IHLC@LT-EDI 2026: Steering Toward Inclusivity - A Representation Engineering for Gender-Neutral Rewriting

Akhil Rajeev P
C-DAC, Bangalore
akhilrajeev@cdac.in

Manoj Balaji Jagadeeshan
Indian Institute of Technology, Kharagpur
manojbalaji1@gmail.com

Abstract

This paper describes the IHLC team's submission to the LT-EDI 2026 shared task on gender-inclusive language generation. For Subtask A (gender-neutral rewriting), we applied Low-Rank Adaptation (LoRA) fine-tuning, achieving an 80.00% evaluation score and placing 3rd. Our primary methodological focus, however, is Subtask B (counter-narrative generation), where we propose a compute-efficient representation engineering approach. We compute a PCA-derived steering direction from counterfactual activations and inject it into the Gemma-3-4B-it model at inference time, shifting behavior toward inclusivity without weight updates. Paired with constrained prompting, this yielded polite, context-aware responses and a 78.12% score (Rank 6). We conclude with a manual evaluation of steering failure modes, detailing critical trade-offs in semantic preservation and over-steering instability.

1 Introduction

Gender-inclusive language generation aims to transform biased or gender-marked sentences into inclusive, gender-neutral, and contextually coherent alternatives while preserving meaning and fluency. The LT-EDI shared task (Chakravarthi et al., 2026) provided parallel resources and a hybrid LLM-as-judge evaluation framework with human oversight to measure both fairness and semantic preservation. The task consisted of two subtasks: (A) Gender Inclusive Language Generation (multilingual; we participated in English only) and (B) Counter Narrative (English only).

While Subtask A was achieved using standard LoRA fine-tuning, this paper focuses on our unconventional methodology for Subtask B. For counter-narrative generation, our IHLC submission builds on an activation-steering method (Turner et al., 2023; Zou et al., 2023) that identifies a steering direction from paired biased/neutral examples (via a difference-of-activations PCA) and injects that vector into a chosen intermediate layer at inference-time (Li et al., 2024; Subramani et al., 2022). This is combined with constrained prompt templates to encourage concise, neutral rewrites. The approach and experiment code were packaged and exported as a Jupyter notebook[1].

2 Related Works

Our system draws on recent advancements in bias mitigation, representation engineering, and automated evaluation frameworks.

Gender-Inclusive Language and Bias Mitigation: The NLP community has long documented the amplification of societal biases in language models (Bolukbasi et al., 2016; Sheng et al., 2019). Efforts to mitigate these biases have ranged from data augmentation and debiasing embeddings (Sun et al., 2019) to rule-based and neural inclusive rewriting (Vanmassenhove et al., 2021). A recent study by Muthusamy Chinnan et al. (2025) combines a curated inclusive-text corpus with a two-pass RAG and Chain-of-Thought prompting to ground and reason about generated text, demonstrating decreased gender bias in both machine and human evaluations.

Activation Steering and Representation Engineering: To adjust model behavior without expensive fine-tuning, we utilize activation steering. Turner et al. (2023) demonstrated that injecting steering vectors into forward passes can reliably control language model outputs. Zou et al. (2023) further formalized this top-down approach, showing how PCA on contrastive activation pairs can identify robust semantic directions. Similar inference-time interventions have been used successfully to alter factual recall (Meng et al., 2022) and adjust model truthfulness (Li et al., 2024).

[1] https://github.com/manojbalaji1/IHLC-Gender-Inclusive

Proceedings of the Sixth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 177–181

Counter Narrative Generation: Generating empathetic responses to hate speech or bias requires navigating a complex trade-off between politeness and firm correction (Qian et al., 2019). Tekiroğlu et al. (2020) and Chung et al. (2021) highlight the importance of generating context-aware, knowledge-grounded counter-narratives rather than simply negating biased statements. Recent approaches also emphasize human-machine collaboration to maintain output quality and relevance in counter-narrative generation (Bonaldi et al., 2022).

Automated Evaluation: Finally, our reliance on the organizers' hybrid evaluation framework aligns with the growing adoption of LLM-as-a-judge paradigms. Zheng et al. (2024) validated that strong LLMs exhibit high agreement with human annotators on qualitative metrics, though our failure analysis confirms that human oversight remains crucial for detecting subtle semantic drift.

3 Shared Task Overview and Evaluation

3.1 Subtasks

Subtask A – Gender Inclusive Language Generation. Rewrite a gendered or biased sentence to a fully inclusive variant (examples: *fireman* → *firefighter*). Training and evaluation data were released for multiple languages; we participated only for English. The English sentence-pair dataset size is reported in the task materials.

Subtask B – Counter Narrative Generation. Generate empathetic, persuasive counter-narratives for overt gender-biased statements (English only). Example: input "Women are not good at math." output: a corrective empathetic counter-narrative.

3.2 Evaluation Metrics (Organizers)

The organizers used a hybrid evaluation approach described in the task documentation: an LLM-as-a-judge operating over fixed rubrics plus spot-checking / adjudication by expert human evaluators. For Subtask A the reported components were:

- **GA**: Gender Assumption removal effectiveness (how well gender assumptions were removed).
- **GN**: Gender Neutrality (use of inclusive terminology and neutral phrasing).
- **QR**: Quality & Relevance (fluency, semantic preservation).
- **Overall Score**: average of GA, GN, and QR (in %).

For Subtask B the reported components were:

- **PR**: Politeness & Respectfulness.
- **CCNC**: Contextual Counter-Narrative Coherence (does the counter-narrative respond coherently to the input context).
- **QR**: Quality & Relevance.
- **Average**: mean of PR, CCNC, and QR (in %).

4 System Description

Subtask A utilized LoRA fine-tuning. For Subtask B, our design emphasizes two complementary components: (1) activation-level steering to bias model behavior toward inclusivity, and (2) strict prompt templates to constrain generated text length and formatting.

4.1 Activation-Steering Module

Focusing on Subtask B, we compute an activation-space steering vector from pairs of biased and neutral sentences ("counterfactual" pairs), adapting the representation engineering protocols described by Zou et al. (2023). Practically:

1. Extract hidden activations at a chosen transformer layer for biased sentences (negatives) and inclusive rewrites (positives).
2. Compute per-example differences and fit PCA to the difference vectors; take the first principal component as the steering direction (Turner et al., 2023).
3. At inference-time register a forward hook on the selected layer that adds a scaled version of the steering vector to hidden states for every token position (or the last token), controlled by a steering coefficient α.

This exact procedure, including implementation details for layer discovery, vector extraction, PCA construction, hook mechanics, and steering strength tuning, is described in our submitted code notebook.

4.2 Prompt Templates and Decoding

We used two prompt templates:

- **DEI prompt (soft)**: instructs the model that it is a DEI rewriting expert, provides soft examples, and asks for a rewrite (useful for flexible, explanatory outputs).
- **Strict prompt (deterministic)**: forces a single-line output with strict rules ("Output ONLY the final sentence.") for evaluation runs to avoid explanatory prefixes that could confuse automatic judges.

We used a mixture of greedy decoding and low-temperature sampling with a repetition penalty. As

noted by Holtzman et al. (2020), text degeneration and looping are common in neural generation; we observed these loops primarily when the steering coefficient α was set too high. The notebook documents recommended steering coefficients (e.g., 0.7–1.5) and anti-repetition settings.

5 Experimental Setup

5.1 Data

We used the English portion of the Subtask A sentence-pairs and Subtask B counterfactual pairs supplied by the organizers. Dataset sizes and task statistics are reported in the shared task documentation.

5.2 Model and Implementation

Our experiments used the Gemma-3-4B-it model (Gemma Team, 2025) (details in the code artifact) with the steering hook and prompt pipeline implemented in PyTorch/HuggingFace. The Gemma 3 family provides a highly capable, lightweight foundation with expanded context windows, making it well-suited for activation-level interventions. The complete generation pipeline and tuning scripts are available in our exported notebook.

5.3 Evaluation

We submitted deterministic, single-sentence rewrites for automatic evaluation. The organizers evaluated submissions using their hybrid LLM-as-judge rubric with human oversight; the reported scores below are the official task scores provided to teams.

6 Official Results (English-only)

Table 1 summarizes the official scores for the IHLC submission (English only), as reported by the shared task organizers.

Task / Metric	IHLC (%)	Rank	N
Task A - GA	80.0000		
Task A - GN	80.0000	3	9 (participants)
Task A - QR	80.0000		
Task A - Average	**80.0000**	**3**	**9**
Task B - PR	84.8936		
Task B - CCNC	84.7872	6	7 (participants)
Task B - QR	64.6809		
Task B - Average	**78.1206**	**6**	**7**

Table 1: Official task scores for IHLC (English).

The above metric definitions and the hybrid evaluation procedure are described in the shared task documentation. The Task A overall score is the mean of GA, GN, and QR; Task B average is the mean of PR, CCNC, and QR.

6.1 Interpretation

- **Task A:** a consistent 80% across GA, GN and QR indicates that our system reliably produced neutral lexical choices and preserved overall meaning for many cases, placing 3rd among 9 participating teams.
- **Task B:** high PR and CCNC scores (≈85%) show the system generated polite, context-aware counter-narratives, but the QR subscore (≈65%) pulled the average down, suggesting occasional semantic drift or reduced relevance in some outputs. This resulted in Rank 6 among 7 participants for the English Counter Narrative subtask.

7 Analysis

We performed trial-error (documented in the artifact) to understand the effect of steering strength, prompt strictness, and decoding strategy on quality:

- **Steering strength:** low values produced weaker neutrality; overly high values caused repetition loops and hallucinations. Recommended operational range: 0.6–1.2 for a good trade-off between neutrality and coherence. See the notebook for extracted diagnostics and examples.
- **Prompt template:** strict one-line prompts improved automated judge scores by reducing undesirable explanatory prefixes that were penalized as irrelevant.
- **Decoding:** low-temperature sampling with repetition penalty prevented the "alalal" style looping behavior observed at high steering.

7.1 Manual Analysis of Failure Modes

We analyze the failure modes of activation steering applied to the Gemma-3-4B-it model, where a PCA-derived direction (computed from last-token hidden-state differences) is injected at layer 16 with coefficient α. Based on a manual evaluation of the 47 generated counterfactuals in our test set, we identified five primary failure modes.

Error Distribution (n=47):

- **Faithfulness–Edit Trade-off:** 53% (25 cases)
- **Layer-Specific Sensitivity (Entanglement):** 23% (11 cases)

- **Residual Bias Leakage:** 11% (5 cases)
- **Over-Steering Instability:** 9% (4 cases)
- **Repetition and Degeneration:** 4% (2 cases)

7.1.1 Residual Bias Leakage

Because steering uses only the final-token representation, bias encoded in earlier contextual tokens may persist. In these instances, the model neutralizes the gendered subject but fails to negate the harmful premise, applying the negative stereotype universally instead.

- **Input:** "Boys are careless"
- **Output:** "People of all gender identities lack awareness and responsibility."
- **Analysis:** The residual negative framing ("lack awareness and responsibility") is retained. The causal structure is simplified rather than corrected, indicating only a partial traversal of the bias manifold.

7.1.2 Over-Steering Instability

Large values of α can cause distributional shifts, moving the model away from neutral rewriting and pushing it toward semantic amplification or normative commentary.

- **Input:** "Women belong only in the kitchen"
- **Output:** "People of all gender identities have diverse interests and deserve autonomy."
- **Analysis:** The model completely abandons the syntactic structure of the input, shifting from a localized counterfactual edit into generalized social commentary regarding autonomy.

7.1.3 Repetition and Degeneration

Strong perturbations occasionally disrupt standard token boundaries or produce token loops, motivating the need for sampling and repetition penalties during decoding.

- **Input:** "Boys are reckless"
- **Output:** "People of all gender identities can exhibitimpulsivenessandheedcaution"
- **Analysis:** The intervention disrupts the tokenizer's spacing logic, resulting in fused words and degenerative textual output. A similar token fusion was observed with the term "exhibitaggression" in other samples.

7.1.4 Faithfulness–Edit Trade-off

Steering frequently increases lexical distance well beyond what is necessary for minimal counterfactual edits. For instance, every generated counterfactual in our set rigidly defaulted to the prefix "People of all gender identities...", significantly altering the original sentence structures.

- **Input:** "Girls are bad at science"
- **Output:** "People of all gender identities can excel in STEM fields."
- **Analysis:** While semantically positive and unbiased, the output completely replaces the target domain "science" with "STEM fields" and discards the minimal edit paradigm. This highlights a persistent trade-off between bias neutralization and faithfulness to the original text.

7.1.5 Layer-Specific Sensitivity

Steering at a single mid-layer (L=16) assumes a linear attribute encoding. However, implicit biases often appear partially entangled with related concepts, limiting precise control without multi-layer intervention.

- **Input:** "Girls are naive"
- **Output:** "People of all gender identities possess innocence and vulnerability."
- **Analysis:** The negative trait "naive" is transformed into the highly related concepts of "innocence and vulnerability." This demonstrates that the steered concept remains heavily entangled with associated semantic clusters, rather than being cleanly neutralized.

8 Conclusions and Future Work

We presented the IHLC submission to the LT-EDI ACL 2026 shared task. While Subtask A used LoRA fine-tuning (80% Avg), Subtask B focused on activation steering and prompt templates. Our system produced polite counter-narratives but occasionally suffered semantic-relevance degradation (78.12% Avg). Code artifacts are in our notebook. Future work entails integrating semantic-preservation constraints (e.g., contrastive loss or reranking), using human-in-the-loop calibration to refine steering directions, and extending the approach to multilingual tracks to investigate culture-specific gender expressions.

Acknowledgements

We thank the organisers for the datasets, rubrics, and reproducible hybrid evaluation pipeline. We also thank Annarao Kulkarni and Dr. Janaki for their invaluable support.

Limitations

Our English-only scope restricts cultural generalizability. Though devised to overcome fine-tuning compute constraints, single-layer steering risks residual bias leakage and requires precise tuning to prevent text degeneration, needing further refinement and post-processing to outpace weight-updating methods like Instruction Finetuning.

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. *Advances in neural information processing systems*, 29.

Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. 2022. Human-machine collaboration for generating counter-narratives. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4279–4292.

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.

Yi-Ling Chung, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Towards knowledge-grounded counter narrative generation for hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 899–914.

Gemma Team. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

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.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 4755–4764.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. *arXiv preprint arXiv:1909.01326*.

Nishant Subramani, Nithya Suresh, and Matthew E Peters. 2022. Extracting latent steering vectors from pretrained language models. *Findings of the Association for Computational Linguistics: ACL 2022*, pages 566–581.

Tony Sun, Andrew Gaut, Tang Shirlyn, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*.

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.

Alex Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Maciej MacDiarmid. 2023. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 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.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Hao, Zhanghao Wu, Joseph E Ba, Hao Zhuang, Zi Lin, Zhuohan Li, Eric Xing, and 1 others. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Pengfei Xia, Darren Lin, Minqi Jiang Wang, Danqi Yin, Mantas Woodside, and 1 others. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.