

Evaluating and Mitigating Misgendering in English-to-Hindi Machine Translation

Shreyas Gantayet¹, Chavi Gupta¹, Samyak Savi¹, Tanay Sodha¹, Dhruv Kumar¹

¹Birla Institute of Technology and Science, Pilani, India

shreyasgantayet@gmail.com

Abstract

Machine translation between gender-neutral languages (those lacking grammatical gender on verbs and adjectives, such as English) and gender-marked Indian languages poses a fundamental challenge: systems must infer and realize grammatical gender in the target language, yet often default to masculine forms, rely on occupational stereotypes, or produce grammatically valid translations that obscure gender altogether. While gender bias in machine translation has been studied extensively for European languages, Indian language models remain under-evaluated, and existing benchmarks do not adequately capture the linguistic and cultural phenomena relevant to Indian settings.

We introduce a 37,345-instance evaluation benchmark for English-Hindi gender bias spanning twelve linguistically motivated categories covering a broad range of gender-signaling phenomena, from stereotype inference and name-based disambiguation to cross-sentence gender tracking and Winograd-style coreference. Paired with a grammar-aware Hindi gender classifier combining rule-based morphological analysis with LLM fallback, the benchmark enables evaluation of six translation systems: three specialized MT models (Helsinki Opus-MT, NLLB-200, IndicTrans2) and three general-purpose LLMs (Sarvam AI, GPT-4o-mini, Claude Haiku 4.5). Across all six systems, we observe a consistent tendency to neutralize gender through Hindi’s ergative construction, a grammatical pattern in which the subject takes the postposition *ne* and the verb agrees with the object, erasing subject gender. In our benchmark, 63% to 76% of translations use such syntax even when source-side gender is unambiguous, causing downstream gender classification accuracy to fall to just 17%–31% on sentences where gender should be recoverable. Accuracy ranges from 48.0% for Helsinki Opus-MT to 59.2% for Claude Haiku 4.5 at baseline. A cross-model prompting sweep across six strategies shows that prompting reduces overt masculine defaulting without eliminating ergative neutralization, with Haiku 4.5 reaching 74.6% at the strongest prompting level. Combining prompting with in-translation reranking further mitigates ergative neutralization, raising accuracy to 88.2% for Haiku 4.5 and 74.5% for GPT-4o-mini; candidate-selection reranking yields consistent gains for the specialized MT models. These findings position Hindi as a qualitatively distinct setting for gender bias evaluation and motivate targeting ergative syntax directly as the primary mitigation lever.

1 Introduction

Neural machine translation has achieved remarkable progress in recent years, enabling real-time multilingual communication across diverse language pairs (Vaswani et al., 2017; NLLB Team et al., 2022). Yet a persistent challenge emerges when translating between gender-neutral languages like English and grammatically gendered languages like Hindi, where verbs, adjectives, and pronouns must carry explicit gender marking. In the absence of clear gender cues, translation models frequently default to masculine forms or draw on occupational stereotypes to resolve ambiguity (Stanovsky et al., 2019). For India’s 1.4 billion

speakers across 22 scheduled languages, such systematic biases risk reinforcing gender discrimination in civic, legal, and educational applications where machine-translated text is increasingly relied upon.

Hindi presents a qualitatively distinct challenge compared to the European languages studied in most prior work. Hindi exhibits ergative alignment in perfective constructions, marking the subject with the postposition *ne* (ने) in perfective aspect, causing the verb to agree with the object rather than the subject. Consider the English sentence “*She completed the project.*” A Hindi MT system can produce *उसने प्रोजेक्ट पूरा किया*, where the verb *kiya* agrees with the masculine object *project*, entirely erasing the feminine subject. Importantly, this is not a translation error: it is grammatically correct Hindi. But it renders downstream gender classification impossible for that sentence, because no morphological trace of the subject’s gender remains. This structural property distinguishes Hindi from European target languages (French, German, Spanish), where gendered morphology is almost always preserved.

Prior research has documented gender bias in MT for European languages (Stanovsky et al., 2019; Attanasio et al., 2023; Gonen et al., 2022). Broader multilingual benchmarks now include Hindi among wider evaluation suites (Currey et al., 2022; Robinson et al., 2024). Hindi-related work has also examined Hindi-English translation bias and Hindi-source evaluation with grammatical gender cues (Ramesh et al., 2021; Singh, 2023), while broader studies of Hindi language technology highlight the need for context-sensitive analysis (Hada et al., 2024). However, prior work has not isolated English-to-Hindi ergative neutralization as the central mechanism of gender loss in a large-scale benchmark setting.

We introduce a 37,345-instance benchmark spanning twelve linguistically motivated categories, constructed following the WinoBias methodology of Zhao et al. (2018) with manually curated cross-stereotype profession pairs. We evaluate three specialized MT systems and three general-purpose LLMs using a grammar-aware Hindi gender classifier with a six-step analysis pipeline and LLM fallback. Our main result is that all six systems frequently neutralize gender through ergative constructions. In our benchmark, 63% to 76% of translations use *ne* syntax that erases subject gender, even for sentences containing explicit English pronouns. The explicit-gender control condition therefore reaches only 17% to 31% accuracy, confirming that most of the loss occurs during translation rather than during downstream classification. We also conduct a cross-model prompting sweep spanning six strategies across Sarvam, GPT-4o-mini, and Claude Haiku 4.5, and show that combining prompting with in-translation reranking substantially reduces ergative neutralization and improves gender preservation.

Contributions.

1. A 37,345-instance evaluation benchmark with zero duplicate test instances, spanning 12 categories, 45 professions across 4 stereotype classes, 30 Indian names, and 50 manually curated profession pairs following WinoBias methodology (Zhao et al., 2018).
2. Identification and quantification of ergative neutralization as a recurring mechanism through which English-to-Hindi translation obscures subject gender, shown to persist across all six evaluated systems.
3. A large-scale comparison of six translation systems—three specialized MT models and three frontier LLMs—together with significance testing and human evaluation.
4. A cross-model prompting sweep across six strategies (including few-shot), showing that prompting reduces overt defaulting without materially reducing ergative neutralization.
5. Three reranking-based mitigation strategies covering all model families, achieving up to 88.2% accuracy (Haiku 4.5) versus 59.2% baseline.

2 Related Work

2.1 Gender Bias in Machine Translation

The problem of gender bias in neural machine translation has received sustained attention since the widespread adoption of attention-based architectures (Vaswani et al., 2017). Stanovsky et al. (2019) conducted one of the earliest large-scale audits, constructing challenge sets from Winogender schemas (Rudinger et al., 2018) and evaluating eight commercial and open-source systems. Attanasio et al. (2023) extended this line of inquiry to pronoun-level misgendering and mitigation. Saunders et al. (2020) showed that translation quality on gender coreference improves when source-side gender information is made more explicit. Most of this literature focuses on European target languages in which subject-verb agreement more directly preserves source gender.

This paper studies a different failure mode. In English-to-Hindi translation, grammatically well-formed outputs can still erase subject gender because perfective ergative constructions shift agreement away from the subject.

2.2 Evaluation Benchmarks

Benchmark construction has been central to progress in this area. Rudinger et al. (2018) introduced Winogender Schemas for controlled coreference-based measurement. Zhao et al. (2018) contributed manually curated profession pairs for stereotype-sensitive evaluation. Currey et al. (2022) introduced MT-GenEval for gender-accurate translation into eight target languages, including Hindi. Robinson et al. (2024) expanded multilingual evaluation further with a 26-language benchmark. These resources provide broad multilingual coverage, but they do not isolate English-to-Hindi ergative neutralization as the central mechanism of gender loss.

2.3 Hindi and Indic Settings

Hindi-related work is more limited and has largely focused on adjacent settings. Ramesh et al. (2021) study gender bias in Hindi-English machine translation, while Singh (2023) argue for evaluation that respects grammatical gender cues in Hindi-source translation. Beyond MT, Hada et al. (2024) show that gender bias in Hindi language technology requires context-sensitive and locally grounded analysis. We build on this direction, but focus specifically on English-to-Hindi translation and on the structural role of ergative neutralization.

2.4 MT Systems for Indian Languages

OPUS-MT (Tiedemann & Thottingal, 2020) provides lightweight Marian-based English-Hindi models trained on the OPUS corpus. NLLB (NLLB Team et al., 2022) extends massively multilingual translation to over 200 languages with dedicated improvements for low-resource Indian languages. IndicTrans2 (Gala et al., 2024) achieves state-of-the-art translation quality for all 22 scheduled Indian languages. Sarvam AI offers a commercial India-focused translation system with domain-specific fine-tuning. To our knowledge, prior work has not provided a large-scale English-to-Hindi comparison across these systems while centering ergative neutralization as the primary failure mechanism.

3 Methodology

3.1 Overview

Our evaluation framework operates through four stages: (1) benchmark construction with controlled gender signals, (2) machine translation into Hindi, (3) grammar-aware gender classification, and (4) evaluation with statistical significance testing. Figure 1 illustrates the complete pipeline.

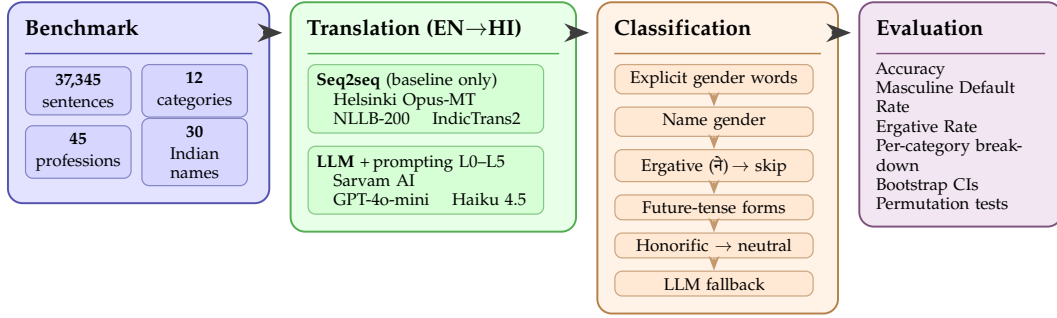


Figure 1: Evaluation pipeline. English sentences with controlled gender cues are translated by six systems (three seq2seq and three LLMs), classified by a grammar-aware six-step rule hierarchy with LLM fallback, then evaluated for gender accuracy, masculine default rate, and ergative rate.

3.2 Benchmark Dataset

We construct a purpose-built evaluation benchmark of 37,345 unique English test instances spanning 12 categories. Each instance is either a single sentence or a multi-sentence passage with controlled gender signals. **Categories.** Table 1 lists all 12 categories. The design follows WinoBias methodology (Zhao et al., 2018), with each category targeting a distinct linguistic phenomenon relevant to gender preservation in translation. Categories include explicit pronoun tests (explicit_gender), profession-only contexts where models must default (neutral_profession), cross-sentence gender tracking (multi_sentence, late_binding), Winograd-style coreference (winograd_coref), counter-stereotypical explicit markers (counter_stereotype), and minimal-signal baselines (minimal_context).

Professions and names. We use 45 professions across four stereotype classes: 12 STEM, 11 CARE, 10 LEADERSHIP, and 12 NEUTRAL, sourced from U.S. Bureau of Labor Statistics data. These stereotype classes are used only to construct controlled contrast pairs and are not intended as claims about Indian labor demographics. Fifty profession pairs are manually curated with cross-stereotype contrasts (e.g., engineer/nurse, CEO/receptionist). The benchmark includes 30 Indian names: 10 male-associated (Ram, Arjun, Vikram, ...), 10 female-associated (Priya, Anita, Lakshmi, ...), and 10 gender-ambiguous (Kiran, Aman, Arya, ...). Name-gender associations are consistent with morphological patterns identified in Indian name classification research (Tripathi & Faruqui, 2011) and with gender distributions observed in large-scale Indian public examination records (Medidoddi et al., 2022). Ambiguous names specifically probe MT behavior under reduced gender cue availability. Illustrative examples for each category are provided in Appendix H.

Ground truth. Each sentence receives a label of *male*, *female*, or *neutral* based on explicit pronouns, name gender, or absence of cues. Table 1 summarizes the benchmark composition.

Category	n	Tests
neutral_profession	7,500	Stereotype default
explicit_gender	7,500	Pronoun preservation
name_profession	6,750	Name+profession
winograd_coref	6,000	Coreference resolution
multi_sentence	2,340	Cross-sentence tracking
late_binding	2,250	Delayed gender reveal
social_role	1,350	Possessive constructions
counter_stereotype	960	Explicit gender markers
temporal_aspect	945	Tense-based gender
name_only	660	Name-only baseline
coreference	550	Ambiguous “they”
minimal_context	540	Minimal-signal baseline
Total	37,345	

Table 1: Benchmark categories, sample counts, and the main linguistic phenomenon targeted by each category.

3.3 Translation Models

We evaluate six English-to-Hindi translation systems spanning specialized MT models and general-purpose LLMs:

- **Helsinki Opus-MT** (Tiedemann & Thottingal, 2020): lightweight sequence-to-sequence model trained on the OPUS English-Hindi corpus, serving as an open-source baseline.
- **NLLB-200** (NLLB Team et al., 2022): Meta’s massively multilingual model covering over 200 languages.
- **IndicTrans2** (Gala et al., 2024): AI4Bharat’s state-of-the-art model for all 22 scheduled Indian languages.
- **Sarvam AI** (Sarvam AI, 2024): a commercial India-focused translation model with domain-specific fine-tuning and system prompt support.
- **GPT-4o-mini** (OpenAI, 2024): a general-purpose LLM evaluated as a translator via the OpenAI API, included to test whether ergative neutralization extends beyond specialized MT models.
- **Claude Haiku 4.5** (Anthropic, 2025): Anthropic’s frontier compact LLM, evaluated via API as a second general-purpose LLM baseline for cross-family comparison.

Prompting strategies. General-purpose LLMs (Sarvam AI, GPT-4o-mini, Haiku 4.5) support system prompts, enabling a six-strategy sweep. Seq2seq MT models (Helsinki, NLLB, IndicTrans2) do not accept system prompts and are evaluated at baseline only. The six strategies are: (1) **Baseline**: “Translate to Hindi”; (2) **Level 1**: “Preserve the grammatical gender of the subject”; (3) **Level 2**: explicit Hindi verb-form rules for masculine ($-\bar{a}$, $-t\bar{a}$, $-g\bar{a}$) and feminine ($-\bar{i}$, $-t\bar{i}$, $-g\bar{i}$) agreement; (4) **Level 3**: “Do not assign gender based on profession; use only explicit gender signals”; (5) **Level 4**: enhanced instruction with stricter source-faithfulness constraints; (6) **Level 5**: four unlabeled few-shot examples spanning `explicit_gender` and `late_binding` without naming either category. Full prompt text is provided in Appendix B.

3.4 Grammar-Aware Hindi Gender Classifier

We develop a rule-based classifier with a six-step decision hierarchy, designed to handle Hindi’s ergative alignment correctly:

1. **Explicit gender words** (पुरुष/महिला): detects explicit markers used in counter-stereotype sentences.
2. **Name detection**: matches gendered Hindi names; ambiguous names are skipped rather than classified.
3. **Sentence splitting with ergative handling**: multi-sentence inputs are split; sentences containing the ergative marker *ne* (ने) are skipped, since their verbs agree with the object rather than the subject.
4. **Singular verb forms**: unambiguous masculine ($-\bar{a}$, $-g\bar{a}$) or feminine ($-\bar{i}$, $-g\bar{i}$) agreement.
5. **Honorific/plural verbs**: classified as neutral, since these forms are ambiguous between honorific singular and genuine plural.
6. **LLM fallback** (Gemini 2.5 Flash): applied only to cases not resolved by rules.

4 Experimental Setup

Dataset. The benchmark comprises 37,345 test instances: 12,677 male (33.9%), 12,963 female (34.7%), and 11,705 neutral (31.3%).

Models. All six translation systems and the Gemini classifier are used inference-only with no fine-tuning. Table 15 in Appendix E lists specific model identifiers.

Metrics. We report four metrics: (1) **Accuracy** ($\text{Acc}\uparrow$): fraction of correct gender predictions; (2) **Masculine Default Rate** ($\text{MDR}\downarrow$): fraction of neutral-ground-truth instances predicted

as male; (3) **Per-class recall**; (4) **Ergative rate (Erg%↓)**: fraction of translations containing the Hindi ergative marker *ne* (ने), indicating gender-neutral verb agreement.

Statistical tests. We use bootstrap 95% confidence intervals (10,000 resamples) for all accuracy figures, permutation tests (10,000 iterations) for pairwise model comparisons, and two-proportion z-tests with Cohen’s *h* effect sizes for ergative rate comparisons.

Implementation. Translation was performed on an NVIDIA GPU with batch sizes of 16–32. Classification used the grammar-aware rule-based classifier with Gemini 2.5 Flash fall-back across all six systems: the three non-general-purpose MT models (Helsinki, NLLB, IndicTrans2) and the general-purpose models (Sarvam, GPT-4o-mini, Haiku 4.5). All experiments used random seed 42.

5 Results

5.1 Overall Performance

Table 2 presents overall performance for all six translation systems. Among the three specialized MT models, IndicTrans2 achieves the highest accuracy (58.7%). Helsinki performs worst at 48.0% but shows the highest masculine default rate (22.5%) alongside the lowest ergative rate (63.2%), suggesting that it preserves more gendered verb forms but defaults to masculine when doing so. NLLB achieves the lowest MDR (5.7%) but the highest ergative rate (76.2%), indicating a strong preference for gender-neutral constructions. Among the general-purpose LLMs, Claude Haiku 4.5 achieves the highest accuracy (59.2%) with the lowest ergative rate (67.1%), narrowly ahead of Sarvam (58.4%, 74.9% ergative) and GPT-4o-mini (57.2%, 76.3% ergative), suggesting that Haiku produces somewhat more morphologically gender-marked output but also incurs more masculine defaults (MDR = 15.0%). Bootstrap 95% confidence intervals are reported in Appendix D.

Model	Acc↑	MDR↓	Erg%↓
Haiku 4.5	59.2	15.0	67.1
IndicTrans2	58.7	7.1	74.7
Sarvam	58.4	10.5	74.9
NLLB	57.2	5.7	76.2
GPT-4o-mini	57.2	10.1	76.3
Helsinki	48.0	22.5	63.2

Table 2: Overall performance across six translation systems. Acc = accuracy (↑ higher is better); MDR = masculine default rate (↓ lower is better); Erg% = ergative rate (↓). Bold = best per column. All pairwise accuracy differences among Sarvam, Helsinki, NLLB, and IndicTrans2 are significant ($p < 0.01$).

5.2 Ergative Neutralization: The Central Finding

The most important result of this study concerns the rate at which translation models employ Hindi’s ergative construction to produce grammatically valid but gender-neutral output. Table 3 shows ergative rates for selected categories.

Category	Hel	NLLB	IT2	Sar	GPT	Haiku
explicit_gender	69.9	68.8	69.8	70.9	73.5	46.7
late_binding	90.7	100.0	99.6	100.0	100.0	100.0
winograd_coref	88.6	98.8	93.3	95.8	100.0	95.0
neutral_prof	82.2	96.5	96.5	94.4	94.2	88.3
name_profession	20.5	45.3	45.5	47.5	46.0	44.0
Overall	63.2	76.2	74.7	74.9	76.3	67.1

Table 3: Ergative (ने) construction rate (%↓) by category and model. Pronoun-driven categories (explicit_gender, late_binding, winograd_coref) show consistently high ergative usage across all models. Haiku 4.5 (67.1% overall) and Helsinki (63.2%) produce fewer ergative constructions than the 74–76% cluster.

The explicit_gender category is especially revealing. These sentences contain unambiguous English pronouns (“He is a skilled engineer,” “She worked as a nurse”) and should, in prin-

ciple, achieve near-perfect accuracy. Instead, accuracy ranges from 16.7% (IndicTrans2) to 31.0% (Helsinki) because 69% to 71% of translations across Sarvam, Helsinki, NLLB, and IndicTrans2 use ergative or honorific constructions that strip the gender signal entirely. This pattern indicates that much of the loss occurs at the *translation stage*, not only during classification. For example, “*He worked as an engineer for many years*” is translated by Sarvam as उसने कई वर्षों तक एक इंजीनियर के रूप में काम किया, where the ergative उसने removes the masculine pronoun. The reverse error also occurs: the gender-neutral “*The engineer arrived.*” is rendered as इंजीनियर आ गया। with the masculine verb गया, injecting gender where none was intended. Further examples appear in Appendix I.

The ergative rate difference between Helsinki (63.2%) and the 74–76% cluster (Sarvam, NLLB, IndicTrans2, GPT-4o-mini) is statistically significant ($z = 34.5$, $p < 10^{-6}$, Cohen’s $h = 0.25$), while those four are indistinguishable from each other ($h < 0.04$). Haiku 4.5 (67.1%) is intermediate ($h \approx 0.08$ – 0.21 versus the other systems).

5.3 Per-Category Analysis

Two distinct clusters emerge from the per-category results (Table 4).

Categories relying on **names** for gender signaling (counter_stereotype, multi_sentence, name_only, name_profession) achieve 86–100% accuracy across Sarvam, NLLB, and IndicTrans2, because Hindi proper nouns survive translation intact and provide a gender signal independent of verb morphology. GPT-4o-mini and Haiku 4.5 show comparable accuracy on name-based categories ($\geq 80\%$), similar to those systems. Categories relying on **pronouns** (explicit_gender, late_binding, winograd_coref) fail at 0–31% across all systems, because English pronouns are converted to ergative constructions. This dichotomy confirms that the failure is specifically tied to how MT models handle English pronouns in Hindi output. Helsinki shows notably lower accuracy even on name-based categories (41.8% on name_only, 52.3% on name_profession). A heatmap visualization across all six models is in Appendix G.

Category	Hel	NLLB	IT2	Sar	GPT	Haiku
counter_stereo	50.7	96.2	100.0	100.0	100.0	100.0
multi_sentence	73.7	100.0	100.0	100.0	99.1	100.0
name_only	41.8	94.8	98.0	97.7	95.9	94.2
name_profession	52.3	86.2	90.4	86.3	80.6	80.5
explicit_gender	31.0	17.8	16.7	21.9	21.1	28.1
late_binding	4.8	0.1	7.2	7.8	7.2	9.0
winograd_coref	14.3	1.3	5.2	4.1	0.0	5.8
Overall	48.0	57.2	58.7	58.4	57.2	59.2

Table 4: Accuracy ($\uparrow$) by selected category across all six systems, showing the top-performing name-based group (top) and the failing pronoun-based group (bottom). Full per-category results are in Appendix G.

5.4 Prompting Strategy Evaluation

Table 5 presents the complete cross-model prompting sweep covering Sarvam, GPT-4o-mini, and Haiku 4.5 across all six strategies.

For **Sarvam** and **GPT-4o-mini**, prompting yields at most 3 pp accuracy gains across all levels. Level 3 (anti-stereotype override) best reduces MDR, halving Sarvam’s MDR from 13.2% to 7.6%, while the ergative rate remains near 75% (Sarvam) and 76% (GPT-4o-mini) regardless of prompt level. This structural stability confirms that prompting can shift *which* gender is assigned but does not change the preference for ergative constructions.

Haiku 4.5 shows much larger prompting gains: Level 5 (few-shot) achieves 74.6%, a +15.4 pp improvement over baseline. Level 2 and Level 5 tie for the lowest ergative rate (63.8%), and Level 3 achieves the lowest MDR (6.9%). Even so, the ergative rate declines only modestly (from 67.1% to 63.8% at best), and Level 5 still falls 13.6 pp short of Haiku’s best reranking

result (88.2%, Section 5.5), confirming that the structural ergative ceiling persists even for the model most responsive to prompting.

Model	Level	Acc $\uparrow$	MDR $\downarrow$	Erg% $\downarrow$
Sarvam	Baseline	58.4	13.2	74.9
	Level 1	58.7	12.1	75.6
	Level 2	56.6	17.1	74.8
	Level 3	60.4	7.6	75.1
	Level 4	57.8	2.8	72.8
	Level 5	59.4	12.3	74.3
GPT-4o-mini	Baseline	57.2	12.5	76.3
	Level 1	56.8	11.5	75.1
	Level 2	56.9	11.7	75.4
	Level 3	57.7	10.9	76.3
	Level 4	56.1	11.3	75.4
	Level 5	60.1	12.8	76.0
Haiku 4.5	Baseline	59.2	15.0	67.1
	Level 1	60.8	14.6	65.6
	Level 2	64.8	10.1	63.8
	Level 3	61.9	6.9	66.1
	Level 4	62.5	8.5	66.6
	Level 5	74.6	14.3	63.8

Table 5: Cross-model prompting sweep across all six strategies. Acc $\uparrow$ = accuracy; MDR $\downarrow$ = masculine default rate; Erg% $\downarrow$ = ergative rate. Bold = best per model per metric.

5.5 Mitigation via Reranking

Since prompting cannot overcome ergative syntax at the structural level, we develop three complementary reranking strategies that target it directly.

In-translation reranking. Applied to the three LLMs (Sarvam, GPT-4o-mini, Haiku 4.5). The model identifies the type of each source sentence and generates $k=5$ diverse candidates using sentence-type-specific prompts, then selects the candidate that best preserves source-side gender.

Post-translation reranking. Applied as an additional refinement step for GPT-4o-mini after in-translation reranking. A second LLM pass re-scores the already-selected best candidate against the source sentence’s gender cues, and may substitute an alternative candidate if one better preserves gender.

Candidate-selection reranking. Applied to the three seq2seq models (Helsinki, NLLB, IndicTrans2), which lack a system-prompt interface. Multiple translation candidates are sampled and the one best preserving source-side gender cues is selected by the rule-based classifier.

Tables 6 and 7 summarize baseline versus best accuracy overall and for the three hardest pronoun-based categories.

System	Baseline	Best	Δ	Strategy
Haiku 4.5	59.2	88.2	+29.0	L5 + In-translation reranking
GPT-4o-mini	57.2	74.5	+17.3	L5 + In-translation + Post-translation
Sarvam	58.4	72.7	+14.3	L3 + In-translation reranking
IndicTrans2	58.7	59.7	+1.0	Candidate-selection
NLLB	57.2	59.0	+1.8	Candidate-selection
Helsinki	48.0	54.3	+6.3	Candidate-selection

Table 6: Accuracy ($\uparrow$) at baseline and best reranking per system. LLMs gain +14 to +29 pp when prompting and reranking are combined; seq2seq models gain +1 to +6 pp from candidate-selection reranking.

System	Method	Expl.↑	Late↑	Wino.↑	Full↑
Sarvam	Baseline	26.7	7.8	5.7	58.4
Sarvam	L3 + In-translation	59.6	27.2	45.7	72.7
GPT-4o-mini	Baseline	21.1	7.2	0.0	57.2
GPT-4o-mini	L5 + In-translation + Post-translation	64.4	49.5	44.0	74.5
Haiku 4.5	Baseline	28.1	9.0	5.8	59.2
Haiku 4.5	L5 + In-translation	99.1	99.0	59.0	88.2

Table 7: Per-category accuracy (↑) for the three hardest pronoun-based categories before and after best reranking. Expl. = explicit_gender; Late = late_binding; Wino. = winograd_coref; Full = overall accuracy.

Results. For LLMs, combining prompting with reranking yields +14 to +29 pp over baseline, substantially mitigating the ergative neutralization that prompting alone cannot overcome. Haiku 4.5 reaches 88.2% when Level 5 prompting is combined with in-translation reranking; explicit_gender and late_binding jump from near-zero to 99.1% and 99.0% respectively. GPT-4o-mini reaches 74.5% with the additional post-translation reranking pass. Seq2seq models show smaller but consistent gains (+1 to +6 pp) from candidate-selection reranking, with Helsinki gaining the most (+6.3 pp). Winograd-style coreference remains the hardest category even after reranking (44% for GPT, 59% for Haiku), suggesting that cross-sentence gender tracking requires more targeted mitigation.

Translation quality. Human annotators rated fluency on a 5-point scale for a 200-instance subset. Candidate-selection reranking achieves a preservation gain of +8.3 pp with negligible fluency cost ($\Delta = -0.01$). In-translation reranking achieves a +71.0 pp preservation gain but with a more noticeable fluency penalty ($\Delta = -0.99$ on a 5-point scale), reflecting that explicit gender insertions occasionally produce less idiomatic Hindi.

5.6 Error Pattern Analysis

Table 8 quantifies the dominant error patterns for pronoun-dependent categories on Sarvam baseline; GPT-4o-mini and Haiku 4.5 show the same neutralization-dominated pattern with comparable gender→neutral rates.

Across all three pronoun-dependent categories, gender-to-neutral conversion accounts for 63–89% of predictions, while active misgendering (male→female or female→male) constitutes only 5–9%. The neutralization rates for male and female ground truths are nearly symmetric (33.1% vs. 29.7% for explicit_gender), indicating the mechanism itself is not gender-biased.

5.7 Human Evaluation

To assess the automatic classifier, two annotators independently labeled 100 stratified test instances drawn from six categories. On the 97 items with valid overlapping gender labels, raw inter-annotator agreement reaches 85.6% and Cohen’s κ reaches 0.776. Classifier-human agreement is 83.8% overall (93.9% when the classifier predicts neutral, 86.4% for male, 71.2% for female), versus only 46.2% human-ground-truth agreement. Annotators judge that 48.0% of translations do not preserve source gender, 51.0% preserve it, and 1.0% are partial – supporting the interpretation that the classifier captures neutralization reliably, while the translations themselves often remove source-side gender information.

Pattern	expl.	late	wino.
male → neutral	33.1	43.7	45.2
female → neutral	29.7	43.6	45.3
correct	20.2	7.8	4.0
female → male	6.5	2.1	0.5
male → female	3.0	2.7	2.1

Table 8: Error patterns for Sarvam on pronoun-dependent categories. Neutralization dominates over active misgendering.

6 Discussion

Hindi as a distinct challenge. Standard accuracy and MDR metrics underestimate gender loss in Hindi, since ergative neutralization produces syntactically valid output that obscures subject gender—a failure mode absent from European MT settings; we recommend reporting the ergative rate alongside standard metrics in future evaluations.

Ergative neutralization extends beyond specialized MT models. All three general-purpose LLMs (Sarvam, 74.9% ergative; GPT-4o-mini, 76.3%; Haiku 4.5, 67.1%) show the same high-ergative pattern as the specialized MT models, with all six systems falling in the 63–76% range.

Reranking as the effective mitigation lever. Prompting reduces overt masculine defaults but cannot overcome ergative syntax (Section 5.4). Reranking targets it directly: in-translation reranking raises Haiku 4.5 from 59.2% to 88.2% and GPT-4o-mini from 57.2% to 74.5%; candidate-selection reranking yields +1 to +6 pp for seq2seq models without prompt access. The fluency penalty ($\Delta = -0.99$ on a 5-point scale) motivates future work on constrained decoding.

Implications for deployed systems. With 63–76% of translations erasing subject gender and Helsinki assigning masculine gender to 22.5% of neutral-ground-truth instances, deployed Hindi MT systems risk systematically misgendering users in civic, legal, and educational contexts.

7 Limitations

Constructed data. The benchmark uses constructed rather than naturally sampled data; additional validation on naturally occurring English-Hindi text would strengthen external validity.

Ambiguous names. The 10 gender-ambiguous names may carry cultural masculine associations not fully captured by our label, though not formally validated against Indian name-gender research (Tripathi & Faruqi, 2011; Medidoddi et al., 2022).

Human evaluation scope. Human evaluation covers 100 doubly annotated classifier instances and a 200-instance reranking subset. Classifier agreement is substantial ($\kappa = 0.776$); preservation judgments are noisier (raw overlap 71.0%, $\kappa = 0.432$).

Reranking coverage. In-translation reranking covers three pronoun-dependent categories (43% of instances); extending to name_profession and neutral_profession would further increase coverage.

8 Conclusion

We evaluate gender preservation in English-to-Hindi translation using a 37,345-instance benchmark spanning twelve categories and six systems. Across all systems, 63% to 76% of outputs adopt ergative syntax that obscures subject gender, driving gender classification accuracy down to 17%–31% on sentences where gender should be recoverable. A cross-model prompting sweep shows consistent but limited gains (≤ 3 pp for Sarvam and GPT-4o-mini, up to +15.4 pp for Haiku 4.5): ergative neutralization is a structural barrier that prompting alone cannot overcome. Combining prompting with in-translation reranking substantially reduces this barrier, raising accuracy to 88.2% for Haiku 4.5 and 74.5% for GPT-4o-mini, with candidate-selection reranking yielding consistent gains for seq2seq models. Future work should explore constrained decoding, cross-linguistic extension to Urdu and Marathi, and more inclusive non-binary gender evaluation.

Ethics Statement

This research evaluates gender bias in machine translation systems with the goal of identifying and measuring systematic failures. The benchmark uses constructed data, avoiding privacy concerns associated with naturally occurring text. All Indian names used are common and non-identifying. The benchmark is designed to test for bias, not to create or amplify it. Our evaluation employs a three-class gender taxonomy (male, female, neutral). We acknowledge that this framework does not capture non-binary, transgender, or other gender identities, which represent an important and underserved population in NLP systems. Extending the evaluation framework to more inclusive gender representations, including honorific and non-binary forms specific to South Asian cultural contexts, is a priority for future work.

Reproducibility Statement

The dataset is generated deterministically with seed 42. All translation models are publicly available on HuggingFace (Helsinki, NLLB, IndicTrans2) or accessible via API (Sarvam AI, GPT-4o-mini, Claude Haiku 4.5). Statistical tests use 10,000 bootstrap/permutation iterations. The benchmark dataset and evaluation code are available at <https://github.com/Y0shicon/Evaluating-Gender-Bias-in-English-to-Hindi-MT-Dataset/>.

Acknowledgments

We thank Shaily Jagat Bhatt (Carnegie Mellon University) for helpful discussions and feedback on the framing, evaluation design, and presentation of this work.

References

- Anthropic. Claude Haiku 4.5. <https://www.anthropic.com/claude/haiku>, 2025. Accessed: August 2026.
- Giuseppe Attanasio, Flor Miriam Plaza del Arco, Debora Nozza, and Anne Lauscher. A tale of pronouns: Interpretability informs gender bias mitigation for fairer instruction-tuned machine translation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3996–4014, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.243. URL <https://aclanthology.org/2023.emnlp-main.243/>.
- Anna Currey, Maria Nadejde, Raghavendra Reddy Pappagari, Mia Mayer, Stanislas Lauly, Xing Niu, Benjamin Hsu, and Georgiana Dinu. Mt-geneval: A counterfactual and contextual dataset for evaluating gender accuracy in machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 4287–4299, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics.
- Jay Gala, Pranjal A. Chitale, A. K. Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M. Khapra, Raj Dabre, and Anoop Kunchukuttan. IndicTrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*, 2024.
- Hila Gonen, Shauli Ravfogel, and Yoav Goldberg. Analyzing gender representation in multilingual models. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pp. 67–77, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.repl4nlp-1.8. URL <https://aclanthology.org/2022.repl4nlp-1.8/>.
- Rishav Hada, Safiya Husain, Varun Gumma, Harshita Diddee, Aditya Yadavalli, Agrima Seth, Nidhi Kulkarni, Ujwal Gadiraju, Aditya Vashistha, Vivek Seshadri, and Kalika Bali. Akal badi ya bias: An exploratory study of gender bias in hindi language technology. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2024.

- Vahini Medidoddi, Jalend Bantupalli, Souvic Chakraborty, and Animesh Mukherjee. Decoding demographic un-fairness from Indian names. In Frank Hopfgartner, Kokil Jaidka, Philipp Mayr, Joemon Jose, and Jan Breitsohl (eds.), *Social Informatics*, pp. 472–489, Cham, 2022. Springer International Publishing. ISBN 978-3-031-19097-1.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.
- OpenAI. GPT-4o mini: Advancing cost-efficient intelligence. <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>, 2024. Accessed: March 2026.
- Krithika Ramesh, Gauri Gupta, and Sanjay Singh. Evaluating gender bias in Hindi-English machine translation. In Marta R. Costa-jussà, Hila Gonen, Christian Hardmeier, and Kellie Webster (eds.), *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pp. 16–23, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.gebnlp-1.3. URL <https://aclanthology.org/2021.gebnlp-1.3/>.
- Kevin Robinson, Sneha Kudugunta, Romina Stella, Sunipa Dev, and Jasmijn Bastings. MiT-TenS: A dataset for evaluating gender mistranslation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4115–4124, Miami, Florida, USA, 2024. Association for Computational Linguistics.
- Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 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*. Association for Computational Linguistics, 2018.
- Sarvam AI. Sarvam translate. <https://www.sarvam.ai/blogs/sarvam-translate>, 2024. Accessed: March 2026.
- Danielle Saunders, Rosie Sallis, and Bill Byrne. Neural machine translation doesn’t translate gender coreference right unless you make it. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pp. 35–43, Barcelona, Spain (Online), 2020. Association for Computational Linguistics.
- Pushpdeep Singh. Gender inflected or bias inflicted: On using grammatical gender cues for bias evaluation in machine translation. In Dongfang Li, Rahmad Mahendra, Zilu Peter Tang, Hyeju Jang, Yugo Murawaki, and Derek Fai Wong (eds.), *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: Student Research Workshop*, pp. 17–23, Nusa Dua, Bali, November 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.ijcnlp-srw.3. URL <https://aclanthology.org/2023.ijcnlp-srw.3/>.
- Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1679–1684. Association for Computational Linguistics, 2019.
- Jörg Tiedemann and Santhosh Thottingal. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pp. 479–480, 2020.
- Anshuman Tripathi and Manaal Faruqui. Gender prediction of Indian names. In *IEEE Technology Students’ Symposium*, pp. 137–141, Kharagpur, 2011. IEEE. doi: 10.1109/TECHSYM.2011.5783842.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 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*, pp. 15–20. Association for Computational Linguistics, 2018.

Appendix

A LLM Usage Disclosure

LLM Disclosure (COLM Policy 1). The authors acknowledge the use of AI tools such as ChatGPT, Claude, and Gemini for improving the presentation and grammar of this paper. All results, analyses, and proposed techniques remain the authors’ own contributions. The authors take full responsibility for the contents of this paper.

B Prompting Strategy Details

Table 9 provides the complete system prompts used for each translation-stage prompting strategy with Sarvam AI.

Level	System Prompt
Baseline	“Translate the text below to Hindi.”
Level 1	“Translate the text below to Hindi. Preserve the grammatical gender of the subject exactly as indicated in the source sentence.”
Level 2	“Translate the text below to Hindi. Follow these rules for grammatical gender agreement: if the subject is male or referred to with he/his, use masculine verb forms ending in -a (past), -ta (habitual), -ga (future); if the subject is female or referred to with she/her, use feminine verb forms ending in -i (past), -ti (habitual), -gi (future); if no gender is specified, use gender-neutral phrasing and avoid defaulting to masculine.”
Level 3	“Translate the text below to Hindi. IMPORTANT: Do not assign gender based on the profession mentioned in the sentence. Only use gender signals that are explicitly present in the source (pronouns he/she, or words like male/female/man/woman). If no explicit gender signal is present, use gender-neutral verb forms and do not default to masculine.”
Level 4	“Translate the text below to Hindi. IMPORTANT: Translate only the user’s sentence. Do not repeat the English sentence. Do not invent content. If the source explicitly says she/her, preserve feminine agreement. If the source explicitly says he/him/his, preserve masculine agreement. If there is no explicit gender signal, use natural Hindi and avoid inferring gender from profession alone.”
Level 5	[Level 3 instruction] + Four unlabeled few-shot examples (two from <code>explicit_gender</code> and two from <code>late_binding</code>), presented without category labels, demonstrating gender-preserving Hindi translations before the test sentence.

Table 9: Complete system prompts for all six translation-stage prompting strategies applied to general-purpose LLMs (Sarvam, GPT-4o-mini, Haiku 4.5). Level 5 uses the Level 3 instruction combined with few-shot examples.

C Full Cross-Model Prompting Results

Table 10 provides the complete six-level prompting sweep for all three general-purpose LLMs.

Model	Strategy	Acc $\uparrow$	MDR $\downarrow$	Erg% $\downarrow$
Sarvam	Baseline	58.4	13.2	74.9
	Level 1	58.7	12.1	75.6
	Level 2	56.6	17.1	74.8
	Level 3	60.4	7.6	75.1
	Level 4	57.8	2.8	72.8
	Level 5	59.4	12.3	74.3
GPT-4o-mini	Baseline	57.2	12.5	76.3
	Level 1	56.8	11.5	75.1
	Level 2	56.9	11.7	75.4
	Level 3	57.7	10.9	76.3
	Level 4	56.1	11.3	75.4
	Level 5	60.1	12.8	76.0
Haiku 4.5	Baseline	59.2	15.0	67.1
	Level 1	60.8	14.6	65.6
	Level 2	64.8	10.1	63.8
	Level 3	61.9	6.9	66.1
	Level 4	62.5	8.5	66.6
	Level 5	74.6	14.3	63.8

Table 10: Full six-level prompting sweep across three general-purpose LLMs. Sarvam and GPT-4o-mini show at most 3 pp accuracy gains across all levels; ergative rates remain stable at 72–76% (Sarvam) and 75–76% (GPT-4o-mini). Haiku 4.5 shows the largest prompting gains (+15.4 pp at Level 5) with ergative rates stable at 64–67%, and still falls 13.6 pp below its best reranking result.

D Statistical Significance Tests

Comparison	Δ Acc	p -value
Sarvam vs. Helsinki	+0.142	<0.001
Sarvam vs. NLLB	+0.017	<0.001
Sarvam vs. IndicTrans2	−0.003	0.003
Helsinki vs. NLLB	−0.125	<0.001
Helsinki vs. IndicTrans2	−0.145	<0.001
NLLB vs. IndicTrans2	−0.020	<0.001

Table 11: Pairwise permutation tests (10,000 iterations) on overall accuracy. All differences are significant at $p < 0.01$.

Model	Acc [95% CI]	Erg Rate [95% CI]
Sarvam	0.568 [0.563, 0.573]	0.749 [0.745, 0.753]
Helsinki	0.426 [0.421, 0.431]	0.632 [0.627, 0.637]
NLLB	0.551 [0.546, 0.557]	0.762 [0.758, 0.767]
IndicTrans2	0.572 [0.567, 0.577]	0.747 [0.742, 0.751]

Table 12: Bootstrap 95% confidence intervals (10,000 resamples) for overall accuracy and ergative rate for Sarvam, Helsinki, NLLB, and IndicTrans2. The other two general-purpose LLMs (GPT-4o-mini, Haiku 4.5) are evaluated across prompting levels; per-level results appear in Appendix C.

Comparison	z	p -value	Cohen’s h
Sarvam vs. Helsinki	34.48	<0.001	0.253 (small)
Sarvam vs. NLLB	−4.25	<0.001	−0.031 (negl.)
Sarvam vs. IT2	0.69	0.490	0.005 (negl.)
Helsinki vs. NLLB	−38.67	<0.001	−0.284 (small)
Helsinki vs. IT2	−33.80	<0.001	−0.248 (small)
NLLB vs. IT2	4.94	<0.001	0.036 (negl.)

Table 13: Ergative rate pairwise comparisons (two-proportion z -test). Helsinki’s lower ergative rate is significantly different from all other models ($h \approx 0.25$, small effect). Sarvam, NLLB, and IndicTrans2 are practically indistinguishable ($h < 0.04$).

E Classification Method Distribution

Method	Sar	Hel	NLLB	IT2
rule_ergative	55.3	52.4	59.7	54.7
rule_name	23.8	8.3	21.9	23.9
rule_verb	12.0	23.0	8.5	10.5
rule_honorific	2.2	2.9	3.1	4.4
rule_explicit	2.7	2.5	2.6	2.6
llm_fallback	3.9	10.8	4.2	3.9

Table 14: Classification method distribution (%) for Sarvam, Helsinki, NLLB, and IndicTrans2. Ergative-based classification is the dominant method (52–60%), reflecting how frequently translations use gender-neutral constructions.

Model	Role	Identifier
Sarvam AI	Translation	sarvam-translate
Helsinki	Translation	Helsinki-NLP/opus-mt-en-hi
NLLB-200	Translation	facebook/nllb-200-distilled-600M
IndicTrans2	Translation	ai4bharat/indictrans2-en-indic-1B
GPT-4o-mini	Translation	openai/gpt-4o-mini
Claude Haiku 4.5	Translation	claude-haiku-4-5
Gemini	LLM fallback	gemini-2.5-flash

Table 15: Model identifiers used in experiments.

F Dataset Category Details

The benchmark follows the WinoBias methodology of [Zhao et al. \(2018\)](#) for profession pair construction. Fifty pairs are manually curated with cross-stereotype contrasts: 20 STEM×CARE pairs (e.g., engineer/nurse, developer/teacher), 15 LEADERSHIP×CARE pairs (e.g., CEO/receptionist, director/secretary), 10 STEM×NEUTRAL pairs (e.g., programmer/designer, scientist/artist), and 5 LEADERSHIP×NEUTRAL pairs (e.g., manager/consultant). All pairs are selected to have at least a 30 percentage-point difference in gender distribution based on BLS labor force statistics.

Ground truth is assigned deterministically: sentences with explicit pronouns receive the pronoun’s gender; sentences with gendered names receive the name’s gender; sentences with ambiguous names or no gender cue receive the neutral label.

G Per-Category Heatmap

Figure 2 visualizes the per-category accuracy pattern across all six models.

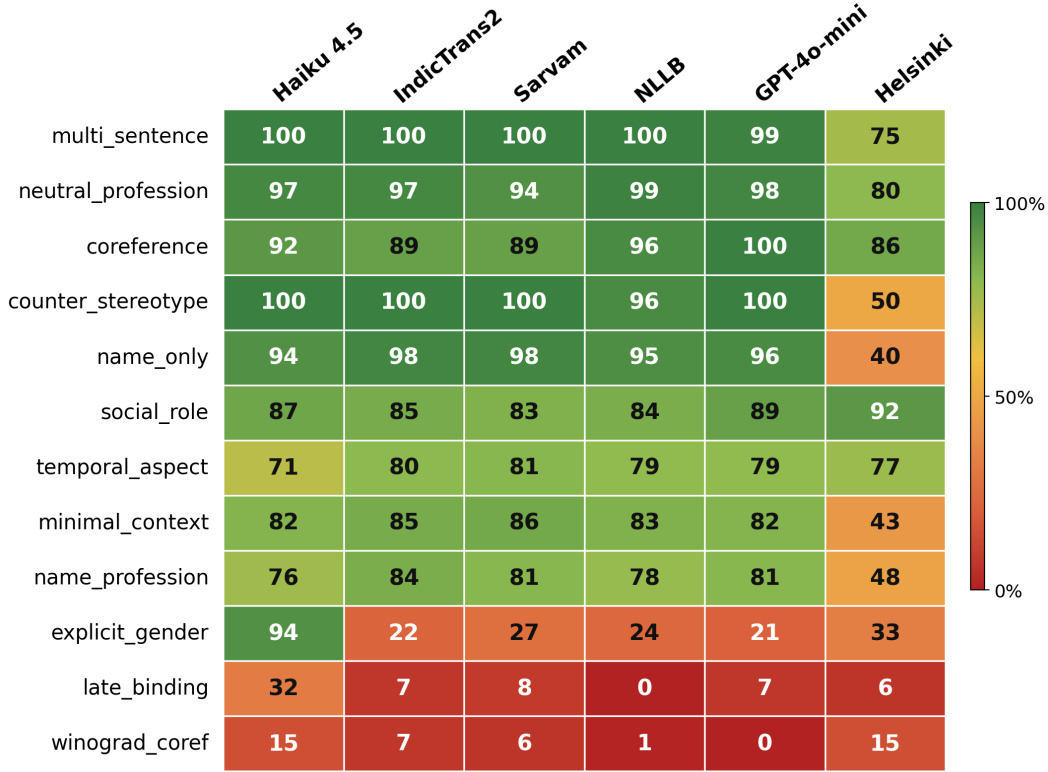


Figure 2: Per-category accuracy (%) across all six MT models. Most categories achieve 75–100% (green), while ergative-heavy categories (*late_binding*, *winograd_coref*) fail at 0–32% (red). Haiku 4.5 uniquely achieves 94% on *explicit_gender*, whereas the remaining systems reach only 21–33%.

H Category Examples

Table 16 provides one illustrative test instance per category, showing the linguistic phenomenon each is designed to probe.

Category	Example Instance	GT
neutral	The engineer completed the task efficiently.	neutral
profession		
explicit gender	She is a skilled engineer.	female
name +	Priya works as an engineer in Mumbai.	female
profession		
winograd coref	The engineer called the nurse since he required assistance.	male
multi sentence	Priya is an engineer. She works at the hospital. She enjoys helping patients.	female
late binding	The engineer arrived at the office. Then he checked the schedule.	male
social role	The engineer’s child attended the school event.	neutral
counter-	The male nurse helped patients professionally.	male
stereotype		
temporal aspect	The engineer will complete the task tomorrow.	neutral
name only	Kiran submitted the report yesterday.	neutral
coreference	The team lead accompanied the student since they traveled together.	neutral
minimal context	The engineer arrived.	neutral

Table 16: One example test instance per benchmark category. GT = ground-truth gender label assigned during construction.

I Qualitative Error Examples

The qualitative cases are separated into *neutralization* and *injection* examples so that the English and Hindi text can be shown at readable width.

Category	English	Hindi Translation	Shift
explicit gender	He worked as an engineer for many years.	उसने कई वर्षों तक एक इंजीनियर के रूप में काम किया।	male → neutral
late binding	The engineer arrived at the office. Then he checked the schedule.	इंजीनियर दफ्तर पहुंचे। फिर उन्होंने शेड्यूल देखा।	male → neutral
winograd coref	The engineer called the nurse since he required assistance.	इंजीनियर ने नर्स को फोन किया क्योंकि उसे मदद की ज़रूरत थी।	male → neutral

Table 17: Representative *neutralization* examples from Sarvam (Level 3). Explicit masculine cues in English disappear in Hindi through ergative or honorific constructions.

Category	English	Hindi Translation	Shift
neutral profession	The programmer arrived late to the conference.	प्रोग्रामर सम्मेलन में देर से पहुंचा।	neutral → male
minimal context	The engineer arrived.	इंजीनियर आ गया।	neutral → male
name only	Aman arrived late to the event.	अमन कार्यक्रम में देर से पहुंचा।	neutral → male

Table 18: Representative *injection* examples from Sarvam (Level 3). Neutral English sentences are rendered with masculine Hindi morphology.