Insights from Multilingual Gender Inclusive Language Generation Shared Task

Bharathi Raja Chakravarthi[1*], Shunmuga Priya Muthusamy Chinnan[1*], Paul Buitelaar[1], Meghann L. Drury-Grogan[3], Miguel Ángel García Cumbreras[4], Salud María Jiménez Zafra[4], Thomas Mandl[5], Sylvia Jaki[6], Rahul Ponnusamy[1], Anand Kumar M[7], Dhanalakshmi V[8], Bharathi B[9], Premjith B[10], Senthil Kumar B[11], Sathiyaraj Thangasamy[12]

[1]Data Science Institute, University of Galway, Ireland
[3]Atlantic Technological University, Ireland. [4]University of Jaén, Spain
[5]University of Hildesheim, Germany. [6]KU Leuven, Belgium.
[7]NITK Surathkal, India
[8]Pondicherry University, India
[9]Sri Sivasubramaniya Nadar College of Engineering, India
[10]Amrita School of Artificial Intelligence, Coimbatore, Amrita Vishwa Vidyapeetham, India
[11]Velammal Institute of Technology, India
[12]Sri Krishna Adithya College of Arts and Science, India

Abstract

We investigate the role of large language models (LLMs) in promoting gender-inclusive language by evaluating their ability to rewrite biased text and generate counterfactual narratives across multiple languages. We introduce a shared task with two subtasks: gender-inclusive rewriting and counterfactual generation. The task covers five languages English, German, Spanish, Tamil, and Kannada reflecting diverse grammatical gender systems and sociocultural contexts. We release curated word-level and sentence-level datasets to support controlled inclusive generation. A total of 50 teams registered for the shared task, and around 8 teams submitted results. Submissions are evaluated using a hybrid framework combining rubric-based automatic scoring with expert human judgment. Finally, we provide an overview of participating systems and discuss key findings and challenges observed across languages.

1 Introduction

Remarkable capabilities of LLMs including (Open AI GPT (OpenAI, 2023) and Gemini (Gemini Team et al., 2023)) have significantly transformed multiple areas by increasing productivity and are becoming ubiquitous in daily life (Naveed et al., 2025; Zhang et al., 2025). Large Language Models, trained on massive and overlapping data sources, risk exhibiting hivemind-like behavior, where dominant societal norms are reinforced and marginalized voices are underrepresented (Wang et al., 2025; Bartl et al., 2025; Jiang et al., 2025). However, their effectiveness in producing socially aware and inclusive language, particularly across diverse cultural and grammatical contexts, remains an open challenge (Gallegos et al., 2024). Addressing such limitations is critical for creating inclusive AI and developing language technologies that are equitable across gender expressions and cultural contexts. To mitigate such risks, we frame gender-inclusive language generation as a problem that requires explicit social scaffolds, including carefully designed prompts, counterfactual rewriting strategies, and culturally grounded evaluation criteria.

Although LLMs are evolving in their ability to handle bias, they still face challenges, particularly in low-resource language context (Buscemi et al., 2025). Addressing these limitations is essential for creating inclusive AI and developing language technologies that are equitable across gender expressions and cultural contexts. The Gender-Inclusive Language Generation Shared Task was designed to address this gap by encouraging the development of NLP systems capable of rewriting and generating text in a gender-neutral and inclusive manner. Prior work in NLP has leveraged these models for bias detection (Luo et al., 2025; Lin et al., 2025), bias mitigation (Kim et al.,

[*]Equal contribution; joint first authorship.

Proceedings of the Sixth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 128–140

2025; Sun et al., 2019), and counterfactual generation. Recent approaches leverage instruction fine-tuning and prompt-based techniques such as Chain-of-Thought (CoT) reasoning. Rather than simple token replacement, LLMs are guided to produce step-by-step rewrites that preserve meaning while removing gender bias. However, these approaches can still propagate subtle stereotypes or fail in low-resource languages, because the manual translation and paraphrasing of prompts to ensure semantic consistency and cultural appropriateness across languages is both time-consuming and difficult to scale (Buscemi et al., 2025). Frameworks that combine retrieval-augmented generation (RAG) with structured reasoning (chain-of-thought prompting) have been proposed to steer LLMs toward less biased outputs by grounding generation in unbiased reference texts and guided reasoning steps (Muthusamy Chinnan et al., 2025). Such approaches demonstrate that systematic debiasing and contextual reasoning can help reduce gender assumptions and enhance semantic quality in generated text.

Despite progress in bias detection and mitigation (Kantharuban et al., 2025), comparatively less work has focused on evaluating and benchmarking inclusive text generation systematically across multiple languages with diverse grammatical structures and sociocultural gender norms. Gender-Inclusive Language Generation Shared Task focused on leveraging large language models (LLMs) to generate and rewrite text in a gender-neutral and inclusive manner. In particular, we define two subtasks on the basis of (a) gender-inclusive rewriting, which involves transforming gender-marked or exclusionary expressions into neutral alternatives, and (b) counterfactual generation, which requires producing empathetic and persuasive counter-narratives for gender-biased statements. Working with curated word-level and sentence-level datasets across five languages English, German, Spanish, Tamil, and Kannada we present the process of defining the task (Section 3), constructing and annotating the datasets (Section 4), introducing our evaluation framework combining rubric-based metrics with expert human judgment (Section 5), and summarizing the approaches, results, and challenges observed across languages (Section 6).

In this overview paper we make the following contributions:

- We introduce two tasks for gender-inclusive language generation: gender-inclusive rewriting and counterfactual generation.
- We describe the process of dataset creation and annotation across five languages: English, German, Spanish, Tamil, and Kannada.
- We provide an overview of the approaches followed by the participating teams, our hybrid evaluation framework combining rubric-based metrics with expert human judgment, and a summary of the results and key challenges observed.

2 Task Definition

The shared task focuses on gender-inclusive language generation and consists of two subtasks. Given a gender-biased input sentence, systems are required to produce outputs that promote inclusivity while remaining linguistically and semantically appropriate.

2.1 Gender-Inclusive Rewriting and Generation

In this subtask, systems are provided with sentences containing gender-marked or exclusionary expressions. The goal is to rewrite the input sentence into a gender-inclusive form while preserving its original meaning, tone, and intent.

2.2 Counterfactual Generation

In this subtask, systems are given gender-biased statements expressing stereotypes or normative claims. The objective is to generate empathetic and persuasive counterfactual narratives that challenge the bias in the input. Unlike inclusive rewriting, this task allows systems to introduce new content, provided it remains respectful, coherent, and relevant to the original statement. Table 1 illustrates example inputs and expected outputs for both subtasks.

3 Dataset Description

The dataset was initially curated in English, covering both gender-inclusive rewriting and counterfactual generation scenarios (Muthusamy Chinnan et al., 2025). To support multilingual, the English dataset was translated into four additional languages including German, Spanish, Tamil, and Kannada using automatic machine translation via GoogleTranslate as shown in Figure 1. Following automatic translation, native speakers and experts of each target language manually verified and

Subtask	Input	Output
Gender-Inclusive Rewriting	Each doctor must discuss *his* diagnosis with the patient.	Each doctor must discuss *their* diagnosis with the patient.
Counterfactual Generation	Women are not good leaders.	Leadership ability is not determined by gender. Many women have demonstrated effective leadership across organizations, communities, and nations.

Table 1: Example inputs and outputs for the two subtasks in the gender-inclusive language generation task.

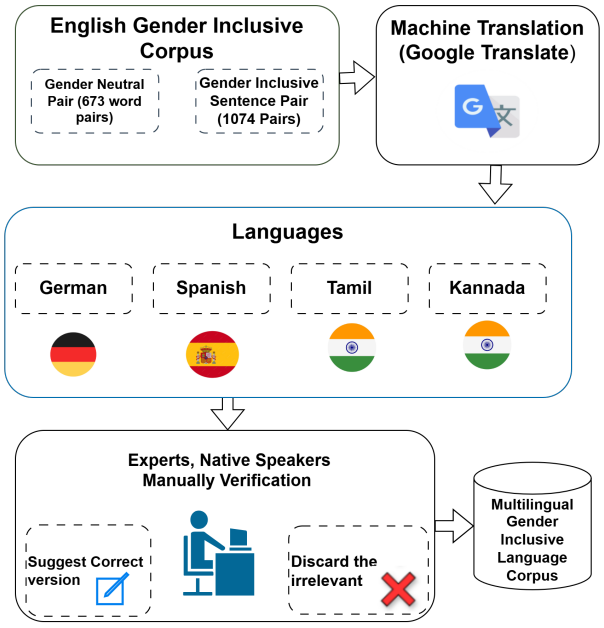

Figure 1: Pipeline for creating a Multilingual Gender Inclusive Language Corpus.

corrected the translated sentences to address translation inaccuracies, grammatical issues, and logical inconsistencies. For each target language, one native-speaking expert with NLP expertise manually verified and corrected the translated sentences. Since each language was reviewed by a single expert annotator, inter-annotator agreement was not computed. We acknowledge this as a limitation and plan to incorporate multiple annotators and agreement analysis in future iterations of the shared task. Translated sentences that were seen irrelevant, culturally inappropriate, were discarded entirely to ensure the overall quality of the multilingual dataset. Even with automated translation and model-based rewrites, expert human validation is crucial for handling gender-neutral cases. Certain terms may appear neutral in one language but carry implicit gender connotations or grammatical nuances in another. For example, while "Chairperson" in English and "Mel Athigari" in Tamil are neutral, other roles or titles may require careful contextual interpretation to avoid bias. Human verification ensures that inclusive language rules are correctly applied, maintains semantic accuracy, and respects cultural and linguistic cues that automated systems alone may not capture. Table 2 summarizes the dataset statistics.

3.1 Subtask A: Gender-Inclusive Language Generation

For Subtask A, two resources were provided:

- **Gender-Neutral Word Pairs:** This resource contains gender-marked terms paired with their inclusive alternatives (*actor* → *performer*). These pairs are intended to support lexical substitution and vocabulary-level neutralization.
- **Gender-Neutral Sentence Pairs:** This dataset consists of sentence-level examples where gendered or biased input sentences are aligned with inclusive rewrites. Table 3 shows few instances of data.

The test set comprised two types of test cases: de-

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

Table 2: Dataset statistics for the Gender Inclusive Language Generation shared task.

Language	Non-Inclusive	Inclusive	Category
German	Der Feuerwehrmann reagierte schnell. *(The fireman responded quickly.)*	Der Feuerwehrkraft reagierte schnell. *(The firefighter responded quickly.)*	Occupational
Spanish	Una enfermera siempre debe ser amable; ella se preocupa de sus pacientes. *(A nurse should always be kind; she cares for her patients.)*	El personal de enfermería siempre debe ser amable; se preocupa de sus pacientes. *(A nurse should always be kind; they care for their patients.)*	Pronoun Assumption
Tamil	நிறுவனத்தின் பெண் பேச்சாளர் புதிய கொள்கையை அறிவித்தார். *(The spokeswoman for the company announced the new policy.)*	நிறுவனத்தின் செய்தித் தொடர்பாளர் புதிய கொள்கையை அறிவித்தார். *(The spokesperson for the company announced the new policy.)*	Gender-Neutral
Kannada	ಮಹಿಳೆಯರು ಸ್ವಾಭಾವಿಕವಾಗಿ ಪುರುಷರಿಗಿಂತ ಹೆಚ್ಚು ತಾಳ್ಮೆಯಿಂದಿರುತ್ತಾರೆ. *(Women are naturally more patient than men.)*	ತಾಳ್ಮೆಯು ವ್ಯಕ್ತಿಗಳಲ್ಲಿ ಬದಲಾಗುತ್ತದೆ ಮತ್ತು ಲಿಂಗದಿಂದ ನಿರ್ಧರಿಸಲ್ಪಡುವುದಿಲ್ಲ. *(Patience varies among individuals and is not determined by gender.)*	Gender Stereotype

Table 3: Examples of Non-Inclusive and Inclusive Language Across Languages

scriptive sentences, which are fully formed prompts requiring gender-neutral description generation, and fill-in-the-blank sentence, which contain explicit placeholders (___) that systems must complete with a gender-neutral term before generating the description. Table 4 illustrates both types.

Type	Example
Descriptive	*Input:* "The nurse enters the hospital ward. Describe their routine and responsibilities." *Output:* "The healthcare worker enters the hospital ward. They check patients' vital signs, administer medications, and document findings in patients' charts."
Fill-in-the-blank	*Input:* " _____ is a skilled nurse in a busy hospital. Describe their daily tasks." *Output:* "The healthcare professional is a skilled nurse. They administer medications, monitor vital signs, and collaborate with other healthcare professionals."

Table 4: Examples of descriptive and fill-in-the-blank test case types in Subtask A.

3.2 Subtask B: Counterfactual Sentence Generation

Subtask B focuses on generating empathetic counter-narratives for gender-biased content. For this subtask, counterfactual sentence pairs were released exclusively in English. Each pair consists of a biased or stereotypical statement and a corresponding inclusive counter response designed to counter the underlying assumption.

4 Participant Methodology

A total of eight teams participated across the two subtasks, exploring a diverse range of modeling strategies for gender-inclusive rewriting and counterfactual generation. Tables 5 and 6 provide a detailed comparison of techniques.

Encoder–Decoder fine-tuning. Several teams adopted supervised sequence-to-sequence fine-tuning using encoder–decoder architectures. Pranav and IReL_IIT(bhu) fine-tuned FLAN-T5-base on paired biased–inclusive examples for Task 1 and counterfactual generation for Task 2. Igniters and cai@tkmce fine-tuned mT5 in a multilingual setting, leveraging cross-lingual transfer for low-resource languages such as Tamil and Kannada.

Instruction tuning and prompt-based adaptation. Instruction tuning and prompt engineering were applied either standalone or with fine-tuning. Pranav, Igniters, and Ihlc used short prompts during training and richer instructional prompts at inference time to improve output alignment and quality.

Team Name	Base Model / LLM	Technique									Language				
		Seq2Seq FT	Instr. Tuning	Prompt Eng.	PEFT	Retrieval / Examples	Bias-aware Rewriting	Post-processing	Data Augmentation	Ctrl. Decoding	English	German	Spanish	Tamil	Kannada
Pranav	FLAN-T5-base	✓	✓	✓			✓				✓				
Igniters (S et al., 2026)	mT5	✓	✓	✓			✓				✓	✓	✓	✓	✓
IHLC (P and Jagadeeshan, 2026)	Gemma (Adapters)		✓	✓	✓						✓				
TheParityLab	Seq2Seq + PEFT	✓	✓	✓	✓		✓	✓			✓	✓	✓	✓	✓
IReL_IIT(BHU) (Mukherjee et al., 2026)	FLAN-T5-base	✓	✓	✓					✓		✓				
JustGen (Adhikary et al., 2026)	RAG-based LLM	✓				✓	✓	✓	✓		✓	✓	✓	✓	✓
CAI@TKMCE (Nair et al., 2026)	mT5	✓	✓	✓			✓	✓		✓	✓	✓	✓	✓	✓
cps (Rawat et al., 2026)	Qwen-2.5		✓	✓	✓		✓		✓		✓	✓	✓	✓	✓

Table 5: Techniques and language coverage used by participating teams for Gender-Inclusive Language Generation (Task 1).

Team Name	Base Model / LLM	Technique								
		Seq2Seq FT	Instr. Tuning	Prompt Eng.	PEFT	Retrieval / Examples	Bias-aware Rewriting	Post-processing	Data-augmentation	Ctrl. Decoding
Pranav	FLAN-T5-base	✓	✓	✓			✓			✓
Igniters (S et al., 2026)	mT5	✓		✓			✓	✓		
IHLC (P and Jagadeeshan, 2026)	Not specified			✓						
TheParityLab	Classical + ML models			✓			✓	✓		✓
TheParityLab	Neural text model	✓		✓		✓	✓	✓	✓	✓
cps (Rawat et al., 2026)	Qwen-2.5		✓	✓	✓		✓			
IReL_IIT(BHU) (Mukherjee et al., 2026)	FLAN-T5-base	✓	✓	✓						
JustGen (Adhikary et al., 2026)	Lightweight rewrite	✓				✓	✓	✓	✓	

Table 6: Techniques used by participating teams for Counter Narrative Generation(Task 2).

The team cps employed Qwen-2.5 with instruction tuning and prompt structuring to guide task-specific responses efficiently.

Parameter-efficient fine-tuning. Adapter-based or low-rank adaptation methods were used to reduce computational overhead while leveraging large LLMs. Ihlc relied on Gemma adapters, cps applied QLoRA to Qwen-2.5, allowing effective specialization without full model fine-tuning.

Retrieval-augmented and bias-aware pipelines. Some teams incorporated retrieval or bias-aware strategies for safer generation. JustGen team combined initial LLM-based rewrites with retrieved examples from curated datasets, refining outputs through automated post-processing. TheParityLab team used a bias-aware pipeline combining auxiliary datasets, prompt normalization, and post-processing to improve inclusivity and reliability.

Data augmentation and controlled decoding. Several submissions expanded training data or applied constrained decoding. IReL_IIT(bhu) team augmented sentence pairs using ChatGPT to create additional training examples. Cai@tkmce applied controlled decoding and gradient-optimized training strategies to maintain quality. TheParityLab and JustGen teams similarly used data augmentation and controlled decoding to improve generalization.

Multilingual handling. Multilingual generalization was addressed through joint multilingual training and language-conditioned prompts. Igniters, Cai@tkmce, and TheParityLab teams relied on

shared multilingual representations. Pranav and JustGen teams applied language-specific prompts or retrieval steps to handle low-resource languages effectively.

Overall, participant approaches reflect three dominant methodological orientations: (1) supervised encoder–decoder fine-tuning, (2) parameter-efficient and instruction-based adaptation of large language models, and (3) retrieval-augmented or bias-aware pipelines. Prompt engineering emerged as a cross-cutting strategy, critical for effective inclusive language generation across multiple languages.

5 Evaluation Metrics

Submitted systems are evaluated using a hybrid LLM-as-a-Judge framework with expert human oversight. A temperature setting of 0 is used to ensure deterministic outputs. The evaluation focuses on both gender-inclusive fairness and semantic quality, using fixed, rubric-based scoring schemes to ensure consistency and reproducibility.

5.1 Task 1: Gender Inclusive Fairness Index (GIFI)

We use the Gender Inclusive Fairness Index (GIFI) to assess the quality of inclusive language generation. GIFI consists of three complementary dimensions: Gender Assumption (GA), Gender Neutrality (GN), and Quality & Contextual Relevance (QR). Each dimension is scored independently using predefined rubrics. Table 7 presents the GA rubric, Table 8 the GN rubric, and Table 9 the QR rubric.

Score	Criterion
0	Explicit gender assumption introduced in the output
1	Mixed or ambiguous gender references present
2	No gender assumption; fully gender-neutral expressions used

Table 7: Rubric for Gender Assumption (GA).

Score	Criterion
0	Gendered or non-inclusive terms retained
1	Appropriate gender-neutral or inclusive terms applied

Table 8: Rubric for Gender Neutrality (GN).

Score	Criterion
0	Incomplete, incoherent, or irrelevant output
1	Partially complete and moderately relevant output
2	Complete, coherent, and contextually relevant output

Table 9: Rubric for Quality and Contextual Relevance (QR).

5.2 Task 2: Counter-Narrative Evaluation

Task 2 submissions are evaluated using three rubric-based scores, each on a 0–100 scale, designed to assess the politeness, coherence, and overall quality of generated counter-narratives. Table 10 presents the PR rubric, Table 11 the CCNC rubric, and Table 12 the QS rubric.

Score Range	Criterion
0–33	Minimal or no polite/respectful language; may include offensive or inappropriate tone
34–66	Attempts polite and respectful framing, but tone is inconsistent or partially unclear
67–100	Clearly and consistently polite and respectful throughout

Table 10: Rubric for Politeness and Respectful Score (PR).

Score Range	Criterion
0–33	Off-topic, incoherent, or fails to address the context of the harmful speech
34–66	Partially coherent or contextually relevant, but lacks clarity or consistency
67–100	Clearly coherent, fully relevant, and context-aware throughout

Table 11: Rubric for Contextual Counter-Narrative Coherence Score (CCNC).

5.3 Human Validation

A team of expert human evaluators performs spot checks, resolves ambiguous cases, and validates final scores to ensure alignment with inclusive language principles and context-sensitive fairness.

6 Results Discussion and Implications

Table 13 and 14 shows the scorecard of task 1 and 2 respectively.

6.1 Task 1: Gender-Inclusive Language Generation

Across all five languages, JustGen emerged as the most consistent top performer, ranking first in English (94.00%), Tamil (95.00%), and Kannada (83.33%), and second in German (80.30%)

Score Range	Criterion
0–33	Poorly written, confusing, or ineffective; includes empty or missing output
34–66	Adequate; conveys the intended message but could be clearer or more impactful
67–100	High-quality; clear, compelling, well-structured, and effective

Table 12: Rubric for Quality Score (QS).

and Spanish (83.33%). This demonstrates that its retrieval-augmented rewriting pipeline, which combines LLM-based rewriting with curated example retrieval via FAISS, generalizes well across typologically diverse languages. The reliability of example-guided generation appears to be a key advantage over finetuning approaches. CPS and TheParityLab were strong competitors, frequently tying for first or second place in German (83.33%) and Spanish (83.33%), and performing competitively in English. CPS's use of QLoRA-based instruction tuning on Qwen-2.5, combined with synthetic data augmentation for low-resource languages, proved highly effective. TheParityLab's bias-aware preprocessing and stratified validation also contributed to robust generalization.

A notable pattern emerges when comparing high-resource vs.low-resource languages. Teams achieved their highest scores in English, but performance dropped considerably in German and Spanish, particularly on the Quality & Contextual Relevance (QR) dimension, where many teams scored only 50.00% suggesting that models struggle to maintain contextual coherence when rewriting in morphologically richer languages. Igniters performed competitively in Tamil (92.07%) due to its joint multilingual mT5 training, while Cai@tkmce also mT5-based consistently underperformed across all languages, pointing to the critical role of data augmentation and prompt design beyond architecture alone.

The Gender Neutrality (GN) dimension was generally the easiest to satisfy, with top teams achieving near-perfect scores, while Quality & Contextual Relevance (QR) was the most challenging dimension. This gap indicates that systems can often substitute inclusive terms at the lexical level but struggle to maintain overall coherence and contextual fidelity, highlighting the need for more sophisticated generation strategies beyond simple lexical substitution. IReL_IIT(bhu) ranked last in English (43.79%), despite using a data augmentation strategy with ChatGPT-generated sentences. This suggests that noisy synthetic data can degrade model performance, and implies the importance of quality-controlled augmentation pipelines.

6.2 Task 2: Counter-Narrative Generation

For Task 2, Igniters and JustGen tied for first place with an average score of 95.83%, both achieving 95.00% on Politeness (PR) and Contextual Counter-Narrative Coherence (CCNC), and 97.50% on Quality Score (QS). This is particularly noteworthy since Igniters used a relatively lean mT5-based approach, while JustGen's retrieval-augmented design again proved effective by grounding generation in curated inclusive examples.The remaining teams clustered closely in the 78–83% range. IReL_IIT(bhu) (82.62%) and CPS (82.09%) performed similarly, while TheParityLab, Pranav, and Ihlc all scored between 78 to 79%, suggesting that once a baseline of politeness and coherence is achieved, further differentiation in quality becomes increasingly difficult. Notably, the QR (Quality) dimension showed the greatest variance across teams, indicating it is the hardest criterion to optimize for in counter-narrative generation. Ihlc relatively modest performance (78.12%), despite its adapter-based instruction tuning on Gemma, may reflect insufficient task-specific fine-tuning data or less expressive prompt templates compared to competitors.

6.3 Human Validation: Human-in-the-Loop Analysis

While the rubric-based LLM-as-a-Judge framework captures qualities such as politeness, coherence, and relevance, expert human evaluation revealed a critical limitation not reflected in the automated scores: a lack of response diversity across test cases. Human evaluators observed that the majority of teams produced counter-narratives that converged on a single structural template of the form:

> *"People of all gender identities can / are capable of [activity/trait]."*

This pattern, while technically inclusive and polite, fails to counter the test cases distinctly. The rubric-based scores (PR, CCNC, QS) reward outputs that are polite, coherent, and relevant, however, they do not penalise response homogeneity.

Human evaluation of Task 1 outputs revealed a critical error where multiple teams left blanks unfilled in sentences that explicitly required pronoun

Table 13: Score Card for Task 1: Gender Inclusive Language Generation - Overall Scores (Average of Gender Assumption (GA), Gender Neutrality (GN), and Quality Relevance (QR)) in % (LLM as a Judge with detailed rubrics under human oversight)

Language	Team Name	GA	GN	QR	Average	Rank
English	JUSTGEN (Adhikary et al., 2026)	94.0000	94.0000	94.0000	94.0000	1
	CPS (Rawat et al., 2026)	92.5000	92.5000	92.5000	92.5000	2
	THE PARITY LAB	92.5000	92.5000	92.5000	92.5000	2
	IHLC (P and Jagadeeshan, 2026)	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 (S et al., 2026)	67.5000	70.0000	43.1250	60.2083	6
	CAI (Nair et al., 2026)	65.0000	58.7500	46.8750	56.8750	7
	IREL_IIT (BHU) (Mukherjee et al., 2026)	43.3750	49.0000	39.0000	43.7917	8
German	CPS (Rawat et al., 2026)	100.0000	100.0000	50.0000	83.3333	1
	THE PARITY LAB	100.0000	100.0000	50.0000	83.3333	1
	JUSTGEN (Adhikary et al., 2026)	96.9697	93.9394	50.0000	80.3030	2
	IGNITERS (S et al., 2026)	69.6970	72.7273	9.0909	50.5051	3
	CAI (Nair et al., 2026)	17.6471	26.4706	0.0000	14.7059	4
Spanish	CPS (Rawat et al., 2026)	100.0000	100.0000	50.0000	83.3333	1
	JUSTGEN (Adhikary et al., 2026)	100.0000	100.0000	50.0000	83.3333	1
	THE PARITY LAB	100.0000	100.0000	50.0000	82.5555	2
	IGNITERS (S et al., 2026)	97.5000	100.0000	47.5000	81.6667	3
	CAI (Nair et al., 2026)	62.5000	70.0000	5.0000	45.8333	4
Tamil	JUSTGEN (Adhikary et al., 2026)	95.0000	95.0000	95.0000	95.0000	1
	THE PARITY LAB	90.9460	95.0000	92.2973	92.7478	2
	IGNITERS (S et al., 2026)	95.0000	93.6487	87.5676	92.0721	3
	CPS (Rawat et al., 2026)	63.2258	83.3871	71.2903	72.6344	4
	CAI (Nair et al., 2026)	45.2703	54.7297	52.0270	50.6757	5
Kannada	JUSTGEN (Adhikary et al., 2026)	100.0000	100.0000	50.0000	83.3333	1
	THE PARITY LAB	100.0000	100.0000	50.0000	83.3333	1
	CPS (Rawat et al., 2026)	95.8333	91.6667	41.6667	76.3889	2
	IGNITERS (S et al., 2026)	96.0000	96.0000	32.0000	74.6667	3
	CAI (Nair et al., 2026)	100.0000	92.0000	0.0000	64.0000	4

insertion. Consider the following example from the test set:

> *"A company hires _____ as a senior manager after a competitive selection process."*

The expected output requires the system to fill the blank with a gender-neutral pronoun (*them* or *the candidate*), thereby producing a grammatically coherent and inclusive sentence. However, several teams returned outputs in which the blank was left empty or the placeholder token was retained verbatim, resulting in semantically incomplete sentences. This error was not captured by the rubric-based scores, as the GN dimension cannot distinguish between a blank that is absent because it has been neutralised and one that is simply left unfilled. Table **??** summarizes the three most frequent error types identified during human validation, along with the discrepancy between LLM judge ratings and human ratings.

Table 14: Score Card fotTask 2: Counter Narrative Generation - Average Scores (Average of Politeness and Respectful (PR), Contextual Counter-Narrative Coherence (CCNC), and Quality and Relevance (QR)) in % (LLM as a Judge with detailed rubrics under human oversight)

Language	Team Name	PR	CCNC	QR	Average	Rank
English	IGNITERS (S et al., 2026)	95.0000	95.0000	97.5000	95.8333	1
	JUSTGEN (Adhikary et al., 2026)	95.0000	95.0000	97.5000	95.8333	1
	IREL_IIT (BHU) (Mukherjee et al., 2026)	88.7766	88.7766	70.3192	82.6241	2
	CPS (Rawat et al., 2026)	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 (P and Jagadeeshan, 2026)	84.8936	84.7872	64.6809	78.1206	6

6.4 Insights

Analysis across both tasks shows that systems are generally good at surface-level inclusivity, such as replacing gendered expressions and maintaining politeness. The main challenge lies in preserving contextual meaning, completeness, and diversity. Top performers like JustGen benefit from retrieval-augmented generation, which grounds outputs in curated examples and reduces errors. cps demonstrates that parameter-efficient instruction tuning with synthetic data can also be effective, though output quality depends on the data used. Gender Neutrality and Politeness are the easiest criteria to satisfy, while Quality and Contextual Relevance remain difficult. Systems often produce neutral and polite outputs that are generic, incomplete, highlighting the limits of simple word-level substitutions.

Performance is relatively better in English when compared with other four languages (German, Spanish, Tamil, and Kannada) showing that multilingual training alone is insufficient without careful data augmentation or prompt design. In Task 2, many systems generate similar counter-narrative templates, which ensures coherence but reduces diversity. These observations suggest that future work should focus on context-aware, meaning-preserving generation, with stronger example grounding, careful synthetic data control, and evaluation metrics that reward complete and diverse outputs.

7 Conclusion

The Shared Task on Gender-Inclusive Language and Counter-Narrative Generation attracted participants from eight research teams, exploring a variety of approaches across five languages: English, German, Spanish, Tamil, and Kannada. The task represents an important step toward creating a multilingual benchmark for inclusive language generation and bias mitigation in text. Participating systems achieved strong performance on rewriting gender-marked sentences and generating counter-narratives while maintaining fluency and semantic coherence.

Ethical Considerations

This shared task was designed to promote fairness and inclusivity in language generation. While we took care to ensure that the datasets and annotations reflect gender-neutral principles, personal assumptions of annotators may still influence what is considered inclusive, especially across diverse languages. Automated evaluation using LLMs as judges carries its own risk, as these models may reward outputs that appear neutral on the surface but miss deeper issues, as confirmed by our human validation. We recommend that any system built on this task be used with human oversight rather than deployed automatically. The datasets contain no personal information, and all participating teams were kept anonymous during evaluation.

Acknowledgments

This work was conducted with the financial support from Research Ireland under Grant Number SFI/12/RC/2289_P2(Insight_2), Research Ireland Centre for Research Training in Artificial Intelligence under Grant No. 18/CRT/6223; and a grant from the College of Science and Engineering, University of Galway, Ireland. The research work of Miguel Ángel García-Cumbreras and Salud María Jiménez-Zafra is part of the ALIA Model

Development project, funded by the Ministry for Digital Transformation and the Civil Service and by the Recovery, Transformation, and Resilience Plan – funded by the European Union – NextGenerationEU. It is also part of the CONSENSO Project (PID2021-122263OB-C21) and the SocialTox Project (PDC2022-133146-C21), funded by MCIN/AEI/10. 13039/501100011033 and by the European Union NextGenerationEU/PRTR, the ROMANET Project (CERV-2024-CHAR-LITI-101215052), funded by the European Union under the Citizens, Equality, Rights, and Values program, the HEART-NLP-UJA project (PID2024-156263OB-C21) and the VERITAS-H project (AIA2025-163322-C64), funded by MICIU/AEI/10.13039/501100011033 and by the ERDF/EU.

References

Nilendu Adhikary, Supriya Chanda, and Sukomal Pal. 2026. JustGen@LT-EDI 2026: Controlled Gender Inclusive and Bias-Aware Language Generation using LLMs. In *Proceedings of the Sixth Workshop on Language Technology for Equality, Diversity, Inclusion*. Association for Computational Linguistics.

Marion Bartl, Abhishek Mandal, Susan Leavy, and Suzanne Little. 2025. Gender bias in natural language processing and computer vision: A comparative survey. *ACM Comput. Surv.*, 57(6).

Alessio Buscemi, Cédric Lothritz, Sergio Morales, Marcos Gomez-Vazquez, Robert Claris'o, Jordi Cabot, and German Castignani. 2025. Mind the language gap: Automated and augmented evaluation of bias in llms for high- and low-resource languages. *ArXiv*, abs/2504.18560.

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.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 208 others. 2023. Gemini: A family of highly capable multimodal models.

Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, Alon Albalak, and Yejin Choi. 2025. Artificial hivemind: The open-ended homogeneity of language models (and beyond). *Preprint*, arXiv:2510.22954.

Anjali Kantharuban, Jeremiah Milbauer, Maarten Sap, Emma Strubell, and Graham Neubig. 2025. Stereotype or personalization? user identity biases chatbot recommendations. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24418–24436, Vienna, Austria. Association for Computational Linguistics.

Taeyoun Kim, Jacob Mitchell Springer, Aditi Raghunathan, and Maarten Sap. 2025. Mitigating bias in RAG: Controlling the embedder. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18999–19024, Vienna, Austria. Association for Computational Linguistics.

Luyang Lin, Lingzhi Wang, Jinsong Guo, and Kam-Fai Wong. 2025. Investigating bias in LLM-based bias detection: Disparities between LLMs and human perception. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10634–10649, Abu Dhabi, UAE. Association for Computational Linguistics.

Xuan Luo, Jing Li, Zhong Wenzhong, Geng Tu, and Ruifeng Xu. 2025. Large language models as reader for bias detection. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 17957–17967, Suzhou, China. Association for Computational Linguistics.

Arjun Mukherjee, Krishna Tewari, Anurag Balaji, and Sukomal Pal. 2026. IReL_IIT(BHU)@LTEDI 2026: Fine-Tuning Instruction-Tuned Transformers for Gender-Inclusive Rewriting and Counterfactual Bias Mitigation. 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*, 328:114092.

Aiswarya P Nair, Sree S Bhagya, and Chinnu Jacob. 2026. CAI@LTEDI 2026: Multilingual Gender Inclusive Language Generation using Instruction-Guided Transformer Model mT5. In *Proceedings of the Sixth Workshop on Language Technology for Equality, Diversity, Inclusion*. Association for Computational Linguistics.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2025. A comprehensive overview of large language models. *ACM Trans. Intell. Syst. Technol.*, 16(5).

OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774.

Akhil Rajeev P and Manoj Balaji Jagadeeshan. 2026. IHLC@LT-EDI 2026: Steering Toward Inclusivity - A Representation Engineering for Gender-Neutral Rewriting. In *Proceedings of the Sixth Workshop on*

Language Technology for Equality, Diversity, Inclusion. Association for Computational Linguistics.

Harsh Rawat, Nitisha Aggarwal, Geetika Jain Saxena, Amit Pundir, and Sanjeev Singh. 2026. CPS@LT-EDI 2026: Parameter-Efficient Fine-Tuning of Qwen2.5-7B for Gender-Inclusive Language Generation. In *Proceedings of the Sixth Workshop on Language Technology for Equality, Diversity, Inclusion*. Association for Computational Linguistics.

Rajendran S, Ramkumar N, and Malarselvi R. 2026. Igniters@LTEDI 2026: Multilingual Gender-Inclusive Language Generation with mT5 and Counter-Narrative Generation Using Llama-3. In *Proceedings of the Sixth Workshop on Language Technology for Equality, Diversity, Inclusion*. Association for Computational Linguistics.

Tony Sun, Andrew Gaut, Shirlyn Tang, 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. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Chenglong Wang, Haoyu Tang, Xiyuan Yang, Yueqi Xie, Jina Suh, Sunayana Sitaram, Junming Huang, Yu Xie, Pengjun Zhao, Zhaoya Gong, Xing Xie, and Fangzhao Wu. 2025. Uncovering inequalities in new knowledge learning by large language models across different languages. *Proceedings of the National Academy of Sciences*, 122(51):e2514626122.

Y. Zhang, S.A. Khan, A. Mahmud, and et al. 2025. Exploring the role of large language models in the scientific method: from hypothesis to discovery. *npj Artificial Intelligence*, 1:14.

A Example Appendix

Table 15: Tamil GIFI Scores by LLM (Task 1)
(CAI@TKMCE vs. TheParityLab)

Test Case	CAI@TKMCE			TheParityLab		
	GA (0–2)	GN (0–1)	QR (0–2)	GA (0–2)	GN (0–1)	QR (0–2)
1	0	0	1	2	1	2
2	0	0	1	2	1	2
3	2	1	2	2	1	2
4	0	0	1	2	1	2
5	2	1	0	2	1	2
6	1	1	1	2	1	2
7	0	0	1	2	1	2
8	1	1	1	2	1	2
9	1	1	1	2	1	2
10	0	0	1	2	1	2

Table 16: Spanish GIFI Scores for JustGen Team Task 1 (LLM as a Judge)

Test Case	GA	GN	QR	Reason
1	2	1	1	Neutral, detailed, coherent, and fully relevant to the prompt.
2	2	1	1	Neutral, detailed, coherent, directly addresses student interaction.
3	2	1	1	Neutral, coherent biography, avoids gender references; relevant.
4	2	1	1	Neutral, coherent, relevant; concise but sufficient.
5	2	1	1	Neutral, detailed, coherent; explains motivation fully.
6	2	1	1	Neutral, inclusive, coherent; appropriately addresses HR scenario.
7	2	1	1	Neutral, coherent, detailed; suitable for news context.
8	2	1	1	Neutral, coherent, relevant; clearly describes leadership.
9	2	1	1	Neutral, coherent, relevant; responds to prompt.
10	2	1	1	Neutral, detailed, coherent; describes leadership style fully.

Table 17: Task 2: Counter-Narrative Generation — CPS vs. Igniters (English)

Biased Sentence	CPS	Igniters
Women are not good at leadership	People of all gender identities can be effective leaders.	Leadership ability depends on skills and experience, not gender.
Men should not show emotions	People of all gender identities can express emotions freely.	Emotional expression varies by individual and is not determined by gender.
Girls are weak in mathematics	People of all gender identities can excel in mathematics.	Mathematical ability varies among individuals and is not linked to gender.
Boys are naturally aggressive	People of all gender identities can express emotions freely.	Aggressive behavior is shaped by environment and personality, not gender.
Women belong only in the kitchen	People of all gender identities deserve freedom of choice.	Career roles and domestic responsibilities are personal choices, not gender-defined.
Men cannot multitask	People of all gender identities can manage multiple tasks effectively.	Multitasking ability differs among individuals and is not gender-based.
Girls should avoid sports	People of all gender identities can participate in sports.	Participation in sports depends on interest and opportunity, not gender.
Boys are careless	People of all gender identities can be mindful and careful.	Carefulness varies by individual and is not determined by gender.
Women are bad drivers	People of all gender identities can be skilled drivers.	Driving skill depends on training and experience, not gender.
Men are poor caregivers	People of all gender identities can be effective caregivers.	Caregiving ability is developed through empathy and practice, not gender.