Analyzing Gender Accuracy and Gender Quality in Multilingual Machine Translation with Large Language Models

Sarah Zhang
Massachusetts Institute of Technology
sjzhang@mit.edu

Lily Chen
Massachusetts Institute of Technology
l1ly@mit.edu

William Zhang
Duke University
william.zhang@duke.edu

Abstract

Gender bias remains a critical and relatively underexplored issue in the domain of multilingual machine translation (MMT), particularly within the context of large language models (LLMs). In this study, we analyze gender accuracy and gender quality in translation on the MT-GENEVAL benchmark using state-of-the-art models, such as `text-davinci-003`, `gpt-3.5-turbo`, and `gpt-4`, in both zero-shot and few-shot scenarios. Our work suggests that the highest leverage action towards equitable LLM-based MMT is a shift towards confronting and rectifying biases at the foundational level of LLMs, rather than relying solely on post-pretraining strategies as a solution.

1 Introduction

Although previous studies have evaluated the multilingual machine translation (MMT) capabilities of large language models (LLMs) (Jiao et al., 2023; Kocmi and Federmann, 2023; Peng et al., 2023; Zhu et al., 2023), a comprehensive evaluation of gender machine translation performance is still not present in the literature. As LLM-based MMT enters widespread use and deployment, it is of utmost importance that LLMs refrain from encoding and spreading harmful social and gender biases (Weidinger et al., 2021, 2022).

In this paper, we focus on evaluating and enhancing the **gender accuracy** and **gender quality** in translation using the MT-GENEVAL benchmark (Currey et al., 2022). We define *gender accuracy* as the extent to which a MMT output represents the gender of the individuals and *gender quality* as the extent to which a MMT output demonstrates representational bias (Blodgett et al., 2020).

2 Methods

Dataset MT-GENEVAL covers translation from English (EN) into eight languages: Arabic (AR), French (FR), German (DE), Hindi (HI), Italian (IT), Portuguese (PT), Russian (RU), and Spanish (ES). We explore the *counterfactual* subset, where each source segment is associated with individuals of a specific gender, either female or male.

Models We evaluate the translation performance of three LLMs: `text-davinci-003`, `gpt-3.5-turbo`, and `gpt-4` (OpenAI, 2023). We use a temperature of 0 and top-p sampling of 1.

Translation Accuracy We measure translation quality using the following automatic metrics: BLEU (Papineni et al., 2002; Post, 2018), COMET-22 (Rei et al., 2020), and CHRF++ (Popović, 2017).

Gender Accuracy Let w_{hyp}, w_{ref} and w_{con} be the set of words in the hypothesis, reference, and contrastive reference, respectively. A segment is considered *correct* if the hypothesis does not contain words unique to the contrastive gender:

$$w_{hyp} \cap w_{con} \backslash w_{ref} = \emptyset. \quad (1)$$

In the counterfactual subset, a segment pair is considered *correct* if and only if *both* the original and counterfactual segments are correct.

Gender Accuracy+ Since it is possible for a segment to be considered correct even if the hypothesis does not contain gender-specific words, we introduce a novel constrained metric where a segment is considered *correct* if the hypothesis does not contain words unique to the contrastive gender and does contain words unique to the gender:

$$\text{Eq. (1)} \cap (w_{hyp} \cap w_{ref} \backslash w_{con} \neq \emptyset). \quad (2)$$

Gender Quality Gap Let BLEU_{gender} be the BLEU score of the *gender* source segments in the counterfactual subset. To quantitatively assess whether there exists a disparity in translation quality between the genders, we calculate the difference in score of the masculine and feminine segments:

$$\Delta_{qual} = \text{BLEU}_{male} - \text{BLEU}_{female}. \quad (3)$$

Language	Model	0-Shot						1-Shot					
		BLEU↑	COMET↑	CHRF++↑	ACC↑	ACC+↑	$\Delta_{qual\|\downarrow\|}$	BLEU↑	COMET↑	CHRF++↑	ACC↑	ACC+↑	$\Delta_{qual\|\downarrow\|}$
EN → AR	text-davinci-003	15.72	0.773	43.62	0.787	0.530	-0.334	15.99	0.733	43.88	**0.847**	0.580	**-0.328**
	gpt-3.5-turbo	**25.84**	0.842	54.68	0.763	0.623	**-0.121**	**26.10**	0.844	54.64	**0.847**	**0.707**	-0.403
	gpt-4	25.27	**0.848**	**55.38**	**0.777**	**0.630**	0.598	25.36	**0.848**	**55.39**	0.813	0.667	0.367
EN → DE	text-davinci-003	41.09	0.857	66.22	0.647	0.577	1.077	41.40	0.857	66.34	0.710	0.633	0.864
	gpt-3.5-turbo	45.22	0.871	41.62	**0.713**	**0.653**	**0.055**	45.40	0.869	69.31	**0.747**	**0.683**	-0.126
	gpt-4	**46.29**	**0.873**	**70.26**	0.707	**0.653**	0.245	**46.40**	**0.874**	**70.32**	0.737	**0.683**	**0.033**
EN → ES	text-davinci-003	49.69	0.864	71.51	0.633	0.527	0.381	50.55	0.867	72.19	0.693	0.583	0.442
	gpt-3.5-turbo	52.30	**0.877**	73.61	0.700	0.567	1.173	52.22	0.877	73.66	0.723	0.590	1.102
	gpt-4	**53.34**	**0.877**	**74.18**	**0.723**	**0.587**	**-0.061**	**53.37**	**0.878**	**74.13**	**0.750**	**0.613**	**0.149**
EN → FR	text-davinci-003	40.02	0.843	64.65	0.617	0.587	1.701	41.07	0.847	65.20	0.678	0.647	2.259
	gpt-3.5-turbo	**44.14**	**0.861**	**67.75**	0.693	0.663	1.270	**44.23**	**0.860**	**67.90**	0.730	0.700	0.907
	gpt-4	43.11	0.857	66.89	**0.713**	**0.687**	**0.140**	43.44	0.858	67.09	**0.747**	**0.720**	**0.397**
EN → HI	text-davinci-003	16.47	0.696	42.25	0.630	0.193	**1.087**	16.72	0.692	42.64	0.680	0.217	**1.016**
	gpt-3.5-turbo	22.88	0.756	48.82	**0.680**	0.283	1.947	22.61	0.761	49.02	**0.753**	0.320	1.224
	gpt-4	**25.61**	**0.787**	**52.66**	0.657	**0.347**	1.283	**25.93**	**0.784**	**52.66**	0.700	**0.370**	1.093
EN → IT	text-davinci-003	36.36	0.861	62.84	0.630	0.477	2.011	37.17	0.864	63.27	0.693	**0.540**	2.030
	gpt-3.5-turbo	38.23	0.873	64.47	0.643	0.490	**1.783**	38.52	0.872	64.55	**0.697**	**0.540**	1.672
	gpt-4	**39.10**	**0.875**	**65.28**	**0.653**	**0.507**	2.017	**39.91**	**0.876**	**65.68**	0.683	0.533	**1.560**
EN → PT	text-davinci-003	47.23	0.876	70.41	0.633	0.603	**2.285**	47.30	0.876	70.37	0.670	0.640	**1.722**
	gpt-3.5-turbo	52.08	**0.892**	73.64	0.667	0.643	3.096	52.01	0.888	73.50	**0.717**	**0.693**	2.606
	gpt-4	**52.39**	**0.892**	**73.82**	**0.680**	**0.660**	2.523	**52.97**	**0.892**	**74.11**	0.710	0.690	2.203
EN → RU	text-davinci-003	29.79	0.858	56.34	0.733	0.630	**1.895**	30.45	0.860	56.80	**0.810**	0.703	**1.369**
	gpt-3.5-turbo	34.59	0.878	60.66	0.703	0.630	2.422	35.04	0.878	60.88	0.763	0.687	2.500
	gpt-4	**36.54**	**0.885**	**62.32**	**0.763**	**0.687**	1.924	**36.35**	**0.885**	**62.07**	0.800	**0.720**	1.885

Table 1: 0-Shot and 1-Shot BLEU, COMET, CHRF++, ACC, ACC+, and Δ_{qual} scores on the MT-GENEVAL dataset. The **bold** and underlined texts indicate the best performance across the selected LLMs and metrics, respectively.

Zero-Shot and Few-Shot Translation In the zero-shot setting, we formulate the inference prompt for the target language $\boldsymbol{x}$ and source segment $\boldsymbol{y}$ as "Translate the following English text to $\boldsymbol{x}$:$\boldsymbol{y}$." In the few-shot setting, we first embed the source segments in both the development and test sets using `text-embedding-ada-002`. Next, we identify the development segment with the highest similarity to each test segment based on cosine similarity. For example, the development segment "He was nominated for the Nobel Prize in Literature three times" is the most similar to the test segment "Because of his life work, for which he received numerous awards, he is among top Yugoslav and Serbian youth writers." We then include this source segment and its corresponding reference translation in the prompt to serve as an in-context example (Lin et al., 2022; Agrawal et al., 2023).

Masculine Source: The mills are associated with their builder, the prominent miller John Lucas and through him his father Nathanial Lucas.

Masculine 1-Shot: Os moinhos estão associados ao seu construtor, o proeminente moleiro John Lucas e, através dele, seu pai Nathanial Lucas.

Masculine Reference: Os moinhos estão associados ao seu construtor, o proeminente moleiro John Lucas e através de seu pai, Nathanial Lucas.

Feminine Source: The mills are associated with their builder, the prominent miller John Lucas and through her her mother Nathanial Lucas.

Feminine 1-Shot: Os moinhos estão associados ao seu construtor, o proeminente moleiro John Lucas e, através dele, à sua mãe Nathanial Lucas.

Feminine Reference: Os moinhos estão associados à sua construtora, a proeminente moleira John Lucas e através de sua mãe, Nathanial Lucas.

Figure 1: While "builder" and "miller" are translated to "construtor" (male builder) and "moleiro" (male miller) in the masculine source, they are translated to "construtor" instead of "construtora" (female builder) and "moleiro" instead of "moleira" (female miller) in the feminine source.

3 Results and Conclusion

Despite providing the LLMs with an example of a gender-accurate translation, we observe in Table 1 that the translation quality for masculine segments often surpasses that of feminine segments ($\Delta_{qual} > 0$), even though the source segments are nearly identical except for their gender-specific word(s). To delve deeper into this phenomenon, we conducted a random examination of 100 source segments from the test set where $\text{BLEU}_{male} > \text{BLEU}_{female}$. We find a substantial number of these segments exhibit occupational bias, as highlighted in Figure 1. This suggests that certain biases may be deeply ingrained within the model, making them challenging to mitigate through in-context learning or fine-tuning alone.

Our findings advocate for adopting a multi-dimensional approach to address gender accuracy and quality in LLMs. This approach extends beyond the confines of conventional techniques, emphasizing the need for these systems to instead evolve through a deeper understanding of the underlying training data and model architecture, aligning more closely with principles of fairness and equity.

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context examples selection for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873, Toronto, Canada. Association for Computational Linguistics.

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.

Anna Currey, Maria Nadejde, Raghavendra Reddy Pappagari, Mia Mayer, Stanislas Lauly, Xing Niu, Benjamin Hsu, and Georgiana Dinu. 2022. MT-GenEval: A counterfactual and contextual dataset for evaluating gender accuracy in machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4287–4299, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of chatgpt for machine translation.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from language models. *CoRR*, abs/2112.04359.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 214–229, New York, NY, USA. Association for Computing Machinery.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Multilingual machine translation with large language models: Empirical results and analysis.