ISSN:2790-167X Volume-13-(2025)

Combating Gender Biases from Source Language in Machine Translation

Jiaxi Xue [1, a]

[1]Guangdong University of Foreign Studies, Guangzhou, China.

[a] xuejiaxi@gdufs.edu.cn

Abstract. Effective Chinese language translations that communicate clear and unbiased messages form the core fabric of peaceful relationships among people in society. Existing machine translations models like Google Translate, DeepL, Youdao Translate or even LLM sometimes generate Chinese-to-English translations filled with prejudice against females. The prejudice-prone Chinese-to-English translations are inherited from some terms in the source texts which are merely norms inherited from Chinese history and may not indicate female discrimination in terms of context. Such a gap must be eradicated to minimize gender biases and gender disparity in Chinese-to-English translation. Addressing the aforementioned gap, this study employs deconstructive translation theory to demonstrate the translational phenomenon of the remaining unnecessary female discrimination expression from the source Chinese to the target English language through the lens of feminism. The study also proposes an improved model of machine translation to hi-jack the gender-biased expression by identifying and modifying the gender-biased expressions in Chinese-to-English texts from a context perspective. The findings and application of the research will improve the well-being of females by preventing the cultural flow of gender biases, and contribute to creating a more amicable language environment for females.

Keywords: gender translation; machine translation; translation studies.

1. Introduction

Translation from Chinese to English plays an important role in the efforts of officials to spread cultural impact, and is progressively emerging as a considerable proportion of global cultural communication [1]. In the process of enhancing Chinese cultural influences in global dialogues, it is sometimes suggested that machine translations are more efficient than human ones since great progresses been achieved in machine translation [2]. While human translators are deeply influenced by cultural contexts and nuanced understandings of gender, Machine Translation systems often lack such awareness [3]. The Machine translation results, however, sometimes contain prejudice against females, which is inherited from some terms in the Chinese source texts, despite the fact that the words with biases are merely remains from history and may not indicate female discrimination in a context angle. If the contexts are not examined, and the translations are not modified, thus remain conventionally loyal to the source text by transferring the literal meaning, phrases with gender biases toward women may send wrong message of discrimination and biases that differ from expected meanings in real life via the translations. However, the so-called loyalty in the condition of translating phrases with biases does not promote communication and is hence unneeded. Truly functional loyalty sometimes requires the violation of certain conventions [4]. Unnecessary loyalty can be arbitrary in machine translations, as there are no translators censoring every word-choice. As a result, the machine translation works of the words with gender biases will maintain the gender disparity from the source texts while flowing into another culture. More related dialogues or statements are formed accordingly, and the meaning of female discriminatory expressions is dynamically established in the différance, a particular connotation by Derrida [5], process across multiple cultures, leading to an unpleasant language environment for females. Not all the machine translation model designers have realized the existence of gender biases in the source language and the necessity to stop the flow of the biases among cultures. Some machine translation models, Google Translate and DeepL for instance, will treat words which contain gender biases implication the same way as the words without them, leading to a preservation of gender biases in

the target text, and providing an involuntary path to the spreading of gender biases. Addressing the aforementioned issue, this study employs deconstructive translation theory to demonstrate the translational phenomenon of the remaining unnecessary female discrimination expression from the source language to the target language through the lens of feminism, and attempts to construct a model of machine translation to hi-jack the gender-biased expression. Eventually, the research will benefit the well-beings of females by preventing the cultural flow of gender biases, and contribute to create a more amicable language environment for females.

The process of différance worked infinitely through cultural communication when a massive net of cultures will be created when using translation to connect one culture and another. Each culture in the net can contain the reflections from the others in this process, as deconstructive translation theory illustrates. The meaning lies not in words of either the target language or source language, but in the infinite intertextuality of différance, underscoring the inherent instability of language and the constant play between signifiers and their associated meanings [6]. “The value of meaning originates in translation, while one inevitable price of translation is always the change of meanings, when freed, interiorized, spiritualized, and elevated above from the body” [7]. The advent of translation, particularly machine translation, has introduced a mechanism by which biases can transcend language boundaries with the potential to affect other cultures, not merely by transferring words but by shaping perceptions and social norms. Given that machine translation systems can transmit these biases on a large scale, their potential impact on global communication must be addressed to prevent the reinforcement of harmful gender stereotypes [8] In some ways, translation acts as a gateway to cultural intertextuality, serving as the foundation for the creation of new meanings.

Rebellions against the source texts will naturally appear if some certain terms do harm to females, and if translators want to interfere this unfriendliness. One issue is whether or not translators are granted the right to modify the literal meaning of original texts to forward their own agendas. There has been a heated debate over the visibility of translators since Venuti [9] emphasized the rightful existence of translators and challenged the unequal power in translation. The creator of the machine translation model, which is the agency, can nevertheless be seen as a translator in this specific instance even though Venuti did not take machine translation into consideration. The theory of deconstructivism also aware the translators of their power, right, and responsibility to create new meanings with their own works [10]. Translators or developers of machine translation models, are not only guaranteed an independent position from the original writer, but also empowered to change the meaning in target languages to serve specific purposes by demonstrating and expressing themselves, which appears as modifying discrimination to female-related words in this study.

Feminism translation [11] thinks that translation is more than just a conversion between two languages; it is also a language-converting activity with political implications. Translation raises the possibility that certain gender equality can be achieved through translation. Considering the fixation of the source text, the translator has to make adjustments to add feminism awareness into the translation, and conduct a creative rebellion of feminism. Though feminism translation has been discussed in a variety of subjects since the day it emerged, most of the relevant researches take their roots into the translation of literature works and eases their focus on other realms of translation studies. Gendered implications of technology development and use emerged as notably underexplored [10]. Existing researches about gender biases in machine translation mainly focus on ensuring the translating of a gender-neutral word remains gender-neutral in the target language, instead of hi-jacking a gender-biased word and convert it into a neutral one.

The potential of the biases in publicly available MT and GPT systems to impact social norms cannot be disregarded considering their massive use [10]. After selecting 10 source language words or phrases with discriminations and biases towards females, this research tested three of the most widely-used machine translation models, namely, Google Translate, DeepL, and Youdao Translate. This study discovered that all three models failed to solve the above problems of female discrimination by running more than ten thousand samples. Thus the research used the

fine-tuning technology to train an LLM model for machine translation which can hi-jack the gender biases towards females and weaken the disharmony elements of female discrimination, making the translation more neutral and friendly to female community. The expressions with gender biases are revised it into gender-neutral one, and deliver translation results with fewer hostile message against female when flowing into another culture. To some extent, this model shed light on solving the problem of unnecessary fluid of female discrimination into intercultural communication and intertextuality, and rebalances the cultural ecology of gender equality.

2. Literature Review

This paper addresses the intersection of two prominent scholarly areas: (i) the relationship between language, translation, and sociology, and (ii) the examination of gender biases within translation studies. A growing amount of research has highlighted gender-related issues which emerge during the development of translation studies itself as well as of the intersections between translation studies and other disciplines. These existing studies often focus on identifying the sources of gender bias in language and investigating their impact on global cultural systems. In addition, research within the fields of Natural Language Processing (NLP) and Large Language Models (LLMs) has sought to assess and mitigate gender biases by fine-tuning computational models to achieve a more neutral language. However, relatively few studies have explored the identification of gender biases inherent in the source language and how to address these biases during translation to prevent their propagation across cultures using NLP and LLMs.

2.1 Gender Studies and Translation Studies: Foundations and Development

The integration of gender studies into translation research has gained prominence since the 1990s, when feminist and gender theories began influencing the discipline [12]. Feminist translators' active participation and efforts to produce changes in translation studies have been widely discussed ever since [13]. Irshad and Yasmin [14] mentioned that scholars like Sherry Simon [15] have underscored the notion that translation is never a neutral or innocent act, and posits the right of female translators to intervene the texts. Simon's work is foundational in recognizing the agency of female translators, urging them to intervene and subvert traditional gendered assumptions in their translations. Building on Simon's work, a nuanced, multidimensional approach was also advocated to translation that accounts for the socio-cultural context of both the source and target languages, allowing for a deeper understanding of how gendered practices in translation mirror societal power structures and contribute to the construction of gendered identities.

Meng [16] further advanced this dialogue by distinguishing between feminist translation and gender translation within a postconstructivist framework and emphasized that translation is not merely a linguistic transfer but an ideological negotiation. This r esearch of Meng [16] highlights the fluidity of both meaning and gender, suggesting that translation is a space where gender norms can be either perpetuated or subverted. By integrating post-structuralist gender theory and critical discourse analysis, Meng [16] opened a critical lens for examining how gender biases are reproduced or resisted in translated texts. Nabinita [17] also contributed to feminist translation studies by articulating the task of feminist translators in the twenty-first century with the arguement that feminist translators must work to recover marginalized voices and reposition them within both national and global literary and cultural histories. This approach aligns with broader feminist goals of challenging patriarchal norms and re-centering women's agency in cultural production [18], suggesting that translators should be attuned to cultural exclusions in the source text and actively seek ways to rectify them in the translation process.

2.2 The Flow and Amplification of Gender Biases Through Translation Technologies

Technological advancements have significantly enhanced the efficiency of translation, with models such as Google Translate, DeepL, and increasingly sophisticated LLMs becoming

ubiquitous. Machine translation techniques can ensure instantaneous communication and promote mutual understanding between cultures [19]. Despite the convenience that machine-aided translations bring, these tools often reproduce gender biases due to limitations in cultural and contextual awareness in machine translation [20], or even exacerbate the gender asymmetries. Moreover, the datasets used to train these models can perpetuate stereotypes, leading to biased translations that reinforce problematic gender norms. As machine translation tools become more widespread, the technology risks amplifying gender biases in ways that human translators might avoid. A powerful tool in analyzing and addressing gender bias in translation is corpus linguistics, which allows for large-scale, data-driven exploration of gendered language use across different corpora. Angouri and Baxter [21] examined how intertextuality and the movement of words across cultures can influence the gendered meanings that emerge in translation, and stressed that the inherent cultural flux in translation can either perpetuate or neutralize gender biases, depending on how translators engage with these shifting meanings. Derrida's [5] concept of différance helps to explain that machine translation, which is more efficient but with biases, can deteriorate the situation of gender biases globally because of the dynamic generation process of meaning. Différance plays a crucial role in understanding the dynamics of translation, particularly in how gendered meanings are negotiated across languages. Vasanathan [6] applied the concept of différance to translation studies, suggesting that female specific expressions may become more gendered in translation, intensifying stereotypes and reinforcing social expectations. This highlights how translation is not simply a transfer of meaning but an act that can either challenge or consolidate gendered norms. Zhu and Chen [22] further examined the role of translation in gender representation, categorizing women and gender minorities as "vulnerable others" who are often marginalized or misrepresented in translated texts. Zhu and Chen [22] argued, however, that women are not merely passive recipients of misrepresentation but active participants in challenging gender biases within translation practices. The need for translators was emphasized to be aware of cultural differences and to adopt strategies that neutralize gendered bias in the translation process, ultimately fostering a more equitable representation in the target language. Such neutralization can be achieved through novel context-inclined smart translation technologies but rarely researched.

2.3 Gaps in Existing Literature

The literature on gender bias in translation underscores the importance of critically examining the complex intersections of language, culture, and ideology. Much of the existing research has focused on the consequences of gender bias in translation, but fewer studies have investigated how to remove these biases during the translation process itself. This gap presents an opportunity to explore how gender bias is embedded in the source language and how translators can mitigate its impact during the translation process, promoting more equitable cultural exchanges.

3. Materials and Methods

3.1 Gender Translation and Machine Translation

To gather relevant translation-related data for the study, a list of gender-biased Chinese words and phrases that are unfavorable to females were extracted from sources including movies, news articles, social media posts and online comments, etc. A list of the sampled gender-biased Chinese words and phrases are reported in Table 1. To demonstrate that the sampled Chinese words and phrases are gender-biased, existing welladopted translation technologies including Google Translate, DeepL and Youdao Translate were applied to translate the targeted words into English. As depicted in Table 1, the Chinese word "妈咪包" translates into "Mummy Bag" by Google Translate, "Mommy Bag" through DeepL and Youdao; all of which contain gender biases unfavorable to females through the use of qualifiers such as "mommy" and "mummy", indicating the nursing job is only for females. Similarly, the translation of "心机婊" using existing technologies is highly

negative towards women with Google Translate, DeepL and Youdao Translate using words such as “bitch” and “woman”.

Table 1. Targeted Chinese Words and English Translation

Gender-biased word	Google Translate	DeepL	Youdao Translate	Gold Standard
波霸奶茶	boba milk tea	boba milk tea	boba milk tea	Pearl milk tea
三八	three-eight	three eights	notala	nosy
八婆	gossip	meddling woman	witch	nosy
妈咪包	mummy bag	mommy Bag	mommy bag	nursing bag
母婴室	baby room	mother's room	mother's Room	nursing room
河东狮	hedong lion	hedong lion	hodonglion	lion
白莲婊	white lotus bitch	white lotus bitch	white lotus bitch	white lotus
绿茶婊	green tea bitch	green tea bitch	green tea bitch	green tea
圣母	saint	holy mother	the Virgin Mary	saint
婆婆妈妈	mother-in-law	mother-in-law	motherly	fussy
狐狸精	vixen	vixen	vixen	seductive
事儿妈	nosy mother	matter-of-fact mom	drama mother	trouble maker
公主病	princess syndrome	princessism	princess disease	Royal syndrome
母老虎	tigress	tigress	tigress	tiger
玛丽苏	Mary Sue	Mary Sue	Mary Sue	overly romantic
妇道人家	womanly family	woman's family	woman's family	a person of the household
妇人之见	woman's view	woman's opinion	woman's view	narrow perspective
狐魅子	foxy	fox-charmer	seductress	seductive
事业线	Career Line	career line	business line	body curve
扶弟魔	brother-supportin g demon	Help me,Devil	Fudi	brother-provider
头发长见识短	long hair but short knowledge	Hair is long and knowledge is short	long hair, short vision	short knowledge
女人心海底针	woman's heart is as deep as the sea	A woman's heart is like a needle under the sea	A woman's heart is a needle	a heart is as deep as the sea
女汉子	female man	femme fatale	tough girl	tough woman
名媛	socialite	young lady of note	debutante	socialite
剩女	leftover woman	leftover woman	leftover woman	single woman
大妈	auntie	auntie	aunt	people
妈妈臀	mother's buttocks	mom's ass	mommy buttock	saggy buttocks
老妈子	old nanny	older women	maidservant	servant
黄脸婆	yellow-faced woman	faded old woman	The woman with the yellow face	old woman
心机婊	scheming bitch	scheming bitch	scheming bitch	scheming person
母鸡司晨	hen crowing	hen in the morning	hen shichen	woman in charge
长舌妇	long-tongued woman	long-tongued woman	yenta	gossip person
水性杨花	fickle	fickle(woman)	a man of easy means	seductive

妇孺皆知	everyone knows	well known to women and children	everyone knows it	everyone know
女博士	female doctor	female doctor	A woman with a doctorate	doctor
女司机	female driver	female driver	woman driver	driver
马子	Ma Zi	brigand	prostitute	girlfriend
破鞋	slut	broken shoes	prostitute	commit adultry
妖艳货	sexy	bitch	sexy goods	seductive person
大雷	big thunder	big thunder	large thunder	breast
得吃	have to eat	gotta eat	have to eat	have sex

Table 1: Humans are dynamic and the meanings people accrue to the gender-biased words or phrases reported in Table 1 could differ by location, person in question and context. To ensure that the dataset accurately reflects real-world language use and the complexities of gender in translation, Chinese sentences were gathered from diverse sources such as contemporary Chinese literature, media, and social interactions. The dataset includes both gender-neutral and gender-specific sentences to highlight how gender can influence translation output. For each Chinese sentence, three different translations were collected from widely used machine translation models: Google Translate, DeepL, and Youdao. Additionally, a "gold standard" translation, as verified by linguists specializing in both Chinese and English, was also included in the dataset. The gold standard translations were chosen to represent nonbiased and contextually accurate English translations that minimize gender bias while preserving the meaning of the original Chinese. The total dataset comprises 3000 sentences, where each sentence contains the following columns:

- Gender-biased words in Chinese; i.e., original source text in Chinese
- Google Translate; i.e., translation produced by Google Translate
- DeepL; i.e., translation produced by DeepL
- Youdao Translate; i.e., translation produced by Youdao, and
- Gold Standard; i.e., linguist-reviewed translation that is free of gender bias.

The gathered dataset consists of 3000 sentences collected over a period of three months, ensuring a comprehensive representation of various contexts, terminologies, and sentence structures. The dataset was constructed to encompass a diverse range of gender-sensitive and neutral expressions found in real-world Chinese-to-English translation tasks. Sources for the dataset include: A. Linguistic Annotations by Experts: where sentences were gathered from professional linguists who provided gender-neutral gold standard translations. B. Crowdsourced Contributions: where native Chinese speakers provided sentences reflecting diverse societal contexts, including historical, colloquial, and professional domains. C. Existing Translation Corpora: here, publicly available bilingual datasets were filtered to include gender-sensitive terms.

The collected sentences were further categorized into three tiers: (a) Explicit Gender Bias: Sentences with obvious gendered expressions. (b) Implicit Gender Bias: Subtle gender connotations or role expectations. (c) Neutral Expressions: Sentences that were inherently gender-neutral.

3.2 Data Cleaning and Preprocessing

To ensure data quality, a data cleaning and preprocessing pipeline that cover the following was applied:

- Text Normalization: Ensuring consistency in formats, such as date styles, punctuations, and spacing.
- Bias Tagging: Annotators tagged sentences with potential gender biases and their associated linguistic contexts.
- Ambiguity Resolution: A panel of three linguists reviewed ambiguity.
- Error Removal: Sentences with typos, grammatical or contextual inaccuracies were discarded.

• Data augmentation techniques were applied, such as paraphrasing and context-switching, to increase the dataset's robustness and diversity.

3.3 Proposed LLM Modeling and Fine-Tuning

The Qwen2.5-0.5B LLM proposed by Alibaba was adopted and fine-tuned for the gender-biased translation research. With Alibaba recognized as among the leading technology companies in China that operates in a nation where Chinese is the native language, it can be argued that the Qwen2.5-0.5B LLM, when fine-tuned, has the propensity to produce improved Chinese-to-English translation results. The Qwen2.5-0.5B model was fine-tuned using the cleaned dataset. The training process utilized training objective, gender-bias mitigation, and optimization of relevant parameters to produce gender biased-free Chinese-to-English translation. To model the training objective, the aim was to minimize the loss function, L, as a function of the input Chinese sentence xi, yi the target English translation (gold standard), and model parameters θ. Next is bias mitigation strategy, where a specialized loss adjustment was applied to penalize gender-biased outputs. This was achieved by introducing a bias weight, β, computed as β = total words in translation count times biased words. The corresponding final loss function L' was modified to $L' = L + \lambda\beta$; where λ is a hyperparameter controlling the penalty's intensity. The cleaned sentence and context-inclined dataset were split into model, training and validation subsets. The finetuned model and training sets were used to train the refined Qwen2.5-0.5B Chinese-to-English translation LLM. Rigorous fine-tuning approaches were applied to derive an optimal fine-tuned Qwen2.5-0.5B Chinese-to-English translation LLM capable of translating Chinese to English with zero gender biases. The key parameters derived from the optimal fine-tuned Qwen2.5-0.5B Chinese-to-English translation LLM include: learning rate: 1×104; batch size: 4; gradient accumulation steps: 4; epochs: 5; and warmup steps: 500. It must be noted that fine-tuning was conducted on NVIDIA GPUs with mixed precision fp16.

4. Result and Discussion

4.1 Results of Fine-tuning

The empirical results from applying the optimal fine-tuned Qwen2.5-0.5B Chinese-to-English translation LLM is depicted Qwen2.5-0.5B LLM translates gender-biased Chinese sentences and contexts into gender biased-free ones, eliminating the biases hitherto unresolved. The improved model is applied to translate 5 gender-biased statements and the results are compared with Google, DeepL and Youdao. As depicted on Table2, the proposed fine-tuned Qwen2.5-0.5B LLM translates the gender-biased Chinese sentence "..." into "..." which is free from the biases provided by Google, DeepL and Youdaoin Table 2. The results show that the proposed fine-tuned Qwen2.5-0.5B LLM translates the gender-biased Chinese sentence "..." into "..." which is free from the biases provided by Google, DeepL and Youdao.

Table 2. Comparison of Chinese-to-English translation models

Gender-biased text	Gold Standard	LLM	Google Translate	DeepL	Youdao	Gender-Neutral-Qwen
在性别敏感培训中，通过角色扮演等方式，让参与者体验"妖艳货"等词汇带来的	In gender sensitivity training, through role-playing and other methods, participants are allowed to	In gender sensitivity training, through role-playing and other methods, participants are allowed to	In gender sensitivity training, let participants experience the harm caused by words such as "flashy goods"	In gender sensitization training, participants should be allowed to experience the harm caused by	In gender sensitivity training, participants were allowed to experience the harm caused by words such as	In gender sensitivity training, participants are allowed to experience the harm caused by

伤害。	experience the harm caused by terms such as "seductive."	experience the harm caused by terms such as "seductive trash."	through role-playing and other methods.	terms such as “slut” through role-playing.	"sexy goods" through role playing and other ways.	terms like 'seductive' through role-playing and other methods.
在某些方言中，“破鞋”被用来形容不忠的人，但这并不意味着它是可接受的。	In some dialects, "commits adultery" is used to describe unfaithful people, but that does not mean it is acceptable.	In some dialects, "slut" is used to describe unfaithful people, but that does not mean it is acceptable.	In some dialects, "broken shoes" is used to describe unfaithful people, but this does not mean that it is acceptable.	In some dialects, “broken shoes” is used to describe unfaithful people, but that does not mean it is acceptable.	In some dialects, "broken shoes" is used to describe someone who is unfaithful, but that doesn't mean it's acceptable.	In some dialects, "破鞋" is used to describe someone who is not faithful, but this does not mean it is acceptable.
那位女司机在乡村道路上行驶，总是礼让行人，赢得村民的好评。	That driver always yields to pedestrians on rural roads, earning praise from villagers.	That female driver always yields to pedestrians on rural roads, earning praise from villagers.	The female driver always gives way to pedestrians when driving on rural roads, winning praise from villagers.	The woman driver always yields to pedestrians when driving on rural roads, winning the villagers' praise.	The woman driver, driving on the country road, always concedes to the pedestrians and wins the praise of the villagers.	The woman driver on the country road always gives way to pedestrians, and she is well-received by the villagers.
朋友们总是劝她：“别太在意‘女博士’这个标签，缘分总会来的。”	Friends always advise her, "Don't worry too much about the label of PhD; fate will come eventually."	Friends always advise her, "Don't worry too much about the label of 'female PhD'; fate will come eventually."	Friends always advise her: "Don't pay too much attention to the label of 'female doctor', fate will always come."	Her friends always advise her, “Don't be too concerned about the label ‘female doctor’, fate will always come.”	Friends always advised her: "Don't pay too much attention to the label 'female doctor', fate will always come."	Friends often advise her to not be too concerned about the label "PhD," as the future
她虽然是妇道人家，但厨艺却十分了得，让人赞不绝口。	Although she's a person of the household, her cooking skills are excellent, earning praise from everyone.	Even though she's just a woman of the household, her cooking skills are extraordinary and deserve praise.	Although she is a woman, her cooking skills are very good and people praise her.	Although she is a womanizer, she is a very good cook, and people are very impressed with her cooking.	Although she is a woman, but the cooking is very good, let people full of praise.	She is from a respectable family, but her cooking is quite impressive, and people are always praising her.

As stated earlier the keywords in Table 1 were tracked in 3000 sentences. The fine-tuned Qwen2.5-0.5B LLM has been applied to translate random realistic gender-biased sentences that capture at least one of the targeted biased Chinese words/phrases. The corresponding gender-biased minimizing results from the fine-tuned Qwen2.5-0.5B LLM were evaluated using qualitative evaluation techniques. The qualitative evaluation included manual inspection where a team of

bilingual linguists reviewed the random realistic samples provided by the fine-tuned Qwen2.5-0.5B LLM for accuracy and gender neutrality in translation. Three checklists that the bilingual linguists applied in checking for accuracy and gender neutrality in translations are (i) whether any gender-biased terms, such as using “he” when the original Chinese sentence was gender-neutral, were present; (ii) the fluency and naturalness of the gender-biased minimizing English translation; and (iii) alignment of the gender-biased minimizing English translation with the gold standard in terms of both gender and contextual meaning.

The reliability of the fine-tuned Qwen2.5-0.5B LLM Chinese-to-English translation results were also tested for the ability to withstand three forms of gender bias categorization, and the proposed model passed them all. The first is no gender bias; here the model was applied to situations where the translation is neutral, using no unnecessary gendered terms. Second is subtle gender bias to capture minor instances of gender bias that are contextually understandable but could be neutralized. The third category is clear gender bias, a category that describes instances where the translation is overtly gender-biased. Unlike the existing Google, DeepL and Youdao translations models that fail with at least one of the three categories, the empirical evidence suggests that the fine-tuned Qwen2.5-0.5B LLM Chinese-to-English translation was successful in withstanding the unreliability that are ascribed to the three classes of bias. Thus, the results suggest that the fine-tuned Qwen model consistently generated translations closely aligned with the gold standard, maintaining semantic and syntactic fidelity. Added, outputs from the fine-tuned Qwen2.5-0.5B LLM Chinese-to-English translation showed a significant reduction in explicit and implicit gender biases compared to the Google, DeepL and Youdao baseline translation models.

For instance, the fine-tuned Qwen2.5-0.5B LLM translates a sentence such as “感觉自己是这条街最靓的仔” as “I feel like the coolest person on this block” - a gender-free translation, rather than “I feel like the most handsome guy on this street” derived from the baseline models. The refined Qwen2.5-0.5B LLM is also able to translate and maintain context. The fine-tuned Qwen2.5-0.5B model demonstrated an improved ability to understand the context and preserve the original meaning while avoiding gender stereotypes.

4.2 Conclusion and Contribution

The growing awareness of biases in language models calls for a proactive approach to ensure these models serve diverse and equitable communities. This research proposes a practical approach to mitigating gender biases in machine translation by fine-tuning models with accurate and gender-neutral translations. By providing alternative, bias-free translations from Chinese to English of commonly gendered phrases, this study aims to reduce the harmful impact of automated translation systems. Furthermore, it calls for a broader cultural shift in which developers, linguists, and translators work collaboratively to identify, address, and prevent all forms of bias in language models. This research aims to address a specific challenge in the translation of gendered language from Chinese to English by fine-tuning machine translation (MT) models with accurate, gender-neutral translation pairs. While this initiative is a step in the right direction, it is only a beginning. It is essential for researchers and developers to recognize that biases are pervasive across languages and cultures, and it is critical to develop the sensitivity needed to identify and mitigate these biases in translation processes.

5. Limitation and Future Research

Despite the significant improvement in the female gender-biased minimizing Chinese-to-English translation recorded by applying the refined Qwen2.5-0.5B LLM, a few limitations are observed and future research can expound on. The limitations are classified into domain-specific bias, ambiguity in linguistic contexts, and dependence on annotated data. On domain-specific bias, the dataset utilized in the study may not comprehensively cover all domains or genres of language use. Domains such as technical, medical or highly specialized fields, may manifest gender biases

differently. Consequently, the performance of the proposed Qwen2.5-0.5B LLM might be less effective in Chinese-English translating texts from such domains. Future research can expand the dataset to include more diverse and domain-specific texts. Next, in cases where gender-specific terms are necessary (e.g., “mother” vs. “father”), the model may neutralize the translation, leading to potential inaccuracies or loss of detail and ambiguity linlinguistic contexts. Here, research can improve the refined Qwen2.5-0.5B LLM model by incorporating dynamic retraining mechanisms to adapt to linguistic and societal changes. Third, the quality of the fine-tuned Qwen2.5-0.5B LLM model heavily relies on the quality of the dataset. Any bias, inconsistency, or error in the gold standard annotations may inadvertently propagate into the Chinese-English translations. The human evaluators, despite their expertise, may introduce subjective biases in assessing the gender neutrality and overall quality of translations, affecting the robustness of qualitative findings. Future studies could develop and leverage more comprehensive evaluation metrics to assess the cultural and contextual quality of translations, and artificial intelligent agents to minimize subjective biases.

The transmission of gender biases occurs not only when translating between Chinese and English, as this study explores, but also in every instance of translation across languages and cultures. Though this research investigated merely on Chinese-English Language pair, such phenomenon is not isolated to a single language or culture but are pervasive across languages worldwide. Gender bias can be embedded in everyday language [23]. Similar instances of gendered language can be found in numerous other cultures and languages, highlighting the universal nature of this problem. As such, it is crucial to extends studies of gender biases in machine translation globally across different language pairs in future studies.

Moreover, gender bias is not the only form of discrimination embedded in language. Linguistic expressions can also reflect biases related to race, age, class, and other social categories. These forms of bias can have significant consequences for individuals’ mental health and societal well-being, as they perpetuate harmful stereotypes and social inequalities [24]. Therefore, it is incumbent upon both human translators and the developers of machine translation models to take future responsibility for identifying and mitigating these biases, not just in gendered terms, but across all forms of discrimination perpetuated through translation models.

Combating gender biases in translation is not only the responsibility of machine translation developers but also of researchers, educators, policymakers, and users who seek to promote a more inclusive linguistic environment. This collaborative effort should include the training of language models, the fine-tuning and ongoing adjustment of these models, and the critical evaluation and refinement of the translations the models generate. By engaging in this process, the transmission of harmful gender biases can be reduced through automated translation, and foster a more respectful and inclusive global communication landscape.

References

[1] T. Dolci, “Fine-tuning language models to mitigate gender bias in sentence encoders,” 2022 IEEE Eighth International Conference on Big Data Computing Service and Applications (BigDataService), pp. 175–176, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:252579643

[2] H. Wang, H. Wu, Z. He, L. Huang, and K. W. Church, “Progress in machine translation,” Engineering, vol. 18, pp. 143–153, 2022.

[3] A. Das, “Neural machine translation (nmt): Inherent inadequacy, misrepresentation, and cultural bias,” International Journal of Translation, vol. 32, pp. 115–145, 2020.

[4] C. Nord, “Scopos, loyalty, and translational conventions,” Target. International Journal of Translation Studies, vol. 3, no. 1, pp. 91–109, 1991.

[5] J. Derrida et al., différance. Carleton University Press Inc., 1982.

[6] R. Vasanathan, “Unveiling conformity and differentiation through derrida’s ‘différance’”, ShodhGyan-NU: Journal of Literature and Culture Studies, vol. 1, no. 1, pp. 6–11, 2023.

[7] J. Derrida and L. Venuti, “What is a” relevant” translation?” Critical Inquiry, vol. 27, no. 2, pp. 174–200, 2001.

[8] N. Suryandari, “Role of stereotyping in intercultural communication,” Journal of Humanities and Social Science, vol. 25, no. 1, pp. 24–30, 2020.

[9] L. Venuti, “The translator’s invisibility,” Criticism, vol. 28, no. 2, pp. 179–212, 1986.

[10] E. Monzó-Nebot and V. Tasa-Fuster, Gendered technology in translation and interpreting: Centering rights in the development of language technology. Taylor & Francis, 2024.

[11] L. Von Flotow, “Feminist translation: contexts, practices and theories,” TTR: traduction, terminologie, rédaction, vol. 4, no. 2, pp. 69–84, 1991.

[12] L. Von Flotow and J. W. Scott, “Connecting the transdisciplines: Translation studies and gender studies,” Border Crossings: Translation Studies and Other Disciplines, pp. 349–374, 2016.

[13] Z. Yu, Translating feminism in China. Taylor & Francis, 2015.

[14] I. Irshad and M. Yasmin, “Feminism and literary translation: A systematic review,” Heliyon, vol. 8, no. 3, 2022.

[15] Simon and L. Von Flotow, “Gender in translation: cultural identity & the politics of transmission,” University of Toronto Quarterly, vol. 67, no. 1, p. 149, 1997.

[16] L. Meng, “From feminism translation to gender translation,” Chinese Translators Journal, vol. 37, no. 5, pp. 23–31, 2016.

[17] S. Nabanita, “Translation and gender,” Litinfinite, vol. 4, no. 2, 2022.

[18] J. Munday, S. R. Pinto, and J. Blakesley, Introducing translation studies: Theories and applications. Routledge, 2022.

[19] S. Kamalesh and S. Jegadeesan, “Smart optimization of machine translation on intercultural communication,” growth, vol. 6, p. 7, 2023.

[20] B. Savoldi, M. Gaido, L. Bentivogli, M. Negri, and M. Turchi, “Gender bias in machine translation,” Transactions of the Association for Computational Linguistics, vol. 9, pp. 845–874, 2021.

[21] J. Angouri and J. Baxter, The routledge handbook of language, gender, and sexuality. Routledge, 2021.

[22] C. Zhu and Z. Chen, “The other, hybridity and translation: A study on the translation of English/Chinese hybrid puns in the postmodern context,” Journal of University of Shanghai for Science and Technology, vol. 46, no. 2, 2024.

[23] H. J. MacArthur, J. L. Cundiff, and M. R. Mehl, “Estimating the prevalence of gender-biased language in undergraduates’ everyday speech,” Sex Roles, vol. 82, no. 1, pp. 81–93, 2020.

[24] M. R. Banaji, R. Bhaskar, and M. Brownstein, “When bias is implicit, how might we think about repairing harm?” Current Opinion in Psychology, vol. 6, pp. 183–188, 2015.