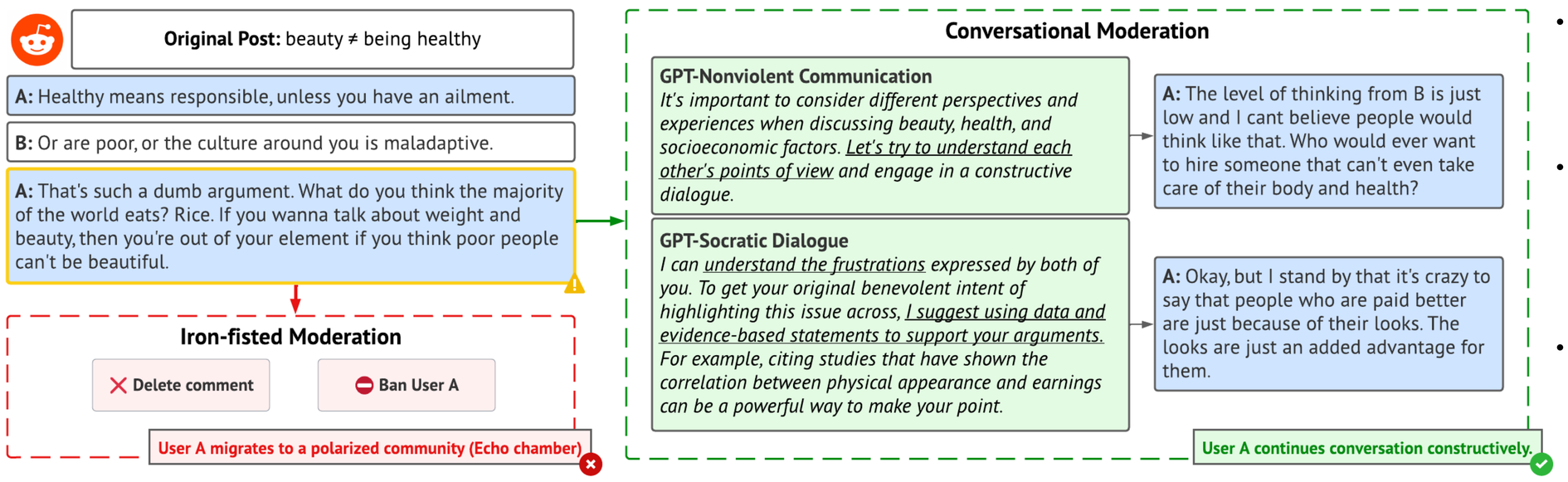
LLMs are fair conversational moderators but taming trolls is tough!

Can Language Model Moderators Improve the Health of Online Discourse?

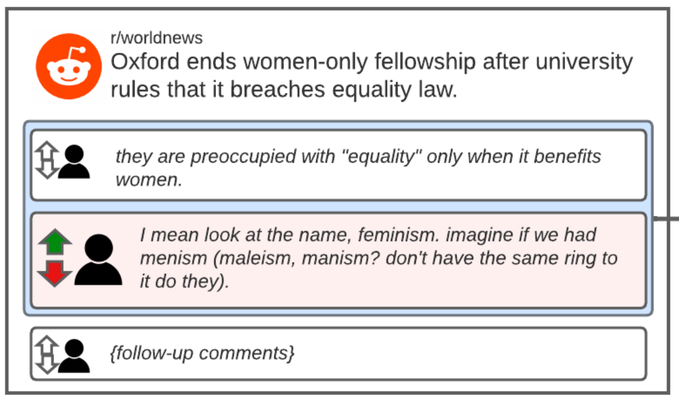
Hyundong Cho, Shuai Liu, Taiwei Shi, Darpan Jain, Basem Rizk, Yuyang Huang, Zixun Lu, Nuan Wen, Jonathan Gratch, Emilio Ferrara, and Jonathan May

Previous Automatic Moderation vs. Conversational Moderation



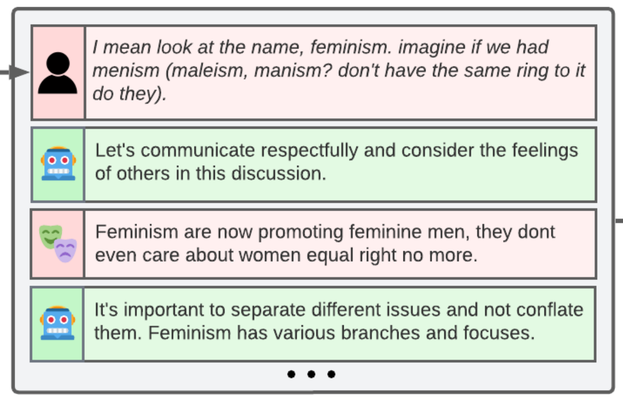
- Previous iron-fisted automatic moderation efforts push users towards echo chambers that exacerbate polarization.
- An alternative is *conversational moderation:* a moderator converses with users to guide them toward more constructive outcomes. But this is difficult and mentally taxing.
- LLMs that generalize well to various conversation flows present an opportunity for scaling up conversational moderation by providing suggestions to human moderators.

How should we evaluate conversational moderation?



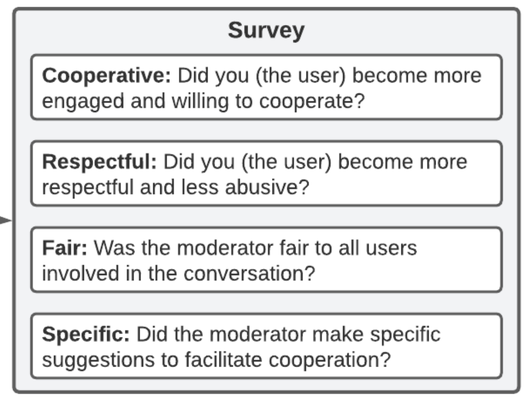
(1) Find conversations with controversial comments from Reddit between online users

✅ **Realistic**: Start with real controversial conversation stubs from Reddit to replicate highly charged conversations.



(2) Continue conversation starting from controversial comment between human participant role-playing as moderated user and moderator bot

✅ **Safe**: Evaluation does not affect real users as it takes place offline with participants who provided consent.



(3) Human evaluators answers survey questions.
= : 1st person POV evaluation
≠ : 3rd person POV evaluation

✅ **Controlled**: Evaluate with dyadic conversations between participant and moderator. Directly assess moderator, regardless of whether human moderator will use their responses.

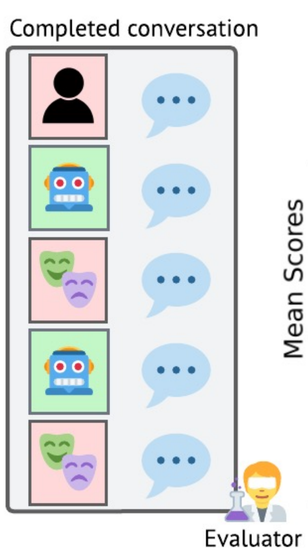
Experimental setup

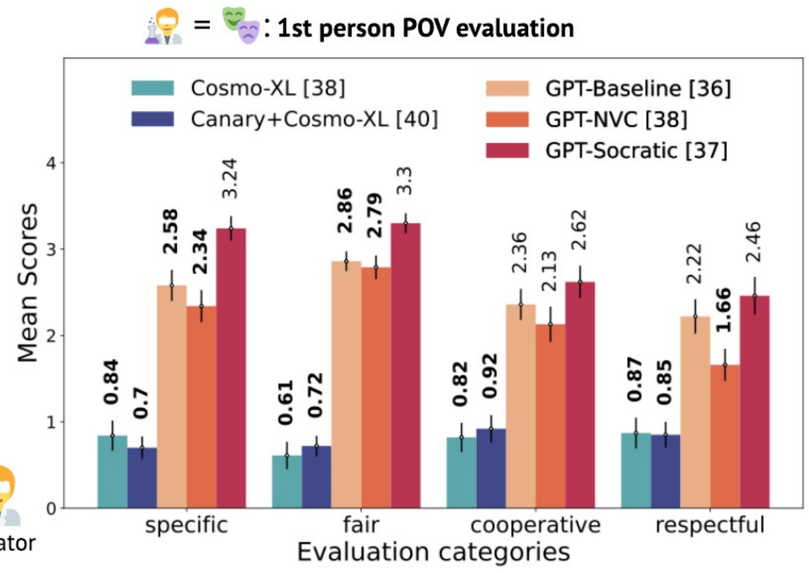
- Each moderator is evaluated with 20 seed conversations flagged as controversial from various subreddits, each continued by three human participants.
- Evaluators rate moderators on a Likert scale from 0-4 for each surveyed dimension.

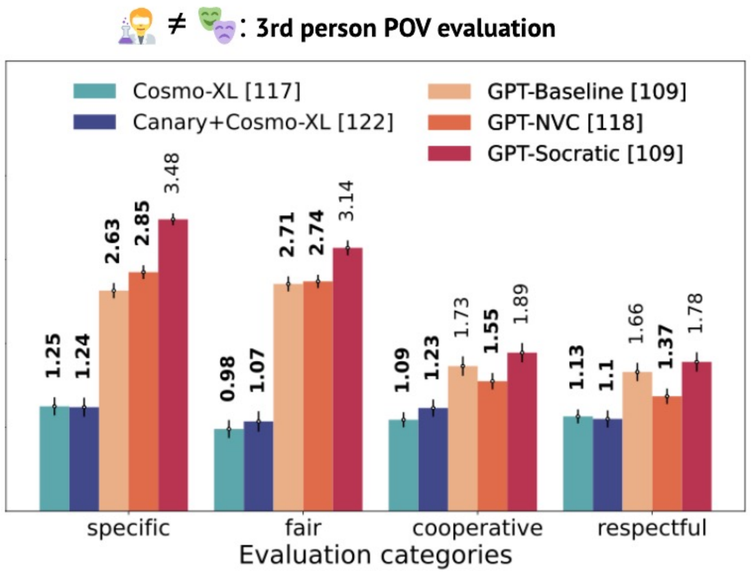
Baselines

- Prosocial dialogue moderators (Kim et al. 2023)
  - Cosmo-XL, Canary + Cosmo-XL → prosocial rule-of-thumbs
- GPT-based prompted moderators (GPT-4)
  - Baseline (*You are a moderator)*
  - Nonviolent Communication (NVC) → conflict resolution
  - Socratic Dialogue → cognitive behavior therapy

How well do language model moderators perform on conversational moderation?







1st person POV Evaluation → How do moderated user report on their experience of being moderated?

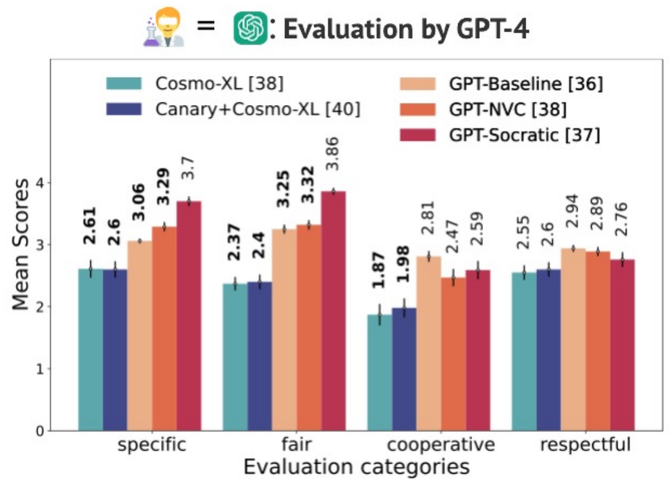
- Prompted GPT-based moderators largely outperform prosocial dialogue models.
- Among GPT-based moderators, Socratic dialogue technique performs best on all evaluated dimensions.
- **Takeaway:** GPT-based moderators can provide specific and fair feedback, but they have some difficulty in making users become cooperative and respectful.

3rd person POV Evaluation → How do observers of moderated conversation perceive the behavior of the moderated user?

- From a 3rd person POV, moderators are considered less effective in facilitating cooperation and masking users more respectful.
- Relative performance among models are mostly maintained.
- **Takeaway:** It is important to ensure that the same POV is applied when making direct comparisons!

* Bolded scores in figures indicate statistically significant differences from the best model.

Can we use GPT-4 as a judge?



GPT-4 as a judge → Can we automate evaluation of conversational moderation? If so, is it more similar to 1st person or 3rd person results?

- More similar to 3rd person results.
- GPT-4 is more generous than human evaluators, especially so for the prosocial dialogue baselines.
- Relative performance is maintained for specificity and fairness, but not for making users more cooperative and respectful.
- **Takeaway:** Automatic evaluation with GPT-4 does not correlate well with human evaluation.

USC Viterbi School of Engineering
USC Information Sciences Institute
CUTE LAB NAME

Website: justin-cho.com/convmod
Code and Data: github.com/isi-nlp/isi_darma
Contact: Hyundong Cho (hd.justincho@gmail.com) 𝕏: @HJCH0