NEUNI@LT-EDI 2026: Counter Narrative Generation on Homophobic and Transphobic Comments

Preethi Gajawada[1], Bhanu Harsha Yanamadala[1], Akankshya Kar[2,‡], Sahil Wadhwa[3,†], Divya Chaudhary[1]
[1]Northeastern University, [2]Apple Inc., [3]Capital One
Correspondence: d.chaudhary@northeastern.edu

Abstract

Counter Narrative (CN) generation via Large Language Models (LLMs) offers a scalable approach to combating hate speech by producing targeted responses that challenge harmful content. However, existing methods typically require costly post-training or fine-tuning to improve narrative diversity and quality. We introduce a fine-tuning-free prompt optimization technique that enhances Counter Narrative effectiveness without additional model training, making it both resource-efficient and readily deployable. We conduct extensive evaluation on hate speech datasets spanning English and Tamil, employing both reference-based metrics and rubric-based LLM-as-a-judge scoring to capture multiple dimensions of narrative quality. Experiments across multiple LLMs demonstrate that our approach consistently outperforms vanilla prompting baselines, exhibits strong transferability across models, and adapts seamlessly to new evaluation metrics—requiring no architectural or procedural changes. Our findings suggest that carefully optimized prompting strategies can match or exceed the performance of more resource-intensive approaches, offering a practical path toward scalable hate speech intervention.

Content Warning: This paper contains content that could be distressing to certain readers.

1 Introduction

The proliferation of LLMs in market has posed a threat to an increase in fake news and hate speeches on social media platforms (Papageorgiou et al., 2024). These hate speeches target individuals, communities, races, religions etc. which makes the process of Counter Narrative generation an indispensable task to ensure equality, diversity and inclusion (LDI) in our society. Researchers have experimented with a variety of ways to make CN balanced, impactful, fact-based while de-escalating hostility and encouraging healthier dialog, directly addressing the biases or misconceptions fueling hate speech.

Large Language Models (LLMs) are trained on massive and diverse corpora, leveraging expansive contextual windows and transformer-based architectures to capture complex linguistic and semantic patterns. This scale of training enables them to generalize across a wide range of tasks, including text generation, reasoning, summarization, and decision support. Recent advancements in their reasoning and instruction-following capabilities (OpenAI et al., 2024; Grattafiori et al., 2024) have significantly improved their reliability, coherence, and performance on multi-step and domain-specific tasks. As a result, LLMs are increasingly being deployed in high-stakes and safety-critical domains such as medicine (Maity and Saikia, 2025; Zhang et al., 2023), where they assist in clinical decision support and medical documentation; law (Chu et al., 2025), where they aid in legal analysis and case understanding; and AI security (Purpura et al., 2025; Rad et al., 2025), where they contribute to vulnerability assessment, policy enforcement, and guardrail design.

In parallel, counter-narrative generation has emerged as an active and socially impactful research direction. By producing constructive, evidence-based responses to harmful or misleading content, LLMs have demonstrated potential in mitigating hate speech, misinformation, and extremist rhetoric (Wadhwa et al., 2025; Wilk et al., 2025). This growing body of work underscores both the transformative promise of LLMs and the importance of robust evaluation, alignment, and safety safeguards in their deployment. Despite these advancements, LLMs continue to exhibit inherent limitations, including biases and hallucinations (Li et al., 2024; Yao et al., 2024; Wadhwa et al., 2026). To address these challenges, prior work has ex-

[†]This work does not relate to the position at Capital One.
[‡]This work does not relate to the position at Apple.

Proceedings of the Sixth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 206–211

plored model fine-tuning strategies (Wadhwa et al., 2025; Furman et al., 2023) and external knowledge grounding techniques (Chung et al., 2021). While effective to some extent, such approaches often incur substantial computational and operational costs, thereby limiting scalability and widespread deployment of LLM-based solutions.

In this paper, we propose a fine-tuning-free approach for counter-narrative (CN) generation targeting homophobic and transphobic comments. We introduce an automatic prompt optimization framework that begins with a seed prompt for CN generation and iteratively refines it using multiple reward functions, including BERTScore, n-gram overlap, politeness, coherence, and overall quality metrics (Kumaresan et al., 2026). Our method is model-agnostic and requires only a task description (CN generation in our case) along with a set of reward functions to guide optimization. We demonstrate that our approach consistently outperforms static prompting strategies across both English and Tamil datasets. We evaluate our framework using GPT-4o-mini pinning the efficacy of our approach. Our code is publicly available at `https://github.com/wadhwasahil/cn-shared-task-acl`.

2 Task and Dataset Descriptions

The shared task[*] LT-EDI@ACL 2026 (Kumaresan et al., 2026; Chakravarthi, 2024; Prasannan et al., 2025; Kumaresan et al., 2025) focused on generating appropriate counter-narratives in response to comments containing homophobia or transphobia. The task provided a dataset containing comments identified as homophobic or transphobic with an aim to generate Counter Narratives that respond constructively to the hateful content. Generated CNs were judged using a combination of automatic (reference-based) metrics and rubric-based evaluation scores. Evaluation was conducted using both reference-based and rubric-based metrics. For reference-based evaluation, the organizer employed BERTScore as the primary metric to measure the semantic similarity between generated counter-narratives and gold references, along with Distinct-2 to assess response diversity. In addition, rubric-based evaluation was performed using an LLM judge that scored each response on a 0–2 scale across three dimensions: Politeness and Respectfulness (PRS), Contextual Counter-Narrative Coherence (CCNC), and overall Quality (QS), assessing tone, contextual relevance, coherence, grammatical correctness, and richness. For final ranking in Task 2, BERTScore, Distinct-2, PRS, CCNC, and QS were each converted to percentage values, and the overall score was computed as the average of these five metrics. Teams were ranked based on this combined average score.

The curated dataset for the task consisted of comments scraped from online platform including those from platforms like YouTube, annotated for the presence of homophobic, transphobic, or non-anti-LGBTQ+ content. Dataset statistics are shown in Table 1.

Language	Split	Homophobia	Transphobia	Total
Tamil	Train	342	458	800
Tamil	Test	73	36	109
English	Train	1,044	756	1,800
English	Test	49	17	66

Table 1: Dataset statistics for the shared task LT-EDI@ACL 2026.

3 Proposed Methodology

In this section, we provide a detailed description of our framework. We adopt a fine-tuning-free prompt optimization approach using DSPy (Khattab et al., 2023), which enables systematic refinement of prompts without modifying model parameters. Our framework begins with a seed prompt designed for counter-narrative generation and iteratively refines it through multiple optimization rounds. Unlike conventional fine-tuning methods that require gradient updates and task-specific training data, our approach operates entirely at the prompt level, making it computationally efficient and readily deployable across different LLMs.

The optimization process directly incorporates the task's evaluation metrics as reward signals. We utilize both reference-based metrics—BERTScore for measuring semantic similarity to gold references and Distinct-2 for assessing lexical diversity—and rubric-based LLM-as-a-judge scoring across three dimensions: Politeness and Respectfulness (PRS), Contextual Counter-Narrative Coherence (CCNC), and overall Quality (QS), which evaluates tone, coherence, grammatical correctness, and richness. By aligning our optimization objective with the final evaluation criteria, we ensure that prompt refinements translate directly to improved task performance. Furthermore, since the optimized prompts require no architectural changes,

[*]https://sites.google.com/view/lt-edi-2026/shared-tasks

You are generating a COUNTER-NARRATIVE to harmful or hateful speech. ***Rules (STRICT):*** *Respond in exactly 2–3 sentences. Be polite, calm, and respectful. Directly address the claim in the hate speech. Do NOT lecture. Do NOT use long explanations. Do NOT include disclaimers. Do NOT repeat the hate speech. Avoid moralizing phrases like "it's important to remember" or "we should all".* ***Goal:*** *Challenge the harmful idea constructively. Promote understanding and empathy. Encourage respectful reflection.*

Figure 1: Baseline prompt for counter-narrative generation.

Your task is to craft an articulate and thoughtful counter-narrative to refute a specific hateful claim about the LGBTQ+ community. Emphasize clarity and structure to address, correct, and conclude with persuasiveness yet in 2–3 comprehensive sentences. Ensure the tone remains composed and friendly, avoiding clichés, and instead harness your argument with rationality without falling into a preachy style, thereby guaranteeing cognitive diversity through skillfully unique and diverse constructive points.

Figure 2: Optimized prompt obtained via COPRO optimization.

they transfer seamlessly to new LLMs and adapt to alternative metrics without modification.

Problem Formulation. Given a hate speech instance h, our goal is to generate a counter-narrative c using a language model $\mathcal{M}$ conditioned on a prompt P. We formulate prompt optimization as:

$$P^* = \arg\max_P E_{h\sim\mathcal{D}}\left[\mathcal{R}(c, c^*)\right], \quad c = \mathcal{M}(P, h) \tag{1}$$

where c^* denotes the gold reference and $\mathcal{R}$ is a composite reward function defined as:

$$\mathcal{R} = \frac{1}{5}\big(\text{BERTScore} + \text{Distinct-2} + \text{PRS} + \text{CCNC} + \text{QS}\big) \tag{2}$$

Here, BERTScore and Distinct-2 are reference-based metrics, while PRS (Politeness and Respectfulness), CCNC (Contextual Counter-Narrative Coherence), and QS (Quality Score) are rubric-based scores obtained via an LLM judge $\mathcal{J}$ (Rad et al., 2025; Zheng et al., 2023).

4 Experiments

Effective counter-narrative generation requires prompts that elicit responses that are simultaneously persuasive, contextually relevant, and linguistically diverse. Manually crafting such prompts is labor-intensive and often suboptimal. To address this, we leverage DSPy[†], an open-source framework for programmatic prompt optimization that enables systematic refinement of prompts through automated search. DSPy treats prompts as modular, optimizable programs, allowing iterative improvement based on task-specific reward signals. Unlike fine-tuning approaches that modify model parameters—requiring gradient computation, large-scale training data, and significant computational overhead—our method operates entirely at the prompt level. This makes our approach both computationally efficient and model-agnostic: optimized prompts transfer directly to any target language model without architectural changes or retraining.

Generation Configuration. We use *GPT-4o-mini* as the base language model for counter-narrative generation. The temperature parameter is set to 0.7 for English and 0.9 for Tamil to balance fluency and diversity across languages. Notably, for Tamil counter-narrative generation, we adopt a direct generation approach without any intermediate language conversion—both the hate speech input and the generated counter-narrative remain entirely in Tamil, preserving linguistic authenticity and avoiding potential translation artifacts.

Optimization Strategy. For prompt optimization, we employ Cooperative Prompt Optimization (COPRO)[‡], which refines prompts through systematic exploration of the configuration space. COPRO evaluates candidate prompts against our composite reward function $\mathcal{R}$, comprising both reference-based metrics (BERTScore, Distinct-2) and rubric-based LLM-as-a-judge scores (PRS, CCNC, QS). The optimizer iteratively improves prompts via cooperative search, selecting configurations that maximize performance across all metrics. *GPT-4o-mini* serves as the LLM judge for rubric-based evaluation.

Results. Table 2 presents the shared task leaderboard, demonstrating that our fine-tuning-free approach achieves competitive performance, ranking 3rd in both English and Tamil tracks. Despite relying solely on prompt optimization without any model parameter updates, our method outperforms several fine-tuned and resource-intensive baselines.

[†] https://github.com/stanfordnlp/dspy
[‡] https://dspy.ai/api/optimizers/COPRO

Team	Run	Reference-Based		Rubric-Based			Overall
		Distinct-2	BERTScore	PRS	QS	CCNC	Avg. (%)
			English				
Team_V	Run 1	73.56	88.78	90.91	90.15	93.94	87.47
SigJBS	Run 1	69.32	86.66	93.18	90.91	91.67	86.35
NEUNI (Ours)	**Run 1**	**64.50**	**86.29**	**91.67**	**86.36**	**86.36**	**83.04**
DLRG	Run 2	74.36	85.55	72.73	69.70	84.09	77.29
Amritha	Run 3	68.16	86.02	100.00	68.18	61.36	76.74
JusticeBots	Run 1	79.11	87.63	76.52	52.27	57.58	70.62
RespectNLP	Run 1	78.56	82.93	53.79	54.55	81.82	70.33
DuoNova	Run 1	58.22	86.04	56.82	37.88	50.00	57.79
			Tamil				
DLRG	Run 3	27.30	85.73	100.00	97.71	91.28	80.40
Amritha	Run 3	20.89	85.27	100.00	100.00	89.45	79.12
NEUNI (Ours)	**Run 2**	**19.16**	**85.09**	**95.41**	**86.24**	**92.66**	**75.71**
JusticeBots	Run 1	27.01	85.67	87.16	66.97	73.39	68.04
TeamV	Run 1	25.61	86.25	87.61	55.50	66.51	64.30
SigJBS	Run 1	25.29	85.29	75.23	72.02	61.01	63.77
DuoNova	Run 1	3.62	86.04	94.50	61.93	64.68	62.15
RespectNLP	Run 1	17.43	80.23	50.92	11.47	7.80	33.57

Table 2: Shared task leaderboard for Counter Narrative Generation (%). Our submission (NEUNI) ranks 3rd in both language tracks. Best results per metric are **bolded**.

Model	Metric	English		Tamil	
		Base	Opt	Base	Opt
GPT-4o-mini	PRS	84.40	91.20	73.47	73.09
	CCNC	48.80	77.40	58.94	90.91
	QS	86.30	88.90	87.27	87.30
	Combined	73.10	86.00	73.22	83.70
	Distinct-2	60.90	98.47	88.28	99.88
	BERTScore	86.32	87.06	94.22	94.22

Table 3: Performance comparison of GPT-4o-mini (COPRO) in (%). *Opt* refers to results on optimized prompt and *base* is the baseline w/o optimization. Best results per language are underlined.

Table 3 provides a detailed comparison between baseline prompts (*Base* as shown in Figure 1) and optimized prompts (*Opt* as shown in Figure 2). The optimized prompts yield consistent improvements across the majority of metrics, with notable gains in CCNC and Distinct-2, indicating enhanced contextual coherence and lexical diversity. These results validate the effectiveness of prompt optimization as a lightweight, practical alternative to model fine-tuning for counter-narrative generation.

5 Conclusion and Future Work

We presented a fine-tuning-free prompt optimization approach for counter-narrative generation against hate speech. By leveraging DSPy's COPRO optimizer, we systematically refined prompts using a composite reward function that combines reference-based metrics (BERTScore, Distinct-2) and rubric-based LLM-as-a-judge scoring (PRS, CCNC, QS). Our approach achieved 3rd place in both English and Tamil tracks of the shared task, demonstrating that carefully optimized prompts can yield competitive performance without the computational overhead of model fine-tuning. Notably, our method operates directly on Tamil without intermediate translation, preserving linguistic authenticity. The consistent improvements observed across diverse metrics validate prompt optimization as a practical, resource-efficient alternative for counter-narrative generation.

Several directions merit further exploration. First, we aim to evaluate our approach on additional languages and hate speech domains to assess cross-lingual and cross-domain transferability. Second, we plan to investigate advanced DSPy optimizers such as MIPROv2, which jointly optimizes instructions and few-shot demonstrations via Bayesian optimization, potentially yielding further performance gains. Third, extending the reward function to incorporate human preference alignment or retrieval-augmented generation could enhance the persuasiveness and factual grounding of generated counter-narratives. Finally, we intend to explore the deployment of optimized prompts across diverse LLM backends to further validate the model-agnostic nature of our approach.

6 Ethical Considerations

Our system is intended solely for constructive counter-speech and should not be misused to generate harmful content. The hate speech examples in our datasets are used strictly for research pur-

poses; we do not endorse the views they express. We recommend human oversight when deploying automated counter-narrative systems.

7 Limitations

Our work has avenues for future improvement. First, the current approach operates purely at the prompt level without incorporating external knowledge bases or retrieval mechanisms. Integrating knowledge grounding could enable the generation of evidence-based counter-narratives that cite verifiable facts, potentially enhancing persuasiveness and credibility. Second, while we utilize reference-based and rubric-based metrics as reward signals for prompt optimization, these same metrics could be employed for reinforcement learning-based model fine-tuning, which may yield complementary or superior performance. Finally, human evaluation of counter-narrative quality and persuasiveness remains an important direction that we leave for future work.

References

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

Yu Ying Chu, Sieh-chuen Huang, and Hsuan-Lei Shao. 2025. Unpacking legal reasoning in LLMs: Chain-of-thought as a key to human-machine alignment in essay-based NLU tasks. In *Proceedings of the 5th Workshop on Natural Logic Meets Machine Learning (NALOMA)*, pages 1–7, Bochum, Germany. Association for Computational Linguistics.

Yi-Ling Chung, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Towards knowledge-grounded counter narrative generation for hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 899–914, Online. Association for Computational Linguistics.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 2 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2023. Dspy: Compiling declarative language model calls into self-improving pipelines. *Preprint*, arXiv:2310.03714.

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.

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. The dawn after the dark: An empirical study on factuality hallucination in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10879–10899, Bangkok, Thailand. Association for Computational Linguistics.

Subhankar Maity and Manob Jyoti Saikia. 2025. Large language models in healthcare and medical applications: A review. *Bioengineering*, 12(6).

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, and Shyamal Anadkat. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Eleftheria Papageorgiou, Christos Chronis, Iraklis Varlamis, and Yassine Himeur. 2024. A survey on the use of large language models (llms) in fake news. *Future Internet*, 16(8).

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.

Alberto Purpura, Sahil Wadhwa, Jesse Zymet, Akshay Gupta, Andy Luo, Melissa Kazemi Rad, Swapnil Shinde, and Mohammad Shahed Sorower. 2025. Building safe GenAI applications: An end-to-end overview of red teaming for large language models. In *Proceedings of the 5th Workshop on Trustworthy*

NLP (TrustNLP 2025), pages 335–350, Albuquerque, New Mexico. Association for Computational Linguistics.

Melissa Kazemi Rad, Huy Nghiem, Andy Luo, Sahil Wadhwa, Mohammad Sorower, and Stephen Rawls. 2025. Refining input guardrails: Enhancing llm-as-a-judge efficiency through chain-of-thought fine-tuning and alignment. *Preprint*, arXiv:2501.13080.

Sahil Wadhwa, Himanshu Kumar, Guanqun Yang, Abbaas Alif Mohamed Nishar, Pranab Mohanty, Swapnil Shinde, and Yue Wu. 2026. Art: Adaptive reasoning trees for explainable claim verification. *Preprint*, arXiv:2601.05455.

Sahil Wadhwa, Chengtian Xu, Haoming Chen, Aakash Mahalingam, Akankshya Kar, and Divya Chaudhary. 2025. Northeastern uni at multilingual counterspeech generation: Enhancing counter speech generation with LLM alignment through direct preference optimization. In *Proceedings of the First Workshop on Multilingual Counterspeech Generation*, pages 19–28, Abu Dhabi, UAE. Association for Computational Linguistics.

Brian Wilk, Homaira Huda Shomee, Suman Kalyan Maity, and Sourav Medya. 2025. Fact-based counter narrative generation to combat hate speech. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 3354–3365, New York, NY, USA. Association for Computing Machinery.

Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, Yu-Yang Liu, and Li Yuan. 2024. Llm lies: Hallucinations are not bugs, but features as adversarial examples. *Preprint*, arXiv:2310.01469.

Haodi Zhang, Jiahong Li, Yichi Wang, and Yuanfeng Song. 2023. Integrating Automated Knowledge Extraction with Large Language Models for Explainable Medical Decision-Making . In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1710–1717, Los Alamitos, CA, USA. IEEE Computer Society.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.