JustGen@LT-EDI 2026: Controlled Gender Inclusive and Bias-Aware Language Generation using LLMs

Nilendu Adhikary
IReL, Dept. of CSE
IIT (BHU) Varanasi

Supriya Chanda
SCSET
Bennett University

Sukomal Pal
IReL, Dept. of CSE
IIT (BHU) Varanasi

Abstract

Over the past decade, the rapid advancement of LLMs has significantly improved natural language generation. However, these models often inherit and amplify gender biases present in large-scale training data, leading to stereotypical associations, androcentric language, and misgendering. Such biases can negatively impact applications in education, healthcare, legal systems, and automated content generation. In this paper, we address this issue as defined in the shared task LT-EDI on Gender-Inclusive Language Generation. The task focuses on rewriting gender-biased sentences into inclusive, gender-neutral alternatives while preserving meaning. We propose a retrieval-augmented framework combining lexical replacement, semantic retrieval, and controlled instruction-tuned generation. An edit-distance constraint and self-evaluation step ensure minimal, coherent, and bias-free outputs. We also present zero-shot adaptation for low-resource languagea. The implementation code is available here https://github.com/SupriyaChanda/gilg-ltedi-acl2026.git.

1 Introduction

Language shapes how society thinks and how knowledge is transferred across generations. It inherently carries values, assumptions, and power structures. In the era of large language models (LLMs) such as Claude[1], DeepSeek[2], ChatGPT[3], and Gemini[4], this concern has become even more significant. Since these models are trained on large-scale real-world data that often inherit and reproduce existing societal stereotypes, particularly gender bias. As a result, biased model outputs can influence ideas, perceptions, and broader social systems.Studies (Burns et al., 2019) have shown that normal terms favors men over women like chairman, spokesman etc. Several datasets like `Crows-pair` (Nangia et al., 2020), `Stereoset` (Nadeem et al., 2021) and `WinoBias` (Zhao et al., 2018) are used to address bias and gender related challenges. Evaluation metrics like `Log Probability Bias Score` (Kurita et al., 2019), `Context Association Test` (Nadeem et al., 2021), `discovery of correlations` (Webster et al., 2021) etc have been proposed along with LLM based approaches (Chen et al., 2024) to measure bias. However, studies (Sitaram et al., 2025) indicate that LLMs do not always align with human judgments in bias detection tasks, highlighting the continued need for human oversight. Existing fine-tuning strategies, such as lexical mapping (Bartl and Leavy, 2024) or embedding-level debiasing methods (Bolukbasi et al., 2016), often remain limited because they do not fully account for contextual meaning, dynamic language usage, or deeper reasoning processes. More recently, (Muthusamy Chinnan et al., 2025) proposed a retrieval-augmented generation (RAG) and chain-of-thought (CoT) based approach to address this task. Building upon these ideas, Team `JustGen` presents our findings and methodological approaches for the LT-EDI 2026 shared task on Gender Inclusive Language Generation[5] (Chakravarthi et al., 2026).

1.1 Problem Statement

Let $\mathcal{D} = \{x_i\}_{i=1}^{N}$ be a set of input sentences that may contain gender bias, gender-marked expressions, exclusionary language, or stereotypical claims. The objective is to construct a controlled transformation function $f : x \rightarrow \hat{y}$ that generates an output $\hat{y}$ satisfying task-specific inclusivity constraints while preserving semantic

[1] https://claude.ai/
[2] https://chat.deepseek.com/
[3] https://chatgpt.com/
[4] https://gemini.google.com/app
[5] https://www.codabench.org/competitions/11336/

Proceedings of the Sixth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 193–197

meaning and fluency . For Sub Task A: *Gender-Inclusive Language Generation*, the goal is to transform a gender-biased or gender-marked sentence into a fully gender-neutral alternative. For Sub Task B: *Counterfactual Generation*, the objective is to generate an empathetic and persuasive counter-narrative that challenges an explicitly biased statement. The transformation function f can be described as lexical replacement mapping $g(x)$, context retrieval function $r(x)$ and an instruction-guided language model generation function $h(x, r(x))$. The final output is therefore defined as $\hat{y} = f(x) = h(x, r(x))$, subject to lexical correction and minimal-edit for Subtask A and constructive counterfactual reframing for Subtask B.

2 Dataset

We were provided with training and test datasets[6] by the organizers for Sub Task A and Sub Task B across multiple languages. The datasets include curated gender-neutral word pairs, gender-inclusive sentence pairs, and counterfactual sentence pairs. For Sub Task A, the dataset is available in English, Spanish, German, Tamil, and Kannada. It consists of (i) gender-neutral word replacement pairs and (ii) gender-neutral sentence pairs. For Sub Task B, counterfactual inclusive sentence pairs are provided in English. The dataset statistics are summarized in Table 1.

Table 1: Dataset statistics across tasks and languages

Task	Category	English	German	Spanish	Tamil	Kannada
A	Gender Neutral Word Pairs	673	–	200	742	693
A	Gender Neutral Sentence Pairs	1074	1002	200	1074	1074
B	Counterfactual Sentence Pairs	726	–	–	–	–

3 Methodology

In this section, we describe the methodologies employed for this task. Different approaches were adopted for various language pairs and sub tasks, and their details are presented below.

3.1 Approach 1

We adopted this approach for English in both the subtasks and for German, Spanish languages in SubTask A. Our framework follows a retrieval-augmented, minimally constrained bias-correction pipeline with deterministic pre-processing and edit-distance controlled generation. The system operates in two phases: (1) offline knowledge indexing and (2) online inference with multi-stage correction (See Figure 1). In offline knowledge indexing, we denote the curated knowledge document $D = D_{lex} \cup D_{sent}$ where D_{lex} denotes the set of curated gendered-to-neutral lexical mappings and D_{sent} denotes the set of counterfactual biased-to-inclusive sentence pairs. The document D is segmented into textual chunks $C = \{c_1, c_2, ..., c_n\}$. Each chunk c_i is embedded using a sentence transformer model. For English we use the model `all-MiniLM-L6-v2`[7], while for Spanish and German we employ the multilingual model `paraphrase-multilingual-MiniLM-L12-v2`[8] to better capture cross-lingual semantic similarity. The embedding function can be defined as $v_i = f(c_i)$ where $f(\cdot)$ produces dense semantic vectors. To enable cosine similarity search, L2 normalization is applied $\hat{v}_i = \frac{v_i}{\|v_i\|}$. All normalized vectors $\hat{v}_i$ are stored in a FAISS `IndexFlatIP`[9](Douze et al., 2025) vector index for efficient nearest-neighbor retrieval.In the online inference pipeline, given an input query Q, the system applies a three-stage progressively controlled correction process. Before that the input query is first preprocessed by normalizing blank placeholders. It is then embedded using the same embedding function $q = f(Q)$ and L2-normalized. Top-k relevant chunks are retrieved using cosine similarity $\text{sim}(\hat{q}, \hat{v}_i) = \hat{q} \cdot \hat{v}_i$ and $RC = \text{argmax}^k_{c_i \in C} \text{sim}(\hat{q}, \hat{v}_i)$. Then the retrieved context $RC = \{c_{i_1}, ..., c_{i_k}\}$ is injected into the generation prompt. The retrieved chunks provide lexical mappings and sentence-level inclusive rewrites that guide the language model toward bias-aware generation through in-context learning.

3.1.1 Stage 1: Deterministic Lexical Correction

Before generative rewriting, a lexical bias map $\mathcal{L} = \{(g_1, n_1), (g_2, n_2), ..., (g_m, n_m)\}$is applied using case-insensitive word-boundary matching. The lexical replacement function is $Q' = \text{Replace}(Q, \mathcal{L})$.If $Q' \neq Q$, the system returns: $R_{final} = \text{Sanitize}(Q')$. This deterministic stage ensures minimal modification for purely lexical bias.

[6]The official dataset can be found here https://www.codabench.org/competitions/11336/.

[7]https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[8]https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

[9]https://faiss.ai/cpp_api/struct/structfaiss_1_1IndexFlatIP.html

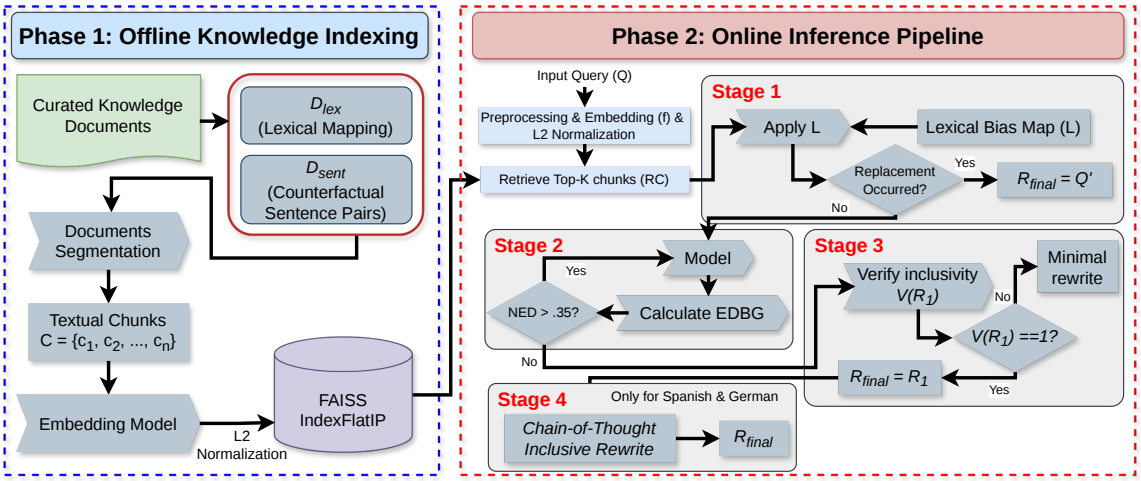

Figure 1: Block Diagram of Approach 1

3.1.2 Stage 2: First-Pass Retrieval-Grounded Generation

If no lexical replacement occurs, a first-pass response is generated using the instruction-tuned language model `mistralai/Mistral-7B-Instruct-v0.2`[10](Jiang et al., 2023). For Spanish and German prompts, generation and verification are performed using `llama-3.1-8b-instant` (Patterson et al., 2022)accessed through the `Groq API`[11] which provides optimized inference for large language models. The generation function is: $R_1 = \mathcal{M}(Q, RC; \theta)$ where $\mathcal{M}$ represents the model and θ represents decoding parameters such as temperature and maximum token length. In English to enforce structural preservation, normalized edit distance is computed $\text{NED}(Q, R_1) = \frac{\text{EditDistance}(Q,R_1)}{\max(|Q|,1)}$, it is called *Edit-Distance Based Guard (EDBG)*. If $\text{NED}(Q, R_1) > \tau$ where $\tau = 0.35$, generation is repeated using stricter decoding. This mechanism constrains semantic drift and prevents excessive rewriting.

3.1.3 Stage 3: Inclusivity Verification and Minimal Correction

The generated response R_1 is evaluated using the same model $\mathcal{M}$ in a verifier prompt: $V(R_1) = 1$ if R_1 is inclusive, and 0 otherwise. If $V(R_1) = 1$ then: $R_{final} = R_1$. Otherwise, a minimal corrective rewrite is triggered: $R_2 = \mathcal{M}(R_1, RC; \theta_{strict})$ where θ_{strict} corresponds to the constrained decoding configuration. The system aims to produce: $R_{final} = \arg\min_R \text{Bias}(R)$.

3.1.4 Stage 4: Chain-of-Thought Inclusive Rewriting

For Spanish and German prompts, an additional corrective reasoning stage is applied when the generated response still contains gender-biased expressions. The model is prompted to perform a structured reasoning process that identifies biased terms and rewrites them using neutral expressions. Formally, $R_{cot} = \mathcal{M}(R_1, RC; \theta_{cot})$ where the model identify gender-biased expressions, replace them with neutral alternatives and remove implicit gender assumptions. The final response is can be described as $R_{final} = (R_{cot})$.

3.2 Approach 2

Low-resource languages such as Tamil and Kannada suffer from limited availability of curated gender-neutral mappings and counterfactual pairs, reducing the effectiveness of RAG-based correction. Sparse vocabulary coverage and insufficient bias-replacement examples increase hallucination risks when using open-source models like LLaMA(Touvron et al., 2023) and Mistral, which have comparatively weaker exposure to inclusivity-aware patterns in these languages. Thus, we adopt a zero-shot prompting strategy using ChatGPT for initial gender-neutral generation and verify the output using a rewrite-based consistency prompt in Google Gemini. No human post-editing was applied in order to avoid manual bias correction. To evaluate semantic fidelity and inclusivity preserva-

[10] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

[11] https://console.groq.com/keys

tion, we employed the Google Translate API[12] as a cross-lingual verification mechanism for randomly selected 4 promts in each language. The zero-shot prompts used for low-resource language generation are provided below.

Prompt 1: ChatGPT Zero-Shot Generation

Your task is to rewrite the following sentence in a gender-neutral and inclusive manner in [TARGET LANGUAGE].
Strict Rules:
1. Preserve the original meaning.
2. Avoid introducing new information.
3. Do not generalize professions.
4. Avoid gender-specific pronouns unless contextually required.
5. Output only the final rewritten sentence.
Input: [INPUT SENTENCE]
Output:

Prompt 2: Google Gemini Rewrite Verification

Rewrite the following sentence in [TARGET LANGUAGE] while strictly preserving its meaning. Ensure the sentence remains gender-neutral and free from bias. Do not add or remove information. Output only the rewritten sentence.
Input: [CHATGPT GENERATED OUTPUT]
Output:

4 Results

The performance of the proposed `JustGen` system for sub task A across multiple languages is presented in Table 2. The evaluation considers three metrics: Gender Accuracy (GA), Gender Neutrality (GN), and Quality of Response (QR) with the final ranking determined based on the overall average score.

Table 2: Results for Sub Task A

Language	GA	GN	QR	Average
English	94.00	94.00	94.00	94.00
German	96.97	93.94	50.00	80.30
Spanish	100.00	100.00	50.00	83.33
Tamil	95.00	95.00	95.00	95.00
Kannada	100.00	100.00	50.00	83.33

Our system demonstrated strong performance across multiple languages in generating gender-inclusive and gender-neutral text. For English, JustGen achieved the highest score of 94.00, securing Rank 1. In Spanish and Kannada, the system shared Rank 1 with an average score of 83.33, while in Tamil it obtained Rank 1 with the highest score of 95.00 among all evaluated languages. For German, JustGen secured Rank 2 with an average score of 80.30, maintaining strong performance across gender agreement and neutrality metrics. These results highlight the robustness of the proposed approach in multilingual gender-inclusive text generation. Table 3 shows the leaderboard results for Subtask B: Counter Narrative Generation. The proposed JustGen system secured Rank 1, sharing the top position with the IGNITERS team, with an average score of 95.83. Our system achieved 95.00 in both Persuasiveness (PR) and Contextual Counter-Narrative Consistency (CCNC), and 97.50 in Quality of Response (QR), demonstrating that the generated responses were persuasive, contextually relevant, and linguistically coherent. Overall, the results highlight the effectiveness of the proposed approach in generating meaningful and empathetic counter-narratives.

Table 3: Results for Sub Task B

Language	PR	CCNC	QR	Average
English	95.00	95.00	97.50	95.83

5 Conclusion

In this work, we addressed gender-inclusive language generation as part of the LT-EDI shared task. We proposed JustGen, a retrieval-augmented framework that combines lexical substitution, semantic retrieval, and controlled generation to transform gender-biased sentences into inclusive alternatives while preserving their original meaning. The framework incorporates edit-distance constraints and self-evaluation to ensure minimal and coherent modifications. Experimental results across multiple languages demonstrate the effectiveness of our approach, achieving competitive performance in the shared task. Future work will focus on improving contextual bias detection and extending the framework to broader multilingual settings.

Limitation

The framework is sensitive to retrieval quality, where irrelevant retrieved context can affect generation accuracy. Additionally, the system does not explicitly model implicit or discourse-level gender bias in multilingual settings. As an example: "The doctor said he will arrive soon." (See the proof[13])

[12] https://translate.google.co.in/?sl=auto&tl=mr&op=translate

[13] https://chatgpt.com/share/6a0213a2-0630-83e9-96e8-b25c5b68fefe

References

Marion Bartl and Susan Leavy. 2024. From 'showgirls' to 'performers': Fine-tuning with gender-inclusive language for bias reduction in LLMs. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 280–294, Bangkok, Thailand. Association for Computational Linguistics.

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. *Preprint*, arXiv:1607.06520.

Kaylee Burns, Lisa Anne Hendricks, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2019. Women also snowboard: Overcoming bias in captioning models. *Preprint*, arXiv:1803.09797.

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.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or LLMs as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327, Miami, Florida, USA. Association for Computational Linguistics.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. The faiss library. *Preprint*, arXiv:2401.08281.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. 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*, 328:114092.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. 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 5356–5371, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

David Patterson, Joseph Gonzalez, Urs Hölzle, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2022. The carbon footprint of machine learning training will plateau, then shrink. *Preprint*, arXiv:2204.05149.

Sunayana Sitaram, Adrian de Wynter, Isobel McCrum, Qilong Gu, and Si-Qing Chen. 2025. A multilingual, culture-first approach to addressing misgendering in llm applications. *Preprint*, arXiv:2503.20302.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. 2021. Measuring and reducing gendered correlations in pre-trained models. *Preprint*, arXiv:2010.06032.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.