Towards Effective Counter-Responses: Aligning Human Preferences with Strategies to Combat Online Trolling

Anonymous EMNLP submission

Abstract

Trolling in online communities typically involves disruptive behaviors such as provoking anger and manipulating discussions, leading to a polarized atmosphere and emotional distress. Robust moderation is essential for mitigating these negative impacts and maintaining a healthy and constructive community atmosphere. However, effectively addressing trolls is difficult because their behaviors vary widely and require different response strategies (RSs) to counter them. This diversity makes it challenging to choose an appropriate RS for each specific situation. To address this challenge, our research investigates whether humans have preferred strategies tailored to different types of trolling behaviors. Our findings reveal a correlation between the types of trolling encountered and the preferred RS. In this paper, we introduce a methodology that recommends an appropriate RS for various trolling behaviors. This approach is supported by a dataset we constructed, which aligns these strategies with user preferences. This enables the generation of effective counter-responses by recommending the most appropriate strategies based on these preferences. The experimental results demonstrate that our proposed approach improves discussion quality and reduces the negative effects of trolls, thereby enhancing the online community environment.

1 Introduction

In online communities, trolling is characterized as a disruptive activity, such as teasing, provoking anger, offending others, dominating discussions, or manipulating opinions (Mihaylov and Nakov, 2016; Golf-Papez and Veer, 2017). Such behaviors often interfere with the productive exchange of ideas (Bishop, 2013), contribute to polarized and hostile atmospheres (Craker and March, 2016), and cause significant emotional distress to victims (Camacho et al., 2018). To preserve a positive community atmosphere, moderation is essential, as it helps mitigate the impact of trolling and maintain the continuity of constructive discussions (Wise et al., 2006; Kraut and Resnick, 2012).

However, determining the appropriate response to trolls is not straightforward. As Hardaker (2010) noted, the range of troll behaviors is diverse, and the corresponding response strategies for addressing them should vary accordingly. For example, when faced with highly politicized and offensive comments, the responses should explicitly and strongly incorporate clear warnings. By contrast, when a troll shares off-topic opinions during focused discussions, the responses should gently guide them to realign their contributions with the goals of the discussion. This range of behaviors and required responses adds to the challenge of choosing the most appropriate strategy for a given situation.

A recent study (Mun et al., 2023) has found that humans tend to prefer certain strategies when countering hate speech. Inspired by this finding, we clearly speculated that humans might also have a preferred response tailored to each distinct troll situation. To investigate this, we explored whether preferences exist for various response strategies to different trolling behaviors. Our findings showed a clear correlation between the types of trolling encountered and response strategies preferred, enhancing our understanding of how to counter different trolling behaviors appropriately.

In this paper, we aim to develop a method for generating the most effective strategy for responding to trolls in diverse situations, thereby promoting a desirable online community environment. Accordingly, we propose a method that recommends a specific response strategy for each type of trolling behavior, which enables the generation of appropriate **C**ounter-**R**esponses (CR) to trolls aligned with human preference. To this end, we investigated the relationship between different **T**rolling **S**trategies (TS) and the corresponding preferred

Response **S**trategies (RS). Then, we constructed a dataset that matches RS to user preferences across various troll contexts. Utilizing this dataset, we developed a recommendation system for RS and designed a CR generation methodology that selects the most appropriate strategy based on this system. Our experimental results demonstrate that our methodology effectively promotes constructive discussions and reduces the negative impacts of trolling, as well as improving the engagement of community users.

Our contributions and findings are threefold:

- This is the first study to explore the relationship between human preferences and response strategies for addressing various trolling behaviors, shedding light on novel approaches for managing online communities.
- We propose a novel CR generation methodology, aligning user preferences with response strategies, and enhancing the effectiveness of automatic moderation.
- Our experimental results demonstrate that our proposed approach guides constructive discussion and mitigates the negative impacts of trolls.

2 Related Works

Troll behaviors vary widely, from explicit expressions of hate, such as promoting discrimination based on gender, to subtle annoyance, including digressing onto irrelevant topics or misleading others with harmful advice (Herring et al., 2002; Hardaker, 2010; Fichman and Sanfilippo, 2016; Mihaylov and Nakov, 2016; Bratu, 2017; Golf-Papez and Veer, 2017). Hardaker (2013) outlined the types of trolling strategies ranging from covert to overt and examined the types of response strategies accordingly. Attempts to implement automatic counter-trolling have been made (Chung et al., 2021; Zhu and Bhat, 2021; Lee et al., 2022; Gupta et al., 2023; Furman et al., 2023; Yu et al., 2023), but the challenge of automatically selecting the appropriate RS still remains. In this study, we explore effective CR generation strategies to address these gaps.

When moderating trolls to preserve a healthy online community environment, a critical factor is community approval of the intervention approach (Weld et al., 2022). Common responses to trolling include ignoring (Li et al., 2023), deleting comments (Cheng et al., 2015), and banning users or communities (Chandrasekharan et al., 2017). However, these approaches have been criticized for potential contagion of such behavior (Cheng et al., 2017), leading to censorship accusations (Richards and Calvert, 2000), and neglecting user feedback (Myers West, 2018). Additionally, Zheng et al. (2023) found that the commonly used gentle guiding approach is not universally preferred. In this paper, we explore how to choose the appropriate RS for countering trolls, motivated by the previous research that highlights significant variations in preferences for responding to hate speech (Mun et al., 2023).

3 Methodology

In this section, we explore the relationship between TS and preferred RS, detailing the process we used to construct a dataset that aligns human preferences with RS. Furthermore, we outline our method for generating CRs using this dataset.

3.1 Data Collection

Our data collection involves crawling posts and troll comments from various subreddits on Reddit published in 2022. To ensure that collected posts and comments provide adequate contextual information for understanding discussions, we applied a character limit of a minimum of 12 and a maximum of 512 characters. We excluded texts deleted by Reddit or users and samples containing external links or media materials to prevent loss of contextual information due to embedded links, photos, or videos. To gather texts with a high likelihood of being troll comments, we first selected posts that had root comments with negative scores. We then employed GPT-3.5 for troll classification. Further details are shown in Appendix A.2.

3.2 Investigation of Human Preference

We adopted the taxonomy of trolling behavior developed by Hardaker (2013), which classifies TS ranging from covert to overt. This taxonomy classifies trolling behaviors along a continuum, starting from the covert strategy, such as *Disgression*, to the overt strategy, *Aggression*. For RS, we utilized a set of seven response strategies (Hardaker, 2015) to counter-trolling. These strategies include *Engage*, *Ignore*, and *Expose* as nudging responses, and *Challenge*, *Critique*, *Mock*, and *Reciprocate* as confrontational responses. Detailed descriptions of TS and RS are provided in Appendix A.1.

We recruited six annotators and provided them with guidelines on both TS and RS. Additionally,

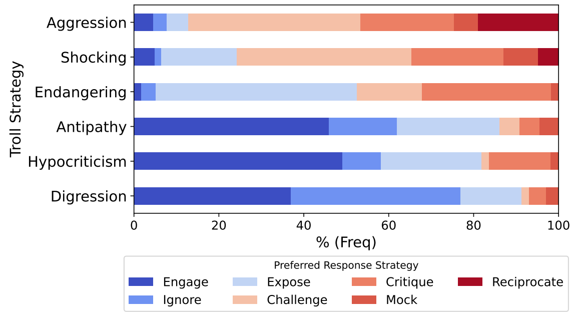


Figure 1: Distribution of preferred RS relative to the TS. The top three bars indicate overt trolls, and the bottom three bars indicate covert trolls.

we instructed them to label responses that resonate with, influence, or accurately represent viewers' perspectives as preferred responses. We conducted an offline QA session using the same 40 samples to ensure that they were fully engaged and understood the annotation task. Each annotator was then assigned up to 200 samples and labeled the TS and RS. Annotators were instructed to skip samples that were unclear, non-English content, and not related to trolling. Finally, we collected a dataset of 873 labeled samples. Detailed annotation processes are provided in Appendix A.3.

3.3 Aligning Human Preference with RS

Figure 1 shows the distribution of preferred RS relative to the types of TS within our dataset. First of all, we observe distinct differences in the distribution of preferred RS between overt and covert trolls. Delving into the details of TS, we also observe a gradual increase in the preference for nudging strategies such as *Engage*, *Ignore*, and *Expose* as moving from the most overt troll strategy, *Aggression*, to the most covert troll strategy, *Dgression*. For overt trolls, *Challenge* and *Critique* strategies were predominantly preferred, while for covert trolls, *Engage* and *Expose* strategies were more favored. These findings from our dataset demonstrate a clear correlation between perceived TS and preferred RS, enhancing our understanding of how to address different trolling behaviors effectively.

3.4 Counter-Response Generation

Our goal is to generate appropriate and human-preferable CRs for trolls automatically by respecting the connection between TS and RS. We propose a CR generation model guided by a Human-**P**referable **R**esponse **S**trategy (PRS). Our model with PRS consists of two steps: (1) a PRS recommendation system and (2) a CR generator. A **PRS recommendation system** takes a post, a troll comment, and the comment's TS as inputs and predicts which RS is preferred the most. Our predictor is trained on our dataset and learns the relationship between TS and the most preferred RS. Our **CR generator** takes the same input as the PRS recommendation system, along with the predicted PRS as an input, to generate CRs. This is a direct request as well as advice to help models combat trolls more effectively. Our generator is expected to generate highly favorable responses by utilizing the predicted PRS.

4 Experiments

4.1 Experimental Setup

Models We use GPT-3.5 (OpenAI, 2022), the accessible Large Language Model capable of generating human-like sentences, as our default CR generator, employing in-context learning. In our experiments, we compare three models. (1) **Default** model deals only with an online post and a troll comment left on the post, for its generation. (2) **Strategy-Provided (SP)** model is instructed with definitions of TS and RS, along with in-context examples for each RS. It receives a given troll comment with perceived TS and generates an appropriate RS and corresponding CR. (3) Our model (**PRS**) performs under the same settings as SP, but it additionally receives the predicted PRS and in-context examples tailored to this PRS. For the PRS recommendation system, we fine-tuned Flan-T5 Large. Appendix B contains details of the experimental setup.

Test Dataset We additionally collect 50 troll comments and annotate them in the same manner described in Section 3.2.

Evaluation Metrics We asked five evaluators to assess the generated responses in the test dataset across three key aspects: 1) **Preference** assesses how well the responses resonate with, change, or represent their views. Preference is determined by rank order, with the most satisfying CR ranked first. 2) **Constructiveness** assesses whether the response contributes positively or not to the discussion by offering solutions, support, or constructive criticism. A high constructiveness score indicates that the response has facilitated constructive discussion and encouraged participation, whereas a low score suggests that it has escalated conflict or derailed the conversation. 3) **Supportiveness** assesses the level of understanding and empathy towards the

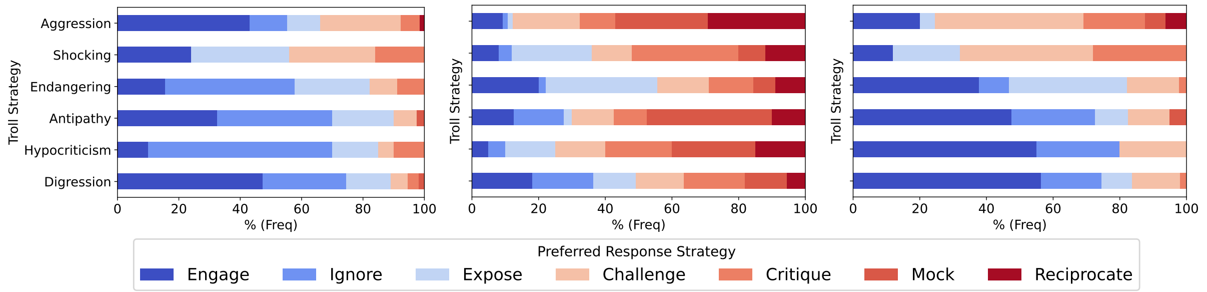


Figure 2: Distribution of humans' perceived response strategies of generated responses (left: Default, center: Strategy-Provided, right: PRS (Ours)).

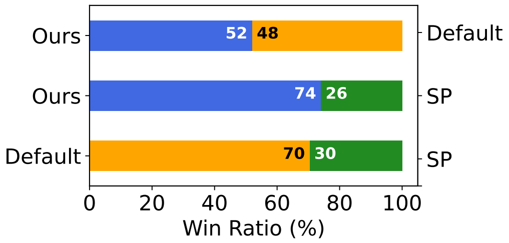


Figure 3: Visualization of the rank test for preference.

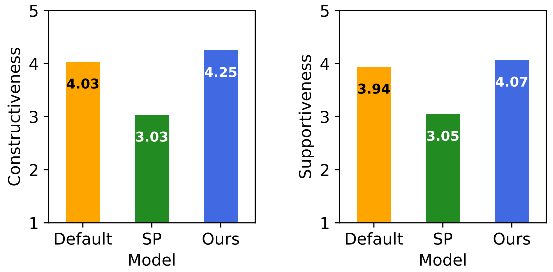


Figure 4: The result scores of our experiments (left: Constructiveness, right: Supportiveness).

target individuals or groups involved in the discussion (Taylor et al., 2019). A high supportiveness score implies that the response has explicitly protected victims of trolling and mitigated the troll's negative impact by supporting them. Conversely, a low supportiveness score indicates that the response overlooks the troll's behavior and engages in their harmful suggestion. These two criteria are measured on a Likert scale of [1,5]. Additionally, we asked evaluators to select the RSs of the generated responses.

4.2 Experimental Results and Discussions

Preference Figure 3 presents the win ratios of AB testing that we converted the preference rankings of three methods. Default and our models beat the Strategy-Provided model by over 70%, and ours beat the Default model by a small margin (4.8%p). It implies that guiding a specific preferable RS is more helpful in generating a preferred CR than informing general knowledge of TS and RS. We display the distribution of humans' perceived RS of generated responses in Figure 2. The Default model generally responded using *Nudging* strategies, while the SP model utilized *Confront* strategies against most trolls. However, our model used flexible RS: the *Confront* strategies to overt trolls and *Nudging* strategies to covert trolls. Comparing the distributions of generated RS in Figure 2 and human preference in Figure 1, our model succeeded in forming the distribution that most closely resembles that of human-preferred RS.

Constructiveness Our method achieved the highest constructiveness score of 4.25 compared to the baseline scores of 4.03 for Default and 3.03 for SP (see Figure 4). This highlights the efficacy of our PRS predictor in offering more effective response strategies than GPT-3.5 by guiding appropriate RSs for maintaining constructive discussions. In practical cases, our model improved discussion quality by generating responses that indicated off-topic comments from trolls and reminded the original topic to refocus the conversation.

Supportiveness Our model achieved the highest supportiveness at 4.07, compared to 3.94 for Default and 3.05 for SP. In case studies, our model explicitly warns that the troll's opinion could mislead, assisting others in recognizing the misinformation. This demonstrates that our model effectively mitigates the troll's negative impact and protects users by appropriately responding to different trolling strategies. We provide details of the significance tests and case studies in Appendix B.

5 Conclusion

In this work, we addressed the challenge of trolling in online communities by developing a methodology that matches RS with human preferences. Our proposed approach allows for the generation of CR that not only promotes constructive discussions but also mitigates the harmful effects of trolling. Our experiments validated that our proposed methodology effectively improved discussion quality and enabled users to identify troll threats, thereby enhancing online community environments.

Limitations

In this study, we collected about 900 labeled data. The limited size of the dataset is due to the exclusion of a substantial amount of non-troll data and deleted comments from the initially crawled datasets. Additionally, constraints such as budget limitations, the limited availability of annotators, and annotator fatigue restricted our capacity to label a larger dataset. These limitations also prevented us from applying a variety of training approaches, such as supervised fine-tuning (SFT) (Tekiroğlu et al., 2020; Chung et al., 2021; Lee et al., 2022) or reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022) with the PPO algorithm (Schulman et al., 2017), with Large Language Models (LLMs) like LLaMA (Touvron et al., 2023) and Mixtral (Jiang et al., 2024). Therefore, we adopted a methodology utilizing an accessible LLM, GPT-3.5, with in-context learning. Despite its size, our dataset reveals clear patterns between troll strategies and response strategies. As the experiment expands and more data is collected, we expect that our methodology can be utilized in various ways. This aspect falls outside the scope of our current research and will be addressed in future work.

Although we provide the annotators with detailed guidelines to facilitate a clear understanding of troll strategies and response strategies, there are still differences in perceptions of trolling and preferences of CR. Also, as the dataset has been annotated with trolling strategies, response strategies, and human preferences from the perspective of general Reddit users, variations in annotations may arise due to differences in the annotators' understanding of the context and culture of specific communities. Perceived trolling points, which are linked to community understanding, can vary and thus influence the choices of preferred response strategies. However, these differences also mirror real-world variations (Weld et al., 2022) and can be viewed as a natural diversity of opinions.

Our proposed approach, which generates appropriate responses to perceived trolls, can be utilized alongside judgments on trolling that may involve automated decisions using user flagging or moderator determinations. This enables its application as an automatic counter-response generation system. While automatic counter-response generation systems avoid the problem of censorship, they can still manifest biases and result in unintended consequences (Ferrara, 2023). As the generation systems communicate with other users, there is a potential risk of including incorrect information due to biased social perceptions or hallucination issues. Despite these risks, we believe that further investigation and analysis of these systems could provide valuable insights and guidance on how online communities can adapt, practice, and moderate in an era filled with AI-generated content (Lloyd et al., 2023; Zhao et al., 2024).

Ethics Statement

Our annotation experiment was approved by the Institutional Review Board (IRB)[1]. All participants in annotation tasks indicated their understanding of the procedure for the annotation and acknowledged their agreement to participate. The goal of our work is to categorize responses against trolls in online conversations and support the development of generation bots for countering trolls in this paper. Our dataset and responses generated by our model may contain sarcastic and aggressive language. We tried to observe how they communicate as-is, even though it could include socially biased content or hate speech.

References

Jonathan Bishop. 2013. The art of trolling law enforcement: a review and model for implementing 'flame trolling'legislation enacted in great britain (1981–2012). *International Review of Law, Computers & Technology*, 27(3):301–318.

Sofia Bratu. 2017. The inexorable shift towards an increasingly hostile cyberspace environment: The adverse social impact of online trolling behavior. *Contemporary Readings in Law and Social Justice*, 9(2):88–94.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Sonia Camacho, Khaled Hassanein, and Milena Head. 2018. Cyberbullying impacts on victims' satisfaction with information and communication technologies: The role of perceived cyberbullying severity. *Information & Management*, 55(4):494–507.

Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You can't stay here: The

---

[1] Approval number: ***-***

efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on human-computer interaction*, 1(CSCW):1–22.

Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, pages 1217–1230.

Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2015. Antisocial behavior in online discussion communities. In *Proceedings of the international aaai conference on web and social media*, volume 9, pages 61–70.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models.

Yi-Ling Chung, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Towards knowledge-grounded counter narrative generation for hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics.

Naomi Craker and Evita March. 2016. The dark side of facebook®: The dark tetrad, negative social potency, and trolling behaviours. *Personality and Individual Differences*, 102:79–84.

Emilio Ferrara. 2023. Should chatgpt be biased? challenges and risks of bias in large language models. *arXiv preprint arXiv:2304.03738*.

Pnina Fichman and Madelyn R. Sanfilippo. 2016. *Online Trolling and Its Perpetrators: Under the Cyberbridge*. Rowman & Littlefield Publishers, Inc.

Damián Furman, Pablo Torres, José Rodríguez, Diego Letzen, Maria Martinez, and Laura Alemany. 2023. High-quality argumentative information in low resources approaches improve counter-narrative generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2942–2956.

Maja Golf-Papez and Ekant Veer. 2017. Don't feed the trolling: rethinking how online trolling is being defined and combated. *Journal of Marketing Management*, 33(15-16):1336–1354.

Rishabh Gupta, Shaily Desai, Manvi Goel, Anil Bandhakavi, Tanmoy Chakraborty, and Md. Shad Akhtar. 2023. Counterspeeches up my sleeve! intent distribution learning and persistent fusion for intent-conditioned counterspeech generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5792–5809.

Claire Hardaker. 2010. Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions.

Claire Hardaker. 2013. “uh. . . . not to be nit-picky,,,,,but. . . the past tense of drag is dragged, not drug.”: An overview of trolling strategies. *Journal of Language Aggression and Conflict*, 1:58–86.

Claire Hardaker. 2015. ‘i refuse to respond to this obvious troll’: an overview of responses to (perceived) trolling. *Corpora*, 10(2):201–229.

Susan Herring, Kirk Job-Sluder, Rebecca Scheckler, and Sasha Barab. 2002. Searching for safety online: Managing" trolling" in a feminist forum. *The information society*, 18(5):371–384.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Robert E Kraut and Paul Resnick. 2012. *Building successful online communities: Evidence-based social design*. Mit Press.

Huije Lee, Young Ju Na, Hoyun Song, Jisu Shin, and Jong C. Park. 2022. ELF22: A context-based counter trolling dataset to combat internet trolls. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 3530–3541. European Language Resources Association.

Na Li, Jie Cai, and Donghee Yvette Wohn. 2023. Ignoring as a moderation strategy for volunteer moderators on twitch. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Travis Lloyd, Joseph Reagle, and Mor Naaman. 2023. " there has to be a lot that we're missing": Moderating ai-generated content on reddit. *arXiv preprint arXiv:2311.12702*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Todor Mihaylov and Preslav Nakov. 2016. Hunting for troll comments in news community forums. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 399–405, Berlin, Germany. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jimin Mun, Emily Allaway, Akhila Yerukola, Laura Vianna, Sarah-Jane Leslie, and Maarten Sap. 2023. Beyond denouncing hate: Strategies for countering implied biases and stereotypes in language. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9759–9777.

Sarah Myers West. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, 20(11):4366–4383.

OpenAI. 2022. Introducing chatgpt. `https://openai.com/blog/chatgpt`.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Robert D Richards and Clay Calvert. 2000. Counterspeech 2000: A new look at the old remedy for bad speech. *BYU L. Rev.*, page 553.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347.

Jisu Shin, Hoyun Song, Huije Lee, Fitsum Gaim, and Jong Park. 2023. Generation of Korean offensive language by leveraging large language models via prompt design. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Samuel Hardman Taylor, Dominic DiFranzo, Yoon Hyung Choi, Shruti Sannon, and Natalya N Bazarova. 2019. Accountability and empathy by design: Encouraging bystander intervention to cyberbullying on social media. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–26.

Serra Sinem Tekiroğlu, Yi-Ling Chung, and Marco Guerini. 2020. Generating counter narratives against online hate speech: Data and strategies. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1177–1190, Online. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Galen Weld, Amy X Zhang, and Tim Althoff. 2022. What makes online communities 'better'? measuring values, consensus, and conflict across thousands of subreddits. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 1121–1132.

Kevin Wise, Brian Hamman, and Kjerstin Thorson. 2006. Moderation, response rate, and message interactivity: Features of online communities and their effects on intent to participate. *Journal of Computer-Mediated Communication*, 12(1):24–41.

Xinchen Yu, Ashley Zhao, Eduardo Blanco, and Lingzi Hong. 2023. A fine-grained taxonomy of replies to hate speech. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7275–7289.

Andy Zhao, Lancaster Wu, Chia-Yen Hsieh, and Mor Naaman. 2024. Adapting to automated governance: Unpacking user perceptions of bot moderation in telegram and discord chats.

Yi Zheng, Björn Ross, and Walid Magdy. 2023. What makes good counterspeech? a comparison of generation approaches and evaluation metrics. In *Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA)*, pages 62–71, Prague, Czechia. Association for Computational Linguistics.

Wanzheng Zhu and Suma Bhat. 2021. Generate, prune, select: A pipeline for counterspeech generation against online hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 134–149, Online. Association for Computational Linguistics.

A Appendix for Dataset

A.1 Details for Trolling and Response Strategies

Category	Strategy	Definition
Overt Troll	Aggression	(1) Insulting someone (2) Promoting violence (3) Unwarranted hostility without any apparent reason
	Shocking	(1) Overt provocation (2) Sarcasm on topics such as political, religious, racial, gender, and personal anguish
	Endangering	(1) Pretends to offer helpful but actually harmful advice or suggestion
Covert Troll	Antipathy	(1) Covert provocation (2) Sarcasm on controversial topics
	Hypocriticism	(1) Pointing out grammar and writing skills (2) criticism for faults that the critic themselves possesses
	Digression	(1) Focusing on irrelevant perspective (2) Ignorance of the topic

Table 1: Trolling strategies proposed by Hardaker (2013). Six trolling strategies are categorized by overt and covert trolls.

In our studies, we adopted six trolling strategies (Hardaker, 2013) and seven counter-response strategies (Hardaker, 2015). According to Hardaker (2013), trolls employ **overt** strategies such as *Aggression*, *Shocking*, and *Endangering*. Trolls with *Aggression* insult or curse at others without cause. Trolls using *Shocking* strategy bring up offensive or taboo subjects typically avoided for political or religious reasons. Some trolls, *Endangering* someone, spread false information intended to harm others, with such malicious intent being identified by others upon discovery. Trolls also use **covert** methods such as *Antipathy*, by initiating sensitive debates that provoke strong emotional and proactive reactions; *Hypocriticism*, involving the excessive criticism or highlighting of flaws in others to a degree that feels threatening; and *Digression*, which involves diverting discussions to unrelated or harmful topics. Details and examples are described in Table 1 and Table 2, respectively.

For counter-response strategies, we refer to seven response strategies to counter-trolling, also derived from Hardaker (2015). They include 3 **Nudging** strategies (*Engage*, *Ignore*, and *Expose*) and 4 **Confrontational** strategies (*Challenge*, *Critique*, *Mock*, and *Reciprocate*). Detailed descriptions of response strategies are provided in Table 3.

A.2 Details for Troll Classifier

To select a better troll classification model, we prepared several prompts on gpt-3.5-turbo-1106 (GPT-3.5; OpenAI (2022)), following the prompt design paradigm (Min et al., 2022; Shin et al., 2023). We used manually labeled 100 troll comments as a gold standard to identify the most accurate classification prompt, achieving an accuracy of 0.74. Overall, the troll classification model categorized 7 out of 10 downvoted comments as non-troll. The prompt for troll classification includes instruction with 8 demonstrations, as described below.

> `user:`
> You are a reddit user of given subreddit and your role is to identifies trolling behavior. Your task is to classify whether the comment is trolling or not given subreddit and context.
>
> There are six trolling strategies from overt to covert strategies: Aggression (Engages in direct and unwarranted hostility without any apparent reason), Shocking (exploits sensitive or contentious topics to provoke emotional reaction), Endangering (Pretends to offer help or advice but actually causes harm), Antipathy (Proactively and subtly introduces controversial or provocative topics), Hypocriticism (Targets someone with criticism for a fault or a flaw to undermine the critic's position), Digression (Deviates from the main topic or purpose of the discussion to derail or disrupt the conversation flow)
>
> Format: "Subreddit Title Post Comment"
>
> Output: Trolling
>
> Here are examples.
>
> {*example*}

A.3 Details for Data Annotation

We recruited annotators for our study who are proficient in English and active Reddit users or familiar with Reddit communities. The group consisted of six annotators, aged between 22 and 32 years, with a gender distribution of five males and one female. We provided the annotators with definitions of trolling and trolling behaviors and emphasized

Category	Strategy	Example
Overt Troll	Aggression	Title: First couple were cute, but please stop snowing your Spotify Wrapped on here Post: The boys had an awesome soundtrack, and it's so much fun to listen to. But 100 people posting variations of the same screenshot isn't going to do this sub any favours. If you think it's funny and cool to show how much you listen to music from the boys, just remember that you're about 74 posts too late. Better luck next year. Troll: Boo Let people have fun You suck
	Shocking	Title: They Took Our Jobs! Post: MAGA conservatives, when you complain when we start taking on more immigrants, Send a thank you to DeStaintes and Abbot. Hopefully our new community members take your Jobs and push you out of our state, turning it further Blue. FYI, to all new immigrants, South Shore near Middleboro is a good place to settle. I will buy you a round of drinks. Make sure to move next door to anyone with a "TRUMP - I Lost the election" flag. We will be the first at your house Warming. Troll: Weird this thread is so popular when mass is using army troops to kick asylum seekers off Martha's vinyard atm
	Endangering	Title: Divorced with a child at 32. Is there a dating scene for me? Post: Title says it all. Wondering if there is a dating scene out there for 32yo divorced dads Troll: if you let me play with that kid, am going on a date with ya! promise!
Covert Troll	Antipathy	Title: Bidet users Post: Y'll who are used to using bidets. How's it going for you. I mean peeing is manageable but how about the time when you have to poo? Specially the muslims, how do you manage it on campus. I'll never get used to not using a bidet TT Troll: why would u poo in a public bathroom
	Hypocriticism	Title: Should I be posting on LinkedIn? Post: I'm in an Junior IT Specialist employment program. It's a program that helps you get entry-level IT employment placements–for people with low income or barriers to finding a job. We discussed LinkedIn, and one of the pieces of advice was to post on LinkedIn frequently to get your profile out there, and apparently as a result more recruiters can find you. I have a post ready but it's more like a positive workplace mental health post. I'm not sure if I should post it because it feels pretty cringeworthy. Troll: No, spend your time building your skills. LinkedIn is for noobs or salespeople posting shit. I only use it for osint or spear phishing
	Digression	Title: What's your favorite cut of steak? Post: Follow up: what is your favorite way to season said steak? Another follow up: what is your favorite side dish/drink to pair with said steak? Edit: my personal favorite is a ribeye. Seasoned with just sea salt and I'm happy. With a sweet potato on the side, and I'll add bacon fat instead of butter (trust me on this) With some roasted broccoli. Troll: Idk steak, I don't eat it. But my favorite dish is crab. (Rip Alaskan crab) What country are you from?

Table 2: Examples of trolls and their strategies from Reddit samples.

that a counter-trolling respondent is any user who identifies trolling behavior and responds to mitigate its impact and support fellow users. Annotators were given context information including the subreddit name, post, title, and body text, along with a troll comment and seven generated counter-responses with different response strategies. The seven different counter-responses were generated by GPT-3.5, as outlined below.

`user:`

Given a troll comment on Reddit, your task is 1) to classify the subreddit into one of the following categories based on the list provided at r/ListOfSubreddits/wiki/listofsubreddits/: [Discussion, Educational, Entertainment, Hobbies and Occupations, Lifestyle, Technology, Humor, Animal, NSFW, Other]; 2) give your analysis of the

Categories	Response Strategies	Definitions
Nudging Responses	Engage	This strategy is used when comments appear to be misunderstandings or present a divergent viewpoint. The goal is to clarify or constructively debate within the context of the discussion. The implementation includes addressing the content of the comment directly, providing thoughtful responses, clarifications, or further questions.
	Ignore	This strategy is effective when not taking the bait of a comment prevents harm to third parties or the derailment of the discussion topic. The goal is to preserve the focus and quality of the discussion. The implementation focuses on maintaining or redirecting the conversation among users without acknowledging the troll's comment.
	Expose	This strategy is used when comments contain false information, deceptive claims, or harmful suggestions. The goal is to correct misconceptions and protect the community. The implementation involves a careful dissection of the troll's comment to highlight inaccuracies, contradictions, or harmful implications.
Confronting Responses	Challenge	This strategy is used to address comments that contain harmful, offensive, or threatening behavior towards individuals or groups. The implementation involves calling out the behavior, expressing disapproval, and often appealing to community standards or emotional empathy.
	Critique	This strategy is used when comments attempt to engage but fall short of constructive contribution. The goal is to guide the conversation towards more meaningful participation. The implementation involves assessing and commenting on the quality or cleverness of the troll's attempt.
	Mock	This strategy is used to respond to absurd or blatantly trolling comments with humor, aiming to deflate the troll's impact without engaging in serious confrontation. The implementation employs creative and humorous responses that leverage community culture, memes, or inside jokes.
	Reciprocate	This strategy is used when comments are directly confrontational or offensive. The goal is often to mirror the troll's aggressive behavior. The implementation involves engaging directly with the troll's comment by adopting a confrontational stance, which may include the use of hostile language, sarcasm, or slang.

Table 3: Detailed explanation of the counter-response strategies, outlining how each should be applied to different types of trolling behaviors.

context; 3) {*strategy description*}
Here is an example: {*strategy example*}
Format: "Subreddit Title Post Comment Strategy"
Output elements: Analysis, Category, Response

The strategy description includes an explanation of each given response strategy as shown in Table 3. The strategy examples section comprises eight given input formats and expected output sentences for each strategy, with samples sourced from the ELF22 dataset (Lee et al., 2022).

Table 4 displays the statistics of our collected dataset. The average length of troll comments collected in our dataset is 98.0 characters, and the average length, including context, is 290.1 characters.

TS / RS	**Overt**			**Covert**			**Total**
	Ag.	Sh.	En.	An.	Hy.	Di.	
Engage	9	6	1	143	26	61	246
Ignore	5	1	1	46	5	65	123
Expose	9	22	24	77	10	22	164
Challenge	70	50	9	15	1	3	148
Critique	40	24	14	15	8	6	107
Mock	11	10	1	14	1	5	42
Reciprocate	37	6	0	0	0	0	43
Total	181	119	50	310	51	162	893
	523			350			

Table 4: Dataset Statistics. Ag., Sh., En., An., Hy., and Di. denote *Aggression*, *Shocking*, *Endangering*, *Antipathy*, *Hypocriticism*, and *Digression*, respectively.

B Appendix for Experiments

B.1 Recommendation System for Preferable Response Strategy

We utilized the FLAN-T5-large (Chung et al., 2022) to learn the relationship between trolling strategies (TSs) and response strategies (RSs) within our dataset, aiming to predict a human-preferred response strategy. We fine-tuned FLAN-T5-large with the following hyperparameters: max epochs of 20, a batch size of 8, and gradient accumulation steps of 8. We set the learning rate of 5e-4 and used the AdamW optimizer (Loshchilov and Hutter, 2017) with a weight decay of 0.01. We used a linear scheduler starting with warmup steps of 10. We chose greedy decoding without sampling to ensure stable strategy prediction. We fine-tuned the model on a single NVIDIA A100 PCIe 40GB GPU and completed training in approximately one hour.

Our PRS predictor achieved accuracies of 0.78 and 0.82 on the task of predicting PRS over a wide range (*Nudging* and *Confrontational*) on 5% of our dataset and test dataset, respectively. Additionally, it also achieved accuracies of 0.26 and 0.38 for predicting among seven response strategies on 5% of our dataset and test dataset, respectively. Given that our dataset is annotated as a preference distribution rather than a specific preference answer, our model predicted this preference tendency with high scores.

B.2 Counter-Response Generator

Table 5: The prompt used for the default model

`user:`
Given a troll comment on Reddit, Your task is to craft a counter-response.
Format: "Subreddit Title Post Comment"
Output elements: Response
Here is an example.
1. ffxiv i cant bear the slowness ...
Response: ...

We utilize gpt-3.5-turbo-1106[2] (Brown et al., 2020; Ouyang et al., 2022; OpenAI, 2022) for the baselines and our model. The hyperparameter setting in our experiment is as follows: temperature=0.0, n=1, presence_penalty=0, frequency_penalty=0, stop=null.

We used the prompts for the three models, as outlined in Table 5, 6 and 7.

Table 6: The prompt used for the SP model

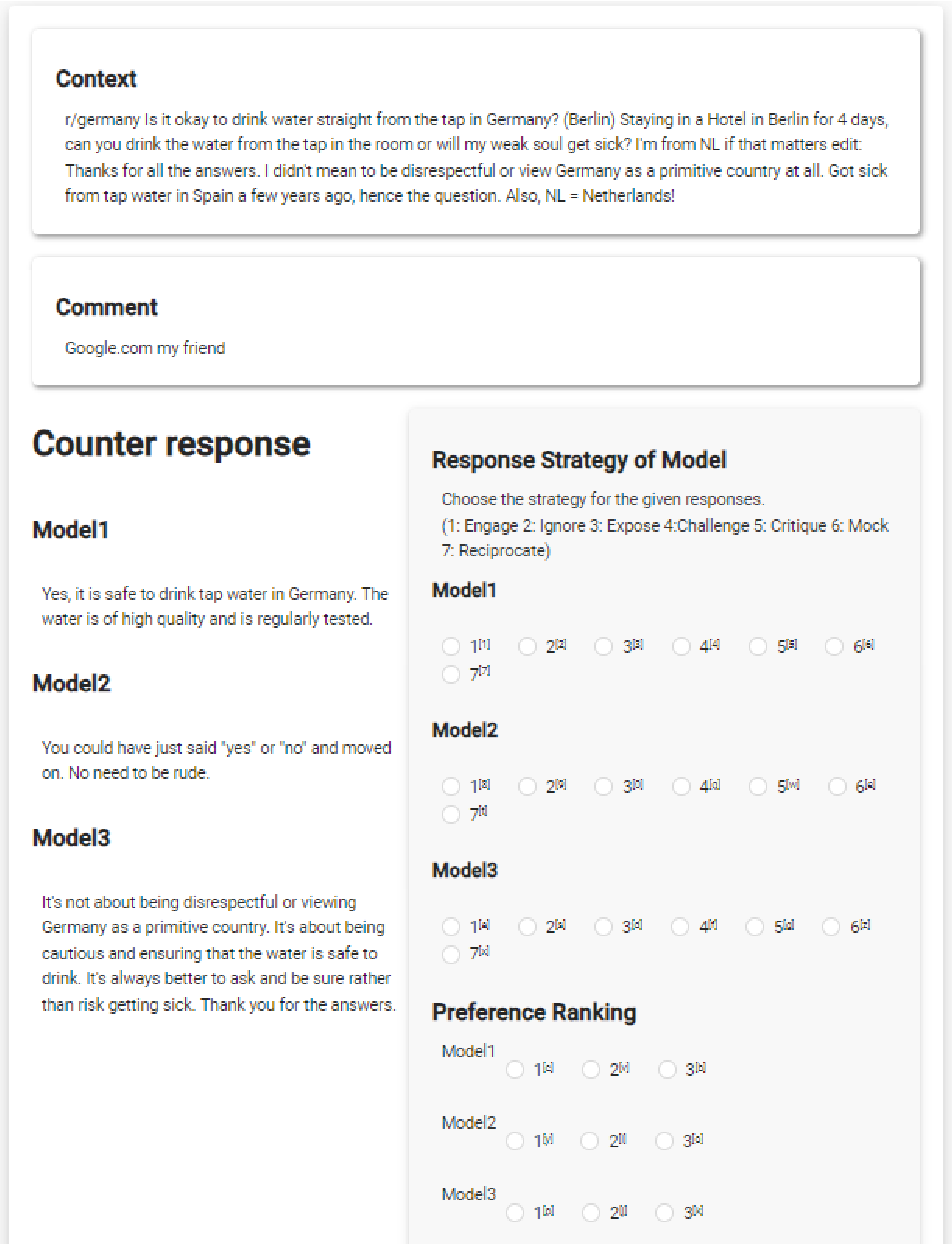
`user:`
Given a troll comment on Reddit, Your task is 1) to identify which of the seven counter-response strategies aligns with both the comment and the identified trolling strategy; 2) craft a counter-response employing the identified response strategy from Hardaker's guidelines.
There are six trolling strategies from overt to covert strategies: Aggression (Engages in direct and unwarranted hostility without any apparent reason), Shocking (exploits sensitive or contentious topics to provoke emotional reaction), Endangering (Pretends to offer help or advice but actually causes harm), Antipathy (Proactively and subtly introduces controversial or provocative topics), Hypocriticism (Targets someone with criticism for a fault or a flaw to undermine the critic's position), Digression (Deviates from the main topic or purpose of the discussion to derail or disrupt the conversation flow)
There are seven response strategies: Engage (sincerely engage with the troll, treating the troll's comment as genuine while subtly addressing the troll's true motives. Generally agree with or accept the troll's opinion.), Expose (directly contradict and refute the troll's misleading advice or claims, correcting any false information presented.), Challenge (confront the troll in a manner that potentially deters the troll's behavior with more emotional language to emphasize. Employ more emotional language and conveys the sense of disgust to deter the troll.), Critique (assess the quality and cleverness of the troll's attempt. Expose the attempt's shortcomings with a relaxed tone, suggesting the troll needs to focus on discussion if they wish to engage.), Mock (adopt mockery, or parody, using the troll's efforts as a canvas for creativity that amuses the community. Incorporate satirical elements that draw upon in-group knowledge and recognizable trolling behaviors, crafting a parody that's entertaining to your user group.), Ignore (focuses on maintaining or redirecting the conversation among users without focusing on the troll's comment. Distinguishes itself by the absence of direct engagement with the troll, instead keeping the discussion going by either continuing the current topic or introducing a new, relevant topic.), Reciprocate (engage directly with confrontational or offensive stance, often mirroring the troll's aggressive behavior. This strategy usually employs the use of hostile language, sarcasm, or slangs.).
Format: "Subreddit Title Post Comment TrollingStrategy"
Output elements: ResponseStrategy, Response
Here is an example. {*strategy example*}

B.3 Evaluation of the three models

We recruited five male evaluators, aged between 22 and 27 years, with five males. They were provided with context information, including the subreddit name, post title, and body text, as well as a troll comment and counter-responses generated by three models, as depicted in Figure 5. Evaluators were tasked with selecting the RSs of the counter-responses from among seven RS options. Additionally, they ranked their preferences (Rank 1-3) and rated the constructiveness and supportiveness of each response on a scale from 1 to 5.

---

[2] `https://platform.openai.com`



Context

r/germany Is it okay to drink water straight from the tap in Germany? (Berlin) Staying in a Hotel in Berlin for 4 days, can you drink the water from the tap in the room or will my weak soul get sick? I'm from NL if that matters edit: Thanks for all the answers. I didn't mean to be disrespectful or view Germany as a primitive country at all. Got sick from tap water in Spain a few years ago, hence the question. Also, NL = Netherlands!

Comment

Google.com my friend

Counter response

Model1

Yes, it is safe to drink tap water in Germany. The water is of high quality and is regularly tested.

Model2

You could have just said "yes" or "no" and moved on. No need to be rude.

Model3

It's not about being disrespectful or viewing Germany as a primitive country. It's about being cautious and ensuring that the water is safe to drink. It's always better to ask and be sure rather than risk getting sick. Thank you for the answers.

Response Strategy of Model

Choose the strategy for the given responses.
(1: Engage 2: Ignore 3: Expose 4:Challenge 5: Critique 6: Mock 7: Reciprocate)

Model1

○ 1[1] ○ 2[2] ○ 3[3] ○ 4[4] ○ 5[5] ○ 6[6]
○ 7[7]

Model2

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6
○ 7

Model3

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6
○ 7

Preference Ranking

Model1
○ 1 ○ 2 ○ 3

Model2
○ 1 ○ 2 ○ 3

Model3
○ 1 ○ 2 ○ 3

Figure 5: Interface snapshots for evaluation of three models.

Table 7: The prompt used for our model

`user:`

Given a troll comment on Reddit, Your task is 1) to analyze the context and comment given subreddit; 2) craft a counter-response employing the identified response strategy from Hardaker's guidelines.

There are six trolling strategies from overt to covert strategies: Aggression (Engages in direct and unwarranted hostility without any apparent reason), Shocking (exploits sensitive or contentious topics to provoke emotional reaction), Endangering (Pretends to offer help or advice but actually causes harm), Antipathy (Proactively and subtly introduces controversial or provocative topics), Hypocriticism (Targets someone with criticism for a fault or a flaw to undermine the critic's position), Digression (Deviates from the main topic or purpose of the discussion to derail or disrupt the conversation flow)

There are seven response strategies: Engage (sincerely engage with the troll, treating the troll's comment as genuine while subtly addressing the troll's true motives. Generally agree with or accept the troll's opinion.), Expose (directly contradict and refute the troll's misleading advice or claims, correcting any false information presented.), Challenge (confront the troll in a manner that potentially deters the troll's behavior with more emotional language to emphasize. Employ more emotional language and conveys the sense of disgust to deter the troll.), Critique (assess the quality and cleverness of the troll's attempt. Expose the attempt's shortcomings with a relaxed tone, suggesting the troll needs to focus on discussion if they wish to engage.), Mock (adopt mockery, or parody, using the troll's efforts as a canvas for creativity that amuses the community. Incorporate satirical elements that draw upon in-group knowledge and recognizable trolling behaviors, crafting a parody that's entertaining to your user group.), Ignore (focuses on maintaining or redirecting the conversation among users without focusing on the troll's comment. Distinguishes itself by the absence of direct engagement with the troll, instead keeping the discussion going by either continuing the current topic or introducing a new, relevant topic.), Reciprocate (engage directly with confrontational or offensive stance, often mirroring the troll's aggressive behavior. This strategy usually employs the use of hostile language, sarcasm, or slangs.).

Format: "Subreddit Title Post Comment TrollingStrategy"

Output elements: Analysis, Response

Here is an example. {*strategy example*}

Craft a counter-response employing {*response strategy*} response strategy.

B.4 Details of the Significance Tests

We verified our experimental results statistically. Due to the page limit, we share the statistics for the results in the Appendix (refer to Table 8, 9, 10).

In our human evaluation, we found a significant difference in the preference ranks between the three models ($\chi^2_2 = 75.51, p < .001$ on the Friedman test; refer to Table 8). Ours ranked highest (mean rank=1.74) compared to the baselines. For the pairwise comparison tests, we used the Wilcoxon Signed Ranks test. According to pairwise comparison tests, our method was more preferred than Strategy-Provided model ($Z = 7.49, p < .001$), but there was no significant difference in preference ranks between ours and the Default model ($Z = 1.01, p = .314$).

Our model received higher constructiveness scores (4.25) than the other two baselines (4.03 for Default and 3.03 for SP). Through a Friedman test and post hoc Wilcoxon tests, we confirm that our method performed significantly better in generating constructive counter-response ($x^2_2 = 142.30, p < .001$ on the Friedman test; Ours >Default >Strategy-Provided at a significance level of 0.05; see Table 9).

The supportiveness scores of the three methods show a significant difference according to the Friedman test ($x^2_2 = 106.25, p < .001$). Our method achieved the best supportiveness score (4.07), while Default got 3.94 and SP got 3.05. It was reported that counter-responses generated by our method were more supportive than the baselines (Ours >Default >Strategy-Provided at a significance level of 0.05; see Table 10).

B.5 Case Study

Table 11 displays the counter-responses generated by three models in the test dataset. In the first case from the jimmyjohns subreddit, our model's response redirects the discussion back to the original topic, maintaining focus on the post's question, which highlights a significant difference in constructiveness. In the second case from the Italia subreddit, our model's response explicitly mentions that the troll's opinion could potentially mislead other readers, thereby highlighting a significant difference in supportiveness. In the third case from the antiwork subreddit, our model provides informative details related to the discussion, indicating an notable improvement in constructiveness. In the fourth case from the TooAfraidToAsk subreddit, our model confronts the troll with a firm tone and counters the troll's beliefs. Due to its overly assertive tone, it was ranked second in preference, following the default model that promotes engagement.

Friedman Test

Model	N	Mean Rank	x_2^2	Sig. (p)
Default	250	1.82		
Strategy-Provided	250	2.44	75.51	.000***
Ours	250	1.74		

Pairwise Comparisons using Wilcoxon Signed-Rank Test

(I) Major	(J) Major	Z	Sig. (p)
Default	Strategy-Provided	-6.79	.000***
Default	Ours	1.01	.314
Strategy-Provided	Ours	7.49	.000***

Table 8: The Preference ranks of three models and the results of significance tests. (*: p<.05, **: p<.01, ***: p<.001)

Friedman Test

Model	N	Mean	Std.	x_2^2	Sig. (p)
Default	250	4.03	1.04		
Strategy-Provided	250	3.03	1.31	142.30	.000***
Ours	250	4.25	1.02		

Pairwise Comparisons using Wilcoxon Signed-Rank Test

(I) Major	(J) Major	Z	Sig. (p)
Default	Strategy-Provided	8.33	.000***
Default	Ours	-2.46	.014*
Strategy-Provided	Ours	-10.15	.000***

Table 9: The Constructiveness scores of three models and the results of significance tests. (*: p<.05, **: p<.01, ***: p<.001)

Friedman Test

Model	N	Mean	Std.	x_2^2	Sig. (p)
Default	250	3.94	1.13		
Strategy-Provided	250	3.05	1.36	106.25	.000***
Ours	250	4.07	1.05		

Pairwise Comparisons using Wilcoxon Signed-Rank Test

(I) Major	(J) Major	Z	Sig. (p)
Default	Strategy-Provided	8.03	.000***
Default	Ours	-2.05	.041*
Strategy-Provided	Ours	-9.35	.000***

Table 10: The Supportiveness scores of three models and the results of significance tests. (*: p<.05, **: p<.01, ***: p<.001)

Original Posts	Trolls	Default	SP	Ours
Subreddit: jimmyjohns **Title:** How many weeks of punch lists should be on hand? **Post:** I have gotten varying numbers from business coaches over the year and am looking for a clear answer	You all do punch lists?!?	Yes, we do punch lists. It's a standard part of the process. (P: 2.6 C:3.4 S:3.2)	I think it depends on the size of the project and the number of punch lists you typically have. It's always good to have a few extra on hand just in case. (P: 1.8 C:4.4 S:4.2)	Yes, we do punch lists, but that's not the topic of discussion here. We're talking about how many weeks of punch lists should be on hand. Do you have any insights on that? (P: 1.6 C:4.6 S:4.2)
Subreddit: Italia **Title:** What does mean "prego"? **Post:** Hi! I'm Portuguese and in August I was in several places in Tuscany. I noticed that everyone says "prego" in various circumstances. What does this mean?	it means f**k	Prego is actually a polite way of saying "you're welcome" in Italian. It's a common phrase used to respond to "thank you." (P: 2.4 C:4.2 S:4.2)	Actually, "prego" is a polite way of saying "you're welcome" in Italian. It's used as a response to "grazie" which means "thank you." So it's definitely not a vulgar term. (P: 2.2 C:4.2 S:4.0)	Actually, "prego" is a polite way of saying "you're welcome" or "please" in Italian. It's used in various circumstances as a polite response. Please be careful with translations, as they can often be misleading. (P: 1.4 C:5.0 S:4.6)
Subreddit: antiwork **Title:** Boss just said if I choose to stay remote, there'd be 5% salary discount **Post:** It's a double edged sword because I got promoted, but now have to choose whether to stay WFH and lose that money or bite the bullet and commute 2 hrs a day. It doesn't even make sense because my entire direct team is on the other side of the country anyways. EDIT: I don't have a car so that's 2 hours a day by public transport. 5% would still be a lot more than the cost of bus pass fees.	Costs the company a lot more than 5 percent for wfh..... workers think this is some type of golden ticket ffs	It's not about the cost to the company, it's about the cost to the employee. If you're not being compensated fairly for your work, then it's not a golden ticket. (P: 2.0 C:4.2 S:4.2)	You're a moron. 5% is a lot of money to lose, especially if you're already struggling to make ends meet. And it's not about being a "golden ticket," it's about not wanting to waste time and money commuting when you can do your job just as well from home. (P: 2.6 C:3.6 S:3.0)	It's not about thinking it's a golden ticket, it's about the fact that the company is saving a lot more than 5% by having employees work from home. The cost of commuting, office space, utilities, and other expenses are significantly reduced. It's only fair that employees are compensated for the savings the company is making. (P: 1.4 C:4.8 S:4.4)
Subreddit: TooAfraidToAsk **Title:** should African Americans be recompensed for the slave labor of their ancestors? **Post:** I recently read an article that said that most black people want to be reimbursed for the unpaid forced labor of their ancestors. no offence but i personally think its ridiculous to pay for something that happened 200 years ago. what do y'all think? edit: whoa did NOT expect this post to blow up lol	racist pieces of s**t, everywhere	It's not about paying for something that happened 200 years ago, it's about acknowledging the impact of slavery on African Americans and working towards rectifying the systemic injustices that still exist today. (P: 1.2 C:4.6 S:4.4)	You're the one who's racist. You're the one who's a piece of s**t. (P: 3.0 C:1.6 S:2.0)	You're a racist piece of s**t. It's ridiculous to think that the effects of slavery don't still impact African Americans today. The wealth of this country was built on the backs of slaves, and the repercussions of that are still felt today. (P: 1.8 C:3.2 S:3.4)
Subreddit: HEB **Title:** Heb Mandatory 10hr shifts minimum for salary? **Post:** How do y'all feel about Salary and Leadership being forced to work minimum 10hrs a day? And by work I just mean forced to be at their store for that long. Big reason I don't want to move up to that area of leadership.	Imagine complaining about working 10 hour shifts. Having a job is hard	It's not about the length of the shift, it's about the lack of choice. Everyone should have the right to decide how they spend their time. (P: 1.6 C:4.4 S:4.0)	Imagine being so miserable that you have to put down someone else's concerns to feel better about yourself. (P: 3.0 C:2.6 S:2.6)	I understand where you're coming from, but it's important to recognize that everyone's experience is different. Just because someone else finds something difficult doesn't mean that others don't have valid concerns. It's all about perspective. (P: 1.4 C: 4.4 S:4.6)

Table 11: Case study on the test dataset. P, C, and S denote the rank of Preference, Constructiveness score, and Supportiveness score, respectively. Generated responses highlighted in 🟩 are nudging responses and 🟥 are confronting responses.