https://doi.org/10.1038/s44387-025-00049-z

Exposure to content written by large language models can reduce stigma around opioid use disorder

Check for updates

Shravika Mittal[1] ✉, Darshi Shah[1], Shin Won Do[1], Mai ElSherief[2], Tanushree Mitra[3] & Munmun De Choudhury[1]

Widespread stigma, both offline and online, hinders harm reduction efforts in the context of opioid use disorder (OUD). This stigma targets clinically approved medications for OUD (MOUD), people with the condition, and the condition itself, among several others. Given the potential of artificial intelligence in promoting health equity, this work examines whether large language models (LLMs) can abate stigmatizing attitudes in virtual healthcare communities. To answer this, we conducted a series of randomized controlled experiments, where participants read LLM-generated, human-written, or no responses to help-seeking OUD-related content. The experiment was conducted under two setups: participants read the responses either once (N = 2, 141) or repeatedly for 14 days (N = 107). Participants reported the least stigmatized attitudes toward MOUD after consuming LLM-generated responses. This study offers insights into strategies that can foster inclusive discourse on OUD. Based on our findings LLMs can serve as an education-based intervention to promote positive attitudes and increase people's propensity toward treatments for OUD.

In the U.S., opioid overdose continues to be a leading cause of death. The National Center for Health Statistics at the U.S. Centers for Disease Control and Prevention reported that 105,007 drug overdose deaths occurred in 2023. Of these, 79,358 were attributed to opioids[1]. Addressing this crisis requires reducing barriers to treatment, which manifest as widespread stigma[2] in the offline settings targeting the condition[3–5], i.e., opioid use disorder (OUD), people with OUD[6–8], and clinically approved medications for opioid use disorder (MOUD)[4,9,10], among several others. To avoid such social, structural, and intervention-based marginalization, people with OUD may seek non-conventional, virtual pathways to recovery[11,12]. Owing to pseudonymity, virtual health communities, such as those on Reddit, are particularly popular among people with OUD[11,13]. Individuals use these forums to freely discuss opioid use[14,15], to explore alternative treatment frameworks[13], and to offer and seek support around issues relating to recovery[11,16].

However, online communities are not necessarily always the safe spaces they aspire to be. They are home to vast amounts of unhelpful advice, misleading, and clinically unverified content,[17–20], and can exacerbate the marginalization of already vulnerable populations[21–23]. Human-written responses on platforms such as Reddit are known to contain stigmatized attitudes toward MOUD[24,25] and can reinforce harmful stereotypes by normalizing language that labels people with OUD as "addicts" or frames their condition as a personal failing[26]. It is imperative to improve the information ecosystem by fostering an environment that reduces stigma, as for people with OUD, such platforms may be their primary source of information, support, and validation.

With advancements in artificial intelligence (AI) based technologies, scholars have started to investigate the potential of AI to augment human capabilities across a wide range of creative, complex, and high-risk tasks—e.g., using text-based large language models (LLMs) to promote health equity[27], to facilitate empathic conversations in peer to peer mental health support[28–30], to provide mental health treatment[31], and to motivate behavior change[32]. Recently, Bouzoubaa et al. used LLMs to re-frame stigmatizing online content directed at people who use substances into more empathetic language[33]. Building on this and other prior work, this paper shifts attention to how LLMs may function as information-providers, specifically when writing responses to help-seeking OUD-related content in online communities. Broadly, we seek to understand if LLMs can support ongoing efforts toward reducing OUD-related stigma in these virtual spaces by examining their impact on attitudes and perceptions of everyday information-seekers.

To do so, we conducted a series of pre-registered randomized controlled experiments. Three prevalent sources known to amplify OUD-related stigma in online communities are negative attitudes toward (a)

[1]College of Computing, Georgia Institute of Technology, Atlanta, GA, USA. [2]Khoury College of Computer Sciences, Northeastern University, Boston, MA, USA. [3]Information School, University of Washington, Seattle, WA, USA. ✉e-mail: smittal87@gatech.edu

MOUD[24,25,34], e.g., they are framed as merely replacing one drug with another; (b) people with OUD[34], e.g., they are portrayed as "dangerous" or "weak"; and (c) OUD itself[35], e.g., full recovery from the condition is at times represented as impossible. Therefore, we operationalized these three attitudes as our dependent variables (DVs). In our experiments, we assessed whether LLM-written content could reduce stigmatized attitudes toward MOUD (DV1), people with OUD (DV2), and OUD (DV3). Using a between-subjects study design, participants were randomly assigned to one of three interventions: (a) LLM, participants read LLM-generated responses to online queries on OUD (sourced from Reddit); (b) Human, participants read human-written responses (again, sourced from Reddit) to the same set of queries as the LLM intervention group; and (c) Control, participants were not provided any content to read. We hypothesized that the LLM intervention would reduce stigmatized attitudes, across all three DVs, more effectively than the no intervention (i.e., Control) and Human intervention conditions. This hypothesis is informed by a recent work[34], which found that, compared to human-written responses, LLM-generated ones were less likely to promote leading myths surrounding patient characteristics and treatment models for OUD, and were more likely to counter them.

The experiment was conducted under two setups: (a) single exposure ($N = 2141$); where participants interacted with the intervention, i.e., read the relevant responses, once, and (b) longitudinal exposure ($N = 107$), where participants interacted with the intervention daily for 14 consecutive days, with the content varying each day. The single exposure setup was inspired by Schleider et al.'s[36] conceptualization of single-session interventions, which are structured programs involving only one encounter with a clinic, provider, or program. Prior work supports the utility of such short-term interventions for promoting behavior change, particularly in the context of anxiety, depression, and substance use[37]. Moreover, the behavioral psychology literature suggests that repeated, longitudinal exposure can further enhance attitude change[38]. By evaluating interventions under these two setups, the paper seeks to gain insights into the differential impacts and capabilities of brief, one-time, and prolonged, repeated interventions, providing valuable guidance for real-world deployments.

Our hypothesis was strongly supported for DV1, under both single and longitudinal exposure setups. Specifically, the LLM intervention was the most effective at reducing stigmatized attitudes toward MOUD compared to the Human and Control interventions. That said, certain interventions had a backfire effect[39]. A single exposure to human-written responses led to worsened attitudes across all three DVs, even when compared to the no-intervention Control. Lastly, the interventions did not affect all participants equally. They had a detrimental impact on participants with highly approving pre-intervention attitudes, but were effective in reducing stigma among those with more stigmatizing pre-intervention attitudes.

Results

Findings of hypotheses testing

H1(a): LLM intervention would reduce stigmatized attitudes toward MOUD (DV1) to a greater extent compared to the no intervention Control. Our study revealed that H1(a) was supported for both the single and longitudinal exposure setups, with statistical significance (Tables 1 and 2). Compared to the no intervention condition, i.e., Control, participants in the LLM intervention reported a more approving change in attitudes, or reduced stigma, toward MOUD (Fig. 1), for both the single (0.248 vs. 0.113, p: 2.64×10^{-9}) and longitudinal (0.435 vs. 0.092, p: 0.00269) exposure setups.

H1(b): LLM intervention would reduce stigmatized attitudes toward MOUD (DV1) to a greater extent compared to the Human intervention. H1(b) was supported for both the single and longitudinal exposure setups, with statistical significance (Tables 1 and 2). Compared to the Human intervention, participants in the LLM condition reported a more approving change in attitudes (Fig. 1) toward MOUD, for both the single (0.248 vs. 0.008, $p < 2.00 \times 10^{-16}$) and longitudinal (0.435 vs. 0.048, $p < 0.00136$) exposure setups.

H2(a): LLM intervention would reduce stigmatized attitudes toward people with OUD (DV2) to a greater extent compared to the no-intervention Control. Although not statistically significant, compared to the Control, participants in the LLM intervention reported a more approving change in attitudes toward people with OUD (Fig. 1, Tables 1 and 2), for both the single (0.058 vs. 0.022, p: 0.488) and longitudinal (−0.247 vs. −0.419, p: 0.131) exposure setups.

H2(b): LLM intervention would reduce stigmatized attitudes toward people with OUD (DV2) to a greater extent compared to the Human intervention. Compared to the Human intervention, participants in the LLM intervention reported significantly more approving change in attitudes toward people with OUD in the single exposure setup (0.058 vs. −0.112, p: 3.40×10^{-12}). In the longitudinal exposure setup, although participants in the LLM intervention showed a numerically greater change, it was not statistically significant (−0.247 vs. −0.303, p: 0.633). Thus, H2(b) was only supported in the single exposure setup (Fig. 1, Tables 1 and 2).

H3(a): LLM intervention would reduce stigmatized attitudes toward OUD (DV3) to a greater extent compared to the no intervention Control. Single exposure to the LLM intervention increased stigmatizing attitudes toward OUD when compared to the Control condition (−0.229 vs. −0.164, p: 0.005; Fig. 1a), based on an aggregate across all six statements. Thus, H3(a) was rejected, with statistical significance, in the single exposure setup (Table 1). At the statement level, this trend was numerically supported across a majority of the six statements (Table 1), particularly for those reflecting fatalistic views (i.e., *full recovery from opioid addiction is not possible*; −0.584 vs. −0.303, p: 7.46×10^{-6}) and individualistic causal attributions (i.e., *moral strength plays a large part in the cause of OUD*; −0.077 vs. −0.065, p: 0.817).

On aggregate, in the longitudinal exposure setup, LLM intervention led to more improvements, though non-significant, in participant attitudes toward OUD, in comparison to the Control (0.193 vs. −0.022, p: 0.055; Fig. 1b). Again, this trend was numerically supported across five of the six statements used to get participants' attitudes toward OUD (Table 2).

H3(b): LLM intervention would reduce stigmatized attitudes toward OUD (DV3) to a greater extent compared to the Human intervention. Compared to the Human intervention, participants in the LLM intervention reported significantly more approving change in attitudes toward OUD in the single exposure setup (−0.229 vs. −0.437, $p < 2.00 \times 10^{-16}$), based on an aggregate score across all six statements. Thus, H3(b) was supported, with statistical significance, in the single exposure setup (Table 1). At the statement level, this trend was numerically supported for five of the six items, with statistically significant differences observed for three (Table 1).

On aggregate, in the longitudinal exposure setup, the LLM intervention led to more improvements, though non-significant, in attitudes toward OUD (0.193 vs. 0.058, p: 0.247). This was also observable for three of the six individual statements, with statistical significance for "*full recovery from opioid addiction is not possible*" (0.181 vs. −0.451, p: 0.035).

Varied impact on post-intervention outcomes by dependent variable, exposure setup, and intervention type. Among the three DVs, the LLM intervention was the most successful at reducing stigmatizing attitudes toward MOUD, i.e., DV1 (Fig. 1). On average, after a single exposure to LLM-generated responses, participants reported an improvement of 6.61% in their perceptions toward MOUD. This approving change in attitudes was even more pronounced in the longitudinal exposure setup (13.64%).

While the longitudinal exposure setup outperformed the single exposure setup in improving participant attitudes toward MOUD and OUD, as seen post both Human (DV1: 0.008 (single) vs. 0.048 (longitudinal); DV3: −0.437 vs. 0.058) and LLM (DV1: 0.248 vs. 0.435; DV3:

Table 1 | Between-condition analysis for the single exposure setup

DV	Control	Human	LLM	*t*, *p*	*d*
DV1 (A)	0.113 (0.047)	0.008 (0.022)	0.248 (0.022)	H < C (−4.713, 2.61×10^{-6}***)	H < C (−0.264)
				L > C (5.979, 2.64×10^{-9}***)	L > C (0.318)
				L > H (10.392, <2.00×10^{-16}***)	L > H (0.502)
DV2 (A)	0.022 (0.042)	−0.112 (0.020)	0.058 (0.020)	H < C (−6.552, 7.17×10^{-11}***)	H < C (−0.345)
				L > C (0.694, 0.488)	L > C (0.023)
				L > H (7.002, 3.40×10^{-12}***)	L > H (0.340)
DV3 (A)	−0.164 (0.049)	−0.437 (0.023)	−0.229 (0.023)	H < C (−11.797, <2.00×10^{-16}***)	H < C (−0.619)
				L < C (−2.780, 0.005 **)	L < C (−0.145)
				L > H (8.674, <2.00×10^{-16}***)	L > H (0.438)
DV3 (S1)	−0.303 (0.127)	−1.141 (0.062)	−0.584 (0.063)	H < C (−13.541, <2.00×10^{-16}***)	H < C (−0.703)
				L < C (−4.491, 7.46×10^{-6}***)	L < C (−0.250)
				L > H (8.710, <2.00×10^{-16}***)	L > H (0.437)
DV3 (S2)	−0.065 (0.107)	−0.251 (0.052)	−0.077 (0.052)	H < C (−3.578, <3.54×10^{-4}***)	H < C (−0.191)
				L < C (−0.232, 0.817)	L < C (−0.014)
				L > H (3.233, <1.25×10^{-3}**)	L > H (0.170)
DV3 (S3)	−0.040 (0.095)	−0.019 (0.046)	−0.034 (0.046)	H > C (0.471, 0.637)	H > C (0.024)
				L > C (0.132, 0.895)	L > C (0.004)
				L < H (−0.327, 0.744)	L < H (−0.019)
DV3 (S4)	−0.257 (0.116)	−0.715 (0.055)	−0.254 (0.005)	H < C (−8.289, <2.00×10^{-16}***)	H < C (−0.443)
				L > C (0.049, 0.961)	L > C (0.010)
				L > H (8.057, 1.31×10^{-15}***)	L > H (0.412)
DV3 (S5)	−0.189 (0.083)	−0.251 (0.040)	−0.213 (0.040)	H < C (−1.534, 0.125)	H < C (−0.079)
				L < C (−0.592, 0.554)	L < C (−0.031)
				L > H (0.906, 0.365)	L > H (0.045)
DV3 (S6)	−0.200 (0.089)	−0.318 (0.042)	−0.289 (0.042)	H < C (−2.824, 0.005**)	H < C (−0.149)
				L < C (−2.105, 0.035*)	L < C (−0.107)
				L > H (0.671, 0.502)	L > H (0.035)

We report the linear mixed-effects model-estimated means (and standard errors) for the raw change in attitudes post-intervention, i.e., $\delta Y = Y_{post} - Y_{pre}$; Y_{post} and Y_{pre} represent the aggregated score post and pre-intervention, respectively. For the three DVs, we report the change in attitudes aggregated (A) across all the survey statements used to measure them. For DV3, we also report statement-level (S) change in attitudes. The two columns on the right summarize pairwise difference analysis of the three intervention conditions: Control (C), Human (H), and LLM (L). Pairs of conditions with statistically significant differences (*p*) under *t*-tests are marked as * ($p < 0.05$), ** ($p < 0.01$), or *** ($p < 0.001$). *d* represents the Cohen's *d* or the effect size measurement.

−0.229 vs. 0.193) interventions (Tables 1 and 2), it was not always better. For DV2, the longitudinal exposure actually worsened stigmatizing attitudes toward people with OUD compared to the single exposure setup, post both Human (−0.112 vs. −0.303) and LLM (0.058 vs. −0.247) interventions.

Interventions can sometimes have a backfire effect[39]. In our case, certain interventions amplified stigmatizing attitudes, post-intervention, in comparison to the no intervention (i.e., Control) baseline (Fig. 1a). Specifically, a single exposure to the Human intervention increased stigmatized perceptions by 0.99%, 4.73%, and 11.58% toward MOUD, people with OUD, and OUD, respectively, which was worse compared to the change in attitudes after no intervention at all (stigmatized perceptions decreased by 1.82% toward MOUD, and increased by 0.28% and 1.26% toward people with OUD and OUD, respectively). Although we observed a change in attitudes for participants in the Control condition, this change was not statistically significant under Kruskal–Wallis *H*-tests (Table S18).

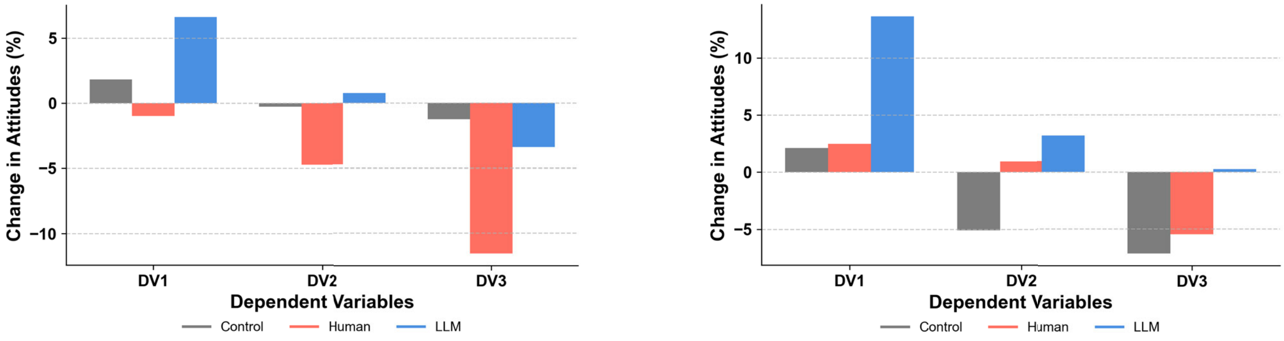
Analysis of the intervention content

In the context of behavioral health conditions, Link and Phelan highlight the potential of simple interventions that elicit hope, optimism, and a shared sense of belonging in reducing stigma[2]. Consequently, we examined the intervention content, read by participants in the LLM and Human interventions, across linguistic dimensions such as readability, emotional appeal, and in-group affinity to provide insights into the rationale behind the effectiveness of the various interventions reported above. We used relevant categories from Empath[40], a lexicon-based tool, to assess emotional appeal (Fig. 2a, b). For both the single and longitudinal exposure setups, the LLM intervention content contained significantly more optimistic ("optimism" category) and supportive ("help" category) linguistic cues compared to the Human intervention content, which are indicative of encouraging, hopeful, and positive discourse. Next, though not significantly, the LLM intervention content contained a greater shared sense of belonging (Fig. 2d), which was quantified using the identity social dimension classifier[41] widely used to determine peer support capabilities of online communities. Again, this provides empirical evidence suggesting that the LLM intervention contained more inclusive linguistic cues, potentially helpful to abate stigma[2]. Finally, we used the Linguistic Inquiry and Word Count (LIWC) tool[42], a psycholinguistic lexicon, to compare the distribution of first- and third-person pronoun usage. While the Human intervention content contained significantly more first-person pronouns (under Wilcoxon signed-rank tests; $W = 267.0$; $p < 0.001$), there were no significant differences in third-person pronoun usage ($W = 1667.5$; $p > 0.05$). This suggests that neutrality within the LLM intervention may have contributed to its effectiveness in reducing stigma. It aligns with prior research showing that stigma can be reduced through objective communication strategies, particularly in public and digital health contexts[43,44]. Post-intervention, participants in the LLM and Human intervention groups self-reported their attitudes toward the intervention content (Table S8), i.e., whether they found the responses to be influential; *offered a different approach to look at OUD*, credible; *were reasonable and trustworthy*, informative; *were knowledgeable*, resourceful; *likely*

Table 2 | Between-condition analysis for the longitudinal exposure setup

DV	Control	Human	LLM	*t*, *p*	*d*
DV1 (A)	0.092 (0.047)	0.048 (0.115)	0.435 (0.111)	H < C (−0.386, 0.700)	H < C (−0.014)
				L > C (3.082, 0.00269**)	L > C (0.908)
				L > H (3.301, <0.00136**)	L > H (0.732)
DV2 (A)	−0.419 (0.234)	−0.303 (0.116)	−0.247 (0.113)	H > C (0.999, 0.320)	H > C (0.256)
				L > C (1.522, 0.131)	L > C (0.377)
				L > H (0.479, 0.633)	L > H (0.149)
DV3 (A)	−0.022 (0.229)	0.058 (0.113)	0.193 (0.111)	H > C (0.706, 0.482)	H > C (0.039)
				L > C (1.941, 0.055)	L > C (0.395)
				L > H (1.166, 0.247)	L > H (0.307)
DV3 (S1)	−0.306 (0.602)	−0.451 (0.294)	0.181 (0.291)	H < C (−0.492, 0.624)	H < C (−0.151)
				L > C (1.674, 0.097)	L > C (0.290)
				L > H (2.095, 0.035*)	L > H (0.421)
DV3 (S2)	0.563 (0.497)	0.489 (0.247)	0.786 (0.241)	H < C (−0.298, 0.766)	H < C (−0.129)
				L > C (0.926, 0.357)	L > C (0.152)
				L > H (1.176, 0.242)	L > H (0.314)
DV3 (S3)	0.144 (0.405)	0.077 (0.194)	−0.07 (0.194)	H < C (−0.348, 0.729)	H < C (−0.099)
				L < C (−1.101, 0.274)	L < C (−0.202)
				L < H (−0.736, 0.464)	L < H (−0.134)
DV3 (S4)	−0.266 (0.577)	−0.168 (0.291)	0.258 (0.281)	H > C (0.336, 0.737)	H > C (0.010)
				L > C (1.859, 0.066)	L > C (0.424)
				L > H (1.434, 0.155)	L > H (0.403)
DV3 (S5)	0.046 (0.385)	0.376 (0.194)	0.298 (0.188)	H > C (1.699, 0.093)	H > C (0.371)
				L > C (1.337, 0.184)	L > C (0.285)
				L < H (−0.394, 0.694)	L < H (−0.070)
DV3 (S6)	−0.394 (0.393)	−0.067 (0.198)	−0.357 (0.192)	H > C (1.646, 0.103)	H > C (0.309)
				L > C (0.193, 0.847)	L > C (0.017)
				L < H (−1.427, 0.157)	L < H (−0.299)

We report the linear mixed-effects model-estimated means (and standard errors) for the raw change in attitudes post-intervention, i.e., δY. For the three DVs, we report the change in attitudes aggregated (A) across all the survey statements used to measure them. For DV3, we also report statement-level (S) change in attitudes. The two columns on the right summarize pairwise difference analysis of the three intervention conditions: Control (C), Human (H), and LLM (L). Pairs of conditions with statistically significant differences (*p*) under *t*-tests are marked as * ($p < 0.05$) or ** ($p < 0.01$). *d* represents the Cohen's *d* or the effect size measurement.

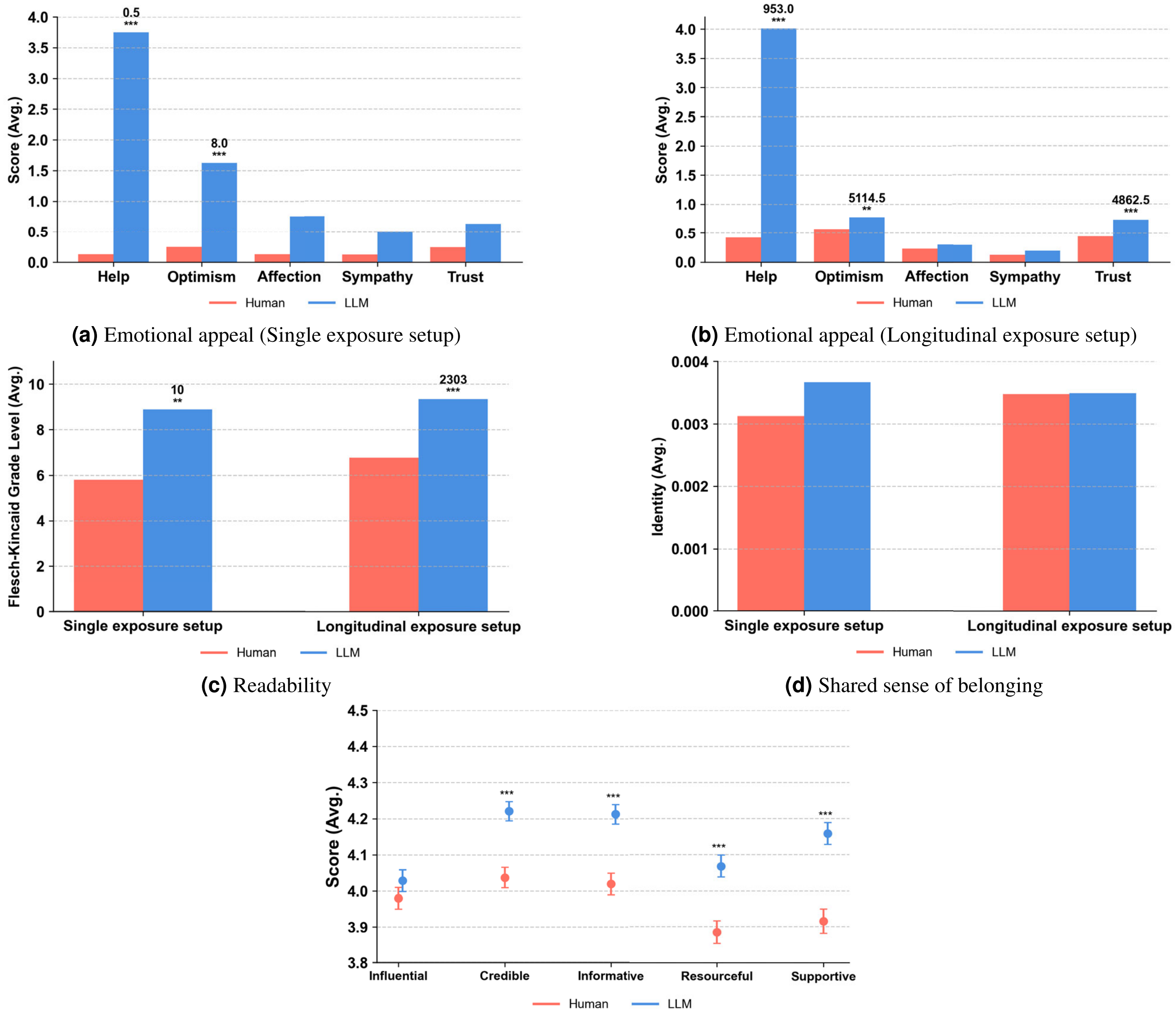


(a) Single exposure setup

(b) Longitudinal exposure setup

Fig. 1 | Raw percentage change in attitudes. Raw percentage change in attitudes toward MOUD (DV1), people with OUD (DV2), and OUD (DV3), averaged across participants, for the Control, Human, and LLM interventions after **a** single and **b** longitudinal exposure setups. Percentage change in attitudes was computed as $\frac{(Y_{post}-Y_{pre})}{Y_{pre}} \times 100$; where Y_{post} and Y_{pre} represent the aggregated DV score post and pre-intervention.

to refer to such responses to gain information about OUD, and supportive; *prefer to receive such responses if one had OUD*. As a post-hoc analysis, on finding no significant differences in ratings across single and longitudinal exposure setups (after applying Bonferroni correction), we combined participant scores for the two setups using a weighted average measure (weighted by the sample size; Fig. 2e). Participants in the LLM intervention rated the LLM-generated responses as more influential, and significantly more credible, informative, resourceful, and supportive compared to how participants in the Human intervention rated the human-written responses.



(a) Emotional appeal (Single exposure setup)

(b) Emotional appeal (Longitudinal exposure setup)

(c) Readability

(d) Shared sense of belonging

(e) Participant attitudes

Fig. 2 | Analysis of the intervention content. Responses read by participants within the LLM and Human intervention groups were evaluated for **a**, **b** emotional appeal, **c** readability, and **d** shared sense of belonging. Emotional appeal is reported using five relevant categories available in Empath[40], a lexicon-based tool; a higher score is indicative of a higher alignment to the category. Readability is reported using the Flesch–Kincaid Grade Level index[80]; a lower score is indicative of simpler, more readable text. Shared sense of belonging is reported using the identity social dimension classifier[41], which quantifies in-group or community forward linguistic cues; the higher the score, the better. Scores are averaged across all the responses read by participants during the single and longitudinal exposure setups. Mann–Whitney U-tests were performed to explore differences in score distributions for responses provided in the LLM and Human interventions. Statistically significant differences are noted with the test statistic and p values (p): * ($p < 0.05$), ** ($p < 0.01$), or *** ($p < 0.001$). **e** At the end of our experiments, participants in the LLM and Human interventions rated the responses consumed during the respective interventions, using a 5-point Likert scale, as: (1) influential, *responses offered a different approach to look at OUD*; (2) credible, *responses were reasonable and trustworthy*; (3) informative, *responses were knowledgeable*; (4) resourceful, *likely to refer to such responses to gain information about OUD*; and (5) supportive, *prefer to receive such responses if one had OUD*. On finding no significant differences in ratings across single and longitudinal exposure setups, we combined participant ratings for the two setups and report a weighted average (weighted by the sample size). Mann–Whitney *U* tests were performed to examine differences in score distributions for ratings provided by participants in the LLM and Human intervention groups. Statistically significant differences are noted with the p values (p): * ($p < 0.05$), ** ($p < 0.01$), or *** ($p < 0.001$).

Participant perspectives on intervention content

Next, we explored thematic patterns within participants' text-based post-intervention reflections, which they provided daily during the longitudinal exposure setup. The reflections were participants' responses to open-ended questions such as "Based on your reading, is Methadone the best route to recovery?" and "Is full recovery possible?" We adopted a mixed-methods approach to manually annotate themes within a handful of reflections, which were then used as ground truth to train LLM-based classifiers for machine-annotating the rest (refer to S10 for details). The two most prominent themes relevant to DV1 in participants' reflections were "Dependence", i.e., MOUD replaces one drug with another, and "Effectiveness," i.e., MOUD is an effective treatment for OUD. Recurring themes related to DV2 were "Blame," meaning people with OUD are responsible for their own condition, and "Labeling", referring to the use of derogatory terms like "addict" and "junkies." For DV3, the two most prominent themes were "Fatality" (i.e., full recovery from OUD is not possible) and its contrast, "Belief in Recovery." Refer to Tables S14–S16 for a complete list of themes.

Per DV1, more reflections from participants in the LLM intervention ($n = 53/102$) supported the effectiveness of MOUD compared to those from participants in the Human intervention ($n = 46/102$). Moreover, fewer reflections from participants in the LLM intervention highlighted dependence on MOUD ($n = 38/102$). Using a generalized linear mixed model (GLMM), we found that participants in the Human intervention were significantly less likely to mention the approving themes in our codebook, related to DV1, in comparison to those in the LLM intervention ($\beta = -0.859$, $p = 0.048$). However, they were more likely to contain the stigmatizing themes, though the differences did not reach statistical significance ($\beta = 0.314$, $p = 0.714$). Though the two most recurring themes for DV2 highlight negative attitudes toward people with OUD, the support for them was a little less pronounced among participants' reflections in the LLM intervention compared to those in the Human intervention ("Blame": $n = 30/136$ (LLM) vs. $n = 31/136$ (Human); "Labeling": 37/136 vs. 50/136). A GLMM further showed that participants in the Human intervention were more likely to mention the above two negative themes ($\beta = 0.359$, $p = 0.697$), and less likely to mention the two approving themes ($\beta = -0.451$, $p = 0.729$) related to DV2 surfaced in our codebook. However, the differences were not statistically significant. In the LLM intervention, many more reflections emphasized the belief in recovery ($n = 80/102$) compared to those rejecting it ($n = 35/102$). Again, this positive support was more evident in reflections from participants in the LLM intervention compared to those in the Human intervention. Only 69/102 reflections from participants in the Human intervention supported the belief in recovery, while 46/102 dismissed it. Lastly, a GLMM revealed that participants in the Human intervention were significantly less likely to mention the three approving themes related to DV3 ($\beta = -0.478$, $p = 0.037$), while they were more likely to mention the two stigmatizing themes ($\beta = 0.365$, $p = 0.184$), although this difference was not statistically significant.

Post-hoc analyses

Impact of pre-intervention attitudes on post-intervention outcomes. For participants within the LLM condition, we further examined whether the intervention impacted them all similarly (Fig. 3). To do so, we divided the participants into three groups, low: [1, 2.33), medium: [2.33, 3.66), and high: [3.66, 5], based on their pre-intervention attitudes toward each DV. Participants in the low pre-intervention attitude category were more likely to report a boost in their post-intervention attitudes. This was consistent across both single (DV1: $N = 14$, paired pre-/post- T-test statistic (t-stat): -2.73, p: 0.017; DV2: $N = 133$, t-stat: -4.64, p: 8.27×10^{-6}; DV3: $N = 100$, t-stat: -4.19, p: 6.17×10^{-5}) and longitudinal (DV2: $N = 8$, t-stat: -1.94, p: 0.09; DV3: $N = 2$, t-stat: -4.00, p: 0.16) exposure setups. Participants in the high pre-intervention score category reported the opposite; they were more likely to exhibit a decrease in their attitudes post-intervention—again, for both single (DV1: $N = 171$, t-stat: 3.55, p: 0.0005; DV2: $N = 51$, t-stat: 5.63, p: 8.27×10^{-7}; DV3: $N = 31$, t-stat: 5.98, p: 1.46×10^{-6}) and longitudinal (DV1: $N = 5$, t-stat: -2.45, p: 0.07; DV2: $N = 2$, t-stat: 1.00, p: 0.49) exposure setups, with one exception. This suggests that participants with stigmatized pre-intervention attitudes are more likely to adopt approving perspectives following the intervention. Contrastively, the intervention can be detrimental for participants with already highly approving pre-intervention attitudes. This remained unchanged when we accounted for demographics such as age, gender, education level, and political leaning (Table S12). The finding also holds true for participants enrolled in the Human intervention (Fig. S3, Table S13).

Impact of external information consumption on post-intervention outcomes. We performed a secondary analysis to examine whether participants' external consumption of OUD-related information impacted their post-intervention attitudes. To assess this, we added an additional fixed effect to our original linear mixed-effects model (S9), which captured participants' self-reported response to whether or not they consumed information on OUD beyond our provided interventions (S11). This external consumption did not have a statistically significant impact on post-intervention change in attitudes across all three DVs and the two exposure setups (Table S19).

Discussion

Derived from the communication literature, framing selects particular aspects of an issue and makes them salient in a communicating text[45]. Consumption of publicly accessible information can impact how people understand issues, attribute responsibility and blame, and endorse possible solutions, thus having major implications for shaping public perception and policy decisions[46]. Our work demonstrates that this is indeed true—engagements with LLM-generated and human-written responses to online queries on OUD influenced participants' perceptions of OUD-related stigma. Notably, perceptions varied by the type of intervention (i.e., whether participants read LLM- or human-written responses), the dependent variable (i.e., attitudes toward MOUD, people with OUD, or OUD), the duration of the intervention (i.e., single vs. longitudinal exposure), and participants' pre-intervention attitudes (Fig. 1, Tables 1 and 2).

As highlighted in our findings, our hypothesis was strongly supported for DV1 under both single and longitudinal exposure setups. Specifically, the LLM intervention was the most effective in reducing stigmatizing attitudes toward MOUD compared to the Human and Control interventions (Tables 1 and 2). Livingston et al. observed that education programs, targeting the general public, clinicians, and professionals, were effective at reducing negative attitudes toward clinically approved treatments, including MOUD[47]. Along similar lines, LLMs can be carefully integrated into online communities, as an education-based intervention, to foster positive attitudes and increase people's propensity toward MOUD. This further stems from our LIWC analysis, which revealed that the LLM-generated responses adopted a more objective, balanced tone—an approach that can be used to reduce stigma by providing a neutral point of view[43,44]. In making this recommendation, we are not advocating for the replacement of human-written responses but rather proposing that LLMs complement the existing online ecosystem by addressing users' MOUD-related queries.

Per our experiments, we found that for some forms of stigmatizing attitudes (DV3), a single exposure to both LLM and Human interventions worsened participants' attitudes (Table 1, Fig. 1a). This finding aligns with prior work, which shows that single, one-shot or brief interventions to correct misperceptions or misinformation can sometimes have a backfire effect[48,49]—an attempt to change someone's beliefs for the better can unintentionally lead to the opposite effect[39,50]. Tully et al.[51] further note that, in the context of online communities, one-shot interventions are insufficient to change attitudes, and sustained, repeated interventions are more effective. As described earlier, this was indeed true for DV3; longitudinal exposure outperformed the single exposure setup in improving participants' attitudes, post both LLM and Human interventions (Tables 1 and 2, Fig. 1). On the other hand, for other attitudes (DV1), a single exposure to the LLM intervention resulted in improvements, while the Human intervention remained detrimental (Table 1, Fig. 1a). These findings imply that future intervention policies should consider not only the extent of exposure (i.e., single or longitudinal) but also the medium used to generate the intervention content. As suggested by our analysis of the intervention content, a single exposure to the LLM intervention may be more effective, as it contained significantly more optimistic, supportive, and inclusive linguistic cues compared to the Human intervention (Fig. 2)—Link and Phelan[2] note the efficacy of optimism, hope, and shared sense of belonging in offsetting stigmatizing attitudes. Each day, following the intervention, participants were asked to recall or comment on the content they were exposed to by responding to a few open-ended questions. These reflections may themselves have contributed to the change in participant attitudes, beyond the effects of consuming the human-written and LLM-generated responses.

Our post-hoc analysis observed that participants with low pre-intervention (i.e., more stigmatizing) attitudes reported a shift toward more approving perspectives following exposure to both Human and LLM interventions. In contrast, those with high pre-intervention attitudes reported a decrease in approval post-intervention. This suggests that

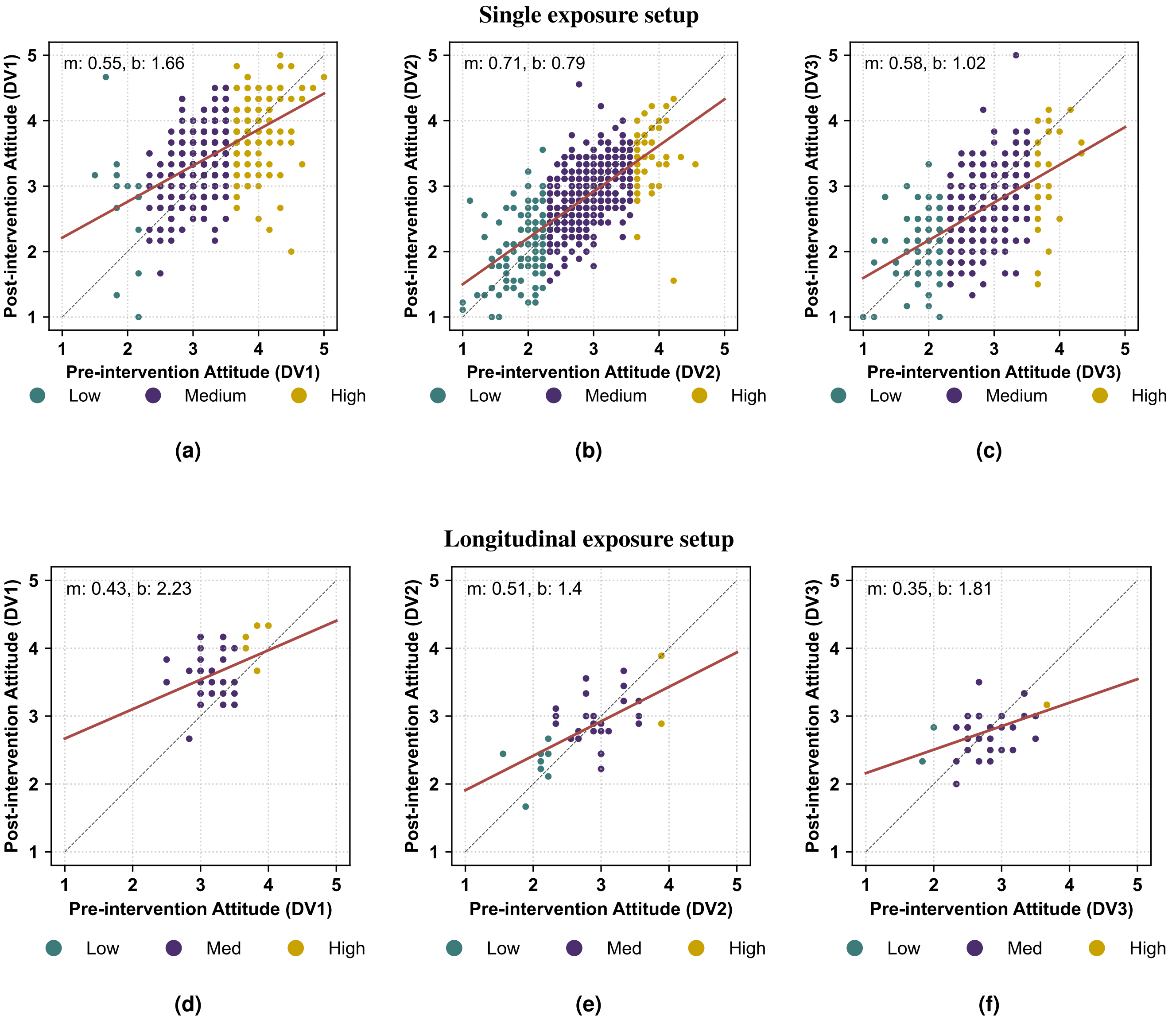


Fig. 3 | Pre- and post-intervention attitudes toward the three DVs for participants within the LLM intervention. a–**c** Single exposure setup; **d**–**f** longitudinal exposure setup. Participants were divided into three groups—low, medium, and high—based on their pre-intervention attitudes. The Likert scale range (1–5) was equally divided into three parts to achieve this. The red lines indicate the regression line or the best linear fit for the pre-/post-intervention score distributions (m: slope, b: intercept). The gray dotted line represents the no-change post-intervention fit, i.e., $Y_{pre} = Y_{post}$. Participants (represented via dots) above (below) the gray line reported a higher (lower) attitude post-intervention. Participants in the low (high) pre-intervention score category were more concentrated in the region above (below) the no-change post-intervention line fit.

participants may have critically reflected on their prior beliefs and adjusted their attitudes accordingly. Moreover, these findings underscore the importance of audience segmentation when designing such interventions, for example, by excluding participants with high (i.e., more approving) baseline attitudes, for whom the interventions may backfire.

Research shows that alleviating stigmatizing attitudes toward people with behavioral health conditions, including OUD, is challenging[52]. In fact, scholars argue that the most effective interventions for addressing these attitudes are those that increase contact between individuals with the condition and the broader population, as such attitudes are deeply ingrained[52,53]. Our interventions did not involve any direct contact, and in certain settings indeed amplified stigmatizing attitudes toward people with OUD (i.e., DV2). Per our findings, post both LLM and Human interventions, a longitudinal exposure worsened stigmatizing attitudes toward people with OUD compared to a single exposure (Tables 1 and 2). We believe that prolonged or persistent engagement with any type of online content, whether LLM-generated or human-written, could potentially reinforce participants' deep-seated stigmatizing attitudes due to psychological reactance[54]; a cognitive response where individuals may react by strengthening their stigmatizing attitudes toward people with OUD when exposed to information that attempts to change their perceptions for the better[52,54].

Finally, we note some limitations that provide potential directions for future work. First, while our dependent variables encoded three dimensions of stigma—towards the clinically approved treatments, people with the condition, and the condition itself—prior work has identified additional dimensions, such as self-stigma[55], structural stigma[25], and stigma shaped by intersecting identities such as gender, race, socioeconomic status, and parenting roles[56–59]. Future studies should investigate changes in participants' attitudes toward these additional dimensions of stigma post-engagement.

Cronbach's alpha across the six statements used to measure DV1 was 0.61. Though moderately acceptable, it falls on the lower end of the reliability range[60]. This could be attributed to the relatively fewer number of items on the scale[61]. For DV2, statements intended to assess stigmatizing attitudes related to the functional impairment and mental health of people with OUD may appear to capture factual knowledge instead. As noted in prior work, it is difficult to disentangle the two. Incorrect and even limited factual knowledge can often shape stigma-related attitudes[62]. Therefore, we asked participants to rate the statements based on their own opinion, both before and after the intervention, regardless of whether they were drawing on factual or incorrect information. We could not find an existing scale to measure attitudes toward OUD. Therefore, we referred to relevant literature to get measurement statements and established content validity via pilots. The purpose of this study was not to develop a psychometric scale, but to evaluate the impact of an intervention. Therefore, as suggested[63], we prioritized content validity over psychometric analysis. Internal consistency across the statements, or Cronbach's alpha (0.5), was below the acceptable range. As a result, we report aggregated and statement-level changes in attitudes. Findings across the two were largely consistent. Future scholars should formally develop and validate a comprehensive scale for measuring attitudes toward OUD.

Next, there may be nuanced variations across cultures and backgrounds in how people perceive information, in general, or on OUD. In our experiment, though we considered several demographic attributes, we recruited participants exclusively residing within the U.S., and from a single crowd-sourcing platform (i.e., Prolific). Expanding the scope of the experiment by including other cultures warrants future investigation. To approximate user interactions on Reddit, we considered the top-most voted comment for the Human intervention. However, we acknowledge that in real-world scenarios, users may go through multiple comments within a thread. Future work should explore the impact of consuming threads or multiple exchanges on participant attitudes. Future work should further inquire the effectiveness of the LLM intervention, e.g., by considering the impact of linguistic differences on attitudes. In this work, we explored the impact of both single and longitudinal exposures to the interventions. Due to resource constraints and participant attrition, we limited the duration of the longitudinal exposure to 14 days. Future work can explore the impact of extended durations of the intervention. Through our work, we examined the potential of LLMs as an intervention to reduce stigmatizing attitudes toward OUD. Though promising, as our findings suggest, LLMs can cause informational[64] and representational[65] harms. When using these as interventions in real-world contexts, a careful expert-driven review, e.g., through content moderators, is needed to minimize potential harms.

Methods

Study design

We adopted a between-subjects experimental setup, in which participants were randomly assigned to one of the LLM, Human, or Control interventions. Participants were asked to read either LLM- (LLM intervention) or human-written (Human intervention) responses to OUD-specific queries. In contrast, participants in the Control group were not provided any content to read. We used a deception-based framework[66], i.e., the participants were not informed if they were reading LLM- or human-written responses to prevent any intended or unintended bias. The experiment was conducted under two setups—Study (a): single exposure setup, where participants interacted with the intervention, i.e., read the relevant responses, once, and Study (b): longitudinal exposure setup, where participants interacted with the intervention daily for 14 consecutive days, with the content varying each day. To ensure consistency, responses provided during the single exposure setup were exactly the same as those provided on the first day of the longitudinal exposure setup. Both studies were approved by the Institutional Review Board of the first author's University and pre-registered (https://osf.io/m8hc2, https://osf.io/f7e4r).

Participant recruitment. We recruited participants through Prolific, a crowd-sourcing platform extensively used in computational social science[67,68]. Our recruitment criteria pre-screened participants to be above 18 (age), located within the United States, with at least 20 completed Prolific submissions (suggested for longitudinal experiments), and with prior experience of using online platforms for healthcare-related information seeking. We enrolled 2400 and 150 participants (S1) for Study (a) and Study (b), respectively. Refer to S8 for participant demographics.

Intervention (Human). Owing to pseudonymity, Reddit enables candid discussions on OUD[13]. Reddit's popularity among people with OUD[11] and the availability of long-form content encouraged us to consider (a) queries on OUD, mentioned within original posts, and (b) corresponding human-written responses, i.e, comments, posted on Reddit to design our Human intervention. Specifically, we used a Reddit Question-Answer dataset made available in prior research[69], hereafter referred to as `Reddit-QA`. It consists of 150,436 posts with an OUD-related query, and the associated comments as a proxy for human-written responses. The posts/comments span across 19 OUD-related subreddits, e.g., `r/Methadone`, `r/OpiatesRecovery`, and `r/Heroin`, from January 2018 to September 2021.

The Human intervention (Fig. S1; Table S2) was designed to mirror users' real-world information consumption experiences on Reddit. Therefore, posts containing an OUD-related query were filtered based on engagement metrics, relevance, and topic diversity. Corresponding to these posts, we then considered the top-most voted comment as a representative human-written response, read by participants in the Human intervention (refer to S2 for more details).

Intervention (LLM). We used `GPT-4`[70] to get LLM-generated responses, read by participants in the LLM intervention (Fig. S1; Table S2), to the filtered `Reddit-QA` posts containing an OUD-related query. In particular, we used `gpt-4-0613` (with a 0.7 temperature, the default for conversational agent interfaces[70]), which was the most stable and capable text generation model available at the time of this work. Referring to well-adopted prompt engineering guidelines, we created a prompt (Table S1) to generate responses for the posts. To assist response generation, we provided (a) simple task-specific instructions, (b) context (an active Reddit user persona, post's subreddit name and description), (c) a question of interest, and (d) the output format (refer to S2 for more details).

Intervention dosage. We chose an empirically driven way to determine the number of responses read by participants during each exposure of the intervention. Through a pilot (refer to S3), we found that each intervention exposure should be limited to a dosage of 8 query-response pairs. This dosage level accounted for factors such as reader attention span and potential exposure to distressing narratives or misinformation. Therefore, we filtered a subset of $(8 \times 14) = 112$ posts from the `Reddit-QA` dataset, obtained the corresponding top-most voted comments (for the Human intervention), and the LLM-generated responses (for the LLM intervention). Refer to S2 for more details. We also decided the duration of our longitudinal exposure setup using the pilot. 14 days was observed as a feasible duration, resulting in a reasonable amount of engagement (accounting for attrition) from our participants.

Dependent variables. We measured three dependent variables (DVs): (1) attitudes toward MOUD (DV1), (2) attitudes toward people with OUD (DV2), and (3) attitudes toward OUD (DV3). Refer to S5 for a detailed description.

Participant attitudes were measured based on their aggregated response to 5-point Likert scale statements (Tables S4 and S7)—ranging from strong agreement (=1) to strong disagreement (=5). For DV3, we also report participants' statement-level change in attitudes. Higher scores were indicative of less stigmatizing attitudes. Statements relevant to each were adopted from prior work in public health, including surveys widely adopted to assess public, structural, and provider-based stigma[71], the Brief Opioid Stigma Scale[72], and the Attitudes Toward Methadone Questionnaire[73]. For DV1, statements measured participants' inclination toward *intervention-based* stigma, i.e., *MOUD is*

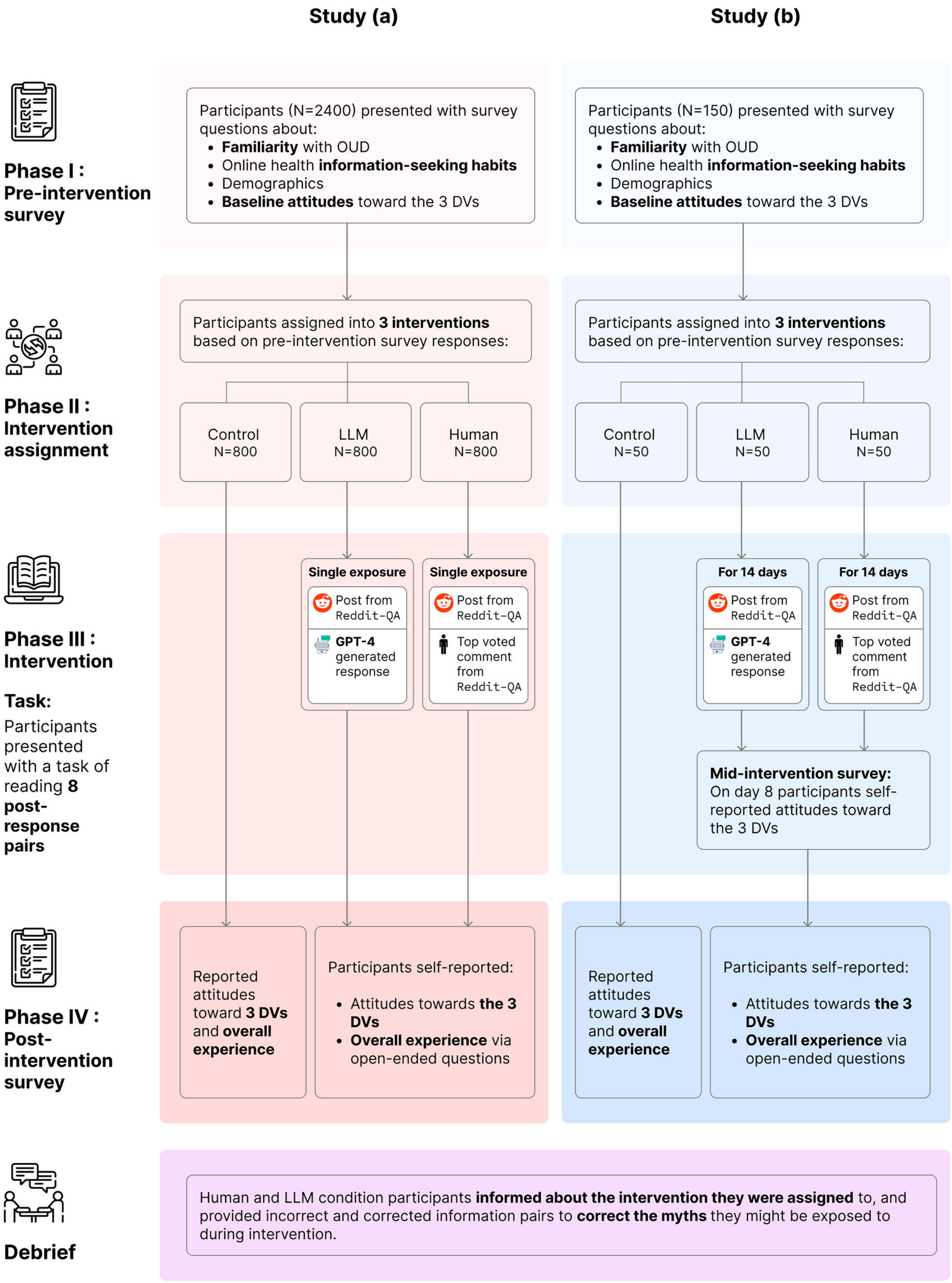

Study (a)
Study (b)
Phase I : Pre-intervention survey
Participants (N=2400) presented with survey questions about:
• Familiarity with OUD
• Online health information-seeking habits
• Demographics
• Baseline attitudes toward the 3 DVs
Participants (N=150) presented with survey questions about:
• Familiarity with OUD
• Online health information-seeking habits
• Demographics
• Baseline attitudes toward the 3 DVs
Phase II : Intervention assignment
Participants assigned into 3 interventions based on pre-intervention survey responses:
Control N=800
LLM N=800
Human N=800
Control N=50
LLM N=50
Human N=50
Phase III : Intervention
Task: Participants presented with a task of reading 8 post-response pairs
Single exposure
Post from Reddit-QA
GPT-4 generated response
Single exposure
Post from Reddit-QA
Top voted comment from Reddit-QA
For 14 days
Post from Reddit-QA
GPT-4 generated response
For 14 days
Post from Reddit-QA
Top voted comment from Reddit-QA
Mid-intervention survey: On day 8 participants self-reported attitudes toward the 3 DVs
Phase IV : Post-intervention survey
Reported attitudes toward 3 DVs and overall experience
Participants self-reported:
• Attitudes towards the 3 DVs
• Overall experience via open-ended questions
Debrief
Human and LLM condition participants informed about the intervention they were assigned to, and provided incorrect and corrected information pairs to correct the myths they might be exposed to during intervention.

Fig. 4 | Overview of our study workflow. Represents the four phases involved in Study (**a**): single exposure setup and Study (**b**): longitudinal exposure setup.

simply replacing one drug with another. For DV2, statements measured perceptions of *dangerousness*, *blame*, and *social distance*, amongst others, directed toward people with OUD, e.g., "It would bother me to live near a person who used opioids." For DV3, statements captured various facets surrounding stigmatizing attitudes toward the condition, such as perspectives on recovery potential and causes[74].

Hypotheses testing. We had three hypotheses, one for each of the three DVs: LLM intervention would reduce stigmatized attitudes toward MOUD (H1), people with OUD (H2), and OUD (H3). Within each hypothesis, we conducted two comparisons: LLM intervention would reduce stigmatized attitudes to a greater extent compared to the (a) Human and (b) Control interventions. To test these, we estimated participants' post-intervention change in attitudes toward the three DVs using a linear mixed-effects model (S9).

Study workflow. We divided the Studies (a) and (b) into the following phases (Fig. 4):

- *Phase I: pre-intervention survey*. All the 2400 and 150 consented participants for Studies (a) and (b), respectively, were asked the same set of survey questions describing their familiarity with OUD (Table S5), online health information seeking habits (Table S5), demographics (Table S6), and baseline or pre-intervention attitudes toward the three DVs (Table S7).
- *Phase II: intervention assignment*. For both the studies, participants were randomly assigned to the three intervention conditions, i.e., LLM, Human, and Control, via stratified randomization[75]. Participants were split into different blocks based on 5 confounders: (1) age, (2) gender, (3) familiarity with OUD, (4) reliance on online platforms for health information seeking, and (5) baseline attitudes toward the three DVs (aggregated measure). Within each block, participants were evenly divided across the three interventions. In case a block had an uneven split, the split was favored for one of the LLM or Human interventions randomly. After this randomization, we evaluated whether the three groups were comparable across other individual characteristics we collected during Phase I—political leaning, occupation, and education level. χ^2 tests found that none of the groups significantly associated with these additional characteristics.
- *Phase III: intervention*. Next, participants performed the main study procedure, a reading task, depending on their intervention group assignment. Participants in the LLM intervention read LLM-generated responses, those in the Human intervention read human-written responses, and those in the Control group were not provided any content to read. Intervention was provided via an interactive interface (S4) once for Study (a) and daily, for 14 days, for Study (b). In Study (b), participants, across all three intervention groups, completed a mid-intervention survey after receiving the intervention for 7 days. In this survey, the participants self-reported their attitudes toward the three DVs; Table S7. Completing this, they continued to receive the intervention for 7 more days.
- *Phase IV: post-intervention survey*. Finally, after the intervention was over, participants across both the studies and intervention groups completed a post-intervention survey. They self-reported their attitudes toward the three DVs; Table S7. Participants in the LLM and Human interventions also rated the responses they read as influential, credible, knowledgeable, resourceful, and supportive using a 5-point Likert scale (Table S8). All the participants answered a few open-ended questions describing their overall experience (Table S9).

Participants who dropped out before completing the studies were removed from our analyses. In total, 2141 (772 in Control, 696 in Human, and 673 in LLM) and 107 (39 in Control, 34 in Human, and 34 in LLM) participants completed Studies (a) and (b). Refer to S7 for details on participant attrition. We provided a compensation of 4 USD and 50 USD for Studies (a) and (b), respectively, via Prolific, prorated based on an hourly wage of 12 USD. Table S10 summarizes time taken by participants to complete the different phases of both studies.

Debrief. As our experiment was based on deception, after the completion of the studies, we informed the participants about the intervention they were assigned to, i.e., whether they read LLM- or human-generated responses. Following best practices[76,77], we also provided mis- and corrected information pairs, through our interactive interface, to correct the myths, misinformation, or misperceptions the participants were exposed to during the intervention. On average, participants across the LLM and Human intervention groups spent 11.12 and 8.23 min going through the debriefs for Study (a) and Study (b), respectively.

Safety, privacy, and ethical considerations

At the beginning, we obtained informed consent from our participants to take part in our IRB-approved experiments. Following best practices[78,79], we worked with de-identified publicly accessible data and refrained from sharing raw and personally identifiable data in any form. In our interventions, the posts and human-generated responses, taken from Reddit, were carefully paraphrased to reduce traceability. Additionally, throughout the intervention, we provided easy access to helpline numbers and resources (S4) as the responses could elicit distress on consumption. Participants were free to withdraw their consent or discontinue participation at any point with no negative consequences. Through Prolific, we were able to anonymously communicate with all our participants. No personally identifiable information was exchanged or used in our study. During the intervention, our participants were exposed to misperceptions and misinformation surrounding OUD. Following recommendations[76,77], at the end of the experiment, we held a debrief session, as described earlier, where misinformation and corresponding corrections were presented in an interactive manner.

Data availability

The online queries and human-written responses from Reddit are taken from an already published work. Due to a change in Reddit's data distribution policy, the Reddit dataset cannot be distributed publicly. The supplementary material provides a detailed description of how the authors obtained the LLM-generated responses (S2). The data collected during the randomized controlled experiments can be made available upon request to the corresponding author with justification.

Received: 31 May 2025; Accepted: 17 October 2025;
Published online: 01 December 2025

References

1. Garnett, M. F. & Miniño, A. M. Drug overdose deaths in the United States, 2003-2023. NCHS Data Brief, No. 522 (National Center for Health Statistics, 2024).
2. Link, B. G. & Phelan, J. C. Conceptualizing stigma. *Annu. Rev. Sociol.* **27**, 363–385 (2001).
3. Garett, R. & Young, S. D. The role of misinformation and stigma in opioid use disorder treatment uptake. *Subst. Use Misuse* **57**, 1332–1336 (2022).
4. Richard, E. L. et al. "You are not clean until you're not on anything": perceptions of medication-assisted treatment in rural Appalachia. *Int. J. Drug Policy* **85**, 102704 (2020).
5. Adams, Z. W. et al. Opioid use disorder stigma, discrimination, and policy attitudes in a national sample of U.S. young adults. *J. Adolesc. Health* **69**, 321–328 (2021).
6. Barry, C. L., McGinty, E. E., Pescosolido, B. A. & Goldman, H. H. Stigma, discrimination, treatment effectiveness, and policy: public views about drug addiction and mental illness. *Psychiatr. Serv.* **65**, 1269–1272 (2014).
7. Taylor, B. G. et al. Social stigma toward persons with opioid use disorder: Results from a nationally representative survey of U.S. adults. *Subst. Use Misuse* **56**, 1752–1764 (2021).
8. Corrigan, P. W., Kuwabara, S. A. & O'shaughnessy, J. The public stigma of mental illness and drug addiction: Findings from a stratified random sample. *J. Soc. Work* **9**, 139–147 (2009).

9. Woo, J. et al. "Don't judge a book by its cover": a qualitative study of methadone patients' experiences of stigma. *Subst. Abuse Res. Treat.* **11**, 1178221816685087(2017).
10. Madden, E. F. Intervention stigma: How medication-assisted treatment marginalizes patients and providers. *Soc. Sci. Med.* **232**, 324–331 (2019).
11. Balsamo, D., Bajardi, P., Morales, G. D. F., Monti, C. & Schifanella, R. The pursuit of peer support for opioid use recovery on Reddit. In *Proc. International AAAI Conference on Web and Social Media* Vol. 17, 12–23 (AAAI Press, 2023).
12. D'Agostino, A. R. et al. Social networking online to recover from opioid use disorder: a study of community interactions. *Drug Alcohol Depend.* **181**, 5–10 (2017).
13. Chancellor, S., Nitzburg, G., Hu, A., Zampieri, F. & De Choudhury, M. Discovering alternative treatments for opioid use recovery using social media. In *Proc. 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)* (Association for Computing Machinery, New York, NY, USA, 2019).
14. Almeida, A. et al. The use of natural language processing methods in Reddit to investigate opioid use: scoping review. *JMIR Infodemiology* **4**, e51156 (2024).
15. Park, A. & Conway, M. et al. Opioid surveillance using social media: How URLs are shared among Reddit members. *Online J. Public Health Inform.* **10**, e62174 (2018).
16. Bunting, A. M. et al. Socially-supportive norms and mutual aid of people who use opioids: an analysis of Reddit during the initial COVID-19 pandemic. *Drug Alcohol Depend.* **222**, 108672 (2021).
17. Chen, X., Lee, W. & Lin, F. Infodemic, institutional trust, and COVID-19 vaccine hesitancy: a cross-national survey. *Int. J. Environ. Res. Public Health* **19**, 8033 (2022).
18. Sager, M. A. et al. Identifying and responding to health misinformation on Reddit dermatology forums with artificially intelligent bots using natural language processing: design and evaluation study. *JMIR Dermatol.* **4**, e20975 (2021).
19. Pollack, C. C. et al. Characterizing the prevalence of obesity misinformation, factual content, stigma, and positivity on the social media platform Reddit between 2011 and 2019: Infodemiology study. *J. Med. Internet Res.* **24**, e36729 (2022).
20. Suarez-Lledo, V. & Alvarez-Galvez, J. Prevalence of health misinformation on social media: systematic review. *J. Med. Internet Res.* **23**, e17187 (2021).
21. Thach, H., Mayworm, S., Delmonaco, D. & Haimson, O. (In)visible moderation: a digital ethnography of marginalized users and content moderation on Twitch and Reddit. *N. Media Soc.* **26**, 4034–4055 (2024).
22. Foriest, J. C., Mittal, S., Bray, K., Tran, A.-T. & De Choudhury, M. A cross community comparison of muting in conversations of gendered violence on Reddit. *Proc. ACM Hum. Comput. Interact*. **8**, https://doi.org/10.1145/3686940 (2024).
23. Wu, Q., Williams, L. K., Simpson, E. & Semaan, B. Conversations about crime: re-enforcing and fighting against platformed racism on Reddit. *Proc. ACM Hum. Comput. Interact*. **6**, https://doi.org/10.1145/3512901 (2022).
24. ElSherief, M. et al. Characterizing and identifying the prevalence of web-based misinformation relating to medication for opioid use disorder: machine learning approach. *J. Med. Internet Res.* **23**, e30753 (2021).
25. Wayne Kepner, M. C. M. & Nobles, A. L. Types and sources of stigma on opioid use treatment and recovery communities on reddit. *Subst. Use Misuse* **57**, 1511–1522 (2022).
26. Giorgi, S. et al. Lived experience matters: automatic detection of stigma toward people who use substances on social media. In *Proc. International AAAI Conference on Web and Social Media.* Vol. 18, 474–487 (Association for Computing Machinery, New York, NY, USA, 2024).
27. Pierson, E. et al. Using large language models to promote health equity. *NEJM AI* **2**, 2 (2025).
28. Sharma, A., Lin, I. W., Miner, A. S., Atkins, D. C. & Althoff, T. Human–ai collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nat. Mach. Intell.* **5**, 46–57 (2023).
29. Liu, T., Zhao, H., Liu, Y., Wang, X. & Peng, Z. Compeer: A generative conversational agent for proactive peer support. In *Proc. 37th Annual ACM Symposium on User Interface Software and Technology, UIST '24*. https://doi.org/10.1145/3654777.3676430 (Association for Computing Machinery, 2024).
30. Hsu, S.-L. et al. Helping the helper: supporting peer counselors via ai-empowered practice and feedback. In *Proc. ACM Hum.-Comput. Interact.* **9**, CSCW095 (2025).
31. Heinz, M. V. et al. Randomized trial of a generative ai chatbot for mental health treatment. *NEJM AI* **2**, AIoa2400802 (2025).
32. Jörke, M. et al. GPTCoach: Towards LLM-Based Physical Activity Coaching. In *Proc. 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. 993, 1–46 (2025) (Association for Computing Machinery, New York, NY, USA, 2025).
33. Bouzoubaa, L., Aghakhani, E. & Rezapour, R. Words matter: Reducing stigma in online conversations about substance use with large language models. In *Proc. 2024 Conference on Empirical Methods in Natural Language Processing* (eds Al-Onaizan, Y., Bansal, M. & Chen, Y.-N.) 9139–9156. https://aclanthology.org/2024.emnlp-main.516/ (Association for Computational Linguistics, 2024).
34. Mittal, S., Jung, H., ElSherief, M., Mitra, T. & De Choudhury, M. Online myths on opioid use disorder: a comparison of reddit and large language model. In *Proc. International AAAI Conference on Web and Social Media.* Vol. 19, 1224–1245 (Association for Computing Machinery, New York, NY, USA, 2025).
35. ElSherief, M. et al. Identification of myths and misinformation about treatment for opioid use disorder on social media. *JMIR Form Res.* **8**, e44726 (2024).
36. Schleider, J. L., Dobias, M., Sung, J., Mumper, E. & Mullarkey, M. C. Acceptability and utility of an open-access, online single-session intervention platform for adolescent mental health. *JMIR Ment. Health* **7**, e20513 (2020).
37. Schleider, J. L. et al. Single-session interventions for mental health problems and service engagement: umbrella review of systematic reviews and meta-analyses. *Annu. Rev. Clin. Psychol.* **21**, 279–303 (2025).
38. Albarracin, D. & Shavitt, S. Attitudes and attitude change. *Annu. Rev. Psychol.* **69**, 299–327 (2018).
39. Swire-Thompson, B., Miklaucic, N., Wihbey, J. P., Lazer, D. & DeGutis, J. The backfire effect after correcting misinformation is strongly associated with reliability. *J. Exp. Psychol. Gen.* **151**, 1655 (2022).
40. Fast, E., Chen, B. & Bernstein, M. S. Empath: understanding topic signals in large-scale text. In *Proc. 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. 4647–4657 (Association for Computing Machinery, New York, NY, USA, 2016).
41. Choi, M., Aiello, L. M., Varga, K. Z. & Quercia, D. Ten social dimensions of conversations and relationships. In *Proc. Web Conference 2020 (WWW '20)*. 1514–1525 (Association for Computing Machinery, New York, NY, USA, 2020).
42. Francis, M. & Booth, R. J. *Linguistic Inquiry and Word Count* (Southern Methodist University, 1993).
43. Corrigan, P. W. & Watson, A. C. Understanding the impact of stigma on people with mental illness. *World Psychiatry* **1**, 16 (2002).
44. Thornicroft, G. et al. Evidence for effective interventions to reduce mental-health-related stigma and discrimination. *Lancet* **387**, 1123–1132 (2016).
45. Entman, R. M. Framing: Toward clarification of a fractured paradigm. *J. Commun.* **43**, 51–58 (1993).
46. Chong, D. & Druckman, J. N. Framing theory. *Annu. Rev. Polit. Sci.* **10**, 103–126 (2007).
47. Livingston, J. D., Milne, T., Fang, M. L. & Amari, E. The effectiveness of interventions for reducing stigma related to substance use disorders: a systematic review. *Addiction* **107**, 39–50 (2012).

48. BADRINATHAN, S. Educative interventions to combat misinformation: evidence from a field experiment in India. *Am. Political Sci. Rev.* **115**, 1325–1341 (2021).
49. Zonoobi, M., Tabatabaee, M. & Amini, H. The effects of an educational intervention on reducing stigma among medical students toward patients with psychiatric disorders. *BMC Med. Educ.* **24**, 1216 (2024).
50. Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N. & Cook, J. Misinformation and its correction: continued influence and successful debiasing. *Psychol. Sci. Public Interest* **13**, 106–131 (2012).
51. Tully, M., Vraga, E. K. & Bode, L. Designing and testing news literacy messages for social media. *Mass Commun. Soc.* **23**, 22–46 (2020).
52. Corrigan, P. W. & Nieweglowski, K. Stigma and the public health agenda for the opioid crisis in america. *Int. J. Drug Policy* **59**, 44–49 (2018).
53. Goffman, E. *Stigma: Notes on the Management of Spoiled Identity* (Simon and Schuster, 2009).
54. Brehm, J. W. *A theory of psychological reactance* (New York, NY: Academic Press, 1966).
55. Eschliman, E. L. et al. First-hand accounts of structural stigma toward people who use opioids on reddit. *Soc. Sci. Med.* **347**, 116772 (2024).
56. Meyers, S. et al. The intersection of gender and drug use-related stigma: a mixed methods systematic review and synthesis of the literature. *Drug Alcohol Depend.* **223**, 108706 (2021).
57. Parlier-Ahmad, A. B., Martin, C. E., Radic, M. & Svikis, D. S. An exploratory study of sex and gender differences in demographic, psychosocial, clinical, and substance use treatment characteristics of patients in outpatient opioid use disorder treatment with buprenorphine. *Transl. Issues Psychol. Sci.* **7**, 141 (2021).
58. Wood, E. & Elliott, M. Opioid addiction stigma: the intersection of race, social class, and gender. *Subst. Use Misuse* **55**, 818–827 (2020).
59. Stone, R. Pregnant women and substance use: fear, stigma, and barriers to care. *Health justice* **3**, 1–15 (2015).
60. Ursachi, G., Horodnic, I. A. & Zait, A. How reliable are measurement scales? External factors with indirect influence on reliability estimators. *Procedia Econ. Financ.* **20**, 679–686 (2015).
61. Loewenthal, K. M. & Lewis, C. A. *An Introduction to Psychological Tests and Scales* (Routledge, 2020).
62. Yoo, J.-M. et al. The role of knowledge and personal experience in shaping stigma associated with covid-19 and mental illness. *Psychiatry Investig.* **22**, 110 (2025).
63. Stewart, B. J. & Archbold, P. G. Nursing intervention studies require outcome measures that are sensitive to change: Part two. *Res. Nurs. Health* **16**, 77–81 (1993).
64. Shelby, R. et al. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proc. 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES '23)*. 723–741 (Association for Computing Machinery, New York, NY, USA, 2023).
65. Corvi, E. et al. Taxonomizing representational harms using speech act theory. *In Findings of the Association for Computational Linguistics: ACL 2025,* pages 3907–3932*,* Vienna, Austria (Association for Computational Linguistics*,* 2025).
66. Distler, V. The influence of context on response to spear-phishing attacks: an in-situ deception study. In *Proc. 2023 CHI Conference on Human Factors in Computing Systems, CHI '23*. https://doi.org/10.1145/3544548.3581170 (Association for Computing Machinery, 2023).
67. Lee, H.-P. H. et al. The impact of generative ai on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proc. ACM CHI Conference on Human Factors in Computing Systems* (ACM, 2025).
68. Raccah, O., Chen, P., Gureckis, T. M., Poeppel, D. & Vo, V. A. The "naturalistic free recall" dataset: four stories, hundreds of participants, and high-fidelity transcriptions. *Sci. Data* **11**, 1–9 (2024).
69. Laud, T. et al. Large-scale analysis of online questions related to opioid use disorder on reddit. In *Proc. International AAAI Conference on Web and Social Media.* Vol. 19, 1068–1084 (2025). (Association for Computing Machinery, New York, NY, USA, 2025).
70. Achiam, J. et al. GPT-4 technical report. Preprint at https://doi.org/10.48550/arXiv.2303.08774 (2024).
71. Kruis, N. E., McLean, K. & Perry, P. Exploring first responders' perceptions of medication for addiction treatment: Does stigma influence attitudes? *J. Subst. Abus. Treat.* **131**, 108485 (2021).
72. Yang, L. H. et al. A new brief opioid stigma scale to assess perceived public attitudes and internalized stigma: evidence for construct validity. *J. Subst. Abus. Treat.* **99**, 44–51 (2019).
73. Brown, B. S., Benn, G. J. & Jansen, D. R. Methadone maintenance: some client opinions. *Am. J. Psychiatry* **132**, 623–626 (1975).
74. Gustin, R., Nichols, J. & Martin, P. Individualizing opioid use disorder (oud) treatment: time to fully embrace a chronic disease model. *J. Reward Defic. Syndr.* **1**, 10–15 (2015).
75. Kang, M., Ragan, B. G. & Park, J.-H. Issues in outcomes research: an overview of randomization techniques for clinical trials. *J. Athl. Train.* **43**, 215–221 (2008).
76. Greene, C. M. & Murphy, G. Debriefing works: Successful retraction of misinformation following a fake news study. *PLoS ONE* **18**, e0280295 (2023).
77. Murphy, G. & Greene, C. M. Conducting ethical misinformation research: deception, dialogue, and debriefing. *Curr. Opin. Psychol.* **54**, 101713 (2023).
78. Weller, K. & Kinder-Kurlanda, K. E. A manifesto for data sharing in social media research. In *Proc. 8th ACM Conference on Web Science (WebSci '16)*. 166–172 (Association for Computing Machinery, New York, NY, USA, 2016).
79. Reddit, Inc. Reddit Data API Update: Changes to Pushshift Access. https://www.reddit.com/r/modnews/comments/134tjpe/reddit_data_api_update_changes_to_pushshift_access/ (Accessed 21 August 2025) (2023).
80. Kincaid, J. P., Fishburne R. P., Jr., Rogers, R. L. & Chissom, B. S. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. No. RBR875 (1975).

Acknowledgements

The authors thank the anonymous participants for their time and effort. The authors also thank Oscar Zhang for building the interactive interface. S.M. and M.D.C. were partly supported through NSF grant #2230692. Along with T.M., S.M., and M.D.C. were also partly supported through NIDA grant DA056725-01A1.

Author contributions

S.M., M.E., T.M., and M.D.C. contributed to the conceptualization of the paper. S.M., D.S., and S.W.D. conducted the study and interpreted the data. S.M. prepared the initial draft of the manuscript. All authors contributed additional feedback on the draft, leading to major revisions, and approved the final version of the paper.

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/s44387-025-00049-z.

Correspondence and requests for materials should be addressed to Shravika Mittal.

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