CAI@LTEDI 2026: Multilingual Gender Inclusive Language Generation using Instruction-Guided mT5 Transformer Model

Aiswarya P Nair[1], Sree S Bhagya[1], Chinnu Jacob[1]
[1]Centre for AI, TKM College of Engineering, Kollam, India.
Correspondence: paiswariya2003@gmail.com, 25072@tkmce.ac.in

Abstract

Gender bias in multilingual language generation systems poses serious ethical and social issues, especially in languages with complex morphology. In this study, we propose a lightweight multilingual approach that employs instruction-guided fine-tuning of the mT5-small transformer model for gender-inclusive language generation. The framework accommodates five languages: English, German, Spanish, Tamil, and Kannada. The approach uses the task-prefix rewriting method to transform gender-specific sentences to their gender-neutral versions. The training data from different languages is combined into a single multilingual dataset for sequence-to-sequence fine-tuning. Beam search decoding with repetition constraints is used during inference to improve the quality of the output. The system's performance is measured using GIFI, semantic similarity, and an overall combined score across all languages. Experimental results show that the system can eliminate gender-biased language while retaining semantic meaning in part across languages.

1 Introduction

Recent advances in Natural Language Processing (NLP) have improved language generation systems, but these models often inherit gender biases from training data, producing stereotypical or exclusionary language. Gender-inclusive language generation aims to rewrite biased or gender-marked expressions into neutral alternatives while preserving meaning and grammatical correctness. This task becomes more challenging in multilingual settings due to differences in grammatical gender and linguistic structure across languages.

The Language Technology for Equality, Diversity, and Inclusion (LT-EDI) shared task focuses on multilingual gender-inclusive language generation (Chakravarthi et al., 2026b). In this work, we propose a lightweight multilingual framework for rewriting gender-biased sentences into inclusive alternatives across English, German, Spanish, Tamil, and Kannada using the mT5-small transformer architecture. Our approach employs unified multilingual training with task-prefix guided rewriting and is evaluated using fairness and semantic similarity metrics, along with multilingual error analysis.

These components enable the model to generate inclusive alternatives that preserve the original sentence's meaning and facilitate the learning of rewriting patterns across multiple languages.

2 Related Work

Recent research in Natural Language Processing (NLP) has highlighted the presence of gender bias in word embeddings and language models. Early studies by Bolukbasi et al. (Bolukbasi et al., 2016) showed that embeddings capture societal stereotypes, leading to biased downstream predictions. Later surveys, such as Stańczak et al. (Stańczak and Augenstein, 2021), reviewed methods for detecting and mitigating such biases in NLP systems.

Gender bias has also been observed in neural text generation and dialogue systems. Sheng et al. (Sheng et al., 2019) demonstrated that language generation models often reproduce stereotypical gender roles, while Costa-jussà et al. (Costa-Jussà and de Jorge, 2020) and Dinan et al. (Dinan et al., 2020) explored gender-aware translation and bias reduction in conversational AI. Other works proposed mitigation strategies, including adversarial learning and contextual stereotype analysis (Liu et al., 2020; Bartl et al., 2020).

Beyond model development, researchers have examined gender representation in real-world text data. Asr et al. (Asr et al., 2021) introduced the Gender Gap Tracker to analyze media representation, while Hovy and Spruit (Hovy and Spruit, 2021) discussed the broader societal impact of bi-

Proceedings of the Sixth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 150–154

ased NLP systems.

More recently, multilingual gender-inclusive language generation has gained attention. Chinnan et al. (Chinnan et al., 2025) proposed reasoning-based inclusive language generation, and the LT-EDI shared task (Chakravarthi et al., 2026a) introduced a multilingual benchmark for gender-inclusive rewriting. Transformer-based rewriting frameworks have also been explored for Portuguese, French, and multilingual settings (Veloso et al., 2023; Lerner and Grouin, 2024; Doyen and Todirascu, 2025).

Our work extends these efforts by developing a lightweight multilingual rewriting framework for English, German, Spanish, Tamil, and Kannada using the mT5 transformer architecture.

3 Task Definition and Dataset

3.1 Task Description

The objective is to produce a gender-inclusive substitute for an input sentence that contains gender-biased or gender-marked language that:

1. Maintain the original meaning of the sentences
2. Uses pronouns and language that is gender-neutral
3. Preserves fluidity and grammatical accuracy
4. Avoid making unwanted gender assumptions

3.2 Dataset Statistics

Table 1: Dataset Statistics for Subtask A

Language	Sentence Pairs
English	1074
German	1002
Spanish	200
Tamil	1074
Kannada	1074

The dataset used in this work (Chakravarthi et al., 2026b) is made up of parallel sentence pairs with an inclusive rewrite for each non-inclusive sentence.

4 Methodology

The proposed approach is based on multilingual sequence-to-sequence modeling. The mT5-small transformer model, which is intended for multilingual text generation, serves as the foundation for the system. With this design, the input sentence is processed by the encoder and the corresponding gender-inclusive rewrite is generated by the decoder.

Multilingual sentence pairs with gender-biased sentences and their inclusive substitutes are used to refine the model. The model learns from these samples to identify patterns of gender-marked expressions and change it into inclusive, neutral variants while maintaining the original semantic meaning. The framework can generate inclusive statements in English, German, Spanish, Tamil, and Kannada due to the model's multilingualism, that allows it to generalize across several languages.

4.1 Dataset Consolidation

English, German, Spanish, Tamil, and Kannada are the five languages for which the dataset for the multilingual gender-inclusive language generation work is supplied as distinct CSV files. Parallel sentence pairs with a gender-biased sentence and a gender-inclusive rewrite are included in every file. The language-specific datasets are merged into a single, cohesive corpus to facilitate multilingual training.

Let $\mathcal{D}_l$ denote the dataset for language $l \in \{EN, DE, ES, TA, KA\}$. Each dataset consists of sentence pairs:

$$\mathcal{D}_l = \{(x_i^l, y_i^l)\}_{i=1}^{N_l}$$

where x_i^l symbolizes the initial biased statement, while the inclusive rewrite is represented by y_i^l. All language datasets are combined to create the final multilingual dataset:

$$\mathcal{D} = \bigcup_{l \in L} \mathcal{D}_l$$

To guarantee uniform data presentation across languages, column names are standardized and incomplete samples are eliminated during consolidation. The multilingual rewriting model is then trained using this combined dataset.

4.2 Data Preprocessing

The following actions are carried out during preprocessing:

- Normalizing columns to distinguish between input and target sentences

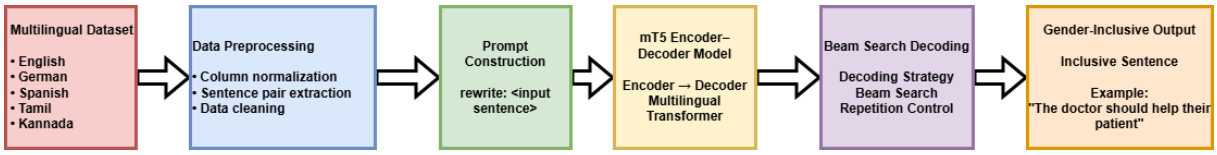

Figure 1: Architecture of the proposed multilingual gender-inclusive language generation framework.

- Elimination of noisy or incomplete samples
- Sentence pair conversion into a single training format

This guarantees that data is represented consistently in all languages.

4.3 Task Prefix Formulation

To guide the multilingual sequence-to-sequence model toward gender-inclusive rewriting, a task-prefix conditioning strategy is employed. A task-specific prefix token, `rewrite:`, is appended before each input sentence to explicitly indicate the rewriting objective.

During training, the model learns to transform gender-marked expressions into more inclusive alternatives while preserving semantic meaning and grammatical structure. This task-prefix formulation enables the multilingual model to learn inclusive rewriting patterns across multiple languages within a unified framework.

4.4 Model Training

The multilingual transformer model mT5-small is employed as the backbone sequence-to-sequence architecture for multilingual gender-inclusive rewriting. The model was selected because of its multilingual coverage, computational efficiency, and suitability for low-resource multilingual generation tasks. Larger variants such as mT5-base were not explored because of computational constraints and the relatively limited size of the shared-task dataset. The multilingual dataset is tokenized using the mT5 tokenizer. During training, the model learns mappings between gender-biased sentences and their corresponding inclusive rewrites using cross-entropy loss optimization. Table 2 summarizes the training configuration used for model fine-tuning. The multilingual fine-tuning setup enables the model to learn shared rewriting patterns across structurally different languages.

4.5 Decoding Strategy

Beam search decoding is employed during inference to improve multilingual generation quality and reduce repetitive outputs. Preliminary experiments using greedy decoding frequently produced incomplete sentences, repetitive token generation, and multilingual interference, particularly for Tamil and Kannada outputs.

Table 2: Training Configuration

Parameter	Value
Model	mT5-small
Optimizer	AdamW
Learning Rate	3e-5
Batch Size	8
Epochs	5
Maximum Sequence Length	128
Beam Width	4
Framework	HuggingFace Transformers

Beam search decoding improves fluency by exploring multiple candidate output sequences during generation. Repetition constraints are additionally applied to reduce duplicated token generation and improve sentence readability.

4.6 Evaluation Metrics

The following metrics are used to evaluate performance:

- **Gender Inclusive Fairness Index (GIFI):**

$$GIFI = 0.6 \cdot BR + 0.4 \cdot IU$$

where BR represents bias removal and IU denotes inclusive term usage.

- **Semantic Similarity:** Cosine similarity between multilingual sentence embeddings is used to calculate semantic similarity between the generated and input sentences:

$$Similarity = \cos(E_{input}, E_{output})$$

- **Overall Score:** The final score is computed as $Score = 0.5 \cdot GIFI + 0.4 \cdot Similarity + 0.1 \cdot Length$.

5 Results and Discussion

Gender-inclusive rewrites in English, German, Spanish, Tamil, and Kannada are partially produced by the multilingual instruction-guided

framework. Compared to open-ended descriptive prompts, the model does better on explicit stereotype-neutralization tests.

Table 3 presents the official shared-task evaluation results obtained by Team CAI using the proposed multilingual gender-inclusive language generation framework across all five languages.

Table 3: Official Shared-Task Evaluation Results

Lang	GA	GN	QR	Avg	Rank
English	65.00	58.75	46.88	56.88	7
German	17.65	26.47	0.00	14.71	4
Spanish	62.50	70.00	5.00	45.83	4
Tamil	45.27	54.73	52.03	50.68	5
Kannada	100.00	92.00	0.00	64.00	4

The results demonstrate that the proposed multilingual framework is capable of generating gender-inclusive rewrites across multiple languages using a unified lightweight architecture. English outputs successfully replace gender-biased statements with neutral alternatives and exhibit comparatively higher semantic retention. For instance, the claim that "men are naturally better leaders" is reinterpreted as "individual differences in leadership skills."A few German and Spanish instances exhibit similar stereotype-neutralization tendencies. Table 4 presents example multilingual rewrites generated by the proposed framework.

5.1 Error Analysis

Several multilingual generation errors were observed during evaluation. Tamil and Kannada often produced incomplete or repetitive outputs due to their morphologically rich structures, while German and Spanish showed grammatical inconsistencies and occasional mixed-language generation. Common failure modes included sentence truncation, repetitive tokens, over-neutralization, mixed-language outputs, and loss of contextual meaning. These results highlight the challenges lightweight multilingual transformers face in generalizing inclusive rewriting across diverse languages.

6 Conclusion

This paper presented a lightweight multilingual framework for gender-inclusive language generation using the mT5-small transformer model. The proposed task-prefix guided rewriting approach was evaluated on the LT-EDI 2026 Shared Task across English, German, Spanish, Tamil, and Kannada. Results showed competitive multilingual performance, particularly for Kannada, English, and Tamil, demonstrating the potential of lightweight transformer models for inclusive rewriting across diverse languages.

The evaluation also revealed challenges such as multilingual interference, grammatical inconsistencies, morphology-aware rewriting issues, and repetitive generation, especially in German, Spanish, Tamil, and Kannada. Despite these limitations, the framework demonstrates the feasibility of unified multilingual gender-inclusive rewriting using sequence-to-sequence modeling. Future work will focus on larger multilingual datasets, improved decoding strategies, morphology-aware modeling, and better semantic preservation for low-resource languages.

Limitations

Gender-biased and gender-inclusive sentence pairings make up the supervised training data used by the proposed framework. As a result, the quality and coverage of the dataset determine how well the model performs, and it might not be able to handle uncommon or unseen linguistic patterns.

Furthermore, gender information is encoded in word forms in languages with rich morphology, like Tamil and Kannada, which might make gender-neutral rewriting more difficult. The existing method ignores the larger discourse context in favor of sentence-level rewriting. To improve gender inclusive language generation, future research should examine larger multilingual datasets and better contextual modeling.

Ethical Considerations

The goal of this research is to reduce gender bias in NLP systems and promote inclusive language generation by rewriting gender-marked sentences into neutral alternatives while preserving semantic meaning. Since training data often contains social biases, careful multilingual modeling is required to avoid reinforcing stereotypes or producing incorrect rewrites across different linguistic and cultural contexts.

The proposed system aims to maintain contextual accuracy while encouraging balanced language use. Proper evaluation is essential to ensure reliable multilingual performance, and future work will focus on larger multilingual datasets and broader fairness considerations for responsible language technology.

Table 4: Examples of multilingual gender-inclusive rewriting generated by the proposed framework

Lang	Input	Generated Output
English	"Men are naturally better leaders."	"Leadership qualities vary by individual."
German	"Männer sind von Natur aus bessere Führungskräfte."	"Leadership qualities vary by individual, not Gender."
Spanish	"Los hombres son naturalmente mejores líderes."	"El liderazgo está influenciado por el individuo, no por el género."
Tamil	ஒரு புதிய மருந்தை கண்டுபிடித்தார்.	கண்டுபிடித்தனர்.
Kannada	ಒಂದು ಪ್ರಮುಖ ಪ್ರಕರಣದಲ್ಲಿ ತೀರ್ಪು ನೀಡಿದರು.	ತೀರ್ಪು ನೀಡಿದರು.

Code Availability

The implementation and datasets used in this work are available in the GitHub repository [1] to ensure research transparency and reproducibility.

References

Faezeh Torabi Asr, Maite Taboada, and Alexandra J. Y. Cheng. 2021. The gender gap tracker: Using natural language processing to measure gender bias in media. *PLOS ONE*.

Marion Bartl, Malvina Nissim, and Albert Gatt. 2020. Unmasking contextual stereotypes: Measuring and mitigating bert's gender bias. In *ACL Workshop on Gender Bias in NLP*.

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, and Meghann Drury-Grogan. 2026a. Insights from multilingual gender inclusive language generation shared task. In *Proceedings of the Sixth Workshop on Language Technology for Equality, Diversity, Inclusion*.

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. 2026b. 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 rag and cot. *Knowledge-Based Systems*.

Marta R. Costa-Jussà and Christian de Jorge. 2020. Fine-tuning neural machine translation on gender-balanced datasets. In *Proceedings of MT Summit*, pages 26–34.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of EMNLP*.

Enzo Doyen and Amalia Todirascu. 2025. Genre: A french gender-neutral rewriting system using collective nouns. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7889–7909.

Dirk Hovy and Shannon L. Spruit. 2021. The social impact of natural language processing. *Computational Linguistics*.

Paul Lerner and Cyril Grouin. 2024. Inclure: a dataset and toolkit for inclusive french translation. In *Proceedings of the 17th Workshop on Building and Using Comparable Corpora (BUCC)@ LREC-COLING 2024*, pages 59–68.

Haochen Liu, Wentao Wang, Yiqi Wang, Hui Liu, Zitao Liu, and Jiliang Tang. 2020. Mitigating gender bias for neural dialogue generation with adversarial learning. In *Proceedings of EMNLP*.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of EMNLP*.

Karolina Stańczak and Isabelle Augenstein. 2021. A survey on gender bias in natural language processing. *Journal of the ACM*.

Leonor Veloso, Luisa Coheur, and Rui Ribeiro. 2023. A rewriting approach for gender inclusivity in portuguese. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8747–8759.

[1] https://github.com/123veno/LTEDI_2026-GENDER-INCLUSIVE-LANGUAGE-GENERATION