

LLMs Homogenize Values in Constructive Arguments on Value-Laden Topics

Farhana Shahid
Cornell University
Ithaca, United States
fs468@cornell.edu

Stella Zhang
Cornell University
Ithaca, United States
jz766@cornell.edu

Aditya Vashistha
Cornell University
Ithaca, United States
adityav@cornell.edu

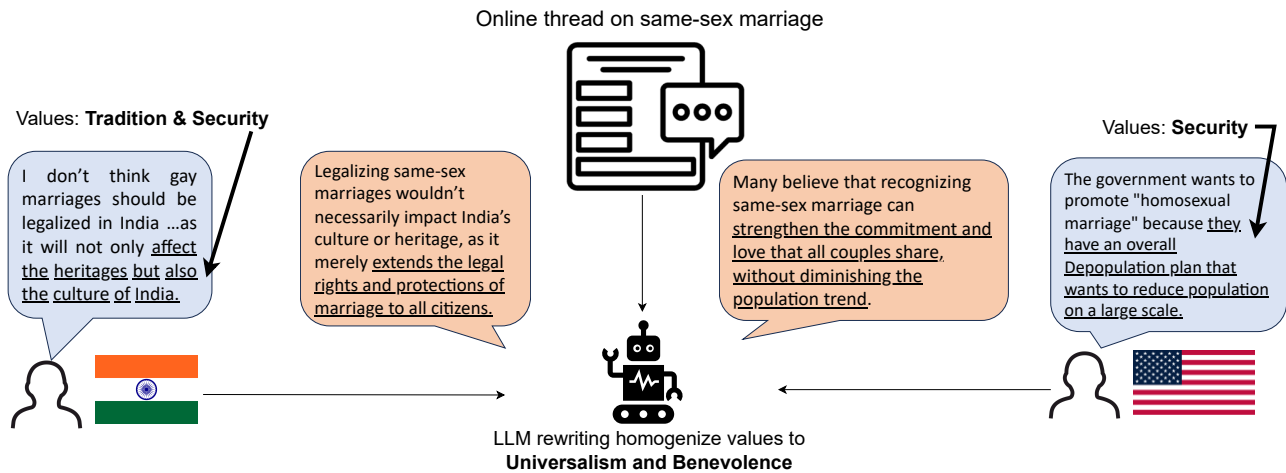


Figure 1: Examples of value homogenization by LLMs when writing constructively on value-laden topics. Comments in blue boxes are human-written and the ones in red boxes are LLM-rewritten versions.

Abstract

Large language models (LLMs) are increasingly used to promote prosocial and constructive discourse online. Yet little is known about how they negotiate and shape underlying values when re-framing people's arguments on value-laden topics. We conducted experiments with 347 participants from India and the United States, who wrote constructive comments on homophobic and Islamophobic threads, and reviewed human-written and LLM-rewritten versions of these comments. Our analysis shows that LLM systematically diminishes Conservative values while elevating prosocial values such as Benevolence and Universalism. When these comments were read by others, participants opposing same-sex marriage or Islam found human-written comments more aligned with their values, whereas those supportive of these communities found LLM-rewritten versions more aligned with their values. These findings suggest that LLM-driven value homogenization can shape how

diverse viewpoints are represented in contentious debates on value-laden topics and may influence the dynamics of online discourse critically.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI.**

Keywords

values, constructive argument, stance, HAI misalignment

ACM Reference Format:

Farhana Shahid, Stella Zhang, and Aditya Vashistha. 2026. LLMs Homogenize Values in Constructive Arguments on Value-Laden Topics. In *Proceedings of CHI conference on Human Factors in Computing (CHI '26)*. ACM, New York, NY, USA, 16 pages. <https://doi.org/XXXXXXX.XXXXXXX>

Content Warning: Some examples in this paper may be offensive or upsetting.

1 Introduction

Disagreements on value-laden topics are especially prone to breakdown, often escalating into toxicity and personal attacks [11, 32]. While researchers define constructive disagreement as civil dialogue oriented toward finding common ground [28, 44], what qualifies as 'constructive' is often contested. Because, people tend to embrace arguments that affirm their own values and resist those that challenge them [48]. Here, values refer to human beliefs regarding

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CHI '26, Barcelona, Spain

© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/2018/06
<https://doi.org/XXXXXXX.XXXXXXX>

desirable end states or modes of conduct [69]. Different value priorities both within and across cultures frequently fuel conflict on divisive issues, which often are value-laden [41, 48, 74].

Recent studies have explored if Large Language Models (LLMs) can help people write better arguments and express their viewpoints constructively on value-laden topics, such as gun regulation, homophobia, and Islamophobia among others [7, 39, 76, 86]. While findings suggest that LLMs can facilitate more constructive discourse, they also reveal misalignment between human's and LLM's preferences for different linguistic styles of constructiveness [76]. For example, humans tend to value direct, logical arguments, while LLMs favor polite, indirect arguments as constructive [76].

Although *values* strongly shape how people frame their arguments on divisive topics [5, 21], little is known about how LLMs handle *underlying values* when rewriting people's comments constructively. Given the growing adoption of LLMs in providing writing support [13, 80, 94, 96] and facilitating civil discourse online [38, 49, 55], it is essential to understand how these models affect the expression of values in value-laden topics. If LLM rewrites comments in ways that shift or distort underlying socio-cultural values, its output may appear constructive but risks alienating the very people it aims to support. Therefore, in this work we aim to explore:

- RQ1:** What values do people from different cultures emphasize when writing constructively on value-laden topics?
- RQ2:** What values do LLMs emphasize when rewriting people's comments constructively?
- RQ3:** How might human-LLM value differences affect which comments people prefer?

To find answer, we conducted a three-phase study with 347 participants from India and the United States, recruited via Prolific. In Phase 1, participants first shared their perceptions of constructiveness and wrote constructive comments on the same homophobic and Islamophobic threads. In Phase 2, we introduced the intervention: we used GPT-4 to rewrite participants' comments from Phase 1 constructively. We also invited additional participants, who wrote an initial constructive comment and then used GPT-4 for rewriting their comments constructively. Together, these steps yielded our pairs of human-written and LLM-rewritten constructive comments. We analyzed the underlying values in these comments based on Schwartz's theory of basic human values [73]. Finally, in Phase 3 we evaluated perceived alignment. A new group of participants, distinct from those in Phases 1 and 2, reviewed the human-written and LLM-rewritten comment pairs and indicated which better aligned with their own values.

In response to RQ1, we found that people's stances on homophobia, rather than culture, influenced what values they expressed when writing constructively on homophobic threads. Participants opposing same-sex marriage framed their arguments through Traditional values more (see Figure 1) than those who supported. Following RQ2, when an LLM was employed to rewrite comments constructively, it systematically suppressed Conservative values while amplifying prosocial values, such as Universalism and Benevolence (i.e., welfare of others and the society; see Figure 1). In this process, LLM changed the stance towards more neutral and supportive direction, especially for comments that were written by opponents of same-sex marriage and Islam. As a result, in RQ3,

when these comments were reviewed by other people, those opposing same-sex marriage and Islam found human-written comments more aligned with their values, while those supportive favored LLM-rewritten versions. We discuss the political tensions and ethical implications of LLM-driven value homogenization when shaping online discourse on value-laden topics. Taken together, our work makes the following contributions:

- We systematically show that LLMs change underlying value framings when rewriting people's comments constructively on value-laden topics.
- We point to the ethical issue of LLMs distorting stances by rewriting comments from individuals who oppose same-sex marriage or Islam as if they were supportive, thereby misrepresenting and conflicting with their value systems.
- We discuss the socio-political complexity of developing Human-AI systems that prioritize prosocial values towards vulnerable populations at the expense of compromising diverse viewpoints and value systems.

2 Related Work

We situate our work by first describing research on using LLMs to facilitate constructive arguments on value-laden topics, outlining the role of values in shaping argumentation. We then present existing work examining human-LLM value alignments.

2.1 Constructive Disagreements and The Role of Values

Constructive disagreements are typically characterized as civil dialogues that are focused, issue-relevant [45, 83] and evidence-based [44]. While other definitions focus on outcomes, such as dispute resolution [24], finding common ground [28], improved team performance [61], or the emergence of new solutions [28]. In online discourse, constructiveness manifests through specific linguistic features, such as longer content [45, 83], richer argumentative markers (e.g., discourse connectives, stance adverbials, reasoning verbs, modals, and root clauses) [44, 59], greater readability [46], greater use of named entities [46, 61], more politeness [24, 59, 95], and less hedging [23, 59, 61].

Given the tension, time, and effort involved in expressing constructive disagreement on value-laden topics [22, 32, 57], researchers have explored if LLMs can support people during this process. For example, Shahid et al. [76] found that LLMs can help people express their opinions on value-laden topics more constructively than they could do on their own, by integrating different linguistic markers of constructiveness in writing. Similarly, Argyle et al. [7] demonstrated that LLMs can help people reframe their arguments by incorporating politeness, restating opposing views, or validating others' sentiments, which improve conversation quality during disagreement. Kambhatla et al. [39] also showed that reframing human-written comments by an LLM to include hedging, acknowledgment, and agreement enabled readers to be more open to alternative perspectives on value-laden topics. Additionally, Govers et al. [31] observed that highly cooperative and persuasive strategies present in LLM-mediated comments were effective in influencing reader's opinions on polarizing value-laden topics. This

growing body of work highlights the promise of LLMs in scaffolding constructive disagreement by leveraging linguistic strategies that individuals often struggle to apply in contentious situations.

Yet constructive disagreement is not only about how arguments are expressed but also about the values they convey. Values shape disagreement by guiding what people emphasize, how they interpret issues, and which recommendations they accept [26]. Research on value-based argumentation shows that persuasiveness of an argument depends on whether the values invoked resonate with the audience [8, 12, 14]. Arguments that combine closely related values (e.g., Tradition and Conformity) are more convincing than those that appeal to conflicting ones (e.g., Freedom and Conformity) [52].

Despite the centrality of values in online argumentation, little is known about how LLMs handle them when rewriting people's comments on value-laden, divisive topics. If an LLM shifts the values expressed in writing, it may satisfy the surface-level linguistic markers of constructiveness but fail to support genuine conflict resolution, since people are more receptive to arguments aligned with their own values [48]. This critical gap motivates our research to systematically examine what values humans and LLMs emphasize when writing constructively on value-laden topics. We now present recent work investigating the value orientations of LLMs.

2.2 Human-LLM Value (Mis)Alignment

With the rapid proliferation of LLMs, researchers have increasingly turned to examining the value orientations of these models for building safe, responsible, and human-centered AI systems [4, 88, 90]. A growing body of work highlights various forms of misalignment between the values of humans and LLMs. For example, research shows that LLMs are more likely to align with prosocial and altruistic values, such as Universalism and Benevolence compared to Security concern or individualistic values, such as Freedom, Power, and Hedonism [50, 70, 75, 79, 92, 98]. These preferences are usually consistent across both short and long-form responses generated by LLMs on value-laden questions [58].

In contrast, few studies [62, 89, 91] found that when presented with role playing or moral scenarios, LLM-generated responses balanced prosocial and individualistic values in ways that mirrored human preferences in those contexts. Despite these apparent alignments, researchers cautioned that while LLMs can produce broadly consensual statements affirming the importance of respecting human values, they lack a deeper understanding of situations when human values are threatened [40]. In fact, there is a critical gap between the values LLMs explicitly state and the actions they choose when asked to act upon those values [78].

Different value priorities across different cultural contexts further complicate the process. For example, Bu et al. [17] identified multiple ways in which LLMs misalign with cultural values, including inaccurate details, cultural misunderstanding, and cultural reductionism in outputs, among others. Cross-cultural evaluations also reveal that most LLMs align more closely with Western cultural values than those from Asian or African countries [4, 37, 78, 79, 82, 85]. Strikingly, even LLMs developed in non-Western regions (e.g., China, India) often converge toward a moderate cultural middle ground, overlooking distinctive traits of their own cultural contexts [3, 75, 82].

Taken together, these findings highlight persistent value misalignments between LLMs and humans across different cultural contexts. Since LLMs are being gradually integrated into online platforms to promote civil discourse [38, 49, 55], it is crucial to examine how these models navigate and shape the underlying values when helping people from different cultures express their opinions constructively on value-laden topics.

3 Methods

To address our research questions, we conducted a three-phase study (see Figure 2) with participants from two different cultures: India and the United States (US), who are known to differ in value priorities [81]. In the first phase, we examined which values participants emphasize when writing constructively on value-laden topics. In the second phase, we analyzed how these underlying values change when LLM rewrites people's comments constructively. In the third phase, we investigated which type of comment: human or LLM-written people find more aligned with their values. The study protocol was approved by the institutional review board at our institution.

3.1 Phase 1: Values in Human-Written Constructive Comments

Study Design. We recruited 58 Indian and 58 American participants on Prolific. Each participant was asked to write constructive comments on value-laden topics. For this writing task, we curated two Reddit threads: one related to homophobia (Those who oppose same-sex marriage, why?) and another related to Islamophobia (Why do people dislike Islam?). We chose these issues because they are highly value-laden and considered socially divisive both in India and the US. Each thread, presented in English, included the original post and four user comments. Overall, 63% of comments in these threads were negative in sentiment (average: -0.69), with several being highly negative and toxic.

We first asked participants about their stance on same-sex marriage and Islam, using adapted instruments from Pew research survey [15, 20] since people's values are associated with their stances on divisive issues [53]. We then randomly showed participants either a homophobic or an Islamophobic thread, and asked them to write a constructive comment in response to the thread. To ensure comparability, we used the same threads for both demographics, so that any differences in values could not be attributed to variation in stimuli. We also asked participants an open-ended question about their perceptions of constructive comments to get more insight into how they framed their writing. We compensated each participant with USD 1.00 for completing the writing task.

Data Analysis. We collected 116 constructive comments written by our participants. We first used GPT-4 to annotate the values expressed in different segments of each comment, following Schwartz's theory of basic human values [73]—which has been used to detect values in arguments in cross-cultural setting [41]. Additionally, several researchers applied Schwartz's theory to measure human-LLM value alignment [50, 75, 78, 79, 89]. Schwartz's theory identifies a set of universal values that guide human behavior, organized into ten broad categories such as Security, Tradition,

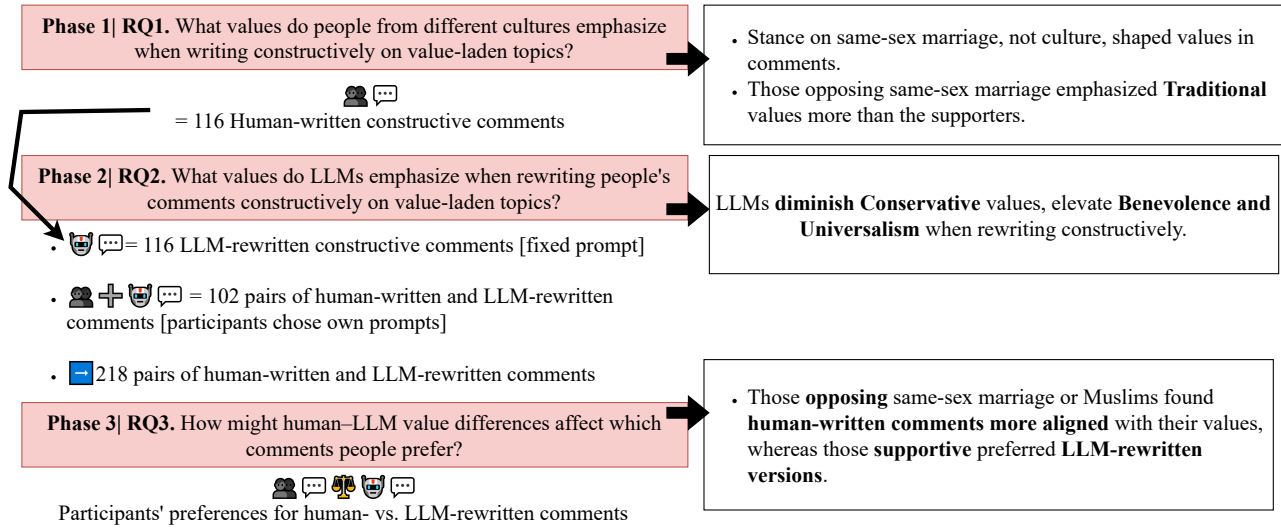


Figure 2: An outline of our data collection methods in each phase and key findings.

Conformity, Universalism, Benevolence, Power, Achievement, Stimulation, Hedonism, and Self-direction [73].

We used GPT-4 to annotate these values in participants' comments. Overall, each comment represented a vector of values. For example, the comment shown in Figure 3 represents six distinct values: Universalism, Hedonism, Self-direction, Tradition, Conformity, and Security. The annotations can either have a positive or negative association with the value. For example, the segment '*why society should interfere*' captures lack of tolerance (Universalism). Some segments can represent multiple values, such as '*man get married to a woman*' signifies both Traditional value and Conformity to rules.

Two authors independently reviewed GPT-4's annotation of values, making updates when we disagreed with the assigned value or when a comment segment reflected multiple values. Inter-rater reliability was high (Cohen's $\kappa = 0.87$). Any remaining disagreements were resolved through discussion. After the annotations were confirmed, we conducted multiple Chi-square tests with Bonferroni correction to examine if there was any cross-cultural difference in the values Indian and Americans expressed in their comments across different topics and stances.

3.2 Phase 2: Values in LLM-Rewritten Constructive Comments

LLM Rewriting. Next, we used GPT-4 to rewrite participants' comments from Phase 1 to make them constructive. To minimize misalignment between human's and LLM's understanding of constructiveness, we relied on the different characteristics of constructive comments that participants provided in Phase 1. We synthesized their open-ended responses into a summary definition: constructive comments are "*respectful, fact-based, balanced, and thoughtful contributions that promote understanding and progress in conversation rather than conflict*"—which also align with the perceptions of constructiveness from prior studies [28, 44, 59, 83, 95]. We used zero-shot, cultural prompting to ensure that LLM rewrites align with

the corresponding country's cultural values [85]. We instructed GPT-4 to rewrite each comment constructively within 100 words, matching the average length of constructive comments as reported in prior studies [45]. We used the following prompt:

```
Constructive comments are respectful, fact-
based, balanced, and thoughtful contribu-
tions that promote understanding and progress
in conversation rather than conflict.
Based on this definition, rewrite the fol-
lowing comment to make it constructive. An
<Indian, American> participant, who <insert
stance> <same-sex marriage, Islam> wrote this
comment. Use at most 100 words when rewrit-
ing.
<insert comment>
```

For each human-written comment from Phase 1, we generated one LLM-rewritten version, resulting in 116 LLM-rewritten constructive comments. Since we used a fixed prompt to generate these rewrites, we wanted to make sure that our findings are not sensitive to the prompt. Therefore, we collected additional comments from participants, where they could choose their own prompts to rewrite constructively using LLM, which we describe below.

Human-LLM Collaborative Writing. Similar to Phase 1, we recruited 51 new participants (US: 26, India: 25) on Prolific for writing constructive comments on value-laden topics. Participants first reported their stance on same-sex marriage and Islam since people's values and stances on value-laden topics are closely related [53]. Then they were asked to write a constructive comment in response to a randomly selected homophobic and an Islamophobic thread. We used different sets of threads (see Table 7 in Appendix) from Phase 1 to ensure that our analyses on value expressions were not sensitive to specific threads.

Universalism (equality) Every people have the fundamental right to live. *Hedonism (pleasure)* If a girl is happy with another girl and vice versa then

Universalism (tolerance) why society should interfere. *Self-direction (freedom of action)* They should live their way in which they like. It's their own choice. Society

Universalism (tolerance) should accept them. *Tradition + Conformity (being compliant)* What if a man interested in man get married to a woman, *Security (having a good life)* both of their lives would be

Self-direction (freedom of action) ruined. So its better to do whatever they want.

Figure 3: Annotated Schwartz’s human values (blue texts) in a comment written by an Indian participant on the homophobic thread.

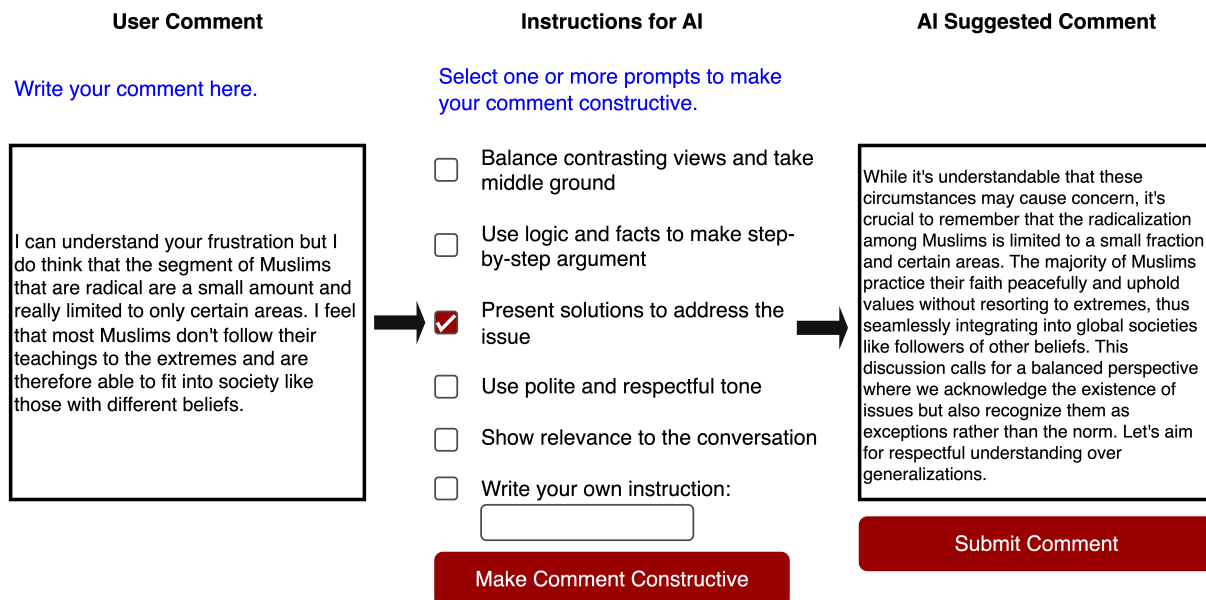


Figure 4: The interface for human-LLM collaborative writing of constructive comments. Participants first entered their comment in the textbox for User Comment and then selected one or more prompts from Instructions for AI section. After they clicked “Make Comment Constructive”, GPT-4 rewrote their comment constructively based on given instructions in real time, which appeared in the AI Suggested Comment box. Participants could repeat the process as many times as they wished. The example shown illustrates a comment written by an American participant in response to an Islamophobic thread.

We asked participants to write a constructive comment on their own first. Then they could prompt an LLM to rewrite their comments constructively (see Figure 4). We provided a set of prompts based on different characteristics of constructive comments that other participants described in Phase 1. Participants could select multiple prompts or write their own prompt. We used GPT-4 to rewrite their comments constructively in real-time using the following instruction:

Constructive comments <insert comma-separated list of prompts selected or written by the participant>
Based on this instruction rewrite the following comment to make it constructive. An <Indian, American> participant who <insert stance> <same-sex marriage, Islam> wrote this

```
comment. Use at most 100 words when rewriting.
<insert participant's comment>
```

After receiving LLM-rewritten version, participants could prompt the LLM again as many times as they wanted. Once they were satisfied with the final output, they could submit it to finish the task. We compensated each participant with USD 1.70 for completing this task. For this study, we logged all the initial human-written comments and the corresponding LLM-rewritten versions.

Data Analysis. We generated 116 LLM-rewritten version of the comments that participants wrote in Phase 1. Additionally, from human-LLM collaborative writing in Phase 2, we collected 102 human-written comments along with LLM-rewritten versions. Overall, this led to 218 pairs of human-written comments and their corresponding LLM-rewritten versions.

First, we wanted to verify if the LLM-rewritten versions were indeed more constructive compared to the original human-written comments. Following prior work [45, 59, 76, 95], we calculated eight linguistic features of constructiveness, including length, readability score, politeness, number of named entities, and argumentative features, such as number of discourse connectives, stance adverbials, reasoning verbs, modals, and root clauses. For length, we calculated the number of words in each comment. We used SMOG index [63] (*textstat* library in Python) to measure readability scores. We evaluated politeness scores using *politeness* package in R. We utilized SpaCy [35] to get number of named entities and root clauses. For discourse connectives, stance adverbials, reasoning verbs, and modals, we adapted the open-source code from Kolhatkar et al. [46]. After calculating the features, we performed Mann-Whitney tests with Bonferroni correction to inspect how these features differed between original human-written and LLM-rewritten comments. Finally, to examine how LLMs handle underlying values when rewriting constructively, we annotated the values in our collected comments based on Schwartz’s theory of basic human values [73]. We followed the same procedure as described above in Phase 1. We performed multiple CHI-square tests with Bonferroni correction to analyze whether the values differed between human-written and LLM-rewritten constructive comments.

3.3 Phase 3: Perceived Value Alignment

Experiment Design. After collecting 218 pairs of human-written and LLM-written comments from both Phase 1 and Phase 2, we conducted a forced-choice experiment to examine what types of comments people perceived as more aligned with their values. We recruited 92 Indian and 92 American crowd workers from Prolific, who had not participated in both Phase 1 and Phase 2. We initially asked participants about their stance on same-sex marriage and Islam. Then they were randomly assigned to review comments either on homophobic or Islamophobic threads. Each participant reviewed four randomly selected pairs of human-written comments and their corresponding LLM-rewritten versions. We ensured that each participant reviewed pairs of comments in which the human-written version was authored by participants from the same country (in Phase 1 and 2) and shared the same stance as the reviewing participant. This is because, a participant supportive of same-sex

marriage or Islam may not find comments written against these topics aligned with their values and vice versa.

We randomized the order of comments within each pair and asked participants to select which comment aligned more with their values. We did not reveal to participants about how these comments were written. We also asked them open-ended questions about how their chosen comments aligned with their values. In total, we received 1122 human evaluations from 180 participants after discarding responses from four participants, who failed the attention check. Participants were compensated with USD 1.00 for completing the task.

Data Analysis. We conducted multiple Chi-square tests with Bonferroni correction to examine whether participants’ preferences for human-written vs. LLM-rewritten comments varied depending on their stance on same-sex marriage and Islam. We also reviewed the open-ended responses they provided for how their chosen comments aligned with their values.

3.4 Ethical Consideration

Given our study required participants to engage with value-laden and potentially harmful content, we undertook several safeguards. When we advertised the task on Prolific, we included an explicit content warning that the study involved sensitive topics, therefore giving potential participants the opportunity to make informed decision before choosing the task. We also utilized Prolific’s Harmful Content Prescreener to ensure the tasks were only shown to participants, who pre-identified themselves as being comfortable with reviewing sensitive content. Moreover, at the beginning of the study, we explicitly informed participants that they would be reviewing online threads containing homophobic and Islamophobic content. We also warned them of the potential negative emotional impact of engaging with such content and assured them that they could quit the task any time. All of the above information was clearly communicated to participants before we collected any data. After collecting participants’ stances on same-sex marriage and Islam, we displayed another content warning before proceeding to show them the online threads. This was to remind participants that they could quit the task any time if they chose to do so. Furthermore, the writing task in Phase 1 and 2 enabled participants to write on these issues from their own stances. We encouraged them to provide their truthful opinions and conveyed that the study did not intend to condone or criticize anyone’s opinions. We also included local resources for participants to manage any potential emotional distress from their participation in the task, and provided contact information if they wished to inform us of any concerns.

4 Findings

We first describe results from Phase 1 to answer what values Indians and Americans expressed when writing constructive comments on homophobic and Islamophobic threads (4.1). Then from Phase 2, we present our analyses of what values LLM emphasized when rewriting comments constructively on value-laden issues (4.2). Finally, we outline what types of comment people found more aligned with their values (4.3).

Table 1: Distribution of participants’ stances on same-sex marriage and Islam (Phase 1)

Same-sex marriage		Islam	
India	US	India	US
Against: 28%	Against: 7%	Against: 24%	Against: 38%
Neutral: 41%	Neutral: 21%	Neutral: 28%	Neutral: 24%
Support: 31%	Support: 72%	Support: 48%	Support: 38%

4.1 RQ1: What Values Do People Emphasize When Writing Constructively on Value-Laden Topics?

Our annotation of 116 human-written constructive comments resulted in 681 segments containing Schwartz’s basic human values. On average, we identified five underlying values ($SD=2.17$) in each comment.

A chi-square test with Yates’ continuity correction revealed significant cross-cultural differences in the distribution of Schwartz’s basic human values between constructive comments written by Indian and American participants; though the effect size was small ($\chi^2(7, N = 675) = 24.32, p < 0.001, \phi = 0.19$). Post hoc analysis with Bonferroni correction showed that Indian participants (13%) expressed significantly more **Traditional** values than the American participants (5%) when writing constructively on the same threads (see Figure 5A). For example, in response to the homophobic threads, some Indian participants invoked concerns about how same-sex marriages might disrupt the socio-cultural, ancestral, and religious traditions. On Islamophobic threads, these comments referred to different traditions practiced by Muslims. This pattern is consistent with prior research showing that Indian society places greater emphasis on traditional values than American society [84].

Cross-Cultural Values Across Topic. When we analyzed the underlying values across topics, we observed that the distribution of values differed significantly between constructive comments written by Indian and American participants; but only for the **homophobic thread** with a small effect size ($\chi^2(7, N = 355) = 18.52, p < 0.001, \phi = 0.23$). Post hoc analysis with Bonferroni correction showed that Indians (19.7%) expressed significantly more **Traditional** values than the Americans (6.4%) when writing constructive comments on the homophobic thread. We did not notice any significant cross-cultural difference for the Islamophobic thread. This indicates that the observed cross-cultural differences in values might be linked to participants’ stances on same-sex marriage, which we examine next.

Cross-Cultural Values Across Stance. Indian and American participants’ stances on same-sex marriage differed significantly; with a medium effect size ($\chi^2(2, N = 58) = 10.4, p < 0.01, \phi = 0.42$). Post hoc analysis with Bonferroni correction showed that Americans supported same-sex marriage (72%) significantly more than the Indians (31%) (see Table 1). In contrast, no significant cultural differences emerged in participants’ stances toward Islam, which may explain the absence of cross-cultural differences in the underlying values expressed in comments on the Islamophobic thread.

To further examine how culture and stance shaped the underlying values in constructive comments on the homophobic thread, we

Table 2: Examples of constructive comments written on homophobic thread by an Indian participant opposing same-sex marriage and an American participant supporting such marriage, with highlighted segments illustrating **Traditional values.**

Stance: opposing same-sex marriage	Stance: supportive of same-sex marriage
<p>Marriage is a sacred practice. It is not that simple like between two body it is much more beyond that it is the unification of two souls as they have to share everything their happiness their sorrow all of these. It is only possible between a man and woman.</p> <p>In today’s generation they think marriage is like a sale whenever they wanted they marry and whenever they want they got divorce. Gay is being normalised but it is not good for the society to function. God has created both men and women so they can share everything between each other.</p> <p>So in my opinion marriage should be only between men and women.</p>	<p>I agree that people should be able to live their lives however they choose as long as they are not hurting anybody else. Don’t we want everyone to be happy? That involves allowing people to marry whoever they want, assuming it is legal in other aspects as well. People who oppose same sex marriage don’t have anything better to do with their lives but judge others and be mad all the time.</p>

fit a multinomial logistic regression model. The analysis revealed only a significant **main effect of stance** on the value expressions ($\chi^2(7, N = 355) = 34.57, p < 0.0001, \phi = 0.37$). Post hoc analysis with Bonferroni correction also showed that participants who opposed same-sex marriage expressed significantly more (31.3%) **Traditional** values in their comments than those who supported (7.2%) such marriage (see Figure 5B). Table 2 gives an example of how opponents of same-sex marriage framed their arguments from a Traditional point of view.

Overall, these findings indicate that people’s stances on same-sex marriage, rather than culture, influenced the values they emphasized when writing constructively on homophobic threads.

4.2 RQ2: What Values Do LLMs Prioritize When Writing Constructively on Value-Laden Topics?

Are LLM-Rewrites Constructive?. We collected 218 pairs of human-written and LLM-rewritten comments both from Phase 1 and Phase 2. Our analyses revealed that these LLM-rewritten versions had significantly more linguistic features of constructiveness than original human-written comments (see Table 8 in Appendix). For example, LLM-rewritten comments were longer, more readable, more polite, more argumentative, i.e., contained more discourse connectives (logical reasoning in argument), reasoning verbs (conclusion in argument), modals (proposition in argument), and root clauses (independent clause) than the original human-written comments. This aligns with prior finding that LLMs can rewrite comments more constructively than humans [76].

Do LLM-Rewrites Have Same/ Different Values?. Our annotation of 218 pairs of human-written and LLM-rewritten constructive comments led to 2218 comment segments containing Schwartz’s basic human values. On average, we identified five values ($SD=2.18$) within each comment. Figure 6 shows the distribution of values

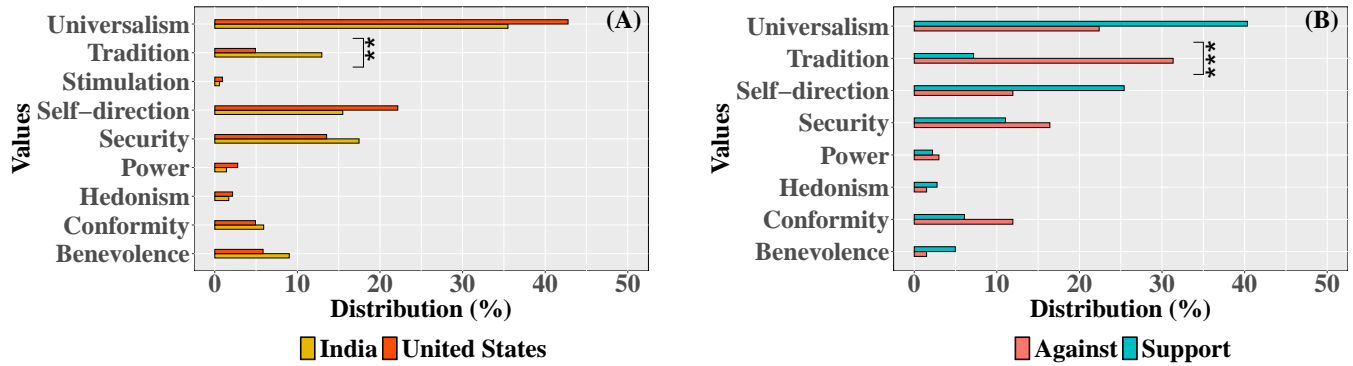


Figure 5: Distribution of Schwartz's human values in (A) all comments written by Indian and American participants, (B) comments written on homophobic thread by participants who either support or oppose same-sex marriage. Statistically significant differences are reported at $p < 0.0001$ (***), $p < 0.001$ (**), and $p < 0.01$ (*) [adjusted P-values after Bonferroni correction].

Table 3: Example of a constructive comment written by an American participant on an Islamophobic thread, alongside the corresponding LLM-rewritten version, with highlighted segments illustrating the values of **Security**, **Conformity**, and **Benevolence**.

Human-written comment	LLM-rewritten comment
There is a lot in the media that portrays Muslims as barbaric and aggressive so it is an understandable viewpoint whether it is true or not. There continues to be media coverage showing hostage situations and brutal killings that would easily paint that picture in people's minds and sway their viewpoint about Muslims in general when that behavior is most likely limited to specific terrorist groups.	The media often highlights violent incidents involving specific terrorist groups, which may inadvertently shape perceptions about Muslims as a whole. It's important to differentiate between these groups and the vast majority of Muslims who do not engage in such actions. Encouraging media outlets to provide a balanced perspective, highlighting the diverse contributions and peaceful practices of Muslim communities, can help counteract these negative portrayals. Open dialogues and educational initiatives can also foster better understanding and reduce stereotypes.

present across human-written and LLM-rewritten constructive comments.

We observed significant differences in the distribution of values between human-written and LLM-rewritten constructive comments, with a small effect size ($\chi^2(9, N = 1329) = 112.53, p < 0.000001, \phi = 0.23$). Post hoc tests with Bonferroni correction showed that LLMs **downplayed** values associated with **Security** (concern for social order and stability; human: 15.62%, LLM: 9.63%) and **Conformity** (adherence to rules; human: 5.95%, LLM: 2.29%), while **emphasizing Benevolence** (concern for others' welfare; human: 9.58%, LLM: 18.78%) and **Universalism** (concern for equality and social justice; human: 33.27%, LLM: 44.57%) when rewriting people's comments. Table 3 gives example of how LLM enhances Benevolence in its rewrite while downplaying Security concern and rule Conformity present in the original comment.

Subsequent Chi-square tests with Bonferroni correction showed that LLM consistently shifted underlying values when rewriting people's comments constructively across different cultures, topics,

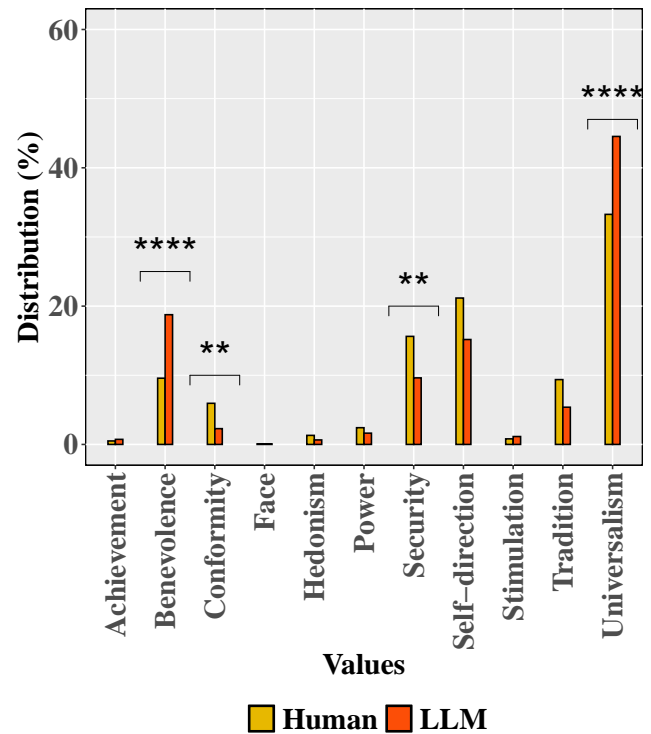


Figure 6: Distribution of Schwartz's human values in original human-written and LLM-rewritten constructive comments. Statistically significant differences are reported at $p < 0.000005$ (****), $p < 0.00005$ (***), $p < 0.0005$ (**), and $p < 0.005$ (*) [adjusted P-values after Bonferroni correction].

and stances; though effect sizes varied from small to medium (see Table 4). For instance, although Indians are known to exhibit greater **Conformity** to rules and social order [6, 84], LLM downplayed this value when rewriting their comments. In contrast, while Americans

Table 4: Results of multiple chi-square tests with Bonferroni corrections examining whether values differ between human-written and LLM-rewritten comments across different cultures, topics, and stances.

Topic	Cases	Statistics	Value Changes
Culture	India	$\chi^2(9, N = 1128) = 67.26, p < 0.000001, \phi = 0.24$	↓ Conformity (human: 5.9%, LLM: 1.6%) ↑ Universalism (human: 33%, LLM: 46.6%)
	US	$\chi^2(9, N = 1089) = 54.2, p < 0.000001, \phi = 0.22$	↓ Self-direction (human: 26.9%, LLM: 17.1%) ↑ Benevolence (human: 8.5%, LLM: 19.5%)
Topic	Homophobia	$\chi^2(9, N = 1126) = 71.12, p < 0.000001, \phi = 0.25$	↑ Benevolence (human: 5.4%, LLM: 16.2%)
	Islamophobia	$\chi^2(9, N = 1089) = 49.44, p < 0.000001, \phi = 0.21$	↓ Security (human: 21.6%, LLM: 12.3%) ↑ Universalism (human: 32.7%, LLM: 44.4%)
Stance	Against same-sex marriage	$\chi^2(6, N = 263) = 43.34, p < 0.000001, \phi = 0.41$	↓ Tradition (human: 27.4%, LLM: 8%) ↑ Universalism (human: 19.5%, LLM: 47.3%)
	Supports same-sex marriage	-	-
	Against Islam	$\chi^2(7, N = 371) = 23.12, p < 0.001, \phi = 0.25$	↑ Universalism (human: 26.5%, LLM: 44.4%)
	Supports Islam	-	-

are known to prioritize **Self-direction** (i.e., freedom of thought and action) [72], LLM reduced emphasis on this value. Similarly, in response to Islamophobic threads, many participants referred to **Security** concern around terrorism and extremism, which LLM deprioritized during rewriting. On the other hand, as observed in Phase 1 (see Figure 5B), opponents of same-sex marriage frequently appealed to **Traditional** values, which the LLM minimized in its rewriting.

According to Schwartz’s theory [73], the values that LLM consistently downplayed—Conformity, Security, and Tradition—fall under the higher-order value of **Conservation**, which emphasizes stability and the preservation of the existing social order. In contrast, Benevolence and Universalism—the values that LLM prioritized—are part of the higher-order value of **Self-Transcendence**, which prioritizes the welfare of others and the society.

Notably, participants who wrote comments supporting same-sex marriage and Islam already emphasized Benevolence or Universalism by appealing to equality and rejecting dehumanization of these groups. This might explain why LLM did not shift underlying values significantly when rewriting comments from participants with prosocial viewpoints (see Table 4). Moreover, our manual inspection revealed that for comments authored by opponents of same-sex marriage and Islam, changes in values in LLM-rewritten versions frequently (89% of cases) led to corresponding changes in stance—moving them toward more neutral or supportive positions. For instance, the example in Table 5 shows that although the participant expressed their views against same-sex marriage, LLM totally altered their arguments when rewriting constructively. In contrast, comments from those supportive did not undergo such changes.

Taken together, these findings indicate that LLM not only changes linguistic expression when writing constructively but also changes underlying values and often stances, systematically steering discourse toward more prosocial direction. While LLM’s emphasis on Benevolence and Universalism may enhance civility and constructive tone, it may risk value homogenization when it comes to representing people’s opinions on value-laden issues.

Table 5: Example of change in stance when LLM rewrote an Indian participant’s comment that opposed same-sex marriage.

Human-written comment	LLM-rewritten comment
having gay marriages or legalizing it against how the nature has made us, biologically there are only two genders male and female. I don’t think gay marriages should be legalized in India as marriages are meant for having a child which is not possible with gay marriages and it will not only affect the heritages but also the culture of India.	It’s important to understand that marriage is not exclusively meant for procreation. Many heterosexual couples also choose not to have children or can’t have them due to biological reasons. However, they are still allowed to marry. Legalizing same-sex marriages wouldn’t necessarily impact India’s culture or heritage, as it merely extends the legal rights and protections of marriage to all citizens. Moreover, gender diversity is biologically proven beyond just male and female categories, adding another dimension to our understanding of human nature.

Table 6: Distribution of participants’ stances on same-sex marriage and Islam (Phase 3)

Same-sex marriage		Islam	
India	US	India	US
Against: 47%	Against: 24%	Against: 58%	Against: 45%
Neutral: 22%	Neutral: 19%	Neutral: 21%	Neutral: 15%
Support: 31%	Support: 57%	Support: 21%	Support: 40%

4.3 RQ3: What Types of Comments Do People Find More Aligned with Their Values?

We collected 1,122 evaluations from a total of 180 participants: who reviewed randomly selected pairs of human-written and LLM-rewritten comments. We found that on average both Indian (59.9%) and American participants (61.1%) reported LLM-rewritten comments to be more aligned with their values than human-written ones. We conducted multiple Chi-square tests with Bonferroni correction to examine if the types of comment participants found aligned with their values differed based on their stances on same-sex marriage and Islam. Table 6 shows the distributions of participants’ stances across these issues.

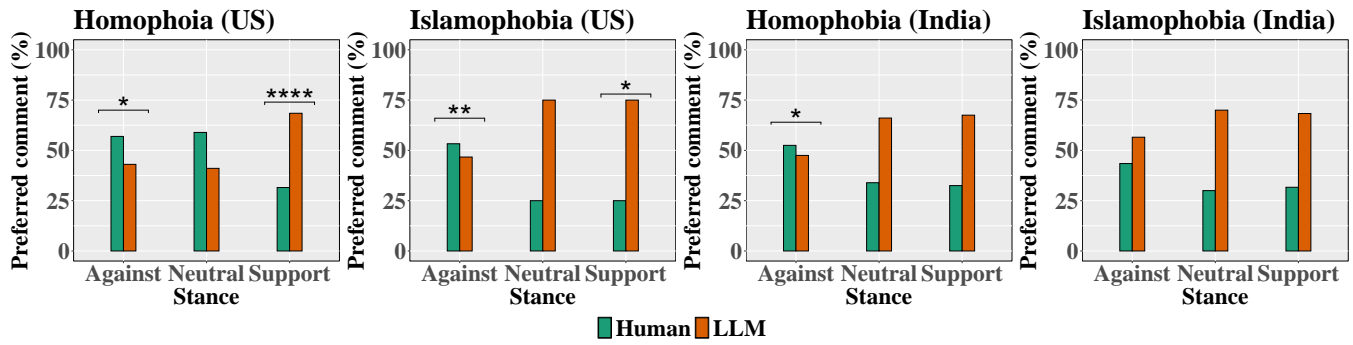


Figure 7: Distribution of Indian and American participants' preferred comments across stances. Statistically significant differences are reported at $p < 0.00001$ (**), $p < 0.0001$ (***), $p < 0.001$ (**), and $p < 0.01$ (*) [adjusted P-values after Bonferroni correction].**

American Participants. We found that the type of comment Americans found more aligned with their values differed significantly based on their stances on homophobia ($\chi^2(2, N = 296) = 20.51, p < 0.0001, \phi = 0.26$); but with a small effect size (see Figure 7). Post hoc analyses with Bonferroni corrections showed that Americans who opposed same-sex marriage (56.9%) found human-written comments significantly more aligned with their values than corresponding LLM-rewritten versions. They shared that their chosen comments better reflected their values around traditional ways of marriage between man and woman. An American participant wrote:

I really don't support same sex marriage. So I chose comments that also oppose same sex marriage or those try to explain why same sex marriage is not okay.

In contrast, those who supported same-sex marriage considered LLM-rewritten comments significantly more aligned with their values (68.4%) than the original human-written ones. Most of the American participants reported that they chose comments which aligned with their values on marriage equality and inclusion. Others chose LLM-rewritten comments because they were better worded, rational, respectful, and less emotional. A participant expressed:

Because the comments I chose show that empathy and respect for consenting adults is the primary focus when considering if gay marriage should be legal or not. The comments I chose show that we should be able to hear opposing thoughts, and we should challenge opponents to think through.

Similarly, we observed a significant difference in American participants' preferences for different comments based on their stances on Islam (see Figure 7), with a small effect only ($\chi^2(2, N = 272) = 13.92, p < 0.001, \phi = 0.23$). Americans who opposed Islam selected human-written comments significantly more than expected (53%) for being aligned with their values. They commented that their chosen comments accurately captured their perceptions of Islam by highlighting issues like terrorism, poor treatment of women and LGBTQ communities, lack of reform, and misalignment with Western liberal values. Some participants also pointed that the comments they did not choose were more woke, rambling, less factual,

less realistic, and read like generic AI. An American participant shared:

I believe Islam does not fit in with Western values, so my answers fully aligned with the answer that leaned more western-valued. They have different religions and culture. The West is a Christianity dominated region.

In contrast, American participants supporting Islam found LLM-rewritten comments significantly more aligned (75%) with their values than expected. They reported that their chosen comments were more constructive, expressed secular, compassionate values around understanding people from different religions, and pointed out that there are both good and bad people in every religion. One of the participants wrote:

The comments I chose align with my values because they are understanding of all religions and emphasize how important it is to understand others and not judge. This is the most important value to me. Also, They make sure everyone feels respected and not hurt.

For both issues, American participants who reported being neutral did not exhibit any significant preference for either type of comment.

Indian Participants. The type of comment Indian participants found more aligned with their values differed significantly based on their stances on homophobia ($\chi^2(2, N = 256) = 9.88, p < 0.01, \phi = 0.20$), with a small effect size (see Figure 7). Post hoc analysis with Bonferroni correction showed that Indian participants who opposed same-sex marriage found human-written comments significantly more aligned with their values (52.5%) than expected. They explained that their chosen comments asserted traditional views of marriage between man and woman, which also aligned with their socio-cultural values. An Indian participant commented:

I hold the view that marriage, by its traditional and biological definition, is a union between a man and a woman, serving not only as a bond of companionship but also as a natural framework for procreation and the raising of children. While I acknowledge that same-sex couples should have the right to live together and form committed partnerships, I do not personally support the concept of redefining marriage to include such unions.

In my understanding, marriage is both a social and cultural institution historically rooted in the complementary union of male and female sexes, encompassing both emotional companionship and the potential for the continuation of family lineage.

Although Indian participants supporting same-sex marriage selected LLM-rewritten comments more (67.5%) than human-written ones, the distribution was not significant. Similarly, participants who held supportive views of Islam also considered LLM-rewritten comments to be more aligned with their values (68.3%); but without significant difference. They shared that they chose comments which reflected equality, harmony, liberal values, and tolerance of diverse groups. One of the Indian participants wrote:

these comments follow equality and harmony which i want in our selves so that we as a society make more good and loveable place for us to live.

On the other hand, Indian participants who were against Islam did not show any significant preference for either types of comments. They shared that although they did not support Islam, they also did not find it appropriate to generalize against a whole community. One such response said:

The comments i selected agree with the fact that Muslims have been connected to a lot of negative behaviors but also recognizes that it is not good to generalize and stereotype but instead look at the person instead of just their behaviour. No point hating someone just because they're muslim but rather look at individual behaviour before forming an opinion.

Again, Indian participants who reported being neutral on these issues did not exhibit any significant preference for either types of comments.

Potential Reasons Behind Participants' Preferences for Different Comments. In 4.2 (see Table 4), we observed that LLM downplayed Traditional values and emphasized equality and tolerance (Universalism) towards same-sex marriage when rewriting comments from participants, who were against such marriage. This shift in value might explain why Indian and American participants, opposing same-sex marriage found human-written comments more representative of their traditional viewpoints on marriage.

Similarly, LLM downplayed Security concern around extremism and highlighted tolerance (Universalism) of diverse religions when rewriting constructive comments on Islamophobic thread (see Table 4). These changes in value framing might have impacted American participants to choose human-written comments that captured their criticism of Islam well. Moreover, as noted in 4.2, for comments opposing same-sex marriage and Islam, the value shifts introduced by LLM often changed their stance, making the rewritten versions more neutral or supportive. This may also explain why participants with opposing views felt human-written comments were more aligned with their values.

On the other hand, LLM did not significantly change underlying value expressions when rewriting comments which were supportive of same-sex marriage and Islam (see Table 4). This might explain why Indian participants with favorable stance on these cases, did not find one type of comment more aligned with their values than

the other. However, the American participants supporting same-sex marriage and Islam preferred LLM-rewritten comments significantly more than the actual human-written comments. This might be because more American participants held supportive views of same-sex marriage and Islam than the Indians (see Table 6).

Taken together, these findings suggest that LLMs' tendency to emphasize prosocial values such as Benevolence and Universalism when rewriting constructive comments may not resonate with individuals who prioritize different values (e.g., Tradition, Security) regarding issues like same-sex marriage and Islam. By contrast, those more tolerant of diverse religious and gender groups may find LLM rewriting more persuasive, since the model reflects values favored by them.

5 Discussion

Through experiments with Indian and American participants, we show that LLMs systematically shift responses toward prosocial values such as Universalism and Benevolence, while downplaying Conservative values (e.g., Security, Tradition, Conformity) originally expressed in people's comments on value-laden topics. This pattern resonates with prior research showing that LLMs prioritize prosocial values in moral dilemmas [70] and decision-making tasks [50, 98]. Our study extends this line of work in several ways. First, we demonstrate that **value homogenization** by LLMs is systemic and extends beyond abstract moral decision-making tasks to practical applications, such as facilitating online dialogue on value-laden topics. Our results indicate that LLMs impose their own value systems, often overriding the values through which people originally framed their arguments. Second, we reveal that this prosocial shift can alter the stance itself. Comments opposing same-sex marriage or Islam are frequently rewritten in ways that sound more neutral or even supportive, highlighting a subtle yet consequential form of value-based reframing not documented in prior work. Third, we show how these shifts affect perceived alignment. Participants opposing same-sex marriage or Islam preferred human-written comments, which aligned more with their Conservative values, while those who were supportive favored LLM rewrites for their emphasis on equality and inclusion. Drawing on these findings, we discuss the tensions and inevitability of value homogenization by LLMs and its broader impact on online discourses on value-laden topics.

5.1 Conservative Bias and Prosocial Leanings of LLMs

Prior research has documented that LLMs consistently lean towards progressive, left-libertarian, and prosocial viewpoints even when they are prompted with Conservative perspectives [34, 64, 68, 71]. These prosocial tendencies persist across diverse models [60], prompting strategies [27], and cultural contexts [10, 56]—suggesting a systemic pattern rather than isolated bias. Alignment strategies, particularly in the post-training phase, are designed to minimize harm [33], yet the very definition of “harm” is deeply value-laden, political, and often contested [29]. Current industry policies frame alignment around reducing bias, prejudice, and exclusion, thereby guiding models toward progressive, inclusive, and human-rights-oriented value systems [9, 30], which critics often

negatively portray as “woke” [42]. In contrast, proponents argue that prosocial leanings by LLMs provide a practical and interpretable proxy for the model safety [93]. Neuman et al. [60] demonstrated that when models are probed about their prosocial leanings, they attribute them to both liberal-leaning training data and fine-tuning by human annotators who prioritize compassion, equality, and inclusion. Thus when confronted with trade-offs, the models typically favor minimizing suffering and protecting the dignity of vulnerable populations [60].

This alignment trajectory reveals a critical trade-off between model safety and value pluralism. On one hand, prosocial alignment advances goals such as preventing harms and bias towards marginalized communities, thereby supporting ethical standards that are widely regarded as necessary for building safe and responsible AI systems [36, 79, 88, 90]. However, this orientation can marginalize Conservative viewpoints, particularly on value-laden issues which LLMs not only deprioritize but also diminish people’s opposing stances when rewriting their comments.

Yet, framing value homogenization simply as bias or underperformance risks obscuring the deeper issue [33, 60]. Alignment is never a purely technical process but an exercise in normative judgment about what counts as harm, fairness, or bias. Current alignment practices tend to privilege ethical frameworks more closely associated with progressive traditions [9, 30], thereby reinforcing a particular vision of “prosociality.” The dilemma, then, is whether alignment should prioritize ethical standards that safeguard vulnerable communities (e.g., racial, gender, religious, or ethnic minorities) or whether it should accommodate a broader spectrum of socio-cultural, political, and moral values, even when those may conflict with the liberal notion of harm-reduction goals.

5.2 Impact on Online Discourse on Value-Laden Topics

Recently, online platforms such as BlueSky [38], Messenger [55], and LinkedIn [49] have begun using LLMs to help users rewrite their messages and reduce toxic content. As these models become rapidly integrated into online communication, it is important to examine how they may shape conversations, particularly on value-laden topics.

Our results show that individuals with prosocial and progressive views perceived LLM-rewritten comments as more aligned with their values, largely because the rewrites emphasized equality and inclusion. This suggests that such models may support these users in responding constructively to problematic speech directed toward religious or gender minorities, and may also encourage empathy and prosocial attitudes among those exposed to such content [47, 67]. In contrast, our participants who opposed same-sex marriage or Islam found human-written comments more aligned with their values since LLM rewrites consistently diminished Conservative values and distorted their position. Thus, conservative users may find LLM rewrites unhelpful if their viewpoints are not accommodated and get misrepresented.

These dynamics become even more complicated in cross-cultural contexts. When tasked with “constructive reframing,” LLMs in our study frequently elevated liberal-prosocial values, such as gender and religious inclusion. These values are more pronounced in the

normative frameworks of average Western societies, who are usually more accepting of same-sex marriage and show less religious hostility [19, 66]. However, in traditional conservative societies where collective obligations, hierarchy, and conformity play a more central role, such reframing can be experienced as alien, illegitimate, or even disrespectful. What is presented as neutral, constructive, or less toxic may therefore operate as a form of moral homogenization.

These asymmetries can contribute to “quality-of-service harms” [77], as models may systematically underserve users with Conservative value systems, potentially deepening mistrust, exclusion, and alienation in digital spaces. In fact, research shows that undesirable exposure to ideologies that do not align with one’s values [9] and experiences of socio-political alienation can drive conservatives toward more polarized positions [1, 2, 54] and far-right platforms such as Truth Social [16]. Thus, LLM-driven prosocial value homogenization might lead to uneven benefits and inadvertently push users to more radical viewpoints.

To minimize such inadvertent risks, first platforms should disclose the value orientations embedded in their models. Transparency about which ethical frameworks guide rewrites; for example, inclusivity or harm reduction—can help users understand why their writing is transformed in particular ways. Such disclosure will not only mitigate perceptions of hidden bias but also foster trust in AI-mediated communication.

Second, instead of relying on one-size-fits-all prosocial models, platforms could experiment with personalized or role-playing LLMs [25, 87, 97, 99]. Prior research shows that in role-playing contexts LLM-generated outputs balance both prosocial and individualistic values in ways that mirror human preferences [62, 89, 91]. Thus, these systems could allow users to express their values and viewpoints constructively without being steered toward a single normative framework that may conflict with their socio-cultural or moral standpoints. However, personalization raises new risks: it may reinforce confirmation biases or deepen echo chambers and polarization [43, 51]. Designers must therefore balance personalization with mechanisms that allow users to configure model outputs based on their preferences within certain safety and legal limits [43].

Third, platforms could give users greater agency in controlling the value framings of rewrites. Adjustable settings or sliders [65] would allow users to calibrate how much emphasis is placed on different values, making alignment an explicit and participatory process rather than a hidden one.

5.3 Limitations and Future Work

Our work has some limitations. First, we focused on two value-laden topics: Islamophobia and homophobia. Thus, our findings may not generalize to other value-laden topics, such as abortion, gun regulation, climate change, or vaccination. Future work should explore if value homogenization by LLMs extends to diverse value-laden topics. Second, in our study we only experimented with GPT-4. Although prosocial leanings have been documented across multiple models [60, 70, 79], further work is needed to examine whether different models also shift values when performing constructive rewriting. Third, our study recruited participants only from India and the United States. The absence of observed cross-cultural differences in constructive comments on value-laden topics may be due

to the limited sample size. Future work should examine how value homogenization unfolds across a wider range of cultural contexts and broad samples. Finally, we only focused on constructive writing and value expressions in English. Research shows that while LLMs tend to produce similar value orientations across many European languages (e.g., French, German), outputs can diverge sharply when comparing languages from more distant cultural and linguistic traditions, such as English and Chinese [18]. Future work should investigate value homogenization across multilingual contexts.

6 Conclusion

Through experiments with 347 participants from India and the US, this work unpacks how LLMs contribute to value homogenization when facilitating constructive discourse on sensitive, value-laden topics. Our findings show that LLMs systematically downplay Conservative values, foreground prosocial values, and at times change the stance in comments from opposing to supporting on value-laden topics. These dynamics raise pressing ethical tensions and moral dilemmas for the deployment of LLMs in cross-cultural, value-sensitive contexts. Addressing these challenges would require our socio-technical systems to be more attuned to and mindful towards diverse human perspectives on contested value-laden issues.

References

- [1] Gilbert Abcarian and Sherman M Stanage. 1965. Alienation and the radical right. *The Journal of Politics* 27, 4 (1965), 776–796.
- [2] Joel D Aberbach. 1969. Alienation and political behavior. *American Political Science Review* 63, 1 (1969), 86–99.
- [3] Dhruv Agarwal, Mor Naaman, and Aditya Vashistha. 2025. AI Suggestions Homogenize Writing Toward Western Styles and Diminish Cultural Nuances. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, Article 1117, 21 pages.
- [4] Dhruv Agarwal, Anya Shukla, Sunayana Sitaram, and Aditya Vashistha. 2025. Fluent but Culturally Distant: Can Regional Training Teach Cultural Understanding? <https://arxiv.org/pdf/2505.21548>
- [5] Yamen Ajjour, Milad Alshomary, Henning Wachsmuth, and Benno Stein. 2019. Modeling Frames in Argumentation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 2922–2932.
- [6] Beverlee B Anderson. 2012. Indian cultural values and economic development: an exploratory study. *World Journal of Entrepreneurship, Management and Sustainable Development* 8, 2/3 (2012), 194–201.
- [7] Lisa P Argyle, Christopher A Bail, Ethan C Busby, Joshua R Gubler, Thomas Howe, Christopher Rytting, Taylor Sorensen, and David Wingate. 2023. Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale. *Proceedings of the National Academy of Sciences* 120, 41 (2023), e2311627120.
- [8] Katie Atkinson and Trevor Bench-Capon. 2021. Value-based argumentation. *Journal of Applied Logics* 8, 6 (2021), 1543–1588.
- [9] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. <https://arxiv.org/pdf/2204.05862>
- [10] Jan Batzner, Volker Stocker, Stefan Schmid, and Gjergji Kasneci. 2024. German-partiesqa: Benchmarking commercial large language models for political bias and sycophancy. <https://arxiv.org/pdf/2407.18008>
- [11] Amanda Baughan, Justin Petelka, Catherine Jaekyung Yoo, Jack Lo, Shiyue Wang, Amulya Paramasivam, Ashley Zhou, and Alexis Hiniker. 2021. Someone Is Wrong on the Internet: Having Hard Conversations in Online Spaces. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 156 (apr 2021), 22 pages.
- [12] Trevor JM Bench-Capon. 2003. Persuasion in practical argument using value-based argumentation frameworks. *Journal of Logic and Computation* 13, 3 (2003), 429–448.
- [13] Shreyan Biswas, Alexander Erlei, and Ujwal Gadiraju. 2025. Mind the Gap! Choice Independence in Using Multilingual LLMs for Persuasive Co-Writing Tasks in Different Languages. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, Article 937, 20 pages.
- [14] Gustavo A Bodanza and Esteban Freidin. 2023. Confronting value-based argumentation frameworks with people’s assessment of argument strength. *Argument & Computation* 14, 3 (2023), 247–273.
- [15] Gabriel Borelli. 2022. About six-in-ten Americans say legalization of same-sex marriage is good for society. Retrieved October 18, 2024 from <https://www.pewresearch.org/short-reads/2022/11/15/about-six-in-ten-americans-say-legalization-of-same-sex-marriage-is-good-for-society/>
- [16] Joseph Brown and Ibrahim Baggili. 2023. Retruth Reconnaissance: A Digital Forensic Analysis of Truth Social. In *International Conference on Digital Forensics and Cyber Crime*. Springer, New York, 96–113.
- [17] Fan Bu, Zheng Wang, Siyi Wang, and Ziyao Liu. 2025. An investigation into value misalignment in LLM-generated texts for cultural heritage. <https://arxiv.org/pdf/2501.02039>
- [18] Samuel Cahyawijaya, Delong Chen, Yejin Bang, Leila Khalatbari, Bryan Wilie, Ziwei Ji, Etsuko Ishii, and Pascale Fung. 2024. High-dimension human value representation in large language models. <https://arxiv.org/pdf/2404.07900>
- [19] Pew Research Center. 2017. How the U.S. general public views Muslims and Islam. Retrieved October 18, 2024 from <https://www.pewresearch.org/religion/2017/07/26/how-the-u-s-general-public-views-muslims-and-islam/>
- [20] Pew Research Center. 2024. Religious Restrictions Around the World. <https://www.pewresearch.org/religion/feature/religious-restrictions-around-the-world/>
- [21] Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing Things from a Different Angle: Discovering Diverse Perspectives about Claims. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 542–557.
- [22] Stephanie Cutler, Yu Xia, and Kacey Beddoes. 2022. A Growth Mindset for Peer Review: Guidelines for writing constructive peer reviews.
- [23] Christine De Kock, Tom Stafford, and Andreas Vlachos. 2022. How to disagree well: Investigating the dispute tactics used on Wikipedia. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 3824–3837.
- [24] Christine De Kock and Andreas Vlachos. 2021. I Beg to Differ: A study of constructive disagreement in online conversations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Paola Merlo, Jorg Tiedemann, and Reut Tsarfay (Eds.). Association for Computational Linguistics, Online, 2017–2027.
- [25] Joel Eapen and VS Adhithyan. 2023. Personalization and customization of llm responses. *International Journal of Research Publication and Reviews* 4, 12 (2023), 2617–2627.
- [26] Robert M Entman. 1993. Framing: Towards clarification of a fractured paradigm. *McQuail’s reader in mass communication theory* 390 (1993), 397.
- [27] Mats Faulborn, Indira Sen, Max Pellert, Andreas Spitz, and David Garcia. 2025. Only a Little to the Left: A Theory-grounded Measure of Political Bias in Large Language Models. <https://arxiv.org/pdf/2503.16148>
- [28] Dennis Friess and Christiane Eilders. 2015. A systematic review of online deliberation research. *Policy & Internet* 7, 3 (2015), 319–339.
- [29] Iason Gabriel. 2020. Artificial intelligence, values, and alignment. *Minds and machines* 30, 3 (2020), 411–437.
- [30] Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. 2022. Improving alignment of dialogue agents via targeted human judgements. <https://arxiv.org/pdf/2209.14375>
- [31] Jarod Govers, Eduardo Velloso, Vassilis Kostakos, and Jorge Goncalves. 2024. AI-Driven Mediation Strategies for Audience Depolarisation in Online Debates. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI ’24)*. Association for Computing Machinery, New York, USA, Article 803, 18 pages.
- [32] Selin Gurgun, Emily Arden-Close, John McAlaney, Keith Phalp, and Raian Ali. 2023. Can We Re-design Social Media to Persuade People to Challenge Misinformation? An Exploratory Study. In *International Conference on Persuasive Technology*. Springer, Germany, 123–141.
- [33] Thilo Hagendorff. 2025. On the Inevitability of Left-Leaning Political Bias in Aligned Language Models. <https://arxiv.org/pdf/2507.15328>
- [34] Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. 2023. The political ideology of conversational AI: Converging evidence on ChatGPT’s pro-environmental, left-libertarian orientation. <https://arxiv.org/pdf/2301.01768>
- [35] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2017. SpaCy: Industrial-strength natural language processing in Python.
- [36] Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. 2023. Ai alignment: A comprehensive survey. <https://arxiv.org/pdf/2310.19852>

- [37] Liwei Jiang, Taylor Sorensen, Sydney Levine, and Yejin Choi. 2024. Can language models reason about individualistic human values and preferences? <https://arxiv.org/pdf/2410.03868>
- [38] Fay Johnson. 2024. CLRSKY Launches, A New AI-Powered Civility Overlay For Bluesky. Retrieved August 26, 2025 from <https://www.prnewswire.com/news-releases/clrsky-launches-a-new-ai-powered-civility-overlay-for-bluesky-302429232.html>
- [39] Gauri Kambhatla, Matthew Lease, and Ashwin Rajadesingan. 2024. Promoting Constructive Deliberation: Reframing for Receptiveness. <https://arxiv.org/pdf/2405.15067>
- [40] Mehdi Khamassi, Marceau Nahon, and Raja Chatila. 2024. Strong and weak alignment of large language models with human values. *Scientific Reports* 14, 1 (2024), 19399.
- [41] Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 2022. Identifying the Human Values behind Arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 4459–4471.
- [42] Tae Wan Kim, Thomas Donaldson, and John Hooker. 2018. Mimetic vs anchored value alignment in artificial intelligence. <https://arxiv.org/pdf/1810.11116>
- [43] Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A Hale. 2024. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence* 6, 4 (2024), 383–392.
- [44] Varada Kolhatkar and Maite Taboada. 2017. Constructive Language in News Comments. In *Proceedings of the First Workshop on Abusive Language Online*, Zeerak Waseem, Wendy Hui Kyong Chung, Dirk Hovy, and Joel Tetreault (Eds.). ACL, Vancouver, Canada, 11–17.
- [45] Varada Kolhatkar and Maite Taboada. 2017. Using New York Times Picks to Identify Constructive Comments. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, Octavian Popescu and Carlo Strapparava (Eds.). ACL, Copenhagen, Denmark, 100–105.
- [46] Varada Kolhatkar, Nithum Thain, Jeffrey Sorensen, Lucas Dixon, and Maite Taboada. 2020. Classifying constructive comments. <https://arxiv.org/pdf/2004.05476>
- [47] Sara Konrath, Emily Falk, Andrea Fuhrel-Forbis, Mary Liu, James Swain, Richard Tolman, Rebecca Cunningham, and Maureen Walton. 2015. Can text messages increase empathy and prosocial behavior? The development and initial validation of text to connect. *PLoS one* 10, 9 (2015), e0137585.
- [48] Marina Kouzakova, Naomi Ellemers, Fieke Harinck, and Daan Scheepers. 2012. The Implications of Value Conflict: How Disagreement on Values Affects Self-Involvement and Perceived Common Ground. *Personality and Social Psychology Bulletin* 38, 6 (2012), 798–807.
- [49] LinkedIn. 2024. Create posts with our AI-powered writing tool. Retrieved August 26, 2025 from <https://www.linkedin.com/help/linkedin/answer/a1517763>
- [50] Xuelin Liu, Pengyuan Liu, and Dong Yu. 2025. What's the most important value? INVP: INvestigating the Value Priorities of LLMs through Decision-making in Social Scenarios. In *Proceedings of the 31st International Conference on Computational Linguistics*. ACL, Abu Dhabi, UAE, 4725–4752.
- [51] Tony Ma. 2024. LLM Echo Chamber: personalized and automated disinformation. <https://arxiv.org/pdf/2409.16241>
- [52] Gregory R Maio, Ulrike Hahn, John-Mark Frost, Toon Kuppens, Nadia Rehman, and Shanmukh Kamble. 2014. Social values as arguments: similar is convincing. *Frontiers in psychology* 5 (2014), 829.
- [53] Lilliana Mason. 2015. "I disrespectfully agree": The differential effects of partisan sorting on social and issue polarization. *American journal of political science* 59, 1 (2015), 128–145.
- [54] Herbert McClosky and Dennis Chong. 1985. Similarities and differences between left-wing and right-wing radicals. *British Journal of Political Science* 15, 3 (1985), 329–363.
- [55] Meta. 2024. Rewrite again and again until you've got it just right with Meta AI. Retrieved August 26, 2025 from <https://www.facebook.com/reel/1809530026241575>
- [56] Fabio YS Motoki, Valdemar Pinho Neto, and Victor Rangel. 2025. Assessing political bias and value misalignment in generative artificial intelligence. *Journal of Economic Behavior & Organization* 234 (2025), 106904.
- [57] Jimin Mun, Cathy Buerger, Jenny T Liang, Joshua Garland, and Maarten Sap. 2024. Counterspeakers' Perspectives: Unveiling Barriers and AI Needs in the Fight against Online Hate. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). ACM, New York, USA, Article 742, 22 pages.
- [58] Inderjeet Nair and Lu Wang. 2025. Do Language Models Think Consistently? A Study of Value Preferences Across Varying Response Lengths. <https://arxiv.org/pdf/2506.02481>
- [59] Courtney Napoles, Joel Tetreault, Aashish Pappu, Enrica Rosato, and Brian Provenza. 2017. Finding Good Conversations Online: The Yahoo News Annotated Comments Corpus. In *Proceedings of the 11th Linguistic Annotation Workshop*, Nathan Schneider and Nianwen Xue (Eds.). Association for Computational Linguistics, Valencia, Spain, 11 pages.
- [60] W Russell Neuman, Chad Coleman, Ali Dasdan, Safinah Ali, Manan Shah, and Kund Meghani. 2025. "Amazing, They All Lean Left"—Analyzing the Political Temperaments of Current LLMs. <https://arxiv.org/pdf/2507.08027>
- [61] Vlad Niculae and Cristian Danescu-Niculescu-Mizil. 2016. Conversational Markers of Constructive Discussions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Kevin Knight, Ani Nenkova, and Owen Rambow (Eds.). Association for Computational Linguistics, San Diego, California, 568–578.
- [62] Hakim Norhashim and Jungpil Hahn. 2024. Measuring human-ai value alignment in large language models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Vol. 7. AAAI, ACM, 1063–1073.
- [63] Inyoung Park, Daeho Lee, and Young June Sah. 2023. Under watching eyes in news comment sections: effects of audience cue on self-awareness and commenting behaviour. *Behaviour & Information Technology* 42, 13 (2023), 2279–2295.
- [64] Pagnarasmy Pit, Xingjun Ma, Mike Conway, Qingyu Chen, James Bailey, Henry Pit, Putrasmy Keo, Watey Diep, and Yu-Gang Jiang. 2024. Whose side are you on? Investigating the political stance of large language models. <https://arxiv.org/pdf/2403.13840>
- [65] Jon Porter. 2021. Today I learned about Intel's AI sliders that filter online gaming abuse. <https://www.theverge.com/2021/4/8/22373290/intel-bleep-ai-powered-abuse-toxicity-gaming-filters>
- [66] Jacob Poushter and Nicholas Kent. 2017. The Global Divide on Homosexuality Persists. Retrieved September 8, 2025 from <https://www.pewresearch.org/global/2020/06/25/global-divide-on-homosexuality-persists/>
- [67] Sara Prot, Douglas A Gentile, Craig A Anderson, Kanae Suzuki, Edward Swing, Kam Ming Lim, Yukiko Horiuchi, Margareta Jelic, Barbara Krahé, Wei Liuqing, et al. 2014. Long-term relations among prosocial-media use, empathy, and prosocial behavior. *Psychological science* 25, 2 (2014), 358–368.
- [68] Luca Rettenberger, Markus Reischl, and Mark Schutera. 2025. Assessing political bias in large language models. *Journal of Computational Social Science* 8, 2 (2025), 1–17.
- [69] Milton Rokeach. 1973. *The nature of human values*. Free press, New York.
- [70] Giuseppe Russo, Debora Nozza, Paul Röttger, and Dirk Hovy. 2025. The Pluralistic Moral Gap: Understanding Judgment and Value Differences between Humans and Large Language Models. <https://arxiv.org/pdf/2507.17216>
- [71] Jérôme Rutinowski, Sven Franke, Jan Endendyk, Ina Dormuth, Moritz Roidl, and Markus Pauly. 2024. The self-perception and political biases of ChatGPT. *Human Behavior and Emerging Technologies* 2024, 1 (2024), 7115633.
- [72] Richard M Ryckman and Diane M Houston. 2003. Value priorities in American and British female and male university students. *The Journal of Social Psychology* 143, 1 (2003), 127–138.
- [73] Shalom H. Schwartz. 1994. Are There Universal Aspects in the Structure and Contents of Human Values? *Journal of Social Issues* 50, 4 (1994), 19–45.
- [74] Shalom H. Schwartz, Gila Melech, Arielle Lehmann, Steven Burgess, Mari Harris, and Vicki Owens. 2001. Extending the Cross-Cultural Validity of the Theory of Basic Human Values with a Different Method of Measurement. *Journal of Cross-Cultural Psychology* 32, 5 (2001), 519–542.
- [75] Robin Segerer. 2025. Cultural Value Alignment in Large Language Models: A Prompt-based Analysis of Schwartz Values in Gemini, ChatGPT, and DeepSeek. <https://arxiv.org/pdf/2505.17112>
- [76] Farhana Shahid, Maximilian Dittgen, Mor Naaman, and Aditya Vashistha. 2025. Examining Human-AI Collaboration for Co-Writing Constructive Comments Online. *Proc. ACM Hum.-Comput. Interact.* 9, CSCW410, Article 410 (2025), 33 pages.
- [77] Renee Shelby, Shalaleh Rismani, Kathryn Henne, Ajung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. 2023. Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (Montréal, QC, Canada) (AI/ES '23). Association for Computing Machinery, New York, NY, USA, 723–741.
- [78] Hua Shen, Nicholas Clark, and Tanushree Mitra. 2025. Mind the Value-Action Gap: Do LLMs Act in Alignment with Their Values? <https://arxiv.org/pdf/2501.15463>
- [79] Hua Shen, Tiffany Kneare, Reshmi Ghosh, Yu-Ju Yang, Nicholas Clark, Tanushree Mitra, and Yun Huang. 2024. ValueCompass: A Framework for Measuring Contextual Value Alignment Between Human and LLMs. <https://arxiv.org/pdf/2409.09586>
- [80] Momin N Siddiqui, Roy D Pea, and Hari Subramonyam. 2025. Script&Shift: A Layered Interface Paradigm for Integrating Content Development and Rhetorical Strategy with LLM Writing Assistants. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, Article 532, 19 pages.
- [81] Paras Nath Singh, Sophia Chang Huang, and George G Thompson. 1962. A comparative study of selected attitudes, values, and personality characteristics of American, Chinese, and Indian students. *The Journal of Social Psychology* 57, 1 (1962), 123–132.
- [82] Nicholas Sukiennik, Chen Gao, Fengli Xu, and Yong Li. 2025. An evaluation of cultural value alignment in llm. <https://arxiv.org/pdf/2504.08863>

- [83] Abhay Sukumaran, Stephanie Vezich, Melanie McHugh, and Clifford Nass. 2011. Normative influences on thoughtful online participation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) (CHI '11). ACM, New York, USA, 3401–3410.
- [84] Norman D Sundberg, Pritam K Rohila, and Leona E Tyler. 1970. Values of Indian and American adolescents. *Journal of Personality and Social Psychology* 16, 3 (1970), 374.
- [85] Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS nexus* 3, 9 (2024), 346.
- [86] Michael Henry Tessler, Michiel A. Bakker, Daniel Jarrett, Hannah Sheahan, Martin J. Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tantum Collins, David C. Parkes, Matthew Botvinick, and Christopher Summerfield. 2024. AI can help humans find common ground in democratic deliberation. *Science* 386, 6719 (2024), eadq2852.
- [87] Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in llms: A survey of role-playing and personalization. <https://arxiv.org/pdf/2406.01171>
- [88] Qiaosi Wang, Michael Madaio, Shaun Kane, Shivani Kapania, Michael Terry, and Lauren Wilcox. 2023. Designing Responsible AI: Adaptations of UX Practice to Meet Responsible AI Challenges. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 249, 16 pages.
- [89] Ye Wang, Tong Li, Meixuan Li, Ziyue Cheng, Ge Wang, Hanyue Kang, Yaling Deng, Hongjiang Xiao, and Yuan Zhang. 2025. RVBench: Role Values Benchmark for role-playing LLMs. *Computers in Human Behavior: Artificial Humans* 5 (2025), 100184.
- [90] Zijie J. Wang, Chinmay Kulkarni, Lauren Wilcox, Michael Terry, and Michael Madaio. 2024. Farsight: Fostering Responsible AI Awareness During AI Application Prototyping. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 976, 40 pages.
- [91] Yuemei Xu, Ling Hu, and Zihan Qiu. 2024. ValueCSV: Evaluating Core Socialist Values Understanding in Large Language Models. In *CCF International Conference on Natural Language Processing and Chinese Computing*. Springer, New York, 346–358.
- [92] Jing Yao, Xiaoyuan Yi, Xiting Wang, Jindong Wang, and Xing Xie. 2023. From Instructions to Intrinsic Human Values—A Survey of Alignment Goals for Big Models. <https://arxiv.org/pdf/2308.12014>
- [93] Haoran Ye, Yuhang Xie, Yuanyi Ren, Hanjun Fang, Xin Zhang, and Guojie Song. 2025. Measuring human and ai values based on generative psychometrics with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. AAAI, Pennsylvania, USA, 26400–26408.
- [94] Chao Zhang, Kexin Ju, Peter Bidoshi, Yu-Chun Grace Yen, and Jeffrey M. Rzeszotarski. 2025. Friction: Deciphering Writing Feedback into Writing Revisions through LLM-Assisted Reflection. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, Article 935, 27 pages.
- [95] Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018. Conversations Gone Awry: Detecting Early Signs of Conversational Failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 1350–1361.
- [96] Zheng Zhang, Jie Gao, Ranjodh Singh Dhaliwal, and Toby Jia-Jun Li. 2023. VISAR: A Human-AI Argumentative Writing Assistant with Visual Programming and Rapid Draft Prototyping. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) (UIST '23). ACM, New York, USA, Article 5, 30 pages.
- [97] Zhehao Zhang, Ryan A Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, et al. 2024. Personalization of large language models: A survey. <https://arxiv.org/pdf/2411.00027>
- [98] Zhaowei Zhang, Ceyao Zhang, Nian Liu, Siyuan Qi, Ziqi Rong, Song-Chun Zhu, and Yaodong Yang. 2025. Heterogeneous value alignment evaluation for large language models. In *International Conference on Artificial General Intelligence*. Springer, New York, 381–392.
- [99] Thomas P Zollo, Andrew Wei Tung Siah, Naimeng Ye, Ang Li, and Hongseok Namkoong. 2024. Personallm: Tailoring llms to individual preferences. <https://arxiv.org/pdf/2409.20296>

A Appendix

Table 7: Title of homophobic and Islamophobic threads that participants reviewed in Phase 2.

Thread Topic	Homophobia	Islamophobia
India	1. Which gay man, without a uterus, has a menstrual cycle? 2. Should gay marriage be legalised in India?	1. Islamic Takeover of India by 2047 2. Why my otherwise liberal family has a problem with Islam
US	1. Speaker Mike Johnson's Obsession With Gay Sex 2. Lib thinks republicans are out to get them because they're gay lol	1. I am Islamophobic 2. Islamophobia is a great thing and I'm tired of being called racist for it

Table 8: Mann-Whitney tests with Bonferroni corrections comparing the linguistic features of constructiveness between human-written and LLM-rewritten comments on value-laden topics.

Feature	Statistics	Human-written comment (Average)	LLM-rewritten comment (Average)
Length	W=3009.5, Z=-9.1, p<0.000001, r=0.44	62.1	80.1
Readability	W=167.5, Z=-12.37, p<0.000001, r=0.6	7.4	15.3
Named entity	W=4621.5, Z=3.95, p<0.00001, r=0.19	1.9	1.5
Discourse connective	W=6811, Z=-2.86, p<0.01, r=0.14	5.99	6.61
Stance adverbial	-	0.59	0.40
Reasoning verbs and modals	W=1355, Z=-10.48, p<0.000001, r=0.51	4.5	7.34
Root clauses	W=2423, Z=-8.34, p<0.000001, r=0.41	3.5	4.6
Politeness	W=6161.5, Z=-4.74, p<0.00001, r=0.23	12.73	14.73