JusticeBots@LT-EDI 2026: Prompt-Based Counter-Narrative Generation for Homophobia and Transphobia Comments

TT Pranesh, KK Thamizhmathi, S Vigneshwaran, B Bharathi
Department of Computer Science and Engineering
Sri Sivasubramania Nadar College of Engineering
pranesh2370060@ssn.edu.in
thamizhmathi2370055@ssn.edu.in
vigneshwaran2370061@ssn.edu.in
bharathib@ssn.edu.in

Abstract

Online platforms increasingly host hate speech targeting marginalized communities, including homophobic and transphobic comments directed at LGBTQ+ individuals. Counter-narratives provide a constructive way to respond to harmful speech by promoting empathy, factual clarification, and respectful dialogue.

In this work, we participate in the Shared Task on Counter-Narrative Generation on Homophobic and Transphobic Comments at LT-EDI @ ACL 2026. We adopt a zero-shot prompting approach using large language models accessed through publicly available AI tools, including GPT-4o, Gemini 1.5 Pro, and Llama-3 Sonar Large via Perplexity AI. Instead of training a task-specific model, we design a structured prompt that guides the models to generate respectful, concise, and contextually appropriate counter-narratives.

Experiments were conducted on English and Tamil comments provided by the organizers. Results demonstrate that prompt-based generation can produce meaningful multilingual counter-narratives without additional training. Our approach highlights the potential of large language models as lightweight tools for counter-speech generation in multilingual online environments.

1 Introduction

Social media platforms have become important spaces for communication and public discussion. However, they are also frequently used to spread hate speech and discriminatory content targeting marginalized communities, particularly LGBTQ+ individuals. Homophobic and transphobic comments often contain prejudice, hostility, and misinformation, negatively affecting online safety and inclusion. Such harmful online interactions can contribute to emotional distress, social exclusion, and the normalization of discriminatory attitudes within digital communities. Addressing these challenges has therefore become an important research problem in natural language processing and online content moderation.

Counter-narratives are constructive responses designed to challenge hateful statements while promoting empathy, factual understanding, and respectful dialogue. Unlike punitive moderation approaches, counter-speech attempts to encourage positive engagement and reduce hostility without suppressing conversation entirely. Effective counter-speech can help reduce the spread and impact of harmful online discourse while promoting healthier and more inclusive online interactions.

The Shared Task on Counter-Narrative Generation on Homophobic and Transphobic Comments at LT-EDI @ ACL 2026 focuses on generating constructive responses to harmful comments in English and Tamil (Kumaresan et al., 2026). The multilingual nature of the shared task highlights the importance of developing systems capable of handling diverse linguistic and cultural contexts.

Recent advances in large language models (LLMs) enable powerful text generation through prompt-based interaction without task-specific training. Large language models have demonstrated strong multilingual capabilities and have increasingly been used for various text generation tasks including summarization, dialogue generation, and content moderation assistance. In this work, we explore a zero-shot prompting framework for multilingual counter-narrative generation using multiple LLM APIs.

Our contributions are summarized as follows:

- We propose a zero-shot prompt-based framework for counter-narrative generation.
- We evaluate multiple LLM APIs including GPT-4o, Gemini 1.5 Pro, and Llama-3 Sonar Large.

Proceedings of the Sixth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 188–192

- We demonstrate multilingual counter-narrative generation for English and Tamil without fine-tuning.

2 Task Description

The shared task aims to develop systems that generate constructive responses to homophobic and transphobic comments collected from social media platforms.

The task contains two subtasks: span detection and counter-narrative generation. In this work, we focus on Task 2, where systems generate respectful responses that encourage empathy and avoid offensive language.

The dataset contains English and Tamil comments annotated for homophobia and transphobia. Table 1 presents the dataset distribution.

Table 1: Dataset distribution

Language	Split	Homophobia	Transphobia
Tamil	Train	342	458
Tamil	Test	73	36
English	Train	1044	756
English	Test	49	17

3 Related Work

The problem of addressing hate speech in online environments has received significant attention in recent years. Research in this area has focused on both detecting harmful content and generating constructive responses to counter such speech. Counter-narratives, also referred to as counter-speech, aim to challenge hateful or discriminatory statements by promoting empathy, factual clarification, and respectful dialogue. These approaches are increasingly viewed as constructive alternatives to content removal strategies, particularly in multilingual online communities.

Early studies demonstrated the effectiveness of counter-speech in mitigating harmful online discourse on social media platforms (Schieb and Preuss, 2016). Their work highlighted how constructive responses can reduce the spread and impact of hateful content while encouraging healthier online interactions. Several works have also focused on detecting homophobic and transphobic language in social media. Detection of such harmful content in YouTube comments was explored by (Chakravarthi, 2024), demonstrating the importance of automated systems for identifying hate speech targeting LGBTQ+ communities. Similarly, span-level identification of homophobic and transphobic content in multilingual low-resource settings was studied by (Kumaresan et al., 2025), highlighting the challenges associated with fine-grained hate speech detection across languages.

To support automated counter-narrative generation, multiple datasets and benchmarks have been introduced. The multilingual CONAN dataset proposed by (Chung et al., 2019) became an important resource for training and evaluating counter-speech generation systems. The dataset was created through a nichesourcing approach in which experts and volunteers generated responses to hateful statements targeting different communities. Neural approaches for automatic counter-narrative generation and analysis of counter-speech strategies were further explored by (Tekiroğlu et al., 2020), who demonstrated that machine learning models can generate constructive responses while maintaining respectful language. Benchmark datasets for intervention generation in online hate speech conversations were introduced by (Qian et al., 2019), enabling the development of systems that can generate interventions aimed at reducing hostility. Similarly, the effectiveness of different counter-speech strategies in combating online hate speech was analyzed by (Mathew et al., 2019), showing that constructive interventions can positively influence online discussions.

More recent works have focused on multilingual and human-centered approaches. A human-in-the-loop framework for collecting counter-narratives targeting multiple forms of hate speech was proposed by (Fanton et al., 2021). Their work emphasized the importance of human expertise in generating culturally sensitive and contextually appropriate counter-speech responses. Counter-speech generation for homophobic and transphobic social media content in Malayalam was studied by (Prasannan et al., 2025), demonstrating the feasibility of developing systems for low-resource languages and highlighting the growing interest in multilingual counter-narrative generation.

In addition to counter-speech generation, several studies have investigated abusive language and hate speech detection. Implicit and explicit abusive language in online communication was analyzed by (Caselli et al., 2020), providing insights into the complexity of identifying harmful language patterns. Dynamically generated datasets for improving hate speech detection models were introduced by (Vidgen et al., 2021), helping improve the robustness and generalization capabilities of hate

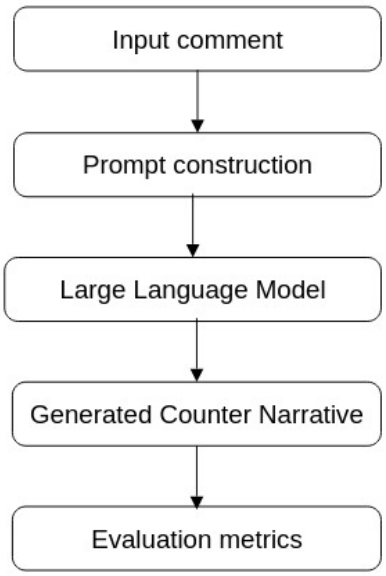

Figure 1: Prompt-based counter-narrative generation pipeline

speech detection systems.

While many existing approaches rely on supervised learning with labeled datasets, recent advances in large language models have enabled prompt-based generation approaches without task-specific fine-tuning. In this work, we explore a zero-shot prompting strategy for generating respectful and constructive counter-narratives for homophobic and transphobic comments in both English and Tamil by leveraging the multilingual capabilities of modern large language models.

4 Methodology

Our approach uses zero-shot prompting with large language models. Instead of training task-specific models, we employ prompt engineering to guide the models in generating respectful and constructive responses.

Figure 1 illustrates the workflow of our system. A social media comment is provided as input, after which a structured prompt is constructed. The prompt is processed by a large language model to generate a counter-narrative response.

4.1 Prompt Design

We designed a structured prompt that instructs the model to:

- Maintain a respectful and constructive tone
- Avoid repeating hateful language
- Encourage empathy and coexistence
- Correct misinformation when necessary
- Generate concise responses
- Produce output in the same language as the input

Task: Generate a counter-narrative response to a homophobic or transphobic comment.
Instructions:

- The response must be calm and respectful.
- Avoid slurs and hateful phrases.
- Encourage empathy and inclusiveness.
- Keep the response concise.
- Write in the same language as the comment.

Comment: {PASTE_COMMENT_HERE}
Counter-narrative:

4.2 Models and Settings

We used the following publicly available LLM APIs:

- GPT-4o (OpenAI)
- Gemini 1.5 Pro
- Llama-3 Sonar Large via Perplexity AI

Responses were generated using the default decoding and generation settings of the respective platforms. The same prompt template was used across all experiments to maintain consistency.

5 Results and Discussion

Table 2 presents evaluation results for different LLMs. Evaluation includes reference-based metrics such as Distinct-2 (D-2) and BERTScore-F1 together with rubric-based metrics including Politeness and Respectfulness Score (PRS), Quality Score (QS), and Contextual Counter-Narrative Coherence (CCNC).

Rubric-based metrics including PRS, QS, and CCNC were computed using GPT-4o as an automated evaluator following the shared task evaluation protocol.

Table 2: Evaluation results for counter-narrative generation

System / Model	Language	F1	D-2	PRS	QS	CCNC
JusticeBots (Official Submission)	Tamil	85.67	27.01	87.16	66.97	73.39
JusticeBots (Official Submission)	English	87.63	79.11	76.52	52.27	57.58
Llama-3 Sonar Large	Tamil	94.93	63.13	96.71	91.84	92.96
Llama-3 Sonar Large	English	86.93	89.57	95.34	90.62	91.48
Gemini 1.5 Pro	Tamil	95.15	91.51	97.08	92.76	93.65
Gemini 1.5 Pro	English	87.27	68.15	96.12	91.34	92.48

The results indicate strong semantic similarity with reference counter-narratives while maintaining politeness and contextual coherence. The lower Distinct-2 score observed for Tamil responses generated using GPT-4o may be due to repetitive lexical patterns and shorter generated outputs. Gemini generated more lexically diverse Tamil responses, resulting in substantially higher diversity scores.

To further examine response quality, we analyzed qualitative examples generated by the Gemini 1.5 Pro model. The generated counter-narratives generally maintained respectful language, encouraged empathy, and avoided aggressive responses even when the input comments were highly offensive. Tamil examples were originally generated in Tamil script and are presented here in transliterated form for ACL-compatible formatting.

For example, for the English comment: "LGBTQ people are ruining society.", the model generated the response: "Every individual deserves dignity and respect regardless of their identity. Diversity does not harm society; discrimination and hatred do. Promoting empathy and understanding helps create a safer community for everyone."

Similarly, for the Tamil comment: "Inime aangal paaliyal balaadhkaaram matrum vankodumai nadakkum", the generated response was: "Paalina adaiyaalam matrum eerppu enbathu ovvoru manithanin thanippatta urimai. Vanmurai enbathu sattappadi kutram, aanaal oruvarin adaiyaalathirkaaga avargalai kaayappaduthuvathu aarokkiyamaana samudhaayathirkku azhagalla."

Another English example includes the comment: "Whoever and whatever against the nature is not acceptable", for which the model generated the response: "Diversity in orientation and identity is recognized as a natural part of the human experience. Fostering empathy and understanding helps us live together more peacefully and respectfully."

Likewise, for the Tamil comment: "Manitha uravugalukkaana punitham ariyaadhavar neengal...", the generated response was: "Manitha uravugal anbin adipadaiyil amaindhavai. Oruvarin thanippatta adaiyaalathai madhippathan moolam naam samudhaayathil pilavugalai kuraiththu, anaivarum samamaaga vaazhum soozhalai uruvaakka mudiyum."

These examples demonstrate that the model is capable of generating constructive multilingual counter-narratives that promote inclusiveness and respectful dialogue.

Overall, the results demonstrate that zero-shot prompting with LLMs can generate respectful and constructive multilingual counter-narratives without additional training.

6 Limitations

Although zero-shot prompting produces constructive responses, some generated outputs remain generic and may lack cultural specificity. Since the approach relies on large language models, responses may occasionally contain factual inaccuracies or overly cautious language. In addition, rubric-based evaluation using LLM judges may introduce evaluation bias. Future work should incorporate human evaluation and culturally grounded assessment strategies.

7 Conclusion

This paper presented a zero-shot prompt-based framework for multilingual counter-narrative generation targeting homophobic and transphobic comments. Using GPT-4o, Gemini 1.5 Pro, and Llama-3 Sonar Large, our system generated respectful responses in both English and Tamil without fine-tuning.

Experimental results demonstrate that prompt-based generation can produce meaningful counter-speech in multilingual settings. Future work may explore few-shot prompting strategies, human evaluation, and improved contextual grounding methods.

8 Ethical Considerations

Counter-narrative generation systems should be used responsibly to promote respectful online interactions. Although large language models can generate constructive responses to harmful content, the outputs may occasionally contain inaccuracies or culturally insensitive responses. Therefore, such systems should be treated as assistive tools rather than fully autonomous moderation systems.

Acknowledgment of Generative AI Usage

Generative AI tools including GPT-4o, Gemini 1.5 Pro, and Llama-3 Sonar Large were used for counter-narrative generation and language refinement during this study. All experiments, evaluations, analysis, and manuscript preparation were conducted and verified by the authors.

References

Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. 2020. I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In *Proceedings of the twelfth language resources and evaluation conference*, pages 6193–6202.

Bharathi Raja Chakravarthi. 2024. Detection of homophobia and transphobia in youtube comments. *International Journal of Data Science and Analytics*, 18(1):49–68.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroğlu, and Marco Guerini. 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.

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. In *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.

Prasanna Kumar Kumaresan, Devendra Deepak Kayande, Ruba Priyadharshini, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2025. Homophobia and transphobia span identification in low-resource languages. *Natural Language Processing Journal*, page 100169.

Prasanna Kumar Kumaresan, Praveen Prasannan, Tanay Singh, Ruba Priyadharshini, Subalalitha Chinnaudayar Navaneethakrishnan, Saranya Rajiakodi, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2026. Findings of the Shared Task on Counter-Narrative Generation on Homophobic and Transphobic Comments. In *Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2019. Thou shalt not hate: Countering online hate speech. In *Proceedings of the international AAAI conference on web and social media*, volume 13, pages 369–380.

Praveen Prasannan, Prasanna Kumar Kumaresan, Saranya Rajiakodi, CN Subalalitha, and Bharathi Raja Chakravarthi. 2025. Counter-speech generation for homophobic and transphobic social media content in malayalam. *Social Network Analysis and Mining*, 15(1):87.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4755–4764.

Carla Schieb and Mike Preuss. 2016. Governing hate speech by means of counterspeech on facebook. In *66th ica annual conference, at fukuoka, japan*, pages 1–23.

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

Bertie Vidgen, Tristan Thrush, Zeerak Talat, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *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 1667–1682.