BRIDGE-MT: A Benchmark for Role Interactions and Dependencies in Machine Translation Gender Evaluation

Neha Gajakos C00315510@setu.ie
South East Technological University, Carlow, Ireland

Christopher Staff christopher.staff@setu.ie
South East Technological University, Carlow, Ireland

Brenda Murphy brenda.murphy@setu.ie
South East Technological University, Carlow, Ireland

John D. Kelleher john.kelleher@adaptcentre.ie
ADAPT Centre, Trinity College Dublin, Dublin, Ireland

Rejwanul Haque rejwanul.haque@setu.ie
South East Technological University, Carlow, Ireland

Abstract

This paper investigates gender behavior in Hindi–English machine translation (MT) within multi-entity settings, especially when two occupational roles appear within the same sentence. Existing benchmarks often focus on single-referenced entities, leaving cross-role dependencies largely unexplored. We define a taxonomy of thirteen role-gender configurations covering masculine (m), feminine (f), and neutral (n) assignments and introduce BRIDGE-MT, a manually created dataset of 351 Hindi–English sentence pairs (594 role-level instances) in order to evaluate dual-role interactions. We evaluated commercial MT systems and multilingual LLMs, and propose neutral-comparison asymmetry metrics and a conditional interaction metric to analyze cross-role dependencies. Our results show that explicitly gendered roles achieve higher $F1$ scores than neutral-labelled roles. We also observe a consistent position effect, where Role B (the second role) tends to have lower accuracy and greater gender asymmetry than Role A (the first role) across all evaluated systems. Conditional interaction analysis further indicates that the gender assigned to one role can influence the translation of the other. These findings highlight the importance of evaluating gender behavior in multi-entity settings to better understand interaction-driven asymmetries in MT.

1 Introduction

Gender bias in MT has been widely studied in occupational contexts and pronoun resolution (Stanovsky et al., 2019; Prates et al., 2019; Savoldi et al., 2021), where systems frequently default to masculine forms or exhibit asymmetric behavior across masculine and feminine entities. This has motivated the development of controlled evaluation benchmarks such as WinoMT (Stanovsky et al., 2019), MuST-SHE (Bentivogli et al., 2020), and MT-GenEval (Currey et al., 2022) for analyzing gender realization in MT systems.

Existing benchmarks, such as WinoMT

(Stanovsky et al., 2019), are primarily designed for single-entity evaluation and lack the framework for a systematic examination of interactions between multiple occupational roles in a single sentence. In these multi-entity settings, MT systems face the added complexity of maintaining grammatical coherence across several roles simultaneously. This raises a critical question regarding cross-role dependency: to what extent does the gender realization of one role influence the translation of other roles within the same sentence? Although such benchmarks may contain more than one occupational entity, they typically evaluate only the role referenced by the gendered pronoun, without systematically varying and evaluating the gender assignment of the co-occurring role. For example, in a sentence containing a masculine scientist and a feminine journalist, evaluating both roles allows examination of whether the gender assigned to one role affects the translation of the other.

To systematically investigate these interactions, we define a taxonomy of thirteen role-gender configurations covering m, f, and n assignments across two occupations, denoted as Role A and Role B. Based on this taxonomy, we present **BRIDGE-MT** (Benchmarking Role Interactions and Dependencies in Gender Evaluation for MT),[1] a manually created dataset of 351 Hindi–English sentence pairs designed for structured evaluation of role co-occurrence. Our contributions are as follows: (i) taxonomy and dataset: we define a comprehensive taxonomy and introduce BRIDGE-MT, which treats multiple roles as independent evaluation units, (ii) novel metrics: standard ΔG (Delta Gender) is defined as the difference between $F1$ scores for m and f instances (Stanovsky et al., 2019). We propose neutral-comparison asymmetry metrics that extend the ΔG formulation to contrast gendered and neutral configurations, alongside a conditional interaction metric (ΔI) to quantify cross-role dependencies. and (iii) empirical analysis and findings: we evaluated a range of commercial NMT systems and multilingual LLMs. Our results show that explicitly gendered roles are generally translated more reliably than neutral-labelled roles. We also reveal a consistent position effect, where Role B exhibits lower accuracy and higher asymmetry than Role A. Furthermore, we provide evidence of directional influence, where the gender of one role can significantly impact the translation of the other; we observe that the extent of this interaction varies significantly across different MT systems.

2 Related Work

The evaluation of gender bias in MT has evolved from identifying simple stereotyping to analyzing complex coreference resolution in structured benchmarks. Foremost among these frameworks is WinoMT, which was derived from the Winogender (Rudinger et al., 2018) and WinoBias (Zhao et al., 2018) coreference test sets. These benchmarks enable researchers to evaluate whether MT systems preserve gender information for occupational entities under controlled coreference ambiguity. Gender asymmetry is typically quantified using ΔG, defined as the difference between $F1$ scores for m and f instances (Stanovsky et al., 2019). This metric enables a structured comparison of system performance across gender classes.

Saunders et al. (2020) studied gender coreference behavior in MT and showed that gender signals can be over-generalized across entities within the same sentence. Their work focused on controlling gender realization during translation through word-level gender tags and tagged adaptation data. In contrast, our work focuses on evaluating gender behavior under co-occurring occupational roles.

Since WinoMT was originally designed with English as the source language, subsequent work adapted it for other languages. Notably, Singh (2023) introduced WinoMT-Hindi[2] to evaluate gender behavior in Hindi–English translation. However, while WinoMT-Hindi facilitates evaluation in a morphologically gendered language, its design does not systematically account for the interactions between co-occurring roles. Our work addresses this limitation by introducing a taxonomy-driven benchmark designed for the

[1] https://github.com/BridgeMT/BridgeMT

[2] https://github.com/iampushpdeep/Gender-Bias-Hi-En-Eval

structured analysis of dual-role interactions and positional dependencies in Hindi–English translation (cf. Section 3.1).

3 The BRIDGE-MT Benchmark

3.1 Limitations of WinoMT-Hindi

Despite its utility in marking source-side gender, the structural design of Singh's adaptation restricts the evaluation of multi-entity settings in several ways. First, although each sentence in WinoMT-Hindi contains two occupational entities, the English translation typically provides an explicit gender cue for only one of the roles, for e.g.:

Hindi: मजदूर होटल की रिसेप्शनिस्ट से अधिक खाया क्योंकि वह आहार पर है।
Transliteration: Mazdoor hotel ki receptionist se adhik khaya kyunki vah aahaar par hai.
English: The laborer eats more than the receptionist as he is on a diet.

In such cases, while both occupations "मजदूर" (*mazdoor, 'laborer'*), "रिसेप्शनिस्ट" (*receptionist, 'receptionist'*) are present, the gendered pronoun in the English translation is associated with only one entity "मजदूर" (*laborer*). As a result, the evaluation focuses primarily on the referenced role "मजदूर" (*laborer*). Other possible configurations of gender assignments across the two occupations are not systematically examined, leaving potential inter-role dependencies unexplored. Second, certain sentences exhibit morphological inconsistencies in gender marking, e.g.:

Hindi: निर्माण मजदूर ने स्कूल की शिक्षक को घूरा क्योंकि उसने निर्माण क्षेत्र में प्रवेश किया।
Transliteration: Nirmaan mazdoor ne school ki shikshak ko ghoora kyunki usne nirmaan kshetra mein pravesh kiya.
English: The construction worker stared at the teacher because she entered the construction zone.

Here, the feminine form "शिक्षिका" (*shikshika, 'teacher.f'*) would be expected instead of the masculine form "शिक्षक" (*shikshak, 'teacher.m'*) when referring to a feminine teacher. Such examples highlight the importance of careful morphological validation when constructing gender-aware evaluation datasets. Third, the occupation stereotype groupings in WinoMT-Hindi are derived from the original English occupation lists in WinoBias. These are based on U.S. Department of Labor statistics, as discussed in Singh (2023). While this provides a structured basis for stereotype categorization, the occupational distributions reflect a U.S.-specific context. Hindi is predominantly spoken in the Indian subcontinent (Kunchukuttan et al., 2018), where certain roles included in the original list, such as "sheriff" or "chiefs" are not widely used or do not directly correspond to culturally prevalent occupational categories. Regional differences in institutional structures and labor distributions may therefore affect how these occupations are interpreted and translated in Hindi.

These observations do not diminish the contribution of WinoMT-Hindi. Instead, they identify opportunities for a more systematic evaluation framework that covers all combinations of gender assignments across co-occurring roles, while ensuring careful lexical and morphological validation. The taxonomy-driven evaluation dataset proposed in this work is designed to address these requirements.

3.2 Taxonomy

We categorize dual-role gender assignments into thirteen distinct configurations, denoted by the grammatical markers assigned to Role A and Role B. This taxonomy provides an exhaustive framework for organizing all possible permutations of m, f, and n labels within a single sentence. In this context, n refers to cases where the source sentence does not provide an explicit gender marker for that role. It represents the absence of gender specification rather than a gender category, and does not capture non-binary or gender-diverse identities. By combining these markers across the two roles, we derive the thirteen configurations presented in Table 1.

ID	Configuration
1	A is masculine
2	A is feminine
3	B is masculine
4	B is feminine
5	A is masculine and B is neutral
6	A is feminine and B is neutral
7	A is neutral and B is masculine
8	A is neutral and B is feminine
9	Both A and B are masculine
10	A is masculine and B is feminine
11	A is feminine and B is masculine
12	Both A and B are feminine
13	Both A and B are neutral

Table 1: Gender-role configurations based on the proposed taxonomy.

This structured approach allows us to move beyond individual role accuracy and quantify how gender assignments for one entity influence the translation of the other. For example, we can examine whether the translation of Role A differs when Role B is m versus when Role B is f or n . Furthermore, symmetric configurations such as "A is m and B is f" and "A is f and B is m" allow for a balanced comparison of cross-role behavior and potential asymmetries.

Unlike benchmarks that focus on a single referenced entity per sentence, our proposed taxonomy treats both roles as independent units of evaluation. This design allows both roles to be evaluated independently and jointly, providing the basis for the conditional interaction metric introduced in Section 4.

3.3 Dataset

Based on the taxonomy described in Section 3.2, we constructed a Hindi–English evaluation dataset, BRIDGE-MT, that instantiates the defined role-gender configurations. The objective is to represent controlled gender assignments across co-occurring occupational roles within grammatically natural Hindi sentences.

We selected 27 pairs of occupational roles, denoted as Role A and Role B. These occupations are chosen because they are commonly used in Hindi-language contexts, have clear lexical realizations in both Hindi and English, and are contextually appropriate within the Indian subcontinent where Hindi is predominantly spoken.

We specifically selected only those occupations that are expressed in gender-neutral forms in Hindi. For example, the role किसान (kisan, 'farmer') can refer to both m and f individuals. By using neutral job titles such as किसान, we ensure that gender is marked through pronouns and verb agreement rather than the occupation name itself. This design choice enabled us to isolate the influence of grammatical gender cues and to measure how accurately an MT model handles them at the sentence level.

For each occupational pairing, sentences are manually constructed to reflect the intended configuration using natural sentence structures. For illustration,

- **A single-role configuration (Role A: m)** is realized as:

  Hindi: लाइब्रेरियन ने बताया कि वह सूची जाँच करेगा।
  Transliteration: Librarian ne bataya ki vah suchi jaanch karega.
  English: The librarian said that he would check the list.
  Here, the occupation "लाइब्रेरियन" (*librarian, 'librarian'*) is marked with m agreement through the verb form "करेगा" (*karega, 'karega'*).

- **A two-role configuration (Role A: m, Role B: f)** is realized as:

  Hindi: वैज्ञानिक ने बताया कि वह उपकरण जाँच करेगा जबकि पत्रकार ने कहा कि वह विश्लेषण लिखेगी।
  Transliteration: Vaigyanik ne bataya ki vah upakaran jaanch karega, jabki patrakar ne kaha ki vah vishleshan likhegi.
  English: The scientist said he would check the equipment, while the journalist said she would write the analysis.
  In this example, gender is encoded separately for each role through verb agreement: m through "करेगा" (*karega, 'karega'*) for "वैज्ञानिक" (*vaigyanik, 'scientist'*) and f through "लिखेगी" (*likhegi, 'write'*) for "पत्रकार" (*patrakar, 'journalist'*).

- **A mixed configuration (Role A: m, Role B: n)** is realized as:

  Hindi: न्यायाधीश ने कहा कि वह निर्णय कल सुनाएगा, तभी वकील गवाही के काग़ज़ जुटाने में व्यस्त था।

Transliteration: Nyayadhish ne kaha ki vah nirnay kal sunayega, tabhi vakil gavahi ke kagaz jutane mein vyast tha.
English: The judge said he would deliver the verdict tomorrow, while the lawyer was busy collecting the evidence papers.

In this example, Role A "न्यायाधीश" (*nyayadhish, 'judge'*) is marked with m agreement through the verb form "सुनाएगा"(*sunayega, 'would deliver'*), while Role B "वकील" (*vakil, 'lawyer'*) appears without an explicit gender-specific pronoun. Therefore, Role B is treated as n within the taxonomy because it does not have a clear gender marking.

The dataset consists of 351 Hindi–English sentence pairs, created from 27 role pairs across 13 configurations. It is designed as a controlled evaluation benchmark rather than a large-scale corpus, with the aim of analyzing interaction-driven gender behavior. Evaluation is conducted at the role level, meaning that sentences containing two occupational roles provide two independent gender predictions. This results in 594 role-level evaluation instances to compare across m, f, and n assignments.

A randomly sampled subset comprising 10% of the source Hindi sentences was independently translated by a second bilingual translator. These translations were then compared with the original references to verify agreement and gender consistency. No discrepancies were observed in the sampled subset.

4 Measuring Interactions and Dependencies

Existing benchmarks such as WinoMT and WinoMT-Hindi evaluate gender behavior primarily through aggregate metrics such as accuracy and ΔG. While this provides insight into overall gender asymmetry, it does not capture dependencies between co-occurring roles within the same sentence.

Since our dataset contains sentences with two occupational roles, it enables conditional analysis. In such settings, the translation behavior for Role A may depend on the gender assignment of Role B, and vice versa. To examine this phenomenon, we propose a conditional interaction metric, denoted as ΔI.

However, interaction can only be meaningfully measured when both roles are explicitly gendered. If one role is neutral or absent, cross-role dependency cannot be evaluated. For this reason, the interaction metric is computed only on the subset of sentences where both Role A and Role B are assigned explicit m or f gender labels.

4.1 Role-Level Evaluation

In our dataset, the intended gender of each role is predefined by the taxonomy configuration (cf. Section 3.2). We refer to this predefined assignment as the gold gender of the role.

Evaluation is conducted at the role level. For each role $r \in \{A, B\}$, we compute $F1$ scores separately for m and f predictions by comparing system outputs against the gold gender label. Following prior work, overall gender asymmetry for a role is quantified using (1):

$$\Delta G_r^{m-f} = F1_r(m) - F1_r(f) \tag{1}$$

While evaluating MT systems, the absolute value of ΔG_r^{m-f} should ideally be close to zero, indicating balanced performance across m and f instances. In addition to the standard measure, ΔG_r^{m-f}, we introduce two neutral-comparison asymmetry measures that incorporate neutral-labelled configurations, as in (2) and (3):

$$\Delta G_r^{m-n} = F1_r(m) - F1_r(\mathrm{n}) \tag{2}$$

$$\Delta G_r^{f-n} = F1_r(f) - F1_r(n) \tag{3}$$

These measures extend the standard ΔG_r^{m-f} formulation by replacing one gender class with n instances. They allow us to quantify how system performance on explicitly gendered roles differs from that in n configurations. To the best of our knowledge, neutral-comparison asymmetry measures have not yet been explicitly studied in MT evaluation. While these additional comparisons provide further insight into role-level behavior, they are not used in the computation of the proposed conditional interaction metric.

4.2 Conditional Interaction Definition

To measure interaction, we condition the bias of one role on the gold gender of the other role, using only the fully gendered configurations: (i) both A and B are masculine, (ii) A is masculine

and B is feminine, (iii) A is feminine and B is masculine, and (iv) both A and B are feminine.

For Role B, we compute conditional bias values using (4) and (5):

$$\Delta G_{B|A=m} = F1_{(B=m|A=m)} - F1_{(B=f|A=m)} \quad (4)$$

$$\Delta G_{B|A=f} = F1_{(B=m|A=f)} - F1_{(B=f|A=f)} \quad (5)$$

Similarly, for Role A, we compute $\Delta G_{A|B=m}$ and $\Delta G_{A|B=f}$.

Then the interaction terms are computed using (6) and (7):

$$\Delta I_B = \Delta G_{B|A=m} - \Delta G_{B|A=f} \quad (6)$$

$$\Delta I_A = \Delta G_{A|B=m} - \Delta G_{A|B=f} \quad (7)$$

These quantities effectively measure how the bias associated with one role influences the gender of the co-occurring role. Our proposed interaction metric is interpreted as follows: (i) (i) if $\Delta I_A = 0$ and $\Delta I_B \neq 0$, Role A influences Role B, (ii) if $\Delta I_A \neq 0$ and $\Delta I_B = 0$, Role B influences Role A, (iii) if both are non-zero, mutual interference is present, and (iv) if both are zero, no measurable interaction is observed. Unlike ΔG, which measures overall asymmetry for a single role, ΔI captures cross-role dependency under controlled dual-gender configurations.

5 Experimental Setups

To evaluate gender behavior within our proposed taxonomy, we examined a range of Hindi–English MT systems representing both commercial production environments and academic research. On the commercial side, we evaluated Google Translate[3] and Amazon Translate,[4] which are production-level NMT systems available via public interfaces and APIs. We also evaluated IndicTrans2,[5](Gala et al., 2023) an open-source multilingual MT system. In addition, we evaluated four instruction-tuned LLMs capable of performing translation via prompting: Llama 3.1 8B Instruct[6] (Grattafiori et al., 2024), Qwen3-4B[7](Yang et al., 2025), TranslateGemma-12B-IT[8](Finkelstein et al., 2026), and Gemma-3-4B-IT[9](Team et al., 2025). This selection allows us to compare traditional encoder-decoder MT systems and LLMs under the same controlled evaluation setting.

Each system translates all 351 Hindi sentences into English. Each sentence is translated independently, and the outputs are used directly for evaluation without post-editing or manual correction. For LLMs, translation is performed in a controlled setting to minimize variations. Models are prompted to produce only the English translation of the Hindi input, without additional instructions or paraphrasing. Sampling-based decoding methods are disabled, and decoding parameters are set to ensure consistent outputs. As a result, each model generates a single translation per sentence. This controlled setup reduces decoding-related variance, making model differences easier to interpret.

5.1 Gender Extraction and Role-Level Scoring with Skipped Cases

Evaluation follows the role-level framework defined in Section 4.1. For each role, the gold gender is determined directly from the taxonomy, while the predicted gender is extracted from the translated English output, and identified using explicit markers (e.g., he, him, and his) for m and (e.g., she, her, and hers) for f in the English translations. If no clear marker is present in configurations where gender is specified, the instance is treated as failing to preserve the intended gender. This includes cases where the system produces gender-neutral paraphrases that omit pronouns entirely.

Note that neutral-labelled configurations were also included in the role-level evaluation using the metrics ΔG_r^{m-n} and ΔG_r^{f-n} defined

[3]https://translate.google.com
[4]https://aws.amazon.com/translate
[5]https://github.com/AI4Bharat/IndicTrans2
[6]https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[7]https://huggingface.co/Qwen/Qwen3-4B
[8]https://huggingface.co/google/translategemma-12b-it
[9]https://huggingface.co/google/gemma-3-4b-it

in Section 4.1. Additionally, if a system's output does not contain a lexical form that corresponds to the gold occupation, the instance is marked as a skipped case and excluded from scoring. These instances typically occur when the target occupation undergoes mistranslation, substitution, or lexical alteration, resulting in a loss of semantic correspondence with the intended role. For instance, consider the following translation:

Source sentence (Hindi): शेफ़ ने संकेत किया कि वह स्वाद-परख पूरी करेगा, तदनंतर सचिव ने बताया कि वह विवरण-पत्र को अंतिम रूप देगी।
Transliteration: Chef ne sanket kiya ki vah swaad-parakh poori karega, tadanantar sachiv ne bataya ki vah vivaran-patra ko antim roop degi.
English (Gold): The chef indicated that he would complete the tasting, after which the secretary said she would finalize the detail sheet.
Qwen3-4B Output: Sheikh indicated that he would complete the taste test, and then the secretary said that she would finalize the details letter.

In this case, Role A ("chef") is mistranslated as "Sheikh". Since the translated word does not correspond to the intended occupation, the gender associated with that role cannot be reliably evaluated. Therefore, Role A is treated as a skipped case and excluded from scoring.

For role-level metrics, valid roles are evaluated independently even if the other role in the sentence is skipped. However, for the conditional interaction metric, both roles must be correctly identified. If either role is skipped, the sentence is excluded from interaction analysis.

5.2 Evaluation Metrics

We measured role-level accuracy and gender asymmetry ΔG^{m-f}, following the standard benchmark evaluation framework. We computed ΔG^{m-n} and ΔG^{f-n} to analyze performance differences between gendered and neutral configurations. The interaction metric (ΔI), defined in Section 4, was computed on the fully gendered dual-role subset. To provide an overall measure of translation quality, we additionally report BLEU (Papineni et al., 2002), chrF++ (Popović, 2017), and COMET (Rei et al., 2020) scores for all evaluated systems.

System	chrF++	BLEU	COMET
Google Translate	**63.61**	**41.95**	**0.8210**
Amazon Translate	59.06	34.42	0.8125
IndicTrans2	60.13	37.16	0.8118
Llama 3.1 8B Inst.	56.67	29.54	0.8020
Qwen3-4B	55.02	29.82	0.8028
TranslateGemma-12B	57.13	29.80	0.8153
Gemma-3-4B	57.91	32.15	0.8099

Table 2: Performance of the MT systems on the BRIDGE-MT dataset.

6 Results and Discussions

We report results for all evaluated systems on the BRIDGE-MT dataset, using evaluation metrics discussed in Section 5.2. Table 2 presents translation quality results and serves as a general performance baseline across systems. While all systems maintain relatively high translation quality, there is notable variance in scores between the top-performing commercial engines and LLMs. Tables 3, 4, and 5 summarise overall role-level performance (combined and role-wise), while Table 6 presents conditional interaction results on the fully gendered subset.

6.1 Overall Role-Level Performance

Table 3 presents overall results across both roles. Among the evaluated MT systems, Google Translate achieves the highest overall accuracy (77.21%), followed by IndicTrans2 (74.16%) and Amazon Translate (70.82%). Google Translate and IndicTrans2 exhibit nearly balanced performance across m and f instances, with ΔG^{m-f} values close to zero (-0.0005 and 0.0022, respectively). In contrast, Llama 3.1 8B Instruct shows a large positive ΔG^{m-f} (0.2336), reflecting substantially higher $F1$ for m instances than f instances. Gemma-3-4B-IT also exhibits noticeable asymmetry ($\Delta G^{m-f} = 0.1293$). Qwen3-4B and TranslateGemma-12B-IT show smaller but still positive ΔG^{m-f} values.

Across all systems, performance on neutral instances is lower than on explicitly gendered cases. This suggests that the absence of explicit gender cues introduces additional difficulty in translation. Skipped cases remain relatively low for most systems, with the exception of Qwen3-4B (71 skipped cases) and Llama 3.1 8B Instruct

System	Acc (%)	$F1$(m)	$F1$(f)	$F1$(n)	ΔG^{m-f}	ΔG^{m-n}	ΔG^{f-n}	Skipped
Google Translate	**77.21**	**0.8438**	**0.8443**	0.4486	**-0.0005**	0.3952	0.3958	6
Amazon Translate	70.82	0.7717	0.7524	0.4579	0.0194	0.3138	0.2944	1
IndicTrans2	74.16	0.8048	0.8026	0.4706	0.0022	0.3342	0.3320	2
Llama 3.1 8B Instruct	51.05	0.6028	0.3691	0.4278	0.2336	**0.1749**	**-0.0587**	26
Qwen3-4B	69.79	0.7561	0.7362	**0.4848**	0.0199	0.2713	0.2514	71
TranslateGemma-12B-IT	68.13	0.7524	0.7378	0.4298	0.0146	0.3227	0.3081	1
Gemma-3-4B-IT	54.82	0.6293	0.5000	0.4240	0.1293	0.2053	0.0760	3

Table 3: Overall role-level results across both roles combined (594 instances). Accuracy, $F1$ scores for masculine (m), feminine (f), and neutral (n) instances, overall gender asymmetry (ΔG), neutral-comparison asymmetry measures (ΔG^{m-n}, ΔG^{f-n}), and number of skipped cases are reported.

System	Acc (%)	$F1$(m)	$F1$(f)	$F1$(n)	ΔG^{m-f}	ΔG^{m-n}	ΔG^{f-n}	Skipped
Google Translate	80.28	**0.8770**	0.8710	0.4285	**0.005**	0.4485	0.4425	1
Amazon Translate	77.49	0.8307	0.8358	0.4587	-0.0051	0.3721	0.3771	0
IndicTrans2	**81.66**	0.8733	**0.8888**	0.4727	-0.0155	0.4006	0.4161	2
Llama 3.1 8B Instruct	57.49	0.6560	0.4888	0.4242	0.1672	**0.2318**	**0.0646**	24
Qwen3-4B	72.96	0.7875	0.7665	**0.4800**	0.0209	0.3075	0.2865	44
TranslateGemma-12B-IT	72.57	0.7973	0.7777	0.4354	0.0196	0.3619	0.3422	1
Gemma-3-4B-IT	56.73	0.6611	0.5192	0.3970	0.1417	0.2639	0.1222	2

Table 4: Role-wise evaluation: Role A only

System	Acc (%)	$F1$(m)	$F1$(f)	$F1$(n)	ΔG^{m-f}	ΔG^{m-n}	ΔG^{f-n}	Skipped
Google Translate	**72.69**	**0.7917**	**0.8022**	0.4706	**-0.0105**	0.3210	0.3316	5
Amazon Translate	61.16	0.6872	0.6053	0.4571	0.0820	0.2301	0.1481	1
IndicTrans2	63.37	0.7029	0.6589	0.4684	0.0440	0.2345	0.1904	0
Llama 3.1 8B Instruct	42.32	0.5267	0.1864	0.4313	0.3402	**0.0953**	-0.2449	2
Qwen3-4B	65.28	0.7106	0.6861	**0.4897**	0.0245	0.2208	0.1963	27
TranslateGemma-12B-IT	61.72	0.6859	0.6708	0.4237	0.0151	0.2622	0.2470	0
Gemma-3-4B-IT	52.07	0.5812	0.4706	0.4561	0.1106	0.1251	**0.0144**	1

Table 5: Role-wise evaluation: Role B only

System	$\Delta G_{B\|A=\text{m}}$	$\Delta G_{B\|A=\text{f}}$	ΔI_B	$\Delta G_{A\|B=\text{m}}$	$\Delta G_{A\|B=\text{f}}$	ΔI_A
Google Translate	**0.0333**	**-0.0014**	**0.0348**	0.0043	**0.0000**	**0.0043**
Amazon Translate	0.6036	0.0168	0.5868	0.0342	0.0231	0.0111
IndicTrans2	0.3105	-0.0387	0.3492	**-0.0077**	-0.0014	-0.0063
Llama 3.1 8B Instruct	0.2778	0.4703	-0.1926	0.1065	0.0768	0.0296
Qwen3-4B	0.0439	0.0873	-0.0435	0.0108	0.0366	-0.0258
TranslateGemma-12B-IT	0.1357	0.0015	0.1342	0.0687	-0.0222	0.0909
Gemma-3-4B-IT	0.2456	0.1392	0.1063	0.2193	0.1576	0.0618

Table 6: Conditional Interaction Results (Fully Gendered Subset)

(26 skipped cases).

6.2 Role-Wise Analysis

Tables 4 and 5 present results for Role A and Role B respectively. This breakdown allows us to examine whether systems behave differently depending on entity position. For Role A (Table 4), MT systems perform strongly. IndicTrans2 achieves the highest accuracy (81.66%), followed closely by Google Translate (80.28%) and Amazon Translate (77.49%). ΔG^{m-f} values for these systems are small in magnitude, indicating balanced *m-f* behavior for the first role. As for Role B (Table 5), performance de-

creases across all systems. Google Translate accuracy drops from 80.28% (Role A) to 72.69% (Role B). Amazon Translate shows a substantial decrease (77.49% to 61.16%), and IndicTrans2 similarly declines (81.66% to 63.37%).

Overall, the role-wise breakdown reveals a consistent position effect across systems. Role B not only achieves lower accuracy than Role A but also, in several cases, exhibits larger gender asymmetry. This observation motivates the need for conditional interaction analysis.

6.3 Conditional Interaction Results

Table 6 presents conditional interaction scores computed on the fully gendered subset. For Google Translate, both ΔI_B (0.0348) and ΔI_A (0.0043) are close to zero. This indicates relatively stable behavior, as the gender assigned to one role has minimal effect on the gender realization of the other role. Amazon Translate, however, shows a very large ΔI_B value (0.5868), while ΔI_A remains small (0.0111). This suggests strong directional interaction, with the gender of Role A substantially influencing the gender realization of Role B, but not vice versa. IndicTrans2 also shows noticeable interaction for Role B ($\Delta I_B = 0.3492$), though smaller than Amazon Translate. Interaction for Role A remains negligible (-0.0063). Among LLM-based systems, interaction patterns vary. Some models show moderate interaction in both directions, while others show limited dependency between roles.

A consistent pattern across several systems is that $|\Delta I_B|$ tends to be larger than $|\Delta I_A|$. This aligns with the earlier role-wise findings, which indicated that Role B showed greater instability and lower accuracy. Overall, the second role in the sentence appears more sensitive to the gender of the first role.

7 Conclusions

In this paper, we investigated gender bias in Hindi–English MT within multi-entity contexts. By introducing a structured taxonomy and the BRIDGE-MT dataset, we moved beyond single-role evaluation to quantify how co-occurring occupational roles influence translation outcomes.

Our empirical analysis showed that MT systems perform more consistently on explicitly gendered configurations than on neutral ones. We also observed a consistent position effect: Role B typically shows lower accuracy and, in several cases, larger gender imbalance than Role A. Conditional interaction analysis further suggests that cross-role effects vary across the MT systems, with some systems exhibiting stable behavior and others showing strong directional influence, most often from Role A to Role B.

Overall, this work highlights the importance of evaluating gender behavior beyond single-entity settings, where dependencies between roles may affect translation outcomes.

As future work, we plan to explore mitigation strategies for reducing interaction-driven asymmetries and to extend this analysis to additional languages, including other gender-marking language pairs and larger datasets.

References

Bentivogli, L., Savoldi, B., Negri, M., Di Gangi, M. A., Cattoni, R., and Turchi, M. (2020). Gender in danger? evaluating speech translation technology on the MuST-SHE corpus. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6923–6933, Online. Association for Computational Linguistics.

Currey, A., Nadejde, M., Pappagari, R. R., Mayer, M., Lauly, S., Niu, X., Hsu, B., and Dinu, G. (2022). MT-GenEval: A counterfactual and contextual dataset for evaluating gender accuracy in machine translation. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4287–4299, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Finkelstein, M., Caswell, I., Domhan, T., Peter, J.-T., Juraska, J., Riley, P., Deutsch, D., Kovacs, G., Dilanni, C., Cherry, C., et al. (2026). Translategemma technical report. *arXiv preprint arXiv:2601.09012*.

Gala, J., Chitale, P. A., Ak, R., Gumma, V., Doddapaneni, S., Kumar, A., Nawale, J., Sujatha, A., Puduppully, R., Raghavan, V., et al.

(2023). Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *arXiv preprint arXiv:2305.16307*.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al. (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Kunchukuttan, A., Mehta, P., and Bhattacharyya, P. (2018). The IIT Bombay English-Hindi parallel corpus. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In Isabelle, P., Charniak, E., and Lin, D., editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Popović, M. (2017). chrF++: words helping character n-grams. In Bojar, O., Buck, C., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., and Kreutzer, J., editors, *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Prates, M. O. R., Avelar, P. H. C., and Lamb, L. (2019). Assessing gender bias in machine translation – a case study with google translate.

Rei, R., Stewart, C., Farinha, A. C., and Lavie, A. (2020). COMET: A neural framework for MT evaluation. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Rudinger, R., Naradowsky, J., Leonard, B., and Van Durme, B. (2018). Gender bias in coreference resolution. In Walker, M., Ji, H., and Stent, A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Saunders, D., Sallis, R., and Byrne, B. (2020). Neural machine translation doesn't translate gender coreference right unless you make it. In Costa-jussà, M. R., Hardmeier, C., Radford, W., and Webster, K., editors, *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 35–43, Barcelona, Spain (Online). Association for Computational Linguistics.

Savoldi, B., Gaido, M., Bentivogli, L., Negri, M., and Turchi, M. (2021). Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.

Singh, P. (2023). Gender inflected or bias inflicted: On using grammatical gender cues for bias evaluation in machine translation. In Li, D., Mahendra, R., Tang, Z. P., Jang, H., Murawaki, Y., and Wong, D. F., editors, *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*, pages 17–23, Nusa Dua, Bali. Association for Computational Linguistics.

Stanovsky, G., Smith, N. A., and Zettlemoyer, L. (2019). Evaluating gender bias in machine translation. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Team, G., Kamath, A., Ferret, J., et al. (2025). Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Yang, A., Li, A., Yang, B., et al. (2025). Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2018). Gender bias in coreference resolution: Evaluation and debiasing methods. In Walker, M., Ji, H., and Stent, A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.