# Implementing Gender-Inclusivity in MT Output using Automatic Post-Editing with LLMs

**Mara Nunziatini**
Welocalize
mara.nunziatini@welocalize.com

**Sara Diego**
Welocalize
sara.diego@welocalize.com

## Abstract

This paper investigates the effectiveness of combining machine translation (MT) systems and large language models (LLMs) to produce gender-inclusive translations from English to Spanish. The study uses a multi-step approach where a translation is first generated by an MT engine and then reviewed by an LLM. The results suggest that while LLMs, particularly GPT-4, are successful in generating gender-inclusive post-edited translations and show potential in enhancing fluency, they often introduce unnecessary changes and inconsistencies. The findings underscore the continued necessity for human review in the translation process, highlighting the current limitations of AI systems in handling nuanced tasks like gender-inclusive translation. Also, the study highlights that while the combined approach can improve translation fluency, the effectiveness and reliability of the post-edited translations can vary based on the language of the prompts used.

## 1 Introduction

This paper aims to explore whether LLMs can be effectively utilized for generating gender-inclusive translations. The goal is to determine if this technology can handle the task, or if the expertise of a linguist is still necessary, and to what extent. The challenge lies in the fact that neural machine translation engines frequently fall short in producing gender-inclusive output. When style guides mandate gender-inclusivity in the final translation, post-editors have to make extensive modifications, therefore the MT output is not beneficial for them.

We are investigating a multi-step approach to machine translation in which the translation is first produced by an MT engine and is then reviewed by an LLM, to make it gender-inclusive. The goal is for the LLMs to streamline this process and reduce the need for extensive human intervention.

## 2 The challenges of inclusive writing

Gender-inclusive writing involves using language that does not reinforce traditional gender stereotypes or exclude individuals based on their gender identity. It aims to promote equality and respect for all genders by adopting inclusive terminology and avoiding gendered language whenever possible. Nowadays, gender-inclusive writing is particularly important as societies worldwide strive for greater gender equality and recognition of diverse gender identities.

In this paper, we decided to focus on the translation from English into Spanish for several reasons. Firstly, this language pair poses several gender bias challenges (as we will see later in the paper). Secondly, it is one of the most relevant language pairs from a business perspective for our company. Thirdly, we have highly-trusted internal linguists who are native Spanish speakers and have experience in the translation and post-editing field. Lastly, we have gender-inclusive language style guides available for this language pair, that we used as a starting point for outlining automatic post-editing guidelines. Still, this work is part of an ongoing effort to include additional languages in this experiment.

As mentioned above, gender-inclusive language presents challenges when translating from English into Spanish. This is mainly due to the grammatical structure and inherent gender marking in the Spanish language. Unlike English, where gender-neutral language is more common,

Spanish is a grammatical gender language (Savoldi et al., 2021), that assigns gender to nouns, adjectives, and pronouns. This gender marking extends to articles ('el' for masculine, 'la' for feminine), and even verb conjugations. For example, the English sentence "The doctor saw the patient" can be translated as "El doctor vio al paciente" (masculine doctor, masculine patient), "El doctor vio a la paciente" (masculine doctor, feminine patient), "La doctora vio al paciente" (feminine doctor, masculine patient) or "La doctora vio a la paciente" (feminine doctor, feminine patient).

This inherent gender marking in Spanish makes it challenging to maintain gender neutrality in translations, especially when dealing with professions, titles, and pronouns. Additionally, Spanish has fewer gender-neutral alternatives compared to English, which further complicates the task of creating inclusive translations. Translators are often requested to navigate these linguistic differences while striving to preserve the intended meaning and promote gender inclusivity in the target language.

### 2.1 Machine Translation and gender-inclusive language

It has been observed that machine translation exacerbates the challenges related to gender-inclusive language since, due to several factors, the raw MT output often contains gender bias (Savoldi et al., 2021). This highlights the need for post-editing and careful consideration of gender-inclusive language.

In our experience, training a machine translation engine to generate gender-inclusive language is challenging due to several reasons:

- MT engines often lack the ability to understand the nuanced context in which gendered language is used, translating based solely on grammar and vocabulary without considering the broader sociocultural implications of gender. This lack of context is often due to the segmentation process that documents go through in order to be translated in Translation Management Systems.
- Different languages have unique grammatical structures and conventions regarding gender. For instance, while English has relatively more gender-neutral options, languages like Spanish assign gender to nouns, adjectives, and pronouns. This variability makes it difficult to create a one-size-fits-all approach to gender inclusivity in machine translation.
- Different clients have different requirements for gender-inclusive language.
- Machine translation models are trained on large datasets of translated texts. However, these datasets may not always include sufficient examples of gender-inclusive language, leading to biases in the generated translations.
- Using gender-inclusive language often means rephrasing, for example: "gays" → "hombres y mujeres homosexuales". This is especially true if the source text itself includes gender-biased language. Rephrasing requires a deep understanding of context and linguistic subtleties, which can be challenging for machine translation systems.
- We often receive very generic (if any) gender-inclusivity guidelines from clients, which are not detailed enough to train a model.

Overall, training a machine translation engine to generate gender-inclusive language requires addressing these complex linguistic, cultural, and contextual challenges, which may necessitate advanced techniques in natural language processing and extensive fine-tuning of algorithms.

### 2.2 LLMs for automatic post-editing of gender-biased translations

It appears that LLMs have the potential to be highly effective tools for post-editing tasks. For example, it has been demonstrated that GPT-4 offers promising results on post-editing (Raunak et al., 2023). Besides, LLMs made by large tech companies go through steps which have a goal of minimizing biases in their outputs (Ouyang et al., 2022). We therefore see the identification and fixing of gender-bias issues (whilst translating text) a challenging and very relevant benchmark for judging and comparing LLMs' performance.

Several experiments have been carried out recently to benchmark MT engines and LLMs, and it has been demonstrated that Neural MT engines keep performing better than LLMs (Welocalize, 2023), especially as for accuracy (Vilar et al.,

2023). We think that by using GPT-4 and PaLM2 for automatic post-editing on the raw MT output, we will take advantage of the accuracy delivered by MT engines while improving the translation's fluency with LLMs.

LLMs' ability to understand the context of a text, thanks to being trained on vast and diverse datasets, allows them to make meaningful and contextually appropriate edits. This, combined with their ability to process and edit large volumes of text relatively quickly, makes them a valuable resource for large-scale projects. Also, LLMs can be fine-tuned according to specific guidelines or style guides, including those for gender-inclusive language. We thought that this could make them a potentially valuable tool for enhancing inclusivity in machine translation outputs.

## 3 Experimental Settings

### 3.1 Producing the initial translations with MT and LLMs

For this test we utilized content shared by a client which is a globally recognized technology company, and mindful of gender-inclusivity. The content we selected includes text about product integration, technical services, customer support, sales inquiries, cloud solutions, and community interactions. We have chosen this content type as it is written in a way that appeals to all genders, making it an ideal candidate for the test. The language to be used in the translation must be professional, informative, and inclusive, avoiding any gender-biased terms or phrases. This makes it an excellent example of gender-inclusive content in the tech industry. The content was previously translated, therefore we owned the reference human translation.

Firstly, we are interested in producing the initial translations and finding out how the outputs from 5 different systems compare against the human reference translation. This will allow us to choose the best output (output most similar to the reference human translation) to be used as a starting point to generate the gender-inclusive post-edited translation.

For producing the initial translations, we experimented with a subset of 1,000 segments (15,307 words). The systems we used for initial translation generation are:

1. DeepL. We chose this engine since in our experience it is one of the best-performing engines for en>es-ES.
2. GPT-4 (OpenAI, 2023). We chose this system as it has been proved that it consistently performs better than GPT-3.5 (Raunak et al., 2023).
3. PaLM2. State-of-the-art language model that has better multilingual and reasoning capabilities and is more compute-efficient than its predecessor PaLM (Anil et al., 2023). For this exercise we are using text-bison@002 model.
4. DeepL output post-edited by GPT-4.
5. DeepL output post-edited by PaLM2.

In order to enhance the consistency of the assessments and more accurately represent the methodology of general users, our present efforts will concentrate on the zero-shot learning scenario for LLMs, in which the model is not presented with any examples provided by humans. The prompts used to generate the initial translations and the post-edited version of the initial translations can be found in the Appendix (Appendix A and B).

To measure the quality of the 5 outputs, we will compare each one of them against the reference human translation. This will be done by computing COMET (Rei et al., 2020), BLEU (Papineni et al., 2002) and Levenshtein Edit Distance (in our analysis, we normalize this value by the number of characters in the MT output), as these are 3 of the most commonly used reference-based state-of-the-art neural MT quality metrics in the translation industry. The results of this comparison can be found in Paragraph 4.1 (Table 1).

### 3.2 Performing automatic post-editing to fix gender bias issues in MT output

Secondly, we will perform automatic post-editing with LLMs (GPT-4 and PaLM2), focused solely on fixing gender-bias issues.

In the context of this study, we created a dummy style guide by merging a generic inclusivity writing manual created by our company and a more detailed inclusive writing style guide provided by our client. We therefore asked GPT-4 to transform the resulting style guide into a list of prompts to be used by GPT-4 itself. The list is appended to this paper and was added to the prompt used to perform the automatic post-editing tasks with both LLMs.

We extracted 200 segments from the initial translations, and annotated gender bias issues. We then generated the post-edited gender inclusive translations with GPT-4 and PaLM2. Our internal linguists then evaluated the effectiveness of GPT-4 and PaLM2 in correcting gender bias errors in both Spanish and English texts at segment level, using a labelling system. Labels were "ALL" if all issues were fixed, "PARTIAL" if only some were addressed, and "NONE" if no issues were corrected. The scores can be found in Paragraph 4.2.

## 4 Results and analysis

### 4.1 Initial translations with MT and LLMs

| Solution | BLEU | PE Distance | COMET |
|---|---|---|---|
| DeepL | 49.70 | 28.00% | 0.89 |
| GPT-4 | 41.47 | 31.00% | 0.88 |
| PaLM2 | 46.48 | 31.00% | 0.89 |
| DeepL+GPT-4 | 45.20 | 30.00% | 0.89 |
| DeepL+PaLM2 | 50.29 | 28.00% | 0.90 |

**Table 1:** Quality scores for the initial translations. The "+" sign in the Solution column is to be interpreted as "post-edited by".

The results in Table 1 suggest that there is no meaningful difference between the 5 different outputs. DeepL+PaLM2 performed the best in terms of translation accuracy and produced an output which is most similar to human reference. However, DeepL alone and the combined approach of DeepL and GPT-4 also performed well.

While GPT-4 and PaLM2 alone performed reasonably well in terms of translation quality, they did not strictly adhere to the prompt we provided for the post-editing step. The internal linguists who carefully reviewed DeepL output post-edited by GPT-4 and DeepL output post-edited by PaLM2 found that both reworked the text more than necessary, to enhance fluency. In many cases, this resulted in the introduction of unnecessary preferential changes, ignoring the part of the prompt stating "Don't change anything if the Proposed Translation is accurate and fluent". In fact, these changes didn't always improve the accuracy or understanding of the text, but rather added a layer of subjective interpretation that was not present in the original text. Moreover, GPT-4 introduced inconsistencies in terminology. For instance, the term "whitepaper", which was consistently translated by DeepL as "libro blanco", was sometimes changed by GPT-4 and PaLM2 into different terms such as "documento técnico", "documento", "informe blanco", "informe técnico" or "documentación técnica". Other times, it was left unchanged ("libro blanco") by both LLMs. These inconsistencies can make the job of the post-editor more difficult, as we believe that it is cognitively less demanding and more time-efficient for a reviewer to rectify a recurring terminology inconsistency in a translation than to deal with a single source term translated into the target language in various ways.

In essence, while GPT-4 and PaLM2 showed potential in enhancing fluency, their tendency to introduce unnecessary changes and inconsistencies in terminology raises concerns about their reliability for consistent and accurate translations. Moreover, GPT-4 frequently added the term "Reviewed" at the beginning of the segments, despite the prompt specifically asking for the reviewed text to be returned alone. A similar behavior was already documented in the literature (Zhang et al., 2023) but it came as unexpected since it did not happen in previous tests performed internally by our teams with a similar prompt. This suggests that GPT-4 may have misinterpreted the instructions or overgeneralized from its training data, leading to unnecessary additions to the translated text. This behavior alone unequivocally underscores the continued necessity for human review in the process.

In our process of selecting the most suitable output for our experiment, we chose DeepL's output. This decision was based on our evaluation of its performance in terms of accuracy, fluency, and consistency of terminology. Furthermore, in a view of adopting this solution in a larger scale scenario, DeepL alone is more cost-effective and time-efficient compared to DeepL reviewed by LLMs. We found that the additional effort required and expense incurred for LLM usage was not justified by a meaningful improvement in quality.

### 4.2 Automatic post-editing to fix gender-bias issues in MT output

We now use GPT-4 and PaLM2 to review DeepL's output and make edits solely aimed to ensure that it is gender-inclusive. This means ensuring that the language used does not favor one gender over another and is respectful and inclusive of all genders, following a series of guidelines added to the prompt.

**Segment selection and error marking in the initial translation:** To ensure an unbiased and

random selection for this experiment, we extracted 200 segments from the initial translation with DeepL. This random extraction ensures a fair and representative sample of the overall text, as it doesn't favor any particular section of the text. The 200 segments were then analyzed by two internal linguists. These individuals are skilled professionals who specialize in language translation and have a keen understanding of gender bias in language.

These internal linguists reviewed each of the 200 segments and marked any gender bias errors, observing the same guidelines that were included in the prompt. These errors could include language that unfairly represents one gender over another, excludes certain genders, or otherwise fails to be inclusive. Out of the 200 segments analyzed, the internal linguists found that 140 of these segments contained one or more gender bias errors. This means that a significant majority of the segments translated by DeepL had issues with gender bias in the translated text. For example, the Spanish equivalent terms for "analyst", "customer", "manager", "developer", were often used in their masculine form. On the other hand, 60 out of the 200 segments were found to be free of any gender bias errors. This means that these segments were considered by the internal linguists to be gender-inclusive, or simply did not include challenges for gender inclusivity.

**Prompting strategy and post-editing by GPT-4 and PaLM2:** Both GPT-4 and PaLM2 were then tasked with editing these segments to make them gender-inclusive. This was done using a specific prompt provided in the Appendix of this paper (Appendix C and D), which would have given GPT-4 and PaLM2 guidance on how to approach this task.

The original gender-inclusive language style guides (which we used as a starting point to create the prompt for post-editing) were written in English but included some examples in Spanish. This created a bit of a dilemma when we were trying to decide the language to use for the prompt. Some research had already been done on this topic (Lai et al., 2023; Zhang et al., 2023), and it appears that LLMs perform better with English prompts even if the task and input texts are intended for other languages. Still, we were curious to see if and how the test outcome differs by changing the prompt language. Therefore, we decided to use two different prompts for gender-inclusive post-editing: first the one in English, and then its translation into Spanish, produced by a professional translator. The reader can find these in Appendix C and D.

**Results and discussion:** After GPT-4 and PaLM2 had made their edits, the revised segments were given back to the internal linguists for review. The internal linguists then evaluated the changes made by GPT-4 and PaLM2 both with the Spanish and with the English prompt and determined how effectively it had fixed the gender bias errors. The internal linguists used a labelling system to indicate the effectiveness of the LLM's edits (Raunak et al. 2023) at segment level:

- If the LLM had successfully fixed all the gender bias issues in a segment, the internal linguists labelled it as "ALL".
- If the LLM had only managed to fix some, but not all, of the gender bias issues, the segment would be labelled as "PARTIAL".
- If the LLM was unable to fix any of the gender bias issues in a segment, the segment was labelled as "NONE".

This scoring system allowed us to evaluate the effectiveness of using LLMs for post-editing to remove gender bias from machine translations with the English and Spanish prompt.

We noticed that there were two different dimensions that are worth commenting on:

- Quantitative: the number of errors found and fixed by each LLM and
- Qualitative: the quality of the resulting translation.

**Quantitative:** By looking at Figure 1, we can notice that GPT-4 is more successful than PaLM2 in fixing gender bias issues. GPT-4 was able to identify and fix the majority of gender bias issues. PaLM2 was not as successful, and almost half of the segments with gender bias issues were not fixed or only partially fixed. The above is true both with the English and Spanish prompt.

In fact, the Spanish and English prompt delivered similar results, with the English prompt delivering slightly better results. In more detail, the test results indicated that:

- PaLM2 – the English prompt delivered a slightly better post-edited translation, as the % of segments with gender bias issues that were not fixed at all ("NONE") is

smaller compared to the post-edited translation delivered with the Spanish prompt.

- GPT-4 – the difference is more meaningful, with almost 80% of the segments with gender-bias issues completely fixed after post-editing with the English prompt, against the 74% with the Spanish prompt.

Based on this data, we can conclude that:

- GPT-4 is more successful than PaLM2 in this task.
- GPT-4 is somewhat more effective at identifying and fixing gender-bias issues when using English prompt compared to Spanish prompt, while changing the prompt language does not make a meaningful difference with PaLM2.
- There is still a clear need for human review as not all segments with gender bias issues were detected and rectified. However, using LLMs helps reducing the number of changes needed.

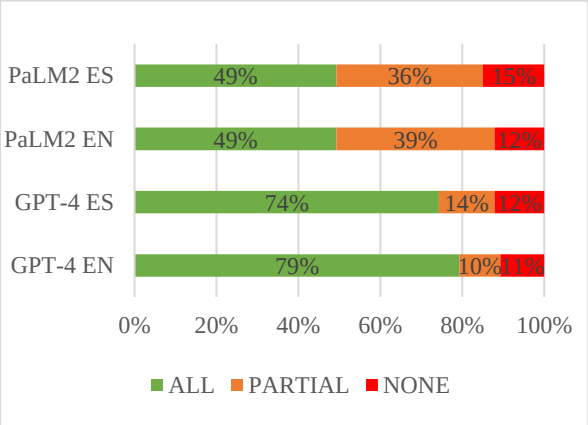


**Figure 1:** percentage of segments with gender bias errors fixed, partially fixed, and not fixed by PaLM2 and GPT-4, with Spanish (ES) and English (EN) prompts.

**Qualitative:** The internal linguists noticed that in some cases, the tone of voice was unnecessarily changed in all the four post-edited translations, varying from a formal tone to an informal one. This was against the client's style guide and was also not requested in the prompt. These unrequested changes can be problematic in a real case scenario, as inconsistencies in tone of voice can complicate the work of the post-editor, who would have to edit much of the text to ensure coherence. Furthermore, in the case of PaLM2, major errors were found with the English prompt, which included neuter forms such as "les desarrolladores". This solution uses the letter "e" as an alternative for "a" (feminine) or "o" (masculine) in articles, nouns, and pronouns. It is a recent linguistic development aimed at promoting gender neutrality. However, this solution is not officially recognized (García, 2021b) and, most importantly, it goes against the instructions included in the prompt. Another example is the addition of "sin importar su género" (which translates into "no matter their gender") in the translation.

Besides, the internal linguists also identified a difference in the quality of the edits between the outputs obtained with the English and Spanish prompts. To address this, we asked our internal linguists to carry out a qualitative ranking of the two translations post-edited by GPT-4 for each segment, judging which gender-inclusive revision was superior from an adequacy and fluency standpoint. We decided to perform this analysis on the translations post-edited by GPT-4 only, without focusing on the translations post-edited by PaLM2, because the former was more successful at this task.

The translator analyzed 140 segments, indicating which between the two post-edited translations demonstrated superior quality for each respective segment. The results indicate that for the greater part of the segments (62%), both translations were comparable from a qualitative standpoint. However, for 22% of the segments, the post-edited translation generated with the English prompt was better, while for the remaining 16%, the post-edited translation generated with the Spanish prompt was better. It was observed that, in those cases where the post-edited translation generated with the English prompt was better, the gender-inclusive solutions proposed were more natural and fluent.

From these results it can be concluded that the choice of the language prompt can have an impact on the quality of the translation, although, in our experiment, in most cases both options delivered similar results.

## 5 Limitations

The analysis predominantly relies on the outcomes generated by three AI systems, leaving out a comprehensive perspective of the broad array of machine translation systems and large language models available. The study's focus on a single content type potentially overlooks variations in language use across diverse contents. By examining a limited subset of segments, the study may risk forming a skewed understanding of AI capabilities. Solely focusing on one

language pair fails to consider the inherent structural, complexity, and nuance differences among languages. A more thorough evaluation would require a diverse range of content types and AI systems, a broader selection of segments, as well as multiple language pairs. Finally, we recognize that a thorough comparison of the solutions we examined should ideally include an analysis of output generation speed and associated costs. However, given the page limitations for this paper, we chose to omit this aspect from our current discussion.

## 6 Conclusions

In conclusion, this study has presented a comprehensive analysis of the performance of LLMs in producing gender-inclusive translations starting from DeepL's raw output. The findings indicate that despite certain potential shown by GPT-4 and PaLM2, the frequent introduction of unnecessary changes, additions, as well as inconsistencies in terminology and tone of voice, raises concerns about their reliability. Furthermore, GPT-4 was found to be more successful than PaLM2 in identifying and fixing gender-bias issues, especially when using an English prompt. This is probably due to the different size of their respective training datasets: GPT-4 was trained on a significantly larger dataset than PaLM2, which means that GPT-4 has "more knowledge" than PaLM2. The study also highlighted the potential impact of the prompt's language on the quality of the translation.

The necessity for human review remains paramount, as not all gender bias issues were detected and rectified by the systems analyzed. Besides, while the use of LLMs to address gender bias issues in translation effectively mitigates the necessity for substantial human intervention in this particular area, it introduces other complications. Specifically, LLMs can create unnecessary alterations in the post-edited translation, such as inconsistencies in terminology and tone of voice. This, in turn, requires further post-editing effort to correct these unintended changes. Therefore, despite the advantages of using LLMs for reducing gender bias, we can't conclusively state that they decrease the overall workload for the post-editor. Further research should delve into the optimization of these systems and their prompts to enhance the accuracy and inclusivity of machine translations.

## References


Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., Chu, E., Clark, J. H., Shafey, L. E., Huang, Y., Meier-Hellstern, K., Mishra, G., Moreira, E., Omernick, M., Robinson, K., . . . Wu, Y. (2023, May 17). PALM 2 Technical Report. arXiv.org. https://arxiv.org/abs/2305.10403

García, C. (2021b, October 11). El nuevo y tajante mensaje de la RAE sobre el lenguaje inclusivo. La Razón. https://www.larazon.es/cultura/20211011/flcl3i4owvcwrpviqqivljy7wq.html

Lai, V. D., Ngo, N. T., Veyseh, A. P. B., Man, H., Dernoncourt, F., Bui, T., & Nguyen, T. H. (2023, April 12). ChatGPT Beyond English: Towards a Comprehensive Evaluation of large Language models in Multilingual Learning. arXiv.org. https://arxiv.org/abs/2304.05613

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Zoph, B. (2023, March 15). GPT-4 Technical Report. arXiv.org. https://arxiv.org/abs/2303.08774

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022, March 4). Training language models to follow instructions with human feedback. arXiv.org. https://arxiv.org/abs/2203.02155

Papineni, Kishore and Roukos, Salim and Ward, Todd and Zhu, Wei-Jing 2002. BLEU: a method for automatic evaluation of machine translation. Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 311-318.

Raunak, V., Sharaf, A., Wang, Y., Awadalla, H. H., & Menezes, A. (2023). Leveraging GPT-4 for automatic Translation Post-Editing. Findings of the Association for Computational Linguistics: EMNLP 2023. https://doi.org/10.18653/v1/2023.findings-emnlp.804

Rei, Ricardo and Stewart, Craig and Farinha, Ana C and Lavie, Alon. 2020. COMET: A neural framework for MT evaluation arXiv preprint arXiv:2009.09025

Savoldi, B., Gaido, M., Bentivogli, L., Negri, M., & Turchi, M. (2021). Gender bias in machine translation. Transactions of the Association for Computational Linguistics, 9, 845–874. https://doi.org/10.1162/tacl_a_00401

Vilar, D., Freitag, M., Cherry, C., Luo, J., Ratnakar, V., & Foster, G. (2022, November 16). Prompting PALM for translation: Assessing strategies and performance. arXiv.org. https://arxiv.org/abs/2211.09102

Welocalize. (2023, August 2). Do LLMs or MT engines perform translation better? - Welocalize. https://www.welocalize.com/do-llms-or-mt-engines-perform-translation-better/

Zhang, B., Haddow, B., & Birch, A. (2023b, January 17). Prompting large language model for machine Translation: A case study. arXiv.org. https://arxiv.org/abs/2301.07069Aho, Alfred V., and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.


### Appendix A. Prompt to generate the Initial Translations

**System message:** You are a professional translator. You are native English and European Spanish speaker. You specialize in technical translations related to computers, servers, data storage devices, software, and other similar products.

**User prompt**: Given Source text in English, only return the Translation in European Spanish. Ensure that the translation is fluent and accurately conveys the Source text meaning.

### Appendix B. Prompt to post-edit the Initial Translations

**System message:** You are a professional post-editor. You are native English and European Spanish speaker. You specialize in technical translations related to computers, servers, data storage devices, software, and other similar products.

**User prompt:** Given Source text in English and its Proposed Translation in Spanish, only return the reviewed translation. Make sure there are no accuracy or fluency issues in the Proposed Translation. If there are, fix them in the reviewed translation. Don't change anything if the Proposed Translation is accurate and fluent.

### Appendix C. English prompt to review the Initial Translations and make them gender-inclusive

**System message:** You are a professional post-editor. You are native English and European Spanish speaker. You specialize in technical translations related to computers, servers, data storage devices, software, and other similar products. You are very interested in inclusive language and always avoid introducing gender bias in your translations.

**User prompt:** Given the source text in English and its translation into Spanish, only return the post-edited translation. Follow these guidelines:

1. "Check for any gendered terms in the text. If found, can you suggest a gender-neutral alternative for these terms?"
2. "Is the language inclusive for both genders? If not, can you add both gender options, such as 'bienvenidos/as' or 'los/as lectores/as'?"
3. "Can the structure or exact wording of the source text be changed to make the language more inclusive without altering the overall meaning?"
4. "If a gendered term like 'empleado' is used, can you think of alternative ways to describe it, such as 'personal' or 'quienes trabajan en…'?"
5. "Is the masculine used as a neutral plural form? If so, can you modify it to avoid sounding awkward?"
6. "Is the 'pasiva refleja' used in the text to emphasize the action rather than the subject?"
7. "Are there binary gender representations in the text? If so, can you rewrite it using gender-neutral language?"
8. "Are 'x', '@' or 'e' used to bypass gender forms? If so, can you suggest an alternative?"
9. "Is a slash (/a) or parentheses (a) used to cover two gender options? If so, can you suggest a different way of doing it?"
10. "Is gender splitting used in the text, i.e., the repetition of masculine and feminine terms? If so, can you suggest a way to avoid it without losing the text's fluency?"

Note: If none of the above guidelines can be implemented, or when their implementation harms the fluency and naturalness, ask yourself: "Is there a way to maintain the fluency and naturalness of the text while seeking gender neutrality?".

**Appendix D. Spanish prompt to review the Initial Translations and make them gender-inclusive**

**System message:** Eres un profesional de la postedición. Hablas inglés y español de forma bilingüe, y estás especializado en traducciones técnicas relativas ordenadores, servidores, programas informáticos, y otros productos tecnológicos. Estás muy interesado en el lenguaje inclusivo y siempre evitas introducir sesgos de género en tus traducciones.

**User prompt:** Dado el texto de origen en inglés y su traducción al español, solo devuelve la traducción post-editada. Sigue estas pautas:

1. "Revisa si hay algún término de género en el texto. Si es así, ¿puedes sugerir una alternativa neutra en género para estos términos?"
2. "¿El lenguaje es inclusivo para ambos géneros? Si no, ¿puedes agregar ambas opciones de género, como 'bienvenidos/as' o 'los/as lectores/as'?"
3. "¿La estructura o el texto exacto del texto fuente pueden ser cambiados para hacer el lenguaje más inclusivo sin alterar el sentido general?"
4. "Si hay un término de género, como 'empleado', ¿puedes pensar en formas alternativas de describirlo, como 'personal' o 'quienes trabajan en…'?"
5. "¿Se utiliza el masculino como forma plural neutra? Si es así, ¿puedes modificarlo para que no parezca incómodo?"
6. "¿Se utiliza la 'pasiva refleja' en el texto para enfatizar la acción y no el sujeto?"
7. "¿Hay representaciones binarias de género en el texto? Si es así, ¿puedes reescribirlo utilizando un lenguaje neutro en cuanto al género?"
8. "¿Se utilizan 'x', '@' o 'e' para eludir las formas de género? Si es así, ¿puedes sugerir una alternativa?"
9. "¿Se utiliza una barra (/a) o un paréntesis (a) para cubrir dos opciones de género? Si es así, ¿puedes sugerir una forma diferente de hacerlo?"
10. "¿Se utiliza el desdoblamiento en el texto, es decir, la repetición de términos masculinos y femeninos? Si es así, ¿puedes sugerir una manera de evitarlo sin perder la fluidez del texto?"

Nota: Si ninguna de las pautas anteriores puede implementarse, o cuando su implementación perjudica la fluidez y la naturalidad, pregúntate: "¿Hay una forma de mantener la

fluidez y naturalidad del texto mientras se busca la neutralidad de género?”.