Towards Equitable Translation: Gender Bias in Large Language Models

Hong Hai Ngo[1] Yunmeng Li[1] Keisuke Sakaguchi[1,2]
[1]Tohoku University [2]RIKEN
ngo.hong.hai.s5@dc.tohoku.ac.jp

Abstract

Machine translation (MT) systems often struggle with handling gender distinctions in languages with grammatical gender. In this paper, we evaluate the performance of 10 large language models (LLMs) in translating sentences from English, using a dataset of sentences structured as *"I am [demonym]."* into masculine/feminine/neuter forms in German, French, Italian, and Spanish. Our results indicate that while most models demonstrated the ability to generate gender-specific translations, they tend to produce masculine forms more frequently than feminine. For entries with non-existing official demonyms models either apply linguistic rules to generate non-standard forms or rely on alternative constructions.

1 Introduction

Machine translation (MT) bridges language barriers, making communication and understanding easier and faster. Thus, these systems have improved significantly over the last decade. However, gender bias remains a significant challenge for MT tools. Gender bias in a text is the use of words or syntactic constructs that connote or imply an inclination or prejudice against one gender [1]. This issue occurs due to contrastive linguistic settings that necessitate disambiguation and explicitness in their representation of gender [2].

This bias is especially pronounced when translating from gender-neutral languages (e.g., English) into those with grammatical gender (e.g., Spanish), where every noun is assigned to a specific category, such as masculine, feminine, or neuter. This categorization affects how related words such as adjectives, pronouns, and verbs agree with the noun in gender and number. The number of grammatical gender classes ranges from two to several tens [3]. In this study, we focus on the previously mentioned three categories, as they represent the most common grammatical gender classifications in Indo-European languages, which comprise the majority of languages with grammatical gender and benefit from relatively strong digital support.

eng:	I am Japanese.	I am Vietnamese.
deu:	Ich bin Japaner. (m) Ich bin Japanerin. (f)	Ich bin Vietnamese. (m) Ich bin Vietnamesin. (f)
fra:	Je suis Japonais. (m) Je suis Japonaise. (f)	Je suis Vietnamien. (m) Je suis Vietnamienne. (f)
it:	Sono Giapponese. (m) Sono Giapponesi. (f)	Sono vietnamita. (n)
spa:	Soy japonés. (m) Soy japonesa. (f)	Soy vietnamita. (n)

Figure 1 Selected examples from a dataset structured as "*I am [demonym].*" with input in **eng** (English) and translations into target languages: **deu** (German), **fra** (French), **it** (Italian), and **spa** (Spanish). Colored text indicate gender-specific forms: masculine (**m**), feminine (**f**), and unmarked for neuter (**n**).

In cases where either gender may be a correct translation, MT systems tend to provide only one option, often due to stereotypical associations [4]. Alternatively, this behavior can arise from grammatical conventions, where defaulting to a specific gender is standard practice when the gender is unknown. However, this approach is unsuitable for first-person sentences, such as *"I am Ukrainian."* where the speaker's gender is inherently known and should be reflected in the translation.

With this in mind, we create a dataset[1)] with the sentence structure *"I am [demonym]."* (Figure 1) designed to evaluate how decoder-only models handle English-to-X translation, where X is a language with grammatical gender. Despite large language models (LLMs) generally lacking the precision of neural machine translation (NMT) systems in traditional MT tasks [5, 6, 7], we aim to explore whether these models can generate all possible gendered

1) https://github.com/cl-tohoku/ngo_hh_gender_bias_dataset

forms in their outputs. Additionally, we use the gender accuracy metric to measure exact matches and manually analyze other possible translations. Using a simple prompt, we demonstrate that modern LLMs can provide both gendered options even without explicitly specifying them in the prompt. Our results show that Claude [8, 9, 10] consistently outperforms GPT [11, 12, 13, 14] models, while Gemini [15] models exhibit competitive performance across most languages. We also show how models behave in cases of ambiguous sentences where the exact translation is undetermined due to the absence of an official demonym. In such instances, models either rely on linguistic inference to construct plausible gendered forms or opt for alternative phrasings, such as *"I am from [country/region].,"* to maintain grammatical correctness and fluency.

2 Related Work

Gender bias in MT has been extensively studied, mostly focusing on English as a source language and high-resource target languages (e.g., Spanish, Arabic). Bias often arises from stereotypical associations or grammatical conventions, leading models to favor one gender over another when ambiguity exists. For example, WinoMT benchmark [16], MuST-SHE [17] and MMHB [18] were introduced to evaluate gender bias in MT systems, revealing the translation tools not only reflect biases present in the training data but also tend to default to one gender more frequently [19].

To address the task of generating equitable translations, one approach involves the use of a post-editing technique. The most popular solution is Google Translate's post-translation gender rewriter [20]. This method creates an initial translation, checks for gender-specific terms, rewrites to include alternative genders, and ensures the only difference is gender. At the moment, this system covers a limited amount of high-resourced languages.

With the advent of LLMs, several studies have evaluated the performance of different models on machine translation tasks and gender bias. These include LLaMa [21], Flor [21], and some commercial products based on GPT such as ChatGPT [22], Gemini [22], and PALM [23]. While base LLMs tend to lag behind NMT models in translation capabilities, recent research has shifted focus toward leveraging prompts to mitigate gender bias rather than solely improving the underlying model. This move is driven by evidence that LLMs allow for more control over output properties, making prompt engineering an effective tool for reducing bias. For instance, prompt structures have been shown to reduce gender bias by up to 12% on the WinoMT evaluation dataset compared to simpler prompts [21]. Another paper demonstrates that LLaMa's gender-specific translation accuracy consistently outperforms NLLB's, with a comparable level of gender bias [24].

3 Methodology

3.1 Dataset

For our experiment, we have created a dataset consisting of sentences structured as *"I am [demonym]."* in English (eng) along with their translations into German (deu), French (fra), Italian (it), and Spanish (spa) for masculine (m) and feminine (f), or neuter (n) forms (Figure 1). The dataset was compiled using the official demonyms of the 193 member states of the United Nations [25], ensuring comprehensive global representation. Translations were sourced from publicly available resources, such as language learning websites. To maintain alignment with international standards and avoid potential geopolitical sensitivities, unrecognized or partially recognized countries as well as observer states were excluded.

While most translations matched the structure *"I am [demonym].,"* a small number of entries could not be translated because the target languages do not have an official demonym. For example, in German, for *"I am Emirati."*, there is no official masculine, feminine, or neutral demonym, leaving these fields blank. These N/A entries are included to analyze how models handle cases of absent

Table 1 Counts of masculine, feminine, neuter, and N/A entries for each language. N/A entries denote cases where the translation cannot be precisely matched due to the absence of an official demonym. In such instances, alternative expressions structured as *"I am from [country/region]"* are commonly used in place of *"I am [demonym]."*

Language	Masculine	Feminine	Neuter	N/A
eng	-	-	193	-
deu	187	187	2	4
fra	170	170	23	-
it	125	125	65	3
spa	144	144	49	-

Table 2 Gender accuracy of GPT, Gemini, and Claude models for each language and their average (avg.) performance across all languages. Results are reported as percentages in the format "`masculine / feminine / neuter`." Highest value per language for each gender is in **bold**, while underlined indicates the lowest score.

Model	deu	fra	it	spa	avg.
gpt-3.5-turbo	75.9 / 7.0 / 50.0	82.4 / 18.8 / 82.6	61.6 / 45.6 / 53.8	88.2 / 57.6 / 67.3	77.6 / 29.6 / 63.3
gpt-4	73.3 / 1.6 / 0.0	85.9 / 0.6 / 82.6	83.2 / 0.0 / 73.8	84.7 / 0.0 / 59.2	81.3 / 0.6 / 69.1
gpt-4-turbo	79.7 / 26.2 / **100.0**	**88.8** / 25.9 / 87.0	86.4 / 65.6 / 78.5	85.4 / 34.7 / 57.1	84.8 / 35.9 / 72.7
gpt-4o	84.0 / 71.1 / **100.0**	86.5 / 81.2 / 82.6	88.0 / 83.2 / 81.5	91.0 / 61.8 / 73.5	87.1 / 74.1 / 79.1
gpt-4o-mini	70.1 / 62.6 / 50.0	85.3 / 72.9 / 87.0	84.8 / 75.2 / 81.5	88.9 / 59.7 / 65.3	81.5 / 67.3 / 76.3
gemini-1.5-flash	72.2 / 55.6 / 0.0	80.6 / 78.8 / 82.6	82.4 / 73.6 / 81.5	86.8 / **71.5** / 67.3	79.9 / 69.2 / 75.5
gemini-1.5-pro	80.7 / 74.9 / **100.0**	**88.8** / 86.5 / 91.3	**89.6** / 80.0 / **84.6**	91.7 / 61.1 / 73.5	87.2 / 75.9 / **82.0**
claude-3-opus	82.9 / 76.5 / 50.0	**88.8** / 86.5 / 91.3	88.0 / 83.2 / 75.4	**93.1** / 68.8 / **79.6**	87.9 / 78.8 / 79.1
claude-3.5-haiku	78.1 / 65.8 / 50.0	86.5 / 84.7 / 73.9	85.6 / 78.4 / 75.4	88.2 / 63.9 / 61.2	84.2 / 73.0 / 69.8
claude-3.5-sonnet	**88.2** / **86.6** / **100.0**	87.6 / **87.1** / **100.0**	**89.6** / **85.6** / 78.5	91.7 / 66.7 / 77.6	**89.1** / **81.9** / **82.0**

translations (Table 1).

This dataset was specifically designed rather than using existing ones to address the following considerations. First, since the gender ratio in the human population is generally close to 50/50, using demonyms offers a more balanced and neutral representation compared to datasets focused on stereotypical and non-stereotypical gender associations for different occupations [26] (e.g., "doctor" "nurse," or "engineer"). Such datasets often reflect societal biases and skewed gender associations toward traditional roles. Additionally, they cover a limited number of languages and are primarily focused on high-resource language pairs. On the other hand, our dataset is simpler, which makes scaling to other languages in future work time-efficient and cost-effective.

3.2 Models

In this paper, we use API-based access to three popular families of state-of-the-art LLMs to evaluate the gender-specific translation task. Specifically, we experiment with OpenAI's[2] GPT-3.5-turbo, GPT-4, GPT-4-turbo, GPT-4o, and GPT-4o-mini. From Google DeepMind,[3] we use Gemini-1.5-Flash and Gemini-1.5-Pro. From Anthropic[4] we include Claude-3-Opus, Claude-3.5-Haiku, and Claude-3.5-Sonnet.

To ensure a fair comparison, we employ the same prompt across all models:

```
Can you translate the following sentence into
<target language>: <sentence in English>
```

We use this scenario because it is probably closer to how an MT-user would prompt since they are not necessarily aware of the fact that the target language might differ from the source in terms of gender marking [2].

3.3 Evaluation

For performance evaluation, we use gender accuracy – the percentage of instances the translation had the correct gender [16]. However, since the input sentence is neuter, there is a lack of information about gender. Therefore, we compute the metric per gender to identify the bias.

For sentences with absent reference translations (4 entries in German and 3 - in Italian), we focus on analyzing the outputs provided by the models rather than evaluating the translations.

4 Results and Analysis

As shown in Table 2, Claude-3.5-sonnet outperformed other models, achieving the highest accuracy for all gender forms in multiple languages. Other Claude models also show high results with a difference between 5% – 10%. In contrast, GPT-3.5-turbo achieved the lowest average score of 77.6/29.6/63.3. Notably, GPT-4 strongly defaulted towards masculine output, leading to significantly lower feminine accuracy. Newer models, GPT-4o and GPT-4o-mini,

2) https://openai.com/about/
3) https://deepmind.google/
4) https://www.anthropic.com/

Table 3 Example of LLMs handling N/A entries (when an official demonym is unavailable, making an exact translation unfeasible) in the English-to-German direction. Colored text indicates gender-specific forms: masculine (**m**), feminine (**f**), and unmarked for neuter or only one output option (**m/n**). The **(m/n)** notation indicates cases where only one translation was provided, making it unclear whether the model intended the output as masculine or neuter based on the ending.
Highlighted text represents alternative translations where "aus" means "from" and "Trinidad" refers to the region name.

Source	I am Trinidadian.
Expected output	-
gpt-3.5-turbo	Ich bin aus Trinidad.
gpt-4	Ich bin Trinidader. (m/n)
gpt-4-turbo	Ich bin Trinidadier. (m/n)
Gpt-4o	Ich bin Trinidadier. (m/n)
Gpt-4o-mini	Ich bin Trinidadier. (m/n)
gemini-1.5-flash	Ich bin Trinidadier. (m) Ich bin Trinidadierin. (f)
gemini-1.5-pro	Ich bin Trinidader. (m) Ich bin Trinidaderin. (f)
claude-3-opus	Ich bin Trinidader. (m/n)
claude-3.5-haiku	Ich bin Trinidadier. (m) Ich bin Trinidadierin. (f)
claude-3.5-sonnet	Ich bin Trinidader. (m) Ich bin Trinidaderin. (f)

showed substantial improvements in providing both gender forms. Their performance is comparable to that of Gemini models. On average, masculine forms still appear in a greater frequency than feminine forms, yet the difference is approximately 12.5%. Despite this, the results demonstrate significant progress in gender-specific tasks.

Analyzing N/A entries, where official gender-specific forms are not approved yet, reveals interesting patterns in model behavior. In these cases, models often rely on grammatical inference to construct outputs, applying linguistic rules to generate non-standard forms (Table 3). While these forms demonstrate the model's ability to generalize and adapt linguistic rules, they also highlight a tendency to prioritize grammatical plausibility over cultural or contextual accuracy.

In other cases, models generate alternative translations that are similar in meaning (Table 3), effectively avoiding the need for a gender-specific term. This approach is often the best strategy when no official demonym exists, ensuring grammatical correctness and fluency.

However, this behavior was also observed in cases where official demonyms exist. For example, for "Djibouti" claude-3.5-sonnet avoided using the officially recognized forms ("Yibutiano" or "Yibutiana" in Spanish) and instead generated output "Soy de Yibuti." (*"I from Djibouti."*).

5 Conclusion and Future Work

In this paper, we explored the capabilities of LLMs to produce gender-specific translations using a purpose-built dataset. Our results demonstrate that Claude-3.5-sonnet consistently achieves the highest accuracy across gender forms and multiple languages, with other Claude and Gemini models also performing strongly. We also recognize that GPT-4 struggled to provide balanced translations, frequently relying only on masculine outputs and neglecting feminine forms, an issue that appears to have been addressed in the later GPT-4o and GPT-4o-mini models.

We observed that the models applied different strategies for handling N/A entries, such as grammatical inference or generating alternative constructions. While these approaches were effective in maintaining grammatical correctness, they sometimes deviated from expected outputs, even when official demonyms were available.

In future work, we aim to expand the dataset by including more languages with grammatical gender, particularly low-resourced ones, to enable broader evaluation and analysis. Furthermore, conducting experiments with other gender-neutral source languages, such as Japanese, would provide valuable insights into how models handle different directions.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers JP22H00524 and JP24K03236, and JST SPRING Grant Number JPMJSP2114.

References

[1] Yasmeen Hitti, Eunbee Jang, Ines Moreno, and Carolyne Pelletier. Proposed taxonomy for gender bias in text; a filtering methodology for the gender generalization subtype. In Marta R. Costa-jussà, Christian Hardmeier, Will Radford, and Kellie Webster, editors, **Proceedings of the First Workshop on Gender Bias in Natural Language Processing**, pp. 8–17, Florence, Italy, August 2019. Association for Computational Linguistics.

[2] Eva Vanmassenhove. Gender bias in machine translation and the era of large language models. **Gendered Technology in Translation and Interpreting: Centering Rights in the Development of Language Technology**, p. 225, 2024.

[3] Greville G. Corbett. **Gender**. Cambridge Textbooks in Linguistics. Cambridge University Press, 1991.

[4] Joel Escudé Font and Marta R. Costa-jussà. Equalizing gender bias in neural machine translation with word embeddings techniques. In Marta R. Costa-jussà, Christian Hardmeier, Will Radford, and Kellie Webster, editors, **Proceedings of the First Workshop on Gender Bias in Natural Language Processing**, pp. 147–154, Florence, Italy, August 2019. Association for Computational Linguistics.

[5] Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. Is chatgpt a good translator? yes with gpt-4 as the engine. **arXiv preprint arXiv:2301.08745**, 2023.

[6] Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. Multilingual machine translation with large language models: Empirical results and analysis. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, **Findings of the Association for Computational Linguistics: NAACL 2024**, pp. 2765–2781, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[7] Rachel Bawden and François Yvon. Investigating the translation performance of a large multilingual language model: the case of BLOOM. In Mary Nurminen, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, Eva Vanmassenhove, Sergi Alvarez Vidal, Nora Aranberri, Mara Nunziatini, Carla Parra Escartín, Mikel Forcada, Maja Popovic, Carolina Scarton, and Helena Moniz, editors, **Proceedings of the 24th Annual Conference of the European Association for Machine Translation**, pp. 157–170, Tampere, Finland, June 2023. European Association for Machine Translation.

[8] Anthropic. Introducing the next generation of claude, 2024. https://www.anthropic.com/news/claude-3-family.

[9] Anthropic. Claude 3.5 sonnet, 2024. https://www.anthropic.com/news/claude-3-5-sonnet.

[10] Anthropic. Claude 3.5 haiku, 2024. https://www.anthropic.com/claude/haiku.

[11] OpenAI. Gpt-3.5 turbo, 2024. https://platform.openai.com/docs/models/gpt-3-5-turbo#gpt-3-5-turbo.

[12] OpenAI. Gpt-4 turbo and gpt-4, 2024. https://platform.openai.com/docs/models/#gpt-4-turbo-and-gpt-4.

[13] OpenAI. Gpt-4o, 2024. https://platform.openai.com/docs/models/#gpt-4o.

[14] OpenAI. Gpt-4o mini, 2024. https://platform.openai.com/docs/models/#gpt-4o-mini.

[15] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. **arXiv preprint arXiv:2403.05530**, 2024.

[16] Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. Evaluating gender bias in machine translation. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 1679–1684, Florence, Italy, July 2019. Association for Computational Linguistics.

[17] Luisa Bentivogli, Beatrice Savoldi, Matteo Negri, Mattia A. Di Gangi, Roldano Cattoni, and Marco Turchi. Gender in danger? evaluating speech translation technology on the MuST-SHE corpus. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 6923–6933, Online, July 2020. Association for Computational Linguistics.

[18] Xiaoqing Ellen Tan, Prangthip Hansanti, Carleigh Wood, Bokai Yu, Christophe Ropers, and Marta R Costa-jussà. Towards massive multilingual holistic bias. **arXiv preprint arXiv:2407.00486**, 2024.

[19] Marcelo OR Prates, Pedro H Avelar, and Luís C Lamb. Assessing gender bias in machine translation: a case study with google translate. **Neural Computing and Applications**, Vol. 32, pp. 6363–6381, 2020.

[20] Melvin Johnson. A scalable approach to reducing gender bias in google translate. In **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, Online, 2020. Association for Computational Linguistics.

[21] Aleix Sant, Carlos Escolano, Audrey Mash, Francesca De Luca Fornaciari, and Maite Melero. The power of prompts: Evaluating and mitigating gender bias in mt with llms. **arXiv preprint arXiv:2407.18786**, 2024.

[22] Faiz Algobaei, Elham Alzain, Ebrahim Naji, and Khalil A Nagi. Gender issues between gemini and chatgpt: The case of english-arabic translation. **World Journal of English Language**, Vol. 15, No. 1, p. 9, 2024.

[23] Mara Nunziatini and Sara Diego. Implementing gender-inclusivity in MT output using automatic post-editing with LLMs. In Carolina Scarton, Charlotte Prescott, Chris Bayliss, Chris Oakley, Joanna Wright, Stuart Wrigley, Xingyi Song, Edward Gow-Smith, Rachel Bawden, Víctor M Sánchez-Cartagena, Patrick Cadwell, Ekaterina Lapshinova-Koltunski, Vera Cabarrão, Konstantinos Chatzitheodorou, Mary Nurminen, Diptesh Kanojia, and Helena Moniz, editors, **Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)**, pp. 580–589, Sheffield, UK, June 2024. European Association for Machine Translation (EAMT).

[24] Eduardo Sánchez, Pierre Andrews, Pontus Stenetorp, Mikel Artetxe, and Marta R. Costa-jussà. Gender-specific machine translation with large language models. In Jonne Sälevä and Abraham Owodunni, editors, **Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)**, pp. 148–158, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[25] United Nations. Member states, 2024. https://www.un.org/en/about-us/member-states#gotoA.

[26] Karolina Stanczak and Isabelle Augenstein. A survey on gender bias in natural language processing. **arXiv preprint arXiv:2112.14168**, 2021.