IReL_IIT(BHU)@LTEDI 2026: Fine-Tuning Instruction-Tuned Transformers for Gender-Inclusive Rewriting and Counterfactual Bias Mitigation

Anurag Balaji **Arjun Mukherjee** **Krishna Tewari**
Sukomal Pal
Department of Computer Science and Engineering
Indian Institute of Technology (BHU)
Varanasi, India
{anurag.balaji.cse23, arjunmukherjee.rs.cse23, krishnatewari.rs.cse24, spal.cse}@itbhu.ac.in

Abstract

This paper presents our submissions to the LT-EDI@ACL 2026 Shared Task on Gender Inclusive Language Generation. The task focuses on controlled text rewriting that reduces gender bias while keeping the original meaning and fluency intact. We participated in both the subtasks and treated them independently, training separate instances of the instruction-tuned encoder–decoder model on the respective training datasets. Scores are calculated based on averages across different rubrics, including Gender Assumption (GA), Gender Neutrality (GN), and Quality Relevance (QR) for Task A, and Politeness and Respectful (PR), Contextual Counter-Narrative Coherence (CCNC), and Quality and Relevance (QR) for Task B. For Subtask A (Gender-Inclusive Language Generation) in the English dataset, an average score of 43.7917 could be achieved. For Subtask B (Counterfactual Generation), we achieved an average score of 82.6241. Overall, the experiments indicate that full finetuning of instruction-tuned transformers provides an effective way to produce sentence in gender-neutral form and also producing counter-factual sentences for biased one, when each subtask is optimized on its own data.

1 Introduction and Related Work

Large language models (LLMs) based on transformer architectures have significantly improved the performance of natural language processing systems across tasks such as machine translation, summarization, dialogue generation, and question answering. Their ability to produce fluent and contextually coherent text has enabled widespread deployment in applications including conversational assistants, educational platforms, and automated moderation tools (Brown et al., 2020; Bommasani et al., 2021; Gallegos et al., 2024). However, despite their strong capabilities, these models often inherit and amplify social biases present in their training data, raising concerns about fairness, accountability, and responsible deployment of language technologies (Weidinger et al., 2022; Gehman et al., 2020).

Gender bias is among the most widely studied forms of bias in language technologies. Bias may appear through occupational stereotypes, gender-marked role nouns, or implicit gender assumptions generated by language models. For instance, certain professions may be disproportionately associated with a particular gender or models may assume gender even when it is not explicitly specified in the input text. Previous studies have shown that statistical associations embedded in large corpora can encode stereotypical gender relationships that subsequently influence language model behavior and generated outputs (Bolukbasi et al., 2016; Rudinger et al., 2018; Sheng et al., 2019; Nangia et al., 2020).

Mitigating such biases has therefore become an important research direction. Controlled text rewriting approaches aim to transform gender-biased or gender-marked expressions into gender-neutral alternatives while preserving the original semantic meaning and grammatical structure. Transformer-based architectures provide strong contextual representations that enable such transformations while maintaining linguistic coherence (Devlin et al., 2019; Raffel et al., 2020). Instruction-tuned models further extend this capability by enabling models to follow explicit natural language instructions describing desired transformations (Chung et al., 2024; Sanh et al., 2022).

Recent research has also explored reasoning-based and prompting-based techniques for bias mitigation in LLM outputs. Retrieval-augmented frameworks integrate external knowledge sources and structured reasoning mechanisms to guide models toward more inclusive responses (Muthusamy Chinnan et al., 2025). Demonstration-based prompting methods select bias-revealing examples and generate structured reasoning that encourages

Proceedings of the Sixth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 182–187

impartial responses while preserving overall model performance (Qiu et al., 2025). Evaluation frameworks such as the Gender Inclusivity Fairness Index (GIFI) provide rubric-based metrics to measure gender neutrality, assumption avoidance, and response quality in generated text (Shan et al., 2025a).

Shared tasks and benchmark datasets have further accelerated research in inclusive language generation by providing standardized evaluation settings. These tasks focus on transforming gender-biased sentences into gender-inclusive alternatives and generating counterfactual responses that challenge biased statements while maintaining respectful and coherent language (Chakravarthi et al., 2026). Counter-narrative generation is particularly important in online discourse and hate speech mitigation, where respectful responses can help counter harmful narratives and promote constructive dialogue (Sap et al., 2020; Dinan et al., 2019; Dixon et al., 2018).

2 Task Overview

The shared task (Chakravarthi et al., 2026) focuses on the controlled transformation of gender-biased, gender-marked sentences into inclusive sentences while preserving the meaning. Descriptions of the two subtasks are given below.

2.1 Subtask A: Gender Inclusive Language Generation

Subtask A requires rewriting a non-inclusive or gendered sentence into a gender-inclusive version while preserving meaning and fluency. Typical transformations include replacing gender-marked roles and pronouns with gender-neutral alternatives (e.g., **policeman** → **policeperson**, **chairman** → **chairperson**). Although the full shared task includes multiple languages, in this paper we reported results for the English subset.

2.2 Subtask B: Counterfactual Generation

Subtask B targets generation of counterfactual, bias-mitigating responses to gender-biased statements, by focussing on giving a counter response. In our implementation, we fine-tune the model on the provided dataset comprising paired examples of gender-biased sentences and their corresponding counterfactual responses. This setup enables the model to learn a direct mapping from biased inputs to neutral outputs using supervised sequence-to-sequence training.

3 Dataset and Preprocessing

This section describes the datasets for each subtask and preprocessing techniques applied.

3.1 Subtask A (English)

The dataset (Chakravarthi et al., 2026) for Subtask A consists of two components: (i) gender-neutral word pairs (e.g., **ballboy** → **ball person**), and (ii) gender-neutral sentence pairs (e.g., **The fireman responded quickly.** → **The firefighter responded quickly.**).

To expand the available training data, we made additional sentence pairs from the word-level pairs. We generated contextualized sentences by prompting ChatGPT(OpenAI, 2024) to create example sentence pairs based on the provided gendered and gender-neutral word mappings. We gave some examples to guide the process. The word-pair dataset was processed in batches of 100 entries, and the generated sentence pairs were combined with the original sentence-pair dataset. After removing duplicate instances, the final combined dataset contained 1,677 sentence pairs.

We construct train/validation/test splits using a 70/20/10 strategy with train split size as 1173, validation as 336 and test as 168.

The dataset enables controlled rewriting in which the target sentence usually varies from the source by only a few substitutions, while preserving the original underlying statement.

3.2 Subtask B (English)

The Subtask B dataset (Chakravarthi et al., 2026) contains biased input sentences paired with counterfactual responses. We split the total of 726 records as 508 for train split, 145 for validation and 73 for test.

4 Methodology

This section describes the methodology employed.

4.1 Base Model

We used Google/flan-t5-base(Chung et al., 2022), an instruction-tuned encoder–decoder transformer from the T5 family. The source code for our system is publicly available[1]. Instruction tuning is really helpful when you want to control what the model generates. This is because the model is already trained to follow instructions that are written in a

[1] https://github.com/anurag2027/Gender-Inclusive-Language-Generation---LT-EDI-ACL

certain way. So it is easy to make the model do what you want by giving it instructions that are consistent for each task. Instruction tuning makes this process straightforward.

4.2 Input Formatting via Instructions

We cast both subtasks as supervised sequence-to-sequence learning with explicit instructions.

Subtask A Prompt
Rewrite the sentence into gender-inclusive language without changing the meaning:
`<sentence>`

Subtask B Prompt
Rewrite the following sentence to remove bias and produce a counterfactual sentence:
`<sentence>`

Table 1: Instruction prompts used for Subtask A and Subtask B.

This design keeps the interface consistent while letting the model learn task-specific transformations from data.

4.3 Training Procedure

We performed full fine-tuning of all the model parameters separately for both subtasks using the Hugging Face Trainer API. Preliminary experiments were conducted with different hyperparameter configurations, and the following setting was found to provide stable convergence and strong performance. The hyperparameters that gave the best results for each tasks are: maximum epochs as 50, learning rate as 2e-4, batch size as 4 for training and evaluation, evaluation frequency of every 100 steps and optimization objective as cross-entropy sequence loss.

For Subtask A, we employed early stopping with a patience of 3, evaluation steps based on validation loss. As a result, training terminated at epoch 12 when no further improvement was observed.

For Subtask B, the initial training run did not include early stopping, and the model achieved a best validation loss of 0.0087. After the submission phase, we conducted an additional experiment following the same early stopping strategy used in Subtask A. This post-submission experiment yielded an improved best validation loss of 0.0057, indicating that early stopping have better results.

5 Evaluation

Submitted systems are evaluated using a metrics-based framework designed to assess gender-inclusive fairness and response quality. The evaluation follows a hybrid LLM-as-a-Judge methodology with human oversight to ensure consistency and reliability.

5.1 Subtask A Metrics

For Subtask A, we adopt the **Gender Inclusive Fairness Index (GIFI)** framework (Shan et al., 2025b), which measures the effectiveness of bias mitigation while preserving contextual relevance. GIFI consists of three rubric-based components:

- **Gender Assumption (GA):** Measures whether the system avoids implicit or explicit gender assumptions when no gender is specified.
- **Gender Neutrality (GN):** Evaluates whether gendered or non-inclusive terms are replaced with appropriate gender-neutral alternatives.
- **Quality and Contextual Relevance (QR):** Assesses completeness, coherence, and contextual appropriateness of the generated output.

Each component is scored using predefined rubrics and normalized for reporting. The overall performance is computed as the average of the three scores.

5.2 Subtask B Metrics

For Subtask B, counterfactual generation outputs are evaluated using three rubric-based criteria that measure politeness, contextual coherence, and overall response quality:

- **Politeness and Respect (PR):** Evaluates whether the generated counter-narrative maintains a respectful and appropriate tone.
- **Contextual Counter-Narrative Coherence (CCNC):** Measures relevance and coherence with respect to the input statement.
- **Quality Score (QS):** Assesses clarity, readability, persuasiveness, and overall effectiveness of the response.

Scores are reported on a 0–100 scale, and the final performance is computed as the average across the three dimensions.

6 Results

This section the results achieved in both the subtasks.

6.1 Subtask A (English) Test Results

Table 2 reports the performance of the proposed system on the Subtask A English test set using the Gender Inclusive Fairness Index (GIFI) evaluation framework.

Team Name	GA	GN	QR	Average	Rank
JUSTGEN	94.0000	94.0000	94.0000	94.0000	1
CPS	92.5000	92.5000	92.5000	92.5000	2
THE PARITY LAB	92.5000	92.5000	92.5000	92.5000	2
IHLC	80.0000	80.0000	80.0000	80.0000	3
ARJUN	51.5000	90.2500	54.6250	65.4583	4
PRANAV	63.1250	62.5000	63.7500	63.1250	5
IGNITERS	67.5000	70.0000	43.1250	60.2083	6
CAI	65.0000	58.7500	46.8750	56.8750	7
IREL_IIT (BHU)	**43.3750**	**49.0000**	**39.0000**	**43.7917**	**8**

Table 2: Performance on Subtask A evaluated using the Gender Inclusive Fairness Index (GIFI).

The results indicate that the model effectively mitigates gender bias while preserving the semantic coherence of the generated responses. The Gender Neutrality score demonstrates that the system predominantly employs inclusive and gender-neutral language in its outputs.

The Gender Assumption score suggests that the model is largely successful in avoiding unwarranted gender assumptions when gender information is not explicitly provided in the input.

Furthermore, the Quality and Contextual Relevance score indicates that the generated responses are generally coherent, contextually appropriate, and aligned with the intent of the input text. The system generally preserves the original semantic meaning during the rewriting process. Additionally, the model tends to employ gender-neutral role nouns, thereby avoiding language that implicitly favors a particular gender. **Our system ranked 8th on the Subtask A leaderboard.**

6.2 Subtask B Results

Table 3 presents the evaluation results for Subtask B using metrics that assess politeness, contextual coherence, and overall response quality.

The results indicate that the model performs effectively in generating appropriate counter-narrative responses, as reflected in the evaluation scores. The slightly lower Quality Score suggests that while the responses are generally coherent and meaningful, there is some variation in terms of persuasiveness and engagement.

Team Name	PR	CCNC	QR	Average	Rank
IGNITERS	95.0000	95.0000	97.5000	95.8333	1
JUSTGEN	95.0000	95.0000	97.5000	95.8333	1
IREL_IIT (BHU)	**88.7766**	**88.7766**	**70.3192**	**82.6241**	**2**
CPS	89.6809	89.4681	67.1277	82.0922	3
THE PARITY LAB	84.8404	84.8404	66.6489	78.7766	4
PRANAV	85.4255	85.5319	64.9468	78.6348	5
IHLC	84.8936	84.7872	64.6809	78.1206	6

Table 3: Leaderboard results of Subtask B.

Overall, the model demonstrates a strong ability to generate responses that are sensitive to potential biases while maintaining contextual relevance.

Since counterfactual or counter-narrative generation tasks may allow multiple valid responses for a single input, evaluation based solely on surface-level similarity can be inadequate. Therefore, rubric-based evaluation criteria provide a more suitable approach for assessing the quality and appropriateness of the generated responses. **The system achieved competitive performance, ranking 2nd in Subtask B.**

7 Conclusion

This paper presents a simple and effective approach for the LT-EDI @ ACL 2026 Shared Task on Gender Inclusive Language Generation using full fine-tuning of `google/flan-t5-base`.

Our results show that task-specific fine-tuning significantly improves the ability of instruction-tuned models to perform gender-inclusive rewriting and counter-narrative generation. The system achieves competitive performance on **Subtask A under the Gender Inclusive Fairness Index (GIFI) evaluation framework with an average score of 43.79**, and produces high-quality counter-narratives for **Subtask B with an average score of 82.62, particularly in terms of politeness and contextual coherence**.

These findings highlight the effectiveness of instruction-tuned encoder–decoder models for bias-aware text transformation using a simple and reproducible fine-tuning pipeline.

Future work will extend experiments to additional languages such as Spanish and Tamil, explore improved decoding strategies for counterfactual generation, and incorporate richer evaluation methodologies to further enhance response quality and inclusiveness.

8 Limitations

This work has several limitations. First, the proposed system relies solely on full fine-tuning of

FLAN-T5 without incorporating retrieval, reasoning, or constrained decoding mechanisms. As a result, the model may struggle with implicit gender bias and context-dependent stereotypes that require deeper semantic understanding.

Second, the Subtask A training data was expanded using synthetic sentence pairs generated with ChatGPT. Although this increased the amount of training data, synthetic examples may introduce stylistic regularities that affect generalization.

In addition, evaluation is primarily based on rubric-driven automatic scoring, which may not fully capture nuanced fairness and linguistic appropriateness. The comparatively lower performance in Subtask A further suggests that sequence-to-sequence fine-tuning alone is insufficient for robust gender-inclusive rewriting.

Future work will explore multilingual evaluation, retrieval-augmented generation, and reasoning-guided approaches for improved bias mitigation.

9 Ethical Considerations

Some minor text editing assistance was obtained using ChatGPT (OpenAI, 2024). We therefore emphasize careful evaluation, particularly for applications involving sensitive identity-related attributes. Future work will incorporate stronger evaluation protocols, including human review, to ensure that generated outputs remain respectful and contextually appropriate.

References

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Sherry Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, and 95 others. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Bharathi Raja Chakravarthi, Shunmuga Priya, and Paul Buitelaar. 2026. Gender inclusive language generation shared task. https://www.codabench.org/competitions/11336/. LT-EDI @ ACL 2026 Shared Task.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tai, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2024. Scaling instruction-finetuned language models. *J. Mach. Learn. Res.*, 25(1).

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, and 12 others. 2022. Scaling instruction-finetuned language models. *arXiv preprint*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4537–4546, Hong Kong, China. Association for Computational Linguistics.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 67–73, New York, NY, USA. Association for Computing Machinery.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration

in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. 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.

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.

OpenAI. 2024. Chatgpt. `https://chat.openai.com`. Large language model.

Hongye Qiu, Yue Xu, Meikang Qiu, and Wenjie Wang. 2025. Dr.gap: Mitigating bias in large language models using gender-aware prompting with demonstration and reasoning. *arXiv preprint arXiv:2502.11603*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. 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 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Adam Stiegler, Teven Le Scao, Arun Raja, and 1 others. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations (ICLR)*.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Zhengyang Shan, Emily Diana, and Jiawei Zhou. 2025a. Gender inclusivity fairness index (GIFI): A multilevel framework for evaluating gender diversity in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2548–2579, Vienna, Austria. Association for Computational Linguistics.

Zhengyang Shan, Emily Diana, and Jiawei Zhou. 2025b. Gender inclusivity fairness index (GIFI): A multilevel framework for evaluating gender diversity in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2548–2579, Vienna, Austria. Association for Computational Linguistics.

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 the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, and 4 others. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 214–229, New York, NY, USA. Association for Computing Machinery.