Did I (she) or I (he) buy this? Or rather I (she/he)? Towards first-person gender neutral translation by LLMs

Maja Popović[1,2], **Ekaterina Lapshinova-Koltunski**[3], **Anastasiia Göldner**[3]
[1] ADAPT Centre, Dublin City University, Ireland
[2] IU University, Berlin, Germany
maja.popovic@adaptcentre.ie, maja.popovic@iu.org
[3] Language and Information Sciences, University of Hildesheim, Germany
lapshinovakoltun@uni-hildesheim.de

Abstract

This paper presents an analysis of gender in first-person mentions translated from English into two Slavic languages with the help of three large language models (LLMs) and two different prompts. We explore if LLMs are able to generate Amazon product reviews with gender neutral first person forms. Apart from the overall question about the ability to produce gender neutral translations, we look into the impact of a prompt with a specific instruction which is supposed to reduce the gender bias in LLMs output translations. Our results show that although we are able to achieve a reduction in gender bias, our specific prompt cause also a number of errors. Analysing those emerging problems qualitatively, we formulate suggestions that could be helpful for the development of better prompting strategies in the future work on gender bias reduction.

1 Introduction

It is known that machine translation (MT) systems, large language models (LLM) and other natural language processing (NLP) applications are prone to bias, for instance gender bias (preference or toward one gender over the other). Gender bias exists not only in training data and embeddings, but also algorithms themselves (Zhao et al., 2018; Prost et al., 2019; Raj et al., 2024) and in human annotations (Hackenbuchner et al., 2024a,b), so that an NLP system can produce gender biased predictions.

Since LLMs perform well for translation-related tasks, we try to test if we can reduce the gender bias by asking for a gender neutral translation in the prompt. We focus on the analysis of first-person gender in different translation variants of Amazon product reviews. These variants were produced by three different large language models using two different prompts: the one approximating a holistic approach in translation, and the other with a specific instruction. The specific instruction is aimed to bring LLMs to use gender neutral forms.

We analyse the LLM translations from English into Croatian and Russian. The underlying texts are product reviews and contain particularly many words in first person singular. In contrast to English, both Slavic target languages under analysis have gender marking not only on pronouns, but also on nouns, adjectives, verbs, determiners and numbers, see Section 3.2 for more details. Therefore, the gender which is not specified in the source needs to be specified in the target. When translating user reviews, the gender form of adjectives and verb past tenses and participles and passive constructions should be specified: купил (masculine for bought) vs. купила (feminine for bought). That is why it is difficult to completely avoid the first-person gender – it can be done only in a very small number of cases. The best way to provide gender neutral text is to use an inclusive form which includes both masculine and feminine gender: купил(а) – bought (m/f).

As it was shown in our previous work (Popovic and Lapshinova-Koltunski, 2024), both human and machine translated texts contain gender bias. Moreover, machine translations tend to contain errors or inconsistencies related to these language contrasts and the need for specification in translation. The authors also reported on the usage of inclusive forms in translations generated with an LLM. Although gender bias has been thoroughly analysed in the recent machine translation studies, especially in translations with LLMs (see Section 2 below), there are still not so many works addressing the impact of prompts onto gender bias and the problems that may additionally emerge.

To the best of our knowledge, this is the first work on this type of languages addressing first person verb forms. For this reason, it is still not clear

Proceedings of the 3rd Workshop on Gender-Inclusive Translation Technologies (GITT 2025), pages 64–73, June 23, 2025

what kind of examples and more elaborated instructions for few-shot (or more complex) prompting would be the best option. Therefore, in this work we investigate two zero-shot prompts: a holistic one and one with instruction to use gender neutral form, to better understand the behaviour of the models.

Furthermore, we examine the nature of systematic errors in the outputs which should help to design better prompt strategies for the future work. An analysis of error types for different models is not in the focus of the work.

In our analysis, we focus on comparing the results of two prompt variants trying to answer the following research questions:

RQ1: Are there differences in the outputs of different LLMs regarding the choice of the first person gender?

RQ2: Can the more specific zero-shot prompt asking for a gender neutral translation decrease the bias?

RQ3: What are the problems with the zero-shot prompt and how the future prompts should be formulated to avoid it?

The remainder of the paper is organised as follows: Section 2 provides an overview of related studies. Section 3 gives details on how translations were generated and how we performed the analysis. We present the results in Section 4 and conclude providing outlook for future work directions in Section 5.

2 Related Work

As already mentioned above, gender bias has attracted attention in machine translation studies in the recent years. For instance, last editions of the EAMT conferences hosted dedicated workshop series (Savoldi et al., 2024a; Vanmassenhove et al., 2023) with many works analysing not only neural machine translation (NMT), but also LLM outputs.

Sant et al. (2024) analyse outputs of various LLMs comparing their performance with neural MT models with a focus on gender bias for the language pairs English-Spanish and English-Catalan. They state that LLMs exhibit a higher degree of bias in comparison to NMT models. Similarly to our aims, they try to eliminate the bias testing various prompting techniques testing the outputs on the WinoMT test set (Stanovsky et al., 2019). In this way, they obtain gender-bias scores for each prompt. The prompt with the most considerable reduction in bias is then evaluated on the remaining test sets. The resulting prompt with the best performance achieves a reduction by 12%.

Savoldi et al. (2024b) use three few-shot prompting strategies with the aim to produce gender-neutral translations. Their prompts include both simple templates with examples not containing verbalised instructions and chain-of-thought step-by-step templates. The authors then perform manual analysis of translations with GPT4 for the English-Italian language pair. Their fine-grained analyses demonstrate promising results in prompting generative models for less gender-biased translations.

The same language pair (English-Italian) is also in focus of gender analysis by Vanmassenhove (2024). The author analyses translations produced with ChatGPT to see how and to what extent ChatGPT can handle gender. She also tests different prompts including one with a specific instruction to produce all possible alternatives in terms of gender. The results show that ChatGPT often fails to generate feminine or binary gender translation alternatives. Besides that, the resulting translations contain various errors.

Some publications are prompting LLMs for gender inclusive forms. Nunziatini and Diego (2024) report on using LLMs for post-editing MT for the English-Spanish language pair in terms of gender inclusiveness. According to their results, GPT4 can be used for this task. Another study (Piergentili et al., 2024) uses gender-inclusive neomorphemes to improve machine translation in terms of gender inclusiveness, again for the English-Italian pair. The authors test several LLMs and various prompting strategies including zero-shot and few-shot templates. It is also interesting to test if LLMs can be prompted for producing specific gender, as it is done by Sánchez et al. (2024) who explore the ability of LLMs to produce gender-specific translations.

In general, some languages or language pairs are more prone to the problem of gender bias. Most of the existing works explore how to totally avoid gender in English, which is easier than for many other languages (Vanmassenhove et al., 2021). The problem already emerges even for closely related target languages such as German (Savoldi et al., 2023). In fact, the more categories with explicit gender marking a language has, the more problematic it is.

Most of the studies cited above deal with nouns

and noun phrases describing professions or animate subjects. There are not so many studies focusing on other categories, for instance, first person constructions. Habash et al. (2019) propose automatic generation of both gender variants for the first person in Arabic NMT translations, before the emergence of LLMs. In our previous work (Popovic and Lapshinova-Koltunski, 2024) we investigated first person gender in the same two target languages, Croatian and Russian, however the focus was predominantly on the differences between human and machine translations. Furthermore, we only included GPT3.5 language model, all other automatic translations were generated by NMT systems. For GPT3.5 translations we used the simple prompt "translate into target language" in order to provide the same instruction as to human translators. The results show that even with this simple and general prompt, GPT3.5 often generated inclusive forms with both variants, especially for the English-Croatian translations. Since some of the previous studies cited above show that prompting LLMs to generate more inclusive forms or gender-neutral forms is promising, in this work, we aim to compare the outputs of various models prompted with either simple or specific instructions for translation of the first person forms from the same English data into Croatian and Russian. We expect that a prompt formulated specifically for gender-neutral translation will result in a higher number of such gender-neutral translations when compared to the outputs of a simple prompt.

3 Methodology

3.1 Data

We use the texts from the publicly available corpus DiHuTra[1] (Lapshinova-Koltunski et al., 2022) containing 196 English Amazon product reviews (14 reviews in each of 14 different product categories) with about 15,000 running words.

The English original texts are then used to generate the Croatian and Russian translations in ChatAI (Doosthosseini et al., 2024)[2], which is a stand-alone LLM web service. For our analyses, we select three LLMS available in ChatAI: GPT4o mini[3], Llama 3.1 Nemotron 70B Instruct (Wang et al., 2024)[4] and Mistral Large Instruct 2407[5] with default settings.

We use two simple zero-shot prompts. Prompt 1 is a simple instruction to translate (a holistic approach in translation), the same one used in previous work for GPT3.5 translations, and Prompt 2 is an instruction to produce gender-neutral translation:

- Prompt 1: Translate into Croatian/Russian: TEXT
- Prompts 2: Translate into Croatian/Russian using gender neutral form for the first person: TEXT

For the sake of comparison, we also include the results of the GPT3.5 translations from our previous study (Popovic and Lapshinova-Koltunski, 2024) into our analyses. The resulting multiple translations of the 196 reviews are then manually analysed as described in 3.3 below.

3.2 Gender in target languages

Croatian and Russian are Slavic languages. Like many others from this language family they have three types of grammatical gender: masculine, feminine, and neuter. The grammatical gender of a noun affects the form of the adjectives, verbs and pronouns which agree or refer to it. The form of adjectives and verbs which agree with a first person subject is determined by the gender of the subject. Since neuter gender is never used for people, only masculine and feminine gender is possible for a writer of user reviews. Table 1 shows an example with a past participle ("received") and an adjective ("upset") agreeing with the first person subject in three forms: feminine, masculine and inclusive (containing both endings).

It should be noted that there are still no non-binary forms in the analysed target languages. Using neuter gender for people would sound awkward, and even possibly offensive, and there are still no attempts to introduce any kind of neopronouns like in some Romance languages.

Contrary to English, a "fully gender neutral" writing in these target languages is possible only very rarely, because rephrasing a text by removing all adjectives and past participles without grammatical errors or awkward styling is difficult. The

[1] http://hdl.handle.net/21.11119/0000-000A-1BA9-A

[2] https://docs.hpc.gwdg.de/services/chat-ai/index.html

[3] https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

[4] https://huggingface.co/nvidia/Llama-3.1-Nemotron-70B-Instruct

[5] https://huggingface.co/mistralai/Mistral-Large-Instruct-2407

	en	this is fake MAC, i just **received** mine and super **upset** to find out it isnt real MAc.
fem.	hr	Ovo je fejk MAC, upravo sam **dobila** svoj i jako sam **ljuta** što nije pravi MAC.
	ru	Это подделка MAC, я только что получила свою косметику и ужасно расстроена, потому что это не настоящая косметика MAC!
masc.	hr	Ovo je fejk MAC, upravo sam **dobio** svoj i jako sam **ljut** što nije pravi MAC.
	ru	Это подделка MAC, я только что получил свою косметику и ужасно расстроен, потому что это не настоящая косметика MAC!
incl.	hr	Ovo je fejk MAC, upravo sam **dobio/la** svoj i jako sam **ljut/a** što nije pravi MAC.
	ru	Это подделка MAC, я только что получил(а) свою косметику и ужасно расстроен(а), потому что это не настоящая косметика MAC!

Table 1: Example of a past participle ("received") and an adjective ("upset") agreeing with the first-person subject in three ways: red=feminine, blue=masculine, orange=inclusive.

United Nations Organisation gives some recommendations how to make gender invisible when it is not relevant for communication[6]. Existing studies describe the use of plural verb forms with first person singular pronouns or gender-gapping (the use of underscore, e.g. студент_ка (Bozhenko et al., 2022). Kirey-Sitnikova (2021) mentions neuter gender along with singular they, gender-gapping, impersonal or indefinite personal structures, plural instead of singular. However, the author points to the drawbacks of the existing forms which results in their limited use and acceptability in the community. However, there is no consistent strategy towards inclusive forms. For our work, we adopt the form that comprise both gender variants in a word as the best inclusive solution.

Table 2 shows some of the possible constructions to correctly avoid the first-person gender.

3.3 Analysis

The analysis process is similar to the one described in (Popovic and Lapshinova-Koltunski, 2024). The annotations are performed on the multiple translations of the 196 reviews by experts, i.e. trained linguists with the native command of the target languages. The annotators carry out the following instructions: for each review, assign a gender label according to the first-person gendered words (adjectives, verb past and passive participles) in it. If the first-person gendered words within a review are consistent (masculine, feminine or inclusive), assign this gender label to the review. If there is a mixture of genders but no inclusive forms, assign the label "mix". If any kind of error related to the first person gender occurs, add the label "e" to the gender label.

The details on the annotation scheme are provided in Table 3, and examples of gender labels are shown in Table 4.

As for the label "x", over 95% of the reviews with this label do not have any adjectives or past tenses in the source language, thus no requiring any gender in the target language. A very small number of reviews is correctly re-written to avoid the gender, for example using present tense instead of conditional form (which requires past participle) for "I would recommend", or rephrasing "I am annoyed" into "it makes me nervous". The attempts to rewrite in a similar way where the meaning of the text is changed or the grammatical structure became incorrect were labelled as errors.

Table 5 shows examples of constructions where it is possible to avoid the first-person gender.

4 Results

The results are presented in Figure 1, where we display distributions of the different gender labels for each of the translation outputs. As previously mentioned, the results for the GPT3.5 translation from (Popovic and Lapshinova-Koltunski, 2024) are also included for the sake of comparison.

Grey colour indicates reviews without first-person gender while orange indicates reviews with a gender inclusive form. Those two categories are considered to be without gender bias and are the focus of the work. For both categories, lighter nuances denote that some error ocurred related to the first-person gender. In red, blue and violet reviews, the system did not generate any word in the inclusive form, but only in single gender form: red in feminine, blue in masculine, and violet with mixed genders or with errors.

[6]See recommendations for Russian under https://www.un.org/ru/gender-inclusive-language/guidelines.shtml, accessed on 22.04.2025.

I was disappointed	⇒	it made me disappointed
I am upset	⇒	it made me upset
I would recommend	⇒	I recommend

Table 2: Examples of possible rephrasing to completely avoid the first person gender: using passive instead of active mode (1 and 2) using indicative instead of conditional mode

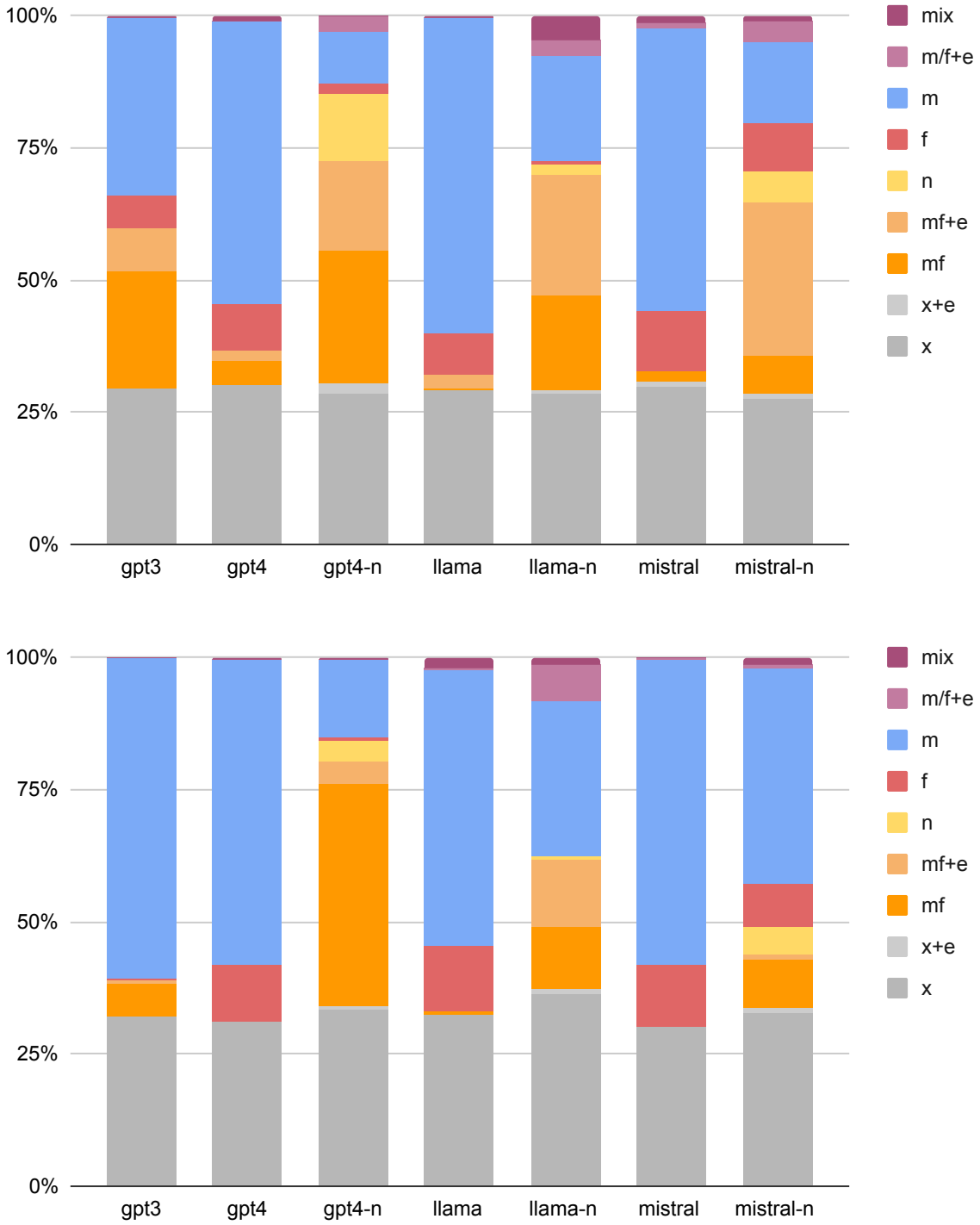

Figure 1: Distribution of first person gender in Croatian (above) and Russian (below) translations for different models. The name of the model with "-n" denotes the neutral prompt.

label	description
x	no first-person gender
x+e	no first-person gender but with errors
mf	inclusive form (contains (both masculine and feminine form)
mf+e	inclusive form and errors
n	neuter gender (error)
m	masculine
f	feminine
mix	mixed genders without inclusive forms

Table 3: Annotaton scheme.

4.1 RQ1: Differences between the outputs

We first compare the distributions across the outputs of LLMs produced either with the holistic prompt or with the prompt instructing to a gender-neutral translation. The general tendencies are same for both language pairs. The amount of completely gender neutral translations (grey) varies only slightly, without notable differences when using a neutral prompt. This is expected, because as mentioned in Section 1, it is difficult to completely avoid gender in the given target languages.

The simple holistic prompt outputs are definitely prone to gender bias, with masculine forms (blue) dominating. Furthermore, we observe more inclusive forms in Croatian and more feminine forms in Russian. An interesting observation is that for all languages, GPT3.5 generated more inclusive forms than GPT4.

The percentage of gender inclusive forms, as hoped for, notably increases when gender-neutral prompt is used. However, the neutral prompt outputs of different LLMs are rather heterogeneous. They also seem to be more prone to errors. The details will be discussed in the next two sections.

4.2 RQ2: Capability of the gender neutral prompt to reduce bias

Now, we analyse category by category the changes that we achieved with the neutral prompt in comparison to the holistic prompt.

Forms without first-person gender As already mentioned, we do not observe any notable changes for this category (marked with x, grey colour in Figure 1), which is expectable, as it is difficult to completely avoid the gender. Apart from that, we note that some of the outputs are written without the first-person gender, however using an inadequate verb form instead of the past participle. These cases are marked as x+e (no first-person gender but with errors) and are displayed in light grey in Figure 1.

Inclusive forms Overall, the number of inclusive forms (both correct as well as erroneous) increases by using the neutral prompt. However there are differences between the models as well as between the two languages.

The largest increase in both languages can be observed for the GPT4 model. About 20% of the outputs still remain in a single gender form, predominantly masculine. While a large portion of Russian inclusive forms is correct, the Croatian model more often ends up with errors.

The other two models generate more inclusive forms in Croatian than in Russian, however also with a larger proportion of errors. It can also be noted that Mistral-neutral translations retains a large portion of feminine forms, contrary to the other two models which almost do not generate any feminine forms with the neutral prompt.

Furthermore, it can be seen that neutral prompt sometimes results in feminine or masculine translations with errors (violet in Figure 1), especially for Llama.

Overall, it can be seen that the zero-shot neutral prompt increases the number of inclusive forms and reduces bias for all models. However, the percentage of fully correct inclusive forms is still relatively low, so the prompts should be improved in order both to reduce the errors as well as to further reduce the number of single gender (mainly masculine) forms.

Qualitative analysis of errors and corresponding suggestions for formulating better prompts is presented in the next section.

4.3 RQ3: Analysis of problems and suggested solutions

While attempting to generate inclusive forms, different types of errors were observed in all models. One particular error is consistent use of the third grammatical gender, namely the neuter form (yellow colour in Figure 1). Despite the fact that the form itself is fully grammatically correct, it can never be used as the first-person gender, and therefore represents an error. It can be noted that GPT4 is mostly inclined to this form, especially for Croatian, while it is rare in the Llama outputs.

The main problem with all models is a mixture of words in the desired form and words with different types of errors (light orange). Qualitative analysis

	en	this is fake MAC, i just **received** mine and super **upset** to find out it isnt real MAc.
fem.	hr	Ovo je fejk MAC, upravo sam **dobila** svoj i jako sam **ljuta** što nije pravi MAC.
	ru	Это подделка MAC, я только что получила свою косметику и ужасно расстроена, потому что это не настоящая косметика MAC!
masc.	hr	Ovo je fejk MAC, upravo sam **dobio** svoj i jako sam **ljut** što nije pravi MAC.
	ru	Это подделка MAC, я только что получил свою косметику и ужасно расстроен, потому что это не настоящая косметика MAC!
incl.	hr	Ovo je fejk MAC, upravo sam **dobio/la** svoj i jako sam **ljut/a** što nije pravi MAC.
	ru	Это подделка MAC, я только что получил(а) свою косметику и ужасно расстроен(а), потому что это не настоящая косметика MAC!
mixed	hr	Ovo je fejk MAC, upravo sam **dobila** svoj i jako sam **ljut** što nije pravi MAC.
	ru	Это подделка MAC, я только что получил свою косметику и ужасно расстроена, потому что это не настоящая косметика MAC!
err.	hr	Ovo je fejk MAC, upravo **smo dobili** svoj i jako sam **ljut/a** što nije **pravi/a** MAC.
	ru	Это подделка MAC, я только что получила свою косметику и ужасно расстроен(а), потому что это не настоящая косметика MAC!

Table 4: Example of gender labels assigned according to first-person gendered words: red=feminine, blue=masculine, orange=inclusive, violet=errors.

	en	I was disappointed	I am upset	I would recommend
mf	hr	bio/la sam razočaran/a	uznemiren/a sam	preporučio/la bih
	ru	я был/а разочарован/а	я был/а расстроен/а	рекоммендую/
x	hr	Bilo mi je žao	uznemirilo me je	preporučujem
	ru	меня разочаровало	меня расстроило	рекоммендую
	gloss for x	it made me sad/disappointed	it made me upset	I recommend

Table 5: Example of constructions with possibility to avoid first-person gender.

of these errors identified the following problems:

- using incorrect verb form (present tense, an incorrect past tense without a past participle, passive form, plural, impersonal form):

  `upravo dobijem svoj i jako sam ljut/a`

  gloss: `I just get (ERROR) mine and I'm very angry (m/f).`

  In some reviews, incorrect verb form was used consistently thus resulting in a genderless erroneous variant labelled as x+e (light gray in Figure 1).

- mixing inclusive forms with single gender forms within a review:

  `upravo sam dobio/la svoj i jako sam ljut`

  gloss: `I have just got (m/f) mine and I am very angry (m).`

  This also includes cases in which gender is explicit in some lexical items. For instance, there are two different words for `married` in Russian – женат for married men and замужем for married women. One of the LLM outputs contained both forms. However using the corresponding verb in mascluline form only:

  и при этом я даже не женат/не замужем! Я купил эту книгу, как только она была опубликована

  gloss: `and yet I'm not even married (m/f)! I bought (m) this book as soon as it was available published.`

- including neuter gender in the inclusive form:

  `upravo sam dobio/la/lo svoj i jako sam ljut/a/o`

  gloss: `I just got (m/f/n) mine and I am very angry (m/f/n).`

  Sometimes, only neuter gender was used as a gender neutral version with the first person, which is not possible from the semantic point of view (apart from some characters in Russian fairy tails):

и я было так взволновано, когда купило это для своего ребёнка

gloss: and I was so excited (n) when I bought (n) this for my child.

- generating incorrect inclusive form (e.g. plural or non-existing suffixes, etc.)

  upravo sam dobio/li svoj i jako sam ljut/a/e

  gloss: I just got (m/pl) mine and I am very angry (m/f/pl).

  Я купил(и) это в формате LP

  gloss: I bought (m/pl) it in LP format.

  Я не понимаю, как пропустил эту важную деталь, но пропустил(я)

  gloss: I don't understand how I missed this important detail, but I did (m/non-existing suffix).

  In some cases, the models generated wrong pronouns changing the first person into the third person plural and using as an inclusive alternative:

  Я (они) высоко рекомендую это всем, кто не хочет вырасти из любимого хобби.

  gloss: I (they) highly recommend this to everyone, who doesn't want to grow out of their favorite hobby.

- generating inclusive form for objects (which have pre-defined gender)

  upravo sam dobio/la svoj/u i jako sam ljut/a

  gloss: I just got (m/f) mine (m/f) and I'm very angry (m/f).

- generating inclusive forms for third person verbs, however, keeping the pronoun in a specific gender:

  Купил(а) ее для ребенка, когда он был(а) в детском саду, и он до сих пор любит в нее играть

  gloss I bought (m/f) this for my child when he (m) was (m/f) in kindergarten and he (m) still loves to play with it.

- changing form of other words (e.g. auxiliary verb, pronoun, noun, etc.)

  upravo sam/smo dobio

  gloss: I have (sing/pl) just got (m) .

5 Conclusion

In this work, we examined three language models and their ability to generate a review translation from English into two Slavic languages, Croatian and Russian, in which first person forms are used in gender-neutral or inclusive manner. We investigated two prompts, i.e. a simple holistic prompt, which corresponds to the translation brief for human translations, as well as a zero-shot prompt with a verbalised instruction to produce a gender-neutral translations for the first person forms.

Our overall result is that the gender-neutral prompt increases the percentage of inclusive variants for both languages and all models, however, with a number of errors. We quantitatively analysed various solutions, as well as the emerging errors. The nature of these errors was then qualitatively examined. As an outcome, we formulate a number of recommendations for more specific prompts that can be used to eliminate the possible errors:

- use few-shot prompts containing examples of desired inclusive word forms;
- explicitly ask not to use other verb tenses and plural forms;
- explicitly ask not to use neuter gender;

We plan to extend our prompting strategies following these recommendations.

Although our research is restricted to two languages only, we believe that these recommendations will be useful for other Slavic languages too. Besides that, these recommendations could also be valuable for other languages with explicit first-person gender marking. We also believe that our findings will be useful not only for translation tasks with LLMs, but also for other language generation tasks for texts and languages with similar properties, e.g. those containing first person mentions in highly gendered languages.

6 Bias statement

With the focus on the analysis of first person forms in translations into languages with grammatical gender marking, this work addresses gender bias problem in product review translations. Testing the ability of LLMs to use gender-neutral forms,

we try to mitigate the existing bias in automated translation that results in stereotypes related to the product types reviewed.

7 Limitations

For a better understanding of the explored phenomena across languages, we need a data set that includes translations into not only Slavic but also other languages, for example the Romance ones.

Also, comparison with a reference human translation would be an asset. However, the human translations available in the dataset (see the description of the corpus DiHuTra, Lapshinova-Koltunski et al., 2022) are not gender-neutral and contain gender-bias that could be also linked to the types of the products reviewed as shown by (Popovic and Lapshinova-Koltunski, 2024).

We are aware of the problems of reproducibility related to the nature of closed-source models. The future results that build upon our findings may differ from those reported by us, as LLMs are regularly updated and are changing.

Acknowledgements

ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology, is funded by Science Foundation Ireland through the SFI Research Centres Programme under Grant Agreement No. 13/RC/2106_P2.

References

Yulia Bozhenko, Lyudmila Em, and Elena Kalinovskaya. 2022. Linguistic Signs of Gender Neutrality in the English and Russian Languages (by the Example of the Internet Publications). *GRAMOTA Publishers*, 15:1543–1547.

Ali Doosthosseini, Jonathan Decker, Hendrik Nolte, and Julian M. Kunkel. 2024. Chat AI: A Seamless Slurm-Native Solution for HPC-Based Services. *Preprint*, arXiv:2407.00110.

Nizar Habash, Houda Bouamor, and Christine Chung. 2019. Automatic gender identification and reinflection in Arabic. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 155–165, Florence, Italy. Association for Computational Linguistics.

Janiça Hackenbuchner, Joke Daems, Arda Tezcan, and Aaron Maladry. 2024a. You shall know a word's gender by the company it keeps: Comparing the role of context in human gender assumptions with MT. In *Proceedings of the 2nd International Workshop on Gender-Inclusive Translation Technologies*, pages 31–41, Sheffield, United Kingdom. European Association for Machine Translation (EAMT).

Janiça Hackenbuchner, Arda Tezcan, and Joke Daems. 2024b. Automatic detection of (potential) factors in the source text leading to gender bias in machine translation. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 2)*, pages 27–28, Sheffield, UK. European Association for Machine Translation (EAMT).

Yana Kirey-Sitnikova. 2021. Prospects and challenges of gender neutralization in Russian. *Russian Linguistics*, 45:143–158.

Ekaterina Lapshinova-Koltunski, Maja Popović, and Maarit Koponen. 2022. DiHuTra: a parallel corpus to analyse differences between human translations. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1751–1760, Marseille, France. European Language Resources Association.

Mara Nunziatini and Sara Diego. 2024. Implementing gender-inclusivity in MT output using automatic post-editing with LLMs. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 580–589, Sheffield, UK. European Association for Machine Translation (EAMT).

Andrea Piergentili, Beatrice Savoldi, Matteo Negri, and Luisa Bentivogli. 2024. Enhancing gender-inclusive machine translation with neomorphemes and large language models. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 300–314, Sheffield, UK. European Association for Machine Translation (EAMT).

Maja Popovic and Ekaterina Lapshinova-Koltunski. 2024. Gender and bias in Amazon review translations: by humans, MT systems and ChatGPT. In *Proceedings of the 2nd International Workshop on Gender-Inclusive Translation Technologies*, pages 22–30, Sheffield, United Kingdom. European Association for Machine Translation (EAMT).

Flavien Prost, Nithum Thain, and Tolga Bolukbasi. 2019. Debiasing embeddings for reduced gender bias in text classification. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 69–75, Florence, Italy. Association for Computational Linguistics.

Chahat Raj, Anjishnu Mukherjee, Aylin Caliskan, Antonios Anastasopoulos, and Ziwei Zhu. 2024. BiasDora: Exploring hidden biased associations in vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10439–10455, Miami, Florida, USA. Association for Computational Linguistics.

Eduardo Sánchez, Pierre Andrews, Pontus Stenetorp, Mikel Artetxe, and Marta R. Costa-jussà. 2024.

Gender-specific machine translation with large language models. In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pages 148–158, Miami, Florida, USA. Association for Computational Linguistics.

Aleix Sant, Carlos Escolano, Audrey Mash, Francesca De Luca Fornaciari, and Maite Melero. 2024. The power of prompts: Evaluating and mitigating gender bias in MT with LLMs. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 94–139, Bangkok, Thailand. Association for Computational Linguistics.

Beatrice Savoldi, Marco Gaido, Matteo Negri, and Luisa Bentivogli. 2023. Test suites task: Evaluation of gender fairness in MT with MuST-SHE and INES. In *Proceedings of the Eighth Conference on Machine Translation*, pages 252–262, Singapore. Association for Computational Linguistics.

Beatrice Savoldi, Janiça Hackenbuchner, Luisa Bentivogli, Joke Daems, Eva Vanmassenhove, and Jasmijn Bastings, editors. 2024a. *Proceedings of the 2nd International Workshop on Gender-Inclusive Translation Technologies*. European Association for Machine Translation (EAMT), Sheffield, United Kingdom.

Beatrice Savoldi, Andrea Piergentili, Dennis Fucci, Matteo Negri, and Luisa Bentivogli. 2024b. A prompt response to the demand for automatic gender-neutral translation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 256–267, St. Julian's, Malta. 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.

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

Eva Vanmassenhove, Chris Emmery, and Dimitar Shterionov. 2021. NeuTral Rewriter: A rule-based and neural approach to automatic rewriting into gender neutral alternatives. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8940–8948, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Eva Vanmassenhove, Beatrice Savoldi, Luisa Bentivogli, Joke Daems, and Janiça Hackenbuchner, editors. 2023. *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*. European Association for Machine Translation, Tampere, Finland.

Zhilin Wang, Alexander Bukharin, Olivier Delalleau, Daniel Egert, Gerald Shen, Jiaqi Zeng, Oleksii Kuchaiev, and Yi Dong. 2024. Helpsteer2-preference: Complementing ratings with preferences. *Preprint*, arXiv:2410.01257.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, Belgium. Association for Computational Linguistics.