



OPEN The impact of generative AI on social media: an experimental study

Anders Giovanni Møller^{1✉}, Daniel M. Romero^{2,3,4}, David Jurgens^{2,4} & Luca Maria Aiello^{1,5}

Generative Artificial Intelligence (AI) tools are increasingly deployed across social media platforms, yet their implications for user behavior and experience remain understudied, particularly regarding two critical dimensions: (1) how AI tools affect the behaviors of content producers in a social media context, and (2) how content generated with AI assistance is perceived by users. To fill this gap, we conduct a controlled experiment with a representative sample of 680 U.S. participants in a realistic social media environment. The participants are randomly assigned to small discussion groups, each consisting of five individuals in one of five distinct experimental conditions: a control group and four treatment groups, each employing a unique AI intervention—*Chat* assistance, *Conversation Starters*, *Feedback* on comment drafts, and reply *Suggestions*. Our findings highlight a complex duality: some AI-tools increase user engagement and volume of generated content, but at the same time decrease the perceived quality and authenticity of discussion, and introduce a negative spill-over effect on conversations. Based on our findings, we propose four design principles and recommendations aimed at social media platforms, policymakers, and stakeholders: ensuring transparent disclosure of AI-generated content, designing tools with user-focused personalization, incorporating context-sensitivity to account for both topic and user intent, and prioritizing intuitive user interfaces. These principles aim to guide an ethical and effective integration of generative AI into social media.

Keywords Generative artificial intelligence, Human-computer interaction, Controlled experiment, Large language models

The rapid integration of artificial intelligence (AI)-driven text generation tools into social media platforms is reshaping how users create and engage with content, raising new questions about their effects on the quality and dynamics of online interactions^{1,2}. AI writing tools notably reduce barriers to content creation by lowering required effort and expertise. Although these technologies are increasingly adopted across sectors such as journalism^{3,4}, education⁵, and creative industries³, their potential impact is particularly pronounced in social media contexts. Social media platforms play a central role in shaping public discourse, influencing democratic engagement, and enabling rapid, large-scale dissemination of information and ideas⁶. Therefore, introducing AI into these platforms may reshape dynamics of interaction, authenticity, and nature of online discussions^{7,8}. AI in social media can be seen not only as a functional writing aid but as an infrastructure that shapes epistemic and normative aspects under which content is produced, amplified, and perceived as legitimate in online settings⁹. From an epistemic perspective, AI models can support generation of misleading content by enabling users to distort or manipulate information, producing plausible but fabricated text⁹. From a normative context, AI models can reshape value considerations, including fairness, bias, equitability, and auditability, thereby influencing how users evaluate credibility and authenticity, and whether they engage with content⁹. This reflects a broader perspective that AI systems are not neutral or separate from human values, but intricately intertwined into society and influencing it^{10,11}. This suggests a duality where AI can improve productivity while triggering negative social consequences¹². In this context, previous studies highlight the promise of AI assistance in advancing human creativity¹³, increasing user engagement¹⁴, and facilitating broader inclusivity in online discussions^{15,16}. Yet, this optimism is tempered by concerns about potential drawbacks, such as declining content quality¹⁷, proliferation of misinformation¹⁸, and diminished authenticity of user interactions¹⁹. AI systems should support human decision-making rather than replace it^{10,20}.

¹Data Science Section, IT University of Copenhagen, Rued Langgaards Vej 7, 2300 Copenhagen, Denmark. ²School of Information, University of Michigan, 2200 Hayward Street, Ann Arbor, MI 48109, USA. ³Center for the Study of Complex Systems, University of Michigan, 500 Church Street, Ann Arbor 48109, USA. ⁴Computer Science and Engineering Division, University of Michigan, 2260 Hayward Street, Ann Arbor 48109, USA. ⁵Pioneer Centre for AI, Øster Voldgade 3, 1350 Copenhagen, Denmark. ✉email: agmo@itu.dk

Empirical evidence quantifying how AI assistance reshapes online participation dynamics, content quality, and user perceptions remains scarce, particularly in realistic, platform-integrated scenarios²¹. Addressing this critical gap, our study provides empirical insights through a controlled experiment conducted on a realistic social media platform. We approach the debate about AI-assisted content creation from two complementary perspectives: first, how AI assistants in social media affect the experience of content *producers*, and second, how these interventions shape *consumers'* perceptions of content. Our experiment fills a crucial gap by directly assessing how AI assistants transform both producer and consumer experiences on social media.

We conduct the experiment using a custom-built platform that closely simulates an online chat room resembling discussions on common social media forums. A representative sample of 680 U.S. participants is partitioned into groups of five people, who are then randomly assigned to one of five conditions: one control (no AI assistance) and four non-overlapping treatment conditions, each featuring distinct AI interventions previously proposed for enhancing online interactions. These interventions include (1) an open-ended *Chat* with an AI assistant²², (2) AI-generated reply *Suggestions* with varying stances (agreeing, neutral, disagreeing)^{7,23}, (3) AI-driven *Feedback* on comment drafts^{5,24}, and (4) AI-generated *Conversation Starters*²⁵; interventions are described in more detail in the Methods section and in “AI Prompts and Settings” in the *Supplementary Information* (see Fig. 1 for platform screenshots). Together, these interventions represent a comprehensive range of AI-based approaches considered by both researchers and industry practitioners for enhancing online interactions. Specifically, the four tools were selected to systematically vary along two dimensions of human-AI interaction^{26,27}. The first dimension is *initiative*: assistance can be reactive, requiring user requests (*Chat* and *Feedback*), or proactive, with the system anticipating user needs given the context (*Suggestions* and *Conversation Starter*). The second dimension is *task orientation*: the AI can generate new content (*Suggestions*, *Conversation Starter*, and *Chat*) or refine user-authored content (*Feedback* and *Chat*). This taxonomy captures meaningful variation in human-AI collaboration, as initiative affects user autonomy and cognitive burden, while task orientation influences content authenticity²⁸.

To assess the heterogeneity of the interventions' effects across topics, participants sequentially discuss three randomly-ordered topics—ranging from conversational (dogs vs. cats), to scientific (health benefits of oats), to political (universal basic income)—with each topic limited to a 10-minute interaction. We assess various proxies for user engagement and quality of experience through questionnaires before and after participation, and further track participants' interactions on the platform—including comments, reactions, and AI usage—to comprehensively evaluate how AI interventions impact both content producers and consumers.

Overall, our findings indicate that AI assistance substantially influences both user-generated content quality and consumer perception, although with notable variation across interventions. AI-supported participants demonstrate increased engagement and content production metrics, but these improvements are associated with nuanced, sometimes negative, shifts in consumer perceptions and reactions. Critically, no single AI tool enhances both producer and consumer experiences, highlighting complex trade-offs.

AI interventions generally increase participants' willingness to engage and improve aspects of content creation from the producer perspective. Participants using the *Chat* and *Suggestions* features notably report that the AI would increase their willingness to participate in online discussions, compared to control (Fig. 2a). Additionally, all AI-supported tools significantly increase the average length of user comments (Fig. 2c). Participation equality among users, measured by normalized Shannon entropy based on the proportion of comments per user in each round, noticeably improves under the *Conversation Starter* intervention, indicating more balanced participation (Fig. 2e).

Regression analysis for each treatment condition indicates that only *Conversation Starter* significantly increased the likelihood of receiving a reply ($\beta = 0.300, p = 0.021$). The *Conversation Starter* is designed to lower barriers to initiating engagement, and is therefore most naturally used early in discussion threads where comments have higher visibility and more time to attract replies. The other AI-tools (*Chat*, *Feedback*, *Suggestions*) showed positive but non-significant signals (see *Supplementary Information*, “Post-hoc Regression Analyses on Treatment Impact of Reply Likelihood”). These non-significant effects may reflect that the tools support content refinement and ideation throughout the discussion rather than encouraging early-stage comments benefiting from greater visibility and response opportunity. Although *Suggestions* is proactive in providing content to use directly, it is designed for continuous use throughout the discussions, whereas *Conversation Starter* is intended specifically to initiate engagement.

From the consumer perspective, none of the AI interventions improve user perceptions compared to the control condition. Participants perceive comments to be less informative and lower quality in both the *Chat* and *Conversation Starter* conditions compared to control (Fig. 2b). Similarly, users rated replies to own comments significantly lower in all but the *Suggestions* condition, which uniquely show positive, though weakly significant, ratings compared to control (Fig. 2d). All AI treatments significantly increase ‘Dislikes’, consistent with the lower quality and informativeness ratings observed. Open-ended feedback supports this pattern, with participants describing content in the AI-assisted conditions as “robotic” and “generic”. *Suggestions* is the only treatment to evoke more ‘Love’ reactions (Fig. 2f).

Together, these findings highlight a critical duality: although AI interventions broaden participation and increase the volume of content, they also risk creating “semantic garbage” perceived as lower quality than human-generated text²⁹ and degrading the quality of subsequent human conversation once AI is introduced to a thread. This pattern aligns with prior work on AI-assisted communication affecting authenticity and effort^{11,12}. Our experiment stands as one of the first direct tests of whether AI can genuinely elevate online discussions or merely amplify low-quality interaction in a way that clouds its potential benefits—an outcome our evidence suggests is more likely. In exploratory regression analyses assessing whether demographics (age, gender, sex, education, and political affiliation) would alter these results, we observe only minor effects that did not reach statistical significance (see *Supplementary Information* for additional details). With 130–140 participants in

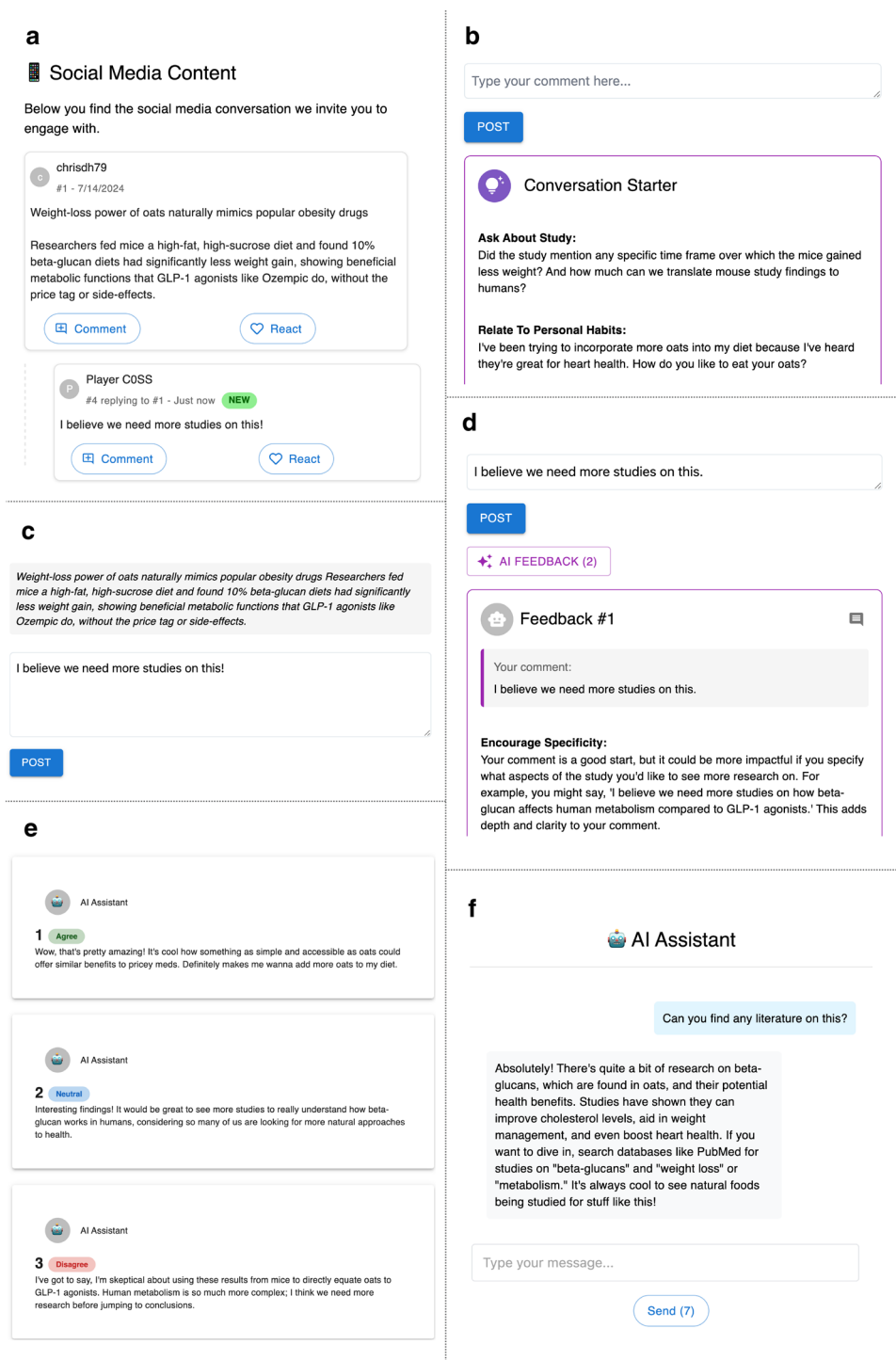


Fig. 1. User interface of the experimental platform. **(a)**, Main discussion thread. **(b)**, Conversation Starter. **(c)**, Comment modal. **(d)**, Feedback tool. **(e)**, Suggestions tool. **(f)**, Chat assistant.

each condition, the number of individuals within demographic subgroups is likely too small to reliably detect moderating effects. Alternatively, we speculate that demographic gaps in the use of AI are narrowing as the technology becomes more accessible and widely adopted. Recent work from OpenAI suggests demographic differences in ChatGPT adoption have decreased over time³⁰, which could contribute to the lack of demographic variation found in our study. Future studies with larger samples would be needed to more accurately assess whether demographic variations moderate AI use and perceptions.

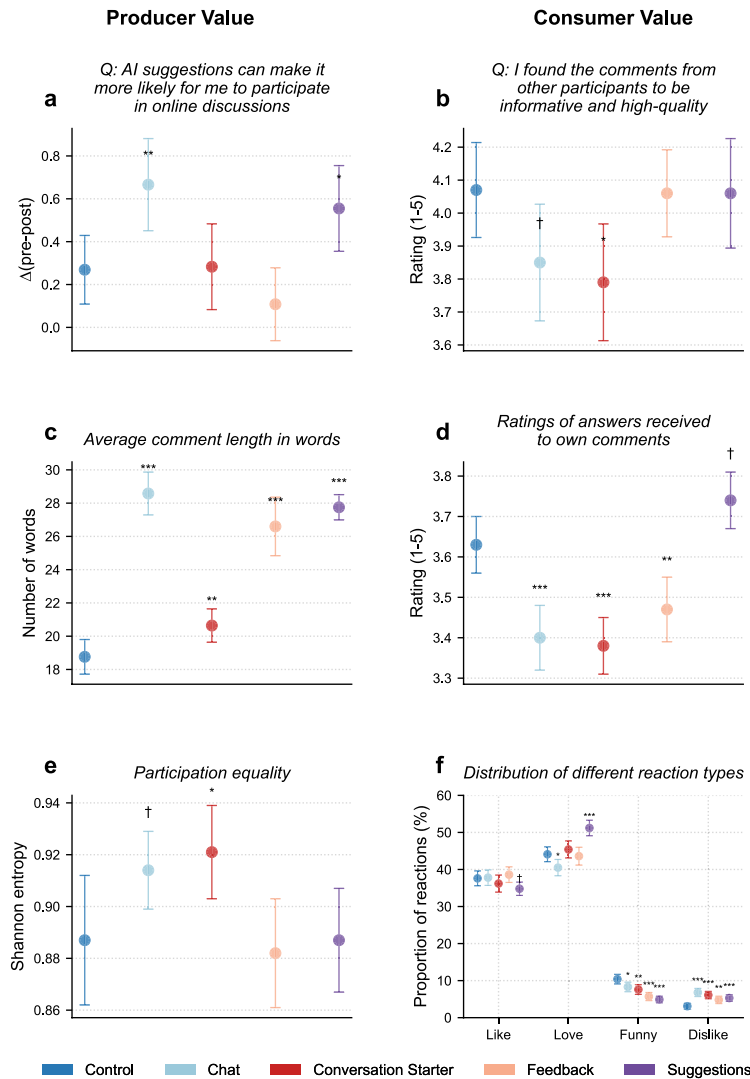


Fig. 2. **Producer value** results (left column) and **consumer value** results (right column). (a), Change in responses to “AI suggestions can make it more likely for me to participate in online discussions” (1–5 Likert scale) before and after the study. (b), Likert-scale responses to “I found the comments from other participants to be informative and high-quality”. (c), Average comment length in words. (d), Individual user ratings of replies received on their own comments. (e), Participation equality measured using normalized Shannon entropy based on participation distribution per round. (f), Distribution of reaction types (Like, Love, Funny, Dislike) across conditions. † $p < 0.10$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$. Effect sizes (Cohen’s d) are reported in Table 1. For details on statistical tests and bootstrapping procedures, see *Supplementary Information*.

How are the AI tools used?

Building on the duality—AI interventions boosting participation yet risking polluting conversation with low-quality content—we also uncover nuanced differences in how participants employed each tool. Usage patterns are far from uniform and often reflect the tools intended design and conversational context. These variations merit closer auditing, both to clarify which AI-aid paradigm holds the most promise for meaningful engagement and to guide future employments and refinements. Figs. 3a–d and 4h–i detail key usage patterns (see the *AI Usage Analysis* section and “AI Usage” in *Supplementary Information* for details on the analyses).

Chat provides flexibility but engagement varies across context

For the *Chat* feature, we find widespread adoption, with 94.4% ($n = 118$) of treatment participants engaging at least once with the AI, resulting in a total of 960 user prompts (Fig. 3a,b). Users predominantly interacted with the AI for topic-related, informal, exploratory dialogue (*Casual Queries*, 34.9%, $n = 335$)—for instance, in the *Cats* discussion: ‘One of my cats getting stuck in a jar’. However, a substantial proportion of queries was used for *Fact Checking* (24.1%, $n = 231$), highlighting the AI’s role as an informational resource. Additionally, smaller proportions of interactions involved *Engagement* (13.0%, $n = 125$), referring to instances where users sought assistance with crafting responses. A similar share of prompts involved *Political Discussions* subjects (11.9%, $n = 114$), reflecting users interest in issues with political or societal implications. Disaggregating

Metric	Figures	Treatment	N	Control μ [95%CI]	Treatment μ [95%CI]	p	Cohen's d [95%CI]
Δ (pre-post)	2a	Chat	123	0.27 [0.10, 0.43]	0.67 [0.46, 0.88]	0.004	0.37 [0.12, 0.63]
Δ (pre-post)	2a	Suggestions	117	0.27 [0.10, 0.43]	0.56 [0.35, 0.76]	0.032	0.28 [0.02, 0.54]
Rating (1–5)	2b	Chat	124	4.08 [3.93, 4.22]	3.85 [3.67, 4.03]	0.060	-0.24 [-0.49, 0.02]
Rating (1–5)	2b	Conv. Starter	124	4.08 [3.93, 4.22]	3.79 [3.61, 3.97]	0.014	-0.31 [-0.56, -0.05]
N words	2c	Chat	1334	18.8 [17.8, 19.8]	28.6 [27.2, 30.00]	< 0.001	0.43 [0.35, 0.51]
N words	2c	Suggestions	1795	18.8 [17.8, 19.8]	27.8 [27.0, 28.5]	< 0.001	0.52 [0.45, 0.60]
Rating (1–5)	2d	Conv. Starter	988	3.63 [3.55, 3.71]	3.39 [3.31, 3.47]	< 0.001	-0.20 [-0.29, -0.11]
Rating (1–5)	2d	Suggestions	1014	3.63 [3.55, 3.71]	3.74 [3.66, 3.82]	0.053	0.09 [0.00, 0.18]
Shannon Entropy	2e	Chat	79	0.89 [0.86, 0.91]	0.91 [0.90, 0.93]	0.064	0.29 [-0.02, 0.61]
Shannon Entropy	2e	Conv. Starter	84	0.89 [0.86, 0.91]	0.92 [0.90, 0.94]	0.028	0.35 [0.04, 0.66]
Proportion, Love	2f	Suggestions	1248	0.44 [0.42, 0.47]	0.51 [0.49, 0.53]	< 0.001	0.14 [0.08, 0.20]
Proportion, Dislike	2f	Chat	126	0.03 [0.02, 0.04]	0.07 [0.06, 0.08]	< 0.001	0.17 [0.10, 0.23]

Table 1. Statistics and standardized effect sizes for key treatment and control comparisons from Fig. 2. For each sub-plot, we report control and treatment means with 95% confidence intervals estimated via bootstrap resampling, p -values, and Cohen's d with 95% confidence intervals. N denotes the number of observations. For a complete table with all statistics and effect sizes, see “Detailed Results and Effect Sizes for Main Text Fig. 2” in *Supplementary Information*.

the usage patterns by topical context, the nature of the engagement with the AI strongly vary: casual queries dominated the lighter subject of *Cats* (46.4%, $n = 159$), whereas participants primarily used the AI for fact checking in the scientific conversation of *Oats* (40.4%, $n = 131$), and for political discussions in the divisive topic of *Politics* (37.5%, $n = 110$). This adaptability suggests that participants tailored their AI-usage to the tone of each conversation. The features' flexibility and the users' likely familiarity with the interface contributed to the observed increases in engagement and production metrics. Yet, this may have induced verbose contributions, as reflected in longer sentence lengths (*Chat*: 28.59 words, vs *Control*: 18.76 words), which could explain the relatively low perceived quality and informativeness of comments, compared to the control condition (see Fig. 2b).

Conversation starters lower barriers to entry but are often misaligned with user intent

The *Conversation Starter* feature, designed to spark initial engagement or enrich ongoing interactions, was used by 71.7% ($n = 91$) of participants at least once, resulting in a total of 345 uses (Fig. 3c,d). When applying these starter suggestions, participants predominantly used open-ended or exploratory hints (*Questions*, 32.3%, $n = 115$). Yet a notable share of user comments (27.5%, $n = 98$) diverged from the AI-generated *Conversation Starters* entirely, suggesting users often dismiss the AI recommendations. This pattern of selective adoption was consistent across topics—usage rates around 52–54% for all subjects—reflecting a stable preference for exploratory, curiosity-driven engagement regardless of the conversational context. Although the *Conversation Starter* did help lower barriers to participation, user behavior and questionnaire answers suggest the aid regularly misaligned with their communication goals.

Feedback refines arguments in high-stakes discussions but is ignored when stakes were low

Participants' integration of AI-generated *Feedback* varied substantially by topic, reflecting diverse user priorities across context (Fig. 4e–g). In the casual *Cats* discussion, users typically made minimal or no textual changes after receiving AI feedback, with 41.7% of cases ($n = 50$) showing no edits to the original comment before submitting. By contrast, in the scientifically grounded *Oats* discussion, users more often revised their comments, often incorporating *Structural Changes* (18.8%, $n = 22$) or making *Informational Updates* (17.8%, $n = 18$). In *Politics*, the most common revisions involved *Argumentation* (17.3%, $n = 19$), highlighting the AI's role in supporting debate-oriented discourse. Overall adoption of the feedback feature was high (74.8%, $n = 98$), with a total of 331 uses, although per-topic usage ranged between 49.6% and 57.9% of the participants. These patterns indicate that users engaged with the *Feedback* tool more frequently on contexts where credibility, persuasion and clarity seemed most important, reflecting an intrinsic motivation for rhetorical strength.

In divisive contexts, users prefer AI suggestions that express agreement

The *Suggestions* feature, which offers context-dependent responses across agreeing, neutral, or disagreeing stances, showed moderate adoption (64.3%, $n = 83$), with a total of 1,197 generated suggestions selected (Fig. 4h,i). Overall, participants predominantly selected agreeing suggestions (48.6%, $n = 582$), with neutral (30.0%, $n = 359$) and disagreeing (21.4%, $n = 256$) responses selected less often. Users' preference for agreement varied by topic sensitivity: participants favored agreeing responses more strongly in higher-stakes or sensitive topics (*Oats*: 51.1%, *Politics*: 51.3%) compared to the less contentious topic (*Cats*: 43.3%). This pattern suggests a tendency to avoid divisive positions on controversial discussions, whereas disagreeing stances were more acceptable in lighter, low-risk conversations.

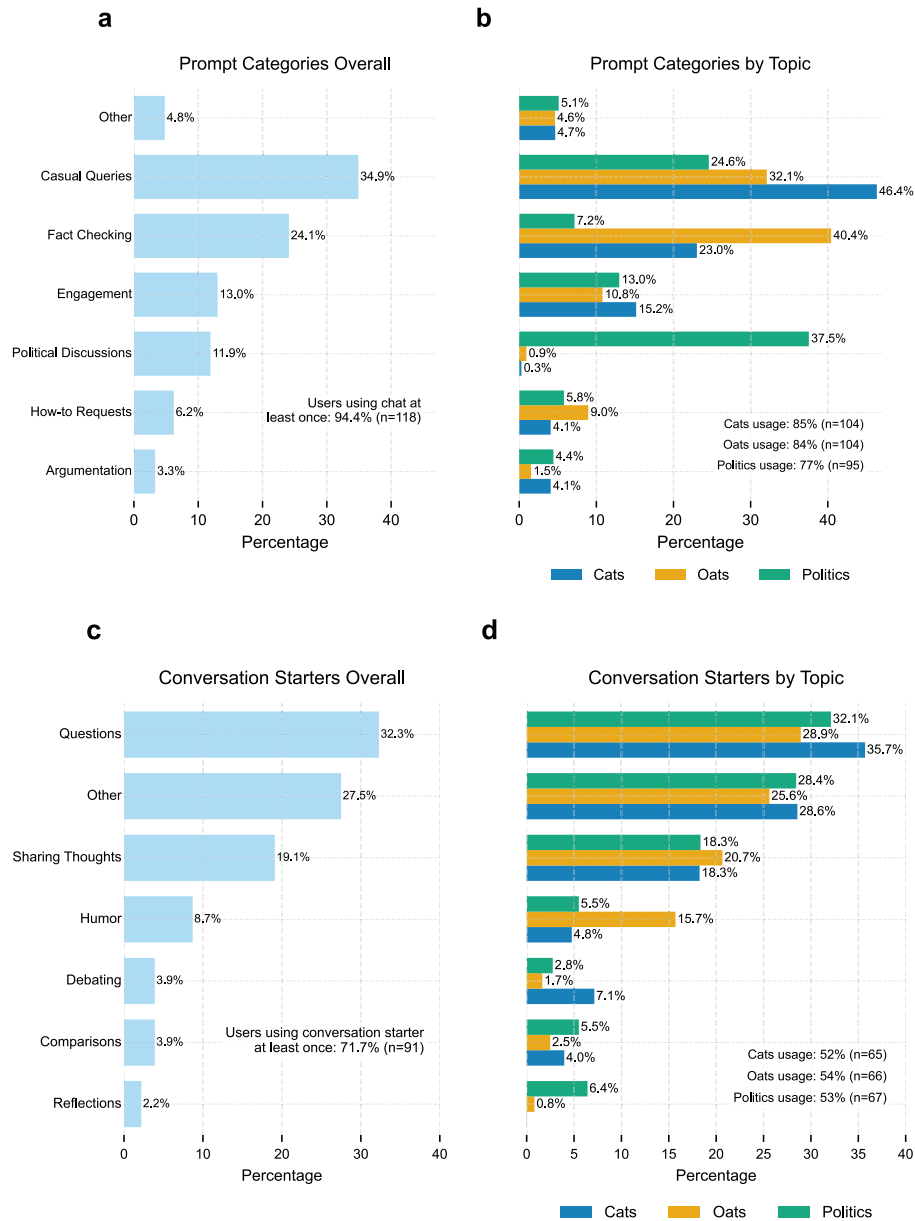


Fig. 3. Treatment usage. *Chat*: (a), Distribution of prompt categories overall. (b), Prompt categories by topics. *Conversation Starter*: (c), Distribution of Conversation Starters overall. (d), Conversation Starters by topic. Usages values indicate the fraction and number of participants who used the AI tool within topics. Overall proportions below 1% are excluded.

Participants value supportive AI but want more personalization

When asking participants for open-ended feedback on the AI tools after the study ($n = 245$ participants submitted written responses in the optional feedback field), users consistently emphasized the value of AI assistance for generating ideas, clarification, and initiating engagement—in particular when they felt “lost for words.” Despite differences across tools, users frequently described the AI as helpful for fact checking, comment reflection, and simplifying the process of contributing to the conversations. The tools would also help keep interactions constructive. In contrast, participants felt the AI content lacked authenticity and came across as overly generic, and requested more personalization of the AI—such as adapting to users’ style and context. Users reported that the *Chat* feature was useful for fact checking or exploring unfamiliar topics, noting that it made them feel more informed and confident in joining discussions. Still, some found the responses unnaturally formal or robotic. The *Conversation Starter* was valued for inspiration, especially for posing questions, but was also described as impersonal. The *Feedback* feature helped users reflect more deeply on their comments, though some noted that this could become exhausting or overly analytical. For *Suggestions*, users appreciated the simplicity and effectiveness in enabling quick replies. This made engagement more accessible, but some users found the suggestions impersonal.

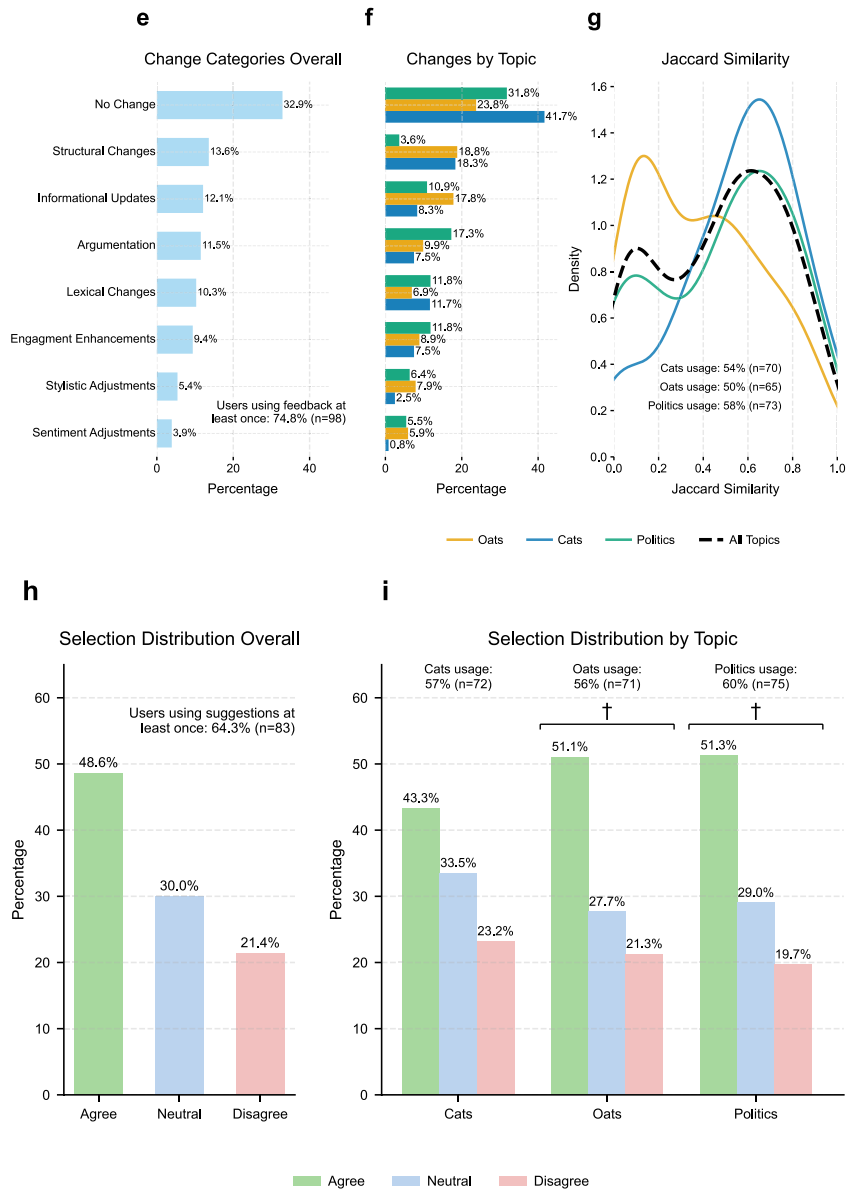


Fig. 4. Treatment usage. Feedback: (e), Distribution of feedback changes overall. (f), Feedback changes by topic. (g), Distribution of Jaccard similarity scores between original and revised comments overall and by topic. **Suggestions:** (h), Overall distribution of selected suggestions. (i), Selected suggestions by topic. Significance markers are based on Chi-square tests of independence comparing selection distributions between Cats and Oats, and Cats and Politics: †p < 0.10; *p < 0.05; **p < 0.01; ***p < 0.001.

How to move forward

These results entail important implications for the deployment of generative AI on social media platforms. Although participants in the treatment conditions produced more content, they often recognized it as generic, impersonal, and of lower quality. Such reduced perceived authenticity diminishes trust and informational value—risks that align with recent concerns about AI-assisted content flooding digital platforms with low-quality text^{31–33}. A central concern is that widespread use of generative AI may saturate platforms with superficial or generic content, diluting the visibility and impact of more original contributions—and, as our findings suggest, lowering the quality of subsequent conversations within threads, even among users not using the AI themselves. With AI systems more deeply integrated into social platforms, elements of dependability, transparency, and ethical aspects must be addressed to cultivate trust in AI¹⁰. One way to increase trust is for users to calibrate their reliance on AI based on the system’s capabilities and limitations³⁴.

Nevertheless, our findings show conditions under which AI tools may enhance public discourse. Participants valued the AI for idea generation, clarity, and initiating engagement—particularly in perceived high-stake or cognitive demanding topics. If tools are developed with greater personalization, contextual awareness, and stylistic nuances, they could support more inclusive and constructive interactions—particularly for individuals

typically hesitant to engage in public discussions. To guide ethical and effective deployment, we propose four design principles:

1. **Appeal of optional use and transparent disclosure:** Our findings suggest that AI tools are most positively received when offered as optional. Across all treatment conditions, usage patterns indicate selective engagement. In the *Chat* condition, participants submitted an average of 3.36 to 3.55 prompts per round, with only 13–16% of their final comments showed direct textual overlap. This suggests that users primarily engaged with the tool for ideation, clarification, or rhetorical guidance rather than direct copying. Similarly, for the *Conversation Starter*, only between 16% and 19% of submitted comments overlapped with AI suggestions—indicating that users often adapted, modified, or disregarded them based on text quality or alignment with their intent. The *Feedback* tool, although more cognitively demanding and less widely adopted, proved highly effective when used: in over 89% of cases where participants used the tool when composing a comment, they subsequently submitted it to the discussion. These behavioral signals are mirrored by open-ended survey responses. Participants reported appreciating AI support for generating ideas, retrieve information, and overcoming initial writer's block. Most treatment participants indicate they would use such tools if integrated into existing social media platforms (mean ratings on a 1-5 Likert scale: *Chat*: 3.97, *Conversation Starter*: 3.54, *Feedback*: 3.82, *Suggestions*: 3.94). At the same time, users noted easy identification of AI-generated content in others' posts—often describing the text as robotic, generic, or impersonal in the open-ended feedback. This perception was also found in the exit survey: the proportion of participants believed by the other users to use AI tools ranged from from 13.8% in the control group to 37.7–43.6% across treatments (*Control*: 13.8%, *Chat*: 38.9%, *Conversation Starter*: 37.7%, *Feedback*: 39.1%, *Suggestions*: 43.6%). This perception of diminished authenticity highlights a core challenge: although AI can reduce cognitive barriers and enhance participation, overuse may erode trust in the overall discourse. Together, these insights motivate a core design principle: transparent disclosure paired with user autonomy. In particular, content directly copied from generative AI tools should be clearly labeled^{35,36}, as such use might affect perceptions of authenticity. However, more nuanced uses—such as revising a draft with AI assistance or drawing inspiration from AI suggestions—may not require explicit labeling. Platform policies should account for these distinctions and support user agency in deciding how and when to engage with AI content. With this, platforms can preserve perceptions of authenticity³⁷, improve trust in human-AI interaction, and accommodate diverse user preferences. This aligns with broader algorithmic accountability where clearly defined responsibility and liability are established¹⁰.
2. **Personalization:** Across conditions, participants reported a desire for greater personalization of AI-generated content, more closely reflecting their individual voice, tone, and communicative intent. In the open-ended feedback, 34 participants explicitly described the AI responses as overly generic, noting it often felt impersonal. These perceptions are reflected in participants' ratings of comment quality and informativeness, as well as their assessments of replies received to their own comments, with treatment groups rated lower than the control (Fig. 2b,d). This suggests that although AI can support content generation, failing to adapt to user-specific context and style may lower perceived value of the interaction. Notably, in the *Feedback* condition—where users are guided in refining their own comments—we observe an increase in perceived value alignment with other participants, an increment from 33.9% to 42.5% (see “Ratings of Users and Comments” in *Supplementary Information*). This may indicate that tools enabling reflective personalization, rather than generic generation, could cultivate better identification and resonance among users. These behavioral and perceptual patterns underline the importance of AI systems that adapt to individual users. Future tools should incorporate stylistic adaptation and learn from prior interactions to provide outputs that is perceived personal, relevant, and authentic.
3. **Contextual awareness, flexibility, and authenticity:** Effective AI support in social media conversations requires responsiveness to the context of the discussion. In our study, participants engaged with the AI in varying ways depending on topic sensitivity. In the *Chat* condition, users most often engaged the AI with casual, topic-related prompts during the *cats vs. dogs* discussion—asking open-ended questions that reflected a conversational, informal, and low-effort use of the tool. In the scientific *Oats* discussion, participants frequently used the AI as an informational resource, asking for facts and clarification. In contrast, political discussion queries were most common in the political topic, where users asked for reflections, viewpoints, and pros and cons—using the AI more analytically to explore perspectives and argumentative reasoning. This adaptive usage suggests that participants naturally attuned expectations for AI based on the conversational subject. In the *Suggestions* condition, users preferred agreement responses in the higher-stakes *oats* and *politics* discussions, but were more willing to express disagreement or neutrality in the *cats* topic. In the *Chat* condition, 12 out of 58 participants who provided open-ended AI-feedback specifically noted that the AI tool was helpful when engaging with unfamiliar topics—supporting that AI can lower barriers to entry in online discussions. However, participants reported lack of personal style and authenticity, mismatching individual demands in the conversation. This points to a central design challenge: fixed-responses that ignore topical and personal risk undermining authenticity and usefulness of AI-assisted content. These findings motivate a design principle for AI tools to embed and understand contextual sensitivity. This imply adjusting conversational intent, adjusting tone, stance, and formality based on topic domain and personal preferences. Informality may help in trivial discussions, but scientific or political contexts appeal toward factual elements, rhetorical nuances, and societal understanding. Systems that fail to accommodate such topical diversity may reduce trust or suppress meaningful engagement. Flexibility and authenticity—grounded in both topic and user intent—should be central to the development of socially integrated AI systems on social media. Platform designers should encourage awareness and transparency about biases within AI systems, recognizing that no system can be truly *objective* or *fair*¹⁰.

4. **Familiar and user-friendly interfaces:** Ease of use is central in the adoption and effectiveness of AI tools for content creation on social media platforms. In the *Chat* condition, designed with a familiar and low-friction sidebar interface, we find the most wide adoption (94.4% used it at least once). In contrast, the *Feedback*, *Conversation Starter*, and *Suggestions* tools, showed lower adoption rates (*Feedback*: 74.81%, *Conversation Starter*: 71.65%, *Suggestions*: 64.34%). This underlines the trade-offs between tool abundance and user effort. These behavioral and perception perspectives affirm the importance of intuitive and convenient interfaces in encouraging tool engagement. Participants in the *Chat* and *Suggestions* conditions find these tools as more supportive overall—reporting that they made participation easier, felt more intuitive to use, and led to higher-quality contributions—compared to those using *Feedback* and *Conversation Starter* (see “Supplementary Figures” in *Supplementary Information*). Altogether, these findings support a key design principle: simplicity and familiarity in user interfaces are vital to enhance adoption and usage. When AI tools are embedded seamlessly into platform workflows—with easy entry points and low interaction costs—they are more regarded as a supportive mechanism rather than disruptive. Future implementations should prioritize clarity and accessibility by embedding AI tools into the natural flow of conversation, minimizing user effort without sacrificing optional depth.

To more deeply understand the dynamics of AI-assisted interactions, future work should build on this experimental foundation by scaling across time, platforms, and population. One critical extension involves capturing temporal AI tool usage—how engagement patterns change over longer exposure, whether the innovation effects diminish, and how personalization adapt over longer interactions. Expanding the sample size would also support more robust cross-treatment comparisons and enable better analysis of subgroup diversity. Deploying similar interventions on other social platforms—within controlled environments—would offer deeper ecological validity and understanding of how AI tools affect different social ecosystems online. Our study establishes a baseline for controlled experimentation with integrated AI support for content-creation, but continuous progress requires research across diverse online environments.

An ethical deployment of AI tools on social media necessitates continuous auditing. This includes not only how user data is processed and how model biases are mitigated, but also how such tools may affect and reshape collective behavior over time. Additionally, transparency in AI usage, guardrails to prevent misinformation and marginalization, and mechanisms for user control must be built into deployments¹⁸. As AI tools are introduced into social media environments already optimized for sustained engagement, they may further influence conversational dynamics. Our controlled experiment captures short-term, user-level interaction dynamics. Long-term effects of AI assistance on social media will emerge through cumulative patterns of use, repeated exposure, and feedback loops. Even subtle interventions—such as nudging users toward agreement or increasing comment length—may gradually change conversational norms, influence tone, speed of engagement, and the inclusiveness of discussions. Recent work demonstrates how prolonged interactions with AI can amplify pre-existing biases, although unbiased AI systems can effectively improve human judgments³⁸, suggesting that long-term effects may depend on bias mitigation. Most of today’s AI systems are black-box algorithms with limited transparency of decision-making. This lack of transparency can undermine trust and challenge accountability, in particular in high-stake contexts¹⁰. As we find, AI lowers barriers to participation, and has also been found to improve individual creativity, though at the cost of collective content diversity, potentially amplifying generic perceptions of AI-assisted content³⁹. Initial negative perceptions might decrease over long-term exposure as users gain familiarity. However, this normalization could alternatively imply reduced vigilance as use and exposure become standard⁴⁰, particularly as authenticity perceptions shift when AI involvement is disclosed⁴¹. From a cognitive perspective, long-term AI assistance can support mental offloading when solving tasks. At the same time, AI limits active recall and problem-solving, which are essential for cognitive development⁴²—a trade-off described as *desirable difficulties*, where elements that are challenging to learn yield better long-term retention⁴³. Societal impact remains uncertain, but AI tools can alter how persuasive, sensitive, or political content disseminates, with potentially impactful consequences for the fragile integrity of public discourse.

AI content-generation tools are naturally becoming a fundamental layer of digital interaction. Our findings highlight both promise and risks. When designed with transparency, personalization, and contextual flexibility, these tools can lower cognitive barriers, broaden participation, and support more inclusive engagement. But if deployed without careful consideration of their emergent effects, social media platforms risk saturating public discourse with generic, inauthentic content, undermining quality and trustworthiness of online conversations. The future of democratic communication online will depend not only on the capabilities of AI tools, but on how thoughtfully they are embedded into the social communicative ecosystem of digital platforms. Ethical integration of AI requires interdisciplinary collaboration ranging philosophy, law social sciences, and computer science to develop systems that prioritize human values and societal benefits^{10,20}.

Methods

In our experiment, participants progressed through a structured experimental flow. After providing informed consent, they (i) completed a pre-survey assessing demographics and attitudes toward social media and AI. Participants were then (ii) onboarded to the platform and the task via instructions and a demonstration video, (iii) engaged in three 10-minute discussions in randomized order, (iv) completed a post-study survey evaluating their experience and rating content and other participants, and (v) finally received their compensation token. Full experimental details, survey questions, distribution of answers, seed content, and evaluation are available in the *Supplementary Information*.

All methods were carried out in accordance with relevant guidelines and regulations. University of Michigan Institutional Review Board (IRB) approved the study protocol (HUM00258995). Participants were compensated \$9.50 USD for an estimated 45-minute session, corresponding to an hourly rate of \$12.66.

Platform design

We built a custom online discussion platform to simulate a typical forum-based social media environment, adopted into Empirica⁴⁴ to handle experiment logistics (see Fig. 1 for platform screenshots). The design was inspired by platforms like Reddit with threaded conversations and lightweight user interaction. Participants could post comments and react using seven emoji-based responses (like, love, dislike, angry, wow, sad, funny), with “love” set as the default. A U.S.-representative sample of 680 participants—recruited via Prolific between January 3 and January 21, 2025—was randomly put into groups of five to mimic small-group online discussions and maintain experimental control (Prolific optionally supports recruitment of U.S.-representative samples based on key demographics such as age, sex, ethnicity, and political affiliation. See <https://researcher-help.prolific.com/en/article/e6555f>). Each group was randomly assigned to one of the five experimental conditions: a control group with no AI assistance, or one of the four AI-assisted treatment conditions. Participants interacted only within their assigned group and had no prior knowledge about the other users. Each group engaged in three 10-minute discussions on different predefined topics on varying social sensitivity and complexity: one trivial (*cats vs. dogs*), one scientific (health benefits of *oats*), and one political (*universal basic income*). To control for order effects, the sequence of topics was randomized across groups. We source the initial seed data—the content shown at the start of each 10-minute discussion—from existing Reddit threads^{45–47}. Each comment displayed a timestamp, interactive buttons for commenting and reacting, visual indentation to indicate thread structure, and a parent ID referring to the comment it responded to. Reactions would be shown to all participants but without identification of who reacted. New comments were highlighted with a brief alert and a “new” tag lasting one minute to support real-time flow.

The platform was designed to minimize distractions and reduce cognitive load, allowing participants to focus on the conversation. The setup ensured controlled conditions while preserving key elements of real-world online discussions. Before entering the discussion platform, participants were shown a brief onboarding interface introducing the platform and their task. This included a written description of the discussion structure and a short demonstration video showing how to navigate the platform, post comments, and react to content. For participants assigned to AI-assisted conditions, additional instructions were provided explaining the specific tool available. The demonstration video included a walkthrough of how to access and use the AI feature. During each 10-minute discussion round, we log user activity at the individual level, including comments posted, reactions given, and use of AI-tools.

AI tools

To test the impact of different paradigms of AI assistance in social media discussions, we integrated four distinct AI tools into the platform. Each tool was designed to reflect approaches to AI-assisted communication found in academic literature and commercial products. All tools were powered by GPT-4o and with a custom prompt tailored to each intervention (see “AI Prompts and Settings” in *Supplementary Information*). Participants were not informed which AI model was used in treatments or whether any content they encountered was AI-generated.

Chat The *Chat* tool allowed for open-ended interaction with an AI assistant through a sidebar window displayed alongside the conversation thread (Fig. 1f). Participants could engage with the assistant up to eight times per discussion topic. The interface supported informal querying, idea generation, and clarification, giving users flexibility to steer the interaction as desired.

Conversation Starter The *Conversation Starter* generated AI-suggested openings for participation, aimed at lowering barriers to entry and stimulating discussion (Fig. 1b). The tool was accessed through a separate button on each comment. The conversation starting suggestions were context dependent but could include follow-up questions, engaging comments, and reflective or contextual statements.

Feedback The *Feedback* tool offered real-time guidance on draft comments prior to submission (Fig. 1d). Once users began typing a comment, they could click on a dedicated “AI Feedback” button to receive tailored suggestions on how to improve their comment. The feedback varied based on context and comment draft, but could include ideas to clarifying arguments, add personal anecdotes, or maintain a balanced tone. The feedback appeared inline below the draft comment, and users could receive three rounds of feedback per comment.

Suggestions The *Suggestions* feature provided three AI-generated replies—each adopting a distinct stance (agree, neutral, disagree)—in response to any selected comment (Fig. 1e). The tool was accessed through the comment modal. Participants could regenerate a new set of suggestions up to three times per comment, allowing them to explore alternatives before selecting a reply.

Each of these tools was embedded into the platform interface to mirror familiar social media interactions while maintaining clarity and minimalism. The tools were designed to be optional, harmoniously integrated into the platform, and supportive of the natural flow of the discussion.

Questionnaires

Pre-study questionnaire The pre-study questionnaire gathered demographic information (age, gender, education level, occupation, and political affiliation) and attitudes toward online discourse and AI. Participants reported their typical engagement with social media, perceived quality of online discussions, trust in user-generated content, and perceived barriers to participation. Responses were recorded using multiple-choice items and five-point Likert scales. Participants were also asked about their attitudes toward the use of AI on social media platforms. This included perceived impact of AI on participation, comfort, content quality, misinformation, toxicity, polarization, and the need for regulation. These questions also used Likert-scale responses. The full set of questions is provided in *Supplementary Information* under “Pre-study Questionnaire”.

Post-study questionnaire Following the experiment, participants answered a second survey capturing their experience on the platform. This included questions about how their participation compared to typical online behavior, their perceptions of quality, and trust in other users. Participants in AI treatment conditions were

asked to evaluate the usefulness, quality, and relevance of the AI tool they used. Users were also invited to provide open-ended feedback on their experience with the AI and the platform. To assess perceptual changes, participants answered the same questions about AI on social media from the pre-study questionnaire. This allowed us to quantify shifts in attitude as a result of the experiment. Finally, participants were asked to evaluate the quality of social interaction in their group. They rated 10 replies received on their own comments using a five-point scale to assess the value in the discussion. They also evaluated the other participants in their group on perceived politeness, engagement, political agreement, shared values, use of AI, and whether they might be a bot. Full set of questionnaire items is provided in the *Supplementary Information* under “Post-study Questionnaire”.

Evaluation

We evaluated both behavioral engagement and perceptual effects across producer and consumer perspectives, using a combination of tracked user behavior and self-reported measures from the questionnaires. We used Shannon entropy to measure participation equality within each group, based on the proportions of comments made by each user in a given round. Details on this calculation are provided in *Supplementary Information* under “Participation Equality (Normalized Shannon Entropy)”. We also tracked AI tool usage, allowing us to assess adoption patterns. To evaluate the perceived quality of the interactions, we analyzed responses from the post-study questionnaire.

For all behavioral and perceptual metrics—including questionnaire responses, entropy, comment length, comment ratings, and distributions of reaction types—we used nonparametric bootstrapping with resampling to estimate uncertainty around group-level means. To test for differences between treatment and control groups, we used permutation tests for Likert-scale and ordinal responses, and two-sided *t*-tests for continuous measures.

We also performed regression analyses to estimate the impact of treatment conditions and participant demographics on behavioral and perceptual measures. First, for each treatment group, we fitted a separate generalized linear model with a binomial distribution to assess the likelihood of a comment receiving a reply. Independent variables capture discussion dynamics at the time of commenting, including topic (categorical), normalized time remaining in the discussion, comment depth, number of active users (defined as users who had posted at least once), number of prior comments, and whether the AI tool was used. Second, we conducted a series of ordinary least squares regressions to examine how demographic characteristics—age (≥ 45), gender (female or not), education (college degree or not), occupation (full-time employment or not), and political affiliation (Republican, Independent, Democrat)—as well as treatment group, predicted two types of outcome variables: (1) differences between pre- and post-study responses to questions on *AI Related to Social Media*, and (2) user ratings of replies to own comments. Detailed information and statistical results are reported under “Supplementary Material” in *Supplementary Information*.

AI usage analyses

To better understand how participants use the AI tools, we developed a structured classification pipeline tailored to each interaction. For *Chat*, *Conversation Starter*, and *Feedback*, we first constructed taxonomies of typical uses based on manual inspection of user interactions. We then used OpenAI’s `o3-mini-2025-01-31` to classify the individual tool uses according to these taxonomies.

For the *Chat* tool, we inspected submitted user-prompts and construct a taxonomy of eight broad prompt types—*casual queries*, *fact checking*, *engagement*, *political discussions*, *how-to requests*, *argumentation*, *sentiment and context analysis*, *conspiracy*—plus an *other* category, each with a short description. The LLM then classified all prompts into these categories.

A similar approach was used for the *Conversation Starter*: an LLM classified which of the AI-generated conversational suggestions directly inspired a user’s submitted comment. The Conversation Starters were categorized into ten broader themes—*practical advice and suggestions*, *personal experiences and anecdotes*, *animal behavior and intelligence*, *research and science discussions*, *reflections*, *debating*, *comparisons*, *humor*, *sharing thoughts*, *questions*, and *other*.

For the *Feedback* tool, we identified how participants revised their comments in response to the AI feedback. We defined a taxonomy of seven revision types—*structural changes*, *informational updates*, *argumentation*, *lexical changes*, *engagement enhancements*, *stylistic adjustments*—plus categories for *other* and *no change*. The LLM classified the type of revision by comparing the user’s initial draft the submitted comment.

Finally, for the *Suggestions* tool, we directly logged which of the three AI suggestion stances (agree, neutral, disagree) the participants selected. This classification framework formed the basis of the usage patterns reported in Figs. 3a-d and 4e-i. Full prompt structures and category definitions are detailed under “Supplementary Materials” in *Supplementary Information*.

Data availability

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Code availability

The code to run the experiment and fully reproduce the analyses described in this work is publicly available as archived releases. The experimental platform code is available at [Zenodo \(DOI: 10.5281/zenodo.18539373\)](https://doi.org/10.5281/zenodo.18539373). The analysis code to reproduce figures and statistical tests is available at [Zenodo \(DOI: 10.5281/zenodo.18537773\)](https://doi.org/10.5281/zenodo.18537773).

Received: 27 October 2025; Accepted: 10 February 2026

Published online: 17 February 2026

References

- Ziems, C. et al. Can large language models transform computational social science?. *Comput. Linguist.* **50**(1), 237–291. https://doi.org/10.1162/coli_a_00502 (2024).
- Xi, Z. et al. The rise and potential of large language model based agents: A survey. *Sci. China Inf. Sci.* **68**(2), 121101. <https://doi.org/10.1007/s11432-024-4222-0> (2025).
- Anantrasirichai, N. & Bull, D. Artificial intelligence in the creative industries: A review. *Artif. Intell. Rev.* **55**(1), 589–656. <https://doi.org/10.1007/s10462-021-10039-7> (2022).
- Pavlik, J. V. Collaborating with chatgpt: Considering the implications of generative artificial intelligence for journalism and media education. *J. Mass Commun. Educat.* **78**(1), 84–93. <https://doi.org/10.1177/10776958221149577> (2023).
- Yan, L., Greiff, S., Teuber, Z. & Gašević, D. Promises and challenges of generative artificial intelligence for human learning. *Nat. Hum. Behav.* **8**(10), 1839–1850. <https://doi.org/10.1038/s41562-024-02004-5> (2024).
- Jennings, F. J., Suzuki, V. P. & Hubbard, A. Social media and democracy: Fostering political deliberation and participation. *West. J. Commun.* **85**(2), 147–167. <https://doi.org/10.1080/10570314.2020.1728369> (2021).
- Jakesch, M., Bhat, A., Buschek, D., Zalmanson, L. & Naaman, M. Co-writing with opinionated language models affects users' views. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. CHI '23, pp. 1–15. Association for Computing Machinery, New York, NY, USA (2023). <https://doi.org/10.1145/3544548.3581196>. <https://dl.acm.org/doi/10.1145/3544548.3581196>.
- Bail, C. A. Can generative AI improve social science?. *Proc. Natl. Acad. Sci.* **121**(21), 2314021121. <https://doi.org/10.1073/pnas.2314021121> (2024).
- Shin, D. Artificial Misinformation: Exploring human-algorithm interaction online. Springer, Cham (2024). <https://doi.org/10.1007/978-3-031-52569-8>. <https://link.springer.com/10.1007/978-3-031-52569-8>.
- Shin, D. Debiasing AI: Rethinking the intersection of innovation and sustainability. Routledge, New York <https://doi.org/10.1201/9781003530244> (2025).
- Hancock, J. T., Naaman, M. & Levy, K. Ai-mediated communication: Definition, research agenda, and ethical considerations. *J. Comput.-Mediat. Commun.* **25**(1), 89–100. <https://doi.org/10.1093/jcmc/zmz022> (2020).
- Reif, J. A., Larrick, R. P. & Soll, J. B. Evidence of a social evaluation penalty for using AI. *Proc. Natl. Acad. Sci.* **122**(19), 2426766122. <https://doi.org/10.1073/pnas.2426766122> (2025).
- Wingström, R., Hautala, J. & Lundman, R. Redefining creativity in the era of AI? perspectives of computer scientists and new media artists. *Creat. Res. J.* **36**(2), 177–193. <https://doi.org/10.1080/10400419.2022.2107850> (2024).
- Rattanaseevee, P., Akarapattananukul, Y. & Chirawat, Y. Direct democracy in the digital age: Opportunities, challenges, and new approaches. *Humanit. Social Sci. Commun.* **11**(1), 1–9. <https://doi.org/10.1057/s41599-024-04245-1> (2024).
- Mikhaylovskaya, A. Enhancing deliberation with digital democratic innovations. *Philos. Technol.* **37**(1), 3. <https://doi.org/10.1007/s13347-023-00692-x> (2024).
- Tessler, M. H. et al. AI can help humans find common ground in democratic deliberation. *Science* **386**(6719), 2852. <https://doi.org/10.1126/science.adq2852> (2024).
- Doshi, A. R. & Hauser, O. P. Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Sci. Adv.* **10**(28), 5290. <https://doi.org/10.1126/sciadv.adn5290> (2024).
- Drolsbach, C. P., Solovev, K. & Pröllöchs, N. Community notes increase trust in fact-checking on social media. *PNAS Nexus* **3**(7), 217. <https://doi.org/10.1093/pnasnexus/pgae217> (2024).
- Osborne, M. R. & Bailey, E. R. Me vs. the machine? subjective evaluations of human- and AI-generated advice. *Sci. Rep.* **15**(1), 3980. <https://doi.org/10.1038/s41598-025-86623-6> (2025).
- Russo, D., Baltés, S., Berkel, N., Avgeriou, P., Calefato, F., Cabrero-Daniel, B., Catolino, G., Cito, J., Ernst, N., Fritz, T., Hata, H., Holmes, R., Izadi, M., Khomh, F., Kjærgaard, M.B., Liebel, G., Lafuente, A.L., Lambiase, S., Maalej, W., Murphy, G., Moe, N.B., O'Brien, G., Paja, E., Pezzè, M., Persson, J.S., Prikladnicki, R., Ralph, P., Robillard, M., Silva, T.R., Stol, K.-J., Storey, M.-A., Stray, V., Tell, P., Treude, C., & Vasilescu, B. Generative AI in software engineering must be human-centered: The copenhagen manifesto. *J. Syst. Softw.* <https://doi.org/10.1016/j.jss.2024.112115> (2024).
- Yang, K., Singh, D., & Menczer, F (2024) Characteristics and prevalence of fake social media profiles with ai-generated faces. *J. Online Trust Safety*, <https://doi.org/10.54501/jots.v2i4.197>.
- Argyle, L. P. et al. Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale. *Proc. Natl. Acad. Sci.* Publisher: Proceedings of the National Academy of Sciences. Accessed 2024-07-24 **120**(41), 2311627120. <https://doi.org/10.1073/pnas.2311627120> (2023).
- Di Fede, G., Rocchesso, D., Dow, S.P. & Andolina, S. The idea machine: Llm-based expansion, rewriting, combination, and suggestion of ideas. In: Proceedings of the 14th Conference on Creativity and Cognition. C&C '22, pp. 623–627. Association for Computing Machinery, New York, NY, USA (2022). <https://doi.org/10.1145/3527927.3535197>. <https://dl.acm.org/doi/10.1145/3527927.3535197>.
- Ziegenbein, T., Skitalinskaya, G., Bayat Makou, A. & Wachsmuth, H. LLM-based Rewriting of Inappropriate Argumentation using Reinforcement Learning from Machine Feedback. In: Ku, L.-W., Martins, A., Srikumar, V. (eds.) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 4455–4476. Association for Computational Linguistics, Bangkok, Thailand (2024). 10.18653/v1/2024.acl-long.244. Accessed 2024-10-30.
- Do, H.J., Kong, H.-K., Lee, J. & Bailey, B.P. How should the agent communicate to the group? communication strategies of a conversational agent in group chat discussions. *Proc. ACM Hum.-Comput. Interact.* **6**(CSCW2), 387–138723 (2022) <https://doi.org/10.1145/3555112>.
- Lee, M., Gero, K.I., Chung, J.J.Y., Shum, S.B., Raheja, V., Shen, H., Venugopalan, S., Wambsgans, T., Zhou, D., Alghamdi, E.A., August, T., Bhat, A., Choksi, M.Z., Dutta, S., Guo, J.L.C., Hoque, M.N., Kim, Y., Knight, S., Neshaei, S.P., Shibani, A., Shrivastava, D., Shroff, L., Sergeyuk, A., Stark, J., Sterman, S., Wang, S., Bosselut, A., Buschek, D., Chang, J.C., Chen, S., Kreminski, M., Park, J., Pea, R., Rho, E.H.R., Shen, Z. & Siangliulue, P. A design space for intelligent and interactive writing assistants. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. CHI '24, pp. 1–35. Association for Computing Machinery, New York, NY, USA <https://doi.org/10.1145/3613904.3642697>. <https://dl.acm.org/doi/10.1145/3613904.3642697> (2024).
- Horvitz, E. Principles of mixed-initiative user interfaces. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '99, pp. 159–166. Association for Computing Machinery, New York, NY, USA (1999). <https://doi.org/10.1145/302979.303030>. <https://dl.acm.org/doi/10.1145/302979.303030>.
- Dhillon, P.S., Molaie, S., Li, J., Golub, M., Zheng, S. & Robert, L.P. Shaping human-AI collaboration: Varied scaffolding levels in co-writing with language models. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. CHI '24, pp. 1–18. Association for Computing Machinery, New York, NY, USA (2024). <https://doi.org/10.1145/3613904.3642134>. <https://dl.acm.org/doi/10.1145/3613904.3642134>.
- Floridi, L. & Chiriatti, M. Gpt-3: Its nature, scope, limits, and consequences. *Mind. Mach.* **30**, 681–694 (2020).

30. Chatterji, A., Cunningham, T., Deming, D.J., Hitzig, Z., Ong, C., Shan, C.Y. & Wadman, K. How people use chatgpt. Working Paper 34255, National Bureau of Economic Research (September 2025). <https://doi.org/10.3386/w34255>. <http://www.nber.org/papers/w34255>.
31. Feuerriegel, S. et al. Research can help to tackle AI-generated disinformation. *Nat. Hum. Behav.* 7(11), 1818–1821. <https://doi.org/10.1038/s41562-023-01726-2> (2023).
32. Yang, K.-C. & Menczer, F. Anatomy of an ai-powered malicious social botnet. *J. Quantit. Descript. : Digital Media* <https://doi.org/10.51685/jqd.2024.icwsm.7> (2024).
33. Wei, Y. & Tyson, G. Understanding the impact of ai-generated content on social media: The pixiv case. In: Proceedings of the 32nd ACM International Conference on Multimedia. MM '24, pp. 6813–6822. Association for Computing Machinery, New York, NY, USA (2024). <https://doi.org/10.1145/3664647.3680631>. <https://doi.org/10.1145/3664647.3680631>.
34. Afroogh, S., Akbari, A., Malone, E., Kargar, M. & Alambeigi, H. Trust in AI: Progress, challenges, and future directions. *Humanit. Social Sci. Commun.* 11(1), 1568. <https://doi.org/10.1057/s41599-024-04044-8> (2024).
35. Gamage, D., Sewwandi, D., Zhang, M. & Bandara, A. Labeling synthetic content: User perceptions of warning label designs for ai-generated content on social media. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, pp. 1–29 (2025). <https://doi.org/10.1145/3706598.3713171>. [arXiv:2503.05711](https://arxiv.org/abs/2503.05711) [cs].
36. Gallego, I.O., Shani, C., Shi, W., Bianchi, F., Gainsburg, I., Jurafsky, D. & Willer, R. Labeling messages as ai-generated does not reduce their persuasive effects ([arXiv:2504.09865](https://arxiv.org/abs/2504.09865)) (2025) [arXiv:2504.09865](https://arxiv.org/abs/2504.09865) [cs].
37. Kirkby, A., Baumgarth, C. & Henseler, J. To disclose or not disclose, is no longer the question - effect of AI-disclosed brand voice on brand authenticity and attitude. *J. Prod. Brand Manag.* 32(7), 1108–1122. <https://doi.org/10.1108/JPB-02-2022-3864> (2023).
38. Glickman, M. & Sharot, T. How human-AI feedback loops alter human perceptual, emotional and social judgements. *Nat. Hum. Behav.* 9(2), 345–359. <https://doi.org/10.1038/s41562-024-02077-2> (2025).
39. Doshi, A. R. & Hauser, O. P. Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Sci. Adv.* 10(28), 5290. <https://doi.org/10.1126/sciadv.adn5290> (2024).
40. Polypartis, A. A longitudinal study on artificial intelligence adoption: Understanding the drivers of chatgpt usage behavior change in higher education. *Front. Artif. Intell.* <https://doi.org/10.3389/frai.2023.1324398> (2024).
41. Jain, G., Pareek, S. & Carlbring, P. Revealing the source: How awareness alters perceptions of AI and human-generated mental health responses. *Intern. Intervent.* 36, 100745. <https://doi.org/10.1016/j.invent.2024.100745> (2024).
42. Jose, B., Cherian, J., Verghis, A.M., Varghise, S.M., S, & M., Joseph, S. The cognitive paradox of AI in education: Between enhancement and erosion 16 <https://doi.org/10.3389/fpsyg.2025.1550621> (2025).
43. Bjork, E. & Bjork, R. Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society*, 56–64 (2011).
44. Almaatouq, A. et al. Empirica: A virtual lab for high-throughput macro-level experiments. *Behav. Res. Methods* 53(5), 2158–2171. <https://doi.org/10.3758/s13428-020-01535-9> (2021).
45. chrisdh79, Sufficient-Cover5956, Anticitizen-Zero: Discussion on cats and dogs. https://www.reddit.com/r/changemyview/comments/106ybx/cmv_cats_are_smarter_than_dogs_on_average/. Accessed: 2024–07–31 (2023).
46. chrisdh79, Sufficient-Cover5956, Anticitizen-Zero: Discussion on health benefits of oats. https://www.reddit.com/r/science/comments/1e9anm6/weightloss_power_of_oats_naturally_mimics_popular/. Accessed: 2024–07–31 (2024).
47. whatisgoingon123422, CutieHeartgoddess: Discussion on Universal Basic Income. https://www.reddit.com/r/changemyview/comments/tdmuae/cmv_universal_basic_income_is_the_way_of_the/. Accessed: 2024–07–31 (2022).

Acknowledgements

We thank the Network, Data, and Society (NERDS) group at IT University of Copenhagen, the Blablablab, and the Romero group at the School of Information, University of Michigan, for valuable feedback during internal testing.

Author contributions

A.G.M, D.R., D.J., and L.M.A. designed the research. A.G.M. developed the platform, and collected and analyzed the data. A.G.M, D.R., D.J., and L.M.A. wrote the paper.

Funding

We acknowledge the support from the Carlsberg Foundation through the COCOONS project (CF21-0432) and the National Science Foundation through Grant No. IIS-2143529.

Declarations

Competing Interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-026-40110-8>.

Correspondence and requests for materials should be addressed to A.G.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026