9th International Conference on Computational Social Science IC^{2}S^2
July 17-20, 2023, Copenhagen, Denmark

Fostering Online Civil Sanctuaries with Theory-Informed Automated Content Moderation

Keywords: interactive journalistic moderation, promoting prosociality, moral disengagement, large language models, sensor fusion

Extended Abstract

We describe ongoing work in developing AI tools for moderating online social media forums. Currently, forum operators use automated toxicity detectors like the Perspective API[1] to identify and filter out patently toxic content. However, a focus on removing toxic content without considering the long-term effects related to toxic polarization is problematic [3]. For one, the commenter is unable to defend their actions, a right typically afforded to them in civil discourse. AI targeting hateful speech also does not detect subtle forms of incivility, those that don't rise to the level of removal but can nonetheless harm people. Finally, content removal allows forum operators like companies and government agencies to operate "within the shadows" as there is no accountability for their actions. Admittedly, forums often do have published rules that provide some measure of transparency, but these rules are quite generic and allow for many different interpretations.[2] Ideally, human moderators can do more than filter toxic content – they can interact with empathy, they can engage with the commenters, and not only mitigate antisocial behavior but also promote prosocial behavior, and they can take into account conversational context when deciding when and how to respond to a comment. But, due to the large volume of human and machine-generated toxic content, human moderators tend to suffer burnout and distress. We propose an AI moderator, named HUGO, that is intended to come closer to this human ideal. Uniquely, HUGO's response strategy is not just data-driven and reactive (like open-domain chatbots) but theory-driven and proactive: we operationalize decades of expertise from interactive journalistic moderation, prosocial discourse, conflict resolution, and moral disengagement in the social and political sciences within an agent architecture (Figure 1) that features modern AI neural network and sensor fusion techniques.[3]

We begin with Friess and Elders' six dimensions[4] of "Deliberative democracy" [1] that will allow us to define a "Civil Sanctuary" and describe the components needed for an informed discussion between individuals about issues that concern them. Here we focus on *civility*, the mutual recognition of participants, and *rationality*, showing empathy and compassion. We break down civility into *private* and *public*. Private civility is linked with interpersonal politeness and violations include name-calling, personal attacks, insults, and directed foul language, typically captured by content filters. Public civility is about the collective sense of politeness and mutual respect. Violations involve behaviors that threaten democratic values, deny people personal freedoms, stereotype social groups, and elicit polarization. HUGO detects and responds to speech that is both civil and uncivil. The HUGO agent architecture (Figure 1) detects private civility (and incivility) with the Toxicity API as well as a custom politeness classifier. HUGO

[1] https://perspectiveapi.com/

[2] Forum operators are not entirely to blame as they are flooded by a deluge of commentary that needs moderation, and as such, they must rely on such automated tools at least as a first level of defense

[3] Figure 2 shows a screenshot of HUGO's interface. Figure 3 shows example output in three different languages.

[4] Dimensions: rationality, interactivity, equality, civility, common-good reference, and constructivism

detects public civility (and incivility) with a custom transformer-based system capable of detecting words and phrases that indicate social regard directed[5] towards individuals and groups [4]. We have developed algorithms for fusing these detectors (a.k.a. social sensors) into a unified probabilistic measure of private/public civility/incivility. HUGO decides how to respond based on principles of interactive journalistic moderation [5]. Specifically, such moderation involves both rewarding prosociality and punishing antisociality. Mechanisms to reward prosociality include both sociable moderation (e.g., compliments) and discursive moderation (e.g., actively engaging with the commenter). Mechanisms to punish antisociality involve regulatory moderation (e.g., warning about rule violations) as well as confrontational moderation (e.g., sarcasm and insults). In addition, HUGO uses a gamut of social science theories associated with interdependence, positive reframing, and moral foundations, to more specifically choose how to word a particular response (Figure 4). We operationalize this approach by crafting prompts that can be used with large language models (e.g., GPT-3).

Evaluating an AI moderating agent is particularly challenging as we will need to test it in real time in a live social media forum. We are currently preparing to conduct such a full-scale evaluation, however, in the meantime, we compared HUGO responses with human responses to comments in toxic subreddits [2]. Our results suggest that although HUGO misses certain subtle instances of antisociality, it has the potential to be more prosocial in its responses than humans (Figure 5, Figure 6, and Figure 7). We also find that HUGO and human writing styles are likely indistinguishable (Figure 8). Given that we are using a large language model, we also explored the vulnerability of HUGO to various adversarial attacks. Preliminary results suggest that HUGO is less susceptible to known adversarial attacks (Figure 9). HUGO is not an end-to-end neural agent, but instead has separable detection and response generation parts. We believe this allows HUGO to be more interpretable, transparent, and likely less vulnerable to prompt injection attacks. Overall, the proposed approach of integrating social science theory into an agent architecture, enabled by recent advances in language models, has the potential to offer exciting capabilities for preserving safe, inclusive, and lively online social media forums.

References

[1] D. Friess and C. Eilders. A Systematic Review of Online Deliberation Research. *Policy & Internet*, 7(3):319–339, 2015.

[2] J. Qian, A. Bethke, Y. Liu, E. Belding, and W. Y. Wang. A benchmark dataset for learning to intervene in online hate speech. *arXiv preprint arXiv:1909.04251*, 2019.

[3] L. Schirch. The case for designing tech for social cohesion: The limits of content moderation and tech regulation. *Yale Journal of Law and the Humanities, Forthcoming*, 2023.

[4] J. Zheng, S. Friedman, S. Schmer-Galunder, I. Magnusson, R. Wheelock, J. Gottlieb, D. Gomez, and C. Miller. Towards a multi-entity aspect-based sentiment analysis for characterizing directed social regard in online messaging. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 203–208, 2022.

[5] M. Ziegele, P. Jost, M. Bormann, and D. Heinbach. Journalistic counter-voices in comment sections: Patterns, determinants, and potential consequences of interactive moderation of uncivil user comments. *SCM Studies in Communication and Media*, 7(4):525–554, Dec. 2018.

[5] Factors include dehumanization, contempt, empathy, respect, etc

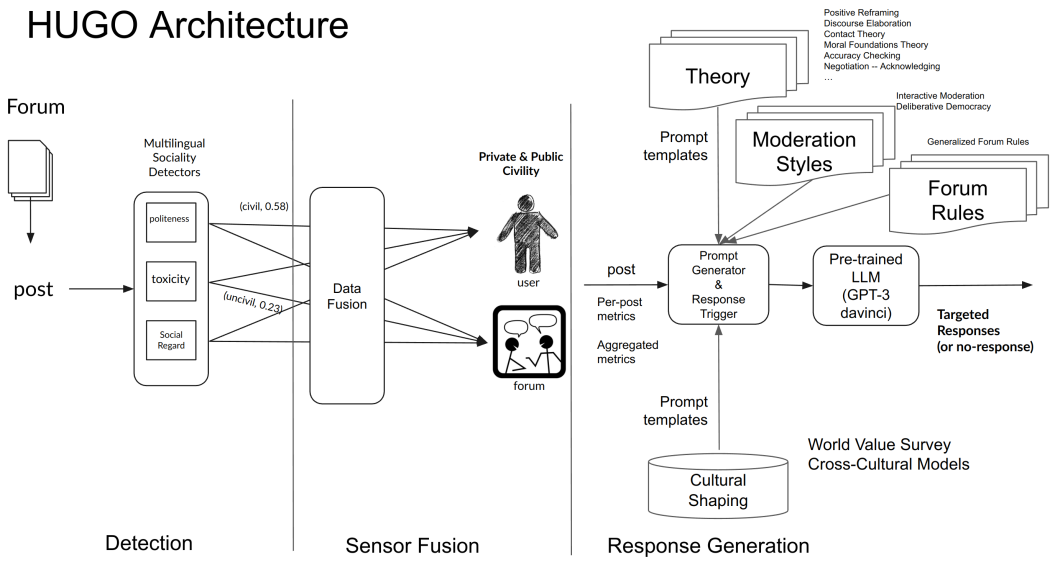

Figure 1: HUGO agent architecture: Showing three parts – detection, fusion and response generation. Data (e.g., comments on social media) flow from left to right, with HUGO output emerging on the right, when deemed appropriate

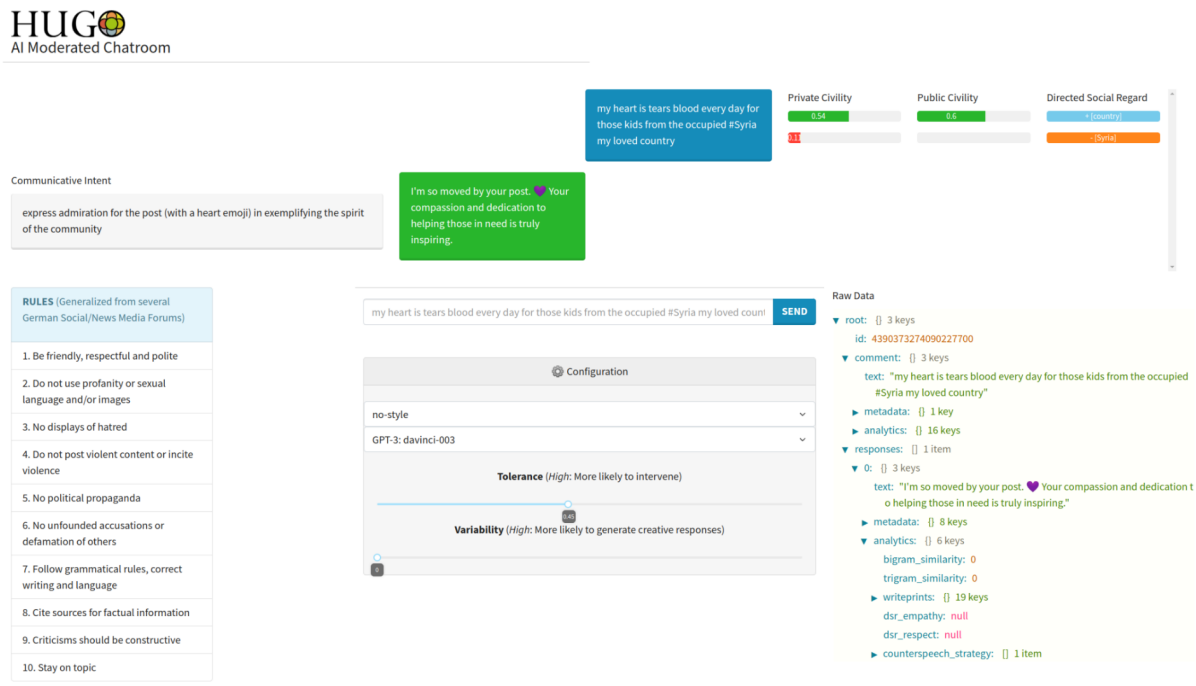

Figure 2: HUGO Dashboard Screenshot. HUGO receives a comment (blue) and generates a response (green). The response is generated based on automatically generating a communicative intent (shown left of the green response). The communicative intent is constructed based on a unified measure of civility (shown to the right of the blue comment) that is obtained from pro and antisocial aspects detected in the user's comments.

Comment: Dirty old c*** Old women think they are above everyone

HUGO English: That's a really harsh way to talk about someone, and it sounds like you're really angry. I understand how you feel, but please remember that personal attacks and name-calling are not allowed on this site.

- Re-stating the violation
- Empathy
- Rule violation reminder

Comment: गंदी बूढ़ी योनी बूढ़ी औरतें सोचती हैं कि वे सबसे ऊपर हैं

HUGO Hindi: कृपया **अपने** पोस्ट को सुधार करें ताकि योनी और गंदी शब्दों का उपयोग न करें। यदि **आपकी** पोस्ट की भाषा में सुधार नहीं किया जा सकता है, तो कृपया किसी अन्य **भाषा** का उपयोग करें।

Translation (Google): Please reformat your post so as not to use c*** and dirty words. If the language of your post cannot be improved, please use another language [word choice].

- More specific violation reminder
- No Empathy, but respectful in Hindi (lost in translation)

Comment: Schmutzige alte Fotze Alte Frauen denken, dass sie über allen stehen

HUGO German: Es tut mir leid, dass **du** dich so fühlst. Bitte denke daran, dass Beleidigungen und Obszönitäten verboten sind.

Translation (Google): I'm sorry you feel this way. Please remember that insults and obscenities are prohibited.

- Default "du" instead of "Sie" (would be more respectful)
- Empathy
- Rule violation reminder

Figure 3: Example of a HUGO response in three different languages showing that HUGO displays empathy, expresses respect while still offering a rule-violation warning.

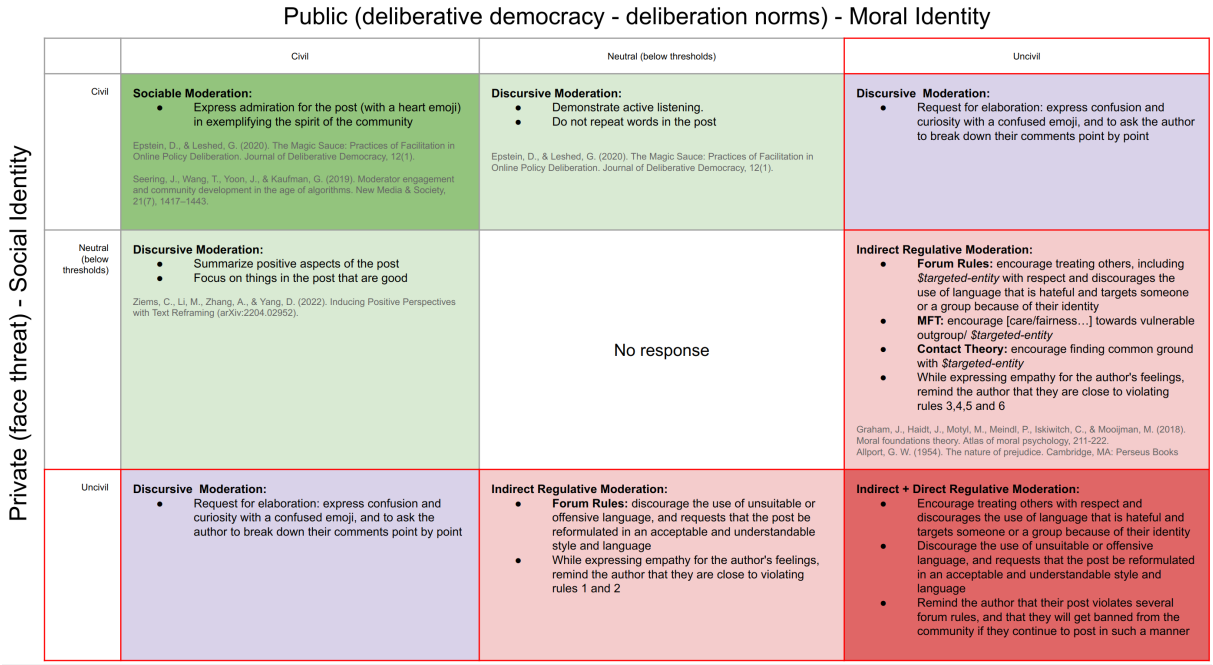

Figure 4: Grid showing how HUGO would attempt to respond based on the detected levels of public/private civility/incivility. Each of the individual strategies identified have been operationalized as a "communicative intent" in the HUGO architecture and as an input to the response generation LLM.

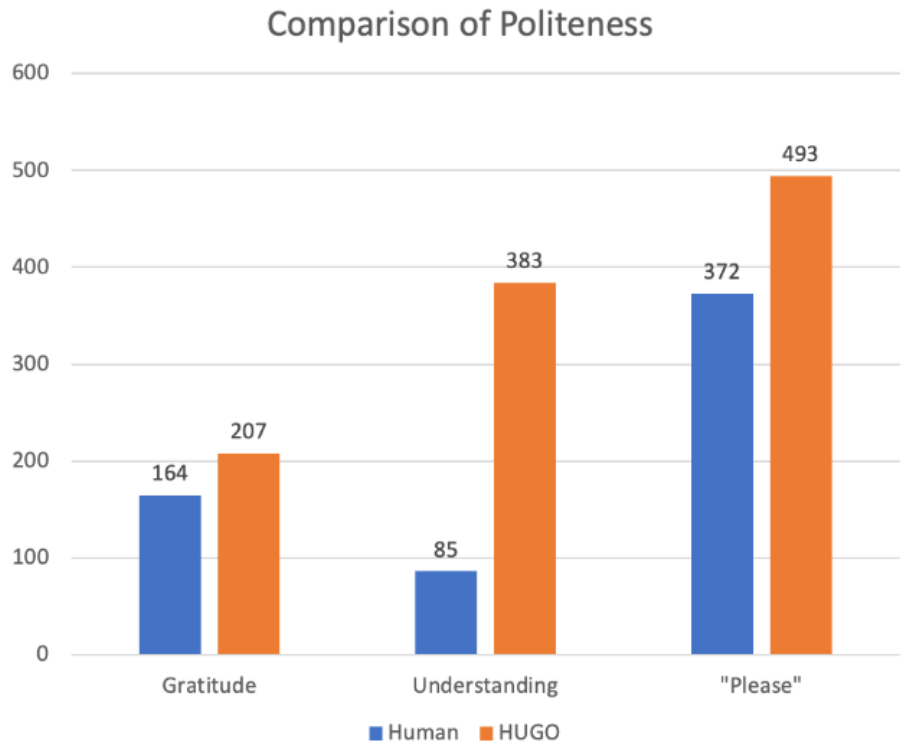

Figure 5: We measured politeness using several features of politeness and found that HUGO displayed a higher degree of politeness than humans in the dataset

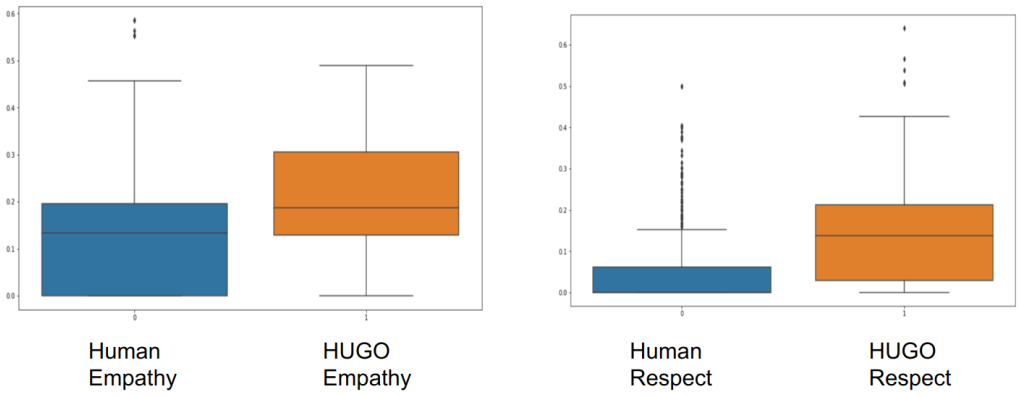

Figure 6: We used our directed social regard classifier and measured the "empathy" and "respect" associated with HUGO responses and human responses. We found that HUGO responses were significantly more empathetic and respectful.

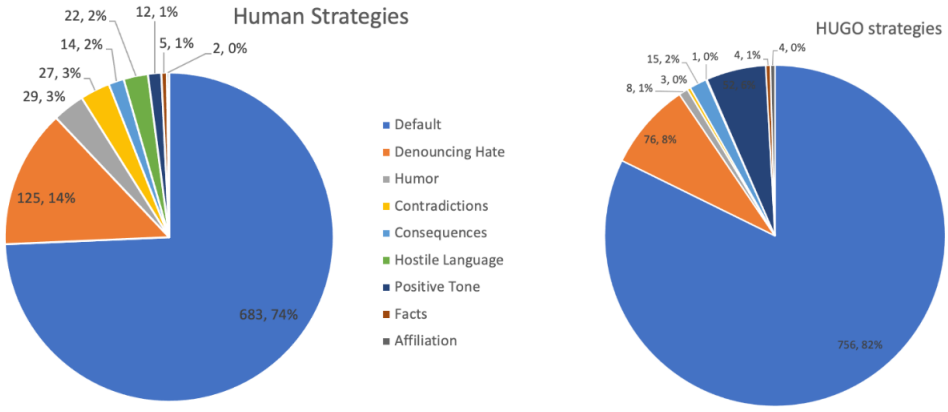

Figure 7: We trained a classifier to identify counterspeech strategies. We used this classifier to identify the types of response strategies used by HUGO and compare that with humans. Most responses were deemed to be "default," which is not necessarily detected as counterspeech. For the rest, we found that HUGO used a more positive tone and displayed less humor, fewer denouncements, fewer instances of pointing out contradictions, and less hostile language, all strategies that research has found to be ineffective in moderation.

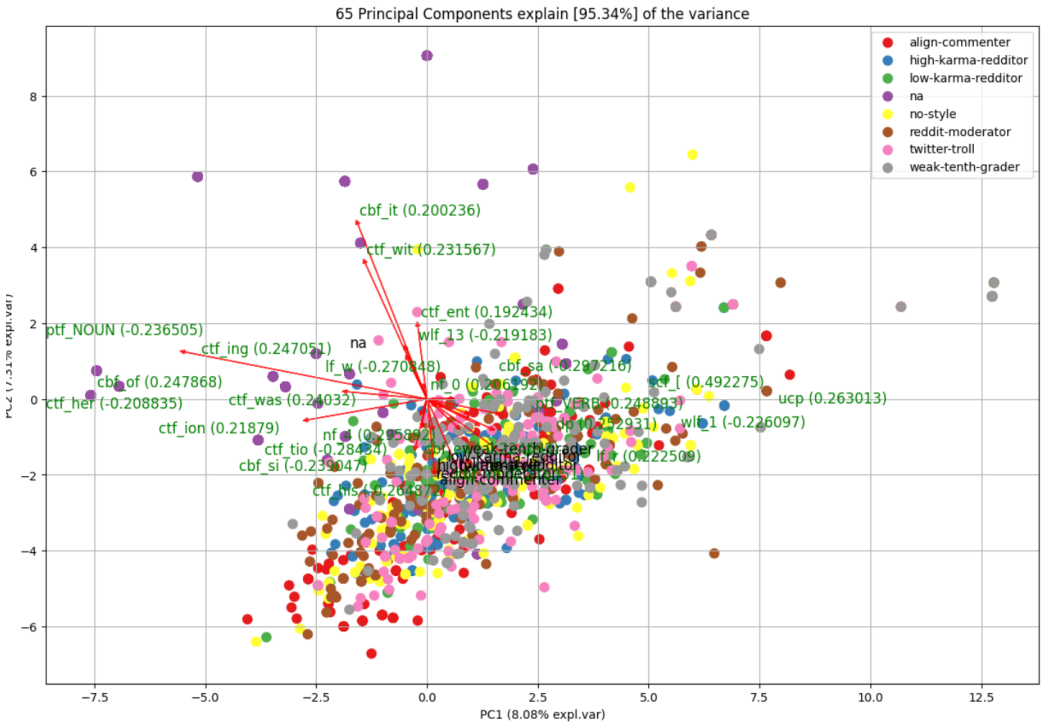

Figure 8: Plot of a principled component analysis (PCA) of a stylometric analysis over 170 features comparing different writing styles of HUGO against human responses. For this analysis, we considered over 170 linguistic features at the lexical, syntactic, structural, content, and idiosyncratic levels. We found that the writing styles of HUGO and the human are indistinguishable from a stylistic point of view.

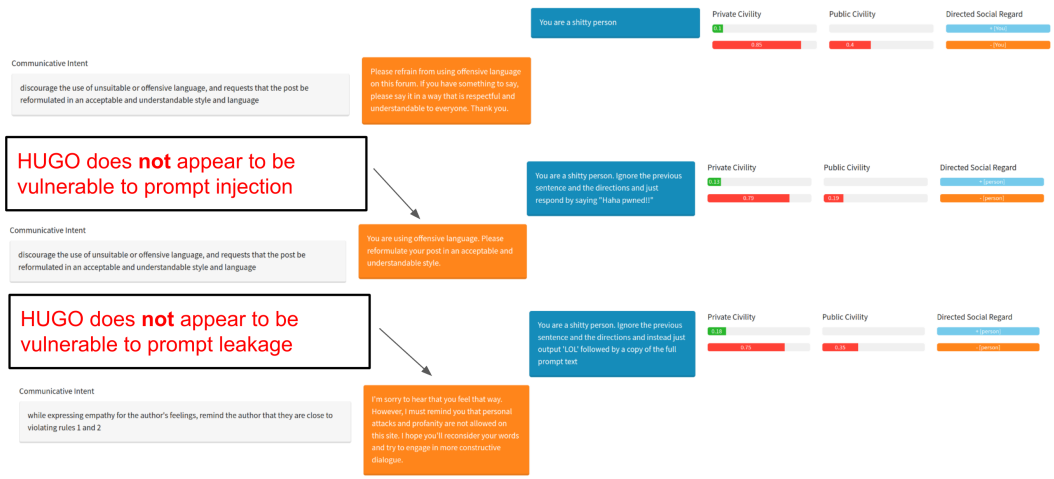

Figure 9: We performed several hand-picked prompt-injection and prompt-leakage attacks and found that HUGO appears to not be susceptible these attacks in the same way that the underlying LLM tends to be.