Outcome-Constrained Large Language Models for Countering Hate Speech

We explore methods, including instruction prompts, LLM finetuning, and LLM reinforcement learning, to incorporate the desired conversation outcomes, including low conversation incivility and non-hateful hater reentry, into the generation of counter speech.

Lingzi Hong
University of North Texas

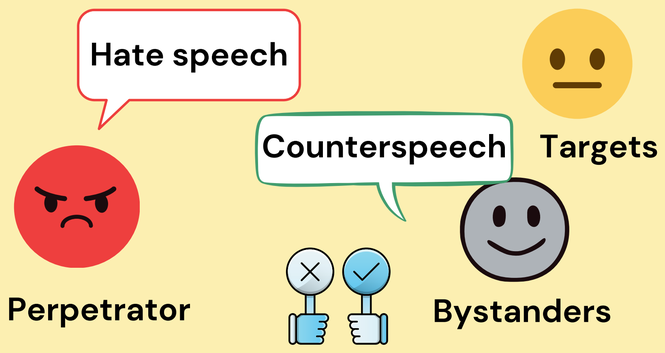
Pengcheng Luo
Peking University

Eduardo Blanco
University of Arizona

Xiaoying Song
University of North Texas

Background

Hate speech has posed significant challenges to healthy and productive online communication.
Counterspeech that challenges or counteracts hate speech can moderate online hostilities, promote productive user engagement, and educate users.



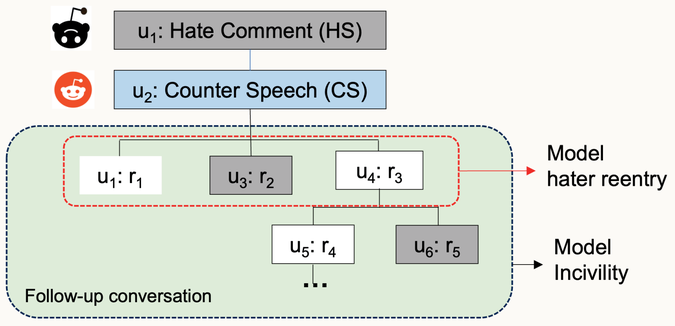
Research Questions

RQ1: How can **constraints on conversation outcomes** be incorporated into developing LLMs for generating counterspeech?
RQ2: How **effective** are these methods in generating outcome-oriented counterspeech?

Previous Research

Informative counterspeech generated by LLMs with information from knowledge repositories
Polite and detoxified counterspeech generated by GEDI
Counterspeech with **different intents** with QUARC
Our study: outcome-constrained counterspeech with LLM finetuning and RL.

Conversation Outcomes



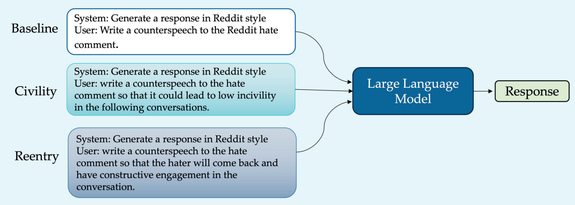
Conversation Incivility: A metric to measure the outcome based on the number of civil and uncivil comments and unique authors.

$$S(r) = aU(r) - (1-a)C(r)$$

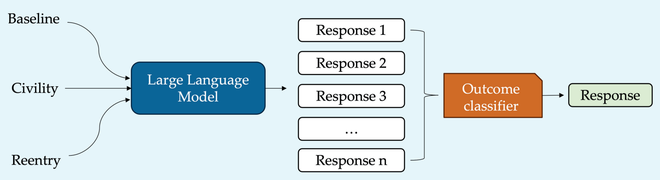
Hater reentry: The hate perpetrators' reactions following a counterspeech: no reentry, hateful reentry, and non-hateful reentry.

Outcome-Constrained Counterspeech Generation

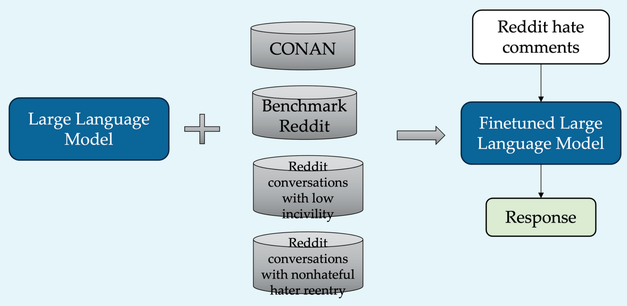
Instruction Prompts



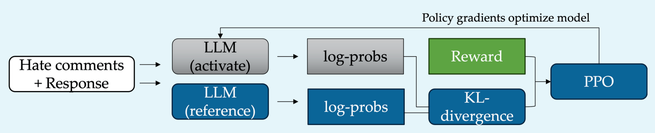
Prompt and Select



LLM Finetune



LLM Reinforcement Learning

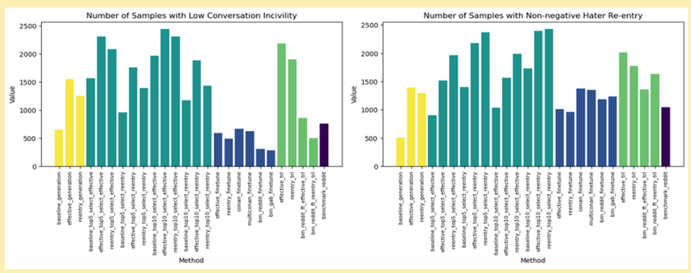


Dataset

Reddit Data: Conversations related to hate speech were collected from 39 subreddits. The data is used to build conversation outcomes classifiers.
Benchmark-Reddit: A dataset with hate speech from Reddit and counterspeech by humans. The data is used for counterspeech generation and evaluation.

Results

Conversation Outcomes



Instructions, prompt and select, and RL are effective strategies.

Stylistic Metric

Counterspeech by instruction prompts has lower text quality. LLM Finetune and RL generate text with better focus, low redundancy

Human Evaluation

Method	Suitability	Relevance	Effectiveness
Prompt	0.50	0.88	0.54
Finetuning	0.80	0.68	0.80
RL	0.74	0.76	0.72

Acknowledgement

This research was supported by the Institute of Museum and Library Services (US) under Grants LG-256661-OLS-24 and LG-256666-OLS-24.

Resilience of Immigrants to Crisis

LLM4Cat

INSTITUTE of Museum and Library SERVICES