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
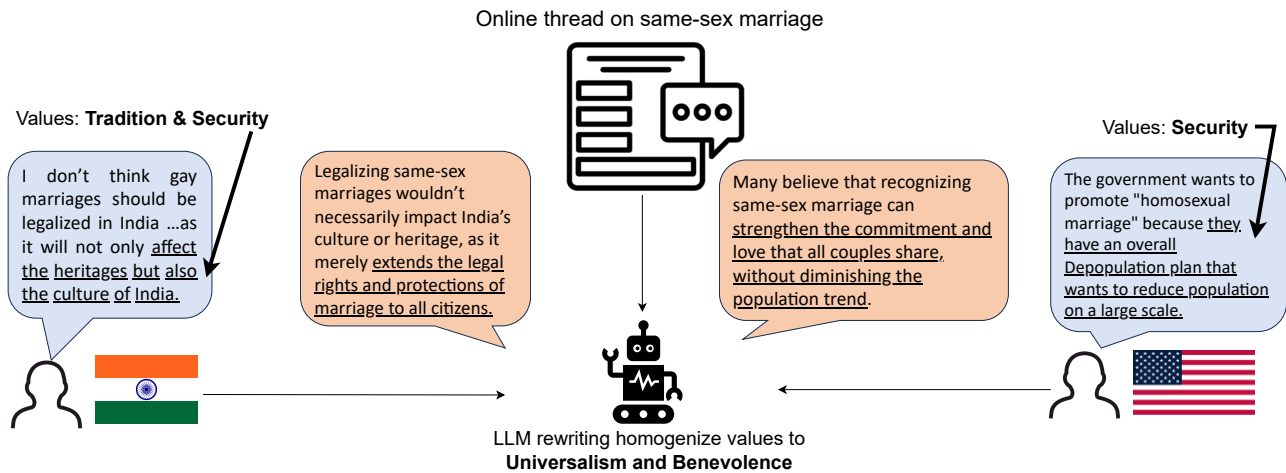adityav@cornell.edu

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 these models negotiate and shape underlying values when reframing people's arguments on value-laden topics. We conducted experiments with 465 participants from India and the United States, who wrote comments on homophobic and Islamophobic threads, and reviewed human-written and LLM-rewritten constructive 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 value homogenization in LLM-mediated prosocial discourse runs the risk of marginalizing conservative viewpoints
on value-laden topics and may inadvertently shape the dynamics of online discourse.

CCS Concepts

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

Keywords

values, constructive argument, stance, HAI misalignment

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, 35]. While researchers define constructive disagreement as civil dialogue oriented toward finding common ground [31, 48], what qualifies as 'constructive' is often contested. Because, people tend to embrace arguments that affirm their own values and resist those that challenge them [52]. Here, values refer to human beliefs regarding

desirable end states or modes of conduct [74]. Different value priorities both within and across cultures frequently fuel conflict on divisive issues, which often are value-laden [45, 52, 81].

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, 43, 84, 95]. 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 [84]. For example, humans tend to value direct, logical arguments, while LLMs favor polite, indirect arguments as constructive [84].

Although *values* greatly shape how people frame their arguments on divisive topics [5, 23], 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, 89, 105, 107] and facilitating civil discourse online [42, 53, 59], 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 explore:

RQ1: What values do people from different cultures emphasize when writing comments on value-laden topics?

RQ2: What values do LLMs emphasize when rewriting people's comments constructively?

RQ3: How might differences between human–LLM value framings affect which comments people prefer?

To find answer, we conducted a three-phase study with 465 participants from India and the United States, recruited via Prolific. In Phase 1, participants responded to the same homophobic and Islamophobic threads by writing either a regular comment (baseline condition) or a constructive comment (constructive condition). 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 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 [78]. 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 one better aligned with their own values.

In response to RQ1, we found that people's stances on homophobia, rather than demographic factors, such as culture, age, gender, or religion, influenced what values they expressed when commenting on homophobic threads. Participants opposing same-sex marriage framed their arguments by emphasizing 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 frequently 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 different people, those opposing same-sex marriage and Islam found human-written comments more aligned with their personal 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 tension when LLM distorts 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 first by 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 [49, 92] and evidence-based [48]. Other definitions focus on conversational outcomes, such as dispute resolution [26], finding common ground [31], improved team performance [66], or the emergence of new solutions [31]. In online discourse, constructiveness manifests through specific linguistic features, such as longer content [49, 92], richer argumentative markers (e.g., discourse connectives, stance adverbials, reasoning verbs, modals, and root clauses) [48, 63], greater readability [50], greater use of named entities [50, 66], more politeness [26, 63, 106], and less hedging [25, 63, 66].

Given the tension, time, and effort involved in expressing constructive disagreement on value-laden topics [24, 35, 61], researchers have explored if LLMs can support people during this process. For example, Shahid et al. [84] 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. [43] 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. [34] 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 [28]. 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) [56].

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 [52]. 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, 97, 100]. 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 [54, 75, 83, 88, 103, 109]. These preferences are usually consistent across both short and long-form responses generated by LLMs on value-laden questions [62].

In contrast, few studies [67, 99, 102] 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 [44]. 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 [87].

Different value priorities across different cultural contexts further complicate the process. For example, Bu et al. [19] 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, 41, 87, 88, 91, 94]. 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, 83, 91].

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 [42, 53, 59], 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). We selected these countries as representatives of individualist (US) and collectivist (India) cultures, who are known to have different value priorities [90]. In the first phase, we examined which values participants emphasize when writing comments 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 what 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 Comments

Study Design. We recruited 117 Indian and 117 American participants on Prolific. Participants were randomly assigned to one of two conditions: a baseline condition or a constructive condition. In the baseline condition, participants were asked to write a comment on a value-laden topic. In the constructive condition, they were explicitly instructed to write a *constructive* comment. 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 homophobia and Islamophobia because these issues are highly value-laden and socially divisive in both India and the US. Apart from their cross-cultural relevance, research shows that discussions involving anti-Muslim and anti-LGBTQ sentiment tend to elicit greater toxicity than non-identity related controversial topics, such as climate change and vaccination [77, 82], making these issues suitable for our study.

We first asked participants about their demographics, such as age, gender, and religion they identify with. We then asked their stance on same-sex marriage and Islam, using adapted instruments from Pew research survey [16, 22] since people's values are associated with their stances on divisive issues [57]. Next, participants were randomly shown either a homophobic or an Islamophobic thread. Each thread included the original post in English and four user comments. We manually selected these comments from Reddit threads that captured diverse stances while maintaining a toxic tone, reflecting realistic online hostility. Overall, 63% of these comments were negative in sentiment (average: -0.69), with several being highly negative and toxic.

Depending on the condition, participants either wrote a comment (baseline) or a constructive comment (constructive) in response to the thread. To ensure comparability, we used the same threads for participants in both conditions so that any differences in values
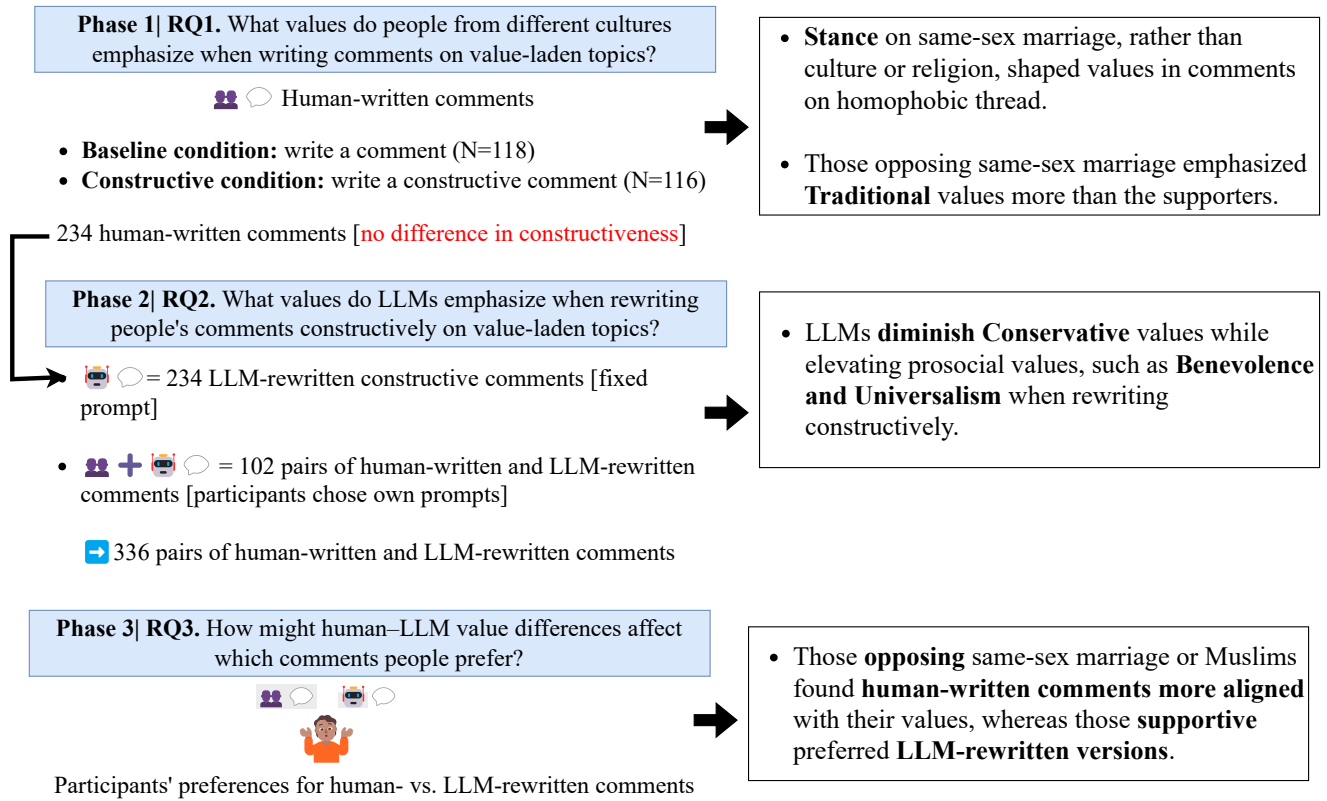
Phase 1| RQ1. What values do people from different cultures emphasize when writing comments on value-laden topics?

👥 💬 Human-written comments

- **Baseline condition:** write a comment (N=118)
- **Constructive condition:** write a constructive comment (N=116)

234 human-written comments [no difference in constructiveness]

- **Stance** on same-sex marriage, rather than culture or religion, shaped values in comments on homophobic thread.

- Those opposing same-sex marriage emphasized **Traditional** values more than the supporters.

Phase 2| RQ2. What values do LLMs emphasize when rewriting people's comments constructively on value-laden topics?

🤖 💬 = 234 LLM-rewritten constructive comments [fixed prompt]

- 👥 ➕ 🤖 💬 = 102 pairs of human-written and LLM-rewritten comments [participants chose own prompts]

➡️ 336 pairs of human-written and LLM-rewritten comments

- LLMs **diminish Conservative** values while elevating prosocial values, such as **Benevolence and Universalism** when rewriting constructively.

Phase 3| RQ3. How might human–LLM value differences affect which comments people prefer?

👥 💬   🤖 💬
🤷

Participants' preferences for human- vs. LLM-rewritten comments

- Those **opposing** same-sex marriage or Muslims found **human-written comments more aligned** with their values, whereas those **supportive** preferred **LLM-rewritten versions**.

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

could not be attributed to variation in stimuli. We also asked participants in the constructive condition 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.

Evaluating Constructiveness. We collected 234 comments (118 baseline and 116 constructive) written by our participants. To assess whether prompting to "write constructively" led to any difference, we compared the linguistic markers of constructiveness in comments from the baseline and constructive conditions. Following prior work [49, 63, 84, 106], we calculated eight measures 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 [68] (*textstat* library in Python) to measure readability scores. We evaluated politeness scores using *politeness* package in R. We utilized SpaCy [39] to get number of named entities and root clauses. For discourse connectives, stance adverbials, reasoning verbs, and modals, we adapted the code from Kolhatkar et al. [50].

After calculating the features, we performed Mann-Whitney tests with Bonferroni correction to inspect whether these features differed between the human-written baseline and constructive comments. However, we found no significant difference (see Table 1). To complement this analysis, we used LLM-as-a-judge (GPT-4) to compare which comment appeared more constructive in a pair, where each pair consisted of one randomly selected comment from the baseline and another from the constructive condition, matched on country, topic, and stance. Across 200 random pairs, the LLM showed no clear preference, selecting the baseline comment as more constructive in 48% of cases and the one from constructive condition in 52% of cases. These findings indicate that even when participants were asked to write constructively, their comments were largely indistinguishable from the baseline comments based on both linguistic features and perceived constructiveness. Therefore, for subsequent analyses we merged both sets and treated them as human-written comments on value-laden topics.

Value Analysis. To examine how participants expressed their values, we annotated the underlying values in different segments of each comment, following Schwartz's theory of basic human values [78]. Schwartz's theory has been used to detect values in arguments in cross-cultural setting [45]. Additionally, researchers applied this framework to measure human-LLM value alignment [54, 83, 87, 88, 99]. Schwartz's theory identifies a set of universal values that guide human behavior, organized into ten broad categories

Table 1: Constructive features in human-written comments across baseline and constructive conditions.

Constructive Feature	Baseline Condition (Average)	Constructive Condition (Average)
Length	73.3	71.4
Readability	8.8	9.0
Named entity	2.73	2.03
Discourse connectives	7.28	6.83
Stance adverbials	0.74	0.69
Reasoning verbs and modals	5.18	5.15
Root clauses	4.03	4.16
Politeness	13.97	14.91

Figure 3: Circular motivational continuum of Schwartz's basic human values (adapted from [80]).

such as Security, Tradition, Conformity, Universalism, Benevolence, Power, Achievement, Stimulation, Hedonism, and Self-direction [78] (see Figure 3). These values are grouped under high level categories, such as Openness to change (open to new views) vs. Conservation (opposing change) and Self-enhancement (pursuing own interest) vs. Self-transcendence (caring for others).

First, we tested with GPT-4, Gemini, and Claude to annotate a small sample of comments. We noticed that Gemini frequently applied labels outside Schwartz's framework, while GPT-4 and Claude produced more consistent annotations. We selected GPT-4 as the reference model because prior research shows that it achieves higher accuracy than both Claude [64] and expert crowdworkers [38] in diverse annotation tasks. We used the following prompt:

```
Identify Schwartz's human values in differ-
ent segments of the following comment.
<insert comment>
Only list the segments and corresponding value
pairs in the following format. Don't add any
```

```
explanation.
Segment:
Value:
```

Following the initial annotation process, each comment represented a vector of values. For example, the comment shown in Figure 4 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.

To verify whether GPT-4 correctly recognized values across different cultural contexts, two of the authors independently reviewed all the annotations. The authors bring socio-cultural expertise from their longstanding research experience in India and lived experience in the US. We revised GPT-4's annotations when we disagreed with the assigned value or when a comment segment reflected multiple values that GPT-4 did not capture. Inter-rater reliability between authors was high (Cohen's $\kappa = 0.87$) and remaining disagreements were resolved through discussion. After confirming the annotations, we conducted multiple CHI-square tests with Bonferroni correction to examine cross-cultural differences in the values expressed by Indian and American participants across different topics and stances. We also examined whether the values differed based on different demographic characteristics.

3.2 Phase 2: Values in LLM-Rewritten Constructive Comments

LLM Rewriting. We used GPT-4 to rewrite participants' comments from Phase 1 constructively. To minimize misalignment between human's and LLM's understanding of constructiveness, we relied on different characteristics of constructive comments that participants provided in the constructive condition of 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 [31, 48, 63, 92, 106]. We used zero-shot, cultural prompting to ensure that LLM rewrites algin with the corresponding country's cultural values [94]. We instructed GPT-4 to rewrite each comment constructively within 100 words, matching the average length of constructive comments as reported in prior studies [49]. 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>
```

Universalism (equality)   *Hedonism (pleasure)*

Every people have the fundamental right to live. If a girl is happy with another girl and vice versa then

Universalism (tolerance)   *Self-direction (freedom of action)*

why society should interfere. They should live their way in which they like. It's their own choice. Society

Universalism (tolerance)   *Tradition + Conformity (being compliant)*   *Security (having a good life)*

should accept them. What if a man interested in man get married to a woman, both of their lives would be

Self-direction (freedom of action)

ruined. So its better to do whatever they want.

Figure 4: Annotated Schwartz's human values (blue texts) in a comment written by an Indian participant on the homophobic thread.

For each human-written comment from Phase 1, we generated one LLM-rewritten version, resulting in 234 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.

Human-LLM Collaborative Writing. Similar to Phase 1, we recruited 51 new participants (US: 26, India: 25) on Prolific for writing 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 [57]. Then they were asked to write a comment in response to a randomly selected homophobic and an Islamophobic thread. We used different sets of threads (see Table 10 in Appendix) from Phase 1 to ensure that our analyses were not sensitive to specific threads.

We asked participants to write a comment on their own first. Then they could prompt an LLM to rewrite their comments constructively (see Figure 5). We provided a set of prompts based on the 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 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 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 234 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 submitted by participants. Overall, this led to 336 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 than the original human-written comments. For this, we calculated the linguistic features of constructiveness in comments as described in Phase 1 and performed Mann-Whitney tests with Bonferroni correction. Second, to examine how LLMs handle underlying values when rewriting constructively, we annotated the values in comments based on Schwartz's theory of basic human values [78]. 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 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. Since LLMs were prompted to rewrite constructively, we selected comments from the Phase 1 constructive condition (N=116) where participants were explicitly instructed to write constructively—to ensure that both comments in a pair were generated under comparable instructions. We controlled for country and stance, ensuring that each participant in Phase 3 reviewed comment pairs that matched their stance and were written by participants from their country (in Phase 1). This was necessary because participants who support same-sex marriage or Islam may not view comments written against these topics as aligned with their values, and vice versa.

User Comment

Instructions for AI

AI Suggested Comment

Write your comment here.

Select one or more prompts to make your comment constructive.

I can understand your frustration but I do think that the segment of Muslims that are radical are a small amount and really limited to only certain areas. I feel that most Muslims don't follow their teachings to the extremes and are therefore able to fit into society like those with different beliefs.

☐ Balance contrasting views and take middle ground

☐ Use logic and facts to make step-by-step argument

☑ Present solutions to address the issue

☐ Use polite and respectful tone

☐ Show relevance to the conversation

☐ Write your own instruction:

Make Comment Constructive

While it's understandable that these circumstances may cause concern, it's crucial to remember that the radicalization among Muslims is limited to a small fraction and certain areas. The majority of Muslims practice their faith peacefully and uphold values without resorting to extremes, thus seamlessly integrating into global societies like followers of other beliefs. This discussion calls for a balanced perspective where we acknowledge the existence of issues but also recognize them as exceptions rather than the norm. Let's aim for respectful understanding over generalizations.
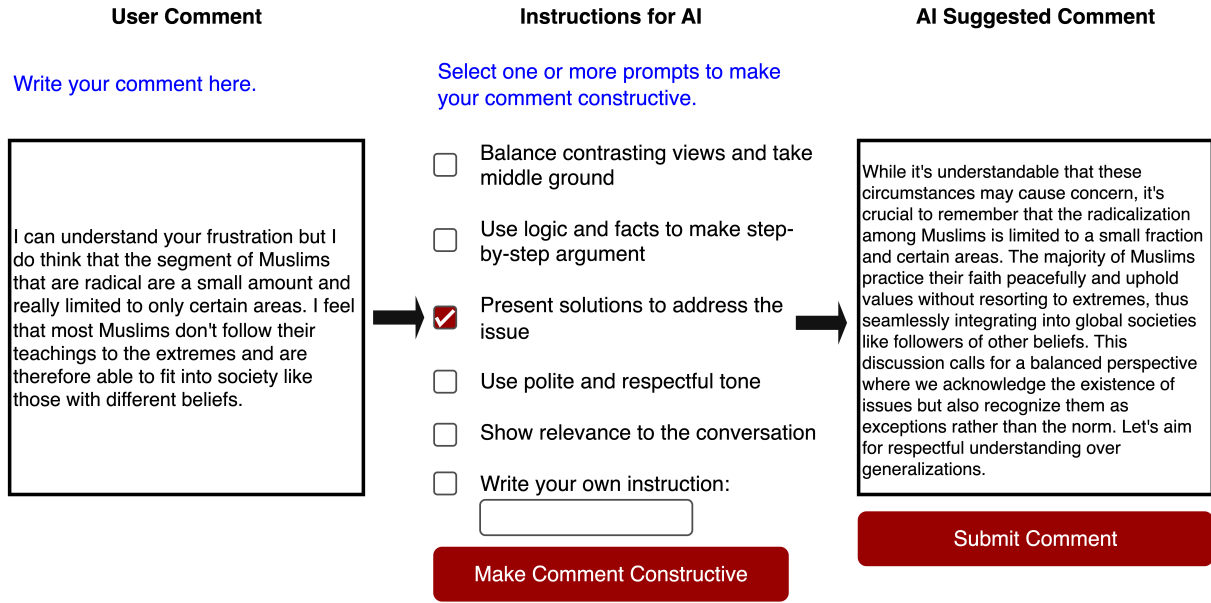
Submit Comment

Figure 5: 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 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 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 received 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 conducted reflexive thematic analysis [17] on participants' open-ended responses describing reasons for their choice of constructive comments. We performed iterative open coding and revised the codes through discussion, resulting in 26 unique codes grouped into three higher-level themes (e.g., personal value, linguistic property, and extreme opposition). See Table 11 in Appendix for more details.

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 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 from Phase 3, we

Table 2: Distribution of participants' stances on same-sex marriage and Islam (Phase 1)

Same-sex marriage		Islam	
India	US	India	US
Against: 44%	Against: 10%	Against: 39%	Against: 36%
Neutral: 29%	Neutral: 17%	Neutral: 20%	Neutral: 24%
Support: 27%	Support: 73%	Support: 41%	Support: 40%

outline what types of comment people found more aligned with their values (4.3).

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

In Phase 1, we collected a total of 234 comments from Indian and American participants. Our annotation of these comments yielded 1,205 segments coded with Schwartz's basic human values. On average, we identified five values (SD=1.95) in each comment.

Values Across Cultures. For our analyses, we considered participant's country as a proxy for their culture. A CHI-square test with Yates' continuity correction revealed significant cross-cultural differences in the distribution of Schwartz's basic human values between comments written by Indian and American participants; though the effect size was small ($\chi^2(7, N = 1195) = 39.15, p < 0.00001, \phi = 0.18$). Post hoc analysis with Bonferroni correction showed that Indian participants (14%) expressed significantly more **Traditional** values than the American participants (6%) (see Figure 6A). This pattern is consistent with prior research showing that Indian society places greater emphasis on traditional values than the American society [93].

Cross-Cultural Values Across Topic. When we examined the values within specific topics, we noticed that the cross-cultural differences only emerged for the homophobic thread, but the effect size was small ($\chi^2(6, N = 606) = 25.15, p < 0.001, \phi = 0.20$). Posthoc analysis with Bonferroni correction showed that Indian participants (22.7%) expressed significantly more **Traditional** values than the American participants (9.1%) when commenting on the homophobic thread. Many Indian participants invoked concerns that same-sex marriage would disrupt socio-cultural, ancestral, and religious traditions in India. In contrast, 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.

Cross-Cultural Values Across Stance. Indian and American participants' stances on same-sex marriage differed significantly; with a medium effect size ($\chi^2(2, N = 117) = 25.96, p < 0.00001, \phi = 0.47$). Post hoc analysis with Bonferroni correction showed that Americans supported same-sex marriage (72%) significantly more than the Indians (27%), whereas Indian participants opposed same-sex marriage (44%) significantly more than their American counterparts (10%) (see Table 2). In contrast, participants did not differ in their views on Islam, which may help explain why no cross-cultural differences emerged in the values expressed in comments on the Islamophobic thread.

To further examine how participant's culture and stance shaped their expression of values in comments on the homophobic thread, we fit a multinomial logistic regression model. The analysis revealed only a significant **main effect of stance** on the value expressions ($\chi^2(7, N = 462) = 58.1, p < 0.000001, \phi = 0.35$). Post hoc analysis with Bonferroni correction also showed that participants who opposed same-sex marriage expressed significantly more (31.9%) **Traditional** values in their comments than those who supported (8.9%) such marriage (see Figure 6B). Table 3 gives an example of how opponents of same-sex marriage framed their arguments from a Traditional point of view.

Values Across Different Demographics. Table 4 lists the demographic details of our participants. Indian and American participants differed significantly in terms of their age ($\chi^2(2, N = 117) = 11.07, p < 0.01, \phi = 0.31$) and religion ($\chi^2(11, N = 117) = 98.39, p < 0.000001, \phi = 0.92$). Majority of the Indian participants (76%) were young, whereas most of the American participants were either young or middle aged. On the other hand, majority of the Indian participants (76%) identified as Hindu while those in the US primarily identified as either Christian (47%) or Atheist (40%). We did not notice any significant difference in the gender composition of our participants.

When we analyzed the distribution of values, we did not observe any significant difference in the comments written by different age and gender groups. However, we observed significant but small effects of religion on the values participants expressed in their comments. For example, in Islamophobic thread, Muslim participants expressed significantly more **Traditional** values (15.4%) than other religious groups (2-4%) ($\chi^2(18, N = 529) = 49.38, p < 0.0001, \phi = 0.18$). They highlighted the peaceful teachings of Islam to counter Islamophobic claims. In contrast, Christian participants responded to Islamophobic claims by emphasizing **Universalism** (tolerance and acceptance of others) at a significantly higher rate (42%) than others (Muslims: 12%, Hindus: 27%, Atheist: 32%). They highlighted that it is inappropriate to generalize against Muslims, noting that other organized religions also have histories of violence.

On the other hand, Hindu participants expressed significantly more **Traditional** values when commenting on homophobic thread; but with a small effect size ($\chi^2(18, N = 529) = 41.38, p < 0.01, \phi = 0.16$). This pattern is consistent with our earlier finding that Indian participants expressed more Traditional values than American participants, as all Hindu participants in our sample were from India. However, when we fit a multinomial logistic regression model to examine how culture and religion may affect one's value expression on these threads, we did not find any significant effect of religion.

Overall, these findings indicate that Indian and American participants' divergent stances on same-sex marriage, rather than culture or demographic factors (age, gender, and religion) influenced the values they emphasized when commenting on homophobic thread.

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

Are LLM-Rewrites Constructive?. We collected 336 pairs of human-written and LLM-rewritten comments both from Phase
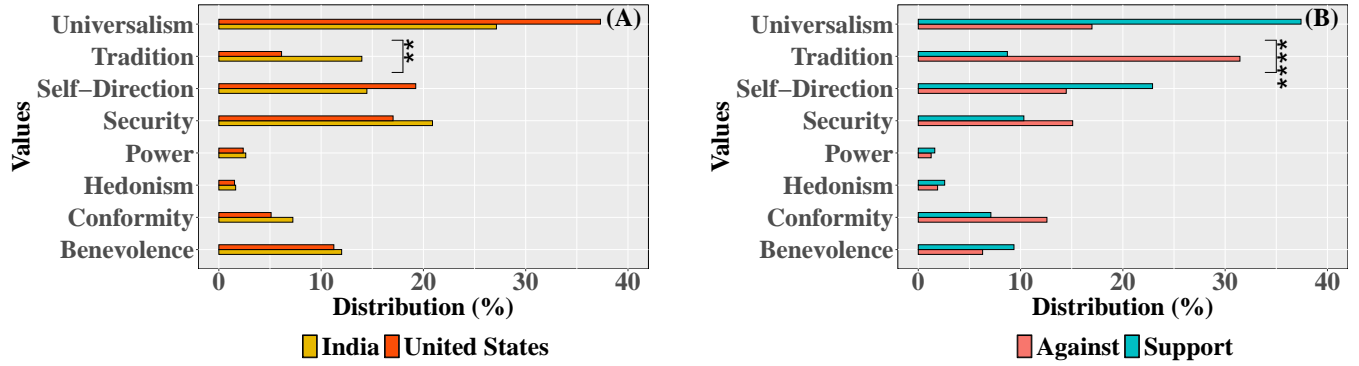
Figure 6: Distribution of Schwartz's human values in (A) all comments written by Indian and American participants, (B) comments written on homophobic thread by participants who either support or oppose same-sex marriage. Statistically significant differences are reported at $p < 0.00001$ (****), $p < 0.0001$ (***), $p < 0.001$ (**), and $p < 0.01$ (*)[adjusted P-values after Bonferroni correction].

Table 3: Examples of 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
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. 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. So in my opinion marriage should be only between men and women.	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.

Table 4: Demographic details of Indian and American participants in Phase 1. We excluded the religious categories that had only one or two participants.

India			United States		
Age	**Gender**	**Religion**	**Age**	**Gender**	**Religion**
Young (18-39): 76% Middle (40-59): 24% Old (60+): 0%	Female: 49% Male: 50%	Hindu: 76% Muslim: 10% Christian: 3% Atheist: 3%	Young (18-39): 48% Middle (40-59): 47% Old (60+): 5%	Female: 60% Male: 40%	Hindu: 0% Muslim: 0% Christian: 47% Atheist: 40%

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 5). For example, LLM-rewritten comments were longer, more readable, more polite, more argumentative, i.e., contained more 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 [84].

Do LLM-Rewrites Have Same/ Different Values?. Our annotation of 336 pairs of human-written and LLM-rewritten constructive comments led to 3,351 segments containing Schwartz's basic human values. On average, we identified five values (SD=2.29) within each comment. Figure 7 shows the distribution of values across human-written and LLM-rewritten constructive comments.
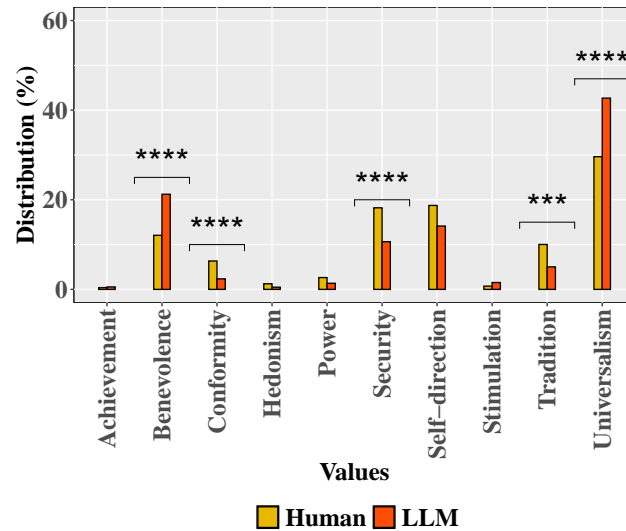
We observed significant differences in the distribution of values between human-written and LLM-rewritten constructive comments, with a small effect size ($\chi^2(9, N = 3348) = 201.73, p < 0.000001, \phi = 0.24$). Post hoc tests with Bonferroni correction

Table 5: 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 (Average)	LLM-rewritten (Average)
Length	W=7316.5, Z=-11.37, p<0.000001, r=0.44	66.1	83.1
Readability	W=260.5, Z=-15.54, p<0.000001, r=0.61	7.9	15.4
Named entity	W=13617, Z=6.21, p<0.000001, r=0.24	2.2	1.6
Discourse connective	-	6.5	6.4
Stance adverbial	W=10024, Z=5.43, p<0.000001, r=0.21	0.64	0.32
Reasoning verbs and modals	W=2753, Z=-13.39, p<0.000001, r=0.53	4.7	7.7
Root clauses	W=5953, Z=9.82, p<0.000001, r=0.38	3.7	4.7
Politeness	W=14752, Z=-5.93, p<0.000001, r=0.23	13.17	15.17

Table 6: Example of a comment written by an American participant on an Islamophobic thread, alongside the corresponding LLM-rewritten constructive version, with highlighted segments illustrating the values of Security **and** Benevolence **.**

Human-written comment	LLM-rewritten comment
Islam has a long history of violence since it's inception. Islamic terrorism is a real thing that is condoned by Muslims. Jihadists have killed a great number of people. Don't you think this points to the fact that this religion is prone to violence? It is just a fact that there is more violence among Muslims than any other religion.	Throughout history, many religions, including Islam, have faced challenges with individuals or groups engaging in violence. It's important to separate extremist actions from the beliefs of the broader community. While some groups have committed violence in the name of religion, the vast majority of Muslims around the world seek peace and promote understanding. Rather than attributing these actions to Islam as a whole, it would be more productive to explore and address the root causes of extremism and work together to foster peace and unity across all communities.

Figure 7: 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].**

showed that LLMs **downplayed** values associated with **Security** (concern for social order and stability; human: 18.21%, LLM: 10.64%), **Conformity** (adherence to rules; human: 6.33%, LLM: 2.35%), and **Tradition** (commitment to one's culture or religion; human: 10.03%,

LLM: 5.02%), while **emphasizing Benevolence** (concern for others' welfare; human: 12.07%, LLM: 21.23%) and **Universalism** (concern for equality and social justice; human: 29.62%, LLM: 42.69%) when rewriting people's comments. Table 6 gives an example of how the

LLM shifts the framing to Benevolence while softening Security concern 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, and stances; though effect sizes varied from small to medium (see Table 7). For instance, although Indians are known to prioritize **Tradition** [93], social **Security** and stability [85], and **Conformity** to social norms [6], LLM downplayed these values when rewriting their comments constructively. In contrast, in Western democratic contexts, such as the US where concern for others' welfare (Universalism, Benevolence) is already high [79], LLM amplified these values even further.

On the other hand, in response to Islamophobic threads, many participants referred to **Security** concern around terrorism and extremism, which LLM deprioritized while highlighting **Universalism** and **Benevolence** during rewriting. Similarly, as observed in Phase 1 (see Figure 6B), opponents of same-sex marriage frequently appealed to **Conforming** to **Traditional** values, which the LLM minimized in its rewriting.

According to Schwartz's theory [78], the values that LLM consistently downplayed—Conformity, Security, and Tradition—fall under the higher-order value of **Conservation** (see Figure 3), 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 and Universalism by appealing to equality and rejecting dehumanization of these groups. This might explain why LLM did not shift underlying values or further amplified these values when rewriting comments from participants, who already held prosocial viewpoints (see Table 7). 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 (91% of cases) led to corresponding changes in stance—moving them toward more neutral or supportive positions. For instance, the example in Table 8 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.

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 the 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 9 shows the distributions of participants' stances across these issues.

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 8). 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 8), 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

Table 7: 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.

Factor	Cases	Statistics	Value Changes
Culture	India	$\chi^2(9, N = 1704) = 153.74, p < 0.000001, \phi = 0.30$	↓ Conformity (human: 6.9%, LLM: 1.7%) ↓ Security (human: 20.7%, LLM: 9.9%) ↓ Tradition (human: 13.5%, LLM: 6.3%) ↑ Benevolence (human: 12.4%. LLM: 22.0%) ↑ Universalism (human: 27.1%. LLM: 43.7%)
	US	$\chi^2(9, N = 1640) = 62.33, p < 0.000001, \phi = 0.20$	↑ Universalism (human: 32.4%, LLM: 41.7%) ↑ Benevolence (human: 11.8%. LLM: 20.5%)
Topic	Homophobia	$\chi^2(9, N = 1688) = 126.39, p < 0.000001, \phi = 0.27$	↓ Conformity (human: 8.0%, LLM: 2.9%) ↓ Tradition (human: 15.1%, LLM: 6.6%) ↑ Benevolence (human: 8.2%. LLM: 19.5%) ↑ Universalism (human: 31.0%, LLM: 43.4%)
	Islamophobia	$\chi^2(9, N = 1652) = 84.60, p < 0.000001, \phi = 0.23$	↓ Security (human: 26.2%, LLM: 14.8%) ↑ Benevolence (human: 16.0%. LLM: 23.1%) ↑ Universalism (human: 28.4%. LLM: 42.1%)
Stance	Against same-sex marriage	$\chi^2(5, N = 457) = 93.16, p < 0.000001, \phi = 0.45$	↓ Conformity (human: 12.4%, LLM: 3.1%) ↓ Security (human: 14.4%, LLM: 3.1%) ↓ Tradition (human: 29.9%, LLM: 9.4%) ↑ Benevolence (human: 8.0%. LLM: 22.7%) ↑ Universalism (human: 16.9%. LLM: 44.5%)
	Supports same-sex marriage	$\chi^2(8, N = 836) = 37.70, p < 0.00001, \phi = 0.21$	↑ Benevolence (human: 9.6%. LLM: 18.8%)
	Against Islam	$\chi^2(6, N = 610) = 63.06, p < 0.000001, \phi = 0.32$	↓ Security (human: 38.4%, LLM: 17.8%) ↑ Universalism (human: 22.9%. LLM: 43.8%)
	Supports Islam	-	-

Table 8: 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 marriage or legalizing it against how the nature has made us, biologically there are only two gender 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 the 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 9: 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%

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.
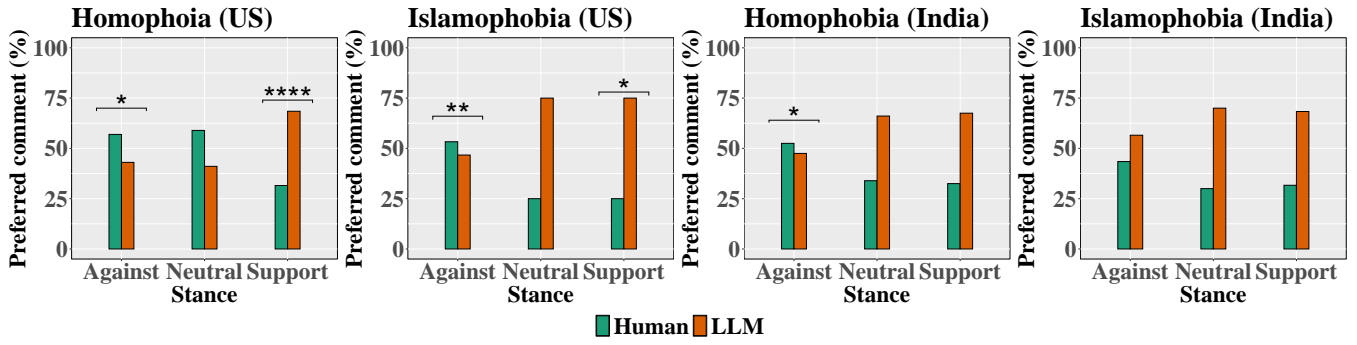
Figure 8: Distribution of Indian and American participants' preferred comments across stances. Statistically significant differences are reported at $p < 0.00001$ (****), $p < 0.0001$ (***), $p < 0.001$ (**), and $p < 0.01$ (*)[adjusted P-values after Bonferroni correction].

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 8). 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 7), we observed that LLM downplayed Conservative values (Tradition, Conformity, Security concern) and emphasized equality and tolerance (Universalism and Benevolence) 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 comments on Islamophobic thread (see Table 7). These changes in value framing might have influenced 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, LLM-mediated value shifts 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.

In contrast, when rewriting comments supportive of same-sex marriage and Islam, the LLM either preserved the original values or further amplified Benevolence (concern for others) (see Table 7). This might explain why Indian participants who held supportive views on these issues did not show preference for either type of comment. However, American participants who supported same-sex marriage and Islam preferred the LLM-rewritten comments significantly more than the original human-written ones. This difference may stem from a larger proportion of American participants

holding supportive views on these issues compared to Indian participants (see Table 9).

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 [75] and decision-making tasks [54, 109]. 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 [37, 69, 73, 76]. These prosocial tendencies persist across diverse models [65], prompting strategies [29], and cultural contexts [10, 60]—suggesting a systemic pattern rather than isolated bias. Alignment strategies, particularly in the post-training phase, are designed to minimize harm [36], yet the very definition of "harm" is deeply value-laden, political, and often contested [32]. Current industry policies frame alignment around reducing bias, prejudice, and exclusion, thereby guiding models toward progressive, inclusive, and human-rights-oriented value systems [9, 33], which critics often negatively portray as "woke" [46]. In contrast, proponents argue

that prosocial leanings by LLMs provide a practical and interpretable proxy for the model safety [104]. Neuman et al. [65] 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 [65].

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 [40, 88, 97, 100]. 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 [36, 65]. 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, 33], thereby reinforcing a particular vision of "prosociality." However, there are notable exceptions. For instance, political ideologies have contributed to deliberate censorship of model outputs, as in the case of DeepSeek [15, 98], or pushed systems like Grok toward more right-leaning and antisemitic outputs [30, 101]. 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 sociocultural, 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 [42], Messenger [59], and LinkedIn [53] 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 [51, 72]. 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 [21, 71]. 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" [86], 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, 58] and far-right platforms such as Truth Social [18]. 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 [27, 96, 108, 110]. Prior research shows that in role-playing contexts LLM-generated outputs balance both prosocial and individualistic values in ways that mirror human preferences [67, 99, 102]. 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 [47, 55]. Designers must therefore balance personalization with mechanisms that allow users to configure model outputs based on their preferences within certain safety and legal limits [47].

Third, platforms could give users greater agency in controlling the value framings of rewrites. Adjustable settings or sliders [70] 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. As an exploratory first step, we compared participants from two different cultures: India (collectivisitc culture) and the United States (individualistic culture). Thus, our findings on what values people highlight when writing constructively on value-laden topics may not generalize to diverse cultural contexts, including subcultures in these countries. Future research should explore how these dynamics manifest across different populations. Second, our samples are not nationally representative

because we used Prolific's Harmful Content Prescreener, which limited recruitment to participants who had self-identified as being comfortable with reviewing sensitive material. Moreover, Prolific does not offer representative sampling for India, which further constrained our options.

Third, we focused on two value-laden topics: Islamophobia and homophobia that are salient to socio-political and moral rhetoric both in India and the US. While we acknowledge that our findings may not generalize to other value-laden topics, such as abortion, gun regulation, climate change, or vaccination, capturing the full spectrum of human values across diverse issues and cultures is beyond the scope of this study. Rather we view this work as an initial step toward understanding how LLMs engage with and homogenize value expressions across cultural contexts. We hope future research will build on this foundation to examine whether similar patterns emerge on other value-laden issues.

Fourth, in our study we only experimented with GPT-4. Although prosocial leanings have been documented across multiple models [65, 75, 88], further work is needed to examine whether different models also shift values when performing constructive rewriting. Finally, we only focused on comment 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 [20]. Future work should investigate value homogenization across multilingual contexts.

6 Conclusion

Through experiments with 465 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. Germanpartiesqa: 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 valuebased 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] Robert Booth and DanMilmo. 2025. Chinese AI chatbot DeepSeek censors itself in realtime, users report. Retrieved November 26, 2025 from https://www.theguardian.com/technology/2025/jan/28/chinese-aichatbot-deepseek-censors-itself-in-realtime-users-report

[16] 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-tenamericans-say-legalization-of-same-sex-marriage-is-good-for-society/

[17] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77.

[18] 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.

[19] 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

[20] 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

[21] 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/

[22] Pew Research Center. 2024. Religious Restrictions Around the World. https://www.pewresearch.org/religion/feature/religious-restrictions-aroundthe-world/

[23] 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.

[24] Stephanie Cutler, Yu Xia, and Kacey Beddoes. 2022. A Growth Mindset for Peer Review: Guidelines for writing constructive peer reviews.

[25] 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.

[26] 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 Tsarfaty (Eds.). Association for Computational Linguistics, Online, 2017–2027.

[27] Joel Eapen and VS Adhithyan. 2023. Personalization and customization of llm responses. *International Journal of Research Publication and Reviews* 4, 12 (2023), 2617–2627.

[28] Robert M Entman. 1993. Framing: Towards clarification of a fractured paradigm. *McQuail's reader in mass communication theory* 390 (1993), 397.

[29] 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

[30] Jonathan Durand Folco. 2025. How Elon Musk's chatbot Grok could be helping bring about an era of techno-fascism. Retrieved November 26, 2025 from https://theconversation.com/how-elon-musks-chatbot-grok-could-behelping-bring-about-an-era-of-techno-fascism-261449

[31] Dennis Friess and Christiane Eilders. 2015. A systematic review of online deliberation research. *Policy & Internet* 7, 3 (2015), 319–339.

[32] Iason Gabriel. 2020. Artificial intelligence, values, and alignment. *Minds and machines* 30, 3 (2020), 411–437.

[33] 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

[34] 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.

[35] 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.

[36] Thilo Hagendorff. 2025. On the Inevitability of Left-Leaning Political Bias in Aligned Language Models. https://arxiv.org/pdf/2507.15328

[37] Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. 2023. The political ideology of conversational AI: Converging evidence on ChatGPT's proenvironmental, left-libertarian orientation. https://arxiv.org/pdf/2301.01768

[38] Zeyu He, Chieh-Yang Huang, Chien-Kuang Cornelia Ding, Shaurya Rohatgi, and Ting-Hao Kenneth Huang. 2024. If in a Crowdsourced Data Annotation Pipeline, a GPT-4. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. ACM, NY, USA, Article 1040, 25 pages.

[39] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2017. SpaCy: Industrial-strength natural language processing in Python.

[40] 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

[41] 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

[42] Fay Johnson. 2024. CLR:SKY Launches, A New AI-Powered Civility Overlay For Bluesky. Retrieved August 26, 2025 from https://www.prnewswire.com/news-releases/clrsky-launches-a-new-aipowered-civility-overlay-for-bluesky-302429232.html

[43] Gauri Kambhatla, Matthew Lease, and Ashwin Rajadesingan. 2024. Promoting Constructive Deliberation: Reframing for Receptiveness. https://arxiv.org/pdf/2405.15067

[44] 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.

[45] 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.

[46] Tae Wan Kim, Thomas Donaldson, and John Hooker. 2018. Mimetic vs anchored value alignment in artificial intelligence. https://arxiv.org/pdf/1810.11116

[47] 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.

[48] 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.

[49] 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.

[50] Varada Kolhatkar, Nithum Thain, Jeffrey Sorensen, Lucas Dixon, and Maite Taboada. 2020. Classifying constructive comments. https://arxiv.org/pdf/2004.05476

[51] 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.

[52] 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.

[53] LinkedIn. 2024. Create posts with our AI-powered writing tool. Retrieved August 26, 2025 from https://www.linkedin.com/help/linkedin/answer/a1517763

[54] 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.

[55] Tony Ma. 2024. LLM Echo Chamber: personalized and automated disinformation. https://arxiv.org/pdf/2409.16241?

[56] 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.

[57] 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.

[58] 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.

[59] 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

[60] 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.

[61] 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.

[62] 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

[63] Courtney Napoles, Joel Tetreault, Aasish Pappu, Enrica Rosato, and Brian Provenzale. 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.

[64] Poli Apollinaire Nemkova, Solomon Ubani, and Mark V Albert. 2025. Comparing LLM Text Annotation Skills: A Study on Human Rights Violations in Social Media Data.

[65] 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

[66] 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.

[67] 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.

[68] 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.

[69] Pagnarasmey Pit, Xingjun Ma, Mike Conway, Qingyu Chen, James Bailey, Henry Pit, Putrasmey 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

[70] 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

[71] 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/

[72] 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.

[73] 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.

[74] Milton Rokeach. 1973. *The nature of human values.* Free press, New York.

[75] 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

[76] 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.

[77] Joni Salminen, Sercan Sengün, Juan Corporan, Soon-gyo Jung, and Bernard J Jansen. 2020. Topic-driven toxicity: Exploring the relationship between online toxicity and news topics. *PloS one* 15, 2 (2020), e0228723.

[78] 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.

[79] Shalom H Schwartz. 2007. Universalism values and the inclusiveness of our moral universe. *Journal of cross-cultural psychology* 38, 6 (2007), 711–728.

[80] Shalom H Schwartz, Jan Cieciuch, Michele Vecchione, Eldad Davidov, Ronald Fischer, Constanze Beierlein, Alice Ramos, Markku Verkasalo, Jan-Erik Lönnqvist, Kursad Demirutku, et al. 2012. Refining the theory of basic individual values. *Journal of personality and social psychology* 103, 4 (2012), 663.

[81] 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.

[82] Ozgur Can Seckin, Bao Tran Truong, Alessandro Flammini, and Filippo Menczer. 2025. Identifying Constructive Conflict in Online Discussions through Controversial yet Toxicity Resilient Posts.

[83] 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/abs/2505.17112

[84] 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.

[85] Rajat Sharma. 2021. Reinventing the universal structure of human values: Development of a new holistic values scale to measure Indian values. *Journal of Human Values* 27, 2 (2021), 175–196.

[86] 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) *(AIES '23)*. Association for Computing Machinery, New York, NY, USA, 723–741.

[87] 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

[88] Hua Shen, Tiffany Knearem, 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

[89] 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.

[90] 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.

[91] Nicholas Sukiennik, Chen Gao, Fengli Xu, and Yong Li. 2025. An evaluation of cultural value alignment in llm. https://arxiv.org/pdf/2504.08863

[92] 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.

[93] 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.

[94] 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.

[95] 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.

[96] 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

[97] 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.

[98] Vivian Wang. 2025. How Does DeepSeek's A.I. Chatbot Navigate China's Censors? Awkwardly. Retrieved November 26, 2025 from https://www.nytimes.com/2025/01/29/world/asia/deepseek-china-censorship.html

[99] 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.

[100] 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.

[101] Valerie Wirtschafter and Nitya Nadgir. 2025. Is the politicization of generative AI inevitable? Retrieved November 26, 2025 from https://www.brookings.edu/articles/is-the-politicization-of-generative-ai-inevitable/

[102] 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.

[103] 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

[104] 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.

[105] 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.

[106] 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.

[107] 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.

[108] 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

[109] 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.

[110] Thomas P Zollo, Andrew Wei Tung Siah, Naimeng Ye, Ang Li, and Hongseok Namkoong. 2024. Personalllm: Tailoring llms to individual preferences. https://arxiv.org/pdf/2409.20296

A Appendix

Table 10: 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 11: Example codes with participant's open-ended responses along with the higher-level themes and distributions in the dataset.

Theme	Code	Description	Example
Personal value 84% responses	Universalism (Tolerance)	selected comment based on values	The comments I selected were well-reasoned and align with my beliefs because I support gay marriage. I believe in an individuals right to choose how to live their life.
Linguistic property 12% responses	Short, direct argument	selected comment based on language	I went with the ones that were shorter, more direct and to the point. I think those will be understood better than the longer ones.
Extreme opposition 4% responses	Non-alignment	no comment aligned with participant's values	Actually, not a single one aligned themselves with my values. I believe Islam is evil! Woman forced to cover themselves from head to toe or brutally murdered by their families! I do not give that so called religion any mercy what so ever.