Evaluating Gender Bias in the Translation of Gender-Neutral Educational Professions from English to Gendered Languages

Anonymous ACL submission

Abstract

This study evaluates the translation of gender-neutral English words to the gendered languages German, French and Italian, using 5 machine translation (MT) models: GPT-3.5 Turbo, LLaMA 2, AWS, SYS, and Google. Focusing on translating educational professions, each model's output was categorized into four gender classifications: unknown (UNK), female (f), male (m), and neutral (n). Error rates were determined through human validation, involving manual review of randomly sampled records. Our findings reveal significant gender bias across all tested MT systems, with a notable over representation of male gender classifications.

1 Introduction

Machine translation (MT) is a dynamic field that automates the conversion of text from one natural language to another (Jurafsky and Martin, 2020). Its applications span diverse domains, including communication (Koehn, 2009), information dissemination (Östling and Tiedemann, 2017), entertainment (Fernández-Martínez et al., 2017), and education (Breen, 2019). Machine translation, however, is far from neutral; it is a complex interplay of factors. The choice of translation model (Han et al., 2021), the richness of training data (Prates et al., 2018), and the evaluation criteria (Han, 2018) all shape the outcome of a translation. These intricacies introduce bias, impacting the accuracy, clarity, and suitability of the translated content. Especially, for human-centered applications of Natural Language Processing such as educational writing support, these challenges might have serious consequences. Better understanding and preventing the prevalence of algorithmic bias, and gender bias in particular, is therefore an important quest in education-relevant domains. Bias in machine translation (MT) refers to deviations from the original meaning or intention of a source text, often influenced by translator preferences, assumptions, or stereotypes (Savoldi et al., 2021a). These deviations can take various forms, including mistranslation, omission, addition, distortion, or polarization of the source text content (Savoldi et al., 2021a). Such biases can have adverse effects, particularly in domains like education, where accuracy and objectivity are crucial (Savoldi et al., 2021a).

In the realm of educational settings, particularly concerning machine translation (MT) and digital tools, combating bias holds paramount significance. Cotelli Kureth et al. (2023) brings attention to key aspects. First, the misuse of machine translation as Online Bilingual Dictionaries is prevalent among language learners, who often resort to it for single-word translations, neglecting broader context and risking inaccuracies in their language acquisition journey. Second, the narrow focus on word-level translation perpetuates this misuse, emphasizing the need to address metalinguistic awareness gaps among MT users. Third, with the wide spread use of Large Language Models through conversational interfaces such as ChatGPT, students increasingly rely on these foundational models to receive tutoring or feedback on pedagogical exercises. Hence, the translation between natural languages seem to play a crucial role when it comes to the learner-centered application of Large Language Models.

Bias in MT has been studied from different perspectives and in different domains. Some studies have focused on specific types of bias, such as gender bias (Saunders and Byrne, 2020), Stanovsky et al. (2019a), racial bias (Font and Costa-jussa, 2019), or political bias (Hovy and Spruit, 2016). Other studies have explored the sources and causes of bias in MT, such as the translation model (Han et al., 2021), the training data (Prates et al., 2018), or the evaluation metrics (Han, 2018), (Freitag et al., 2020). Despite the growing attention to bias in MT, there remains a significant gap in the literature concerning the examination of bias specifically in the translation of educational texts.

This study aims to investigate gender bias in machine translation (MT) by analyzing the gender classification performance of five prominent MT models: GPT-3.5 turbo, LLaMA 2, AWS, SYS, and Google. By translating educational profession-related texts into German, French, and Italian, and categorizing the results into unknown (UNK), female (f), male (m), and neutral (n) genders, we seek to identify patterns of bias and inaccuracies.

2 Related work

2.1 Bias Detection and Mitigation in Machine translation (MT)

Bias detection and mitigation in machine translation (MT) involve identifying and correcting unfair or skewed elements in machine-generated translations. These bias can include factors such as gender, race, culture, or other social aspects. The goal is to ensure that translations are impartial and do not exhibit favoritism toward any particular group.

Various methods and frameworks have been proposed to identify, quantify, and mitigate bias in MT systems across different dimensions and levels. One pioneering study by Vanmassenhove et al. (2018) addressed gender bias in neural machine translation (NMT) systems. The study found that translations often reinforce gender stereotypes and proposed a data augmentation technique to enhance gender accuracy without compromising quality. Similarly, Stanovsky et al. (2019a) focused on evaluating gender bias in machine translation systems. They introduced a new dataset and methodology to measure how often and to what extent machine translation systems produce gender-biased outputs. Behnke et al. (2022) investigated bias in NMT quality estimation (QE), highlighting partial input bias and proposing informative and adversarial approaches for bias mitigation, thereby improving QE performance. In a sociolinguistic-aware framework introduced by Hall-Lew et al. (2021), considerations for accommodating social and linguistic variation in MT were discussed, suggesting methods such as metadata utilization and interactive techniques to address bias and enhance translation appropriateness.

While these studies provide valuable insights into bias detection and mitigation in MT, there is currently no specific research examining bias in text translation within the education domain. This gap includes understanding how translations respond to non-gender stereotypes in sentences involving educational professions or roles. By investigating how translation models handle gender-neutral language in educational contexts, this study aims to explore and address biases in educational translations.

2.2 Machine Translation in Education

Despite significant advancements in Machine Translation (MT), largely driven by neural network models and extensive parallel corpora (Jurafsky and Martin, 2020; Vanmassenhove, 2024), concerns regarding biases that affect translation quality and fairness remain persistent. Gender bias, in particular, is frequently highlighted as a significant issue in MT (Savoldi et al., 2021b). The complexity of algorithmic bias and linguistic diversity underscores the urgent need for comprehensive evaluation and mitigation strategies (Vanmassenhove et al., 2021).

Bias in MT can originate from various sources, including data, models, and evaluation methodologies (Behnke et al., 2022; Vanmassenhove, 2024; Vanmassenhove et al., 2021). Data bias arises from imbalances or deficiencies in the training datasets, while model bias is attributed to the inherent limitations or assumptions within MT models. These biases can negatively impact translation quality by under-representing certain languages or domains, thereby restricting adaptability across different contexts.

In the field of education, MT is a valuable tool for language pedagogy, facilitating communication between international students and educators and enhancing global accessibility to educational resources (Abimbola, 2023; High, 2023). Particularly in the realm of foreign language teaching and learning, Neural Machine Translation (NMT) shows promise for improving language skills, despite ongoing concerns about translation accuracy and quality (Urlaub and Dessein, 2022; Macketanz et al., 2018).

2.3 Bias in Education technology

Research on bias in educational technology dates back to the 1960s, and many contemporary studies on algorithmic bias and fairness build on these early foundations (Baker and Hawn, 2021). To effectively investigate bias, it is crucial to establish a clear perspective, as the term "bias" encompasses a range of definitions across different research domains (Hutchinson and Mitchell, 2019; Baker and

Hawn, 2021). In our study, we define algorithmic bias as "situations where model performance is substantially better or worse across mutually exclusive groups" (Baker and Hawn, 2021, p. 4). Such bias can lead to errors, misuse, and unfair outcomes, either directly or indirectly. Since biases are not explicitly encoded or stated, their presence and harmful effects are often challenging to detect.

One context where bias can manifest and impact users is in translation engines. Students in language education frequently use translation tools like Google Translate and DeepL. Although Groves and Mundt (2015) found that these tools provide comprehensible and sometimes impressive translations of students' texts, few studies have examined the biases introduced by translation engines. Our research aims to investigate the biases present in these tools, focusing specifically on gender bias in educational downstream tasks, as suggested by Lee et al. (2022).

3 Methodology

Our aim is to quantify the use of male and female forms in machine translations of gender-neutral sentences from English. For example, "the professor" can be translated as "der Professor" to refer to a male person, or "die Professorin" when referring to a female person.

In this study, we evaluated the gender translation performance of five different translation models: GPT-3.5 turbo, LLaMA 2, AWS, SYS, and Google. The analysis was conducted across three languages: German (de), French (fr), and Italian (it). Each model was tasked with translating a dataset containing education professions and the translated text underwent gender classifications into four categories: unknown (UNK), female (f), male (m), and neutral (n). For each translation model, we compiled occurrences of each gender category. The counts and percentages of each gender were recorded, as well as the error rates for gender classification. Error rates were determined based on human-annotated samples, which involved manually checking the gender classification accuracy for a subset of the records.

3.1 Dataset

To collect real-world data, we utilized the Cornell Conversational Analysis Toolkit (ConvoKit)[1]. Our objective was to construct a dataset of gender-neutral sentences containing a single human entity. We employed a meticulously tuned Named Entity Recognition (NER) model to identify human entities associated with professions in the educational sector. Education roles and professions considered include teacher, tutor, coach, mentor, and instructor (Wikipedia contributors, 2024). A comprehensive list of some educational professions can be found in Table 4 in the appendix A. Sentences with more than one professional title or any gender-specific terms were excluded. The resulting dataset comprises naturally occurring sentences, free from templated constructs, ensuring a diverse array of sentence structures. Examples include "a teacher," "my teacher," and "the teacher," which guarantees the inclusion of varied linguistic patterns, enhancing the dataset's applicability and robustness. A sample of the sentences include are shown in Table 1.

3.2 Experimental Setup

MT systems We test five widely used MT models Amazon Translate [2] Google Translate [3] SYSTRAN [4] GPT-3.5 Turbo [5], LLaMA 2 [6], representing the state of the art in both commercial and academic research to translate the collected data into three languages: German (de), French (fr), and Italian (it). Following experiments involving four different languages, we utilized the multilingual variant of LLaMA 2, which was pretrained and fine-tuned specifically for multilingual tasks (Tang et al., 2020). We selected three languages with grammatical gender that exhibit a wide range of other linguistic properties (e.g., alphabet, word order, grammar), while still allowing for highly accurate automatic morphological analysis. These languages belong to different language families: Romance languages (French, Italian) and Germanic languages (German), all of which have gendered noun-determiner agreement.

We then use the GPT-3.5 Turbo API by OpenAI (OpenAI, 2024) to identify matching educational professional words between the original and translated sentences to determine the gender of the translated word, thereby labeling the translated text

[1] https://convokit.cornell.edu/documentation/subreddit.html

[2] https://aws.amazon.com/translate

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

[4] http://www.systransoft.com

[5] https://openai.com/api/

[6] https://huggingface.co/SnypzZz/Llama2-13b-Language-translate

Sentences
I had a conversation recently with a superintendent of a public school.
My professor was killed over a year and a half ago.
I should note that I intend to become a professor.
I work as an instructional coach.
I am an aspiring educator, have been reading up on different educational philosophies.
Professional and qualified home tutor visits home on flexible time period and gives simple learning process based on advanced research.
I've always thought about becoming a professor.
What are some good ways to become a professor?
I want to know what my fellow Redditors think.
I am currently an administrator and director of programs for a school district.
I only want to look into any real science behind grading philosophies, that I might leave my professor with something to think about going forward.
How can I voice my concern to my professor politely and constructively?
UNC President Erskine Bowles has announced that Willie J. Gilchrist, superintendent of Halifax County Schools since 1994.
So, what would my fellow education professionals do in my situation regarding graduate programs?
Some background quickly: A friend of mine is an adjunct professor at a prestigious, well known university.
A professor's success is defined by research.
I have applied to grad school to get my credential to become a school counselor.
Willie Gilchrist was Elizabeth City State University Chancellor.
State law requires school boards to appoint a "schools transportation safety director."

Table 1: Sample of sentences contained in the dataset

according to the gender of the translated profession. We initially tried using the alignment method developed by (Dyer et al., 2013) used by (Stanovsky et al., 2019b) to determine the gender of the translated professions detected through morphological tagging. However, this approach did not yield satisfactory results due to numerous misaligned professions, resulting in significant errors in gender determination. The primary reason for this failure was the complexity and high variability of the generated text, which contrasted with the more straightforward nature of templated text typically used in such alignments.

3.3 Classification and Quantitative Analysis

The classification process involved parsing the output of each translation model and categorizing the detected genders into four predefined categories: unknown (UNK), female (f), male (m), and neutral (n). Each set of sentence pairs includes the first sentence in English and the second the translated text in the target language (German, French, or Italian). The prompt given to the model to label the translated text were: 1) Identify the one pair of human-occupational nouns related to the education field in these two sentences, and 2) Distinguish the gender of the human-occupational noun in the translated language, marking them with 'm', 'f', or 'n'. If the model was unable to find the pair or distinguish the gender, it returned 'UNK' for all results.

For each translation model, the occurrences of each gender category were tallied and converted into percentages to allow for comparative analysis across models and languages. The data were analyzed to determine the total count of occurrences for each gender category (UNK, f, m, n), the percentage of total classifications for each gender category within each language, and the error rates for gender classification, calculated based on a human-annotated sample. This systematic approach enabled a clear comparison of how different models translated non gendered text in various languages.

The quantitative analysis focused on identifying patterns and discrepancies in gender classification across different models and languages. The high occurrence of male gender classifications and the variability in the unknown and female categories were of particular interest. Additionally, the representation of neutral gender classifications was assessed, given its consistently low occurrence across all models and languages. This analysis provided insights into how each model performed in handling translation into gendered languages.

3.4 Human Validation

To estimate the accuracy of our gender bias evaluation method, we conducted a human validation process. This involved selecting a random sample of 50 male and 50 female records from most of the model's output in each language except in the case of GPT-3.5 turbo french where the total was 92 due to the limited number of label 'f'. One human annotator reviewed these records to verify the gender classifications, with the error rate calculated as the proportion of misclassified records. Errors were categorized into issues such as assigning gender to non-human entities, ambiguity with the word "fellow," unclear gender from sentences, contextually indeterminate gender, and incorrect alignment of translations. This validation process was crucial for identifying and quantifying the limitations and biases in each model, particularly in handling gender-related data, and informed improvements for future data processing.

4 Results

Our primary findings are summarized in Tables 2 and 3. We evaluated five different machine translation (MT) systems (GPT-3.5 turbo, LLaMA 2, AWS, SYS, and Google) across three languages (German, French, and Italian), calculating the proportions of unknown (UNK), female (f), male (m), and neutral (n) labels, along with their estimated errors. This analysis assessed the effectiveness of each system in conveying the correct gender in the target language. The results indicate that all tested MT systems exhibit gender bias in clearly preferring the male translation over the female one.

5 Discussion

The analysis of gender occurrences and their respective percentages across various language models highlights several key trends and discrepancies as shown in the results 4. Firstly, the dominance of the male gender ('m') across all models and languages is evident, with occurrences typically exceeding 90% of the total. This trend underscores a significant bias towards male gender representation in the text outputs of these language models.

Secondly, the unknown gender category ('UNK') shows considerable variability across models and languages. For instance, GPT-3.5 has relatively low

Table 2: Gender Occurrences and Percentages for AWS, SYS, and Google (* The error rate was calculated by a human-annotated sample of 50 male and 50 female records (see section 3.4 for details))

Model	Language	Gender	Count	Percentage (%)	Estimated Error Rate (%)*
AWS	de	UNK	122	2.95	42.00
		f	156	3.78	
		m	**3850**	**93.79**	
		n	15	0.37	
	fr	UNK	79	1.96	9.00
		f	235	5.88	
		m	**3808**	**95.16**	
		n	21	0.52	
	it	UNK	235	5.81	-
		f	80	1.98	
		m	**3807**	**94.88**	
		n	21	0.52	
SYS	de	UNK	122	3.05	23.00
		f	230	5.75	
		m	**3771**	**94.01**	
		n	20	0.50	
	fr	UNK	69	1.72	7.00
		f	236	5.88	
		m	**3818**	**95.08**	
		n	20	0.50	
	it	UNK	217	5.45	-
		f	148	3.71	
		m	**3763**	**94.32**	
		n	15	0.38	
Google	de	UNK	158	3.96	19.00
		f	111	2.78	
		m	**3851**	**96.13**	
		n	23	0.57	
	fr	UNK	71	1.76	12.00
		f	234	5.78	
		m	**3815**	**94.76**	
		n	23	0.57	
	it	UNK	240	6.09	-
		f	24	0.61	
		m	**3855**	**97.38**	
		n	24	0.61	

Table 3: Gender Occurrences and Percentages for GPT-3.5 turbo and LLaMA 2 Models (* The error rate was calculated by a human-annotated sample of 50 male and 50 female records (see section 3.4 for details))

Model	Language	Gender	Count	Percentage (%)	Estimated Error Rate (%)*
GPT-3.5 Turbo	de	UNK	52	**1.28**	26.00
		f	135	**3.34**	
		m	**3904**	**96.09**	
		n	52	**1.28**	
	fr	UNK	34	**0.84**	22.83
		f	42	**1.04**	
		m	**4035**	**98.88**	
		n	28	**0.68**	
	it	UNK	102	**2.55**	-
		f	86	**2.15**	
		m	**3942**	**98.13**	
		n	9	**0.22**	
LLaMA 2	de	UNK	192	**4.79**	24.00
		f	124	**3.09**	
		m	**3796**	**94.59**	
		n	31	**0.77**	
	fr	UNK	89	**2.21**	8.00
		f	203	**5.04**	
		m	**3814**	**94.52**	
		n	37	**0.92**	
	it	UNK	88	**2.18**	-
		f	208	**5.16**	
		m	**3785**	**93.89**	
		n	62	**1.54**	

occurrences of unknown gender in French (0.84%) and higher in Italian (2.55%). LLaMA 2 shows the highest variability, with German having 4.79%. This inconsistency suggests that models handle or classify uncertain gender information differently, which might depend on the training data or the inherent biases of each model. This variability indicates the need for standardizing how gender-unknown cases are managed across different models.

The occurrences and percentages of the female gender ('f') vary significantly among the models. For example, in the French language, AWS (5.88%) and SYS (5.88%) have higher female gender representation compared to GPT-3.5 (1.04%). Similarly, LLaMA 2 shows higher female occurrences in Italian (5.16%) compared to GPT-3.5 (2.15%). These variations indicate differences in how models were trained and the datasets used, affecting the classification and representation of female gender. Such discrepancies highlight the presence of gender bias in translation.

We did not manually verify the neutral classification, so it should be noted that there might be instances where it is not possible to phrase something in a gender-neutral manner. So, as seen in the results 4 , the neutral gender ('n') has the lowest occurrences across all models and languages, with percentages typically below 1%. The highest percentage observed is in LLaMA 2 for Italian (1.54%). This minimal representation suggests that neutral gender is either underrepresented in the training data or not well handled by current models, indicating an area for improvement to ensure inclusivity. Addressing this issue could involve incorporating more neutral-gendered data into the training sets and refining the models to better recognize and classify neutral gender.

Error rates were calculated by human-annotated samples of 50 male and 50 female records. These rates also vary among models, with AWS showing the highest error rate in German (42.00%) and the lowest in French (9.00%). These error rates indicate the proportion of misclassified or ambiguous gender occurrences, reflecting the models' performance and reliability in gender classification.

Lastly, model-specific trends are notable. GPT-3.5 generally shows lower female and unknown gender occurrences compared to other models. LLaMA 2 and SYS tend to have higher female gender representation, especially in French. AWS and Google show a similar pattern with relatively high male gender occurrences but differ in their handling of unknown and female genders, particularly in Italian. These differences underscore the importance of continuous evaluation and improvement of translation models.

6 Conclusion

Our analysis revealed a prevalent gender bias across all tested models, with a significant overrepresentation of male gender classifications. The evaluation process, including human validation, highlighted various sources of errors such as non-human entity gender assignments, ambiguous terms like "tutor," and unclear context.

Despite the insights gained, the study faced limitations such as the restricted language scope, potential biases in human annotation, and the specific, limited dataset sourced from Reddit. These findings underscore the need for more comprehensive and clean datasets , broader language inclusion, and continuous updates to model evaluations to better understand and mitigate gender bias in machine translation systems.

Future research should address these limitations by incorporating a wider variety of languages, expanding the range of evaluated contexts, and leveraging the latest advancements in translation technology to enhance the accuracy and fairness of gender representation in machine translations.

Limitations

While our study provides valuable insights into the gender translation performance of various machine translation models, several limitations should be considered.

One limitation of this study is the inherent bias present in the training datasets of the evaluated machine translation models. The task-specific text set was collected from Reddit, which is limited in quantity and varied in context. This dataset includes educational terms that may appear in non-human contexts, such as "tutor" which is both a noun and a verb and "fellow," affecting their overall translation accuracy. Consequently, the variability and context-specific nature of the data could introduce inconsistencies and biases in gender classification, impacting the study's results.

Another limitation is the variability in the handling of unknown (UNK) and neutral (n) gender categories across different models and languages. The models showed inconsistency in classifying gender when the context was ambiguous or when the words did not explicitly indicate gender. This inconsistency highlights the need for more sophisticated algorithms that can better manage gender-neutral and ambiguous contexts.

Additionally, the error rates determined from human validation reveal that certain types of errors, such as misalignment of translations and ambiguity in the source text, are prevalent. These errors indicate that the models sometimes struggle with accurately aligning gender-specific words between the source and target languages, particularly when the gender is not clearly defined or contextually apparent.

Lastly, the study focused on only three languages with grammatical gender (German, French, and Italian), which limits the generalizability of the findings. Other languages with different grammatical structures and gender rules might present different challenges and outcomes, and further research is needed to evaluate the performance of these models across a broader range of languages.

References

Lawal Abimbola. 2023. The impact of machine translation on vocabulary acquisition and reading comprehension in esl learners. *Pulchra Lingua: A Journal of Language Study, Literature & Linguistics*, 2:80–93.

Ryan S Baker and Aaron Hawn. 2021. Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, pages 1–41.

Hanna Behnke, Marina Fomicheva, and Lucia Specia. 2022. Bias mitigation in machine translation quality estimation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1475–1487, Dublin, Ireland. Association for Computational Linguistics.

Mark Breen. 2019. *Advances in Machine Translation: Contemporary Approaches in Education*, pages 146–169. IGI Global.

Sara Cotelli Kureth, Alice Delorme Benites, Mara Haller, Hasti Noghrechi, and Elizabeth Steele. 2023. *"I Looked It Up in DeepL": Machine Translation and Digital Tools in the Language Classroom*.

Chris Dyer, Lili He, Marie-Catherine de Marneffe, and Noah A. Smith. 2013. A simple, fast, and effective re-implementation of Fast Align. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*. Association for Computational Linguistics.

Marcos Fernández-Martínez, Antonio Toral, and Raquel Martín-Morales. 2017. *Automatic Translation in the Entertainment Industry: How Online Subtitling Is Transforming Audiovisual Translation*, pages 167–190. Springer, Cham.

Joel Font and Marta Costa-jussa. 2019. Equalizing gender bias in neural machine translation with word embeddings techniques. pages 147–154.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online. Association for Computational Linguistics.

Michael Groves and Klaus Mundt. 2015. Friend or foe? google translate in language for academic purposes. *English for Specific Purposes*, 37:112–121.

Lauren Hall-Lew, Emma Moore, and Robert J. Podesva. 2021. *Social meaning and linguistic variation: Theoretical foundations*, pages 1–24. Cambridge University Press, United States.

Lifeng Han. 2018. Machine translation evaluation resources and methods: A survey. *Preprint*, arXiv:1605.04515.

Lifeng Han, Alan Smeaton, and Gareth Jones. 2021. Translation quality assessment: A brief survey on manual and automatic methods. In *Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital Age*, pages 15–33, online. Association for Computational Linguistics.

Michael David High. 2023. *The Perils and Potential Benefits of Machine Translation in Transnational Higher Education*, pages 115–135. IGI Global.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Ben Hutchinson and Margaret Mitchell. 2019. 50 years of test (un)fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 49–58, New York, NY, USA. Association for Computing Machinery.

Dan Jurafsky and James H. Martin. 2020. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Pearson.

Philipp Koehn. 2009. *Statistical Machine Translation*.

Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, et al. 2022. Evaluating human-language model interaction. *arXiv preprint arXiv:2212.09746*.

Vivien Macketanz, Aljoscha Burchardt, and Hans Uszkoreit. 2018. Tq-autotest : Novel analytical quality measure confirms that deepl is better than google translate by.

OpenAI. 2024. Gpt-3.5 turbo. Accessed: 2024-06-13.

Robert Östling and Jörg Tiedemann. 2017. Neural machine translation for low-resource languages. *CoRR*, abs/1708.05729.

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.

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.

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

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

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019a. 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.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019b. Evaluating gender bias in machine translation. *CoRR*, abs/1906.00591.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.

Per Urlaub and Eva Dessein. 2022. Machine translation and foreign language education. *Frontiers in Artificial Intelligence*, 5:936111.

Eva Vanmassenhove. 2024. Gender bias in machine translation and the era of large language models. *ArXiv*, abs/2401.10016.

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

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213, Online. Association for Computational Linguistics.

Wikipedia contributors. 2024. Category: Education and training occupations. `https://en.wikipedia.org/wiki/Category:Education_and_training_occupations`. Accessed: 2024-06-13.

A Appendix A: Additional Results

Figures 1 and 2 show the results of the gender occurrences for the different translation engines.

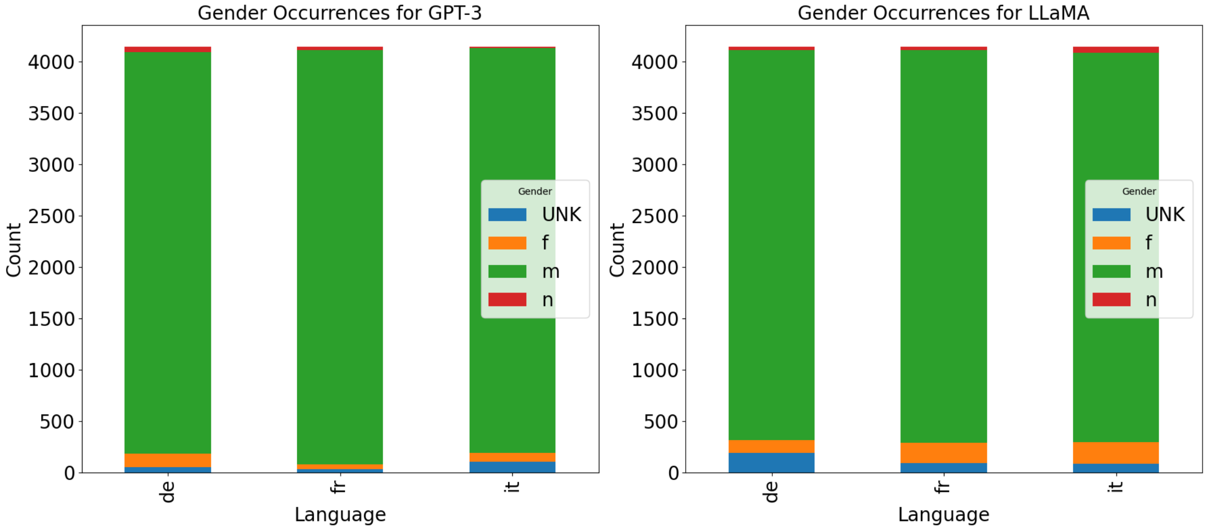

Figure 1: Gender Occurrences for GPT-3.5 turbo and LLaMA 2 Models

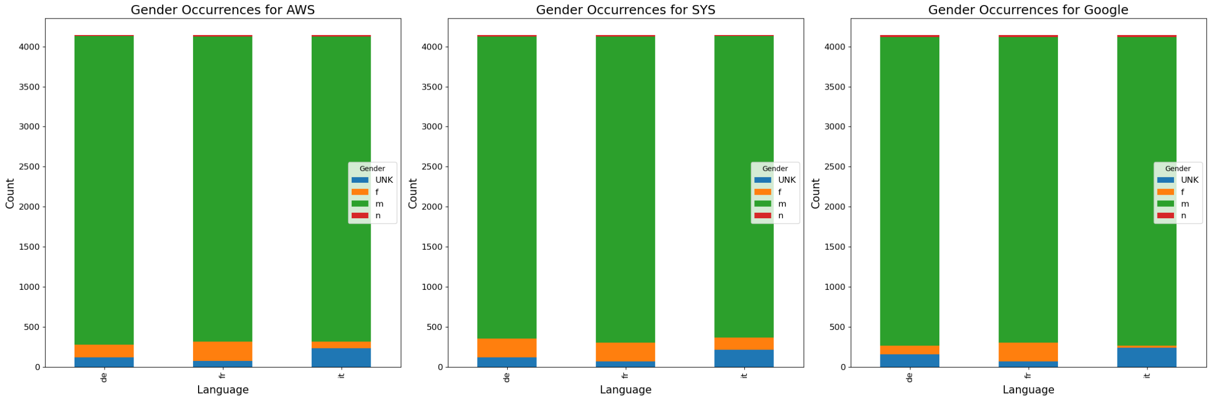

Figure 2: Gender Occurrences for Google, SYSTRAN and AWS Models

Column 1	**Column 2**	**Column 3**	**Column 4**	**Column 5**
teacher	tutor	coach	mentor	instructor
professor	lecturer	counselor	principal	dean
provost	librarian	curator	educator	trainer
superintendent	regent	director	chancellor	bursar
fellow	student	learner	administrator	researcher
curriculum-developer	educational-psychologist	education-consultant	school-psychologist	special-education-teacher
school-nurse	academic-advisor	educational-technologist	assistant-language-teacher	foreign-language-assistant
bear-leader	deputy-head-teacher	employment-counsellor	exam-invigilator	global-career-development-facilitator

Table 4: List of Educational Professions (Wikipedia contributors, 2024)