Media and Communication
2026 • Volume 14 • Article 12692
https://doi.org/10.17645/mac.12692

ARTICLE

Open Access Journal

Counterspeech or Disagreement? Comparing Oppositional Replies in China's AI Controversies on Social Media

Lishuang Wei and Chuanlin Ning

School of Media and Communication, Shanghai Jiao Tong University, China

Correspondence: Chuanlin Ning (ningchuanlin@sjtu.edu.cn)

Submitted: 4 May 2026 **Accepted:** 6 October 2026 **Published:** 12 November 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

Online technological controversies often involve both ordinary disagreement and counterspeech, yet these forms of opposition are rarely compared within the same discussion environment. Drawing on argumentation theories, this study compares disagreement replies to non-hateful comments with counterspeech replies to hateful comments in Douyin discussions of China's AI controversies, where hate speech may include stigmatizing generalizations about social groups. Using 9,407 oppositional parent–response dyads, we compared like counts, toxicity, subsequent toxic replies, and the distribution and reception of six response types. Counterspeech and disagreement did not differ significantly in like counts, but counterspeech exhibited higher toxicity and was related to greater odds of toxicity among direct replies posted within seven days. Despite sharing the same six response types, the two response contexts varied in their type distributions and associations with likes. These findings situate counterspeech within a broader spectrum of online opposition, revealing contextual differences in its argumentative expression and subsequent interaction across hateful and non-hateful contexts.

Keywords

artificial intelligence; China; counterspeech; disagreement; social media; stigmatization; toxicity

1. Introduction

Knowledge distillation is an established machine-learning technique. Developed to compress complex models, it enables a "student" model to approximate the outputs, representations, or reasoning patterns of a more capable "teacher" (Beyer et al., 2022; Gou et al., 2021; Hinton et al., 2015). In the era of large language models (LLMs), distillation reduces computational costs, improves deployment efficiency, and transfers capabilities across architectures (Fang et al., 2026; C. Yang et al., 2025). Yet its public meaning is less settled.

© 2026 by the author(s), licensed under a Creative Commons Attribution 4.0 International License (CC BY).

When model development intersects with commercial competition, intellectual property disputes, and geopolitical rivalry, distillation may be recast as imitation, copying, or theft. Recent controversies surrounding Chinese LLMs illustrate this transformation, with DeepSeek the most visible case. In January 2025, Microsoft and OpenAI reportedly investigated whether actors linked to DeepSeek had used OpenAI outputs for model development ("Microsoft probes if," 2025). In February 2026, Anthropic accused DeepSeek, Moonshot AI, and MiniMax of large-scale distillation aimed at extracting capabilities from Claude ("Chinese AI companies," 2026). Irrespective of legal substantiation, the debate became a dispute over what counts as innovation or plagiarism and whether accusations against Chinese AI constitute evidence-based criticism or a geopolitical narrative.

Douyin provided a visible arena for these controversies. As a major short-video platform in China, it brings technology news, influencer commentary, nationalist discourse, and user debate together through comment sections (Ning, 2026). Users did not simply repeat either accusations of copying or unconditional defenses of domestic AI. They questioned sources, demanded evidence, explained model training, compared AI systems, and situated allegations within China–US competition. Other comments extended judgments about models into derogatory claims about national groups or collective competence. Research likewise shows that factual disputes over science and technology can become intertwined with identity conflict, nationalism, and negative group evaluations (Chen et al., 2022; Li et al., 2026; Tian & Ge, 2024). Discussions of Chinese LLMs therefore contain disagreement and responses opposing hateful generalizations.

Counterspeech and disagreement lend insight into the forms of discursive opposition that unfold in such controversies and the social feedback these responses garner. Argumentation theories conceptualize opposition as communicative moves challenging a standpoint, its premises or evidence, or the credibility of its advocate (van Eemeren & Grootendorst, 2004). This premise underpins the argument-mining division between source-oriented and content-oriented rebuttals (Habernal et al., 2018; Jo et al., 2020; Wachsmuth et al., 2018). In online disputes, author-directed responses encompass "personal attack" and "affiliation accusation," while content-directed responses comprise "contradiction," "justification," "request," and "counterclaim" (Muntigl & Turnbull, 1998; X. Yu et al., 2023). This taxonomy clarifies how users contest a speaker's competence or interests, reject a claim, request evidence, articulate reasons, or advance an alternative account. Counterspeech and disagreement may share this repertoire, although their antecedents are not equivalent. Counterspeech opposes hateful content, while disagreement opposes non-hateful content (Cepollaro et al., 2023; Pomerantz, 1984). Scholarship on these response contexts has largely proceeded along separate trajectories, leaving their response types, audience feedback, toxicity, and subsequent interactions underexplored within a comparison that holds opposition constant.

Against this background, the study analyzes counterspeech and disagreement in Douyin discussions of controversies surrounding Chinese LLMs, with DeepSeek serving as a representative case within the controversy. The analysis compares the two response contexts in like counts, toxicity, and response types in 9,407 oppositional parent–response dyads. By restricting the comparison to oppositional replies while varying whether the antecedent was hateful, this study clarifies the boundary between counterspeech and disagreement and integrates them within a shared argumentation taxonomy. It further reveals how response context and response type relate to social-media feedback and subsequent toxicity in public controversies surrounding Chinese LLMs, providing a relational understanding of online opposition that connects antecedent contexts, argumentative responses, and subsequent interactions.

2. Literature Review

2.1. Boundaries Between Counterspeech and Disagreement

Counterspeech is defined by its function of responding to harms associated with prior speech. Cepollaro et al. (2023) describe it as "communication that seeks to counteract potential harm that is brought about by other speech," centering the relation between speech and potential harm. This boundary distinguishes counterspeech from ordinary disagreement. Disputing a claim or expressing an opposing position does not make a response counterspeech, since disagreement is common in public discussion and does not require a harmful antecedent (Angouri & Locher, 2012; Pomerantz, 1984; Sifianou, 2012). In online hate research, counterspeech is understood as an oppositional response to hate speech. Its identification depends on the antecedent and the stance of reply. The preceding comment contains hate speech, while the reply challenges, rejects, or opposes that hateful content (Benesch et al., 2016; Garland et al., 2020). Counterspeech and disagreement share an oppositional relation to parent discourse, although this study limits disagreement replies to those directed at non-hateful comments.

Existing research has increasingly recognized disagreement as central to counterspeech, but has rarely examined counterspeech alongside disagreement in non-hateful contexts. X. Yu et al. (2023) use disagreement as the primary criterion for classifying countering from non-countering replies within hate-speech conversations. Replies that disagree with a hateful comment are classified as countering, while replies that agree with the hate or remain neutral are non-countering. Disagreement therefore identifies opposition to hate speech within their framework; it is not established as a separate comparison condition. This approach connects counterspeech with argumentation research, but all antecedents remain hateful comments, leaving disagreement directed at non-hateful comments outside the comparison. A similar limitation applies to studies contrasting counterspeech with non-counterspeech. Non-counterspeech can include agreement, neutrality, continued hate, and other responses without an oppositional stance, so it cannot represent ordinary disagreement (Mathew et al., 2019; X. Yu et al., 2023). Evidence therefore offers limited insight into whether counterspeech exhibits response patterns distinct from those of other oppositional replies once opposition itself is held constant.

Technological controversies surrounding plagiarism allegations against Chinese LLMs provide a useful public-discussion setting for examining the relationship between counterspeech and disagreement. Research on contentious scientific discussion suggests that disagreements over factual claims can develop alongside identity-based divisions and negative group evaluations (Chen et al., 2022). A similar pattern appears in Chinese social media discussions of the Apple 5G controversy, where disputes over technical performance and the credibility of technological claims became connected with nationalist, class, and gender discourse, including derogatory evaluations of particular user groups (Tian & Ge, 2024). These studies suggest that technological controversy can accommodate both ordinary public disagreement and disputes involving hateful expression within the same discussion environment. Controversies surrounding Chinese LLMs offer such a setting, where technical and commercial debates intersect with geopolitical narratives, national identity, and more antagonistic forms of exchange (Li et al., 2026). This coexistence makes the controversy suitable for examining how counterspeech relates to disagreement under a shared topical and platform context. Accordingly, this study uses this controversy as an analytical setting to explore the similarities and differences between counterspeech replies to hateful comments and disagreement replies to non-hateful comments.

2.2. Response Types Across Counterspeech and Disagreement

Argumentation theories provide the basis for classifying oppositional replies into response types. Argumentation becomes relevant when a standpoint is questioned, rejected, or confronted with a competing position. Participants may deny a proposition, challenge its premises or evidence, question the link between reasons and conclusions, provide reasons for an opposing position, or advance an alternative claim. Pragma-dialectical theory conceptualizes these as communicative moves that advance, challenge, or support standpoints (van Eemeren & Grootendorst, 2004). Argument mining further specifies the object of opposition. Content-oriented rebuttals may address premises, conclusions, or reasoning, whereas source-oriented rebuttals may question the speaker's credibility (Alshomary et al., 2021; Habernal et al., 2018; Jo et al., 2020; Wachsmuth et al., 2018).

Research on disagreement gives empirical form to this repertoire. Muntigl and Turnbull (1998) identify irrelevancy claims, challenges, contradictions, and counterclaims, showing that opposition may reject a proposition, question its grounds or relevance, or introduce a competing position. Pragmatic research indicates that disagreement can range from mitigated contestation to explicit confrontation, with its expression contingent on social relations and communicative norms (Angouri & Locher, 2012; Rees-Miller, 2000). Similar patterns appear online. Reddit research places disagreement alongside elaboration and negative reaction, suggesting that contesting a position and expressing hostility toward another participant represent related but non-identical functions (Zhang et al., 2017). Online debate corpora also capture questioning versus asserting, factual versus emotional, and reasonable versus attacking replies, indicating that response form and tone constitute separate facets (Sridhar & Getoor, 2019; Walker et al., 2012).

Counterspeech research reveals a broad repertoire when opposition is directed at harmful speech. Existing taxonomies include factual correction, exposure of inconsistency, warnings, affiliation, condemnation, humor, positive tone, and hostile language (Benesch et al., 2016; Mathew et al., 2019). Argumentative resources remain salient, with reasoning, historical information, statistics, and examples often coexisting with hostility or negative emotion (Baider, 2023). More recent frameworks align counterspeech with general argumentation by separating author-directed replies such as insults and accusations from content-directed claims, justifications, and requests (X. Yu et al., 2023). Discourse-informed work likewise identifies correction, clarification, explanation, probing questions, and elaboration (Hassan & Alikhani, 2023). Therefore, disagreement and counterspeech may share a common classification framework, but evidence on the actual deployment of these response types across the two contexts remains limited.

2.3. Users' Reactions to Counterspeech and Disagreement

Likes constitute a lightweight form of social-media feedback and an observable dimension of user engagement, although the meanings attached to liking behavior are multiple and context-dependent (Hayes et al., 2016). In interactions surrounding online hate, such feedback also provides a visible means for bystanders to participate in the exchange, as users can endorse existing counterspeech through popularity cues such as likes (Obermaier, 2024). Existing evidence further suggests that counterspeech and other comments may elicit different levels of such platform reactions. On YouTube, for example, counterspeech received more likes overall than non-counterspeech comments (Mathew et al., 2019). Likes thus provide an observable platform-level indicator for examining whether counterspeech replies and disagreement replies garner different levels of like-based engagement.

Toxicity provides a lens for understanding how opposition is expressed and how the interaction subsequently unfolds. An oppositional stance does not necessarily correspond to civil or benign language. Counterspeech may contain insults, hostile expressions, and attacks on the speaker, while replies to hate speech can also convey opposition through accusations and other aggressive formulations (Mathew et al., 2019; X. Yu et al., 2023). Comparable forms of hostility can occur in ordinary disagreement, making the toxicity of the focal reply a relevant linguistic characteristic beyond its oppositional stance. Toxicity may also extend into subsequent interaction. Research has linked the linguistic character of counterspeech to later incivility, negative conversational tone, and continued hostility, with more toxic counterspeech more likely to be followed by hostile responses within the same discussion (Baider, 2023; Hickey et al., 2026). Social responses to online hate have been associated with changes in the toxicity of users' subsequent messages, showing that reactions within an exchange can correspond with later hateful expression (Y. Yang et al., 2025). More general research on online conversations similarly finds that toxic posts are more likely to elicit toxic replies and that early conversational characteristics are related to the toxicity of subsequent responses (Saveski et al., 2021). Examining both the toxicity of oppositional replies and that of subsequent replies therefore captures how counterspeech and disagreement are expressed at the focal turn and how the ensuing interaction develops. The literature reviewed above indicates that counterspeech and disagreement share an oppositional relation to prior discourse and appear to draw on overlapping response forms, yet have rarely been directly compared as distinct response contexts. Specifically, existing research has documented user reactions and conversational outcomes surrounding counterspeech and disagreement separately, but offers limited evidence on how these two response contexts compare in the reactions they elicit and the interactions that follow. Likewise, although their taxonomies point to overlap in response types, little is known about the similarities and differences in how these response types are used and received across the two contexts. These gaps motivate the following research questions:

> RQ1: In online discussions of technological controversies, how do counterspeech replies and disagreement replies differ in the like counts they receive?
>
> RQ2: In online discussions of technological controversies, how do counterspeech replies and disagreement replies differ in their toxicity levels and in the likelihood that subsequent direct replies are toxic?
>
> RQ3: In online discussions of technological controversies, what types of responses characterize counterspeech and disagreement, and how do the associations between response type and like counts differ between the two?

3. Methods

3.1. Data Collection

This study investigates publicly available Douyin comment threads related to controversies over Chinese AI, particularly DeepSeek. Since the Douyin API was not accessible, data were collected using Playwright from publicly accessible Douyin video pages and their comment sections. To capture the controversy, we paired AI-related search terms, including "DeepSeek" and "domestic AI," with controversy-related terms such as "plagiarism," "wrapper model," "theft," and "distillation." Under each search query, the 100 videos with the highest like counts were selected, and their comment threads were subsequently collected for analysis.

The initial dataset contained 53,735 comments. We excluded duplicate and empty entries, as well as comments composed solely of emojis, @-mentions, or pictures, which produced a corpus of 334 videos and 43,887 valid comments. We then reconstructed direct reply relationships and organized comments into parent–response dyads. The parent comment refers to the comment being addressed and the response comment to its direct reply. Dyads in which both comments were posted by the same user were deleted. The resulting corpus contained 22,815 parent–response dyads.

3.2. Measures

3.2.1. Response Context

Response context was measured as a binary categorical variable distinguishing counterspeech replies from disagreement replies according to the type of parent comment they opposed. Counterspeech replies responded to hate-speech parent comments, whereas disagreement replies responded to non-hateful parent comments.

Coding proceeded in two stages. First, we used DeepSeek V4 Pro to classify each response comment as opposition, support, or neutral toward its parent comment. The full parent comment was provided as context, while the label applied only to the response, since the preceding message can shape the interpretation of counterspeech (X. Yu et al., 2022). The LLM annotation instructions followed a codebook structure used in human content analysis, with explicit category definitions, coding criteria, and examples (Farjam et al., 2026). We further incorporated boundary conditions for ambiguous cases and few-shot examples adapted to the linguistic features of our corpus. An overview of the annotation instructions is provided in the Supplementary File.

Following the hybrid validation procedure outlined by Larsson (2026), two human coders independently coded 400 parent–response dyads after one week of training involving coding-rule review, practice coding, discussion of disagreements, and examination of boundary cases. Intercoder reliability was Krippendorff's $\alpha = .80$. After disagreements were adjudicated through discussion, agreement between the adjudicated human labels and the model classifications was $\alpha = .78$. Subsequent analyses retained the 9,407 dyads classified as opposition.

Second, parent comments in these oppositional dyads were classified for hate speech using a Chinese toxicity classifier trained on ToxiCN (Lu et al., 2023), a Chinese toxic-language benchmark with annotations for general offensiveness, hate speech, target groups, and explicit or implicit toxic expression. The knowledge-enhanced RoBERTa model achieved a weighted F1 of approximately .81 for binary toxicity identification and .87 for distinguishing hate speech from general offensive language in our test. The classifier additionally produces both a binary label and a continuous toxicity probability from 0 to 1, with higher scores indicating a greater probability of toxic expression.

Consistent with the definition of counterspeech as a direct response opposing hateful content (Chung et al., 2024; Y. Yang et al., 2025), an oppositional response directed at a parent comment classified as hate speech was defined as a counterspeech reply. An oppositional response to a non-hateful parent comment was defined as a disagreement reply. Finally, the dataset contained 776 counterspeech replies and 8,631 disagreement replies.

3.2.2. Response Type

After examining recurring expressions in our corpus, we developed a taxonomy combining Muntigl and Turnbull's (1998) typology of disagreement with X. Yu et al.'s (2023) division between responses addressing the author and those addressing the content. Counter author included personal attack and affiliation accusation. Personal attack disparaged the author's knowledge, competence, credibility, or judgment, while affiliation accusation attributed the author's position to presumed motives, interests, affiliations, or evaluative bias. Counter content comprised contradiction, justification, request, and counterclaim. Contradiction directly rejected or corrected the parent claim without providing supporting reasons or evidence. Justification supported opposition with reasons or evidence. Request challenged the parent claim by seeking clarification, evidence, or justification. Counterclaim advanced an alternative claim or perspective, including a different causal explanation or evaluative standard. The same taxonomy was applied to counterspeech and disagreement replies.

The annotation instructions followed the same general structure used for stance annotation, specifying the coding task, category system, decision criteria, and output requirements. Additional boundary conditions addressed overlaps among response types, and few-shot examples illustrated recurrent and ambiguous cases in the corpus. DeepSeek V4 Pro assigned each response one primary type and, where applicable, one secondary type based on the parent–response dyad. RQ3 uses the primary type as the mutually exclusive categorical measure. Reliability was assessed using a stratified sample of 600 dyads covering both response contexts and all six types. Krippendorff's α was .79 between the two human coders and .75 between the adjudicated human labels and model classifications.

3.2.3. Like Counts

We measured engagement using the number of likes received by each response comment. Douyin provides cumulative comment-level like counts but does not provide timestamps for individual likes. Therefore, comment exposure time (days) was included as a control variable to account for differences in the time available to accumulate likes.

3.2.4. Toxicity

Response-comment toxicity was measured using the ToxiCN classifier's continuous toxicity probability. Subsequent interaction was measured at the level of individual direct replies posted within seven days of the response comment. Each observed direct reply was assigned a binary toxicity label indicating whether it was classified as toxic.

3.2.5. Covariates

All regression models included four covariates: parent comment likes, parent–response posting interval, comment exposure time, and response comment length. The temporal measures were recorded in days, and response length was measured in non-whitespace characters. The four covariates were log-transformed using $\ln(1 + x)$ and then z-standardized.

3.3. Analytical Approach

For RQ1, a hurdle negative binomial model was estimated for likes. The logit component estimated the probability that a response received any like, while the zero-truncated negative binomial component modeled positive like counts. Response context—counterspeech versus disagreement—was the principal predictor, with the four covariates described above included.

For RQ2, the distribution of continuous toxicity scores was first compared between counterspeech and disagreement replies using a Mann–Whitney *U* test because the scores were strongly skewed. Toxicity in subsequent interaction was examined with a mixed-effects logistic regression predicting whether a subsequent direct reply was toxic. The model included response context and the same covariates, with random intercepts for parent comments to account for dependence among replies within the same conversational environment.

For RQ3, a Pearson χ^2 test assessed whether the distribution of the six response types differed between counterspeech and disagreement replies. Hurdle negative binomial models then examined likes as a function of response type, response context, and their interaction, controlling for the same covariates. Because responses posted under the same parent comment may share unobserved conversational and exposure conditions, standard errors were estimated using cluster-robust adjustment at the parent-comment level.

4. Results

The parent comments identified through our classification showed that technological controversies involved originality, intellectual property, model performance, corporate rivalry, and national competition. Hate speech appeared when these disputes became identity-based judgments, linking technological conduct to collective morality, national competence, or political loyalty, sometimes extending to prejudice based on region, gender, race or ethnicity, and sexual orientation.

To address RQ1, diagnostic checks supported the use of the hurdle negative binomial model for like counts, with a maximum variance inflation factor of 1.052 indicating no multicollinearity concerns. The results showed no significant difference between counterspeech and disagreement replies in either component of the model. Counterspeech and disagreement replies did not differ significantly in the probability of receiving any likes ($b = .068$, $SE = .11$, odds ratio [*OR*] = 1.070, $p = .536$). Among replies receiving at least one like, their positive like counts also did not differ significantly ($b = -.202$, $SE = .159$, incidence rate ratio [*IRR*] = 0.817, $p = .203$).

RQ2 examined whether counterspeech and disagreement replies differed in their own toxicity and in the toxicity of subsequent direct replies. A Mann–Whitney *U* test showed that counterspeech replies had significantly higher ToxiCN toxicity probabilities than disagreement replies, $U = 3{,}514{,}453.5$, $p < .05$.

A similar contrast appeared in subsequent interactions. Among direct replies, 15.32% of those following counterspeech were classified as toxic, compared with 9.68% of those following disagreement. Logistic regression confirmed this pattern, showing that within seven days, direct replies to counterspeech had significantly higher odds of being toxic than those following disagreement replies ($b = .531$, $SE = .163$, $OR = 1.700$, $p < .01$). As depicted in Figure 1, the model-predicted probability of a toxic direct reply

increased from 9.0% for disagreement replies (95% CI [7.9%, 10.3%]) to 14.4% for counterspeech replies (95% CI [11.0%, 18.8%]).

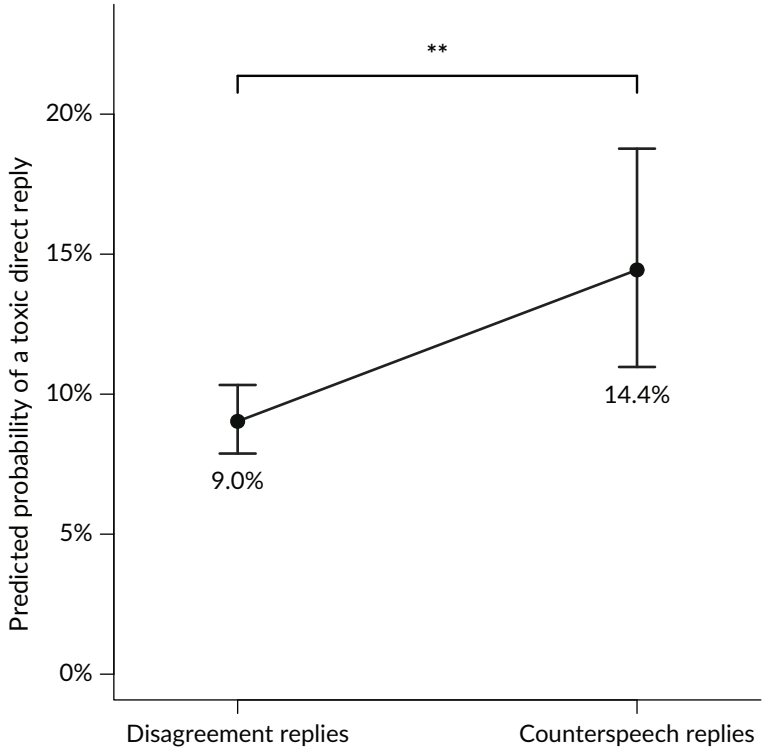


Figure 1. Predicted probability of toxicity in subsequent direct replies by response context. Notes: Error bars represent 95% confidence intervals; ** $p < .01$.

For RQ3, as shown in Table 1, the distribution of response types differed significantly between counterspeech and disagreement replies, $\chi^2(5) = 16.16$, $p < .01$. Post hoc tests using adjusted standardized residuals showed that justification, the most common type in both contexts (27.23% vs. 27.19%, $p_{adj} = 1.000$), was typically characterized by technical definitions, performance comparisons, concrete examples, and causal reasoning. Personal attack (19.81% vs. 19.97%, $p_{adj} = 1.000$) commonly questioned the other user's knowledge, competence, or credibility. Compared with disagreement replies, counterspeech replies contained a higher proportion of requests (20.87% vs. 25.77%, $p_{adj} < .01$), but a lower proportion of contradictions (18.82% vs. 14.43%, $p_{adj} < .05$). Affiliation accusation (5.90% vs. 5.15%, $p_{adj} = 1.000$) and

Table 1. Percentage distribution of response types in disagreement and counterspeech replies.

Response Target	Response Type	Disagreement (%)	Counterspeech (%)	Difference (pp)
Counter Author	Personal Attack	19.81	19.97	+0.16
	Affiliation Accusation	5.90	5.15	−0.74
Counter Content	Contradiction	18.82	14.43	−4.38*
	Justification	27.23	27.19	−0.04
	Request	20.87	25.77	+4.91**
	Counterclaim	7.38	7.47	+0.09

Notes: Difference (pp) represents counterspeech minus disagreement; significance markers are based on post hoc tests using adjusted standardized residuals with Holm-adjusted p-values; * $p_{adj} < .05$, ** $p_{adj} < .01$, *** $p_{adj} < .001$.

counterclaim (7.38% vs. 7.47%, p_{adj} = 1.000) occurred at similar rates across the two contexts, with the former attributing positions to presumed motives, interests, or affiliations and the latter advancing alternative causal or evaluative accounts involving markets, institutions, policy, or technological competition.

Among disagreement replies (see Table 2), affiliation accusation showed the clearest advantage over justification, both in the likelihood of receiving at least one like (*OR* = 1.719, $p < .001$) and in the number of likes received among replies with at least one like (*IRR* = 2.827, $p < .001$). Examples of affiliation accusation replies included comments that attributed the parent commenter's position to presumed interests, affiliations, or evaluative biases, such as "You're exactly the kind of person they love—anything for clicks and attention," and comments that framed the parent commenter's comparisons of Chinese and foreign technologies as reflecting a pro-US bias, as in "America wins again—could you come up with a new way for it to 'win'?"

By contrast, contradiction and request were associated with both lower odds of receiving likes (*OR* = 0.793, $p < .01$; *OR* = 0.872, $p < .05$, respectively) and fewer likes among replies that received at least one like (*IRR* = 0.726, $p < .05$; *IRR* = 0.737, $p < .05$, respectively) than justification. These replies were often brief, offering direct denial or challenge, such as "That's way off!" or "Which model has copied a Chinese one?" Justification, in contrast, more often developed an explicit argumentative chain, for instance by distinguishing distillation from reinforcement learning and using European and Japanese cases as comparative evidence.

Table 2. Hurdle negative binomial models predicting like counts by response type among disagreement replies.

	Stage 1: Binary Component			Stage 2: Count Component		
	b	*SE*	*OR*	*b*	*SE*	*IRR*
Personal Attack	−0.090	0.071	0.914	0.150	0.153	1.162
Affiliation Accusation	0.542***	0.110	1.719	1.039***	0.263	2.827
Contradiction	−0.232**	0.075	0.793	−0.320*	0.147	0.726
Request	−0.137*	0.069	0.872	−0.305*	0.144	0.737
Counterclaim	0.036	0.094	1.037	0.075	0.223	1.078
Parent comment likes (log)	0.409***	0.091	1.506	1.193***	0.077	3.296
Parent–response posting interval (log)	−0.384***	0.045	0.681	−0.563***	0.075	0.569
Comment exposure time (log)	0.105**	0.039	1.111	0.116*	0.049	1.123
Comment length (log)	0.011	0.026	1.011	0.004	0.061	1.004
McFadden R^2			.048			.088

Notes: Justification is the reference category; * $p < .05$, ** $p < .01$, *** $p < .001$.

Table 3 displays a different pattern for counterspeech replies. None of the response types significantly predicted whether a reply received any likes relative to justification. Differences emerged only in the number of likes received among replies with at least one like. Personal attack was associated with substantially fewer likes than justification (*IRR* = 0.357, $p < .01$), as was counterclaim (*IRR* = 0.342, $p < .05$). Affiliation accusation, contradiction, and request did not differ significantly from justification.

Personal attack targeted the parent commenter's intelligence, knowledge, or judgment. Examples included "This has to be the dumbest take yet. Do you even know what open source means?" "What degree do you

even have?" and "You think throwing money at everything magically solves it. You're so smart." Counterclaim more often reframed key terms in the parent comment. For example, "Bloodlines are for pets" recast the notion of an "AI bloodline" as the language of pet breeding, satirizing bloodline-based technological nationalism.

Table 3. Hurdle negative binomial models predicting like counts by response type among counterspeech replies.

	Stage 1: Binary Component			Stage 2: Count Component		
	b	*SE*	*OR*	*b*	*SE*	*IRR*
Personal Attack	0.196	0.259	1.216	−1.031**	0.374	0.357
Affiliation Accusation	0.364	0.367	1.440	0.365	0.607	1.440
Contradiction	−0.236	0.287	0.790	−0.153	0.477	0.858
Request	0.008	0.263	1.008	−0.273	0.382	0.761
Counterclaim	0.179	0.301	1.196	−1.074*	0.515	0.342
Parent comment likes (log)	0.728***	0.214	2.071	1.084***	0.152	2.956
Parent–response posting interval (log)	−0.287**	0.104	0.751	−0.374**	0.141	0.688
Comment exposure time (log)	0.007	0.085	1.007	0.244*	0.122	1.276
Comment length (log)	0.163	0.097	1.177	0.030	0.170	1.030
McFadden R^2			.077			.056

Notes: Justification is the reference category; * $p < .05$, ** $p < .01$, *** $p < .001$.

Finally, interaction tests directly examined whether the associations between response type and likes differed between counterspeech and disagreement replies. The overall interaction between response type and response context was not significant in the binary component, $\chi^2(5) = 2.07$, $p = .839$. In contrast, the interaction was significant in the count component, $\chi^2(5) = 21.71$, $p < .001$. Relative to justification, the relative disadvantage of personal attack was significantly greater in the counterspeech replies than in the disagreement replies ($b = -1.158$, $p < .01$), and counterclaim showed a similar interaction ($b = -1.238$, $p < .05$).

5. Discussion

This study investigated oppositional replies in Douyin discussions of AI controversies in China. It compared counterspeech directed at hate speech with disagreement directed at non-hateful comments, focusing on likes received, linguistic toxicity, subsequent toxic replies, and response types. The results showed no significant overall variation in likes between the two forms of oppositional reply. Counterspeech, however, displayed higher levels of toxicity, and its direct replies were more likely to contain toxic expressions. Both reply contexts shared the same six-type repertoire, with justification the most common type in both contexts, although the prevalence of these response types and their relationships with like counts varied across the two contexts. Taken together, the findings reveal that a hateful antecedent was not associated with an overall liking advantage for oppositional replies. It nonetheless was associated with variation in how opposition was expressed, the toxicity of subsequent interactions, and the relationships between selected response strategies and like counts.

RQ1 provided no evidence of a significant association between response context and like counts. Counterspeech and disagreement showed no significant gap in either the probability of receiving a like or

positive like counts. This pattern departs from Mathew et al. (2019), whose comparison mainly included comments agreeing with the hateful content or expressing hate toward the target community. By restricting the comparison to oppositional replies, the design lends insight into the possibility that replies in the counterspeech context may not receive additional likes simply because they respond to hateful antecedents. In technological controversies, disputes over national competition and technological autonomy (Li et al., 2026), as well as corporate interests, evaluative bias, and fairness (Marres et al., 2024), often involve clearly opposed issue positions. Because competing scientific claims are often evaluated in line with preexisting issue attitudes (Wang, 2021), and source and tone cues also shape the perceived believability of corrective messages (W. Yu et al., 2022), issue-position congruence may contribute to audience engagement in both response contexts, leaving relatively little additional advantage for counterspeech in attracting likes.

With respect to RQ2, toxicity varied between the two response contexts in both the focal replies and subsequent direct replies. Counterspeech replies had higher toxicity probabilities than disagreement replies, consistent with prior taxonomies that characterize counterspeech as combining reason-giving and factual rebuttal with condemnation, insults, and author-directed attacks (Mathew et al., 2019; X. Yu et al., 2023). Moreover, direct replies following counterspeech were more likely to be toxic within seven days, although toxic replies remained a minority in both contexts, accounting for 15.32% of replies following counterspeech and 9.68% following disagreement. This relative increase echoes prior research on the tendency for hostile cues to lower restraints on aggressive expression and for toxicity in an ongoing conversation to make subsequent toxic replies more likely (Cheng et al., 2017; Saveski et al., 2021).

RQ3 identifies a shared argumentative repertoire across counterspeech and disagreement. Justification, personal attack, affiliation accusation, contradiction, request, and counterclaim appeared in both contexts, with justification the most common type in each. Its use of technical definitions, comparative cases, concrete evidence, and causal reasoning suggests that reason-giving may remain central to opposition in technological controversy, consistent with established disagreement and counterspeech taxonomies (Muntigl & Turnbull, 1998; X. Yu et al., 2023). The distribution of these resources nonetheless varied across contexts. Counterspeech contained a higher proportion of request responses but a lower proportion of contradiction responses. No statistically significant differences were found for the other four response types. Specifically, disagreement mainly contested technical evidence, responsibility, and technological legitimacy, whereas counterspeech additionally challenged the extension of technical judgments into group denigration. Such generalizations may invite challenges to the grounds linking an individual case to a group-level conclusion, helping to explain the greater use of requests and the lower reliance on direct contradiction (Mun et al., 2023).

The association between response type and like-based feedback also varied across response contexts. In disagreement, affiliation accusation received more likes than justification, whereas contradiction and request received fewer. Compared with reason-giving through facts, examples, and causal links, more confrontational criticism may be more likely to attract attention and elicit audience engagement, given prior evidence that incivility can increase user interaction (Masullo Chen et al., 2019; Muddiman & Stroud, 2017). However, this association emerged for affiliation accusation instead of personal attack. In our corpus, personal attack mainly disparaged the speaker's knowledge, competence, or credibility, whereas affiliation accusation attributed the disputed position to commercial interests, attention-seeking motives, evaluative bias, or double standards. This distinction resonates with evidence that attacks on researchers' motivations

can lower perceived integrity, even when perceived expertise remains unaffected (Gierth & Bromme, 2020). Contradiction and request, by comparison, often denied or questioned a claim with limited elaboration, while justification supplied facts, examples, and causal reasoning, which may have made the latter more favorably evaluated by readers (Kim & Gonzales, 2022).

Counterspeech yielded a different pattern. No response type differed significantly from justification in the likelihood of receiving any likes, but among replies that received at least one like, personal attack and counterclaim accumulated fewer likes than justification. When opposition addresses hateful content, reason-giving may be particularly compatible with positive audience feedback because it provides bystanders with grounds for evaluating the disputed claim. Collective counterspeech similarly prioritizes factual and respectful contributions intended to reach the "movable middle" (Buerger, 2021), and experimental evidence shows that logical appeals are evaluated more favorably than reputational attacks, which are more likely to elicit negative audience reactions (Zhao et al., 2024). Counterclaim may also receive less positive feedback because it redirects the argumentative frame by advancing an alternative causal or evaluative position without necessarily providing the same evidentiary or explanatory grounds directly as justification. Importantly, these relative disadvantages were stronger in counterspeech than in disagreement. The stronger pattern in counterspeech may be related to norm enforcement, as responses to hateful speech can signal and reinforce norms against such conduct, potentially increasing the value placed on constructive reason-giving over personal attacks or alternative counterclaims (Gagrčin & Milzner, 2023; Zapata et al., 2024).

This study advances theory in three ways. First, it extends research on technological controversies by showing how issue-based disagreement and identity-based hostility can coexist within the same platformed debate, linking disputes over technical evidence and legitimacy with broader social and geopolitical tensions. Second, grounded in argumentation theories, it connects counterspeech typologies with general disagreement research and elucidates the use of shared oppositional forms and the social-media feedback they receive across response contexts. Third, by comparing two forms of opposition, it provides an empirical understanding of counterspeech by revealing its similarities to and differences from ordinary disagreement in like counts, reply toxicity, and subsequent toxicity.

In practice, these findings highlight practical steps that public institutions, platforms, and users can take to foster healthier social media environments. Government and technology communication agencies should provide verifiable information on distillation, open-source practices, intellectual property, and model performance, and correct claims that extend technical disputes into judgments about national or social groups. Platforms could consider monitoring follow-up chains of online debates, flagging emerging toxicity, prompting users to revise personal attacks, and providing templates for evidence-based rebuttal. Users can challenge claims, request evidence, and question group generalizations while avoiding personal attacks and other unnecessarily hostile formulations.

6. Limitations and Future Research

This study presents several limitations that offer opportunities for future research. First, the data represent a group of Douyin users who actively commented on AI-related content, rather than Douyin users or the Chinese public more broadly. Platform privacy restrictions precluded collecting impression-level exposure data, measures of algorithmic visibility, and users' demographic and behavioral characteristics, including age,

technology engagement, and commenting frequency. Consequently, we cannot determine how these commenters differ from less active users, why they posted or liked comments, or what psychological effects the exchanges had. Future research could combine platform traces with surveys, interviews, or experiments and compare active commenters with passive viewers across platforms. Second, the taxonomy captured response function but not tone. Sarcasm, irony, humor, ridicule, and empathy may alter how similar counterspeech is interpreted and received (Hangartner et al., 2021; Lasser et al., 2025). Moreover, assigning one primary response type compressed hybrid comments that justified a position, challenged evidence, accused an author, or used sarcasm. Multi-label coding that jointly models strategy and tone would better represent this complexity (Mathew et al., 2019). Finally, the study was not preregistered. Preregistration of the sampling and analytic procedures could have further strengthened the transparency of the research process, particularly for future confirmatory studies building on the present findings.

7. Conclusion

This study situates counterspeech within broader public disagreement by comparing oppositional replies to hateful and non-hateful comments in Douyin discussions of Chinese AI controversies. The findings show that counterspeech and disagreement did not differ significantly in the likes they received and drew on a largely shared repertoire of argumentative responses, while differing in several aspects. Counterspeech exhibited higher toxicity and was associated with a greater likelihood of toxic direct replies. Although both contexts relied most heavily on justification, counterspeech contained more requests but fewer contradictions. Response type was also associated with likes in different ways across the two settings. Affiliation accusation received more positive feedback in disagreement relative to justification, whereas contradiction and request were associated with less; in counterspeech, personal attack and counterclaim accumulated fewer likes than justification among comments that received at least one like.

Taken together, this comparison places counterspeech within a spectrum of online dispute without framing it as an isolated response category. It further highlights variation in response types, their associations with platform feedback, and subsequent interaction across the two response contexts. This perspective connects counterspeech research with scholarship on disagreement and foregrounds the relational structure linking antecedent discourse, oppositional responses, and subsequent conversations in platformed public debate.

Acknowledgments

We would like to thank the anonymous reviewers and academic editors for their valuable feedback.

Funding

This work was supported by the National Social Science Fund of China (Grant No. 25CKX032).

Conflict of Interests

The authors declare no conflict of interests.

LLMs Disclosure

The authors used DeepSeek V4 Pro for response annotation as described in the Methods section. GPT-5.6 Sol was used only for grammar checking and language polishing of a manuscript originally written by the authors.

Supplementary Material

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

References

Alshomary, M., Syed, S., Dhar, A., Potthast, M., & Wachsmuth, H. (2021). Counter-argument generation by attacking weak premises. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (pp. 1816–1827). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.159

Angouri, J., & Locher, M. A. (2012). Theorising disagreement. *Journal of Pragmatics*, *44*(12), 1549–1553. https://doi.org/10.1016/j.pragma.2012.06.011

Baider, F. (2023). Accountability issues, online covert hate speech, and the efficacy of counter-speech. *Politics and Governance*, *11*(2), 249–260. https://doi.org/10.17645/pag.v11i2.6465

Benesch, S., Ruths, D., Dillon, K. P., Saleem, H. M., & Wright, L. (2016). *Counterspeech on Twitter: A field study*. Dangerous Speech Project. https://www.dangerousspeech.org/libraries/counterspeech-on-twitter-a-field-study

Beyer, L., Zhai, X., Royer, A., Markeeva, L., Anil, R., & Kolesnikov, A. (2022). Knowledge distillation: A good teacher is patient and consistent. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 10925–10934). IEEE. https://doi.org/10.1109/CVPR52688.2022.01065

Buerger, C. (2021). #iamhere: Collective counterspeech and the quest to improve online discourse. *Social Media + Society*, *7*(4). https://doi.org/10.1177/20563051211063843

Cepollaro, B., Lepoutre, M., & Simpson, R. M. (2023). Counterspeech. *Philosophy Compass*, *18*(1), Article e12890. https://doi.org/10.1111/phc3.12890

Chen, K., Jin, Y., & Shao, A. (2022). Science factionalism: How group identity language affects public engagement with misinformation and debunking narratives on a popular Q&A platform in China. *Social Media + Society*, *8*(1). https://doi.org/10.1177/20563051221077019

Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C., & Leskovec, J. (2017). Anyone can become a troll: Causes of trolling behavior in online discussions. In *CSCW'17: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (pp. 1217–1230). Association for Computing Machinery. https://doi.org/10.1145/2998181.2998213

Chinese AI companies 'distilled' Claude to improve own models, Anthropic says. (2026, February 23). *Reuters*. https://www.reuters.com/world/china/chinese-companies-used-claude-improve-own-models-anthropic-says-2026-02-23

Chung, Y.-L., Abercrombie, G., Enock, F., Bright, J., & Rieser, V. (2024). Understanding counterspeech for online harm mitigation. *Northern European Journal of Language Technology*, *10*(1), 30–49. https://doi.org/10.3384/nejlt.2000-1533.2024.5203

Fang, L., Yu, X., Cai, J., Chen, Y., Wu, S., Liu, Z., Yang, Z., Lu, H., Gong, X., Liu, Y., Ma, T., Ruan, W., Abbasi, A., Zhang, J., Wang, T., Latif, E., Liu, W., Zhang, W., Kolouri, S., . . . Ma, P. (2026). Knowledge distillation and dataset distillation of large language models: Emerging trends, challenges, and future directions. *Artificial Intelligence Review*, *59*(1), Article 17. https://doi.org/10.1007/s10462-025-11423-3

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

Gagrčin, E., & Milzner, M. (2023). "Intervening is a good thing but . . .": The role of social norms in users' justifications of (non-)intervention against incivility. *Social Media + Society*, *9*(3). https://doi.org/10.1177/20563051231186561

Garland, J., Ghazi-Zahedi, K., Young, J.-G., Hébert-Dufresne, L., & Galesic, M. (2020). Countering hate on social media: Large-scale classification of hate and counter speech. In S. Akiwowo, B. Vidgen, V. Prabhakaran, & Z. Waseem (Eds.), *Proceedings of the Fourth Workshop on Online Abuse and Harms* (pp. 102–112). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.alw-1.13

Gierth, L., & Bromme, R. (2020). Attacking science on social media: How user comments affect perceived trustworthiness and credibility. *Public Understanding of Science*, *29*(2), 230–247. https://doi.org/10.1177/0963662519889275

Gou, J., Yu, B., Maybank, S. J., & Tao, D. (2021). Knowledge distillation: A survey. *International Journal of Computer Vision*, *129*(6), 1789–1819. https://doi.org/10.1007/s11263-021-01453-z

Habernal, I., Wachsmuth, H., Gurevych, I., & Stein, B. (2018). Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. In M. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 386–396). Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1036

Hangartner, D., Gennaro, G., Alasiri, S., Bahrich, N., Bornhoft, A., Boucher, J., Demirci, B. B., Derksen, L., Hall, A., Jochum, M., Murias Munoz, M., Richter, M., Vogel, F., Wittwer, S., Wüthrich, F., Gilardi, F., & Donnay, K. (2021). Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proceedings of the National Academy of Sciences*, *118*(50), Article e2116310118. https://doi.org/10.1073/pnas.2116310118

Hassan, S., & Alikhani, M. (2023). DisCGen: A framework for discourse-informed counterspeech generation. In J. C. Park, Y. Arase, B. Hu, W. Lu, D. Wijaya, A. Purwarianti, & A. Alfa Krisnadhi (Eds.), *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 420–429). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.ijcnlp-main.28

Hayes, R. A., Carr, C. T., & Wohn, D. Y. (2016). One click, many meanings: Interpreting paralinguistic digital affordances in social media. *Journal of Broadcasting & Electronic Media*, *60*(1), 171–187. https://doi.org/10.1080/08838151.2015.1127248

Hickey, D., Fessler, D. M. T., Schmitz, M., Smaldino, P., Lerman, K., Murić, G., & Burghardt, K. (2026). Assessing how hate, counterspeech, and toxicity affect hate group newcomers. *Proceedings of the International AAAI Conference on Web and Social Media*, *20*(1), 1007–1026. https://doi.org/10.1609/icwsm.v20i1.42679

Hinton, G., Vinyals, O., & Dean, J. (2015). *Distilling the knowledge in a neural network*. arXiv. https://doi.org/10.48550/arXiv.1503.02531

Jo, Y., Bang, S., Manzoor, E., Hovy, E., & Reed, C. (2020). Detecting attackable sentences in arguments. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1–23). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.1

Kim, Y., & Gonzales, A. L. (2022). When we tolerate online incivility: Dual-process effects of argument strength and heuristic cues in uncivil user comments. *Computers in Human Behavior*, *131*, Article 107235. https://doi.org/10.1016/j.chb.2022.107235

Larsson, A. O. (2026). Going negative across platforms and time: Assessing the supply and demand of negative campaigning. *New Media & Society*. Advance online publication. https://doi.org/10.1177/14614448261430714

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

Li, Y., Ma, C., & Yu, L. (2026). Artificial intelligence and digital nationalism: A social media discourse analysis. *Technology in Society*, *85*, Article 103197. https://doi.org/10.1016/j.techsoc.2025.103197

Lu, J., Xu, B., Zhang, X., Min, C., Yang, L., & Lin, H. (2023). Facilitating fine-grained detection of Chinese toxic language: Hierarchical taxonomy, resources, and benchmarks. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 16235–16250). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.898

Marres, N., Castelle, M., Gobbo, B., Poletti, C., & Tripp, J. (2024). AI as super-controversy: Eliciting AI and society controversies with an extended expert community in the UK. *Big Data & Society*, *11*(2). https://doi.org/10.1177/20539517241255103

Masullo Chen, G., Muddiman, A., Wilner, T., Pariser, E., & Stroud, N. J. (2019). We should not get rid of incivility online. *Social Media + Society*, *5*(3). https://doi.org/10.1177/2056305119862641

Mathew, B., Saha, P., Tharad, H., Rajgaria, S., Singhania, P., Maity, S. K., Goyal, P., & Mukherjee, A. (2019). Thou shalt not hate: Countering online hate speech. *Proceedings of the International AAAI Conference on Web and Social Media*, *13*(1), 369–380. https://doi.org/10.1609/icwsm.v13i01.3237

Microsoft probes if DeepSeek-linked group improperly obtained OpenAI data, Bloomberg News reports. (2025, January 29). *Reuters*. https://www.reuters.com/technology/microsoft-probing-if-deepseek-linked-group-improperly-obtained-openai-data-2025-01-29

Muddiman, A., & Stroud, N. J. (2017). News values, cognitive biases, and partisan incivility in comment sections. *Journal of Communication*, *67*(4), 586–609. https://doi.org/10.1111/jcom.12312

Mun, J., Allaway, E., Yerukola, A., Vianna, L., Leslie, S.-J., & Sap, M. (2023). Beyond denouncing hate: Strategies for countering implied biases and stereotypes in language. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 9759–9777). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.653

Muntigl, P., & Turnbull, W. (1998). Conversational structure and facework in arguing. *Journal of Pragmatics*, *29*(3), 225–256. https://doi.org/10.1016/S0378-2166(97)00048-9

Ning, C. (2026). Blending positivity energy and fun: Dominant discourse patterns of popular short videos in China's mobile media communication. *Journal of Information Technology & Politics*, *23*(2), 136–154. https://doi.org/10.1080/19331681.2024.2428701

Obermaier, M. (2024). Youth on standby? Explaining adolescent and young adult bystanders' intervention against online hate speech. *New Media & Society*, *26*(8), 4785–4807. https://doi.org/10.1177/14614448221125417

Pomerantz, A. (1984). Agreeing and disagreeing with assessments: Some features of preferred/dispreferred turn shapes. In J. M. Atkinson & J. Heritage (Eds.), *Structures of social action: Studies in conversation analysis* (pp. 57–101). Cambridge University Press.

Rees-Miller, J. (2000). Power, severity, and context in disagreement. *Journal of Pragmatics*, *32*(8), 1087–1111. https://doi.org/10.1016/S0378-2166(99)00088-0

Saveski, M., Roy, B., & Roy, D. (2021). The structure of toxic conversations on Twitter. In *WWW '21: Proceedings of the Web Conference 2021* (pp. 1086–1097). Association for Computing Machinery. https://doi.org/10.1145/3442381.3449861

Sifianou, M. (2012). Disagreements, face and politeness. *Journal of Pragmatics*, *44*(12), 1554–1564. https://doi.org/10.1016/j.pragma.2012.03.009

Sridhar, D., & Getoor, L. (2019). Estimating causal effects of tone in online debates. In S. Kraus (Ed.), *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence* (pp. 1872–1878). International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2019/259

Tian, W., & Ge, J. (2024). Decoding the Apple paradox: A critical discourse analysis of gender, technology, and nationalism in China's digital space. *Humanities and Social Sciences Communications*, *11*, Article 1269. https://doi.org/10.1057/s41599-024-03819-3

van Eemeren, F. H., & Grootendorst, R. (2004). *A systematic theory of argumentation: The pragma-dialectical approach*. Cambridge University Press.

Wachsmuth, H., Syed, S., & Stein, B. (2018). Retrieval of the best counterargument without prior topic knowledge. In I. Gurevych & Y. Miyao (Eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 241–251). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1023

Walker, M. A., Fox Tree, J. E., Anand, P., Abbott, R., & King, J. (2012). A corpus for research on deliberation and debate. In N. Calzolari, K. Choukri, T. Declerck, M. Uğur Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (pp. 812–817). European Language Resources Association. https://aclanthology.org/L12-1643

Wang, Y. (2021). Debunking misinformation about genetically modified food safety on social media: Can heuristic cues mitigate biased assimilation? *Science Communication*, *43*(4), 460–485. https://doi.org/10.1177/10755470211022024

Yang, C., Zhu, Y., Lu, W., Wang, Y., Chen, Q., Gao, C., Yan, B., & Chen, Y. (2025). Survey on knowledge distillation for large language models: Methods, evaluation, and application. *ACM Transactions on Intelligent Systems and Technology*, *16*(6), Article 143. https://doi.org/10.1145/3699518

Yang, Y., Lange, R., Zhang, Y., & Walther, J. B. (2025). How social responses to online hate messages affect hatefulness. *New Media & Society*. Advance online publication. https://doi.org/10.1177/14614448251396951

Yu, W., Shen, F., & Min, C. (2022). Correcting science misinformation in an authoritarian country: An experiment from China. *Telematics and Informatics*, *66*, Article 101749. https://doi.org/10.1016/j.tele.2021.101749

Yu, X., Blanco, E., & Hong, L. (2022). Hate speech and counter speech detection: Conversational context does matter. In M. Carpuat, M.-C. de Marneffe, & I. V. Meza Ruiz (Eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 5918–5930). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.433

Yu, X., Zhao, A., Blanco, E., & Hong, L. (2023). A fine-grained taxonomy of replies to hate speech. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 7275–7289). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.450

Zapata, J., Sulik, J., von Wulffen, C., & Deroy, O. (2024). Bystanders' collective responses set the norm against hate speech. *Humanities and Social Sciences Communications*, *11*(1), Article 335. https://doi.org/10.1057/s41599-024-02761-8

Zhang, A. X., Culbertson, B., & Paritosh, P. (2017). Characterizing online discussion using coarse discourse sequences. *Proceedings of the International AAAI Conference on Web and Social Media*, *11*(1), 357–366. https://doi.org/10.1609/icwsm.v11i1.14886

Zhao, P., Bazarova, N. N., DiFranzo, D., Hui, W., Kizilcec, R. F., & Margolin, D. (2024). Standing up to problematic content on social media: Which objection strategies draw the audience's approval? *Journal of Computer-Mediated Communication*, *29*(1), Article zmad046. https://doi.org/10.1093/jcmc/zmad046

About the Authors



Lishuang Wei is a doctoral candidate at the School of Media and Communication, Shanghai Jiao Tong University. Her research explores human–AI communication, science communication, and computational social science, with a focus on the societal implications and governance of generative AI.



Chuanlin Ning is currently a research assistant professor at the School of Media and Communication, Shanghai Jiao Tong University. He earned his PhD in journalism and communication from Shanghai Jiao Tong University. His research focuses on sports communication and international communication, with particular attention to their complex intersections with technological development, health inequalities, youth culture, and broader social change in contemporary media environments.