SigJBS@LT-EDI 2026: QLoRA-Tuned Homophobic and Transphobic Counter Narrative Generation

Gaurangi Sinha[1] **Rajarajeswari Palacharla**[1] **Manoj Balaji Jagadeeshan**[2]

[1]Department of Computer Science and Engineering, Texas A&M University
[2]Department of Computer Science and Engineering, Indian Institute of Technology, Kharagpur

Correspondence: gaurangisinha@tamu.edu

Abstract

We present our approach to LT-EDI@ACL 2026 on counter-narrative generation for homophobic and transphobic comments. Generating high-quality counter-narratives in multilingual and low-resource settings remains challenging, particularly when data imbalance and script variation affect model performance. To address these issues, we explore multiple modeling strategies built around Gemma 3 12B with QLoRA fine-tuning, including data rebalancing and alternative input strategies for Tamil. Our findings show that task-specific fine-tuning combined with native-script Tamil produces more stable and higher-quality outputs than large few-shot prompts or transliteration-based inputs. On the official leaderboard, our system ranks second in English with an overall score of 86.35% and sixth in Tamil with 63.77%, highlighting both the effectiveness of targeted fine-tuning and the challenges of low-resource counter-narrative generation.

1 Introduction

As LGBTQ+ individuals gain greater visibility in online spaces, they are frequently met with homophobic and transphobic hostility (Balaji and Chinmaya, 2022). Prior work has documented the prevalence and linguistic characteristics of such abuse across platforms, including YouTube and other social media, highlighting both language-specific and cross-lingual patterns (Chakravarthi, 2024; Kumaresan et al., 2025). These findings frame the problem not merely as content moderation but as a broader social and moral challenge.

Beyond detection, recent research has explored counter-speech generation as a constructive intervention strategy (Tekiroğlu et al., 2020)(Chung et al., 2019). Counter-narratives aim to respond to hateful content with corrective, empathetic, and non-escalatory language (LT-EDI 2026 Organizers, 2026; Prasannan et al., 2025). Unlike detection-only pipelines, counter-narrative generation requires models to simultaneously maintain safety, fluency, contextual relevance, and moral clarity. This challenge is particularly acute for anti-LGBTQ+ content, which often blends misinformation, moral condemnation, and dehumanizing rhetoric.

Although large language models (LLMs) have shown promise in both detection and generation settings, existing approaches remain predominantly English-centric (Pendurkar and Sharon, 2025). Extending such systems to multilingual and low-resource contexts introduces additional complexity (Ling et al., 2025). In languages such as Tamil, limited annotated data and sociolinguistic variation make it substantially harder to produce fluent, culturally grounded counter-narratives. The LT-EDI@ACL 2026 shared task (LT-EDI 2026 Organizers, 2026) situates this challenge in a bilingual English–Tamil setting focused on counter-narrative generation.

Methodologically, our system follows recent parameter-efficient adaptation work. LoRA (Hu et al., 2021) and QLoRA (Dettmers et al., 2023) enable practical fine-tuning of large instruction-tuned models under limited computational resources, making them well suited to shared-task settings. We combine this adaptation strategy with a multilingual prompting setup and language rebalancing rather than training a new model from scratch.

We present the SigJBS system for Task 2 of LT-EDI@ACL 2026 (LT-EDI 2026 Organizers, 2026),[1] built on a quantized Gemma 3 12B instruction model. We evaluate prompting-based baselines against supervised fine-tuning and introduce a Tamil-focused rebalancing stage to address the pronounced English–Tamil data imbalance. Our work

[1]Code repository: https://github.com/gaurangisinha-tamu/LTEDI-Counter-Narrative-Generation.

Proceedings of the Sixth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 234–238

contributes: (1) a memory-efficient bilingual system combining 4-bit quantization with parameter-efficient adaptation, (2) a structured comparison across prompting, fine-tuning, and transliteration strategies, and (3) evidence that Tamil oversampling yields measurable improvements in generation quality.

2 Data and Metric

The training data distribution is shown in Table 1. The imbalance between English and Tamil directly shaped our final training recipe.

Language	Train	Test
English	1,800	66
Tamil	800	109

Table 1: Task 2 data released by the organizers.

Formally, we write the bilingual training collection as

$$\begin{aligned} \mathcal{D} &= \mathcal{D}_{en} \cup \mathcal{D}_{ta}, \\ \mathcal{D}_\ell &= \{(x_i^\ell, y_i^\ell)\}_{i=1}^{N_\ell}. \end{aligned} \tag{1}$$

where x_i^ℓ is a hateful comment in language $\ell \in \{en, ta\}$ and y_i^ℓ is the corresponding gold counter-narrative. The released data is imbalanced, with $N_{en} > N_{ta}$, so bilingual fine-tuning without reweighting naturally favors English.

The official Task 2 score is the average of five percentage-scaled components: BERTScore (Zhang et al., 2020), Distinct-2, politeness and respectful score (PRS), quality score (QS), and contextual counter-narrative coherence (CCNC). For a submission s, the shared-task score is

$$\text{Score}(s) = \frac{1}{5}\big(\text{D2} + \text{BS} + \text{PRS} + \text{QS} + \text{CCNC}\big). \tag{2}$$

3 Method

3.1 Backbone and Prompting

Our backbone is `unsloth/gemma-3-12b-it-unsloth-bnb-4bit`, (Daniel Han and team, 2023) a 4-bit Gemma 3 12B model loaded with Unsloth for efficient inference and QLoRA fine-tuning. We first evaluate prompting-only baselines with $k \in \{0, 1, 3, 5, 10\}$ in-context examples. The prompt asks for a respectful, non-toxic counter-narrative in 1–3 sentences, and for Tamil inputs it explicitly requires Tamil output. (Team et al., 2025)

Prompt construction follows the same chat format in both inference and fine-tuning. Each instance begins with a task instruction describing tone and output constraints, followed by a short assistant acknowledgment, and then the user comment to be answered. When few-shot examples are enabled, we insert them as additional user–assistant turns before the final test comment. Few-shot examples are sampled within language and approximately balanced across available hate labels so that a single type of abusive framing does not dominate the context.

Let $\mathcal{E}_k^\ell(x)$ denote the k in-context examples selected for language ℓ. The resulting prompt context can be written as

$$\mathcal{C}_k(x, \ell) = \text{Template}\Big(s, \mathcal{E}_k^\ell(x), x, \ell\Big), \tag{3}$$

where s is the shared system instruction. Generation then follows

$$\hat{y} = \text{Decode}_\theta\big(\mathcal{C}_k(x, \ell)\big), \tag{4}$$

with $k = 0$ corresponding to zero-shot prompting.

To keep the few-shot prompts label-diverse, we approximately allocate

$$n_c = \left\lceil \frac{k}{|\mathcal{Y}_\ell|} \right\rceil \tag{5}$$

examples per label $c \in \mathcal{Y}_\ell$ before trimming back to k total examples. This is a simple heuristic, but it avoids few-shot prompts from being dominated by one label-specific pattern.

We use deterministic decoding for the final implementation rather than temperature sampling. This decision was pragmatic: in early experiments, sampling increased stylistic variation but also increased formatting drift and occasional off-task outputs. Deterministic decoding gave more stable generations and made behavior easier to compare across shot settings and fine-tuning variants.

3.2 Supervised Fine-Tuning

We then convert each labeled training pair into a chat-style supervision instance and optimize the standard autoregressive objective

$$\mathcal{L}_{\text{SFT}} = -\sum_{t=1}^{T} \log p_\theta(y_t \mid x, y_{<t}), \tag{6}$$

where x denotes the prompt context and y the gold counter-narrative. We train LoRA adapters (Hu

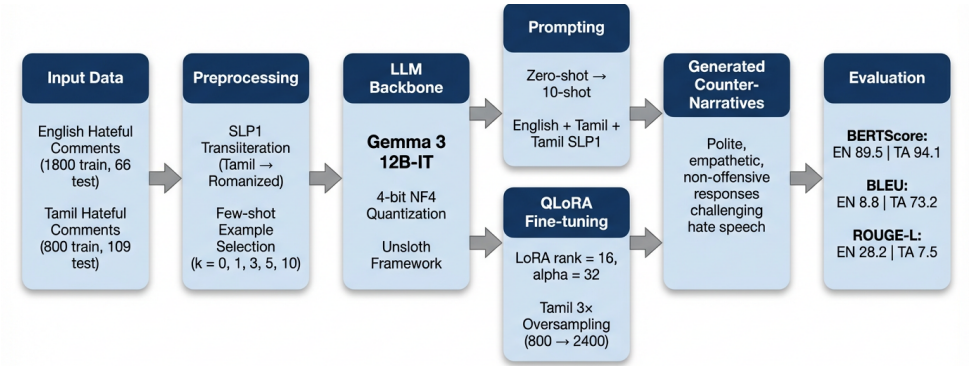

Figure 1: Pipeline for QLoRA Tuned Homophobic and Transphobic Counter Narrative Generation. The system starts from the English and Tamil shared-task inputs, applies preprocessing and prompt construction, adapts a quantized Gemma 3 12B backbone through prompting and QLoRA fine-tuning, and produces bilingual counter-narratives that are evaluated using the official shared-task metrics.

et al., 2021) under QLoRA (Dettmers et al., 2023) with rank $r = 16$, $\alpha = 32$, zero dropout, and target modules on both attention and MLP projections.

With LoRA, a frozen linear map W_0 is adapted through a low-rank update

$$\begin{aligned} W &= W_0 + \Delta W, \\ \Delta W &= BA, \end{aligned} \tag{7}$$

where $A \in \mathbb{R}^{r\times d}$ and $B \in \mathbb{R}^{d'\times r}$ with $r \ll \min(d, d')$. In our setup, only the LoRA parameters are updated while the quantized backbone remains frozen.

The initial fine-tuning stage uses the combined English and Tamil data for 3 epochs. This stage adapts the model to the task format itself: hateful comment in, constructive counter-narrative out. Using the same chat template for both supervision and inference reduces train–test mismatch and encourages the model to internalize not only the semantic goal of the task but also its stylistic constraints.

3.3 Tamil Rebalancing and Transliteration Variant

Because English training data is much larger, we add a second Tamil-focused stage that oversamples Tamil 3×:

$$\tilde{\mathcal{D}} = \mathcal{D}_{en} \cup \bigcup_{i=1}^{3} \mathcal{D}_{ta}^{(i)}. \tag{8}$$

This produces a 4,200-example balanced set and is followed by one more training epoch. The goal is not to remove English supervision, which remains useful, but to prevent the adapter from being dominated by the larger English subset. We also increase Tamil generation length from 128 to 256 tokens in this stage because some Tamil generations were overly short during initial analysis.

This rebalanced stage can also be viewed as modifying the effective bilingual objective to place greater mass on Tamil examples:

$$\begin{aligned} \mathcal{L}_{\text{bal}} = \lambda_{en}\, \mathbb{E}_{(x,y)\sim\mathcal{D}_{en}}\big[-\log p_\theta(y \mid x)\big] \\ + \lambda_{ta}\, \mathbb{E}_{(x,y)\sim\mathcal{D}_{ta}}\big[-\log p_\theta(y \mid x)\big], \end{aligned} \tag{9}$$

with $\lambda_{ta} > \lambda_{en}$ induced by oversampling rather than explicit loss weights.

In a separate variant, we transliterate Tamil into SLP1 for both training and inference. The motivation was to test whether romanization would better match the backbone model's prior exposure to Latin-script text. In practice, however, this variant was weaker in our exploratory diagnostics and produced less satisfactory Tamil outputs, so it was not used in the final submission.

4 Experimental Setup

All experiments were run in Google Colab on a single NVIDIA A100 40GB GPU. We loaded the quantized model with a maximum sequence length of 4096 and fine-tuned at a sequence length of 1024. Training uses per-device batch size 2, gradient accumulation 4, AdamW 8-bit, learning rate 2×10^{-4}, cosine decay, and warmup ratio 0.1. Inference uses batch size 4 with deterministic decoding.

5 Results and Discussion

5.1 Official Shared-Task Results

Figure 1 summarizes the training and inference workflow, while Table 2 reports the official organizer scores. Our final system ranks second in English and sixth in Tamil.

Metric	English	Tamil
Distinct-2	69.32	25.29
BERTScore-F1	86.66	85.29
PRS	93.18	75.23
QS	90.91	72.02
CCNC	91.67	61.01
Overall Avg.	86.35	63.77
Rank	2	6

Table 2: Official shared-task results for the SigJBS submission. Higher is better for all reported metrics.

The English run is strong on both reference-based and rubric-based dimensions, with PRS, QS, and CCNC all above 90%. Tamil is substantially harder: despite competitive BERTScore-F1, rubric-based quality and coherence are lower, indicating that semantic overlap alone is not sufficient for high-quality Tamil counter-narratives.

5.2 Exploratory Diagnostic Results

To better understand model behavior, we also ran post-hoc diagnostics on sampled data to compare variants within the same pipeline (not official held-out scores).

Variant	BLEU	R-1	BS
EN fine-tuned (v1)	8.82	31.58	89.52
TA fine-tuned (v1)	16.75	2.50	80.80
TA balanced (v2)	73.21	7.50	94.09
TA SLP1	1.43	0.00	70.41

Table 3: Post-hoc diagnostic comparison across major modeling variants. ‘R-1‘ denotes ROUGE-1 and ‘BS‘ denotes BERTScore. These are exploratory measurements, not official shared-task scores.

These comparisons, summarized in Table 3, help explain the official English–Tamil gap. First, larger few-shot prompts were not reliably helpful: average output length fell sharply as k increased, from 53.3 to 17.9 words in English and from 25.7 to 11.6 words in Tamil between 0-shot and 10-shot prompting. This indicates that long in-context prompts often encouraged shorter, more template-like completions rather than richer counter-narratives. Second, native Tamil script clearly outperformed the SLP1 transliteration variant in our diagnostic analysis, which justified retaining native script in the submitted system. Third, the Tamil oversampling stage improved the behavior of generated Tamil outputs, particularly on BERTScore and response adequacy, supporting the decision to rebalance training even though the final official Tamil rank remained below the top systems.

5.3 Qualitative Error Patterns

Manual inspection of generations revealed consistent patterns, illustrated in Table 4.

Pattern	**Input (hateful comment)**	**Model output**
EN: Generic	*“This guy was very bad and not used in any society. Born naturally good but artificially changed...”*	*“LGBTQ+ individuals aren’t ‘artificially changed’—they live authentically as themselves. Being true to one’s identity isn’t wrong.”* [Fluent but relies on broad inclusion language without addressing the ‘society’ framing.]
TA: Brief (v1)	*“Names used as slurs to mock people...”* (Tamil input)	*“Slurs increase social discrimination; they should be avoided.”* (translated, 7 words in Tamil) [Correct stance but too short to engage with the specific claim.]
TA: Improved (v2)	*“Gays and lesbians are known for sexual crimes...”* (Tamil input)	*“Linking sexual crimes to homosexuality is dangerous. Views that erode social trust should be avoided.”* (translated) [After rebalancing: longer, directly addresses the false crime association.]

Table 4: Representative error patterns (Tamil shown in English translation).

English outputs were usually fluent and well aligned with the task prompt, but they sometimes became slightly generic when the model over-relied on broad inclusion language rather than addressing the specific framing of the hateful comment.

Tamil outputs showed a different failure mode: the model often stayed polite but produced overly brief responses that stated a correct position without engaging with the specific hateful claim. After the rebalancing stage, Tamil generations became longer and more claim-specific.

6 Conclusion

In conclusion, we find that our best system uses QLoRA fine-tuning of a 4-bit Gemma 3 12B model with a Tamil oversampling stage to mitigate language imbalance. Overall, our experiments indicate that supervised adaptation is more effective than larger few-shot prompts and that native-script Tamil is preferable to transliterated Tamil for this task.

7 Ethical Considerations

This work addresses harmful content targeting LGBTQ+ communities. Our aim is to generate respectful counter-narratives, not to reproduce abuse. However, automated counter-speech can still fail through tone mismatch, oversimplification, or contextual misunderstanding. We therefore do not view such systems as substitutes for human moderation or community-led interventions, and any practical deployment should include human oversight.

8 Limitations

Our study has several limitations. First, the training data is relatively small and imbalanced across languages, which likely contributed to stronger English performance and weaker Tamil generalization. Second, the system is not explicitly stress-tested for code-mixed text, adversarial phrasing, or evolving slang, which are common in online abuse. Finally, we do not conduct large-scale human evaluation, so safety and quality conclusions should be interpreted as task-specific rather than deployment-ready.

9 Acknowledgments

We thank Hugging Face for open access to pre-trained model weights and the Transformers library, the Unsloth team for efficient LoRA training tooling, and the shared task organizers for providing the datasets.

References

Manoj J Balaji and HS Chinmaya. 2022. A study on sentimental analysis, homophobia-transphobia detection for dravidian languages. In *CEUR Workshop Proceedings. https://ceur-ws.org*, volume 3395, pages T2–7.

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 Tekiroglu, 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, Florence, Italy. Association for Computational Linguistics.

Michael Han Daniel Han and Unsloth team. 2023. Unsloth.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. *arXiv preprint arXiv:2305.14314*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

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*, 12:100169.

Hongyi Ling, Shubham Parashar, Sambhav Khurana, Blake Olson, Anwesha Basu, Gaurangi Sinha, Zhengzhong Tu, James Caverlee, and Shuiwang Ji. 2025. Complex llm planning via automated heuristics discovery. *Preprint*, arXiv:2502.19295.

LT-EDI 2026 Organizers. 2026. Counter-narrative generation on homophobic and transphobic comments - LT-EDI@ACL 2026. `https://www.codabench.org/competitions/11333/`. Shared task overview and evaluation page, accessed March 3, 2026.

Sumedh Pendurkar and Guni Sharon. 2025. Policy-guided search on tree-of-thoughts for efficient problem solving with bounded language model queries. *Transactions on Machine Learning Research*.

Praveen Prasannan, Prasanna Kumar Kumaresan, Saranya Rajiakodi, C. N. 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.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

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, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *Proceedings of the International Conference on Learning Representations*.