



Counter-speech generation for homophobic and transphobic social media content in Malayalam

Title	Counter-speech generation for homophobic and transphobic social media content in Malayalam
Author(s)	Prasannan, Praveen;Kumaresan, Prasanna Kumar;Rajiakodi, Saranya;Subalalitha, C. N.;Chakravarthi, Bharathi Raja
Publication Date	2025-08-12
Publisher	Springer
Repository DOI	https://doi.org/10.1007/s13278-025-01507-x



Counter-speech generation for homophobic and transphobic social media content in Malayalam

Praveen Prasannan¹ · Prasanna Kumar Kumaresan¹ · Saranya Rajiakodi² · C. N. Subalalitha³ · Bharathi Raja Chakravarthi¹

Received: 30 May 2025 / Revised: 18 July 2025 / Accepted: 19 July 2025
© The Author(s) 2025

Abstract

The growing prevalence of hate speech online has amplified acts of discrimination against marginalized populations, with the LGBTQIA+ community being particularly affected. In areas where under-resourced languages such as Malayalam are used, the issue grows more complex because of the absence of localized resources. This research offers an in-depth analysis of the production of counter-speech to address transphobia and homophobia in Malayalam. Our work incorporates both native Malayalam script and Malayalam written in Latin script, addressing the diverse linguistic practices of online users in Kerala. This paper introduces a two-stage pipeline to counter such online abuse. The first stage focuses on dataset creation through a human-in-the-loop process, beginning with 100 seed pairs of hate speech and their corresponding counter-speech manually curated. This set is expanded iteratively using language models culminating in 5,000 validated pairs. In the second stage, we propose a method to generate counter speech in Malayalam that leverages the Retrieval-Augmented Generation framework enhanced by REFINE (Retrieval Evaluation via Fluency, Inversion, and NEarness) for knowledge retrieval and constrained decoding. Evaluation metrics for both dataset quality and model performance demonstrate the effectiveness of this approach in producing diverse, fluent, and target-specific counter-speech. This research provides a foundational resource and scalable strategy for countering hate in low-resource regional languages. **GitHub Link:** <https://github.com/Bharathi-AI-for-Social-Good/CN-Malayalam>.

Keywords Homophobia and transphobia · Counter speech · Low-resource language · Human-in-the-loop · Generative AI · Retrieval augmented generation

1 Introduction

The rise of online hate speech has become a significant worldwide concern, especially due to the rapid increase of digital communication platforms (Williams 2019; Zhou et al. 2024). These platforms, although promoting connectivity and community, have simultaneously created an environment conducive to the spread of hatred, discrimination, and abuse. Hate speech poses a significant global threat to digital safety and social cohesion, targeting a wide spectrum of marginalized communities (Guiora and Park 2017; Chetty and Alathur 2018; Castaño-Pulgarín et al. 2021). These include racial and ethnic minorities, religious groups, women, migrants, caste-oppressed populations, and individuals with disabilities (Arcila Calderón et al. 2024). One of the most affected groups includes those identified as lesbian, gay, bisexual, transgender, queer, intersex, asexual, and others (LGBTQIA+), who frequently encounter transphobic

✉ Bharathi Raja Chakravarthi
bharathi.raja@universityofgalway.ie

Praveen Prasannan
p.prasannan1@universityofgalway.ie

Prasanna Kumar Kumaresan
P.Kumaresan1@universityofgalway.ie

Saranya Rajiakodi
saranya.rajiakodi@cutn.ac.in

C. N. Subalalitha
subalalitha.navaneethakrishnan@srmist.edu.in

¹ School of Computer Science, University of Galway, Galway, Ireland

² Central University of Tamil Nadu, Thiruvarur, India

³ SRM Institute of Science and Technology, Chennai, India

and homophobic material (Chakravarthi et al. 2022). This specific mistreatment not only harms the dignity and safety of these people but also fosters a larger culture of exclusion and aggression.

Hate speech encompasses any form of communication—verbal, written, or behavioral that demeans or provokes violence, discrimination, or bias against people or groups due to characteristics such as race, religion, ethnicity, sexual orientation, gender identity, disability, or nationality (Tareen et al. 2021). In digital environments, hate speech frequently appears as insulting remarks, threats, false information, and dehumanizing language. The World Wide Web magnifies such harmful material, quickly disseminating it across platforms and sustaining its adverse effects (Paz et al. 2020; Rivers 2015).

Transphobia is the unfounded fear, dislike, or prejudice towards transgender people (Bettcher 2014). It can manifest as verbal insults, mislabeling, social exclusion, or acts of violence. Homophobia signifies adverse feelings, bias, or animosity towards people who are attracted to those of the same gender. Transphobia and homophobia are deeply embedded in cultural, religious, and societal norms, frequently resulting in systemic marginalization and violence (Rollè et al. 2024). Regrettably, online platforms have turned into primary spaces where these types of hate are conveyed and disseminated (Chakravarthi 2024; Kumaresan et al. 2023).

Counter speech is a non-violent, positive reaction to hate speech (Cepollaro et al. 2023). It includes responding to hateful messages with information, understanding, backing for affected groups, or appeals for acceptance and respect for one another. In contrast to censorship or deletion, counter speech maintains the conversation while seeking to diminish harmful rhetoric and lessen its impact (Garland et al. 2020, 2022). Impactful counter speech can address false information, present different perspectives, or make the individuals or groups at risk more relatable. This method has been strongly supported by scholars and human rights groups as a means to encourage better online communication (Yadav 2018).

In India, hate speech in local languages presents distinct challenges because of the nation's linguistic variety and socio-cultural intricacies.¹ Malayalam (ISO 639-3: mal) is a language from the Dravidian family, mainly spoken in the Indian state of Kerala and in the union territories of Lakshadweep and Mahe (a district in Puducherry) (Thottingal 2019). Recognized as one of India's 22 scheduled languages, Malayalam also serves as the official language of Kerala (Nandi et al. 2024). One of the most remarkable achievements of Malayalam is its recognition as a Classical

Language of India in 2013. This honor was granted by the Government of India in acknowledgment of its long-standing literary tradition, historical depth, and original ancient texts. The classical status affirms the antiquity and cultural significance of Malayalam. The Malayalam script traces its origins to the ancient Pallava Grantha script, which was historically used in southern India to write Sanskrit. With over 35 million speakers worldwide, Malayalam continues to thrive in education, media, cinema, and digital platforms, serving as a strong cultural identity and a vibrant mode of communication for millions but is one of the under-resourced languages in the digital space (Chakravarthi et al. 2020).

Hate speech in Kerala often appears in both standard Malayalam script and in Malayalam expressed using Latin (Roman) characters (Ghanghor et al. 2021). This dual-language format presents a unique challenge for natural language processing systems and content moderation tools. Although Malayalam possesses a rich literary and cultural tradition, it remains significantly underrepresented in natural language processing (NLP) research. This scarcity is further complicated when the language is rendered in Latin (Roman) script, where informal usage, orthographic inconsistency, and code-switching introduce additional linguistic challenges (Baiju et al. 2025). Consequently, the systems for recognizing, examining, and combating online hate in Malayalam are considerably less advanced than those accessible for global languages such as English (Singh and Thakur 2024).

Additionally, Kerala, in spite of its advanced literacy rates and social metrics, still struggles with deeply rooted biases against gender and sexual minorities. Such biases tend to be heightened in online conversations, as anonymity and the absence of accountability promote the voicing of hate without concern for social or legal repercussions (Sharma et al. 2023). In this context, the spread of transphobic and homophobic rhetoric poses a significant issue, endangering the mental health and physical safety of marginalized groups.

Traditional responses to hate speech have largely relied on reactive measures such as flagging, content removal, and user bans. While these methods are necessary for platform regulation, they often draw criticism for infringing on freedom of speech and failing to address the root causes of hate (Chakravarthi 2020). Moreover, the manual moderation of content is labor-intensive and not feasible at scale. This has led to an increased interest in proactive strategies like counter speech, which involves responding to hateful messages with factual, empathetic, and constructive narratives. Counter speech aims not only to refute the harmful message but also to educate and shift public discourse towards tolerance and inclusivity (Sahoo et al. 2024a).

The concept of counter speech aligns well with democratic ideals and is gaining traction globally as a sustainable and rights-respecting solution to online hate. However,

¹ <https://lawcommissionofindia.nic.in/report267/>

crafting effective counter speech requires linguistic sensitivity, cultural awareness, and strategic communication skills qualities that are hard to embed in automated systems without appropriate training data and feedback mechanisms (Bonaldi et al. 2024).

Recent advances in artificial intelligence, particularly large language models (LLMs) such as GPT, have opened new avenues for generating counter-narratives at scale (Park et al. 2025). These models can learn to produce human-like text given sufficient examples. Yet, their application in low-resource languages like Malayalam remains limited due to the scarcity of annotated datasets and domain-specific corpora. Malayalam is considered an under-resourced language in NLP because it lacks large annotated datasets, pre-trained language models, and foundational tools like tokenizers and parsers (Hande et al. 2022). Its complex script and the widespread use of informal Latin-script writing pose additional challenges for data standardization. Compared to global or even other Indian languages like Hindi or Tamil, Malayalam has a smaller digital footprint, less academic and commercial investment, and limited availability of structured online content. These factors, combined with dialectal variation and lower prioritization in tech development, have contributed to its under-resourced status (Baiju et al. 2025). Furthermore, LLMs are prone to generating generic or repetitive content, and without proper oversight, they risk reinforcing stereotypes or introducing inaccuracies (Nguyen et al. 2025).

The motivation for creating a counter-speech dataset and generation framework focused on Malayalam arises from both social urgency and linguistic complexity. Online hate targeting LGBTQIA+ individuals is particularly

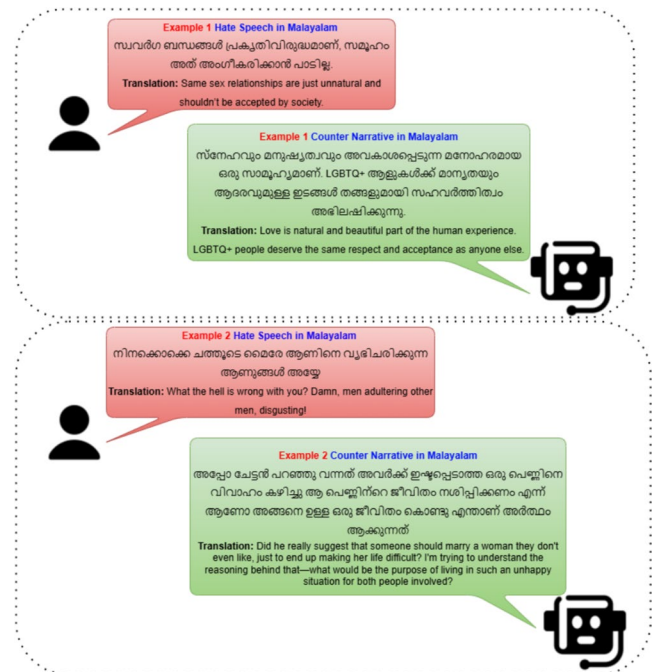
visible and harmful in Malayalam-speaking digital spaces, where moderation tools and affirmative content remain scarce (Chakravarthi 2020). Figure.1 shows the sample chat between user (Homophobic/Transphobic) and bot (counter-speech).

The main contributions of this paper are as follows:

- We developed a manually annotated seed dataset of counter speech responses in Malayalam specifically targeting transphobic and homophobic hate speech, one of the first of its kind for this under-resourced language.
- We introduced an iterative human-in-the-loop framework for data expansion and validation, demonstrating its superiority in producing contextually relevant, diverse, and high-quality counter narratives.
- We integrated the Retrieval-Augmented Generation (RAG) framework with REFINE (Retrieval Evaluation via Fluency, Inversion, and NEarness) -based retrieval and constrained decoding to produce counter speech for unseen hate speech inputs.
- We applied distinct evaluation metrics for both dataset construction and chatbot performance. These include Acceptance Rate, Human-targeted Translation Edit Rate (HTER), Repetition Rate, and Vocabulary Growth for dataset evaluation, and Relevance, Novelty, Stance, Fluency and Toxicity for assessing RAG-generated counter speech.

We have structured this paper as follows: Sect. 2 reviews the existing literature, focusing on previous methodologies and their limitations. Section 3 explores the benchmark

Fig. 1 Examples of homophobic & transphobic hate speech and associated counter-narratives (calm and reasoned) by a bot in Malayalam



methodologies, explaining the baseline setting, model fine-tuning, and comparative results. It includes the one-shot generation method, as well as the human-in-the-loop iterative framework using ChatGPT-4o and Gemini 2.0. Additionally, this section provides a detailed explanation of the RAG framework, highlighting the REFINE (Retrieval Evaluation via Fluency, Inversion, and NEariness) -based retrieval strategy and the decoding constraints employed for generating high-quality counter narratives. Section 4 and 5 describes the procedural setup and evaluation, detailing the objectives and methodologies used during iterative generation and result discussion. The paper concludes in Sect. 6, including the future outlook and possible extensions of this work. Section 7 includes the possible extensions of this work.

2 Related work

The proliferation of online hate speech has prompted extensive research into counter speech generation, particularly leveraging advancements in Natural Language Processing (NLP) and machine learning. Initial methods for managing hate speech online often depended on people manually reviewing and responding to harmful content. However, given the massive volume of posts generated across digital platforms each day, this approach quickly became unmanageable. The limited speed and scalability of human oversight make it an impractical long-term solution, driving the need for automated systems that can respond more efficiently and consistently (Tekiroğlu et al. 2020).

Bonaldi et al. (2022) introduced DIALOCONAN, a dataset comprising multi-turn dialogues between a hater and an NGO operator. Their human-machine collaboration approach combined machine-generated dialogues with expert annotations, capturing the dynamic nature of real-life interactions and providing a valuable resource for training dialogue-based counter speech systems. (Chung et al. 2019) presented the CONAN dataset, a multilingual collection of hate speech and counter narrative pairs, focusing on Islamophobia in English, French and Italian. Their work laid the foundation for subsequent studies in multilingual counter speech generation.

Fanton et al. (2021) introduced a human-in-the-loop data collection methodology, resulting in the Multi-Target Counter Narrative (CONAN) dataset for English language. This approach involved iterative refinement of a generative language model, with expert reviewers post-editing the outputs to ensure quality and diversity. The CONAN dataset encompasses counter narratives targeting various forms of hate speech, including those directed at the LGBTQIA+ community. Building upon this, Sahoo et al. (2024b) developed IndicCONAN, a multilingual dataset aimed at combating hate speech in Indian contexts, specifically in

Hindi and Indian English. Their methodology combined autoregressive language models with human-in-the-loop strategies to generate culturally relevant counter narratives. However, several Indian languages remain under-represented in such initiatives, especially low-resource languages like Malayalam. In response to this gap, our work focuses on the development of counter narratives specifically tailored for the Malayalam language.

Wilk et al. (2025) proposed a fact-based counter narrative generation framework that integrates web search and document retrieval to enhance the factual accuracy and informativeness of generated responses. Their approach demonstrated superior performance in terms of factual consistency and persuasive quality compared to existing baselines. Vallecillo Rodríguez et al. (2024) focused on the Spanish language, creating the CONAN-MT-SP corpus by translating the English CONAN dataset using the DeepL API and generating counter narratives with GPT-based models. Their evaluation highlighted the potential of large language models in producing effective counter speech in under-resourced languages.

Ashida and Komachi (2022) explored the automatic generation of messages countering online hate speech and microaggressions using prompting techniques with pretrained language models like GPT-2 and GPT-3. Their study emphasized the efficacy of prompting over fine-tuning, especially when data and computational resources are limited. Human evaluations assessed the generated messages based on offensiveness, stance, and informativeness, indicating the viability of this approach.

Jiang et al. (2025) proposed the Retrieval-Augmented Zero-shot Generation (ReZG) framework for counter speech generation. Unlike traditional generative approaches, ReZG retrieves relevant counter-knowledge from a curated repository using a Stance-Semantic-Fitness (SSF) filtering strategy. This ensures that retrieved content not only aligns semantically with the hate speech but also provides an opposing and contextually appropriate stance. ReZG applies constrained decoding to enforce knowledge consistency, counter stance, and fluency in the generated response. This makes ReZG especially effective for producing counter narratives in domains or languages where training data is sparse or unavailable.

The previous studies collectively underscore the importance of culturally and linguistically tailored counter speech generation methods. However, there remains a notable gap in resources and methodologies for under-resourced languages like Malayalam. Our work aims to address this by developing a Malayalam dataset and employing human-in-the-loop techniques and retrieval-augmented generation enhanced by REFINE to generate effective counter-narratives against homophobic and transphobic speech.

3 Methodology

The proposed methodology for counter speech generation follows a structured pipeline combining manual data curation, iterative model generation, human validation, and retrieval-based augmentation. This multi-phase strategy was designed to capture real-world linguistic features, script variation, and socio-cultural sensitivity essential for effective counter speech in Malayalam and Malayalam rendered in Latin script.

3.1 Seed dataset preparation

The dataset construction for this study was designed to reflect the real-world linguistic diversity of Malayalam users online. We built a seed dataset of 100 hate-counter speech pairs. These were selected from a larger dataset, originally compiled from various social media platforms shared by Kumaresan et al. (2023). Figure 2 shows the sample from the dataset where the first two columns contain the text and corresponding category respectively. The text field included hate speeches annotated as homophobic, transphobic, hope speech and counter speech.

To create the seed dataset, 100 hate speech samples focused on homophobia and transphobia were selected. Each hate speech example was manually paired with a counter speech from the dataset by two voluntary Malayalam speakers (annotators) familiar with LGBTQIA+ advocacy and digital discourse patterns. Annotators were instructed to ensure that counter-speech was polite, nonaggressive,

Malayalam Comments	Category	English Translation
നശിച്ച ജന്തുക്കൾ നാണമില്ലേ ഇവർക്ക് പടച്ചോൻ നമ്മെ ഏവരെയും കാത്ത രക്ഷിക്കട്ടെ അല്ലാഹുവേ ഇതിന്റെ പേരിൽ ഞങ്ങളെ ശിക്ഷിക്കല്ലേ നാഥാ അവർക്ക് മനസ്സിലാക്കാനുള്ള ബോധം കൊടുക്ക് റബ്ബേ	Homophobic -derogation	Don't these ruined creatures have any shame? O Creator, please protect and save all of us. O Allah, do not punish us because of them. Lord, grant them the awareness to understand.
ഓരോ സ്വവർഗ്ഗനുരാഗിയുടെയും ജീവിതം അവസാനം ഇത്തരത്തിൽ ആണെന്ന് നിങ്ങൾ പിന്നെയും പിന്നെയും പറഞ്ഞു വെക്കുകയാണല്ലോ ഇത് മാറ്റാനായില്ലേ	Counter-Speech	You keep saying again and again that the life of every homosexual ends like this — hasn't this changed yet?
Mumbail poyal ivarude thani niram kaanamPaisa chodhichal koduthillel adikkukayum theri parayukayum cheyyum athukokanthe thanneya veruppu (Malayalam rendered in Latin Script)	Transphobic-derogation	When you go to Mumbai, you see their true colors. If you don't give them money when asked, they'll hit you and hurl abuse — that's exactly why there's so much hatred toward them
Kundan Spotted (Malayalam rendered in Latin Script)	Homophobic -derogation	Gay is spotted

Fig. 2 Sample data from Kumaresan et al. (2023) showing homophobic, transphobic and counter speeches in Malayalam and Malayalam rendered in Latin script. English translations are provided for each comment for readers

informed, and directly responded to the hate content. Figure 3 represents the workflow of seed dataset creation. These responses included various approaches, such as correcting misinformation, promoting empathy, and affirming the dignity of LGBTQIA+ individuals. In addition, a second round of review was conducted by two additional volunteers of Malayalam origin to ensure the quality of the dataset. Each annotated pair was evaluated for linguistic correctness, relevance, emotional tone, and cultural sensitivity. Any pair that failed to meet these standards was either revised or removed. The whole process of creating the seed dataset took a week to complete.

The Homophobic/Transphobic speech samples were also categorized by type (e.g., derogation, threats) and context (e.g., targeting transgender people, mocking same-sex relationships) to help in selecting the most effective counter strategy. The diversity of sources also ensured that the dataset included both formal hate messages (e.g., longer rants on forums) and informal or sarcastic comments. Malayalam written in Latin script inputs were especially diverse in spelling and syntax, which posed challenges during pre-processing and annotation, but reflected the authenticity of real-world usage. This dataset of 100 manually matched pairs served as the starting point for our human-in-the-loop (HITL) expansion process, which later generated more than

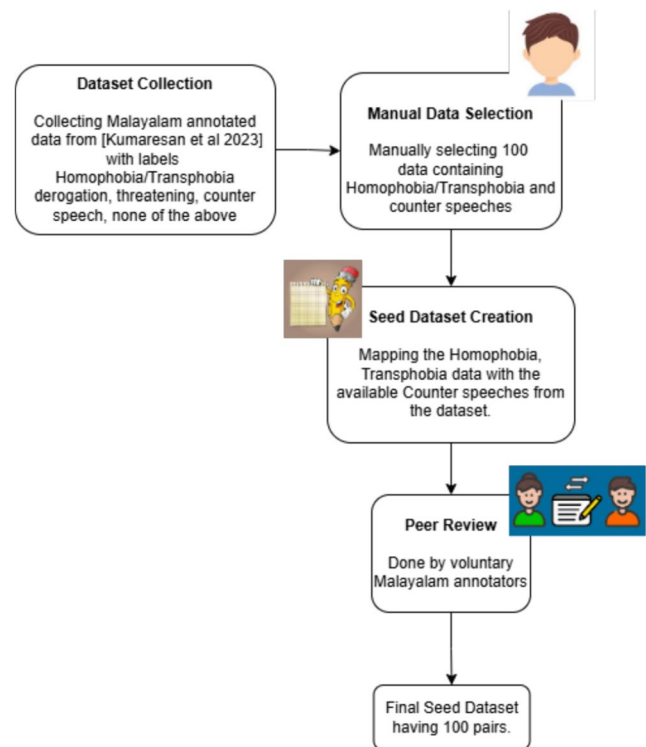


Fig. 3 Comprehensive workflow illustrating the creation of the seed dataset for Homophobia, Transphobia, and counter speech pairs in Malayalam

5,000 examples using ChatGPT and Gemini. The dataset included a mix of short, informal comments, sarcastic remarks, and longer texts. Malayalam in Latin script entries showed a wide range of spelling and structure, reflecting how people communicate online and presenting realistic challenges for model training. This 100-pair seed dataset is version 1 (V1).

Table 1 outlines key statistics from the finalized seed dataset, which underwent validation by a group of voluntary annotators. The dataset consists of 100 entries categorized under homophobic or transphobic content. In parallel, the dataset also includes 100 counter-speech responses aimed at addressing the hate speech instances.

3.2 Annotation guidelines

- *Empathy First*: Responses must prioritize compassion and avoid any retaliatory or inflammatory tone. Annotators were instructed to “respond with care” rather than confront.
- *Direct Relevance*: Counter speech should directly address the hateful content, ideally correcting misinformation or challenging stereotypes presented in the original message.
- *Script Consistency*: Annotators were advised to maintain the same script (Malayalam or Latin) as the original hate speech comment for contextual coherence.
- *Non-aggression Principle*: All responses were required to avoid name-calling, sarcasm, or any aggressive language that might escalate the interaction.
- *Tone and Cultural Sensitivity*: Annotators were guided to maintain a tone that would be socially and culturally acceptable in the Malayalam-speaking community, avoiding religious or political triggers.
- *Clarity and Readability*: Messages were required to be easily understandable, avoiding excessive jargon or overly academic phrasing.
- *Inclusiveness in Language*: Annotators were provided with lists of inclusive and affirming terms to replace derogatory language without mirroring or reinforcing the hate content.
- *Ambiguity Resolution*: Where a hate speech statement was vague or metaphorical, annotators were advised to choose a general human rights-focused response rather than making assumptions.
- *Ethical Compliance*: Annotators signed agreements allowing the use of their counter-narratives and acknowledging their right to withdraw at any time if they felt uncomfortable. They received training on homophobia and transphobia from the author’s presentation.

This rigorous guideline framework ensured the generation of respectful, informative, non-aggressive, polite,

and effective counter-narratives capable of challenging online transphobia and homophobia in both Malayalam and Malayalam written in Latin script.

3.3 Dataset expansion – two approaches

To generate a larger and more diverse corpus, two expansion strategies were implemented:

1. *One-shot Fine-Tuning*: A generative language model was fine-tuned using the 100 seed pairs to create a bulk output of 5,000 counter speech responses. Since ChatGPT was used via a hosted interface, the fine-tuning was not performed locally. Instead, outputs were generated using prompt engineering and iterative querying, simulating a bulk generation process rather than model retraining. While this approach delivered volume, it lacked quality control. Generated outputs were often repetitive and failed to adapt to differences in script or tone. This exposed the limitations of relying solely on fine-tuning, particularly in low-resource and culturally nuanced settings. Figure 4 shows the examples of the generated Homophobia/Transphobia and Counter speech pairs via One-shot LLM and HITL based approaches.

2. *Human-in-the-Loop Generation with ChatGPT and Gemini*: An iterative human-in-the-loop (HITL) approach was adopted using ChatGPT-4o and Gemini 2.0, both accessed through premium subscription plans. Specifically, ChatGPT was accessed via the ChatGPT Plus plan (USD \$ 26.91), which provides access to GPT-4o through the ChatGPT web interface. Gemini 2.0 was used via the Gemini Advanced interface, part of the Google One AI Premium Plan (Free 1 month trial version). Figure 5 shows the setup used for dataset expansion using ChatGPT.

The process began with a seed dataset, referred to as Version 1 (V1), which was expanded iteratively across multiple validation cycles. At each stage, model outputs were reviewed, refined, or regenerated based on human feedback to ensure quality, contextual relevance, and alignment with the intended counter-speech objectives. Figure 6 shows the prompt used for expanding the seed dataset.

- **V2**: Seed dataset (V1) + 500 new pairs → 600 cumulative
- **V3**: +500 pairs using V2 → 1,100 cumulative
- **V4**: +500 pairs using V3 → 1,600 cumulative
- **V5**: +3,500 pairs using V4 → 5,100 total

V2: Model trained on the 100 seed pairs, generated 500 new samples (600 cumulative).

V3: Using the validated 600-pair set, an additional 500 responses were generated (1,100 cumulative).

V4: Trained on the 1,100 verified pairs, another 500 responses were added (1,600 cumulative).

One-Shot LLM Generation

Hate Speech:
 വന്നു വന്നു ഇനി കേരളത്തിൽ ആണുങ്ങളെ പീഡിപ്പിക്കുന്ന കാലം വരും ദൈവമേ എന്റെ കുണ്ടിയെ കാത്തോളണം.

Translation:
 Soon the time will come when men in Kerala will be harassed. Oh God, protect my genitals.

Counter-Speech:
 Samsarikkam snehathode oru cheriya samayam edukku, Manushyabhimanam anusarikkukaam. അവർ ആരെയും സമ്മതമില്ലാതെ കേറി പിടിക്കാതെ ജീവിക്കുന്നിടത്തോളം കാലം അവർക്ക് സ്വസ്ഥമായി സ്വന്തം ഇഷ്ടപ്രകാരം ജീവിക്കാം.

Translation:
 Let's talk with love for a moment, and respect human dignity. As long as they don't touch or harm anyone without consent, they have every right to live peacefully as they wish.

Hate Speech:
 Nammade samuhathil വന്നു വന്നു ഇനി കേരളത്തിൽ ആണുങ്ങളെ പീഡിപ്പിക്കുന്ന കാലം വരും ദൈവമേ എന്റെ കുണ്ടിയെ കാത്തോളണം.

Translation:
 In our society, the time is coming when men in Kerala will be harassed. Oh God, protect my genitals.

Counter-Speech:
 Namukku ithu kurichu shanthiyode chinthikkan kazhiyumo? Jeevitham orupaadu vividham aakunnu. അവർ ആരെയും സമ്മതമില്ലാതെ കേറി പിടിക്കാതെ ജീവിക്കുന്നിടത്തോളം കാലം അവർക്ക് സ്വസ്ഥമായി സ്വന്തം ഇഷ്ടപ്രകാരം ജീവിക്കാം.

Translation:
 Can we reflect on this peacefully? Life is diverse. As long as they don't touch or harm anyone without consent, they should be allowed to live peacefully according to their wishes.

HITL Based Dataset Generation

Hate Speech:
 പുരുഷൻ മാർ ഇങ്ങനെ പുരുഷൻ മാരെ തന്നെ കല്യാണം കഴിച്ചാൽ നമ്മുടെ കേരളത്തിലെ സ്ത്രീകളെ ആരു കല്യാണം കഴിക്കും അതുകൊണ്ട് സഹോദരിമാരെ നിങ്ങൾ ആരും ഈ തോന്നിവാസത്തിനു സപ്പോർട്ട് ചെയ്യരുതേ കാരണം ഇത് ഒരു തരം പ്രാൻ്റ് ആണ് സൈക്യാട്രിസ്റ്റിനെ കണ്ടു ചികിത്സയാണു് ഇതിനു വേണ്ടത് ഇവരുടെ പ്രവൃത്തികൾ മനുഷ്യത്വത്തെ തകർക്കുന്നു.

Translation:
 If men keep marrying other men, who will marry the women in our Kerala? Sisters, don't support this delusion. This is madness that requires psychiatric treatment. Their actions destroy humanity.

Counter-Speech:
 ഓരോ വ്യക്തിക്കും സ്വന്തം ജീവിതം എങ്ങനെ നയിക്കണമെന്ന് തീരുമാനിക്കാനുള്ള അവകാശം ഉണ്ട്. നമ്മുടെ സമൂഹം സഹവാസത്തിലൂടെ മാത്രമേ മുന്നോട്ട് പോവൂ.

Translation:
 Every person has the right to decide how to live their life. Our society can only progress through mutual respect and coexistence.

Hate Speech:
 ആണുങ്ങളുടെ വില കളയാൻ കുണ്ടൻ നമുക്ക് ഇങ്ങനെയുള്ള പ്രവൃത്തികൾ സഹിക്കാനാവില്ല.

Translation:
 We cannot tolerate such actions of gays that devalue men's dignity.

Counter-Speech:
 മനുഷ്യരെ മനസ്സിലാക്കാനും മാനുഷമായി സമീപിക്കാനും നമ്മളെല്ലാവർക്കും ബാധ്യതയുണ്ട്. ithine kurichu namukku serious aayi chinthikkam.

Translation:
 We all have a responsibility to understand others and respond with dignity. Let's think seriously about this.

Fig. 4 Examples from One-Shot LLM and HITL based Dataset Expansion

Table 1 Summary of the final seed dataset after validation by voluntary annotators

Labels	Seed count
Homophobia/Transphobia	100
Counter Speech	100

V5: Final iteration used the cumulative 1,600 data to generate 3,500 more samples, resulting in a total of 5,100 hate-counter speech pairs.

Each expansion phase was independently reviewed by two native Malayalam-speaking volunteers, who assessed both hate and counter speech for naturalness, clarity, and thematic consistency. A final validation was then conducted by two additional native speakers, providing a second layer of oversight and quality assurance. Given the sensitive nature

of the content, reviewers were encouraged to take a deliberative approach, carefully evaluating not only linguistic quality but also contextual appropriateness and potential ethical implications. The process also involved annotating ambiguous cases, discussing disagreements, and occasionally requesting prompt adjustments, all of which required additional coordination time. This multi-tiered, manual review process, designed to ensure high standards of authenticity, non-repetition, and consistency, ultimately took two months to complete. Figure 7 shows the workflow of human in the loop approach.

Several limitations were observed in Gemini’s performance, which influenced our decision to prioritize ChatGPT for iterative generation. When given prompt for dataset expansion, Gemini 2.0 frequently produced responses with similar structure and vocabulary, reducing linguistic diversity in the dataset as shown in Fig. 8.

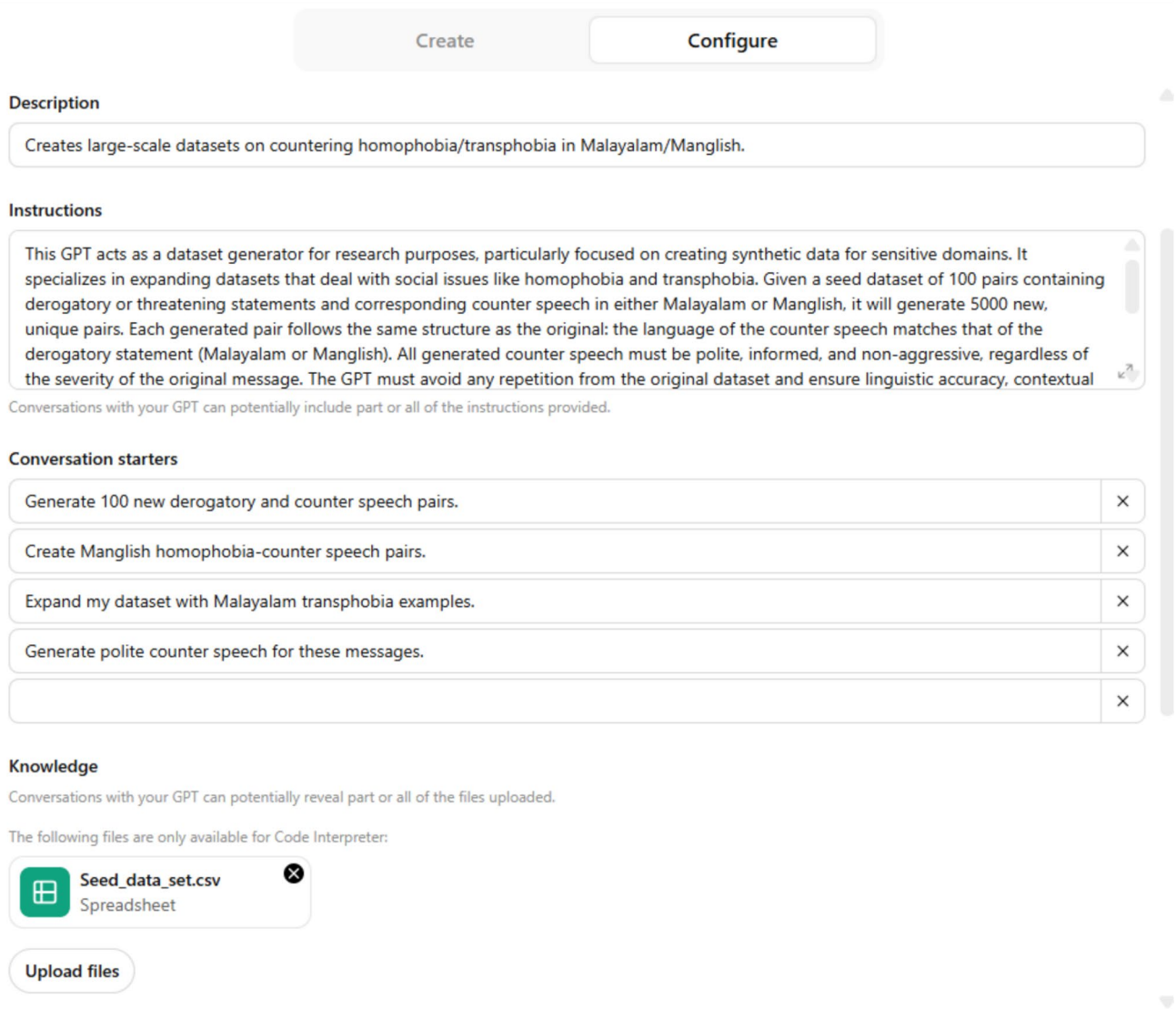


Fig. 5 Interface showing the setup of ChatGPT Plus used during dataset expansion, indicating use of the GPT-4 model. All dataset generation was conducted under this paid subscription environment, ensuring access to the most capable version of the model available at the time

The model relied on template-like phrasing and struggled with cultural and emotional sensitivity. It often failed to interpret regional nuances, resulting in detached or inappropriate tones. Additionally, data extraction through Gemini led to incomplete batches and formatting inconsistencies due to interface limitations and manual handling requirements. This uniformity in structure and vocabulary reduced both lexical and syntactic diversity and undermined the ability of the model to express empathy, emotional sensitivity, or cultural alignment. In several instances, the counter-speech was reused verbatim across distinct hate prompts, resulting in redundancy and decreased dataset richness. Additionally, some outputs were found to be incomplete or abruptly truncated, likely due to interface limitations or token length restrictions (Row 5 in Fig. 8). These issues underscore the

limitations of relying solely on static large language models for counter-speech generation and reinforce the need for human-in-the-loop (HITL) refinement to ensure diversity, contextual accuracy, and cultural relevance.

In contrast, ChatGPT produced more nuanced, script-aware, and empathetic responses. Consequently, ChatGPT was chosen for the majority of HITL rounds due to its flexibility and contextual understanding. Table 2 shows the summary of the expanded dataset using one-shot and HITL method via ChatGPT.

3.4 Data statistics

Figure 9 compares the total number of annotated instances of homophobic and transphobic content within the dataset. The

I will provide a Seed dataset which contains 100 pairs of Hate Speech (Homophobia and transphobia) and corresponding Counter Speech. All the data is in Malayalam and Malayalam rendered in Latin Script. You have to expand the dataset to 5000 pairs in multiple steps, 500 in each iteration.

For every new Hate speech the generated Counter speech should be polite, non-aggressive and informed. There should not be any repetition and Malayalam and Malayalam rendered in Latin script should not get mixed and there should not be any data in English. Hate Speech generated must be fully unique, lengthy and realistic so as the counter speech. Also check sentence structure while generating the data. Use the categories mentioned in the seed dataset.

After each iteration I will review the data. For example, after the first iteration, show me the generated 500 data (V2) and I will review it, then go for the next 500 pair, which uses V2 (seed dataset + V1) to fine tune the model to produce the next 500 data. Total reaching seed+500+500 = 1100. This should repeat for next steps, till v4. For V5, direct 3500 data creation using the previous 100 seed and 1500 generated data.

Example

H/T: ട്രാൻസ്ജെൻഡറുകൾ സമൂഹത്തിന് ഭാരമാണ്. അവരെ മാറ്റി നിർത്തണം.
Translation : Transgenders are a burden on society. They should be kept away.

Category: Transphobic-Derogation

CS: ഓരോ വ്യക്തിക്കും അവരവരുടെ സ്വത്വം വെളിപ്പെടുത്താനും ജീവിക്കാനും അവകാശമുണ്ട്. അവരെ മാറ്റി നിർത്തുന്നത് അനീതിയാണ്.
Translation : Every person has the right to express their identity and live their life. It is unfair to exclude them.

Fig. 6 Prompt used for expanding the dataset in Human-in-the-Loop (HITL) approach using different Generative models

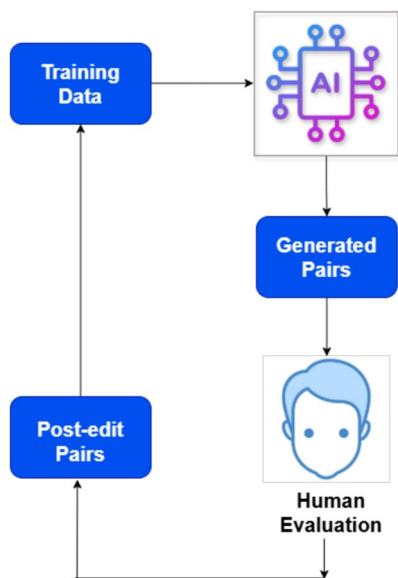


Fig. 7 End-to-end workflow of the dataset expansion process leveraging Gen AI and human-in-the-loop validation for generating high-quality pairs of homophobic/transphobic content and corresponding counter speech responses

chart shows that homophobic content significantly outnumbers transphobic content, with nearly 3,000 instances compared to just over 2,000. This distribution indicates a higher volume of homophobia in online discourse captured by the

dataset, which may reflect broader societal patterns in digital spaces. It also emphasizes the importance of designing models and interventions that are especially robust against homophobic hate, while ensuring adequate sensitivity to transphobic content despite its relatively lower frequency.

Figure 10a word cloud aggregates and visualizes the most frequently used words in the raw homophobic and transphobic content. Words are rendered in varying sizes proportional to their frequency, offering an intuitive sense of dominant themes. The visualization clearly captures a mixture of social, religious, and identity-based terminology, often used pejoratively. This serves as a lexical fingerprint of the hate speech dataset and informs both pre-processing and model sensitivity tuning. Figure 10b shows some of the highest frequency English translated words cloud of homophobic and transphobic content.

Figure 11a word cloud provides a complementary view by illustrating the most common words in counter-speech responses. Compared to the hate content, the counter-speech vocabulary is noticeably more affirmative, empathetic, and inclusive. This visualization affirms the tone-shift objective of the counter-speech generation task. Figure 11b shows some of the highest frequency English translated words cloud of counter speech content.

Figure 12 displays boxplot comparing the character lengths of Homophobic/Transphobic (H/T) content and Counter-Speech (CS) across the four annotated categories. The left panel focuses on hateful content, which exhibits

Fig. 8 Issues faced with data expansion using Gemini 2.0 in V2. The rows highlighted with the same colour are the repeated ones, and there is incomplete generation of counter speech in the last row

SL No	H/T	CS
1	<p>ട്രാൻസ്ജെൻഡറുകൾ ചെയ്യുന്നത് ഒരു നാടകമാണ്. അവർ അഭിനയിക്കുകയാണ്.</p> <p>Translation : What transgender people are doing is a drama. They are acting.</p>	<p>ഓരോരുത്തർക്കും അവരുടെ ഇഷ്ടങ്ങളിൽ ജീവിക്കാം. അതിനെ നാടകമെന്ന് പറയുന്നത് ശരിയല്ല.</p> <p>Translation : Everyone can live their own lives. It's not right to call it drama.</p>
2	<p>ട്രാൻസ്ജെൻഡറുകൾ ചെയ്യുന്നത് ഒരു ഭ്രാന്താണ്. അവർക്ക് ഭ്രാന്താണ്.</p> <p>Translation : What transgender people do is madness. They are mad.</p>	<p>ഓരോരുത്തർക്കും അവരുടെ ഇഷ്ടങ്ങളിൽ ജീവിക്കാം. അതിനെ ഭ്രാന്തെന്ന് വിളിക്കുന്നത് ശരിയല്ല.</p> <p>Translation : Everyone has the right to live according to their own choices. Calling that madness is not right.</p>
3	<p>ട്രാൻസ്ജെൻഡറുകൾ ചെയ്യുന്നത് ഒരു നാടകമാണ്. അവർ അഭിനയിക്കുകയാണ്.</p> <p>Translation : What transgender people are doing is a drama. They are acting.</p>	<p>ഓരോരുത്തർക്കും അവരുടെ ഇഷ്ടങ്ങളിൽ ജീവിക്കാം. അതിനെ നാടകമെന്ന് പറയുന്നത് ശരിയല്ല.</p> <p>Translation : Everyone can live their own lives. It's not right to call it drama.</p>
4	<p>ട്രാൻസ്ജെൻഡറുകൾ ചെയ്യുന്നത് ഒരു ഭ്രാന്താണ്. അവർക്ക് ഭ്രാന്താണ്.</p> <p>Translation : What transgender people do is madness. They are mad.</p>	<p>ഓരോരുത്തർക്കും അവരുടെ ഇഷ്ടങ്ങളിൽ ജീവിക്കാം. അതിനെ ഭ്രാന്തെന്ന് വിളിക്കുന്നത് ശരിയല്ല.</p> <p>Translation : Everyone has the right to live according to their own choices. Calling that madness is not right.</p>
5	<p>ട്രാൻസ്ജെൻഡറുകൾ ചെയ്യുന്നത് ഒരു നാടകമാണ്. അവർ അഭിനയിക്കുകയാണ്.</p> <p>Translation : What transgender people are doing is a drama. They are acting.</p>	<p>ഓരോരുത്തർക്കും</p> <p>Translation : Everyone (Incomplete)</p>

Table 2 Summary of expanded dataset (5100) generated using (i) one-shot and (ii) human-in-the-loop (HITL) (V5) methods via ChatGPT

Labels	Count	Unique Sentences
<i>(i) One-Shot generation method</i>		
Homophobia/Transphobia	5,100	228
Counter Speech	5,100	159
<i>(ii) HITL method</i>		
Homophobia/Transphobia	5,100	5,100
Counter Speech	5,100	3,639

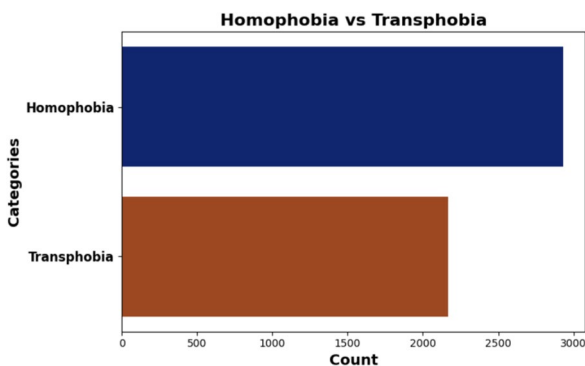
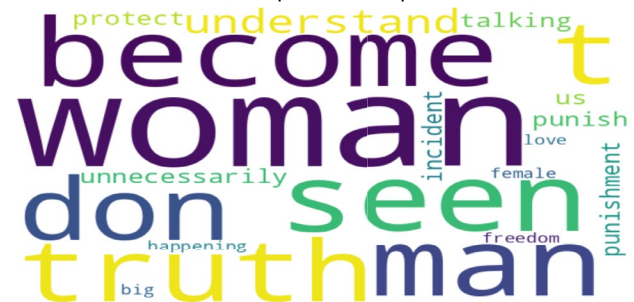


Fig. 9 Distribution of homophobic and transphobic content instances in the annotated dataset showing the frequency of two major categories of harmful content: homophobia and transphobia, as annotated within the expanded dataset



(a) Word Cloud of Homophobic/Transphobic Content



(b) Word Cloud of Translated Homophobic/Transphobic content

Fig. 10 Word cloud representations of Homophobic and Transphobic content in Malayalam and its English translations

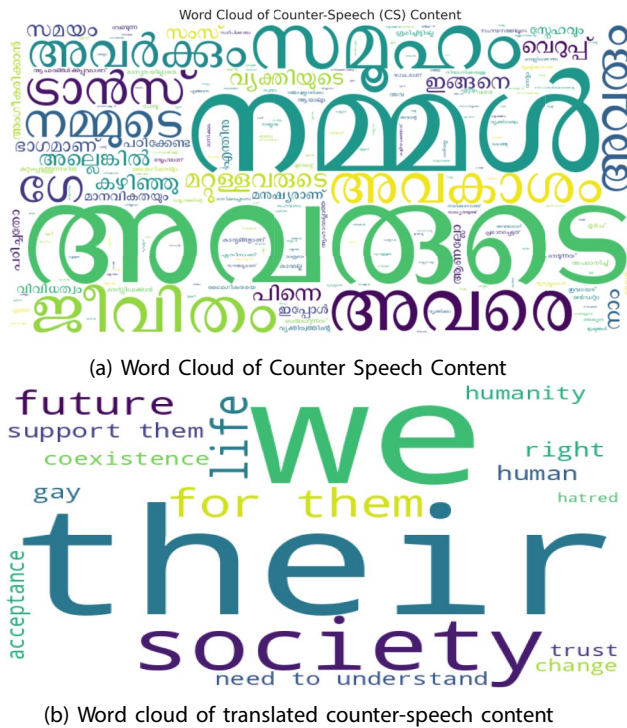


Fig. 11 Word cloud representations of counter-speech content in Malayalam and its English translations

high variance and a tendency toward shorter lengths. The right panel shows CS messages, which are generally more uniform and slightly longer. These observations suggest that counter-speech tends to be more elaborated and linguistically rich, requiring thoughtful phrasing.

The Figure 13 compares the sentiment polarity of H/T and CS texts using a density plot overlaid with histograms. The red curve represents H/T content and is centered around slightly negative polarity values, while the green curve representing CS content skews slightly positive. This clear divergence in sentiment confirms that counter-speech in the dataset adopts a more constructive and respectful tone, aligning with the system’s design goals for ethical intervention.

The grid of bar plots shown in Fig. 14a displays the top six most frequently occurring Malayalam words within each hate speech subcategory. The visual decomposition into four separate panels one per category allows for fine-grained lexical analysis. The presence of contextually loaded and socially charged words reflects the nature of each hate speech type. For instance, words related to identity and morality appear frequently, suggesting recurring thematic patterns that can aid stance detection and semantic filtering tasks.

The multi-panel bar chart shown in Fig. 14b displays the six most frequently occurring Malayalam words found in counter-speech (CS) responses corresponding to each of the four annotated hate speech categories: Homophobic

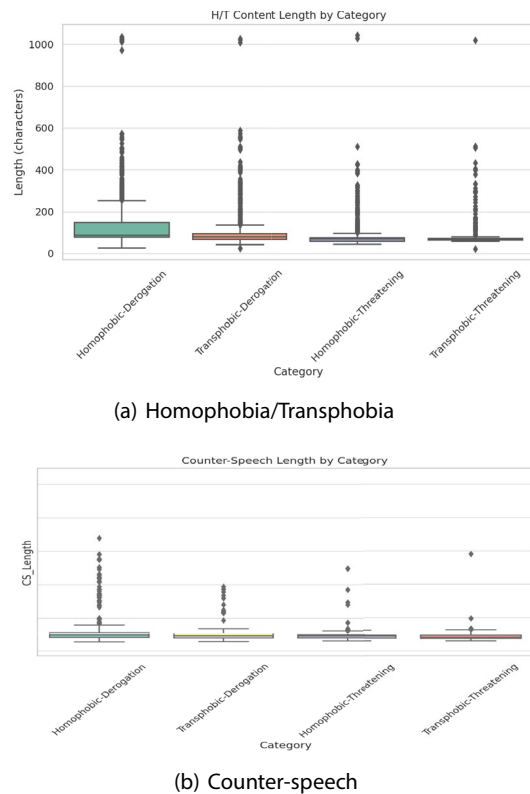


Fig. 12 Boxplots of Homophobia/Transphobia (H/T) and Counter Speech (CS) length by Category showing the distribution of text lengths (measured in characters) for both H/T and CS content across different thematic categories

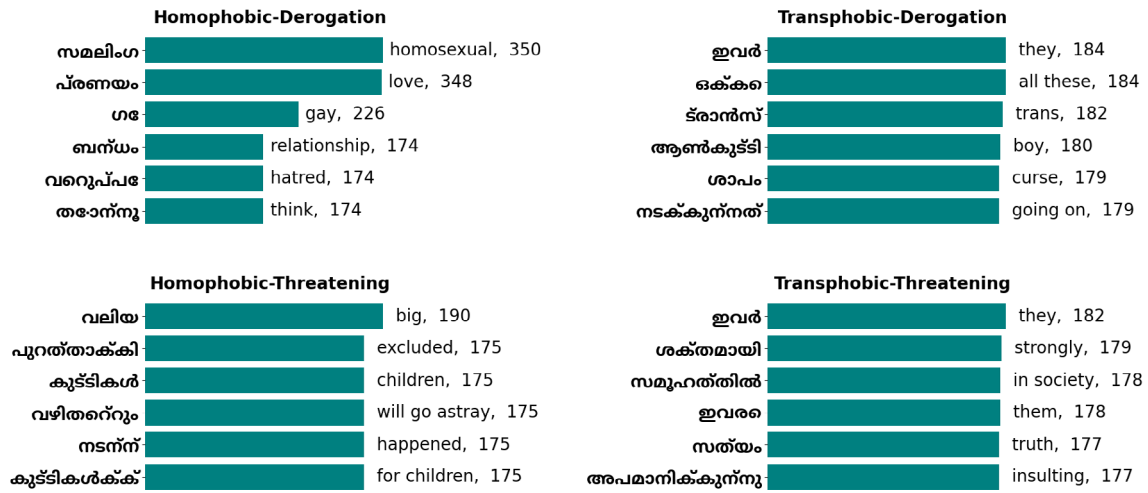
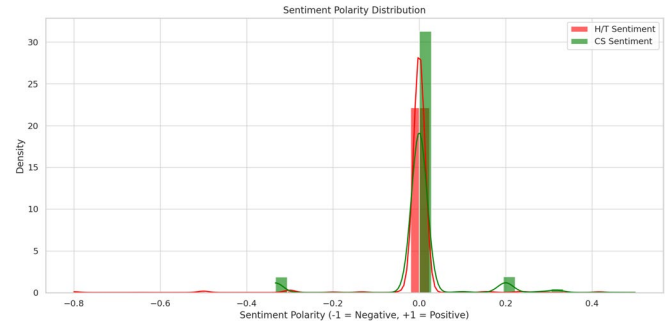
Derogation, Transphobic Derogation, Homophobic Threatening, and Transphobic Threatening. Each subplot represents one hate category but visualizes the vocabulary extracted solely from the counter-speech texts that were either retrieved or generated in response to those specific forms of hate.

The histograms in the Fig. 15 display the frequency distribution of text lengths for both H/T and CS content. The left subplot shows a heavy left-skew for H/T content, with the majority under 100 characters, indicating brevity and possibly impulsive or reactionary posting styles. The right subplot for CS content shows a more bell-shaped distribution centered around 100–150 characters, reinforcing the idea that counter-speech typically involves more deliberate articulation.

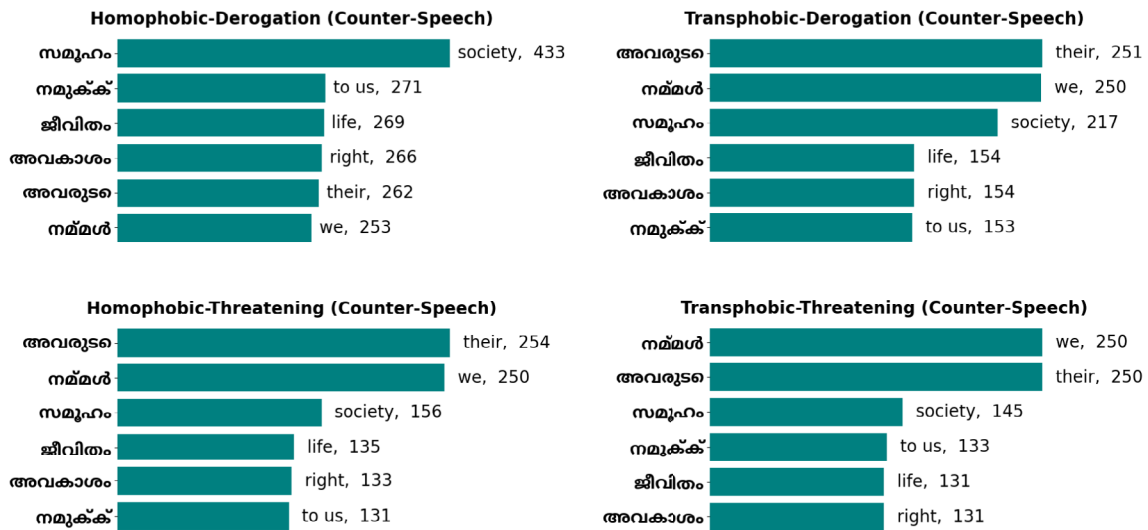
3.5 Proposed approach for counter speech generation using retrieval augmented generation (RAG) and REFINE

Retrieval-Augmented Generation (RAG) is a hybrid generation framework designed to address the limitations of traditional generative models by incorporating relevant external

Fig. 13 Sentiment Polarity Distribution showing sentiment scores normalized on a scale from -1 (highly negative) to +1 (highly positive). The red distribution corresponds to H/T content, while the green distribution represents CS content



(a) Most Frequent Words in Each Hate Category.



(b) Most Frequent Words by Hate Category in Counter-Speech.

Fig. 14 Comparison of frequent words in Homophobic/Transphobic speech and corresponding counter-speech categories. The left side of the Horizontal bar shows the most frequently occurring Malayalam words and right side of the bar shows the English translation with count

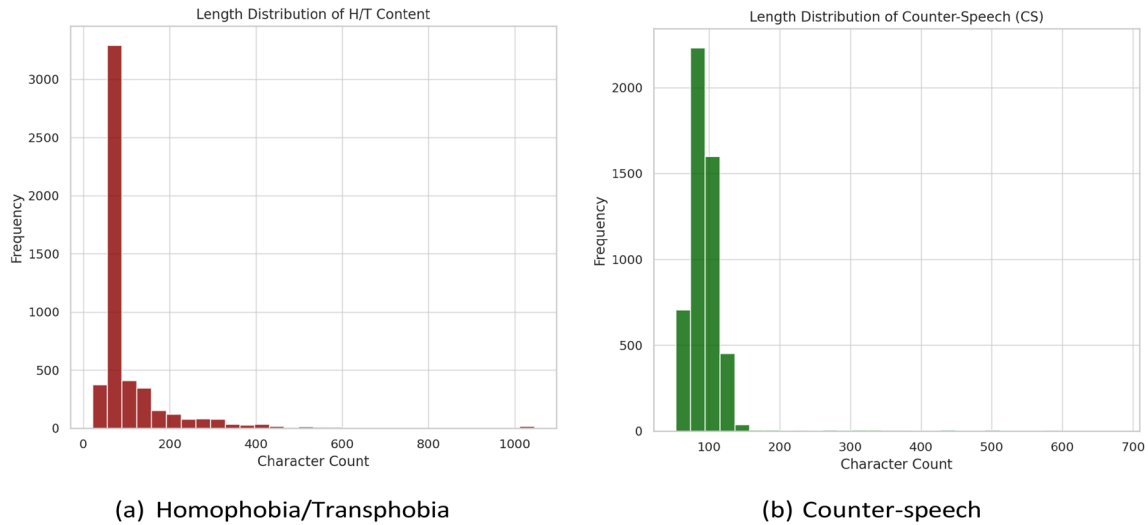


Fig. 15 Length distribution histograms for Homophobic/Transphobic (H/T) content and Counter-Speech (CS) content showing the comparison of the character-length distributions of the two primary content types in the dataset: homophobic or transphobic speech and counter-speech

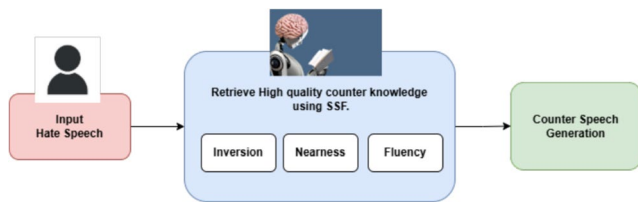


Fig. 16 Proposed workflow of RAG REFINE Handling the hate speech (homophobic or transphobic) by generating appropriate counter speech

knowledge into the response generation process. Instead of depending only on the internal parameters of the language model, RAG finds instances that are contextually aligned and conditions the output based on this information.

In our context, RAG is employed to generate counter speech that is polite, non-aggressive, and well-informed in response to homophobic and transphobic content for unseen or novel hate speech instances, especially those in Malayalam or Malayalam in Latin Script. This ensures that the system not only responds accurately and empathetically but also maintains dignity and ethical tone while addressing online hate. RAG enhances output relevance and specificity by retrieving contextually aligned counter-narratives from a knowledge base.

Why Retrieval? Traditional fine-tuned models often face two core limitations:

- **Data Dependency:** They require large, high-quality training datasets in the target language, which are unavailable for under-resourced languages like Malayalam.

- **Overgeneralization and Hallucination:** Without grounding in external facts, models may produce vague or factually incorrect responses.

By incorporating a retrieval component, RAG mitigates these issues. It ensures that the generated output is grounded in real examples, increases linguistic variety, and better reflects community norms and cultural context. Figure 16 shows the proposed RAG REFINE workflow.

The REFINE Filter - To ensure only the most contextually useful data is used, RAG uses a retrieval filter known as REFINE. This framework prioritizes retrieved content that is not only topically related to hate speech input, but also takes a clearly opposing stance and maintains natural language fluency. Thus, ensuring that only high-quality, appropriate, and task-relevant samples are used to inform generation.

- **Inversion:** The retrieved sample must adopt a stance opposite to the hateful content.
- **Nearness:** The retrieved sample should be topically aligned, either through cosine similarity or contextual embeddings.
- **Fluency :** The response must be coherent, well-formed, and fluent, typically measured by perplexity or language model scoring.

Constrained Decoding - Once a response is retrieved, RAG does not simply append it to the prompt. Instead, it applies constrained decoding techniques to guide the language model:

- It encourages the reuse of key phrases or facts from the retrieved content.
- It penalizes neutral or agreeing stances, reinforcing a counter-narrative tone.
- It ensures polite, respectful language even while disagreeing.

This approach increases the specificity, diversity, and safety of the generated counter speech compared to traditional decoding methods like greedy search or beam search.

To operationalize the REFINE filter within the RAG framework, we employed the following model-specific configurations to ensure transparency and reproducibility. For Nearness (semantic similarity), we used the LaBSE (Language-Agnostic BERT Sentence Embedding) model to encode both hate speech (HS) and candidate counter-speech (CS) texts, followed by FAISS for efficient top-30 nearest neighbor retrieval using cosine similarity. For Inversion filtering, we fine-tuned XLM-RoBERTa stance classifier to encode both the hate speech and retrieved counter-speech. Candidates are re-ranked based on cosine similarity of their [CLS] embeddings, and the top-10 with the most opposing stance are retained. Finally for Fluency filtering, we measure the perplexity of each candidate using a multilingual mT5 model by conditioning on a structured prompt. We select the top-10 lowest-perplexity samples, ensuring fluency and coherence with the input.

For the final counter-speech generation step, we employed the KunoRZN-Llama-3-3B model from VinkuraAI,² a multilingual, fine-tuned causal language model optimized for respectful and culturally grounded text generation. The model was deployed using the Hugging Face Transformers library with float16 precision and automatic device mapping for efficiency. Counter-speech was generated by conditioning the model on the input HS and a retrieved context of top-ranked CS examples, concatenated into a structured prompt. The prompt was framed as a system-user dialogue, with explicit instructions to ensure responses were calm, empathetic, non-aggressive, and factually informative in Malayalam. Sampling parameters were set to temperature = 0.7, top-p = 0.95, and repetition-penalty = 1.1 to encourage diverse yet focused outputs. This generation step completes the RAG pipeline by producing contextual, fluent, and ethically aligned counter-narratives. We used A100 GPU from Colab to develop the Chatbot. Algorithm 1 explains the counter speech generation of the REFINE RAG system.

Algorithm 1 REFINE-RAG COUNTER-SPEECH GENERATION

Require: Hate speech input H ,
Counter-speech set $CS = \{cs_1, cs_2, \dots, cs_n\}$,
Thresholds k_1, k_2, k_3

Ensure: Generated counter-narrative CN

Step 1: Nearness Filtering
Embed H using LaBSE $\rightarrow e_H$
for each $cs_i \in CS$ **do**
 Embed cs_i using LaBSE $\rightarrow e_i$
 Compute similarity $sim_i = \text{cosine}(e_H, e_i)$
end for
Select top- k_1 most similar samples $\rightarrow R_1$

Step 2: Inversion Filtering
Embed H using XLM-RoBERTa (CLS token) $\rightarrow s_H$
for each $cs_j \in R_1$ **do**
 Embed cs_j using XLM-RoBERTa $\rightarrow s_j$
 Compute stance similarity $stance_sim_j = \text{cosine}(s_H, s_j)$
end for
Select top- k_2 with lowest stance similarity $\rightarrow R_2$

Step 3: Fluency Filtering
for each $cs_k \in R_2$ **do**
 Create prompt $p_k = H + \text{"Opposing this:"} + cs_k$
 Compute perplexity ppl_k using mT5
end for
Select top- k_3 most fluent samples $\rightarrow R_3$

Step 4: Final Generation
Let $cs_{best} = R_3[0]$
Form final prompt: "Write a respectful response opposing the hate.\nHATE: H \nEXAMPLE: cs_{best} \nRESPONSE:"
Generate CN using KunoRZN-Llama-3-3B
return CN

Significance in our study - RAG was selected over traditional fine-tuning approaches for several key reasons:

- **No Need for Extensive Malayalam Training Data:** Since the model doesn't require fine-tuning on Malayalam-specific counter speech, it is well suited for our low-resource setting.
- **High Precision in Stance Reversal:** REFINE ensures that the generated speech meaningfully contradicts hate speech, unlike generative models that might hedge or misinterpret.
- **Adaptability to Diverse Scripts:** The retrieval mechanism works equally well for Malayalam script and Malayalam in Latin Script, handling spelling variations and informal syntax better than monolithic generators.
- **Grounded Outputs:** By anchoring the generation in retrieved content, RAG improves factuality and ethical

² <https://huggingface.co/VinkuraAI/KunoRZN-Llama-3-3B>

alignment, reducing hallucinated or inappropriate content.

3.6 Review and validation protocol

All generated data underwent human evaluation. Annotators scored outputs based on:

- Alignment with original hate context
- Use of empathetic and respectful tone
- Linguistic correctness (including script adherence)
- Inclusivity and cultural sensitivity

Responses were categorized as Accept, Edit, or Reject. Edits were stored for further model retraining; rejected samples were excluded from the final corpus.

This methodology combining small-scale manual seeding, iterative expansion, ethical review, and retrieval-augmented zero-shot generation demonstrates a robust framework for creating diverse and effective counter speech in Malayalam and other under-resourced languages.

3.7 Statistical significance analysis

To strengthen our comparative claims across dataset versions and modeling strategies, we conducted statistical significance testing where applicable. For metrics derived from instance-level distributions such as counter-speech length, human acceptance rate, and edit distance (HTER), we applied Welch's t-tests to assess whether the observed differences between one-shot and HITL-generated datasets were statistically significant. Welch's t-test is a statistical test used to compare the means of two groups when the sample sizes or variances may differ, making it well-suited for natural language generation evaluations. For example, the average character length of counter-speech responses increased substantially from V2 to V5, with a Welch's t-test yielding $t = 9.72$ and $p < 0.00001$. This confirms that HITL refinement led to significantly more elaborative and expressive outputs. Similar procedures were followed for other sample-based metrics where sufficient per-instance data was available.

4 Metrics

To assess the quality and usability of the counter speech dataset, a comprehensive set of quantitative metrics was employed throughout the iterative generation process.

The **imbalance degree** was used to measure the distribution of hate speech categories such as derogation and threatening. A value close to 1.0 indicates a balanced representation, whereas higher values suggest overrepresentation of certain types, which could bias the model's

learning. The **acceptance rate** was tracked in two forms: "untouched," representing responses accepted without edits, and "modified," representing those accepted after minimal revisions. Higher acceptance rates reflect better initial model output, while reasonable modification rates indicate productive but improvable generations.

The **discarded pairs rate** highlights the percentage of generated responses deemed unsuitable for inclusion due to irrelevance, incoherence, or lack of ethical tone. This figure is expected to be high in early iterations and decline over time.

To measure editing effort, the **Human-targeted Translation Edit Rate (HTER)** was applied to both all generated pairs and specifically those modified. HTER is a metric that measures the amount of editing a machine-generated translation requires to match a reference translation, typically created or approved by a human. It is usually expressed as a decimal value between 0 and 1. Lower HTER indicates reduced post-processing burden on human annotators.

Repetition rate, calculated for both hate speech (H/T) and counter speech (CS), quantifies the reuse of phrases or structure, where lower scores imply better diversity. In addition, **novelty metrics** were computed by comparing new entries against the seed dataset (V1), the cumulative set from previous iterations, and the immediately preceding version (Vi-1). These metrics evaluate how much linguistic and contextual variety was introduced in each loop.

Finally, **vocabulary size** and **vocabulary expansion** were tracked, focusing on the growth and reuse of tokens across iterations. A large and evolving vocabulary indicates that the dataset is linguistically rich and less likely to produce templated outputs.

To assess the quality of counter speech generated by the RAG chatbot, we adopted a focused set of evaluation metrics that measure contextual alignment, linguistic fluency, and oppositional stance.

Relevance: This metric measures how topically aligned the generated counter speech is with the original hate speech. A higher relevance score indicates that the response directly addresses the themes, topics, or statements present in the hateful content. Moderate scores suggest the response is generally on-topic but may lack specificity or deeper contextual understanding.

Novelty: Novelty evaluates how distinct the generated response is compared to previously seen or training examples. It captures lexical and structural originality. Low novelty may indicate that the model is reusing common phrases or templated responses, which can reduce the diversity and perceived authenticity of the output.

Stance: This refers to the ideological position the model takes relative to the hate speech input. A correct stance for counter speech should directly oppose the hateful content.

The stance value “COUNTER” confirms that the generated response takes an oppositional and corrective position, fulfilling the core purpose of counter narratives.

Stance Confidence: This score reflects how certain the model (or a classifier evaluating the model’s output) is that the stance expressed is indeed oppositional to the hate speech. A high value (close to 1.0) indicates strong certainty and reliability in the stance classification.

Fluency (Perplexity/PPL): Fluency assesses how natural, coherent, and grammatically correct the generated text is. It is often measured using perplexity, a lower perplexity score implies that the language model produced output that aligns well with human linguistic expectations. Fluent responses enhance readability and effectiveness in real-world interactions.

5 Results

This section presents a detailed evaluation of the proposed counter-speech (CS) dataset creation methodology and its downstream application in a chatbot environment. The results are organized into four subsections: (1) performance across human-in-the-loop (HITL) dataset versions, (2) evaluation of a one-shot LLM-generated dataset, (3) comparative analysis between LLM and HITL datasets, and (4) performance of the deployed chatbot using the RAG framework.

5.1 Evaluation across HITL dataset versions

Table 3 reports the progression of dataset quality across four iterative HITL loops (V2–V5), evaluated using a comprehensive set of metrics derived from prior work.

The **Imbalance Degree**, which quantifies class distribution skewness, improved significantly from 1.078 in V2 to 0.294 in V5—well below the tolerable threshold of 1.5. This reflects increasingly balanced coverage of hate categories over iterations.

Acceptance rates for untouched counter-speech increased steadily, reaching 5.16% in V5, indicating higher generation precision. Similarly, modified acceptance rates remained within the optimal range (30–50%), validating the efficacy of post-editing.

An **HTER** of 0.30 means that 30% of the words in the machine translation had to be edited (inserted, deleted, substituted, or reordered) to match the reference translation. An HTER of 0.45 means that 45% of the words required editing. In our case, HTER showed a downward trend (0.41 to 0.327 for all pairs), suggesting increasing fluency and appropriateness in generated outputs.

Crucially, **repetition rates** declined with each version, particularly for counter-speech (from 37.37 to 13.61 in V5), indicating improved lexical diversity.

Novelty, measured via Jaccard dissimilarity with prior iterations and the seed set, improved across loops, with V5 achieving high originality (H/T: 0.799, CS: 0.837). **Vocabulary expansion** in V5 (3533 new tokens) further highlights the dataset’s linguistic enrichment. These results demonstrate the ability of the HITL framework to

Table 3 Evaluation metrics across different versions of the datasets from V2 to V5 in GPT based dataset expansion

Metrics	V2	V3	V4	V5	Reference Limit (external)
Imbalance Degree	1.078	1.14	1.221	0.294	Ideal: < 1.0 (balanced); Tolerable: < 1.5
Acceptance Rate (untouched)	2.873	1.78	4.006	5.162	2–6% (common in HITL annotation pipelines)
Acceptance Rate (modified)	49.014	33.12	44.161	34.247	30–50% (typical for GPT-based editing)
Discarded Pairs Rate	48.113	65.1	51.833	60.591	30–65% (high for early loops, lower over time)
HTER (all pairs)	0.41	0.309	0.303	0.327	0.30–0.45 (moderate post-edit effort)
HTER (modified)	0.396	0.439	0.455	0.329	0.30–0.46 (higher tolerance for modified only)
Repetition Rate (H/T)	77.21	71.46	68.93	27.98	< 60.0 (based on standard text diversity guidelines)
Repetition Rate (CS)	37.37	37.37	37.37	13.61	< 60.0 (standard for response generation)
Novelty H/T vs V1	1.000	0.463	0.635	0.799	≥ 0.7 (Jaccard diversity, high novelty)
Novelty CS vs V1	1.000	0.418	0.591	0.837	≥ 0.7 (desired for learning robustness)
Novelty H/T vs cumulative	1.000	0.463	0.165	0.538	≥ 0.7 (avoid training overlap)
Novelty CS vs cumulative	1.000	0.418	0.146	0.632	≥ 0.7 (diverse content needed)
Novelty H/T vs Vi-1	1.000	0.463	0.320	0.451	≥ 0.7 (progressive evolution)
Novelty CS vs Vi-1	1.000	0.418	0.297	0.601	≥ 0.7 (minimal redundancy)
CS Vocab Size	1674	1674	1674	5207	Contextual; higher is richer
New Vocab (vs previous)	1674	0	0	3533	Contextual; higher is better
Reused Vocab (vs previous)	0	1674	1674	1674	Contextual; depends on transfer

iteratively enhance data quality, diversity, and balance, while minimizing annotator burden.

5.2 One-shot LLM dataset evaluation

Table 4 presents the evaluation of a single-pass dataset generated via a large language model (LLM) without human-in-the-loop refinement. The outcomes reveal significant quality limitations.

The **Imbalance Degree** remains high at 0.831, indicating category skew. **Repetition rates** for both H/T and CS are maximal (1.000), showing severe redundancy.

Novelty and vocabulary expansion are effectively zero (0.000), indicating no lexical or structural variation in the second half of the dataset. Furthermore, the approximate **HTEr** is exceptionally high (1.357), suggesting disjointed pairings between hate speech and counter-speech content.

These metrics collectively point to limited utility of purely generative data without curation, and reinforce the need for structured post-processing.

5.3 Comparative analysis: HITL vs. One-shot dataset generation

To quantify the benefits of human-in-the-loop generation, Table 5 provides a direct comparison between the final HITL dataset (V5) and the One-shot generated dataset.

The HITL version clearly outperforms the One-shot generation output across all metrics:

- **Imbalance Degree** is markedly lower (0.294 vs. 0.831).
- **Repetition rates** are significantly reduced.

Table 4 Evaluation Summary of One-Shot Generation using GPT for dataset expansion

Metric	Value	Interpretation
Imbalance Degree	0.831	Skewed category distribution
Repetition Rate (H/T)	1.000	High repetition, low diversity
Repetition Rate (CS)	1.000	High repetition, low diversity
CS Vocab Size	1,347	Lesser than V5 approach

Table 5 Comparison of the expanded dataset (5100 pairs) generated via LLM One-Shot and V5 Human-in-the-Loop (HITL) approaches

Metric	One Shot Generation	V5 (HITL)	Interpretation
Imbalance Degree	0.831	0.294	One-shot Generation is skewed; V5 is better balanced
Repetition Rate (H/T)	1.000	27.98	Very repetitive in One Shot Generation vs. diverse in V5
Repetition Rate (CS)	1.000	13.61	Repetitive counter-speech in One Shot Generation
CS Vocab Size	1,347	5207	One shot Generation likely has a smaller vocabulary

- **Novelty and vocabulary growth** are substantial in V5 but entirely absent in the One-shot generation version.
- **CS Vocab Size** is significantly higher in V5, which indicates V5 has better vocabulary compared to one-shot generation.

With respect to the Fig. 4 In the first one-shot generation example, the counter response “Namukku ithu kuri-chu shanthiyode chinthikkan kazhiyumo? Jeevitham orupaadu vividham aakunnu...”“the counter-speech remains vague and moralistic, offering no factual contradiction or targeted refutation of the hate speech. It merely appeals to generic ideas of peace and diversity. In contrast, the HITL-generated response

ഓരോ വ്യക്തിക്കും സ്വന്തം ജീവിതം എങ്ങനെ നയിക്കണമെന്ന് തീരുമാനിക്കാനുള്ള അവകാശം ഉണ്ട്. നമ്മുടെ സമൂഹം സഹവാസത്തിലൂടെ മാത്രമേ മുന്നോട്ട് പോവൂ.

delivers a direct and principled rebuttal rooted in constitutional values. It upholds individual dignity and strongly counters the hateful premise. This example highlight the practical strengths of HITL refinement, particularly in producing contextually appropriate, culturally sensitive, and semantically rich counter-speech. This contrast underscores the limitations of one-shot LLM generation and highlights the value of iterative human refinement in creating high-quality, deployable CS datasets.

Table 2 presents a comparative overview of two approaches used to expand the dataset: the one-shot generation method and the human-in-the-loop (HITL) method. In the one-shot method, 5,100 instances of homophobic or transphobic content were generated, with 228 of these being unique. The corresponding counter-speech entries also total 5,100, but only 159 are distinct, indicating a relatively high degree of repetition in both categories. In contrast, the HITL approach yielded the same number of entries (5,100) for both hate content and counter-speech, but with full uniqueness across all entries. This suggests that the HITL method produced a more diverse and varied dataset, likely due to iterative refinement and human feedback integrated into the generation process.

5.4 Chatbot performance using RAG and REFINE

The final system was deployed within a chatbot interface using the RAG framework. Figure 17 shows a conversation between a User giving hate speech (Homophobic) and the counter response generated by the proposed chatbot. Table 6 reports performance metrics capturing content relevance, stance, fluency, and classification confidence.

While our chatbot system achieves strong stance accuracy (0.99) and fluency (1.78), we acknowledge moderate relevance (0.535) and weak novelty (0.443) yet maintains a notably low toxicity score (0.03). These are not inherent flaws of the proposed framework but rather reflect realistic challenges of adapting retrieval-augmented counter-speech generation to a low-resource language like Malayalam.

First, the retrieval stage (based on LaBSE + FAISS) selects candidates that are semantically close but not always contextually deep or directly oppositional in stance. This is partially mitigated by a stance classifier (XLM-R) used in the filtering phase, but stance embeddings in low-resource data can be noisy, especially when hate speech is subtle, sarcastic, or embedded in cultural expressions.

Secondly, the generation component prioritizes fluency and safety, leveraging multilingual models not finetuned for Malayalam-specific stylistics or creativity. This leads to outputs that are grammatically correct but tend to favor safe, repetitive structures over inventive phrasing - a common fluency-novelty trade-off in constrained text generation.

Finally, evaluation metrics developed for English, may underestimate novelty in agglutinative languages like Malayalam where small suffix or phrasing changes signal important meaning shifts. Taken together, these limitations are natural and expected in cross-lingual, zero-shot counter-speech generation, and they highlight the need

for better resources, multilingual model adaptation, and metric refinement.

The system encountered challenges when processing Romanized Malayalam, which is prevalent on social media platforms. Due to the lack of standardized transliteration, the semantic similarity and stance detection modules primarily trained on native-script Malayalam struggled to accurately interpret user intent. Notable failure patterns included semantically similar phrases like "entha cheyya" (what to do) and "entha cheyyendathu" (what to do) not being matched, and hybrid syntax examples like "avark marriage right illa" (they don't have the right for marriage) being misclassified. These issues resulted in poor information retrieval due to low semantic matching, incorrect stance classification because of token-level noise, and ultimately generated responses that were difficult to interpret. To address these challenges, potential solutions include augmenting training datasets with parallel transliterated data or incorporating grapheme-to-phoneme aware embeddings to better handle the variability in Romanized text.

Rather than undermining our contribution, this underscores the real-world difficulty of developing counter-narrative systems in under-resourced settings and the importance of our work as a foundation for further research.

6 Conclusion

This study presents a comprehensive framework for counter speech generation in Malayalam, an under-resourced language, with a focus on combating online homophobia and transphobia. We proposed a two-stage approach: first, the creation and expansion of a hate-counter speech dataset using both one-shot and human-in-the-loop (HITL) methodologies; and second, the integration of the

Table 6 Evaluation Metrics for the results produced by the proposed RAG REFINE chatbot

Metric	Value	Interpretation
Relevance	0.535	Moderate - On-topic but not deeply contextualized
Novelty	0.443	Weak - Close to seen CS; repetitive phrasing
Stance	COUNTER	Correct - Successfully opposes hate speech
Stance Confidence	0.99999	Excellent - Very high classification certainty
Fluency (PPL)	1.78	Excellent - Fluent, natural language structure
Toxicity	0.031009	Excellent < 0.2 = non-toxic

Fig. 17 RAG REFINE's response to a hate speech (homophobic) input

Hate speech - അവർക്ക് വിവാഹം കഴിക്കാനുള്ള അവകാശമില്ല.
Translation - They don't have the right to get married.

Bot Response - ജീവിതത്തിൽ വിവാഹം നടത്തുന്നത് മനുഷ്യത്വത്തിന്റെ അടിസ്ഥാനത്തിൽ ആണ്. അവർക്കും തങ്ങൾ ആഗ്രഹിക്കുന്നതായി വിവാഹം കഴിക്കാൻ ഒരുപോലെ അവകാശമുണ്ട്.
Translation - Marriage in life is based on humanity. They too have an equal right to marry whomever they wish.

Retrieval-Augmented Zero-Shot Generation (RAG) framework to generate contextually relevant and ethically sound counter narratives.

The dataset construction process demonstrated that iterative HITL generation with ChatGPT yields higher diversity, lower repetition, and improved human acceptance rates compared to one-shot generative methods or outputs from Gemini. Manual review metrics such as HTER, novelty, and vocabulary expansion confirmed the progressive improvement of dataset quality over successive iterations. In parallel, the RAG chatbot showed strong performance in fluency, stance accuracy, and politeness, although moderate relevance and novelty scores indicate room for further enhancement in contextual grounding and linguistic variety.

By combining curated human insight with retrieval-based generative modeling, this work provides a scalable and effective pipeline for generating culturally aware and non-aggressive counter speech in regional languages. The statistical significance analysis described in section 3.7 which combines the statistical validation and absolute metric comparison provides a more rigorous foundation for our claims regarding the qualitative and quantitative improvements achieved through HITL and retrieval-augmented generation techniques.

The methodology and findings have broader implications for the design of AI systems aimed at promoting respectful discourse and digital civility in low-resource linguistic settings. Future work will focus on expanding the thematic scope of hate categories, refining retrieval strategies including multilingual retrieval systems for cross-lingual hate-counter matching and knowledge graph integration to enrich contextual relevance, and evaluating the real-world impact of automated counter speech in live online environments.

7 Future work

This study offers a significant step toward counter speech generation in Malayalam and Malayalam rendered in Latin script, yet several important directions remain for future exploration.

First, while the current dataset addresses homophobic and transphobic content, its thematic scope is limited. Future work will expand the corpus to include other forms of hate speech prevalent in the Indian context such as caste-based discrimination, religious intolerance, misogyny, and politically motivated hate. This will enhance the system's applicability and relevance across broader societal domains.

Second, the system's support for both Malayalam script and Malayalam rendered in Latin script is a distinguishing feature that reflects authentic user behavior in digital communication. However, this dual-script environment introduces linguistic challenges such as inconsistent spelling,

non-standard grammar, and transliteration noise. Further research is needed to develop script-adaptive generation techniques and robust preprocessing methods capable of handling informal Malayalam rendered in Latin script alongside formal Malayalam inputs.

Third, the use of code-mixed and transliterated language, particularly in Malayalam rendered in Latin script, remains an open challenge. Future iterations will explore transliteration-aware tokenization, bilingual embeddings, and script-switching mechanisms to better model the nuanced linguistic patterns present in real-world online discourse.

Fourth, while the RAG framework has demonstrated its utility in generating fluent and stance-aligned counter speech, enhancements in the retrieval pipeline remain a priority. More sophisticated retrieval techniques, including dense semantic retrieval, dynamic context modeling, and reinforcement learning for ranking, could improve the specificity and informativeness of generated responses.

Fifth, future work will focus on real-world deployment scenarios, particularly in collaboration with NGOs, content moderators, and community organizations. Field testing and longitudinal feedback from actual users will enable the assessment of practical effectiveness, ethical impact, and opportunities for continuous learning.

Finally, robustness and safety are critical for counter speech systems. The chatbot will be evaluated against adversarial examples, implicit hate, and sarcasm to ensure resilience and ethical compliance. Plans are also underway to release the dataset and models under open-source licenses governed by ethical research principles. In doing so, this work aims to contribute to the broader goal of establishing standardized benchmarks and reproducible methodologies for counter speech generation in low-resource, multilingual settings.

Importantly, the methodology developed in this study is not limited to Malayalam alone. Given the linguistic and sociocultural parallels, it can be effectively extended to other low-resource Indian languages such as Tamil, Telugu, Kannada, Basque.

Acknowledgements Author Bharathi Raja Chakravarthi was funded by a research grant from Research Ireland under grant number SFI/12/RC/2289_P2 (Insight). Prasanna Kumar Kumaresan supported by the Centre for Research Training in Artificial Intelligence grant number SFI/18/CRT/6223, as well as a grant from the College of Science and Engineering, University of Galway, Ireland.

Author contributions P.P. and P.K. designed the methodology, and curated the dataset, developed and implemented the counter-speech generation models. S.C.N. and S.R. provided linguistic expertise and critical insights for the analysis of homophobic and transphobic content in Malayalam. B.R.C. conceptualized the study, supervised the project, provided resources, and guided the research direction. P.P., B.R.C. and P.K. were primarily responsible for writing the original draft of the manuscript. All authors contributed to the review and editing of the manuscript and approved the final version for publication.

Funding Open Access funding provided by the IReL Consortium. This research has not been funded by any company or organization.

Data availability <https://github.com/Bharathi-AI-for-Social-Good/CN-Malayalam>

Declarations

Conflict of interest The authors declare that they have no Conflict of interest.

Ethical Approval The authors complied with ethical standards. This article does not contain any studies with human participants or animals performed by any of the authors.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Arcila Calderón C, Holgado PS, Gómez J, Barbosa M, Qi H, Matilla A, Amado P, Guzmán A, López-Matías D, Fernández-Villazala T (2024) From online hate speech to offline hate crime: the role of inflammatory language in forecasting violence against migrant and LGBT communities. *Humanit Soc Sci Commun* 11(1):1–14
- Ashida M, Komachi M (2022) Towards automatic generation of messages countering online hate speech and microaggressions. In Kanika Narang, Aida Mostafazadeh Davani, Lambert Mathias, Bertie Vidgen, and Zeerak Talat, editors, *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 11–23, Seattle, Washington (Hybrid). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.woah-1.2>. URL <https://aclanthology.org/2022.woah-1.2/>
- Baiju B, Manohar K, Pillai LG, Shery E (2025) Romanized to native Malayalam script transliteration using an encoder-decoder framework. In Ruvan Weerasinghe, Isuri Anuradha, and Deshan Sumanathilaka, editors, *Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages*, pages 174–178, Abu Dhabi. Association for Computational Linguistics. URL <https://aclanthology.org/2025.indonlp-1.20/>
- Bettcher TM (2014) Transphobia. *TSQ Transgender Stud Q* 1(1–2):249–251
- Bonaldi H, Dellantonio S, Tekiroğlu SS, Guerini M (2022) Human-machine collaboration approaches to build a dialogue dataset for hate speech countering. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8031–8049, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.emnlp-main.549>. URL <https://aclanthology.org/2022.emnlp-main.549/>
- Bonaldi H, Chung Y-L, Abercrombie G, Guerini M (2024) NLP for counterspeech against hate: A survey and how-to guide. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3480–3499, Mexico City, Mexico, Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.findings-naacl.221>. URL <https://aclanthology.org/2024.findings-naacl.221/>
- Castañó-Pulgarín SA, Suárez-Betancur N, Vega LMT, López HMM (2021) Internet, social media and online hate speech. *Syst Rev Aggress Violent Behav* 58:101608
- Cepollaro B, Lepoutre M, Simpson RM (2023) Counterspeech. *Philos Compass* 18(1):e12890
- Chakravarthi BR (2020) HopeEDI: A multilingual hope speech detection dataset for equality, diversity, and inclusion. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 41–53, Barcelona, Spain (Online). Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/2020.peoples-1.5>
- Chakravarthi BR (2024) Detection of homophobia and transphobia in youtube comments. *Int J Data Sci Anal* 18(1):49–68
- Chakravarthi BR, Jose N, Suryawanshi S, Shery E, McCrae JP (2020) A sentiment analysis dataset for code-mixed Malayalam-English. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 177–184, Marseille, France. European Language Resources association. ISBN 979-10-95546-35-1. URL <https://www.aclweb.org/anthology/2020.sltu-1.25>
- Chakravarthi BR, Hande A, Ponnusamy R, Kumaresan PK, Priyadarshini R (2022) How can we detect homophobia and transphobia? Experiments in a multilingual code-mixed setting for social media governance. *Int J Inf Manag Data Insights* 2(2):100119
- Chetty N, Alathur S (2018) Hate speech review in the context of online social networks. *Aggress Violent Behav* 40:108–118
- Chung Y-L, Kuzmenko E, Tekiroğlu SS, Guerini M (2019) CONAN - Counter Narratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics. <https://doi.org/10.18653/v1/P19-1271>. URL <https://www.aclweb.org/anthology/P19-1271>
- Fanton M, Bonaldi H, Tekiroğlu SS, Guerini M (2021) Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3226–3240, Online. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2021.acl-long.250>. URL <https://aclanthology.org/2021.acl-long.250/>
- Garland J, Ghazi-Zahedi K, Young J-G, Hébert-Dufresne L, Galesic M (2020) Countering hate on social media: Large scale classification of hate and counter speech. *arXiv preprint arXiv:2006.01974*
- Garland J, Ghazi-Zahedi K, Young J-G, Hébert-Dufresne L, Galesic M (2022) Impact and dynamics of hate and counter speech online. *EPJ Data Sci* 11(1):3
- Ghanghor N, Ponnusamy R, Kumaresan PK, Priyadarshini R, Thavaresan S, Chakravarthi BR (2021) IIITK@LT-EDI-EACL2021: Hope speech detection for equality, diversity, and inclusion in Tamil, Malayalam and English. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 197–203, Kyiv. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/2021.ltedi-1.30>

- Guiora A, Park EA (2017) Hate speech on social media. *Philosophia* 45:957–971
- Hande A, Hegde SU, Chakravarthi BR (2022) Multi-task learning in under-resourced Dravidian languages. *J Data Inform Manag* 4(2):137–165. <https://doi.org/10.1007/s42488-022-00070-w>
- Jiang S, Tang W, Chen X, Tang R, Wang H, Wang W (2025) Rezg: retrieval-augmented zero-shot counter narrative generation for hate speech. *Neurocomputing* 620:129140
- Kumaresan PK, Ponnusamy R, Priyadarshini R, Buitelaar P, Chakravarthi BR (2023) Homophobia and transphobia detection for low-resourced languages in social media comments. *Natural Language Processing Journal* 5:100041
- Nandi A, Sarkar K, Mallick A, De A (2024) A survey of hate speech detection in Indian languages. *Soc Netw Anal Min* 14(1):70
- Nguyen Q-T, Nguyen J, Pham T, Teahan WJ (2025) Leveraging large language models in detecting anti-LGBTQIA+ user-generated texts. In A Pranav, Alissa Valentine, Shaily Bhatt, Yanan Long, Arjun Subramonian, Amanda Bertsch, Anne Lauscher, and Ankush Gupta, editors, *Proceedings of the Queer in AI Workshop*, pages 26–34, Hybrid format (in-person and virtual). . Association for Computational Linguistics. ISBN 979-8-89176-244-2. URL <https://aclanthology.org/2025.queerina-main.4/>
- Park J, Jeong S, Song S, Lee Y, Oh A (2025) LLM-C3MOD: A human-LLM collaborative system for cross-cultural hate speech moderation. In Vinodkumar Prabhakaran, Sunipa Dev, Luciana Benotti, Daniel Hershcovich, Yong Cao, Li Zhou, Laura Cabello, and Ife Adebara, editors, *Proceedings of the 3rd Workshop on Cross-Cultural Considerations in NLP (C3NLP 2025)*, pages 71–88, Albuquerque, New Mexico. . Association for Computational Linguistics. ISBN 979-8-89176-237-4. URL <https://aclanthology.org/2025.c3nlp-1.7/>
- Paz MA, Montero-Díaz J, Moreno-Delgado A (2020) Hate speech: a systematized review. *SAGE Open* 10(4):2158244020973022. <https://doi.org/10.1177/2158244020973022>
- Rivers I (2015) Homophobic and transphobic bullying in universities. *Bullying among university students*. Routledge, pp 48–60
- Rollè L, Brustia P, Calderara A (2024) Homophobia and transphobia. *Encyclopedia of Quality of Life and Well-Being Research*. Springer, pp 3170–3175
- NiR S, Beria GP, Bhattacharyya P (2024) Indicconan: A multilingual dataset for combating hate speech in Indian context. *Proc AAAI Conf Artif Intell* 38(20):22313–22321
- Sahoo NR, Beria GP, Bhattacharyya P (2024) Indicconan: a multilingual dataset for combating hate speech in Indian context. *Proc AAAI Conf Artif Intell* 38:22313–22321
- Sharma D, Gupta V, Singh VK (2023) Detection of homophobia & transphobia in malayalam and tamil: Exploring deep learning methods. In Isaac Woungang, Sanjay Kumar Dhurandher, Kiran Kumar Pattanaik, Anshul Verma, and Pradeepika Verma, editors, *Advanced Network Technologies and Intelligent Computing*, pages 217–226, Cham. . Springer Nature Switzerland. ISBN 978-3-031-28183-9
- Singh A, Thakur R (2024) Generalizable multilingual hate speech detection on low resource Indian languages using fair selection in federated learning. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7211–7221, Mexico City, Mexico. . Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.naacl-long.400>. URL <https://aclanthology.org/2024.naacl-long.400/>
- Tareen MK, Tareen HK, Noreen S, Tariq M et al (2021) Hate speech and social media: a systematic review. *Turkish Online J Qualitative Inquiry*, 12 (8)
- Tekiroğlu SS, Chung Y-L, Guerini M (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. <https://doi.org/10.18653/v1/2020.acl-main.110>. URL <https://www.aclweb.org/anthology/2020.acl-main.110>
- Thottingal S (2019) Finite state transducer based morphology analysis for Malayalam language. In *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*, pages 1–5, Dublin, Ireland. . European Association for Machine Translation. URL <https://www.aclweb.org/anthology/W19-6801>
- Vallecillo Rodríguez MEV, Romero MVC, De Castro IC, Ráez AM, Valdivia MTM (2024) CONAN-MT-SP: A Spanish corpus for counternarrative using GPT models. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3677–3688, Torino, Italia. . ELRA and ICCL. URL <https://aclanthology.org/2024.lrec-main.326/>
- Wilk B, Shomee HH, Maity SK, Medya S (2025) Fact-based counter narrative generation to combat hate speech. In *Proceedings of the ACM on Web Conference 2025*, pages 3354–3365
- Williams M (2019) Hatred behind the screens: A report on the rise of online hate speech
- Yadav A (2018) Counterspeech: an alternative policy to combat hate speech in India. *India J Law Human Behaviour* 4(2):169–78
- Zhou S, Guan Q, Yang H, Cao Y (2024) Navigating the social media landscape: unraveling the intricacies of safety perceptions. *Hum Soc Sci Commun* 11(1):1–10

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.