions. This approach was chosen because initial pilot tests indicated that it provided a suitable trade-off between accuracy and efficiency (cf. Sun et al., 2026). To identify the optimal prompting strategy, we evaluated zero-shot and one-shot prompting, both with and without the inclusion of label distribution information. The results of this comparison indicated that a one-shot approach augmented with label distribution information yielded the most accurate classifications. Hints on label distribution were included to mitigate the increased false positive rates observed during early testing, as prior research (K. Gupta et al., 2023) demonstrates that skewed label distributions in in-context prompts can systematically bias LLM predictions toward majority classes. The label distribution information was derived from the validation data obtained from the randomly sampled subset in the first stage of human coding (see Section 4). A separate validation conducted on the more balanced, stratified sample confirmed that adding distribution information to the prompt did not harm classification accuracy, suggesting that the label information remains appropriate and informative for the analysis of the full dataset.

For each deliberative dimension, the pipeline consisted of four steps: prompt construction, LLM-based labeling, post-processing, and evaluation. Following evidence that LLMs benefit from focused, standardized, and clearly structured instructions (e.g., Farjam et al., 2026), we used GPT-5.2 (Singh et al., 2025) to condense the detailed manual codebook into a simplified prompt. We then manually revised the category descriptions and examples to align them with the codebook and iteratively refined the prompt based on initial coding results to improve performance (e.g., Törnberg, 2024). The final prompt consists of: a role instruction; the target post and its conversational context as input; a category definition; coding instructions for assigning labels; one manually selected in-context example per label; and output constraints specifying the allowed labels and format. Labels were generated via prompting with temperature set to zero and a fixed random seed to ensure reproducibility.

To explore different model sizes and both standard and thinking variants, we chose and evaluated eight widely adopted open-weight models (see Appendix C, Table C.6 for quantization information). Based on initial tests, which revealed difficulties for the LLMs in distinguishing nuanced labels for external justification and emotions, we decided to merge the multi-class problems into binary categories. For external justification, the class labels distinguishing between “high-quality” and “low-quality” sources were merged into the label “present” (Class 1). The labels “no emotion” and “unclassifiable emotion” were merged and converted to “absent” (Class 0).

5.2. LLM Performance for Measuring Deliberative Dimensions

Table 1 reports average classification performance across all 13 deliberative quality (sub-)categories. As the validation dataset exhibits strong class imbalance for some categories, accuracy can be misleading, and macro-averages of precision, recall, and $F1$ are the more appropriate performance indicators (see Appendix C, Table C.4 for performance metrics of the “present” minority label per category for the selected model). Although the validation data are notably imbalanced across categories, stratification ensured a stronger representation of “present” labels, particularly for rare Type II categories, compared with our initial random sampling approach (see Section 4). Results revealed substantial variation across different model sizes as well as thinking and non-thinking models. The smallest model, Llama3.1-8B-instruct, performed poorly ($F1 = .54$), plausibly confirming that small models struggle with the nuances of codebook-level annotation. Some mid-sized models showed moderate results, including Qwen3:30B-instruct ($F1 = .63$) and Gemma3:27B ($F1 = .67$). Qwen3:30B-thinking ($F1 = .74$) demonstrated better results, indicating that “thinking,” i.e., chain-of-thought reasoning before answer generation, partially compensates for scale. The strongest overall performance was achieved by GPT-OSS:120B ($F1 = .77$), suggesting robust generalization even under strong class imbalance. Notably, the much smaller “thinking” models GPT-OSS:20B ($F1 = .75$) and Gemma4:26B ($F1 = .75$) came close to the performance of the best model, while also outperforming the larger Llama3.3:70B ($F1 = .71$).

Table 1. Model selection.

Model	Thinking	Accuracy	Precision	Recall	$F1$
Gemma3:27B (Kamath et al., 2025)	No	.77	.69	.77	.67
Gemma4:26B-a4B (Abd et al., 2026)	Yes	.87	.76	.78	.75
GPT-OSS:120B (Agarwal et al., 2025)	Yes	.87	.77	.80*	.77*
GPT-OSS:20B (Agarwal et al., 2025)	Yes	.78	.75	.76	.75
Llama3.3:70B (Grattafiori et al., 2024)	No	.82	.73	.76	.71
Llama3.1:8B instruct (Grattafiori et al., 2024)	No	.63	.63	.64	.54
Qwen3:30B-a3B thinking (Yang et al., 2025)	Yes	.89*	.80*	.73	.74
Qwen3:30B-a3B instruct (Yang et al., 2025)	No	.76	.68	.69	.63

Notes: Average performance of the models across all categories (macro average); * = the highest performance; accuracy—proportion of all instances correctly classified; precision—of all instances labelled as positive by the model, the proportion that was truly positive; recall—of all truly positive instances, the proportion the model correctly identified; $F1$ -score—the harmonic mean of precision and recall, providing a single balanced measure particularly useful when classes are imbalanced.

Considering the substantial computational costs associated with LLM-based classification, particularly for very large models, Gemma4:26B was selected as a practical compromise. Although GPT-OSS:120B achieved the highest overall performance, Gemma4:26B demonstrated comparable results with a significantly smaller model size and substantially faster processing.

Table 2 presents the per-category performance of Gemma4:26B, which offered the best overall balance between predictive performance, model size, and computational efficiency. Performance varied across categories, with particularly strong results for reflexivity ($F1 = .89$), external justification ($F1 = .84$), and empathy ($F1 = .83$). Internal justification ($F1 = .81$) and follow-up questions ($F1 = .76$) also showed strong performance, suggesting that the model handles reasoning-based categories well. Conversely,

interaction-based dimensions proved more difficult, as formal dialogic exchange ($F1 = .72$) demonstrated slightly weaker results, while the model struggled most with substantial dialogic exchange ($F1 = .51$). For the latter category, this likely indicates that the model is challenged by the annotation task, which requires considering both the comment and the content of the previous comment to which it replies, which is often only partially or implicitly referenced.

Table 2. Performance of selected model Gemma4:26B for data analysis (macro average).

Model	Deliberation Type	Category	Accuracy	Precision	Recall	F1-score
Gemma-4-26B	Type I	External Justification	.94	.81	.87	.84
		Follow-up Questions	.92	.92	.70	.76
		Internal Justification	.82	.80	.83	.81
		Formal Dialogic Exchange	.74	.73	.74	.72
		Substantial Dialogic Exchange	.85	.59	.52	.51
		Civility	.92	.70	.87	.75
		Constructiveness	.74	.71	.79	.71
		Common Good Orientation	.83	.74	.89	.77
	Type II	Emotions	.88	.65	.63	.63
		Empathy	.92	.84	.82	.83
		Storytelling	.89	.74	.82	.77
		Reflexivity	.97	.93	.86	.89
		Humor	.87	.75	.81	.78

For the remaining Type I dimensions, Gemma4:26B achieved robust performance on civility ($F1 = .75$), constructiveness ($F1 = .71$), and common good orientation ($F1 = .77$). Among the remaining Type II dimensions, storytelling ($F1 = .77$) was comparatively easy for the model to classify, while humor ($F1 = .78$) was also handled well. In contrast, emotions ($F1 = .63$) posed a significant challenge, contrasting with prior research showing that fine-tuned transformer models like BERT achieve high performance in emotion detection, including in political texts, often outperforming larger generative models in zero- or few-shot settings (Bucher & Martini, 2024).

For the two low-performing categories, a more detailed picture emerges when examining the per-class metrics for the positive label (see Appendix C, Table C.4). For substantial dialogic exchange, the low macro-averaged $F1$ (.51) masks strong performance for the present label (precision = .85, recall = .98), suggesting that the low macro score is driven by poor classification of the minority class. For emotions, per-class metrics reveal an imbalance, as the “absent” class is reliably detected (precision = .93, recall = .92), while both emotional valence categories perform poorly, with positive emotions showing particularly low precision (.46) and recall (.35).

5.3. Prevalence of Deliberative Dimensions in Climate Change Discourse

Automated content analysis using Gemma4:26B provides insights into the distribution of Type I and Type II deliberative dimensions in contributions by public speakers and debating users (non-public speakers). To illustrate the practical application of our classification approach, we measure the prevalence of these deliberative dimensions in the full sample of X comments discussing climate change. Of the 13 categories,

11 exhibit significant distributional differences by speaker type. Our analysis (see Figure 1) highlights that Type II deliberative dimensions are rare across both user groups, ranging from 0.36% for reflexivity to 16.70% for humor. This suggests that alternative deliberative communication practices are considerably less common compared with classic deliberative dimensions, which range from 4.37% for follow-up questions to 95.67% for substantial dialogic exchange across all users. This finding is largely aligned with previous research (e.g., Esau et al., 2023).

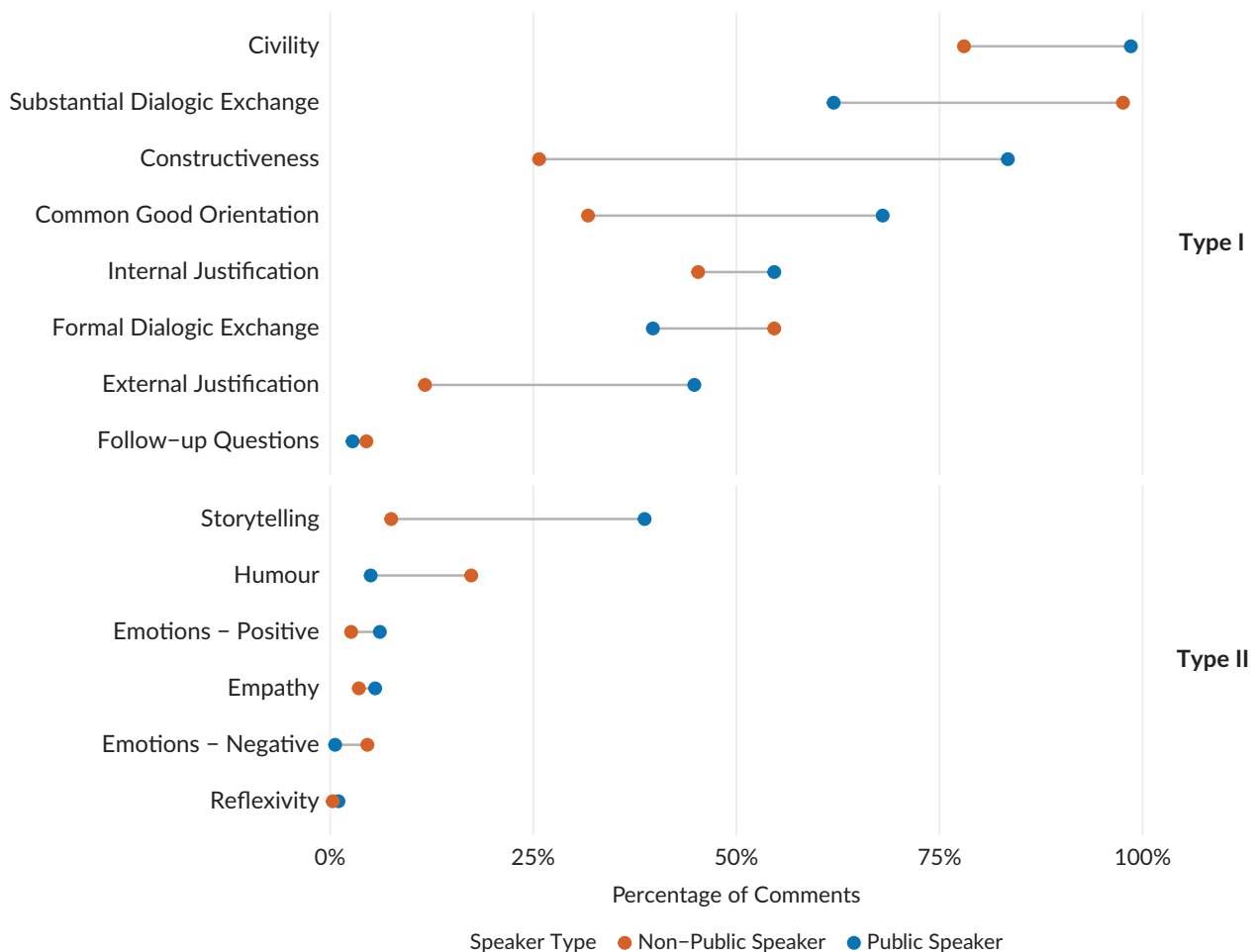


Figure 1. Prevalence of deliberative dimensions in climate change discourse. Note: Percentage of comments signaling the presence of the dimension (Class 1).

Concerning Type I dimensions, the groups show some similarities in their ways of providing justification, as they often employ their own reasoning (internal justification) to support their positions. Nevertheless, public speakers refer more often to traceable evidence (external justification) when backing up their statements, whereas debating users ask follow-up questions slightly more often, although this remains rare overall. Overall, these patterns suggest that both groups, at least partly, adhere to the deliberative norm of rational discourse, yet public speakers are more strongly oriented towards empirically grounded forms of legitimation. This may reflect their public role, in which references to scientific evidence are particularly relevant in the context of climate change as a highly contested issue. Although storytelling as an alternative form of justification is rare, it appears more often among public speakers (38.7%). This suggests that narrative communication may be perceived as a potential way of presenting views or news on climate change topics in some cases.

A more distinct pattern emerges for reciprocity. Debating users not only more frequently address other users directly (formal dialogic exchange), but also show substantially higher levels of substantial dialogic exchange; i.e., they engage more often with the content of prior arguments. However, this difference should be interpreted with caution, as public speaker comments often initiate discussions. Thus, the lower prevalence of reciprocity may reflect their conversational position rather than a unique communicative style.

In their interactions with other users, both groups show relatively high levels of civility, understood here as the absence of uncivil language. However, incivility should not merely be seen as problematic, as it can serve as a legitimate form of democratic contestation by oppressed groups (Bormann & Ziegele, 2023), a distinction reflected in our codebook. The pattern is particularly strong among public speakers, who remain civil in almost all cases (98.55%), whereas debating users demonstrate incivility in roughly one quarter of their comments.

Interestingly, both groups show low levels of emotions, while debating users slightly more often express negative emotions (4.58% vs. 0.61%), whereas public speakers more often show positive emotions (6.12% vs. 2.58%). Debating users also use humorous expressions (17.36% vs. 4.99%) more frequently. Furthermore, public speakers' comments are considerably more often constructive and oriented toward the common good, as they frequently contribute to X discussions in a productive way and state their positions with regard to collective concerns and well-being, both of which appear particularly relevant in climate discourse.

6. Discussion

6.1. RQ1

Our findings address a critical gap in the literature on the normative ideals of AI in deliberation. As Friess et al. (2025) note, while AI excels at detecting incivility and structuring arguments (Type I), efforts to address reciprocity and equality (Type II) remain sparse, often lacking grounding in normative democratic standards. By successfully operationalizing and measuring classical and alternative dimensions such as empathy and storytelling at scale, our approach provides a pathway for leveraging AI to foster more equitable and inclusive dialogue (RQ1). This is encouraging because it suggests that LLMs can successfully capture Type II dimensions, which have historically been more time-consuming to assess and more difficult to reliably measure than traditional dimensions such as rationality (e.g., Kies, 2022). Moreover, by comparing multiple open-weight models, we find that larger models, as well as some smaller models equipped with chain-of-thought reasoning capabilities, are able to navigate relatively subtle linguistic cues and strong class imbalances.

Our contribution extends beyond methodological innovation. Measuring deliberative quality and identifying the contexts in which deliberation breaks down through uncivil or otherwise norm-violating comments—and, even more importantly, in which such comments co-occur with deliberative contributions by other users—is crucial for devising effective counter-strategies against incivility and the deterioration of constructive dialogue. While the automated identification of uncivil comments is the most straightforward way to support counterspeakers using LLMs, counterspeech targets and relies on bystander activation, which is influenced by contextual cues such as the presence of communication norms within an online discussion (e.g., Obermaier & Schmid, 2026). Therefore, targeted intervention benefits from a nuanced evaluation of the deliberative quality of conversations, as this helps identify discourse environments in which

counterspeech is not only most needed but also likely to activate bystander support. As uncivil online communication poses a pressing challenge, with professional communicators being among those particularly affected (e.g., Reich & Bachl, 2026), our approach provides a foundational tool for identifying these incidents at scale, enabling more coordinated and timely interventions.

However, counterspeech is a fundamentally human process in which AI should not substitute counterspeakers (e.g., Ping et al., 2025). Recent research shows that counterspeakers harbor valid concerns about AI's potential negative effects on authenticity, agency, and intervention effectiveness (Mun et al., 2024). Our approach thus provides an initial basis for human-AI collaboration that can address these concerns while lowering barriers to counterspeech, such as the difficulty of identifying promising comment sections for effective interventions, while preserving ultimate human agency over the decision to intervene.

6.2. RQ2

Our analysis of climate-related discussions on X, comparing public speakers and debating users, indicates distinct communicative patterns: While public speakers appear more strongly aligned with civil, rational, constructive, and common-good-oriented forms of Type I deliberative communication, debating users seem to focus more on reciprocal engagement, justification through their own reasoning, and, at times, emotional and humorous styles of alternative deliberative communication (RQ2). Our findings are consistent with the existing literature, which indicates that comments by debating users often do not fully adhere to classical deliberative norms (Lasser et al., 2025; Naab et al., 2025).

This underlines that particular attention is required for debating user comments, as they often contain more non-deliberative communication (e.g., incivility), which may disproportionately affect public speakers. Although public speakers are more often targeted by uncivil communication (see also Reich & Bachl, 2026), they may also be seen as less vulnerable, which can reduce bystanders' sense of responsibility to intervene (Obermaier & Schmid, 2026). This highlights a central challenge: While both debating users and public speakers can be affected by uncivil comments and deliberative norm violations, our results indicate that counterspeech in support of public speakers appears particularly necessary, yet less likely to receive bystander support, given their perceived lower vulnerability.

7. Conclusion

We have sought to present a comprehensive strategy for measuring deliberative quality across both Type I and Type II dimensions using LLMs. Our results are promising, demonstrating that these tools allow us to assess complex, multidimensional constructs such as deliberative quality in a reliable and scalable fashion. Furthermore, we contribute substantive findings regarding how deliberative quality varies by selected contextual factors in online discussions on climate change, which are frequently characterized by intense politicization and affective polarization. Our analysis reveals distinct communicative patterns in how public and non-public speakers engage in this discourse, with public speakers more often aligning with Type I deliberative norms, while debating users more frequently display reciprocal and alternative deliberative styles and are more often uncivil. Beyond its methodological contributions, the LLM-based evaluation of deliberation on social media prepares the ground for automated tools to assist counterspeakers, such as civil society organizations, who need to monitor vast digital public spheres. This is achieved by identifying

discussion contexts in which uncivil communication occurs, but also prosocial, deliberative cues are present, thereby making it easier and less time-consuming to target interventions where both counterspeech is needed and bystander support may be possible. Ultimately, because the effectiveness of counterspeech acts is affected by multiple factors, our reliable, scalable, and theoretically grounded approach, which preserves human agency in the counterspeech process, should be seen as one complementary piece of this broader puzzle rather than as a stand-alone solution.

These findings need to be considered in light of certain limitations. First, the decision to perform stratified sampling in the second round of validation sampling probably led to an over-representation of examples that are easier for the model to detect than those in the full sample. Still, dealing with label imbalance is a necessary step in skewed label scenarios, and stratified sampling from label candidates is an established practice to mitigate its negative effects (Fedtke et al., 2026; Sun et al., 2026). Second, while the inclusion of dataset-specific label distribution information in the prompt proved beneficial for our classification in the present study, future work should explore the transferability of this strategy to datasets with different class distributions. Third, a small number of categories, such as emotions, proved more difficult to classify reliably for our selected model. This indicates, particularly in light of the better performance of pre-LLM approaches for this dimension (Bucher & Martini, 2024), that further refinement of LLM prompts and testing of different approaches are needed. Fourth, the findings on speaker-type-related communicative patterns remain descriptive for our broadly defined user groups, and emotions and the reciprocal sub-dimension of substantial dialogic exchange cannot be reliably identified by our model.

Looking forward, the deployment of AI tools must be guided by robust ethical frameworks and bottom-up approaches critically involving the perspectives of different user groups in order to address the negative effects of LLM biases and make their use in sensitive areas, such as deliberative discourse, safer and more trustworthy. Moreover, further research is needed to examine not only how and to what extent AI tools can be used effectively and ethically in counterspeech practices, but also how they are perceived by and affect particular user groups.

Acknowledgments

We sincerely thank the reviewers and editors for their valuable feedback, Raquel Silva from Cogitation Press for her continuous support, and our student assistants Karima Küster and Niklas Berger for their thoughtful support with manual data annotation.

Funding

The authors disclose receipt of the following financial support for the research, authorship, and publication of this article: this article is based on research conducted in the DFG/FWF funded Research Unit 5656 Communicative AI, project Political Discourse: ComAI and Deliberative Quality (TP 07), funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—516511468.

Conflict of Interests

The authors declare no conflict of interests.

Data Availability

The code used for the LLM-based content analysis will be made publicly available in a GitHub repository: <https://github.com/Leibniz-HBI/Can-AI-Gauge-Deliberative-Quality-Evaluating-Deliberation-on-Social-Media-with-Large-Language-Models.git>

LLMs Disclosure

For prompt generation and prompt optimization for the codebook for LLM-based content analysis: GPT 5.2; Qwen/Qwen3-235B-A22B-Thinking-2507 via SAIA (by GWDG Göttingen). Models for LLM-based content analysis: GPT-OSS:120B; GPT-OSS:20B; Llama3.3:70B; Llama3.1-8B-instruct; Qwen3:30B-thinking; Qwen3:30B-instruct; Gemma3-27B; Gemma4-26B. For bug fixing: Claude. For grammar correction and initial proofreading: Grammarly; DeepL pro version; GPT 5.6. For translation and qualitative analysis of German texts: DeepL pro version; Google Translate.

Supplementary Material

Supplementary material for this article is available online in the format provided by the author (unedited).

References

- Abd, S. E., Aggarwal, V., Algayres, R., Andreev, A., Bachem, O., Ballantyne, I., Brick, C., Cărbune, V., Casbon, M., Chaturvedi, M., Chawla, A., Cotruta, V., Coucke, A., Culliton, P., Dadashi, R., Dixon, L., Elhawaty, M., Evci, U., Farabet, C., . . . Zou, C. (2026). *Gemma 4 technical report* (version 2). arXiv. <https://doi.org/10.48550/arXiv.2607.02770>
- Agarwal, S., Ahmad, L., Ai, J., Altman, S., Applebaum, A., Arbus, E., Arora, R. K., Bai, Y., Baker, B., Bao, H., Barak, B., Bennett, A., Bertao, T., Brett, N., Brevdo, E., Brockman, G., Bubeck, S., Chang, C., Chen, K., . . . Zhao, S. (2025). *Gpt-oss-120b & gpt-oss-20b model card*. arXiv. <https://doi.org/10.48550/arXiv.2508.10925>
- Argyle, L. P., Bail, C. A., Busby, E. C., Gubler, J. R., Howe, T., Rytting, C., Sorensen, T., & Wingate, D. (2023). Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale. *Proceedings of the National Academy of Sciences*, 120(41), Article e2311627120. <https://doi.org/10.1073/pnas.2311627120>
- Bächtiger, A., Shikano, S., Pedrini, S., & Ryser, M. (2009, September 10-12). *Measuring deliberation 2.0: Standards, discourse types, and sequentialization* [Paper presentation]. ECPR General Conference, Potsdam, Germany.
- Bär, D., Maarouf, A., & Feuerriegel, S. (2024). *Generative AI may backfire for counterspeech*. arXiv. <https://doi.org/10.48550/arXiv.2411.14986>
- Behrendt, M., Wagner, S. S., Weinmann, C., Bormann, M., Warne, M., & Harmeling, S. (2026). *Machine learning for enhancing deliberation in online political discussions and participatory processes: A survey*. arXiv. <https://doi.org/10.48550/arXiv.2506.02533>
- Behrendt, M., Wagner, S. S., Ziegele, M., Wilms, L., Stoll, A., Heinbach, D., & Harmeling, S. (2024). *AQuA—Combining experts' and non-experts' views to assess deliberation quality in online discussions using LLMs*. arXiv. <https://doi.org/10.48550/arXiv.2404.02761>
- Bilewicz, M., Tempska, P., Leliwa, G., Dowgiałło, M., Tańska, M., Urbaniak, R., & Wroczyński, M. (2021). Artificial intelligence against hate: Intervention reducing verbal aggression in the social network environment. *Aggressive Behavior*, 47(3), 260–266. <https://doi.org/10.1002/ab.21948>
- Bohrunova, S. S. (2024). The concept of cumulative deliberation: Linking systemic approaches to healthier

- normativity in assessing opinion formation in online discussions. *Journal of the Association for Information Science and Technology*, 75(10), 1202–1215. <https://doi.org/10.1002/asi.24850>
- Bormann, M., & Ziegele, M. (2023). Incivility. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 199–217). Digital Communication Research. <https://doi.org/10.48541/dcr.v12.12>
- Bosley, M. (2025). Towards qualitative measurement at scale: A prompt-engineering framework for large-scale analysis of deliberative quality in parliamentary debates. *Journal of Political Institutions and Political Economy*, 6(3/4), 355–383. <https://doi.org/10.1561/113.00000128>
- Boulianne, S., & Theocharis, Y. (2020). Young people, digital media, and engagement: A meta-analysis of research. *Social Science Computer Review*, 38(2), 111–127. <https://doi.org/10.1177/0894439318814190>
- Brennan, R. L., & Prediger, D. J. (1981). Coefficient Kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement*, 41(3), 687–699. <https://doi.org/10.1177/001316448104100307>
- Bucher, M. J. J., & Martini, M. (2024). Fine-tuned 'small' LLMs (still) significantly outperform zero-shot generative AI models in text classification. arXiv. <https://doi.org/10.48550/ARXIV.2406.08660>
- Buerger, C. (2021a). #iamhere: Collective counterspeech and the quest to improve online discourse. *Social Media + Society*, 7(4). <https://doi.org/10.1177/20563051211063843>
- Buerger, C. (2021b). *Counterspeech: A literature review*. SSRN. <http://doi.org/10.2139/ssrn.4066882>
- Cypris, N. F., Engelmann, S., Sasse, J., Grossklags, J., & Baumert, A. (2022). *Intervening against online hate speech: A case for automated counterspeech* [Research brief]. Institute for Ethics in Artificial Intelligence. https://ieai.sot.tum.de/wp-content/uploads/2022/05/Research-Brief_Intervening-Against-Online-Hate-Speech_April2022_FINAL.pdf
- Duncan, S., Dwyer, L., Smith, H., Vallesi, D., Zeller, F., & Davis, C. (2026). Toward a computational mixed methods framework to measure online deliberative discourse. *Communication and the Public*, 11(2), 119–136. <https://doi.org/10.1177/20570473241284759>
- Esau, K. (2022). *Kommunikationsformen und Deliberationsdynamik: Eine relationale Inhalts- und Sequenzanalyse politischer Online-Diskussionen auf Beteiligungsplattformen*. Nomos. <https://doi.org/10.5771/9783748934226>
- Esau, K., & Friess, D. (2022). What creates listening online? Exploring reciprocity in online political discussions with relational content analysis. *Journal of Deliberative Democracy*, 18(1), 1–16. <https://doi.org/10.16997/jdd.1021>
- Esau, K., Wilms, L., Baleis, J., & Keller, B. (2023). For deliberation sake, show some constructive emotion! How different types of emotions affect the deliberative quality of interactive user comments. *Javnost—The Public*, 30(4), 472–495. <https://doi.org/10.1080/13183222.2023.2171217>
- Esteve-del-Valle, M. (2025). Social media, political polarization, and climate change: A systematic literature review. *Review of Communication Research*, 13, 106–121. <https://doi.org/10.52152/RCR.V13.S4>
- Farjam, M., Meyer, H., & Lohkamp, M. (2026). A practical guide and case study on how to instruct LLMs for automated coding during content analysis. *Social Science Computer Review*, 44(3), 488–502. <https://doi.org/10.1177/08944393251349541>
- Fedtke, C., Wiedemann, G., & Besio, C. (2026). Argument mining for organizational research: A computer-aided analysis of organizational talk. *Organizational Research Methods*. Advance online publication. <https://doi.org/10.1177/10944281261453948>
- Friess, D., & Eilders, C. (2015). A systematic review of online deliberation research: A review of online deliberation research. *Policy & Internet*, 7(3), 319–339. <https://doi.org/10.1002/poi3.95>
- Friess, D., Weinmann, C., & Warné, M. (2025). AI and deliberation: Normative ideals in the light of current AI research—A review. *Journal of Deliberative Democracy*, 21(1). <https://doi.org/10.16997/jdd.1805>

- Friess, D., Ziegele, M., & Heinbach, D. (2021). Collective civic moderation for deliberation? Exploring the links between citizens' organized engagement in comment sections and the deliberative quality of online discussions. *Political Communication*, 38(5), 624–646. <https://doi.org/10.1080/10584609.2020.1830322>
- Garland, J., Ghazi-Zahedi, K., Young, J.-G., Hébert-Dufresne, L., & Galesic, M. (2022). Impact and dynamics of hate and counter speech online. *EPJ Data Science*, 11(1), Article 3. <https://doi.org/10.1140/epjds/s13688-021-00314-6>
- Graham, T. (2010). The use of expressives in online political talk: Impeding or facilitating the normative goals of deliberation? In E. Tambouris, A. Macintosh, & O. Glassey (Eds.), *Electronic participation* (Vol. 6229, pp. 26–41). Springer. https://doi.org/10.1007/978-3-642-15158-3_3
- Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., . . . Ma, Z. (2024). *The Llama 3 herd of models*. arXiv. <https://doi.org/10.48550/arXiv.2407.21783>
- Gupta, K., Roychowdhury, S., Kasa, S. R., Kasa, S. K., Bhanushali, A., Pattisapu, N., & Murthy, P. S. (2023). *How robust are LLMs to in-context majority label bias?* arXiv. <https://doi.org/10.48550/ARXIV.2312.16549>
- Gupta, R., Desai, S., Goel, M., Bandhakavi, A., Chakraborty, T., & Akhtar, M. S. (2023). *Counterspeeches up my sleeve! Intent distribution learning and persistent fusion for intent-conditioned counterspeech generation*. arXiv. <https://doi.org/10.48550/arXiv.2305.13776>
- Habermas, J. (1996). *Between facts and norms: Contributions to a discourse theory of law and democracy*. MIT Press.
- Hickey, D., Fessler, D. M. T., Lerman, K., & Burghardt, K. (2025). X under Musk's leadership: Substantial hate and no reduction in inauthentic activity. *PLoS ONE*, 20(2), Article e0313293. <https://doi.org/10.1371/journal.pone.0313293>
- Jungherr, A., & Rauchfleisch, A. (2025). *Artificial intelligence in deliberation: The AI penalty and the emergence of a new deliberative divide*. arXiv. <https://doi.org/10.48550/arXiv.2503.07690>
- Jungherr, A., & Schroeder, R. (2023). Artificial intelligence and the public arena. *Communication Theory*, 33(2/3), 164–173. <https://doi.org/10.1093/ct/qtad006>
- Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., Rouillard, L., Mesnard, T., Cideron, G., Grill, J., Ramos, S., Yvinec, E., Casbon, M., Pot, E., Penchev, I., Liu, G., . . . Hussenot, L. (2025). *Gemma 3 technical report*. arXiv. <https://doi.org/10.48550/arXiv.2503.19786>
- Kathirgamalingam, A., Lind, F., & Boomgaarden, H. G. (2025). Measuring racism and related concepts using computational text-as-data approaches: A systematic literature review. *Annals of the International Communication Association*, 49(3), 241–256. <https://doi.org/10.1093/anncom/wlaf013>
- Kies, R. (2022). Online deliberative matrix. In S. A. Ercan, H. Asenbaum, N. Curato, & R. F. Mendonça (Eds.), *Research methods in deliberative democracy* (pp. 148–162). Oxford University Press. <https://doi.org/10.1093/oso/9780192848925.003.0010>
- Korre, K., Tsirmpas, D., Gkoumas, N., Cabalé, E., Kontarinis, D., Myrtzani, D., Evgeniou, T., Androutsopoulos, I., & Pavlopoulos, J. (2025). *Evaluation and facilitation of online discussions in the LLM era: A survey*. arXiv. <https://doi.org/10.48550/arXiv.2503.01513>
- Krippendorff, K. (2004). *Content analysis: An introduction to its methodology*. Sage.
- Lasser, J., Herderich, A., Garland, J., Aroyehun, S. T., Garcia, D., & Galesic, M. (2025). Collective moderation of hate, toxicity, and extremity in online discussions. *PNAS Nexus*, 4(11), Article pgaf369. <https://doi.org/10.1093/pnasnexus/pgaf369>
- Masullo, G. M., Riedl, M. J., & Huang, Q. E. (2022). Engagement moderation: What journalists should say to

- improve online discussions. *Journalism Practice*, 16(4), 738–754. <https://doi.org/10.1080/17512786.2020.1808858>
- Meyer, H., Peach, A. K., Guenther, L., Kedar, H. E., & Brüggemann, M. (2023). Between calls for action and narratives of denial: Climate change attention structures on Twitter. *Media and Communication*, 11(1), 278–292. <https://doi.org/10.17645/mac.v11i1.6111>
- Mun, J., Buerger, C., Liang, J. T., Garland, J., & Sap, M. (2024). Counterspeakers' perspectives: Unveiling barriers and AI needs in the fight against online hate. In F. F. Mueller, P. Kyburz, J. R. Williamson, C. Sas, M. L. Wilson, P. T. Dugas, & I. Shklovski (Eds.), *CHI '24: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Article 742). ACM. <https://doi.org/10.1145/3613904.3642025>
- Naab, T. K., Ruess, H.-S., & Küchler, C. (2025). The influence of the deliberative quality of user comments on the number and quality of their reply comments. *New Media & Society*, 27(1), 62–83. <https://doi.org/10.1177/14614448231172168>
- Obermaier, M., Buerger, C., Frischlich, L., Bund, L., Carathanassis, F., Clausen, A., Doseva, S., Eissfeldt, J., Garland, J., Ghazi-Zahedi, K., Grimme, C., Haim, M., Herderich, A., Li, Y., Oetting, H., Puschmann, C., Rho, E., Stoll, A., Wenninger, A., & Ziegele, M. (2026). *The role of AI in counterspeech: Assessing risks and potential* (bidt Analyses and Studies No. 21). Bavarian Research Institute for Digital Transformation. <https://doi.org/10.35067/XYPQ-KN79>
- Obermaier, M., & Schmid, U. K. (2026). Pay it forward? How target group and social norms affect online bystander intervention. *Political Communication*, 43(4), 650–671. <https://doi.org/10.1080/10584609.2025.2608808>
- Ping, K., Kumar, A., Ding, X., & Rho, E. H. (2025). Behind the counter: Exploring motivations and barriers of online counterspeech writing. *ACM Transactions on Computer-Human Interaction*, 32(5), Article 55. <https://doi.org/10.1145/3745769>
- Reich, S., & Bachl, M. (2026). When sexism becomes the norm: The effect of sexism on women's participation in political online discussions. *Communication Research*, 53(6), 844–866. <https://doi.org/10.1177/00936502251343988>
- Schäfer, M. S., & Painter, J. (2021). Climate journalism in a changing media ecosystem: Assessing the production of climate change-related news around the world. *WIREs Climate Change*, 12(1), Article e675. <https://doi.org/10.1002/wcc.675>
- Schmid, U. K., & Greipl, S. (2025). The thin line between harmful and benign: Perceptions of humorous and non-humorous hate speech in humorous and neutral social media contexts. *Media Psychology*. Advance online publication. <https://doi.org/10.1080/15213269.2025.2580291>
- Schmidt, J.-H., Merten, L., & Münch, F. V. (2023). *Die „Datenbank Öffentlicher Sprecher:innen“ (DBÖS)* [Data set]. OSF. <https://doi.org/10.17605/OSF.IO/SK6T5>
- Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A. J., Ananthram, A., Nathan, A., Luo, A., Helyar, A., Madry, A., Efremov, A., Spyra, A., Baker-Whitcomb, A., Beutel, A., Karpenko, A., . . . Wang, Z. (2025). *OpenAI GPT-5 system card*. arXiv. <https://doi.org/10.48550/arXiv.2601.03267>
- Steenbergen, M. R., Bächtiger, A., Spörndli, M., & Steiner, J. (2003). Measuring political deliberation: A discourse quality index. *Comparative European Politics*, 1, 21–48. <https://doi.org/10.1057/palgrave.cep.6110002>
- Stoll, A., Yu, J., Andrich, A., & Domahidi, E. (2025). Classification bias of LLMs in detecting incivility towards female and male politicians in German social media discourse. *Communication Methods and Measures*, 19(4), 350–368. <https://doi.org/10.1080/19312458.2025.2551693>

- Stolwijk, S. B., Boukes, M., Yeung, W. N., Liao, Y., Munker, S., Kroon, A. C., & Trilling, D. (2025). Can we use automated approaches to measure the quality of online political discussion? How to (not) measure interactivity, diversity, rationality, and incivility in online comments to the news. *Communication Methods and Measures*, 20(1), 1–25. <https://doi.org/10.1080/19312458.2025.2553300>
- Summers, E., Brigadir, I., Hames, S., van Kemenade, H., Binkley, P., tinafigueroa, Ruest, N., Walmir, Chudnov, D., Thiel, D., Betsy, Chartier, R., celeste, Lin, H., Alice, Chosak, A., Lenz, M., McCain, R. M., Milligan, I., . . . Elwert, F. (2023). *DocNow/twarc: V2.14.0* [Computer software]. Zenodo. <https://doi.org/10.5281/zenodo.7799050>
- Sun, L., Pendyala, V., Chuang, Y.-S., Yang, S., Feldman, J., Zhao, A., Choudhury, M. D., Yang, S., & Shah, D. V. (2026). When more shots don't help: LLM sensitivity and variability in social media annotation and stance detection of health information. *Computational Communication Research*, 8(1), 1–30. <https://doi.org/10.5117/CCR2026.1.4.SUN>
- Theocharis, Y., Barberá, P., Fazekas, Z., & Popa, S. A. (2020). The dynamics of political incivility on Twitter. *Sage Open*, 10(2). <https://doi.org/10.1177/2158244020919447>
- Törnberg, P. (2023). *How to use LLMs for text analysis*. arXiv. <https://doi.org/10.48550/arXiv.2307.13106>
- Törnberg, P. (2024). *Best practices for text annotation with large language models*. arXiv. <https://doi.org/10.48550/arXiv.2402.05129>
- Wachsmuth, H., Lapesa, G., Cabrio, E., Lauscher, A., Park, J., Vecchi, E. M., Villata, S., & Ziegenbein, T. (2024). *Argument quality assessment in the age of instruction-following large language models*. arXiv. <https://doi.org/10.48550/arXiv.2403.16084>
- Wu, C., Wang, Y., Zhang, Y., Wang, H., & Pang, Y. (2025). Confront hate with AI: How AI-generated counter speech helps against hate speech on social Media? *Telematics and Informatics*, 101, Article 102304. <https://doi.org/10.1016/j.tele.2025.102304>
- Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., . . . Qiu, Z. (2025). *Qwen3 technical report*. arXiv. <https://doi.org/10.48550/arXiv.2505.09388>
- Ziegele, M., & Eilders, C. (2022). I like Deliberation? Zusammenhänge zwischen der deliberativen Qualität von Nutzerbeiträgen und ihrer Bewertung mittels Likes und Dislikes durch andere Nutzer:innen. In F. Gerlach & C. Eilders (Eds.), *#meinfernsehen 2021* (pp. 197–216). Nomos. <https://doi.org/10.5771/9783748928690-197>
- Ziegele, M., Jost, P., Bormann, M., & Heinbach, D. (2018). Journalistic counter-voices in comment sections: Patterns, determinants, and potential consequences of interactive moderation of uncivil user comments. *Studies in Communication and Media*, 7(4), 525–554. <https://doi.org/10.5771/2192-4007-2018-4-525>

About the Authors



Patrick Frey is a PhD researcher in communication science at the Centre for Media, Communication and Information Research (ZeMKI) at the University of Bremen. He works as a research assistant in the project Political Discourse: ComAI and Deliberative Quality in the ComAI Research Unit.



Smita Bhattacharya is a PhD researcher in natural language processing at the Media Research Methods Lab of Leibniz Institute for Media Research | Hans-Bredow-Institut. She works as a research assistant in the project Political Discourse: ComAI and Deliberative Quality in the ComAI Research Unit.



Gregor Wiedemann (PhD) is a senior researcher in computational social science at the Leibniz Institute for Media Research | Hans-Bredow-Institut, where he heads the Media Research Methods Lab. He co-leads the project Political Discourse: ComAI and Deliberative Quality in the ComAI Research Unit, leading social-bot development.



Cornelius Puschmann is professor of communication and media studies with a focus on digital communication at ZeMKI and leader of the Digital Communication and Information Diversity (DCID) lab. He co-leads the project Political Discourse: ComAI and Deliberative Quality in the ComAI Research Unit, leading AI-based content analysis.