# Igniters@LTEDI 2026: Multilingual Gender-Inclusive Language Generation with mT5 and Counter-Narrative Generation Using Llama-3

**Rajendran S**[1] **Ramkumar N**[2] **Malarselvi R**[3]
[1,3]Undergraduate Student, Coimbatore Institute of Technology, India
[2]Assistant Professor, Coimbatore Institute of Technology, India
asrajendrayadav@gmail.com[1], ramkumar@cit.edu.in[2], malarrajamani24@gmail.com[3]

## Abstract

The deployment of Large Language Models (LLMs) has intensified concerns regarding the propagation of societal stereotypes encoded with web-scale training corpora. This paper presents a dual-paradigm framework specially designed to address multilingual gender-inclusivity and counterfactual generation. For multilingual gender-neutral text transformation, a fine-tuned mT5 encoder–decoder model performs controlled sentence rewriting with minimal edits while preserving semantic fidelity and grammatical fluency. For counter-narrative generation, the Llama-3 8B decoder-only model is employed to generate empathetic and persuasive responses through structured prompt-based generation. The framework is evaluated using datasets from the LT-EDI ACL 2026 shared task across multiple languages, including English, Tamil, Kannada, German, and Spanish. Experimental results demonstrate strong effectiveness in identifying and neutralizing gender markers, particularly in morphologically rich languages, while the counter-narrative component achieves high performance in politeness, coherence, and relevance. Overall, the proposed approach contributes toward the development of responsible and inclusive multilingual NLP systems.

## 1 Introduction

The pursuit of fairness and inclusivity in natural language processing (NLP) has become increasingly important with the tremendous growth in digital text generation. Large Language Models (LLMs) are trained on vast corpora - including the internet, books, and social media, which inherently reflect and encode human prejudices (Bolukbasi et al., 2016; Sun et al., 2019; Zhao et al., 2018). Consequently, the models often return gender-biased predictions, stereotypically aligning occupations with gender markers rather than maintaining neutral contexts. Furthermore, studies show unequal representation in sentiment and descriptions, where male-coded text often emphasizes leadership and strength, while female-coded text frequently centers on emotion or appearance. Research in abusive language detection and counter-speech generation has highlighted the importance of developing systems that mitigate harmful narratives and promote constructive responses (Mathew et al., 2019; Dinan et al., 2019; Davidson et al., 2017; Founta et al., 2018).

Traditional sequence-to-sequence (Seq2Seq) models are often suboptimal for inclusive language generation, as they regenerate the entire sentence from scratch, leading to unnecessary token copying (Vanmassenhove et al., 2021; Chinnan et al., 2025; Piergentili et al., 2025). Consequently, the literature highlights a shift toward sequence-to-edit (Seq2Edit) frameworks like LaserTagger and Felix, which utilize tagging or localized mask-infilling to preserve original fluency and mitigate the over-correction phenomenon (Nozza et al., 2019; Watson et al., 2024).

## 2 Methodology

### 2.1 Dataset Description

The primary data used for the research is sourced from the LT-EDI ACL 2026 Shared Task, which consists of two objectives: Subtask A (Multilingual Gender-Inclusive Generation) and Subtask B (English Counterfactual Generation) (Chakravarthi et al., 2026). Subtask A focuses on generating inclusive sentences by applying correct gender-neutral terminology, replacing gender-marked nouns, roles, and pronouns with inclusive alternatives. Subtask B focuses on generating counter-narratives for gender-biased sentences in an empathetic and persuasive way. The overall objective is to transform gender-biased, gender-marked, or exclusionary sentences into inclusive, gender-neutral, and contextually coherent alternatives while pre-



*Proceedings of the Sixth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 172–176

serving the original meaning and fluency.

The dataset used for Subtask A comprises sentence pairs in five different languages: English, Tamil, Kannada, German, and Spanish. It includes both gender-neutral word pairs and gender-neutral sentence pairs, which allows the model to learn lexical substitutions and contextual sentence-level transformations. The dataset used for Subtask B consists of counterfactual sentence pairs in English. Languages such as English, German, and Spanish use pronouns and occupational nouns to express gender. Dravidian languages such as Tamil and Kannada often encode gender morphologically through suffixes and inflectional endings. This diversity introduces additional complexity, especially for agglutinative languages where gender markers are embedded within words. A summary of the dataset distribution across languages and subtasks is presented in Table 1.

### 2.2 Model Description

The system uses a dual-paradigm architectural strategy which utilizes the strengths of encoder-decoder and decoder-only models to address the different linguistic and psychological requirements of text transformation for fairness.

#### 2.2.1 Gender-Inclusive Generation

Subtask A is framed as a controlled sequence-to-sequence rewriting problem. To handle this task, a multilingual transformer-based encoder-decoder model, mT5 has been fine-tuned (**?**). mT5 is a massively multilingual pre-trained language model based on T5 architecture. The encoder-decoder structure is particularly well-suited for mapping source-to-target transformations while making minimal edits to the sentences. The corruption and mask-infilling objective of the model helps to identify specific biased spans and generate neutral replacements without regenerating the entire sentence (Muthusamy Chinnan et al., 2025). This approach ensures that the majority of input tokens are copied directly, preserving the original fluency and structural nuances. The multilingual nature of mT5 allows the model to learn shared semantic representations across different languages (Xue et al., 2021).

#### 2.2.2 Chain-of-Thought Prompting Strategy

The proposed system uses a Chain-of-Thought (CoT)-inspired prompting strategy to improve contextual understanding during gender-inclusive rewriting. Instead of performing direct word substitution, the model is guided to interpret the surrounding semantic context before predicting an appropriate inclusive term. This helps the model infer relationships between entities, pronouns, and sentence structure, producing more contextually accurate and socially inclusive outputs.

The reframed CoT-style prompt used for gender-inclusive generation is shown below:

> **Instruction:** You are an inclusive language assistant. Read the context, identify the missing or gendered term, determine its semantic relationship with the subject, and replace it with an appropriate gender-neutral or inclusive term.
>
> **Input:** {sentence_pair}
>
> **Output:** {inclusive_sentence}

For example, given the input "The doctor entered the examination room. _____ reviewed the patient's report carefully," the model identifies that the blank refers to the doctor and generates the gender-neutral output: "They reviewed the patient's report carefully." This CoT-inspired strategy helps reduce incorrect substitutions and improves the fluency of generated gender-inclusive text.

#### 2.2.3 Counter Narrative Generation

The Subtask B requires generating empathetic, persuasive counter-narratives, for which the Llama-3 8B model was utilized (Dubey et al., 2024). Llama-3 is a decoder-only transformer-based autoregressive model trained on large-scale multilingual and English-dominant corpora, enabling strong contextual reasoning and natural language generation capabilities. The task is formulated as a controlled prompt-based generation problem, where each input sentence is embedded within a structured prompt that instructs the model to produce a constructive counterfactual response that highlights alternative perspectives or evidence. The Llama-3 8B model utilizes its large-scale pretraining knowledge to move from restrictive binary claims to universal, practice-based explanations.

#### 2.2.4 Prompt Template for Counter-Narrative Generation

For Subtask B, a structured prompt template was used to guide the model toward generating concise, empathetic, and non-adversarial counter-narratives. The template used for counter-narrative generation is shown below:

| Task | Category | English | German | Spanish | Tamil | Kannada | Total |
|---|---|---|---|---|---|---|---|
| Subtask A | Gender Neutral Word Pairs | 673 | – | 200 | 742 | 693 | 2308 |
| Subtask A | Gender Neutral Sentence Pairs | 1074 | 1002 | 200 | 1074 | 1074 | 4424 |
| Subtask B | Counterfactual Sentence Pairs | 726 | – | – | – | – | 726 |
| **Total** | | 2473 | 1002 | 400 | 1816 | 1767 | 7458 |

Table 1: Dataset statistics for LT-EDI Shared Task subtasks across languages.

**Instruction:** Generate an empathetic and persuasive counterfactual response to the following gender-biased statement. Do not insult the speaker. Do not add new facts. Keep it concise.

**Input:** "Women are not suitable for leadership roles."

**Output:** "Leadership ability depends on skills, experience, and opportunity rather than gender."

This prompt design encouraged the model to produce responses that directly addressed the underlying gender bias while maintaining a respectful and constructive tone.

### 2.3 Experimental Setup

The models were implemented using the Hugging Face Transformers library and fine-tuned on the dataset using a single NVIDIA RTX 3090 GPU. To address the computational constraints associated with fine-tuning large-scale models like Llama-3 8B and the multilingual breadth of mT5, Parameter-Efficient Fine-Tuning (PEFT) via Low-Rank Adaptation (LoRA) is employed. This approach allows the model to learn task-specific nuances for gender-inclusive and counterfactual generation by updating only a small fraction of the total parameters, thereby preventing catastrophic forgetting and maintaining the integrity of the pre-trained weights.

#### 2.3.1 Model Configuration

The preprocessing for Subtask A was performed using the mT5 tokenizer, which converts input and target sentences into subword tokens compatible with the multilingual encoder-decoder architecture. The Chain-of-Thought (CoT) reasoning paths in the instructions guide the model to identify gendered inflections and handle agglutinative languages. For Subtask B, Llama-3 8B was used with prompt-based generation to produce concise and empathetic counter-narratives. The key training and generation settings are summarized in Table 2.

| Parameter | mT5 | Llama-3 8B |
|---|---|---|
| Framework | HF Transformers | HF Transformers |
| Hardware | RTX 3090 | RTX 3090 |
| Method | Fine-tuning | Prompt inference |
| Batch size | 4 | – |
| Optimizer | AdamW | – |
| Learning rate | $5 \times 10^{-5}$ | – |
| Weight decay | 0.01 | – |
| Epochs | 5 | – |
| Warmup steps | 100 | – |
| Max. train steps | 600 | – |
| Max. gen. length | – | 128 |
| Temperature | – | 0.7 |
| Top-p / Top-k | – | 0.9 / 50 |
| Repetition penalty | – | 1.1 |
| Decoding | – | Autoregressive |

Table 2: Training and generation configuration for mT5 and Llama-3 8B.

### 2.4 Evaluation Metrics

The performance of the system is assessed using an LLM-as-a-Judge evaluation framework which utilized advanced models to provide scalable assessment for low-resource languages. The final performance score for each system is computed as the average of multiple task-specific dimensions.

The Subtask A employs three metrics to evaluate the system for inclusive transformation: Gender Assumption (GA), Gender Neutrality (GN), and Quality and Relevance (QR). The final score for Subtask A is the arithmetic average of the GA, GN, and QR scores. The Subtask B prioritizes the effectiveness of the model in challenging biased claims through a functional cognitive lens. The model has been evaluated with metrics such as Politeness and Respectfulness (PR), Contextual Counter-Narrative Coherence (CCNC), and Quality and Relevance (QR). The overall performance for Subtask B is calculated as the average of the PR, CCNC, and QR scores.

## 3 Results and Discussion

The performance of the proposed dual-paradigm system was evaluated using rubric-based qualitative metrics assessed through an LLM-as-a-Judge framework.

### 3.1 Gender-Inclusive Generation Results

The results for Subtask A, summarized in Table 3, demonstrate the versatility of the mT5 architecture in handling diverse morphological typologies.

| Language | GA | GN | QR | Average |
|---|---|---|---|---|
| English | 67.50 | 70.00 | 43.13 | 60.21 |
| German | 69.70 | 72.73 | 9.09 | 50.51 |
| Spanish | 97.50 | 100.00 | 47.50 | 81.67 |
| Tamil | 95.00 | 93.65 | 87.57 | 92.07 |
| Kannada | 96.00 | 96.00 | 32.00 | 74.67 |

Table 3: Evaluation Results for Subtask A (mT5)

The system achieved its highest performance in Tamil (92.07) and Spanish (81.67), indicating that the model effectively navigated the strict rules of gender neutralization, such as de-gendering occupations and pronouns. In the agglutinative languages (Tamil and Kannada), the model demonstrated high precision in identifying gender-marked suffixes.

### 3.2 Counter-Narrative Generation Results

Llama-3 8B exhibited superior performance across all dimensions, with a near-perfect Quality and Relevance (QR). The success is attributed to the model's 8-billion parameter scale and advanced instruction-tuning, which allowed it to successfully adopt a functional view of counterfactuals. The results for Subtask B are summarized in Table 4.

| Metric | Score |
|---|---|
| PR | 95.00 |
| CCNC | 95.00 |
| QR | 97.50 |
| Average | 95.83 |

Table 4: Evaluation Results for Subtask B (Llama-3 8B)

Table 5 presents representative biased inputs and the corresponding generated counter-narratives.

| Biased Sentence | Generated Counter-Narrative |
|---|---|
| Women are not good at leadership. | People of all genders can be great leaders. |
| Men should not show emotions. | People of all genders should be able to express emotions openly. |
| Girls are weak in Mathematics. | People of all genders can excel in Mathematics. |
| Boys are naturally aggressive. | People of all genders can be gentle and strong in different ways. |

Table 5: Qualitative examples of counter-narrative generation for gender-biased statements.

## 4 Limitations

Despite the strong multilingual performance, the proposed framework consists several limitations. The system relies heavily on the quality and size of the training dataset, which may limit generalization for low-resource languages and culturally nuanced gender expressions. In some cases, the mT5 model introduces semantic drift and over-neutralization, particularly in morphologically rich languages such as Tamil and Kannada. Additionally, the LLM-as-a-Judge evaluation framework introduces a degree of subjectivity in assessing politeness, coherence, and relevance.

## 5 Conclusion

This paper presented a dual-model framework for multilingual gender-inclusive rewriting and counter-narrative generation. The mT5 model was used for controlled gender-neutral text transformation across English, Tamil, Kannada, German, and Spanish, while Llama-3 8B was used for prompt-guided counter-narrative generation. The use of CoT-inspired prompting helped the system interpret contextual relationships before generating inclusive replacements, particularly for morphologically rich languages.

Experimental results showed strong performance in counter-narrative generation and competitive performance in multilingual gender-inclusive rewriting. The findings suggest that combining multilingual encoder-decoder models with prompt-guided decoder-only models can support inclusive and respectful text generation. Future work may include stronger baseline comparisons, ablation studies, human evaluation, and extensions to multimodal content such as speech and visual memes.

## AI Usage Statement

The authors used generative AI tools only for auxiliary writing support, including grammar correction, and wording refinement. The technical contributions, experimental setup, analysis, and conclusions were developed and validated by the authors. The authors take full responsibility for the content of the paper.

## References


Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*.

Bharathi Raja Chakravarthi, Shunmuga Priya Muthusamy Chinnan, Paul Buitelaar, Meghann L. Drury-Grogan, Miguel Ángel García Cumbreras, Salud María Jiménez Zafra, Thomas Mandl, Sylvia Jaki, Rahul Ponnusamy, Anand Kumar M, Dhanalakshmi V, Bharathi B, Premjith B, Senthil Kumar B, and Sathiyaraj T. 2026. Insights from Multilingual Gender Inclusive Language Generation Shared Task. In *Proceedings of the Sixth Workshop on Language Technology for Equality, Diversity, Inclusion*. Association for Computational Linguistics.

Shunmuga Priya Muthusamy Chinnan, Meghann Drury-Grogan, and Bharathi Raja Chakravarthi. 2025. Gender inclusive language generation framework: A reasoning approach with retrieval augmented generation and chain-of-thought. *Knowledge-Based Systems*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *ICWSM*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. In *EMNLP*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aishwarya Rao, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Antigoni-Maria Founta, Constantinos Djouvas, and Despoina Chatzakou. 2018. Large scale crowdsourcing and characterization of twitter hate speech. In *ICWSM*.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2019. Thou shall not hate: Countering online hate speech. In *ICWSM*.

Shunmuga Priya Muthusamy Chinnan, Meghann Drury-Grogan, and Bharathi Raja Chakravarthi. 2025. Gender inclusive language generation framework: A reasoning approach with rag and cot. *Knowledge-Based Systems*, 328:114092.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2019. What the mask? making sense of gender bias in bert. In *ACL Workshop on Gender Bias in NLP*.

Andrea Piergentili, Beatrice Savoldi, Matteo Negri, and Luisa Bentivogli. 2025. Gender-neutral rewriting in italian: Models, approaches, and trade-offs. *arXiv preprint arXiv:2509.13480*.

Tony Sun, Andrew Gaut, Shirlyn Tang, and Yuxin Huang. 2019. Mitigating gender bias in natural language processing. In *ACL*.

Eva Vanmassenhove, Chris Emmery, and Dimitar Shterionov. 2021. Neutral rewriter: A rule-based and neural approach to automatic rewriting into gender-neutral alternatives. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8940–8948.

Tom Watson and 1 others. 2024. Fine-tuning with gender-inclusive language for bias mitigation in large language models. *arXiv preprint arXiv:2407.04434*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. *Proceedings of NAACL-HLT*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution. In *NAACL*.