Gender-specific Machine Translation with Large Language Models

Eduardo Sánchez[⋆,†] **Pierre Andrews**[⋆] **Pontus Stenetorp**[†]
Mikel Artetxe[‡] **Marta R. Costa-jussà**[⋆]
[⋆]Meta [†]University College London
[‡]University of the Basque Country (UPV/EHU)
{eduardosanchez, mortimer, costajussa}@meta.com
p.stenetorp@cs.ucl.ac.uk mikel.artetxe@ehu.eus

Abstract

While machine translation (MT) systems have seen significant improvements, it is still common for translations to reflect societal biases, such as gender bias. Decoder-only language models (LLMs) have demonstrated potential in MT, albeit with performance slightly lagging behind traditional encoder-decoder neural machine translation (NMT) systems. However, LLMs offer a unique advantage: the ability to control the properties of the output through prompting. In this study, we leverage this flexibility to explore Llama's capability to produce gender-specific translations. Our results indicate that Llama can generate gender-specific translations with translation quality and gender bias comparable to NLLB, a state-of-the-art multilingual NMT system.

```
Translate the following English sentences into Spanish.

[English] I have friends who are Hispanic people.
[Spanish] Tengo amigos que son personas hispanas.

[English] What do you think about ginger children?
[Spanish] ¿Qué piensas de las niñas pelirrojas?

[English] I have friends who are orphans.
[Spanish] Tengo amigos huérfanos.
```

(a) Standard MT template.

```
Translate the following English sentences into both
grammatical genders of Spanish. Provide two Spanish
sentences for each English source sentence.

[English] I have friends who are Hispanic people.
[Spanish] Tengo amigos que son personas hispanas.
[Spanish] Tengo amigas que son personas hispanas.

[English] What do you think about ginger children?
[Spanish] ¿Qué piensas de los niños pelirrojos?
[Spanish] ¿Qué piensas de las niñas pelirrojas?

[English] I have friends who are orphans.
[Spanish] Tengo amigos huérfanos.
[Spanish] Tengo amigas huérfanas.
```

(b) Gender-specific template.

Figure 1: Prompt templates employed. Languages like Spanish (displayed here) present morphological differences between masculine (red) and female (purple) grammatical genders.

1 Introduction

Over the last few years, machine translation (MT) systems have seen significant improvements with the introduction of Neural Machine Translation (NMT). Despite these advances, MT can reflect societal biases, such as gender bias. A prominent instance of this problem occurs when the target language marks the grammatical gender, but the source language does not (Fig. 1b). In such instances, translating into either gender can be correct, but MT systems tend to pick the gender that corresponds to stereotypical associations (e.g., associating certain professions to males and others to females (Escudé Font and Costa-jussà, 2019)). Instead, it would be preferable to generate both options, and/or let the user control the gender.

Decoder-only Large Language Models (LLMs) have shown MT capabilities inferior to but competitive with encoder-decoder Neural Machine Translation (NMT) systems (Chowdhery et al., 2022; Agrawal et al., 2023; Zhang et al., 2023; Bawden and Yvon, 2023; Zhu et al., 2023; Jiao et al., 2023; Hendy et al., 2023). However, LLMs have been proven to allow for more control over the properties of the output (Garcia et al., 2023; Moslem et al., 2023; Pilault et al., 2023). While NMT models are trained to accept a single sequence and output its translation, LLMs make it possible to condition the output format with a prompt.

The task of inferring gender-specific translations from a gender-neutral source has been addressed mainly through post-editing, the most popular solution being Google Translate's post-translation gender rewriter (Johnson, 2020). The proposed system produces a single sentence that is then switched into the opposite gender using a second language-specific model. This approach is limited by having to train language-specific gender-switching models

Lang	Source Sentence	Masculine	Feminine	Neutral/Generic
est	What do you think about Hasidic children?	-	-	Mida arvate lastest, kes on hassiidid?
spa	What do you think about Hasidic children?	¿Qué piensas de los niños judíos jasídicos?	¿Qué piensas de las niñas judías jasídicas?	-
kat	What do you think about Hasidic children?	-	-	რას ფიქრობ სასიდ ბავშვებზე? (ras pikrob sasid bavshvebze?)

Table 1: Example of entries in the full MULTILINGUALHOLISTICBIAS dataset. From the 50 languages present in MULTILINGUALHOLISTICBIAS, we selected those that exhibit grammatical gender differences, like Spanish (spa). The sentence in Georgian (kat) has been transcribed into Latin script by the authors.

and the breadth of patterns it can cover.

Given the flexibility of prompting, we explore the capacity of LLMs to produce gender-specific translations for languages with grammatical gender from gender-neutral sources without significant losses in translation quality or increases in gender bias.

We use in-context examples (ICEs) to elicit the task of translation from a gender-neutral source to two gender-specific targets (Figure 1b). Additionally, we evaluate the quality of the gender-specific translations on two aspects: gender bias (measured against coreference resolution accuracy) and translation quality (measured in BLEU).

We show that it is possible to generate gender-specific translations with translation quality and gender bias competitive with NLLB, with a slightly better performance than Llama for masculine/both references evaluation and over 10 BLEU points for the feminine reference. We also demonstrate the reliance on coreference resolution of the gender-specific translation method, showing steep decreases in performance when using the opposite gender as an evaluation reference in a gender-focused dataset (MULTILINGUALHOLISTICBIAS), but exhibiting lesser variance in a general translation dataset (FLoRes).

2 Related Work

MT and controlled output with LLMs A few papers have evaluated the quality of MT using different models and GPT-based commercial products, such as PALM (Chowdhery et al., 2022), XGLM (Agrawal et al., 2023), GLM (Zhang et al., 2023), BLOOM (Bawden and Yvon, 2023), OPT (Zhu et al., 2023) or ChatGPT (Jiao et al., 2023; Hendy et al., 2023). They conclude that the translation quality comes close but remains behind the performance of NMTs (Kocmi et al., 2023). Using LLMs can, however, allow for more control over the properties of the output without further finetuning, such as specifying the language variety and style of the translation (Garcia et al., 2023), producing terminology-constrained translations (Moslem et al., 2023) or using an iterative prompting process to clarify ambiguities in the source sentence (Pilault et al., 2023). Challenges persist in the area of hallucinations (Zhang et al., 2023; Guerreiro et al., 2023) and in performance in low-resource languages (Bawden and Yvon, 2023; Zhu et al., 2023). This work revisits these ideas, taking gender specificity as a controllable feature.

Gender Bias in MT Some authors have worked in analyzing and mitigating gender bias in MT. Prates et al. (2018) studied the bias of the commercial translation system Google Translate and found that it yields male defaults much more frequently than what would be expected from US demographic data. Costa-jussà et al. (2022) investigate the role of model architecture in the level of gender bias, while Měchura (2022) looks at the source sentences and elaborates a taxonomy of the features that induce gender bias into the translations. Others have looked more closely at the challenge of gender bias mitigation. Stafanovičs et al. (2020) assume that it's not always possible to infer all the necessary information from the source sentence alone and a method that uses word-level annotations containing information about the subject's gender to decouple the task of performing an unbiased translation from the task of acquiring gender-specific information. Saunders and Byrne (2020) treat the mitigation as a domain adaptation problem, using transfer learning on a small set of trusted, gender-balanced examples to achieve considerable gains with a fraction of the from-scratch

	cat	deu	fra	ita	nld	por	rus	spa	swe	ukr	avg
nllb	45.81	43.38	53.43	36.34	33.96	53.05	38.40	32.99	47.58	36.31	42.13
unsp.	46.05	41.79	52.24	34.70	32.54	51.76	36.17	31.34	47.74	36.02	41.04
masc.	46.06	42.18	52.05	34.46	32.36	51.68	36.23	31.25	47.90	36.05	41.02
fem.	43.83	41.02	50.25	33.25	31.43	49.29	34.57	29.72	47.63	35.38	39.64
Δ_F	2.23	1.16	1.80	1.21	0.93	2.39	1.66	1.53	0.27	0.67	1.39

Table 2: BLEU scores for each output of Llama's gender-specific translation on FLoRes's testset. Δ_F denotes the difference between male and female translations. Since FLoRes's sentences are not expected to contain a high rate of ambiguity, a correct translation should tend to be identical in both outputs.

training costs. Fleisig and Fellbaum (2022) develop a framework to make NMT systems suitable for gender bias mitigation through adversarial learning, adjusting the training objective at fine-tuning time. Finally, Wang et al. (2022) focus on existing biases in person name translation, applying a data augmentation technique consisting of randomly switching entities, obtaining satisfactory results. Given this work's focus area, we aim not only at producing accurate gender-specific translations, but also at ensuring selecting an output gender does not increase reproduction of underlying gender biases.

3 Experimental Framework

Data For our main experiments, we use the MULTILINGUALHOLISTICBIAS dataset (Costa-jussà et al., 2023), a multilingual subset of Holistic Bias (Smith et al., 2022) with separate translations for each noun class or grammatical gender for those languages that make use of them[1]. An example of an entry of the dataset can be found in Table 1. We also filtered out the languages which are not explicitly present in the Llama-2 pre-training set (Touvron et al., 2023). Since MHB was created translating a limited number of templates, we exclude entries with a similar template when performing ICL. A complete list of languages used from the MULTILINGUALHOLISTICBIAS dataset can be found in Appendix A. Additionally, we use a subset of BUG's (Levy et al., 2021) gold (human-annotated) set for gender bias analysis and the FLoRes (NLLB Team et al., 2022; Goyal et al., 2021a; Guzmán et al., 2019) devtest set to reproduce our results in the general domain.

[1]For this study, we selected the subset of languages that make use of grammatical genders or noun classes and for which there is correlation between grammatical gender and natural gender, allowing us to establish a relationship between gender bias and the accuracy of coreference resolution in a model.

Models We use Llama-2 (Touvron et al., 2023), a decoder-only model, and NLLB (NLLB Team et al., 2022), an encoder-decoder model. We use the NLLB-200 version with 3 billion parameters. For Llama-2 we use the 70 billion parameter version. We prompt Llama-2 with ICEs (Figure 1b) to elicit the gender-specific translation task. To facilitate comparisons, we also prompt Llama-2 with a standard MT in-context learning (ICL) prompt template (Figure 1a).

Evaluation Following the work of Costa-jussà et al. (2023), we use the sacrebleu implementation of spBLEU (Goyal et al., 2021b) to compute the translation quality with `'add-k = 1'` smoothing. We also provide evaluations in chrF (Popović, 2015), COMET (Rei et al., 2020), BLEURT (Sellam et al., 2020) and BLASER (Chen et al., 2023) as alternative metrics. For gender bias evaluation, we use Stanovsky et al. (2019)'s reference-less coreference resolution metric.

Experimental Setup We investigate the capability of Llama to produce gender-specific translations. We prompt Llama with 8 ICEs comprised by source, masculine and feminine translations from MULTILINGUALHOLISTICBIAS (Fig. 1b). We also prompt Llama with a standard MT template, randomly selecting among the available translations when there's more than one option (Fig. 1a). Hereinafter all experiments are performed with these settings. For NLLB, we calculate three BLEU scores on the output: one with the masculine reference, one with the feminine reference and one with both. In the case of Llama, we calculate two BLEU scores for each gender-specific output: one with the corresponding gender's reference and one with both references, for a total of four BLEU scores per generation.

		masc	fem	both
NLLB	unsp	40.07	28.67	40.41
Llama	unsp	41.57	30.92	42.43
	masc	**41.63**	30.12	42.08
	fem	31.84	**39.55**	**43.37**

Table 3: BLEU scores of the unspecified, masculine and feminine outputs of NLLB and Llama evaluated on masculine, feminine, and both references of MULTILINGUALHOLISTICBIAS

```
Source:    I have friends who are Hispanic people.
Masc:      Tengo amigos que son personas hispanas.
Fem:       Tengo amigos que son personas hispanos.
```

(a) Error generating Spanish language.

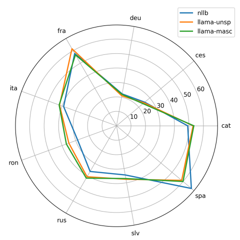
```
Source:    I have friends who are American-born people.
Masc:      J'ai des amis qui sont des Américains nés aux États-Unis.
Fem:       J'ai des amis qui sont des Américaines nées aux États-Unis.
```

(b) Error generating French language.

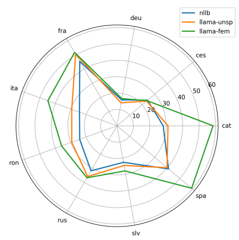
Figure 3: Two examples of the sort of errors encountered while generating gender-specific translations. In some cases, masculine gender marker is used in both outputs, particularly when there's already a female gender marker in another part of the sentence. In other cases, agreement errors are encountered in the feminine gender output.

4 Results

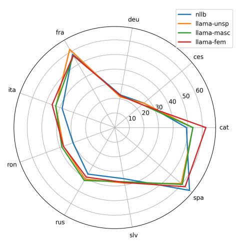
Gender-specific MT results in MULTILINGUALHOLISTICBIAS As Table 3 shows, on average Llama outperforms NLLB on all three references. While the differences between masculine/both references are moderate (Figs. 2a & 2c), Llama outperforms NLLB by an average of over 10 BLEU points for the feminine reference (Fig. 2b), highlighting the capacity of gender-specific MT to provide comparable results for masculine and feminine outputs. Some of the most common errors encountered while generating gender-specific translations can be found in Figure 3.



(a) Masculine reference



(b) Feminine reference



(c) Both references

Figure 2: BLEU scores of the unspecified, masculine and feminine outputs of NLLB and Llama-2 evaluated on masculine, feminine, and both references of MULTILINGUALHOLISTICBIAS.

We also assessed the capacity of the BLEU evaluation metric to capture gender nuances. We inverted masculine and feminine references and found steep decreases, indicating the effectiveness of our approach in specifying the grammatical gender. We provide evaluations in chrF, COMET, BLEURT and BLASER, which show consistency with BLEU scores. Full results can be found in Appendix B. Additionally, we included a comparison of results between LLama-2 and GPT-4o to validate whether our results are model-specific or can be generalized. We also find satisfactory results for GPT-4o (Table 4).

Gender bias MT results in BUG Besides translation accuracy, we're interested in verifying the incidence of gender bias in gender-specific translations with respect to unspecified translation. We translate BUG's gold set, reusing MULTILINGUALHOLISTICBIAS examples for ICL. BUG's gold set is made of English sentences that require unambiguous coreference resolution or grammatical gender utilization to produce correct translations, regardless of stereotypical associations. To ensure fairness in our analysis, we sampled four subsets of 90 sentences from BUG gold, each subset corresponding to a combination of stereotypical/antistereotypical correferences and male/female nouns. Stanovsky et al. (2019) and Levy et al. (2021) found that several (encoder-decoder) NMTs are significantly prone to translate based on gender stereotypes rather than more meaningful context. We verify to which degree these errors are reproduced by Llama in gender-specific translations. When performing the translation of BUG, we noticed that the phenomenon of empty or incomplete outputs occasionally occurs (i.e., either only one output or no output at all is produced).

Language		Llama	GPT-4o
cat	Masc.	53.36	58.44
	Fem.	58.56	60.63
ces	Masc.	23.85	21.83
	Fem.	23.88	30.54
deu	Masc.	22.04	35.93
	Fem.	16.88	36.89
fra	Masc.	56.69	57.52
	Fem.	51.76	58.82
ita	Masc.	42.16	39.61
	Fem.	44.86	40.45
ron	Masc.	36.85	34.92
	Fem.	35.96	35.17
rus	Masc.	41.81	42.49
	Fem.	36.67	43.82
slv	Masc.	37.07	38.55
	Fem.	27.98	35.42
spa	Masc.	59.94	61.84
	Fem.	59.36	62.61
avg	Masc.	41.53	43.46
	Fem.	39.55	44.93

Table 4: BLEU score comparison between LLama-2 and GPT-4o. Results remain competitive, further supporting the potential of LLMs to produce gender-specific translations.

	NLLB		Llama					
	unsp		unsp		masc		fem	
	acc.(↑)	Δ_B(↓)	acc.(↑)	Δ_B(↓)	acc.	Δ_B(↓)	acc.(↑)	Δ_B (↓)
ces	59.3	<u>6.5</u>	57.2	11.3	**61.7**	10.1	48.4	8.8
deu	66.4	11.8	67.8	10.8	**70.6**	9.5	52.4	<u>8.6</u>
ita	46.2	<u>12.5</u>	45.4	13.7	**46.5**	14.4	38.9	14.2
spa	**52.5**	<u>10.1</u>	50.0	11.4	49.4	14.4	34.2	29.4
rus	36.6	25.0	**39.5**	23.8	38.1	27.5	36.9	<u>16.7</u>
ukr	41.2	11.1	42.1	10.1	**43.2**	8.8	39.0	<u>1.0</u>

Table 5: Noun gender prediction accuracy on the subset of BUG's gold dataset's fully generated gender-specific translations with Llama, compared to NLLB's prediction accuracy. Llama results are presented for male (m.), female (f.), and unspecified (unsp.) genders. We also show the differences in accuracy between male nouns and female nouns for each case (Δ_B)

Since a gender bias analysis is not defined over an empty sentence, for each language we evaluate all models in the subset that has been correctly generated by Llama both in the unspecified and the gender-specific modalities.

Table 5 shows that Llama's masculine output's noun gender prediction accuracy outperforms NLLB's for almost every language, but underperforms NLLB for feminine outputs. Difference of accuracy between genders for the same type of output (Δ_B) is comparable across models.

General domain MT results in FLoRes A possible concern about previous results is that they are produced by the system forcing a specific gender instead of performing coreference resolution to determine the correct gender. To study whether this is the case, we assess the difference in performance for each produced gender when there aren't major gender ambiguities to translate. In this case, a robust model should not have significant differences between both genders. We translate FLoRes's devtest set into ten languages included in Llama's training corpus. Given that FLoRes is a general domain dataset, ambiguities should not be prevalent and both outputs should tend to converge. We use MULTILINGUALHOLISTICBIAS as ICEs and compare the BLEU scores of both outputs. The list of languages we translate into for this experiment can be found in Table 6 (Appendix A).

The results show minor differences between both genders, suggesting a coreference resolution-based gender-specific generation rather than on mechanically switching the grammatical gender of the words of the sentence.

5 Conclusions

In this paper, we explored the capabilities and limitations a decoder-only LLM to produce gender-specific translations. We observed that Llama's gender-specific translations' accuracy is consistently above NLLB's. We also showed that Llama's gender-specific translations' gender bias is comparable to NLLB's. These results indicate that it is possible to use LLMs to produce gender-specific translations without compromising on lower translation accuracy or higher gender bias. Our experiments also reveal that Llama's translations rely on coreference resolution to determine gender, showing significant performance drops when evaluated against opposite-gender references in gender-ambiguous datasets, but maintaining consistency in less ambiguous contexts.

While these results are promising indicator of the flexibility of the output in the task of MT for languages present in Llama's training set, the limited multilinguality of currently available LLMs limits the application of this approach to a subset of the languages present in state-of-the-art NMT models. More work is needed to bring LLMs' multilingual capabilities on par with NMTs.

Limitations

Even though we performed a diverse set of experiments, some limitations arise due to the vastness of the research space we're dealing with. The study heavily relies on the effectiveness of prompt engineering, specifically in providing accurate ICEs. The conclusions drawn are thus constrained by the quality and relevance of the prompts used. Variations in prompt structure or content could yield different results. Moreover, the study focuses on a particular model, Llama-2, leaving out an exploration of alternative LLMs that could yield different results.

MULTILINGUALHOLISTICBIAS's small number of templates and their simplicity limit the scope of our results. An exploration with a more diverse dataset could bring additional insights to our conclusions.

Ethics Statement

The understanding of nuanced gender contexts is intricate and can be challenging even for humans. The study tends to approach gender in a binary manner, which might not account for social perceptions among some of the users of these languages. This limitation is inherent in the current state of the field and warrants future investigations into better representation and handling of gender-related nuances.

Furthermore, the stereotypical and non-stereotypical datasets were built based on the US Department of Labor data. Since we work with a variety of world languages, the proportions stated on these datasets might not reflect the realities of the users of the wide range of languages employed in this study.

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.

Rachel Bawden and François Yvon. 2023. Investigating the translation performance of a large multilingual language model: the case of bloom.

Mingda Chen, Paul-Ambroise Duquenne, Pierre Andrews, Justine Kao, Alexandre Mourachko, Holger Schwenk, and Marta R. Costa-jussà. 2023. BLASER: A text-free speech-to-speech translation evaluation metric. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9064–9079, Toronto, Canada. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Marta R. Costa-jussà, Pierre Andrews, Eric Smith, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Daniel Licht, and Carleigh Wood. 2023. Multilingual holistic bias: Extending descriptors and patterns to unveil demographic biases in languages at scale.

Marta R. Costa-jussà, Carlos Escolano, Christine Basta, Javier Ferrando, Roser Batlle, and Ksenia Kharitonova. 2022. Interpreting gender bias in neural machine translation: Multilingual architecture matters. Proceedings of the AAAI Conference on Artificial Intelligence, 36(11):11855–11863.

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

Eve Fleisig and Christiane Fellbaum. 2022. Mitigating gender bias in machine translation through adversarial learning.

Xavier Garcia, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Fangxiaoyu Feng, Melvin Johnson, and Orhan Firat. 2023. The unreasonable effectiveness of few-shot learning for machine translation.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2021a. The flores-101 evaluation benchmark for low-resource and multilingual machine translation.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2021b. The flores-101 evaluation benchmark for low-resource and multilingual machine translation.

Nuno M. Guerreiro, Duarte Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023. Hallucinations in large multilingual translation models.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. Two new evaluation datasets for low-resource machine translation: Nepali-english and sinhala-english.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation.

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.

Melvin Johnson. 2020. A scalable approach to reducing gender bias in google translate. Accessed: September 5th, 2023.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In Proceedings of the Eighth Conference on Machine Translation, pages 1–42, Singapore. Association for Computational Linguistics.

Shahar Levy, Koren Lazar, and Gabriel Stanovsky. 2021. Collecting a large-scale gender bias dataset for coreference resolution and machine translation. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2470–2480, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michal Měchura. 2022. A taxonomy of bias-causing ambiguities in machine translation. In Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 168–173, Seattle, Washington. Association for Computational Linguistics.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Jonathan Pilault, Xavier Garcia, Arthur Bražinskas, and Orhan Firat. 2023. Interactive-chain-prompting: Ambiguity resolution for crosslingual conditional generation with interaction.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Marcelo O. R. Prates, Pedro H. C. Avelar, and Luís C. Lamb. 2018. Assessing gender bias in machine translation - A case study with google translate. CoRR, abs/1809.02208.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation.

Danielle Saunders and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7724–7736, Online. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7881–7892, Online. Association for Computational Linguistics.

Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. "I'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 9180–9211, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Artūrs Stafanovičs, Toms Bergmanis, and Mārcis Pinnis. 2020. Mitigating gender bias in machine translation with target gender annotations. In Proceedings of the Fifth Conference on Machine Translation, pages 629–638, Online. Association for Computational Linguistics.

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.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Jun Wang, Benjamin Rubinstein, and Trevor Cohn. 2022. Measuring and mitigating name biases in neural machine translation. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2576–2590, Dublin, Ireland. Association for Computational Linguistics.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study.

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.

A Languages

Language code	Name	Script	MULTILINGUALHOLISTICBIAS	BUG	FLoRes
arb	Modern Standard Arabic	Arabic		✓	
cat	Catalan	Latin	✓		✓
ces	Czech	Latin	✓	✓	
deu	German	Latin	✓	✓	✓
fra	French	Latin	✓		✓
ita	Italian	Latin	✓	✓	✓
nld	Dutch	Latin			✓
por	Portugese	Latin			✓
ron	Romanian	Latin	✓		
rus	Russian	Cyrillic	✓	✓	✓
slv	Slovenian	Latin	✓		
spa	Spanish	Latin	✓	✓	✓
swe	Swedish	Latin			✓
ukr	Ukrainian	Cyrillic		✓	✓

Table 6: List of languages analyzed in this work by dataset

B Full Results

			Reference		
Language	Model	Type	masc	fem	both
cat	NLLB	unsp	49.13	28.14	49.14
	Llama	unsp	52.86	31.08	53.56
		masc	**53.36**	30.59	53.52
		fem	33.07	**58.56**	**62.44**
ces	**NLLB**	unsp	**25.41**	**24.32**	**26.05**
	Llama	unsp	24.74	23.53	26.00
		masc	23.85	21.11	24.44
		fem	20.23	23.88	24.38
deu	NLLB	unsp	**22.40**	16.05	**22.63**
	Llama	unsp	21.03	14.24	21.35
		masc	22.04	15.74	22.29
		fem	20.37	**16.88**	22.20
fra	NLLB	unsp	57.79	45.47	57.90
	Llama	unsp	**61.56**	50.47	**61.78**
		masc	56.69	45.44	56.77
		fem	49.68	**51.76**	56.99
ita	NLLB	unsp	38.87	24.37	38.38
	Llama	unsp	41.88	29.39	42.99
		masc	**42.16**	29.03	43.10
		fem	26.74	**44.86**	**45.68**

			Reference		
Language	Model	Type	masc	fem	both
ron	NLLB	unsp	28.61	24.23	30.47
	Llama	unsp	35.04	29.38	37.39
		masc	**36.85**	29.89	**38.62**
		fem	26.27	**35.96**	37.47
rus	NLLB	unsp	36.48	31.75	36.78
	Llama	unsp	40.71	35.80	40.71
		masc	**41.81**	36.88	**41.80**
		fem	35.72	**36.67**	39.12
slv	NLLB	unsp	34.53	22.66	35.51
	Llama	unsp	**37.55**	24.58	**38.57**
		masc	37.07	23.26	37.66
		fem	33.07	**27.98**	38.17
spa	NLLB	unsp	**67.46**	41.00	**66.87**
	Llama	unsp	58.72	39.83	59.56
		masc	59.94	39.13	60.50
		fem	41.42	**59.36**	62.98
avg	NLLB	unsp	40.07	28.67	40.41
	Llama	unsp	41.57	30.92	42.43
		masc	**41.63**	30.12	42.08
		fem	31.84	**39.55**	**43.37**

Table 7: BLEU scores on MULTILINGUALHOLISTICBIAS with masculine, feminine, and both references.

			Reference		
Language	Model	Type	masc	fem	both
cat	NLLB	unsp	68.76	57.33	68.85
		unsp	71.08	59.62	71.40
	Llama	masc	**71.24**	59.41	71.44
		fem	62.11	**72.81**	**72.98**
ces	**NLLB**	unsp	**50.21**	**48.72**	**50.54**
		unsp	49.68	47.95	50.15
	Llama	masc	48.44	46.09	48.60
		fem	47.28	47.83	48.86
deu	NLLB	unsp	50.14	43.45	50.25
		unsp	50.17	43.37	50.30
	Llama	masc	**51.63**	44.88	**51.77**
		fem	50.65	**46.16**	51.08
fra	NLLB	unsp	69.68	65.81	69.79
		unsp	**76.77**	**72.81**	**76.85**
	Llama	masc	73.63	69.64	73.66
		fem	71.77	71.95	73.68
ita	NLLB	unsp	62.34	53.45	62.65
		unsp	**65.55**	57.44	66.17
	Llama	masc	64.76	56.55	65.29
		fem	55.70	**66.39**	**66.71**
ron	NLLB	unsp	61.24	57.88	61.60
		unsp	63.98	60.50	64.51
	Llama	masc	**64.82**	61.14	**65.22**
		fem	61.27	**63.75**	64.56
rus	NLLB	unsp	55.58	50.59	55.78
		unsp	58.32	**53.07**	58.43
	Llama	masc	**58.94**	53.66	**59.06**
		fem	53.53	52.83	55.79
slv	NLLB	unsp	56.80	51.33	**57.35**
		unsp	**57.01**	50.88	57.33
	Llama	masc	56.66	50.37	56.88
		fem	54.81	**51.93**	55.80
spa	NLLB	unsp	**79.81**	68.44	**79.84**
		unsp	76.36	65.66	76.61
	Llama	masc	77.21	66.03	77.33
		fem	67.91	**75.55**	77.26
avg	NLLB	unsp	61.62	55.22	61.85
		unsp	**63.21**	56.81	**63.53**
	Llama	masc	63.04	56.42	63.25
		fem	58.34	**61.02**	62.97

Table 8: chrF scores on MULTILINGUALHOLISTICBIAS with masculine, feminine, and both references.

			Reference		
Language	Model	Type	masc	fem	both
cat	NLLB	unsp	0.87	0.85	-
		unsp	0.88	0.86	-
	Llama	masc	**0.89**	0.87	-
		fem	0.86	**0.88**	-
ces	NLLB	unsp	**0.88**	0.86	-
		unsp	**0.88**	**0.87**	-
	Llama	masc	**0.88**	0.86	-
		fem	0.84	0.84	-
deu	NLLB	unsp	**0.72**	**0.71**	-
		unsp	**0.72**	0.70	-
	Llama	masc	**0.72**	0.71	-
		fem	0.71	**0.71**	-
fra	NLLB	unsp	0.87	0.85	-
		unsp	**0.89**	**0.88**	-
	Llama	masc	0.88	0.87	-
		fem	0.87	0.87	-
ita	NLLB	unsp	0.86	0.82	-
		unsp	**0.88**	0.84	-
	Llama	masc	**0.88**	0.84	-
		fem	0.83	**0.85**	-
ron	NLLB	unsp	**0.89**	0.87	-
		unsp	**0.89**	0.87	-
	Llama	masc	**0.89**	0.87	-
		fem	0.86	**0.88**	-
rus	NLLB	unsp	0.88	0.87	-
		unsp	0.88	0.86	-
	Llama	masc	**0.89**	0.87	-
		fem	0.86	**0.88**	-
slv	NLLB	unsp	**0.85**	**0.84**	-
		unsp	**0.85**	0.83	-
	Llama	masc	**0.85**	0.83	-
		fem	0.81	0.82	-
spa	NLLB	unsp	**0.91**	0.88	-
		unsp	**0.91**	0.88	-
	Llama	masc	**0.91**	0.88	-
		fem	0.88	**0.90**	-
avg	NLLB	unsp	0.86	0.84	-
		unsp	0.86	0.84	-
	Llama	masc	**0.87**	0.84	-
		fem	0.84	**0.85**	-

Table 9: COMET scores on MULTILINGUALHOLISTICBIAS with masculine, feminine, and both references.

Language	Model	Type	Reference masc	Reference fem	Reference both
cat	NLLB	unsp	0.83	0.77	-
		unsp	0.84	0.78	-
	Llama	masc	**0.85**	0.79	-
		fem	0.77	**0.82**	-
ces	NLLB	unsp	**0.81**	**0.80**	-
		unsp	**0.81**	**0.80**	-
	Llama	masc	**0.81**	0.78	-
		fem	0.76	0.79	-
deu	NLLB	unsp	**0.54**	**0.53**	-
		unsp	**0.54**	**0.53**	-
	Llama	masc	**0.54**	0.53	-
		fem	0.52	0.52	-
fra	NLLB	unsp	0.77	0.75	-
		unsp	**0.80**	**0.78**	-
	Llama	masc	0.78	0.76	-
		fem	0.76	0.76	-
ita	NLLB	unsp	0.79	0.76	-
		unsp	**0.81**	0.78	-
	Llama	masc	**0.81**	0.78	-
		fem	0.76	**0.81**	-
ron	NLLB	unsp	0.80	0.79	-
		unsp	0.82	**0.81**	-
	Llama	masc	**0.83**	0.81	-
		fem	0.77	0.80	-
rus	NLLB	unsp	0.77	**0.76**	-
		unsp	**0.78**	**0.76**	-
	Llama	masc	0.78	0.77	-
		fem	0.73	0.74	-
slv	NLLB	unsp	0.76	**0.76**	-
		unsp	0.77	0.75	-
	Llama	masc	0.77	0.76	-
		fem	0.73	**0.76**	-
spa	NLLB	unsp	0.85	0.79	-
		unsp	0.85	0.80	-
	Llama	masc	**0.86**	0.80	-
		fem	0.80	**0.84**	-
avg	NLLB	unsp	0.77	0.75	-
		unsp	**0.78**	0.75	-
	Llama	masc	**0.78**	0.75	-
		fem	0.73	**0.76**	-

Table 10: BLEURT scores on MULTILINGUALHOLISTICBIAS with masculine, feminine, and both references.

Language	Model	Type	Reference masc	Reference fem	Reference both
cat	NLLB	unsp	4.32	4.27	-
		unsp	4.35	**4.30**	-
	Llama	masc	**4.36**	4.30	-
		fem	4.27	**4.30**	-
ces	**NLLB**	unsp	**4.31**	**4.27**	-
		unsp	4.24	4.20	-
	Llama	masc	4.24	4.20	-
		fem	4.20	4.18	-
deu	**NLLB**	unsp	**4.15**	**4.11**	-
		unsp	4.14	4.10	-
	Llama	masc	4.14	4.10	-
		fem	4.11	4.08	-
fra	NLLB	unsp	4.44	4.41	-
		unsp	**4.48**	**4.45**	-
	Llama	masc	**4.48**	4.10	-
		fem	4.11	4.08	-
ita	NLLB	unsp	4.46	4.39	-
		unsp	**4.48**	**4.42**	-
	Llama	masc	**4.48**	4.41	-
		fem	4.35	4.38	-
ron	**NLLB**	unsp	**4.38**	**4.34**	-
		unsp	4.35	4.30	-
	Llama	masc	4.34	4.29	-
		fem	4.28	4.28	-
rus	**NLLB**	unsp	**4.47**	**4.43**	-
		unsp	4.33	4.30	-
	Llama	masc	4.39	4.35	-
		fem	4.29	4.28	-
slv	**NLLB**	unsp	**4.14**	**4.08**	-
		unsp	4.08	4.02	-
	Llama	masc	4.08	4.01	-
		fem	4.04	4.01	-
spa	NLLB	unsp	**4.56**	**4.47**	-
		unsp	4.53	4.45	-
	Llama	masc	**4.56**	4.48	-
		fem	4.43	4.46	-
avg	**NLLB**	unsp	**4.36**	**4.31**	-
		unsp	4.33	4.28	-
	Llama	masc	4.34	4.25	-
		fem	4.23	4.22	-

Table 11: BLASER scores on MULTILINGUALHOLISTICBIAS with masculine, feminine, and both references.