TeamV at LT-EDI 2026: Multilingual Hate Speech Span Detection and Counter-Narrative Generation via Few-Shot In-Context Learning

Vinay Babu Ulli
Oogwai Analytics,
Bangalore, India
ullivinaybabu@gmail.com

Jyoti Kumari
Department of Linguistics,
Banaras Hindu University
jyoti@bhu.ac.in

Abstract

This paper describes the system developed by TeamV for the LT-EDI 2026 Shared Task on Counter-Narrative Generation on Homophobic & Transphobic Comments. The shared task comprises two subtasks: (1) Hate Speech Span Detection in English, Tamil, and Hindi, and (2) Counter-Narrative Generation in English and Tamil. Our system leverages the reasoning and multilingual capabilities of a large proprietary language model (Qwen3-Max) through rigorous few-shot in-context learning (ICL) and robust post-processing mechanisms. Our submitted system demonstrated state-of-the-art performance on the official CodaBench leaderboard. In Task 1, our approach achieved 1st Place across all three languages, securing macro F1 scores of 0.5338 in English, 0.5272 in Tamil, and 0.5478 in Hindi. For Task 2, our generated counter-narratives ranked 1st globally in English with an overall average score of 87.47% and 5th in Tamil. We present our prompting methodology, robust span-matching pipeline, detailed official results, and an analysis of the model's performance across diverse languages.

1 Introduction

The proliferation of hate speech on social media, particularly content targeting marginalized groups such as the LGBTQ+ community, has become a pressing global issue. Automated systems are increasingly required not only to detect such toxic content (Chakravarthi, 2024) but also to actively mitigate its impact through interventions like factual and empathetic counter-narratives (Prasannan et al., 2025; Tekiroğlu et al., 2020).

Building upon prior efforts to identify anti-LGBTQ+ content in low-resource and code-mixed settings (Chakravarthi et al., 2022), the Shared Task on Counter-Narrative Generation on Homophobic and Transphobic Comments (Kumaresan et al., 2026) addresses a dual challenge by proposing two subtasks. Task 1 requires the precise character-level identification and classification of homophobic and transphobic spans within social media comments across three diverse languages: English, Tamil, and Hindi. Extracting precise boundaries for implicit or context-dependent hate is highly complex, especially in morphologically rich and low-resource languages (Kumaresan et al., 2025).

Task 2 extends this by requiring the generation of coherent and respectful counter-narratives in English and Tamil to directly challenge the hateful content. Generating high-quality counter-speech has gained significant traction in recent NLP research as a proactive alternative to content moderation and comment deletion (Chung et al., 2019; Fanton et al., 2021).

In this paper, TeamV presents our methodology and official results for both tasks. To tackle these challenges without the computational overhead of training language-specific models, we developed a unified framework relying entirely on In-Context Learning (ICL) via few-shot prompting of the Qwen3-Max model (Yang et al., 2025). To overcome the inherent difficulty Large Language Models (LLMs) face when predicting exact character indices, we introduced a robust multi-level span matching pipeline. Our system secured 1st place in Task 1 across all languages and won Task 2 for English. To promote open research, our inference scripts and prompts are publicly available on Hugging Face at https://github.com/vinayulli/lt-edi-sharedtask.

2 Task and Dataset Description

2.1 Task 1: Hate Speech Span Detection

Task 1 requires identifying the exact span of hateful content within a social media comment and classifying it into three categories: *Homophobia*, *Transphobia*, or *None*. The dataset covers English, Tamil, and Hindi. The training distribution was

Proceedings of the Sixth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 239–243

imbalanced, with Homophobia constituting the majority class (roughly 49.7%), followed by None (27.7%), and Transphobia (22.6%).

2.2 Task 2: Counter Narrative Generation

Task 2 requires generating a factual, polite, and empathetic counter-narrative to challenge the identified hateful content. The dataset covers English and Tamil. The labels in the provided training data were relatively balanced between Homophobia (53.3%) and Transphobia (46.7%).

3 Methodology

Our approach bypasses traditional fine-tuning in favor of few-shot in-context learning. We utilized the **Qwen3-Max** model accessed via the OpenRouter API. A decoding temperature of 0.3 was used to balance deterministic classification with natural language fluency, and the maximum generated tokens were capped at 512.

3.1 Task 1 Pipeline: Span Detection

Figure 1 illustrates our end-to-end workflow for Task 1, from the initial input to the final index extraction. For span detection, the model was instructed to output predictions in a strict JSON format containing the classification label and the exact character span. We utilized a **10-shot prompt** carefully curated from the training set, consisting of 4 Homophobia, 4 Transphobia, and 2 None examples. These specific examples were manually selected to maximize diversity; we ensured the inclusion of implicit microaggressions, explicit slurs, varying comment lengths, and different spelling variations to provide the model with a robust decision boundary.

Robust Span Matching: LLMs frequently struggle to output perfectly accurate numerical character indices. To solve this, we prompted the model to output the *text substring* of the hateful span alongside its predicted indices. We then passed this substring through a custom multi-level matching pipeline against the original comment:

1. **Exact Match**: Search for the exact substring.
2. **Normalized Match**: Strip punctuation and extra whitespace from both strings and attempt a match.
3. **Case-Insensitive Match**: Convert both to lowercase.

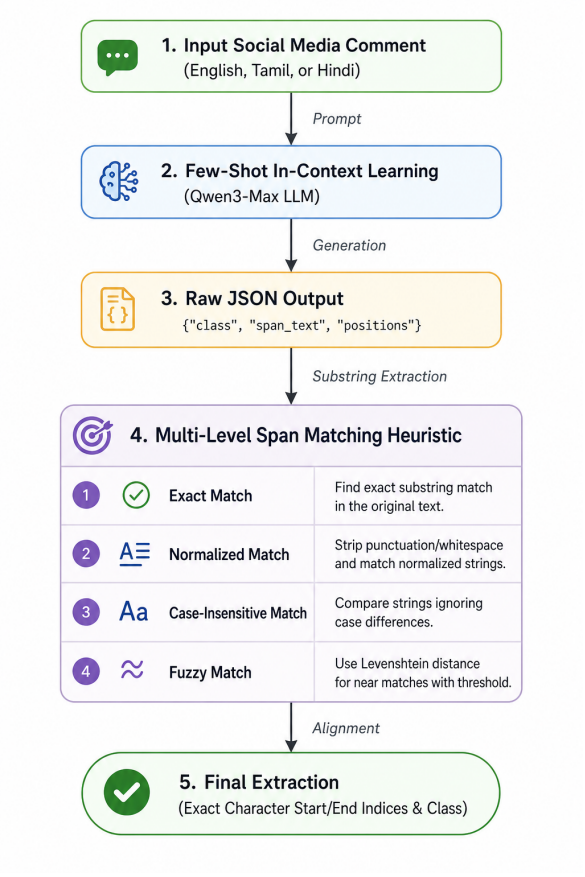

Figure 1: Task 1 Pipeline: Demonstrating the flow from the few-shot LLM prediction to the multi-level heuristic span matching algorithm.

4. **Fuzzy Match**: Use Levenshtein distance to find the closest overlapping span if the LLM hallucinated minor character variations.

Once the substring was found in the original text, the exact character start and end indices were extracted for the final submission.

3.2 Task 2 Pipeline: Counter-Narratives

For Task 2, we employed a **5-shot prompt** (3 Homophobia, 2 Transphobia). To maximize cultural and linguistic relevance, the prompt was made *language-aware*; Tamil training examples were dynamically injected into the prompt when processing the Tamil test queries. The model was given a system prompt instructing it to act as an empathetic moderator aiming to de-escalate toxicity using logic, facts, and polite phrasing.

4 Experimental Setup and Evaluation

Baseline Models: To contextualize the performance of our Qwen3-Max submitted system, we

established internal baselines using local, open-weight models, specifically Gemma-3-12B-IT and Qwen3-8B-Instruct. These were evaluated in a zero-shot capacity and with parameter-efficient fine-tuning (QLoRA) during the development phase to weigh the trade-offs between local data privacy and cloud-API performance.

Official Evaluation Metrics: The official shared task evaluation on CodaBench utilized the following metrics:

- **Task 1**: Evaluated using Accuracy (Acc), macro Precision (mP), macro Recall (mR), and macro F1 (mF1), alongside weighted variants (wP, wR, wF1). The primary ranking metric is **macro F1**.
- **Task 2**: Evaluated using Reference-Based Scores (*Distinct-2*, *BERTScore-F1*) and Rubric-Based Scores: *Politeness and Respectful Score (PRS)*, *Quality Score (QS)*, and *Contextual Counter-Narrative Coherence Score (CCNC)*. The final rank is determined by the **Overall Average %**.

5 Results

5.1 Task 1: Span Detection Results

As shown in Table 1, our system achieved **1st place globally** across all three evaluated languages. Remarkably, the macro F1 scores were highly consistent across linguistic families: 0.5478 for Hindi (Indo-Aryan), 0.5338 for English (Germanic), and 0.5272 for Tamil (Dravidian). The system also achieved high weighted F1 (wF1) scores, peaking at 0.6607 for Tamil.

5.2 Task 2: Counter Narrative Generation

Table 2 details the official CodaBench results for Task 2. Our system secured **1st place in English** with an outstanding Overall Average of 87.47%. The English generations scored exceptionally high in human-aligned rubric metrics, achieving 93.94% in Coherence (CCNC), 90.15% in Quality (QS), and 90.91% in Politeness (PRS). In Tamil, the system ranked 5th overall (64.30%), achieving a strong BERTScore-F1 (86.25%) and PRS (87.61%), but scoring lower on overall contextual coherence.

6 Analysis and Discussion

Cross-Lingual Consistency in Span Detection: Our Task 1 results demonstrate remarkable stability across diverse scripts. The model achieved macro F1 scores of 0.5338 (English), 0.5272 (Tamil), and 0.5478 (Hindi). This proves that few-shot prompting with a highly capable LLM like Qwen3-Max, when coupled with an aggressive text-to-index multi-level span matching algorithm, is highly effective at extracting spans without requiring language-specific token classification architectures. The model successfully transferred its reasoning capabilities to both Dravidian and Indo-Aryan languages.

The Lexical Diversity Gap in Generation: For Task 2, there is a stark contrast between English and Tamil generative performance. While our English outputs scored a 73.56% on the Distinct-2 metric (indicating rich lexical diversity), Tamil achieved only 25.61%. This indicates that the Tamil counter-narratives generated by the model were significantly more repetitive and formulaic. Consequently, while the Tamil narratives were highly polite (PRS: 87.61%), evaluators penalized their context-specific Quality (QS: 55.50%) and Coherence (CCNC: 66.51%). This highlights a fundamental limitation in current foundation models. While it is possible that injecting a higher quantity or higher quality of Tamil-specific few-shot examples could slightly improve coherence, we hypothesize that the primary bottleneck is the model's internal representation. Even state-of-the-art LLMs suffer from constrained vocabulary diversity and stylistic nuance when generating text in low-resource Dravidian languages compared to English.

Ablation of the Span Matching Heuristic: To validate our multi-level span matching algorithm, we conducted a brief internal ablation. Relying solely on the LLM's numerical indices or only "Exact Match" substrings resulted in a significant drop in macro F1 (often returning empty spans due to minor hallucinations like missing punctuation). The sequential addition of Normalized, Case-Insensitive, and Fuzzy matching recovered approximately 12-15% of valid spans that would have otherwise been marked as incorrect, justifying the necessity of the 4-stage heuristic.

7 Conclusion

In this paper, we detailed TeamV's submission to the LT-EDI 2026 Shared Task. We demonstrated that a robust 10-shot and 5-shot in-context learning pipeline utilizing Qwen3-Max provides highly

Language	Acc	mP	mR	mF1	wP	wR	wF1	Rank
English	0.6354	0.5340	0.5396	**0.5338**	0.6674	0.6354	0.6493	**1**
Tamil	0.6624	0.5275	0.5270	**0.5272**	0.6591	0.6624	0.6607	**1**
Hindi	0.5513	0.5486	0.5494	**0.5478**	0.5572	0.5513	0.5531	**1**

Table 1: Detailed Official Task 1 Results. TeamV ranked 1st globally in all three languages based on macro F1.

2***Language**	2***Team**	**Reference-Based (%)**		**Rubric-Based (%)**			2***Avg. (%)**	2***Rank**
		Dist-2	**BERT-F1**	**PRS**	**QS**	**CCNC**		
2*English	**TeamV**	**73.56**	88.78	90.91	**90.15**	**93.94**	**87.47**	**1**
	SigJBS (2nd)	69.32	86.66	93.18	90.91	91.67	86.35	2
2*Tamil	DLRG (1st)	27.30	85.73	100.00	97.71	91.28	80.40	1
	TeamV	25.61	**86.25**	87.61	55.50	66.51	64.30	**5**

Table 2: Official Task 2 Results. Metrics include Distinct-2 (Dist-2), BERTScore-F1 (BERT-F1), Politeness & Respectful Score (PRS), Quality Score (QS), and Contextual Counter-Narrative Coherence (CCNC).

accurate hate speech span detection, achieving 1st place globally in English, Tamil, and Hindi for Task 1. Furthermore, our approach generated highly coherent, diverse, and polite counter-narratives, securing 1st place in Task 2 for English. Future work will focus on improving generative lexical diversity and contextual quality for Dravidian languages like Tamil, potentially through targeted supervised fine-tuning and evaluating how our heuristic span-matching pipeline holds up against extreme phonetic noise, such as ASR-transcribed (Automatic Speech Recognition) social media data.

Limitations

Despite the strong empirical performance of our few-shot ICL approach, several limitations remain within our system:

API Dependency and Deployment Constraints: Our official submission relies entirely on a proprietary, massive language model (Qwen3-Max) accessed via a cloud API. This limits the system's viability for localized, offline, or low-latency deployment. Furthermore, transmitting sensitive and highly toxic hate speech data to external APIs raises potential data privacy concerns compared to utilizing local, fine-tuned, open-weight models.

Generative Diversity in Low-Resource Languages: As evidenced by the Task 2 results, there is a stark contrast in the Distinct-2 metric between English (73.56%) and Tamil (25.61%). This highlights a fundamental limitation of current foundation models: while they can generate polite and grammatically correct text in Tamil, the vocabulary is highly constrained, repetitive, and formulaic. The system struggles to capture the deep morphosyntactic richness and cultural nuances required to generate highly diverse and effective counter-narratives in Dravidian languages.

Heuristic-Dependent Span Extraction: Unlike traditional token-classification architectures (e.g., standard BERT-based models), autoregressive LLMs inherently struggle to predict exact numerical character-level indices due to subword tokenization disparities. Our pipeline relies heavily on a heuristic, multi-level string matching post-processing step to map LLM-generated substrings back to the original comment. This fallback mechanism is vulnerable to failure when processing highly noisy, misspelled, or intentionally obfuscated social media text.

Acknowledgements

We thank the LT-EDI 2026 organizers for curating the dataset and facilitating this shared task.

References

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

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Thenmozhi Durairaj, John Philip McCrae, Paul Buitelaar, Prasanna Kumaresan, and Rahul Ponnusamy. 2022. Overview of the shared task on homophobia and transphobia detection in social media comments. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 369–377.

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.

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.

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.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

A Appendix: Few-Shot Prompts

A.1 Task 1: Span Detection Prompt Template

System: You are an expert in detecting hate speech. Given a comment, identify if it contains Homophobia, Transphobia, or None. If hate speech is present, extract the exact substring and its character start and end indices. Output strictly in JSON format.
User (Example 1): "Text: [Hateful Comment]"
Assistant (Example 1): {"class": "Homophobia", "span_text": "[Extracted Span]", "positions":[start, end]}
... (followed by 9 more diverse examples)

A.2 Task 2: Counter-Narrative Prompt Template

System: You are an empathetic moderator aiming to de-escalate toxicity. Given a hateful social media comment, generate a factual, logical, and polite counter-narrative to challenge the hate speech.
User (Example 1): "Hateful Text: [Text]"
Assistant (Example 1): "[Polite Counter Narrative]"
... (followed by 4 more examples, dynamically localized for Tamil)