Gender Bias in Nepali-English Machine Translation: A Comparison of LLMs and Existing MT Systems

Supriya Khadka[*]
Coventry University
Coventry, United Kingdom
khadkas25@uni.coventry.ac.uk

Bijayan Bhattarai
Diyo.AI
Jwagal, Lalitpur, Nepal
bijayan.bhattarai@diyo.ai

Abstract

Bias in Nepali NLP is rarely addressed, as the language is classified as low-resource, which leads to the perpetuation of biases in downstream systems. Our research focuses on gender bias in Nepali-English machine translation, an area that has seen little exploration. With the emergence of Large Language Models (LLMs), there is a unique opportunity to mitigate these biases. In this study, we quantify and evaluate gender bias by constructing an occupation corpus and adapting three gender-bias challenge sets for Nepali. Our findings reveal that gender bias is prevalent in existing translation systems, with translations often reinforcing stereotypes and misrepresenting gender-specific roles. However, LLMs perform significantly better in both gender-neutral and gender-specific contexts, demonstrating less bias compared to traditional machine translation systems. Despite some quirks, LLMs offer a promising alternative for culture-rich, low-resource languages like Nepali. We also explore how LLMs can improve gender accuracy and mitigate biases in occupational terms, providing a more equitable translation experience. Our work contributes to the growing effort to reduce biases in machine translation and highlights the potential of LLMs to address bias in low-resource languages, paving the way for more inclusive and accurate translation systems.

1 Introduction

Based on Stahlberg et al. (2011), Nepali is a grammatical gender language, unlike English, which is a notional gender language. In Nepali, verbs and adjectives carry gender inflections, while pronouns indicate formality, affecting the verb form. For example, "He/She is tall" translates to उनी अग्ली छिन् (*oo-ni uglee chhinn*) for females and उनी अग्ला छन् (*oo-ni uglaa chhann*) for males in a familiar setting. The pronoun changes for different levels of formality, altering the verb and adjective accordingly. The most formal third-person pronoun, उहाँ (*oo-haan*), uses a gender-neutral verb, while other pronouns use gendered verbs. There have been extensive studies on gender bias in translation for grammatical gender languages (Stanovsky et al., 2019; Vanmassenhove and Monti, 2021; Ghosh and Caliskan, 2023), but Nepali remains unexplored. Due to Nepali's low-resource status (Shahi and Sitaula, 2022), the focus has traditionally been on improving translation accuracy, often neglecting issues of bias. This can result in fluent yet biased outputs, reinforcing stereotypes and prejudices over time (Savoldi et al., 2021).

We define "bias" as the systematic and unfair representation of one gender over another in translation outputs. In this study, we consider only two genders: male and female. The inclusion of other genders is beyond the scope of this work. Our experiments identify bias in three ways: reinforcement of gender stereotypes, incorrect gender assignments to neutral and opposite-gendered terms, and unequal translation accuracy across genders. As highlighted by Blodgett et al. (2020), these biases can cause significant harm, particularly by reinforcing stereotypes. In Nepali-English translation, this is evident in how systems often associate occupations with specific genders, use respectful pronouns predominantly for men, and fail to properly represent women in high-ranking positions.

Our work aims to study and evaluate these biases in Nepali-English machine translation. Our major contributions are:

- Adapting three benchmarks to evaluate gender bias in Ne-En machine translation and creating a Nepali occupations corpus.
- Assessing gender bias in Ne-En machine translation for gender-neutral and gender-specific contexts.

[*]Work done while at Diyo.AI

Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 75–82

- Highlighting how LLMs are promising alternatives to existing MT systems.

Data and code are publicly available.[1]

2 Experimental Setup

MT Systems

We begin our test with two Ne-En MT systems: Google Translate (GT)[2], a proprietary MT system, and IndicTrans2 (IT2), an open-source MT system (Gala et al., 2023). We selected IT2 as the open-source representative because it is specifically trained for Indic languages, including Nepali. Additionally, we include LLMs: OpenAI's GPT-3.5, GPT-4o (an advanced version of GPT-4 (Achiam et al., 2023)), and BigScience's BLOOM (Le Scao et al., 2023). We select BLOOM, a multilingual LLM trained on a diverse set of languages, for its ability to understand and generate Nepali text. Due to our limited computational resources, we use its 7b variant. OpenAI's models are accessed via API. To convert LLMs into translators, we use the instruction:

You are a translator who translates the user input from Nepali to English.

We evaluate systems using BLEU scores on the FLORES200 (Costa-jussà et al., 2022), IN22-Gen (Gala et al., 2023), and IN22-Conv (Gala et al., 2023) benchmarks and observe below par performance for BLOOM-7b and GPT-3.5 as reported in Table 1. Due to this, for rest of the experiments, GT, IT2 and GPT-4o translator are selected.

3 Approach

3.1 Gender Neutral Approach

The Translation Gender Bias Index (TGBI), introduced by Cho et al. (2019) for Korean-English translation, evaluates bias in gender-neutral pronouns using phrase sets with positive/negative expressions and occupations. Ramesh et al. (2021) adapted TGBI for Hindi-English translation using gender-neutral third-person pronouns वह (*vah*), वे (*ve*), and वो (*vo*). Similarly, in Nepali, we use third-person pronouns उहाँ (*oo-haan*), उनी (*oo-ni*), and ऊ (*oo*) to build our TGBI dataset, corresponding to formal polite (honorary), formal impolite (familiar), and informal (colloquial) settings.[3]

[1] https://github.com/anon-sketch/En-Ne_GenderBiasEval

[2] https://translate.google.com/

[3] Hereafter we will refer *formal polite* as *formal*, *formal impolite* as *familiar* and *informal* as it is.

	FLORES200	IN22-G	IN22-C
GT	**46.51***	**46.82***	**43.14***
IT2	**46.29**	**45.13**	**42.38**
GPT-3.5	26.11	27.30	28.42
GPT-4o	**41.57**	**43.71**	**41.02**
bloom-7b	15.51	15.42	21.24

Table 1: BLEU score evaluation on 3 Ne-En benchmarks: Bold indicates the top three highest scores and the selected translators. * denotes the highest score.

Unlike Hindi, Nepali verbs vary by formality. For example, "She is a farmer" translates to उहाँ किसान हुनुहुन्छ (*oo-haan kisaan hunu-hunchha*), उनी किसान हुन् (*oo-ni kisaan hunn*), and ऊ किसान हो (*oo kisaan ho*) for formal, familiar, and informal contexts, respectively. We used these variations and a corpus of sentiment words and occupations to build the Equity Evaluation Corpus-Nepali (EEC-Nepali).

3.1.1 Corpus Construction

Sentiment Word Corpus

To create the sentiment word corpus, we translated 600 negative and 533 positive sentiment words from Ramesh et al. (2021) in Hindi to Nepali using Google Translate. These translations were then manually checked for errors and mis-translations by the authors, who are native Nepali speakers fluent in Hindi.

Occupation Corpus

The occupation corpus was generated through three methods. First, we translated the list from Cho et al. (2019) to Nepali and manually checked for errors by the authors, yielding 955 unique occupations. Since this list, derived from an official Korean employment site, wasn't fully relevant to the Nepali context, we supplemented it by creating our own employment corpus from two additional sources.

We constructed our initial employment corpus by extracting data from the *finance*, *forestry*, *agriculture*, *education*, and *miscellaneous* divisions of the Public Service Commission (PSC)[4] in Nepal. Due to Unicode font incompatibilities in Nepali official documents, we used OCR for text extraction. Paudel et al. (2024) demonstrated that Pytesseract[5] provides the best results for Nepali documents, so we chose it. We also incorporated job titles and ranks from the Nepal Army and Nepal Armed Po-

[4] https://psc.gov.np

[5] https://pypi.org/project/pytesseract/

	PSC Corpus	NTO Corpus
GT	14.64	22.86
IT2	15.26	24.13
GPT-4o	**5.60**	**9.52**

Table 2: Translation Error Rate for Nepali Occupations

lice Force, yielding a corpus of 321 unique occupations (PSC Corpus).

Apart from official job titles, Nepal boasts a rich array of traditional occupations spanning centuries. Many people adopted family names based on these roles, such as ताम्रकार (*taamra-kaar* - coppersmith) and स्वर्णकार (*swarna-kaar* - goldsmith). Nepali has also borrowed occupation names from various languages spoken within Nepal. For instance, मजदुर (*majdur*) and ज्यामी (*jyaami*) both denote daily-wage laborers, with the latter originating from the Newar (*Nepalbhasa)* language. The same occupation can have multiple names based on historical periods, cultural contexts, and linguistic backgrounds. For instance, a carpenter can be referred to as सिकर्मी (*sikarmi*), तक्षक (*takshak*), दारु (*daaru*), or काष्ठकर्मी (*kaastha-karmi*). Nepal's diverse religious history has led to various names for different types of priests: महन्त (*mahanta*) serves as the chief priest, सूत (*soot*) historically performed rituals for the king, and धामी (*dhaami*) refers to shamans and priests of the Dhimal caste. Attempting to classify all these occupations under a single term like "priest" would oversimplify and diminish their rich contextual nuances. We compiled a distinct corpus of these traditional Nepali occupations, totaling 314 unique entries (NTO Corpus), sourced from the Nepali Brihat Shabhakosh.[6]

EEC-Nepali Compilation

To ensure accurate evaluation of gender bias, we tested selected MT systems to determine their ability to translate various Nepali occupations. This preliminary test included both the PSC-corpus and NTO-corpus. We manually reviewed the translations and reported error rates for each translator in Table 2.

GPT-4o consistently outperformed GT and IT2 across both corpora. One significant advantage it offered is contextual understanding. For instance, the occupation लाहुरे (*laahure*) from the NTO corpus was not translated by GT and IT2, but GPT-4o provided a translation with additional context:

लाहुरे - *Soldier (specifically referring to those who served in the British/Indian armies)*

To ensure consistency in our gender bias assessment, we only included words recognized by all translators. This resulted in 261 commonly recognized words in the PSC corpus and 221 in the NTO corpus. The final EEC-Nepali corpus consists of six sets of gender-neutral sentences: positive (S1), negative (S2), occupation (S3), informal (S4), familiar (S5), and formal (S6).

3.1.2 TGBI Metric Modification

The Translation Gender Bias Index (TGBI) measures how sentences in a set S are translated as masculine (p_m), feminine (p_f), or neutral (p_n) in the target language. Here, p represents the proportion of sentences translated into each gender category. In this context, "neutral" includes terms such as "the person". The formula for P_S, as proposed by Cho et al. (2019) is

$$P_S = \sqrt{p_m * p_f + p_n} \tag{1}$$

where

$$\begin{aligned} p_m + p_f + p_n &= 1 \\ 0 \leq p_m, p_f, p_n &\leq 1 \end{aligned} \tag{2}$$

With the rise of LLMs, translating gender-neutral terms into both masculine and feminine forms has become more feasible. While Google Translate has provided both feminine and masculine translations since 2018 for some gender-neutral languages (not including Nepali yet) (Kuczmarski, 2018; Johnson, 2020), LLMs like GPT-4o can handle this task effectively. To adapt the TGBI formula to accommodate both he/she aspects, we modify it as follows:

$$p_m^{'} + p_f^{'} + p_n = 1 \tag{3}$$

Here, $p_m^{'}$ and $p_f^{'}$ cover all mentions of males and females, including instances where both are mentioned.

$$(p_m + p_f) - p_{both} + p_n = 1 \tag{4}$$

Hence, p_{both} representing sentences containing both genders, is calculated as:

$$p_{both} = p_m + p_f + p_n - 1 \tag{5}$$

[6] https://archive.org/download/nepali-brihat-sabdkosh/

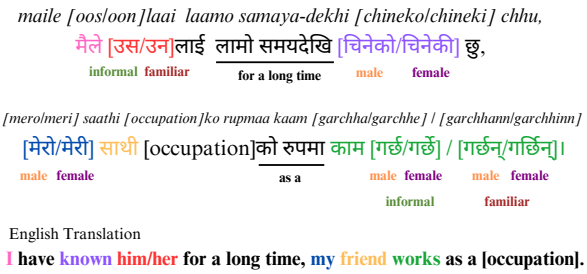

Figure 1: OTSC-Nepali Creation Process

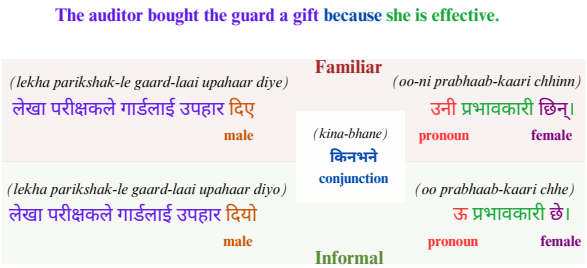

Figure 2: WinoMT-Nepali Creation Process

3.2 Simple Gender-Specific Context

Escudé Font and Costa-jussà (2019) introduced a test set using custom sentences to assess gender bias in English-Spanish translation with the pattern: *"I've known {her, him, <proper noun>} for a long time, my friend works as {a, an} <occupation>."* across various professional fields. Building on this, Singh (2023) adapted the approach for Hindi, incorporating gender-inflected possessive pronouns. In Nepali, a similar pattern is observed, but with an additional nuance: the formality of the third-person pronoun influences the action verb.

To address these nuances, we propose *OTSC-Nepali*, featuring eight sets of sentences. These sets include variations using familiar and informal third-person pronouns in four combinations of male and female for both the speaker and the friend. The formal third-person pronoun is excluded because it employs the same verb form for all genders, making it unsuitable for measuring gender-specific context. We used the filtered occupation list created in Section 3.1.1. Each of these occupations contributes to constructing the eight sets, with 1296 sentences in each set, where we analyze the percentage of sentences translating the speaker's friend as male or female as p_m and p_f respectively. The detailed creation process is shown in Figure 1.

3.3 Complex Gender-Specific Context

Stanovsky et al. (2019) introduced the *WinoMT* challenge set, pioneering gender bias analysis in machine translation. It combines *Winogender* (Rudinger et al., 2018) and *WinoBias* (Zhao et al., 2018) coreference resolution datasets. *WinoMT* includes two sets of sentences balanced across male and female genders, as well as stereotypical and non-stereotypical gender-role assignments. Adapting *WinoMT* for Nepali, we developed the *WinoMT-Nepali* challenge set to assess bias in Ne-En MT systems.

To create our challenge set, direct translation of *WinoMT* into Nepali was not feasible due to existing MT systems' limitations in handling complex English sentences accurately and tendency to translate towards more stereotypical roles, undermining our study's purpose. Therefore, for *WinoMT-Nepali*, each sentence was divided at the conjunction. Both halves were first automatically translated using Google Translate, then manually checked for grammatical consistency and gender mismatches against the original *WinoMT*. Similar to *OTSC-Nepali*, the challenge set includes familiar and informal third-person pronouns, as illustrated in Figure 2.

We generated four sets of sentences: anti and pro-stereotypical for familiar and informal contexts, each containing 1497 sentences. For gender bias evaluation, we use the same metrics proposed by Stanovsky et al. (2019): Acc measures correctness of gender labels post-translation, Δ_G indicates performance differences (F_1 score) between male and female translations, and Δ_S measures differences between stereotypical and non-stereotypical gender roles. In adapting *WinoMT* for Hi-En MT, Singh (2023) noted some sentences translated into gender-neutral forms. Our experiments with GPT-4o revealed a notable percentage of gender-neutral translations, detailed in Section 4.3. We report the percentage of gender-neutral sentences as N.

4 Results and Discussion

4.1 Evaluation using EEC-Nepali

We presented three scores from the EEC-Nepali corpus evaluation in Table 3: the average P_S for each sentence set (TGBI), the fraction of sentences

Sentence	Size	GT	IT2	GPT-4o
		$P_S(p_f, p_{both})$	$P_S(p_f, p_{both})$	$P_S(p_f, p_{both})$
Positive (S1)	1732	0.308 (0.098, 0.001)	0.205 (0.022, 0.004)	**0.571 (0.380, 0.159)**
Negative (S2)	1802	0.294 (0.085, 0.000)	0.176 (0.007, 0.003)	**0.509 (0.277, 0.098)**
Occupation (S3)	2994	0.278 (0.081, 0.000)	0.173 (0.023, 0.001)	**0.470 (0.278, 0.042)**
Informal (S4)	2176	0.123 (0.008, 0.000)	0.195 (0.013, 0.004)	**0.362 (0.129, 0.108)**
Familiar (S5)	2176	0.436 (0.248, 0.000)	0.230 (0.039, 0.011)	**0.531 (0.646, 0.038)**
Formal (S6)	2176	0.098 (0.004, 0.000)	0.093 (0.003, 0.004)	**0.373 (0.139, 0.120)**
Average		0.256	0.179	**0.469**

Table 3: Evaluation on EEC-Nepali test set. Here $P_S(p_f, p_{both})$ are TGBI value (fraction of feminine sentences, fraction of sentences with both masculine and feminine words) respectively. The average TGBI is calculated in the last row. Bold represents highest P_S for each sentence set. Underline represents highest P_S for each translator.

translated as feminine (p_f), and and the fraction translated as both (p_{both}). GT and IT2 demonstrate stronger biases towards masculine translations, whereas GPT-4o shows a higher proportion of gender-neutral translations. Our result indicates that GPT-4o is the least biased system overall, particularly in positive, negative, and occupational sentence sets, suggesting a more balanced gender representation.

A notable observation is the bias in occupational terms. Stereotypically female professions (e.g., "nurse") are often translated with feminine pronouns, while technical and high-ranking roles (e.g., "engineer" or "minister") are predominantly assigned masculine pronouns. We will see this bias highlighted more prominently in our third experiment (Section 4.3), but the results here also aligns with findings in prior studies on gender bias in MT for various other languages, where translation systems reinforce occupational stereotypes rather than providing balanced representations.

Additionally, formality plays a role in gender bias. In the *familiar* sentence set (S5), GPT-4o achieves the highest P_S score, with a particularly high p_f indicating common usage of उनी (*oo-ni*) for females in Nepal. Conversely, the honorary pronoun उहाँ (*oo-haan*) overwhelmingly defaults to male translations. This suggests that existing MT systems, including GPT-4o, are more likely to associate higher-status roles with men, reinforcing societal hierarchies in language.

4.2 Evaluation using OTSC-Nepali

The OTSC-Nepali test set (Table 4) provides further insight into gender-specific translation biases. We have presented the percentage of sentences where the speaker's friend is translated as male or female across our eight distinct sentence sets. Across the *familiar* sentence set, all translators perform well except for the case of a female speaker with a male friend using GPT-4o, which shows this pattern in the informal sentence set as well. Notably, GPT-4o tends to translate the friend as female when the speaker is female.

Interestingly, IT2 exhibits the least bias in the familiar sentence set, correctly distinguishing gender roles in most cases. However, in the informal sentence set, both GT and IT2 default to masculine translations, failing to leverage the given gender cues. This pattern suggests that existing MT systems struggle with informal pronoun variations in Nepali, reinforcing masculine defaults. GPT-4o generally performs adequately in the informal set, with the exception of instances involving a female speaker and a male friend.

4.3 Evaluation using WinoMT-Nepali

The WinoMT-Nepali evaluation (Table 5) reveals further complexities in gender bias, particularly in ambiguous or multi-clause sentences. GT and IT2 achieve higher accuracy (Acc) scores in gender labeling, but this comes at the cost of reinforcing stereotypical translations. Conversely, GPT-4o produce a significantly higher proportion of gender-neutral translations (N score), often using "they" or repeating the noun rather than assigning a gender. We also observed that GPT-3.5 displayed similar behavior, generating a large number of neutral sentences, which is why we included it in this experiment.

If we consider gender-neutral translations as correct, GPT-4o's accuracy improves to 71.36% (familiar) and 68.09% (informal). This suggests that LLMs, particularly GPT-4o, are more capable of avoiding gender misclassification but at the expense of erasing gender-specific distinctions. Prior

	GT		IT2		GPT-4o	
Familiar	p_m	p_f	p_m	p_f	p_m	p_f
Female Speaker Female Friend	0.00	**100.00***	0.10	**99.90***	0.00	**100.00***
Female Speaker Male Friend	**78.00***	22.00	**97.53***	2.47	3.42*	**96.13**
Male Speaker Female Friend	0.10	**99.90***	0.10	**99.90***	0.10	**99.90***
Male Speaker Male Friend	**89.70***	10.30	**98.50***	1.50	**89.52***	6.00
Informal	p_m	p_f	p_m	p_f		
Female Speaker Female Friend	**88.40**	11.60*	**99.80**	0.20*	0.10	**99.90***
Female Speaker Male Friend	**97.80***	2.20	**99.80***	0.20	26.63*	**71.79**
Male Speaker Female Friend	**87.42**	12.64*	**99.80**	0.20*	0.32	**99.62***
Male Speaker Male Friend	**98.50***	1.50	**99.80***	0.20	**97.68***	1.72

Table 4: Evaluation using the *OTSC-Nepali* test set. * corresponds to the percentage of sentences translated into the correct label for each set. Bold values show the highest percentage translated into a single gender class. Our desired case is when the same items are both bolded and marked with an asterisk.

Familiar Sentence Set				
	Acc	Δ_G	Δ_S	N
GT	61.18	6.80	18.65	4.11
IT2	**61.48**	17.57	**10.90**	4.51
GPT-4o	48.04*	**0.22**	26.29	**23.35**
GPT-3.5	30.07*	33.92	6.24	39.46
Informal Sentence Set				
	Acc	Δ_G	Δ_S	N
GT	**57.67**	29.08	8.38	3.91
IT2	51.69	47.94	**3.49**	5.05
GPT-4o	49.95*	**22.59**	18.35	**18.14**
GPT-3.5	35.12*	37.991	8.26	23.35

Table 5: Evaluation using the WinoMT-Nepali test set on Acc, Δ_G, Δ_S, N measures. Bold indicates the best value for each metric. $*$ indicates anomaly seen in LLMs' accuracy due to high neutral score.

research (Vanmassenhove et al., 2018; Mirkin et al., 2015; Rabinovich et al., 2017) has shown that neutralizing gender in translations can sometimes reduce bias, but it also removes important linguistic and contextual details, which may not always be desirable.

Notably, IT2 sometimes defaults to "he or she", a strategy that provides more explicit gender representation while mitigating bias. This hybrid approach, offering multiple gendered translations, has also been explored in commercial systems, as we discussed in Section 3.1.2, but has yet to be fully implemented for Nepali.

4.4 Implications and Future Direction

These findings highlight important considerations for improving gender bias in Nepali-English MT systems. While LLMs like GPT-4o show promise in reducing bias, their tendency to neutralize gender can lead to information loss in translations. This raises an important question: should strategies to mitigate bias focus on fairness even if it means less specific context, or should they aim for explicit, dual-gender outputs similar to IndicTrans2 and other proprietary systems?

In addition, the role of formality in gender bias needs more attention, specially in the context of Nepali language. The strong association between honorific pronouns and masculinity suggests that MT systems may be influenced by cultural norms embedded in training data. Future research could explore debiasing strategies that explicitly adjust for formality-based gender skew.

Our study provides a Nepali-specific benchmark for gender bias evaluation, contributing to broader efforts in low-resource language fairness. While LLMs offer improvements over traditional MT systems, their behavior in gender-specific contexts suggests that additional refinements, such as context-aware prompting (Vanmassenhove, 2024) or multi-gender output options, could further enhance translation fairness and accuracy.

5 Bias Statement

This study investigates gender bias in Nepali-English machine translation, specifically how MT systems and LLMs reinforce or mitigate gendered stereotypes. We define bias as the systematic and unfair representation of one gender over another, which manifests in three key ways: (1) reinforcement of gender stereotypes, (2) incorrect gender assignments to neutral or opposite-gendered terms,

and (3) unequal translation accuracy across genders.

Our evaluation focuses on binary gender representation (male and female) due to linguistic constraints of the Nepali language and the scope of available benchmark datasets. While this approach provides a structured analysis, it does not encompass the full spectrum of gender identities. By highlighting these biases, our work aims to contribute to more equitable and inclusive MT systems, particularly for low-resource languages like Nepali, where gender bias and its mitigation has been largely overlooked.

6 Conclusion

In conclusion, we assessed gender bias in Nepali-English machine translation in existing MT systems and LLMs. We developed a Nepali-specific occupation corpus and adapted three challenge sets for a gender-neutral and two gender-specific contexts. Our findings show that traditional MT systems reinforce stereotypes, while LLMs reduce bias but often neutralize gender distinctions. As LLMs continue to evolve, incorporating context-aware prompting and multi-gender translation strategies could help strike a balance between gender neutrality and accurate representation. By refining both MT and LLM strategies, we can develop fairer translation systems for low-resource languages like Nepali.

7 Limitations

Our study is limited to two existing MT systems: one proprietary and one open-source system, which limits the scope of our findings. We could have also experimented with other proprietary systems, such as Amazon Translate and Microsoft Translator, as well as open-source alternatives like NLLB to get a more comprehensive assessment. Similarly, our evaluation of LLMs was restricted to two proprietary models from the same company, which may not fully represent the diversity of capabilities across different LLM architectures. We could have strengthened our analysis by including a broader range of models.

We also acknowledge limitations in our corpus construction. Our occupation corpus was derived from only five categories of the PSC database, which may not fully capture the diversity of occupations in Nepal. Additionally, the WinoMT-Nepali challenge set is a direct translation of the English WinoMT dataset, preventing us from incorporating occupations specific to our corpus, thereby limiting its contextual relevance.

Our study focuses exclusively on translations from Nepali to English. While we could have included English to Nepali translations, doing so would introduce significant ambiguity and limit the scope for bias evaluation. For example, the English sentence "She is a minister" could be translated as उहाँ मन्त्री हुनुहुन्छ (*oo-haan mantri hunuhunchha*), उनी मन्त्री हुन् (*oo-ni mantri hunn*) or ऊ मन्त्री हो (*oo mantri ho*)" in Nepali corresponding to formal, familiar or informal context respectively. Although it would be interesting to analyze which honorific pronoun MT systems prefer, this would not be particularly relevant for evaluating gender bias. Although alternative criteria could have been devised to assess bias in English-to-Nepali translations, this was not the focus of the present study. Nonetheless, this study marks the initial step in evaluating gender bias and other forms of bias in Nepali NLP, with potential for further improvements in the future.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Won Ik Cho, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. 2019. On measuring gender bias in translation of gender-neutral pronouns. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 173–181, Florence, Italy. Association for Computational Linguistics.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Joel Escudé Font and Marta R. Costa-jussà. 2019. Equalizing gender bias in neural machine translation with word embeddings techniques. In *Proceedings of the First Workshop on Gender Bias in Natu-*

ral Language Processing, pages 147–154, Florence, Italy. Association for Computational Linguistics.

Jay Gala, Pranjal A Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, et al. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *arXiv preprint arXiv:2305.16307*.

Sourojit Ghosh and Aylin Caliskan. 2023. Chatgpt perpetuates gender bias in machine translation and ignores non-gendered pronouns: Findings across bengali and five other low-resource languages. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 901–912.

Melvin Johnson. 2020. A scalable approach to reducing gender bias in google translate.

James Kuczmarski. 2018. Reducing gender bias in google translate.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Shachar Mirkin, Scott Nowson, Caroline Brun, and Julien Perez. 2015. Motivating personality-aware machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1102–1108, Lisbon, Portugal. Association for Computational Linguistics.

Prabin Paudel, Supriya Khadka, Rahul Shah, et al. 2024. Optimizing nepali pdf extraction: A comparative study of parser and ocr technologies. *arXiv preprint arXiv:2407.04577*.

Ella Rabinovich, Raj Nath Patel, Shachar Mirkin, Lucia Specia, and Shuly Wintner. 2017. Personalized machine translation: Preserving original author traits. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1074–1084, Valencia, Spain. Association for Computational Linguistics.

Krithika Ramesh, Gauri Gupta, and Sanjay Singh. 2021. Evaluating gender bias in Hindi-English machine translation. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 16–23, Online. Association for Computational Linguistics.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.

Tej Bahadur Shahi and Chiranjibi Sitaula. 2022. Natural language processing for nepali text: a review. *Artificial Intelligence Review*, 55(4):3401–3429.

Pushpdeep Singh. 2023. Gender inflected or bias inflicted: On using grammatical gender cues for bias evaluation in machine translation. In *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.

Dagmar Stahlberg, Friederike Braun, Lisa Irmen, and Sabine Sczesny. 2011. Representation of the sexes in language. In *Social communication*, pages 163–187. Psychology Press.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Eva Vanmassenhove. 2024. Gender bias in machine translation and the era of large language models. In *Gendered Technology in Translation and Interpreting*, pages 225–252. Routledge.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium. Association for Computational Linguistics.

Eva Vanmassenhove and Johanna Monti. 2021. gENder-IT: An annotated English-Italian parallel challenge set for cross-linguistic natural gender phenomena. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 1–7, Online. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. 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, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.