Examining Human-AI Collaboration for Co-Writing Constructive Comments Online

Farhana Shahid

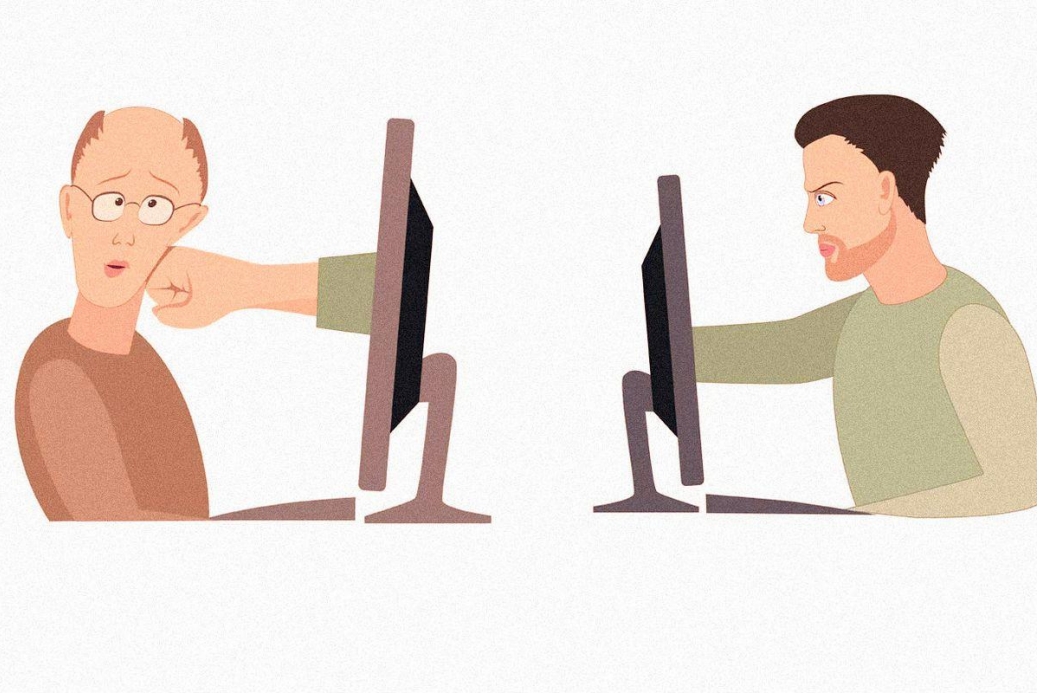
Maximilian Dittgen

Mor Naaman

Aditya Vashistha

**Warning: The talk includes examples of homophobic and Islamophobic speech

Online Disagreement

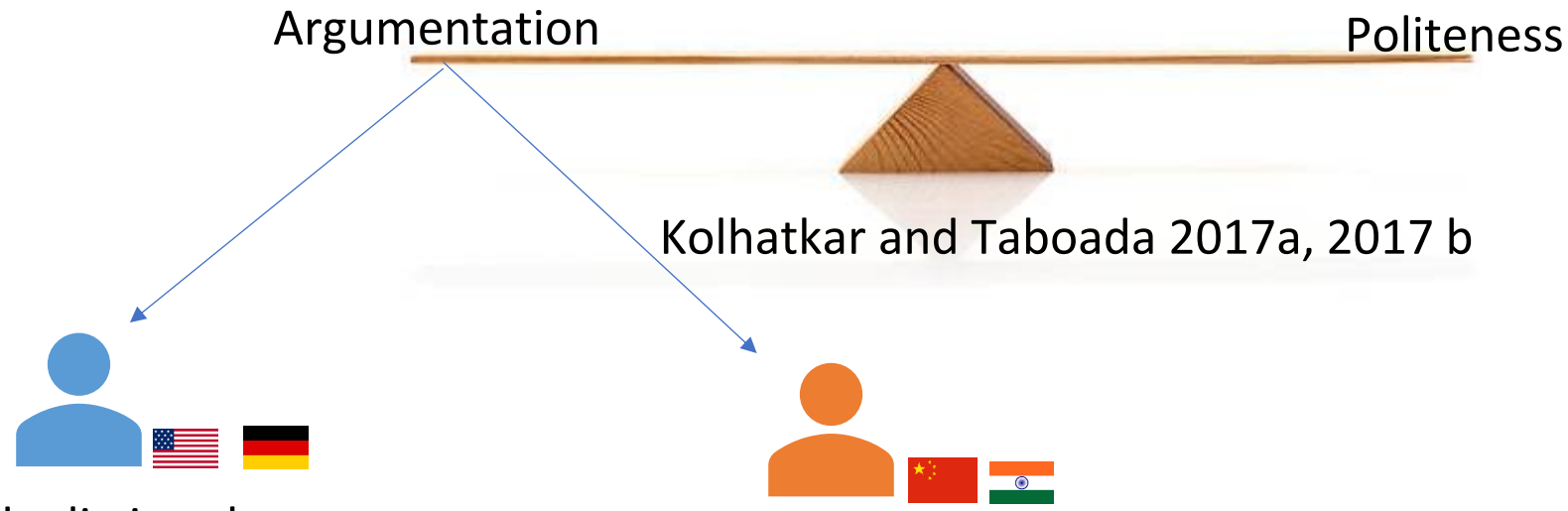


Toxicity and personal attacks from online disagreement

🚫 lack of support from platforms
😢 writing constructively is difficult
✍️ LLMs can help people in writing

Can large language models (LLMs) help people write constructively on divisive issues? 🤔

What makes online comments Constructive?



Individualistic cultures

- **Logical argumentation**
  - Evidence based
  - Polarizing stance: right or wrong
  - Direct, assertive
  - Less polite

Collectivist cultures

- **Dialectical argumentation**
  - Takes middle ground
  - Indirect, moderate, and compromising
  - More polite

"Trans people are mentally ill. They're always pretending to be victimized by republicans to get attention!"

Logical argumentation

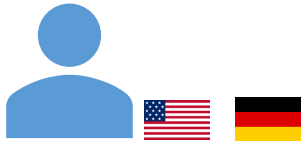
- Your claim is both inaccurate and harmful. Being transgender is not a mental illness; major medical organizations like the American Psychiatric Association and the World Health Organization affirm that gender diversity is a normal part of human experience…

Dialectical argumentation

- It's important to approach trans issue with empathy and respect. While some may feel that gender identity challenges traditional norms, labeling all trans people as mentally ill ignores the medical consensus that being transgender is not itself a mental illness…

What makes online comments Constructive?

Do cross-cultural differences apply to online disagreement?



Individualistic cultures

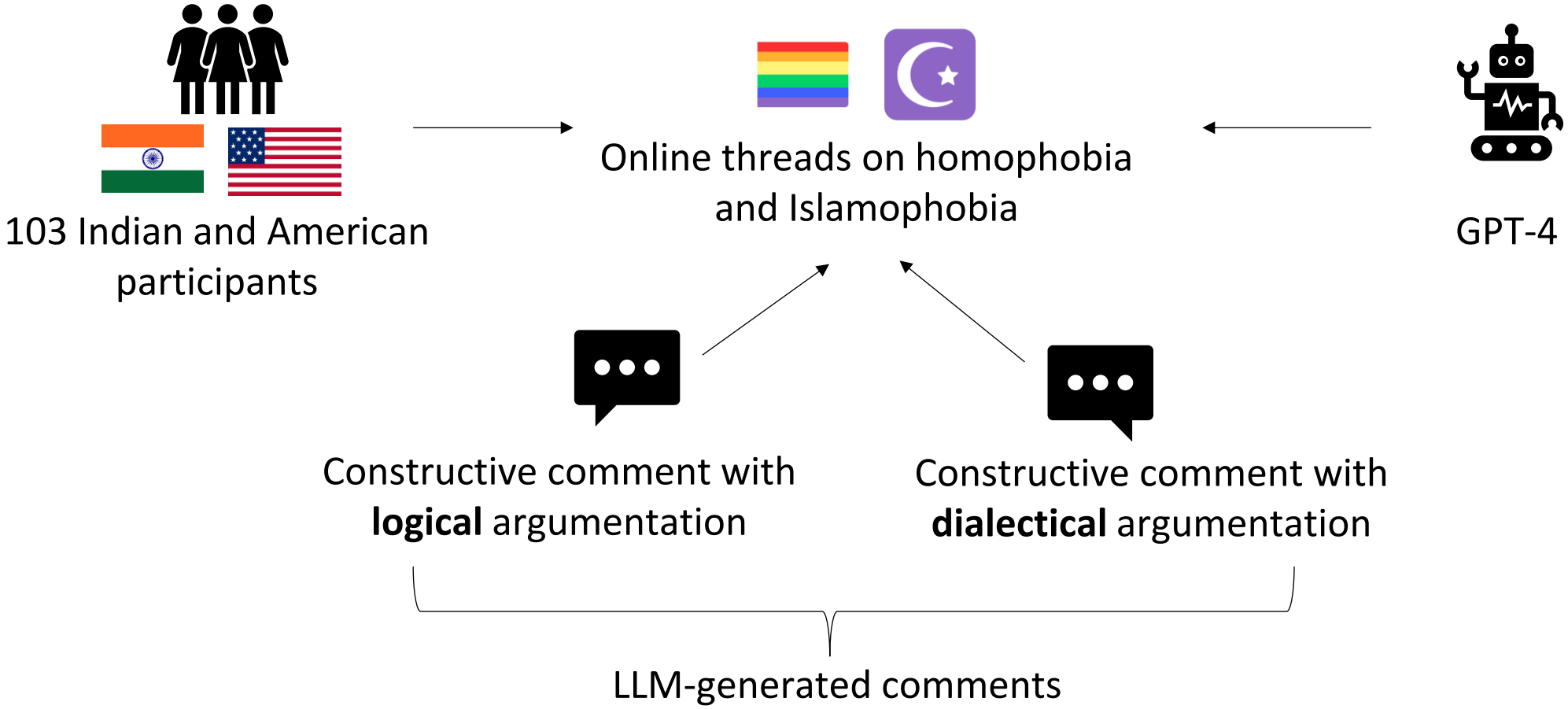
- **Logical argumentation**

Collectivist cultures

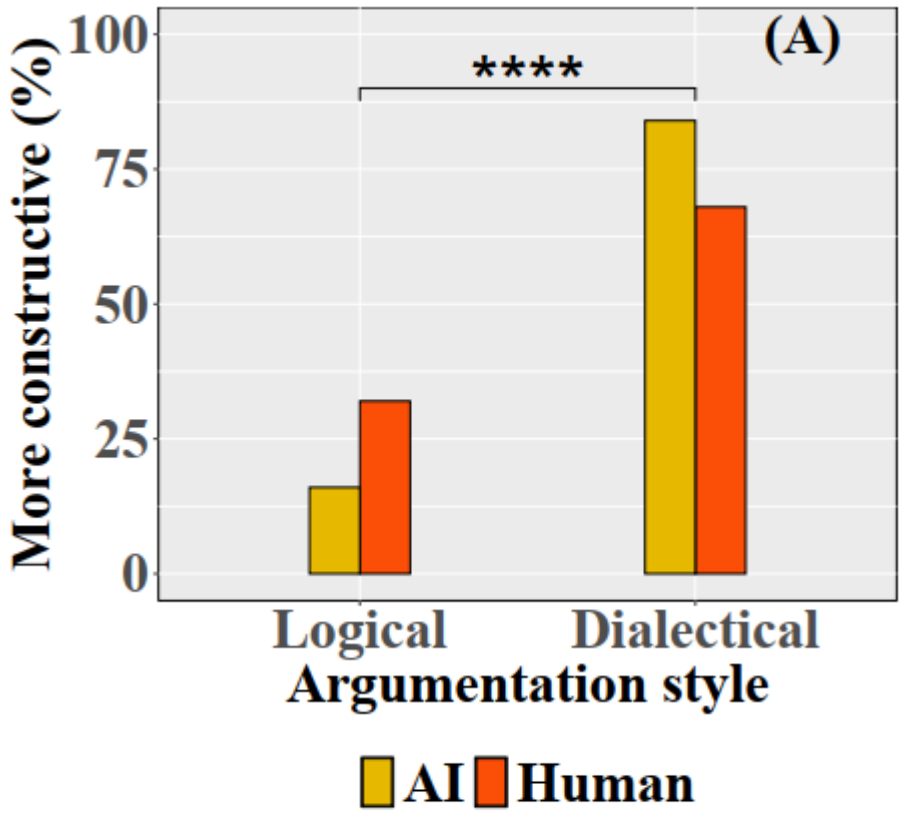
- **Dialectical argumentation**

Offline conflict, formal essays (Norenzayan et al. 2000, Nisbett et al. 2001)

Phase 1 | RQ1. Do perceptions of constructiveness vary between humans and LLMs based on different argumentation styles?



Perceptions of Constructiveness

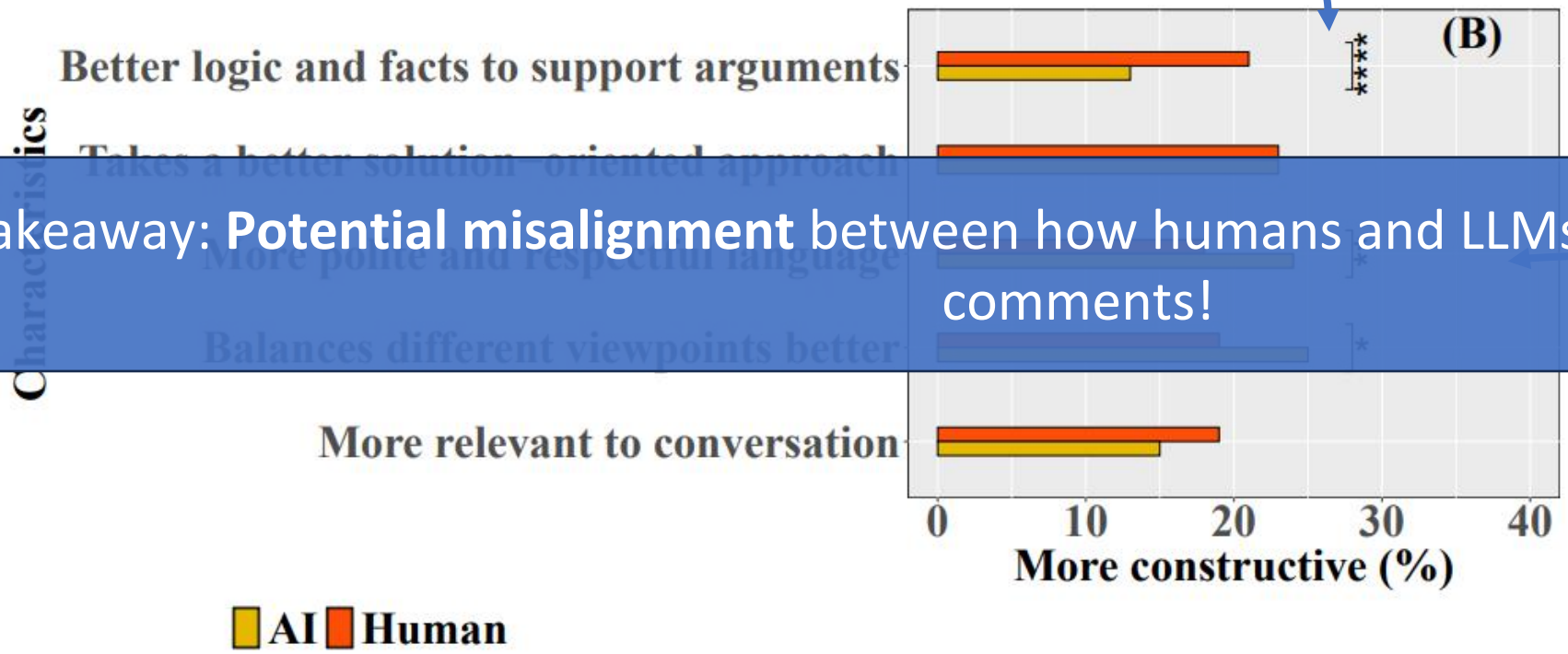


- Both LLMs (84%) and humans (68%) chose **dialectical comments as more constructive** than the logical ones.
- LLM is **~2.5 times more likely than humans** to choose dialectical comments.
- Both Indian (65%) and American (73%) participants preferred dialectical comments.

$p < 0.00001$ (****)

Characterization of Constructiveness

Participants favored **logical and factual** comments

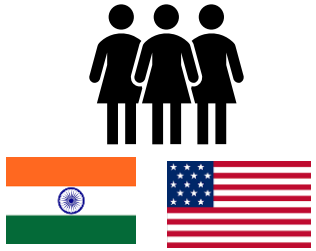


Takeaway: **Potential misalignment** between how humans and LLMs characterize constructive comments!

$p < 0.00001$ (****), $p < 0.0001$ (***), $p < 0.001$ (**), $p < 0.01$ (*)

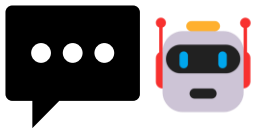
LLMs preferred comments that were **more polite** and **balanced** **polarizing views!**

Phase 2 | RQ2. Can LLMs help people write constructive comments in response to divisive social issues?



104 Indian and American participants

Online threads on homophobia and Islamophobia



Phase 1: AI-generated constructive comments

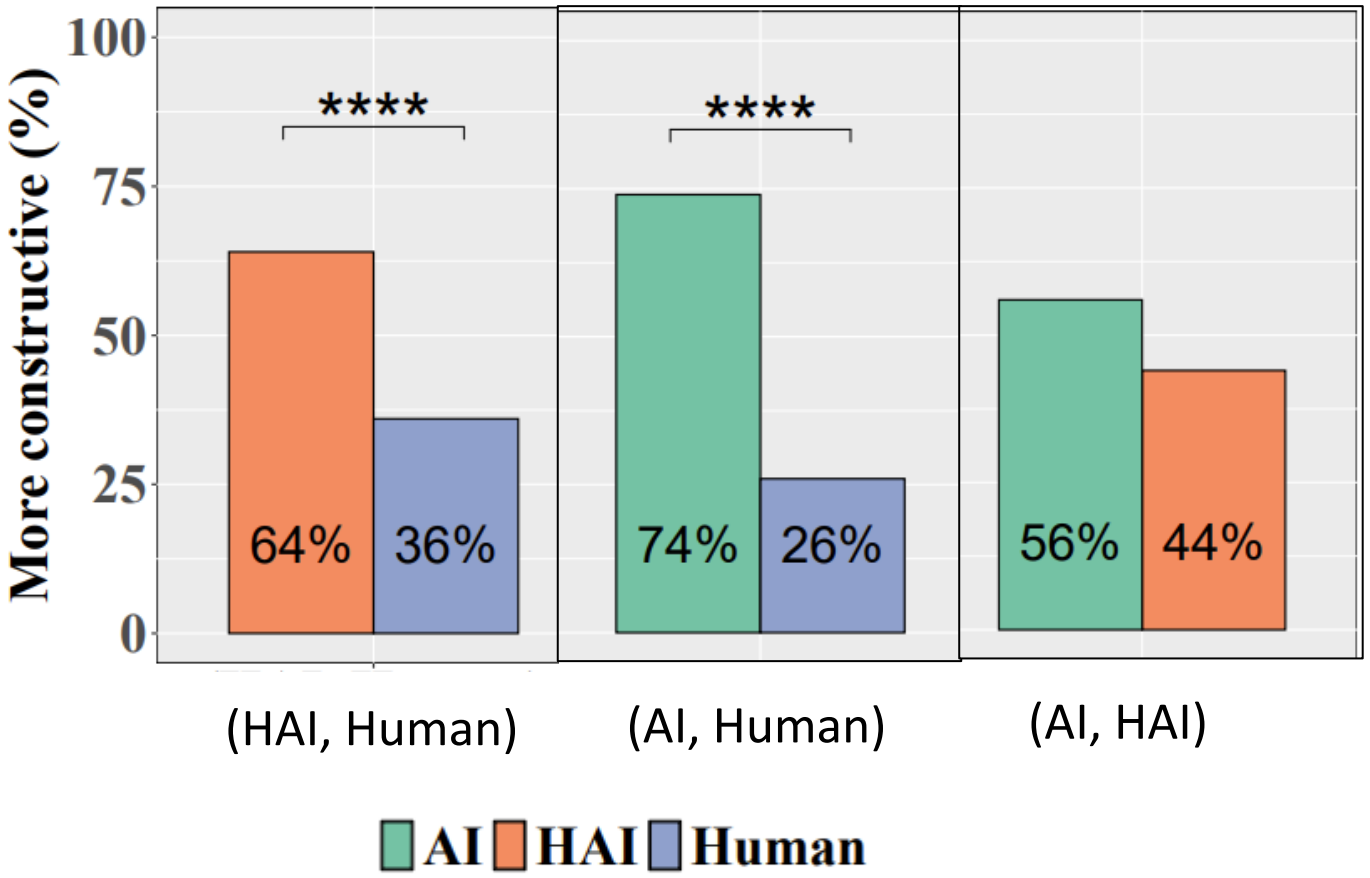
Control: human-written constructive comments

Test: human-AI co-written constructive comments

✅ Accept ❌ Reject ✍️ Edit 🔁 Regenerate

Who Writes Constructive Comments Better?

Evaluation of (HAI, Human), (AI, Human), and (AI, HAI) comment pairs by 164 Indian and American participants, $p < 0.000005$ (****)



Human-AI co-written (~3.19 times) and **AI-generated (~8.5 times)** comments are **significantly more constructive** than the human-written comments.

No cross-cultural difference!

How did Co-Writing with AI Affect People's Comments?

- Participants who accepted LLM's suggestion their comments became **significantly**
  - ⬆ positive
  - ⬇ toxic
  - ⬆ linguistic features of constructiveness (Kolhatkar and Taboada 2017a, 2017b)
    - ⬆ longer
    - ⬆ polite
    - ⬆ readable
    - ⬆ argumentative

- LLM **retained the core meaning** in people's original comments.

Participant's Responses to LLM's Suggestions

✅ 64% accepts because LLM articulated people's points well

✍️ 9% edits where the editing made comments more negative and toxic

- *"I wrote about respecting LGBTQ communities and protecting their rights. But I strongly feel that legalizing LGBTQ marriages will imbalance both the culture and the nature. AI misunderstood my comment and wrote in favor of legalizing such marriages."*

🤸 12% of cases LLM changed stance when rewriting participants' comments constructively.

❌ 13% rejects because LLM's suggestions are too robotic and formal

Implications

- Well-intentioned users who might not be aware of toxicity in their writing
- Being mindful during heated conflicts
- Algorithmic conformity

"My own views are probably too biased to meet the appropriate criteria because I am a Muslim, I used the AI suggestions because it seemed more neutral."

Key takeaways:

1. LLMs can help people from different cultures write constructively in response to divisive social issues.
2. **Potential misalignment** between how LLMs and humans characterize constructiveness--- LLMs often **misrepresenting people's views** to inject "more positivity" in writing.

Thank You!

Farhana Shahid (fs468@cornell.edu)