The power of Prompts: Evaluating and Mitigating Gender Bias in MT with LLMs

Aleix Sant, Carlos Escolano, Audrey Mash,
Francesca De Luca Fornaciari, Maite Melero
Barcelona Supercomputing Center (BSC)
{aleix.santsavall,carlos.escolano,audrey.mash
francesca.delucafornaciari,maite.melero}@bsc.es

Abstract

This paper studies gender bias in machine translation through the lens of Large Language Models (LLMs). Four widely-used test sets are employed to benchmark various base LLMs, comparing their translation quality and gender bias against state-of-the-art Neural Machine Translation (NMT) models for English to Catalan (En → Ca) and English to Spanish (En → Es) translation directions. Our findings reveal pervasive gender bias across all models, with base LLMs exhibiting a higher degree of bias compared to NMT models.

To combat this bias, we explore prompting engineering techniques applied to an instruction-tuned LLM. We identify a prompt structure that significantly reduces gender bias by up to 12% on the WinoMT evaluation dataset compared to more straightforward prompts. These results significantly reduce the gender bias accuracy gap between LLMs and traditional NMT systems.

1 Introduction

Within the domain of machine translation, gender bias is defined as the tendency of MT systems to produce translations that reflect or perpetuate gender stereotypes, inequalities, or assumptions based on cultural and societal biases (Friedman and Nissenbaum, 1996; Savoldi et al., 2021). Given that the presence of such bias can lead to harmful consequences for certain groups — either in representational (i.e., misrepresentation or underrepresentation of social groups and their identities) or allocational harms (i.e., allocation or withholding of opportunities or resources to certain groups) — (Levesque, 2011; Crawford, 2017; Lal Zimman and Meyerhoff, 2017; Régner et al., 2019), it becomes paramount to thoroughly investigate and mitigate its occurrence. Nevertheless, addressing gender bias is a multi-faceted task.

Gender bias is a pervasive issue in all generative NLP models, and LLMs are no exception to this situation. LLMs have gained significant popularity in recent years and are being used for many NLP tasks, including machine translation. While gender bias in machine translation has been extensively studied for Neural Machine Translation models, little attention has been paid to this type of bias in LLMs. This paper aims to address this gap by examining and trying to mitigate this bias in the translations generated by the LLMs.

The aim of this work is twofold. First, a comprehensive benchmarking process is conducted to compare various base LLMs with some state-of-the-art NMT models. The directions of the translations under study are English → Catalan and English → Spanish. Distinct popular test sets such as FLoRes-200 (NLLB Team, 2022), WinoMT (Stanovsky et al., 2019), Gold BUG (Levy et al., 2021), and MuST-SHE (Bentivogli et al., 2020) are used to assess the translation quality and the gender bias of the models.

Following the benchmarking, an investigation into the effectiveness of prompts in mitigating this bias in LLMs is conducted. The purpose of this research is to determine whether well-designed prompts can serve as a useful strategy in addressing bias. While existing literature has explored various approaches to mitigating this bias in Neural Machine Translation models (Costa-jussà et al., 2020; Stafanovičs et al., 2020; Saunders and Byrne, 2020), we specifically focus on the realm of LLMs, probing the role of prompts. In this phase of the study, an instruction-tuned LLM is employed, and several prompt engineering techniques are experimented with, including few-shot (Radford et al., 2019; Zhao et al., 2021; Chowdhery et al., 2023), context-supplying, and chain of thought (Wei et al., 2022).

The relevance of this work lies in several insightful findings. Firstly, we demonstrate that base LLMs tend to lag behind NMT models in terms of translation capabilities and gender-bias scores.

Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 94–139

Afterwards, through an extensive trial-and-error examination into prompting, we present a prompt that, when applied to an instructed LLM, achieves impressive bias mitigation across gender-bias test sets, resulting in an increase of 12.4 and 11.7 in the respective Catalan and Spanish WinoMT scores. Finally, we study how gender-bias mitigation through prompting impacts LLMs translation performance.

The rest of the paper is organized as follows: Section 3 reviews relevant research in the field. Section 4 details the methodology, including the datasets, models, and evaluation metrics employed. Section 5 focuses on the benchmarking, while Section 6 explores the investigation into prompting to mitigate gender bias. Section 7 presents the results. Finally, Section 8 provides a discussion and Section 9 highlights the conclusions of this work.

2 Gender Bias Statement

As previously stated, gender bias may lead to inequalities and harmful consequences. In the context of machine translation, we easily come up with two different motivations to consider this issue seriously. First, the presence of gender bias may affect the representation of genders in certain communities. On the other hand, the majority of users of a machine translation system may not be proficient in at least one of the languages involved in the translation. Producing incorrect gender translations provides inaccurate information, misleading users who are trying to understand the original text from a translation, or causing them to convey a different meaning when relying on MT engines to communicate.

The presence and extent of gender bias in machine translation can vary depending on the languages involved, as gender is manifested differently across languages (Dagmar Stahlberg and Sczesny., 2007). When translating from a language with fewer gender cues to a language with more explicit gender markings, the issue of gender bias can arise. This is precisely the case in our study: we translate from a language with notional gender (English) to languages with grammatical gender (Catalan and Spanish). In this context, certain professions may be stereotypically associated with certain genders. Examples of this phenomenon are *engineers*, who are often translated as masculine, while *nurses* are translated as feminine (Parmy Olson, 2018). Additionally, adjectives may be gendered as masculine or feminine based on these stereotypes, rather than relying on gender cues. Gender pronouns may also be overlooked in favor of or against certain genders. Let's consider a typical example (Figure 1).

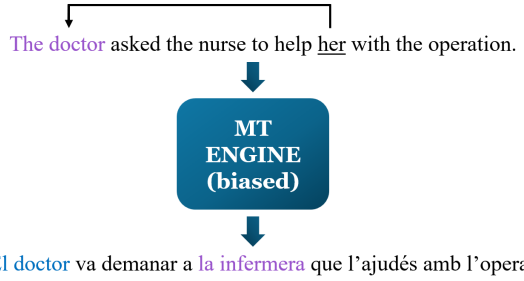

Figure 1: Example of Gender Bias in MT

In our study, we analyze gender bias in two distinct ways, which we will refer to as gender-bias tasks: *Gender Coreference Resolution* and *Gender Terms Detection*. In both tasks, models must utilize contextual gender information (i.e., gender cues) to accurately translate, providing the correct gender terms in the translation.

Gender Coreference Resolution In this task, we assess whether an MT engine correctly predicts the gender of a human entity in the translation based on its corresponding coreference pronoun in the source sentence. We address this task using POS tagging, focusing solely on the gender of specific human entities in the translation.

Gender Terms Detection In this other task, we evaluate whether an MT engine generates translations that include all the correct gender terms based on the gender cues of the source sentence. These clues for disambiguating gender terms include coreference pronouns, proper nouns, and semantic meaning, among others. Detection of the correct gender terms (or their incorrect counterparts) relies on textual comparison of reference terms.

Both gender-bias problems are approached as classification problems since they involve determining the correct gender labels, allowing for the derivation of typical ML scores. Devoid of gender context, we only pay attention to the proportion of male and female terms generated in the translations. As evaluation benchmarks, WinoMT and Gold BUG focus on *Gender Coreference Resolution*, whereas MuST-SHE in *Gender Terms Detection*. Check Figure 2 for an illustration of the gender-bias tasks just explained.

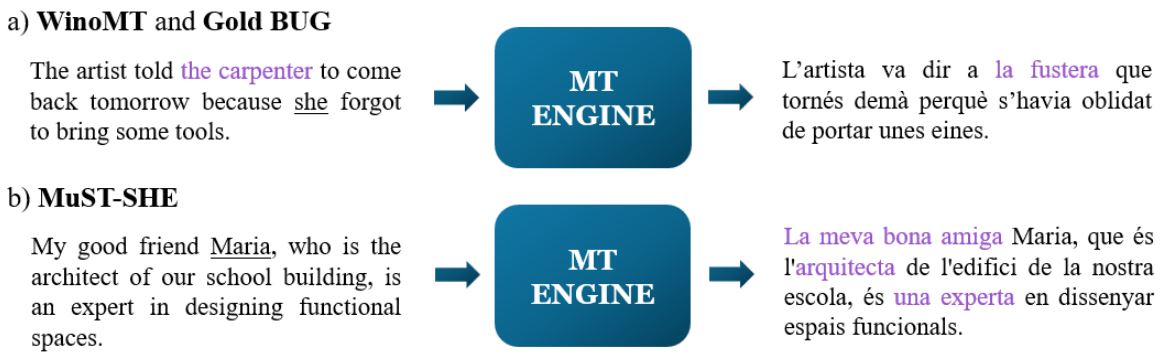

Figure 2: Examples of *Gender Coreference Resolution* (a) and *Gender Terms Detection* (b) in En → Ca

3 Related work

Large Language Models are advanced AI models designed to understand and generate language (Yang et al., 2023; Zhao et al., 2023). These models typically employ a decoder-only architecture and are characterized by their enormous size, often containing billions of parameters (Brown et al., 2020; Thoppilan et al., 2022; OpenAI, 2023). The scale and capacity of LLMs enable them to capture intricate linguistic nuances and handle a wide range of language-related tasks, despite not being explicitly trained for each specific task (Sun et al., 2023; Wei et al., 2023; Li et al., 2023; Gao et al., 2023a; Yao et al., 2023; Yang et al., 2022; Gao et al., 2023b; Ning et al., 2023). The training process for LLMs typically consists of two steps. First, they undergo self-supervised pretraining using vast amounts of text data, which allows them to develop a general understanding of language (i.e., base LLMs). Subsequently, they are fine-tuned on specific supervised tasks to specialize in various applications (Chung et al., 2022; Sanh et al., 2022). One of the key features of LLMs is the prompting mechanism. A prompt serves as the input or activation signal provided to the model. Through this input, we specify to the model the NLP task we want it to perform, such as translation in our case.

By leveraging the ability to guide the model with prompts, instruction-tuned LLMs are created (Zhang et al., 2023b). These are base LLMs that have undergone additional fine-tuning using datasets of instructions, containing explicit instructions or prompts to enhance their performance on various tasks. Instruction-tuning is a subsequent step that tailors the model's behavior and output according to specific instructions or guidelines (Mishra et al., 2022; Muennighoff et al., 2023; Longpre et al., 2023).

Moving away from LLMs, we find Neural Machine Translation models (Cho et al., 2014; Bahdanau et al., 2015; Luong et al., 2015; Johnson et al., 2017; Wu et al., 2016; Fan et al., 2020; NLLB Team, 2022). These models represent the state-of-the-art in machine translation, consistently achieving the highest translation performance. They typically leverage an encoder-decoder transformer (Vaswani et al., 2017) trained with parallel data in a supervised manner, intended solely for the task of translation. Unlike LLMs, NMT models are relatively smaller in size and present unique challenges when it comes to scaling (e.g., bidirectional processing, the attention mechanism complexity...). However, besides their size, the main distinction between LLMs and NMT models lies in the prompting method. LLMs necessitate a prompt to operate, making them entirely dependent on context. This is precisely the aspect we aim to explore: whether we can use the prompting mechanism, absent in NMT models, to alleviate gender bias.

To date, significant research has been conducted on the translation capabilities of LLMs, as extensively documented in the literature (Chowdhery et al., 2023; Jiao et al., 2023a; Zhu et al., 2023; Agrawal et al., 2023; Jiao et al., 2023b; Zhang et al., 2023a; Bawden and Yvon, 2023; Hendy et al., 2023). Furthermore, several efforts have been made to identify and address bias in LLMs (Ernst et al., 2023; Su et al., 2023; Cai et al., 2024). However, the exploration of gender bias in the realm of MT and LLMs remains relatively scarce. This encompasses (Sánchez et al., 2023), which sought to leverage LLMs for gender-specific translations, and (Vanmassenhove, 2024), which experimented

with En → It translation in ChatGPT, revealing how GPT models perpetuate biases even when explicitly prompted to provide alternative translations. Additionally, (Ghosh and Caliskan, 2023) examines bias between English and languages that exclusively use gender-neutral pronouns, and (Savoldi et al., 2024) demonstrate through extensive manual analysis the potential of GPT-4 to produce gender-neutral translations for En → It.

4 Methodology

4.1 Models

Llama-2-7B A base model that belongs to a family of state-of-the-art LLMs openly released by Meta (Touvron et al., 2023). This family of models outperforms open-source models on popular benchmarks and has demonstrated high efficacy and safety based on human evaluations. Llama-2-7B was trained on a combination of publicly available data, primarily in English. Catalan and Spanish (among other languages) were also included to a lesser extent. However, any use of the model in languages other than English is explicitly declared out of scope by the developers.

Ăguila-7B An open-source base LLM from Barcelona Supercomputing Center (BSC) that was trained on a combination of Spanish, Catalan, and English data, resulting in a total of 26 billion tokens. The model was built upon the Falcon-7B model, which is a highly advanced English language model.

Flor-6.3B Another publicly available base LLM tailored for Catalan, Spanish, and English, published by the BSC. This model is derived from the language adaptation process applied to Bloom-7.1B, involving adjustments to the vocabulary and embedding layer. Additionally, the model underwent continuous pre-training with 140 billion tokens specific to Catalan and Spanish.

M2M-100-1.2B A multilingual NMT model released by Meta in October 2020 (Fan et al., 2020) that can directly translate between the 9,900 directions of 100 languages, including our languages of interest (i.e., English, Catalan, and Spanish). It was considered the first AI model that could translate between 100 languages without relying on English.

NLLB-200-1.3B The following multilingual NMT model released by Meta in July 2022 (Costa-jussà et al., 2022) enabling translation across 200 languages, including less commonly spoken languages. It also incorporates the languages we are concentrating on, namely English, Catalan, and Spanish.

Mt-aina-en-ca The only parallel NMT model assessed in this work, functioning exclusively for English → Catalan translation. Developed at BSC, it was trained from scratch employing a combination of English-Catalan datasets consisting of approximately 11 million sentences.

Google Translate It is widely acknowledged in the literature as one of the leading translation models of today. This multilingual NMT model encompasses 133 languages, with English, Catalan and Spanish among them.

Llama-2-7B-chat It is the refined iteration of Llama-2-7B, optimized specifically for dialogue applications. This version underwent supervised instruction-tuning as well as Reinforcement Learning from Human Feedback (RLHF). Opting for this instructed version for the investigation into prompting is preferable over the base model, as it is more robust to prompt variations and better comprehends complex prompts and nuances. Selecting the base model along with its instructed version allows us to make insightful comparisons between these models.

4.2 Test sets

All test sets comprise English sentences (or paragraphs) aimed to be translated into either Catalan or Spanish. After obtaining translations in their respective grammatical languages, the evaluation frameworks are applied to derive the metrics (either MT or gender scores).

4.2.1 Machine Translation

FLoRes-200 It is a massively multilingual general domain dataset. Initially presented by (Guzmán et al., 2019; Goyal et al., 2021), it has been further developed and expanded by the (Goyal et al., 2022). The most recent version of this dataset encompasses 200 languages (NLLB Team, 2022). This dataset[1] includes two subsets: FLoRes-200 dev (997) and FLoRes-200 devtest (1,012).

4.2.2 Gender Bias

WinoMT Developed by (Stanovsky et al., 2019), this test set is intended to evaluate the presence of

[1] https://github.com/facebookresearch/flores/tree/main/flores200

gender bias in translations from English to various gender-inflected languages. The corpus[2] consists of 3,888 sentences in the schema of Winograd. Each sentence in the corpus presents two human entities defined by their roles, along with a subsequent pronoun that must be correctly resolved to one of the entities (Levesque et al., 2012). One of the main limitations of this dataset is its synthetic nature, as it is built on templates.

Gold BUG The previous limitation of WinoMT could be addressed through the introduction of BUG[3] (Levy et al., 2021), the first publicly accessible large-scale corpus designed for gender-bias evaluation, comprising 108,000 real-world English sentences. BUG was built by crawling text according to specific syntactic patterns, offering a more diverse and realistic dataset than WinoMT. The Gold BUG version used in our evaluation consists of a gold-quality, human-validated set extracted from BUG, totaling 1,717 instances.

MuST-SHE This test set, initially introduced by (Bentivogli et al., 2020) for English-French, English-Italian, and English-Spanish, serves as a valuable benchmark for evaluating gender bias in the context of speech translation. This dataset[4] is constructed using TED talks data, as described by (Cattoni et al., 2021), lending it a more natural and realistic tone. Recently, (Mash et al., 2024) created an English-Catalan[5] version of the dataset tailored for the machine translation domain, resulting in 1,046 sentences. For our analysis, we adapted the original English-Spanish version for machine translation following the same steps as in the Catalan version, resulting in 1,164 instances. Both datasets, English-Spanish and English-Catalan, contain two types of instances: those with and without cues to disambiguate the gender of certain terms. In instances where gender cues are present, the task to be addressed is *Gender Terms Detection*; otherwise, we are solely interested in the proportion of male and female terms generated in the translations.

Furthermore, both WinoMT and Gold BUG contain pro- and anti-stereotypical sets based on US labor statistics (Zhao et al., 2018). A pro-stereotypical set comprises sentences with stereotypical gender-role assignments (e.g., male doctors, female housekeepers), while an anti-stereotypical set includes sentences with non-stereotypical gender-role assignments (e.g., female doctors, male housekeepers). These sets facilitate the investigation of whether the translation performance of models correlates with gender stereotypes. Specifically, they help determine whether models exhibit better or worse gender scores when translating sentences that align (or do not align) with their pre-established biases.

4.3 Metrics

4.3.1 Machine Translation

To measure the MT capabilities of the models, we employ two widely-used metrics: BLEU (Papineni et al., 2002), which is based on comparing n-grams and is computed using the SacreBLEU library[6] (Post, 2018), and COMET (Rei et al., 2020), a more recent metric that relies on sentence embeddings.

4.3.2 Gender Bias

When the source sentence contains gender cues to disambiguate the gender of certain terms, meaning we have a known gender reference or ground truth, the translation problem is treated as a typical classification task. Consequently, in the context of gender bias, we evaluate models using Gender Accuracy (in %), F1-male, and F1-female scores. For WinoMT and Gold BUG these scores are computed directly[7], whereas for MuST-SHE we obtain first the confusion matrix[8] and then we compute the scores using scikit-learn library. Additionally, we get standard metrics such as ΔG, which indicates the performance difference between correctly predicting male and female terms, and ΔS, which requires both a pro- and anti-stereotypical sets to assess whether a model relies on gender-stereotypes to generate translations. Conversely, when no gender cues are available in the source sentence, we simply analyze the proportion of predicted male and female terms in the translations.

[2] https://github.com/gabrielStanovsky/mt_gender

[3] https://github.com/SLAB-NLP/BUG

[4] https://mt.fbk.eu/must-she/

[5] https://huggingface.co/datasets/projecte-aina/MuST-SHE_en-ca

[6] Version 1.5.1

[7] https://github.com/gabrielStanovsky/mt_gender/blob/master/scripts/evaluate_all_languages.sh

[8] https://github.com/audreyvm/tfm_gender_bias/blob/main/mustshe_acc_v1.2.py

5 Benchmarking

5.1 Prompting LLMs

In our benchmarking, we employ a 5-shot approach for our LLMs. This ensures that the LLMs better comprehend the requested task (i.e., machine translation) and potentially produce higher-quality translations, as demonstrated in existing literature (Vilar et al., 2023; Garcia et al., 2023; Zhang et al., 2023c). Additionally, during our experimentation with prompts, we observe that incorporating the language label followed by a colon (e.g., "English:", "Catalan:","Spanish:") before the sentence to be translated and its corresponding translation is an effective strategy for our LLMs. Furthermore, *beginning* and *end of sentence* tokens are added to delimit the source and translation examples in the shots, enhancing the models' understanding and facilitating the extraction of the output translations. Flor-6.3B and Llama-2-7B work with "<BOS>" and "<EOS>", while Ăguila-7B uses "<s>" and "</s>".

When evaluating the FLoRes-200 dev set, we use 5 shots from the FLoRes-200 devtest set in the prompt. Conversely, when assessing the FLoRes-200 devtest set, we incorporate 5 shots from the FLoRes-200 dev set into the prompt. For the remaining gender-bias test sets (WinoMT, Gold BUG, and MuST-SHE), we utilize the same prompt employed during testing of the FLoRes-200 dev test set, consisting of the same 5 shots from the FLoRes-200 devtest. When selecting these 5 instances to serve as shots, we ensure diversity in content, length, and structure to provide a broader range of examples to the model. The specific prompts created are detailed in Section C of the Appendix.

5.2 Configurations

Since we are only performing inference, we adjust only two parameters: the *top_k*, which is set to 1 to ensure a deterministic process, and the limit of *maximum tokens* to generate, which is adjusted depending on the test sets. We use greedy decoding for all models since beam search in LLMs demands significant time and resources. These choices are made to ensure the comparability of the results.

5.3 Key takeaways

Based on the benchmarking evaluation, the following findings emerge:

- Base LLMs fall behind NMT models in terms of MT in both En → Ca and En → Es directions (check Table 1 to see the results).
- All models exhibit gender bias in the assessed directions, with LLMs showing a more pronounced bias compared to NMT models (check Tables 3, 4, and 5).
- The performance of all studied models correlates with gender stereotypes, achieving better gender metrics for the pro-stereotypical set rather than the anti-stereotypical set (check Section D in the Appendices).
- In the absence of contextual gender cues, all models predict mostly male terms (∼75%-94%). The corresponding (∼6%-25%) mainly relates to female-stereotypical examples (check Section E in the Appendices).

6 Gender Bias mitigation through prompting

After observing that LLMs exhibit more gender bias than NMT models, we found it necessary to address this bias in LLMs. Consequently, we have chosen to leverage prompting, as it is a distinctive feature of these models. Therefore, the second stage of our work involves conducting exploratory research in a trial-and-error manner, aiming to identify a prompt that effectively mitigates bias in LLMs. For this experiment, we have selected the instruction-tuned model Llama-2-7B-chat since it is more robust to complex prompts than its base version. In addition, in this stage, we have decided to focus solely on the *Gender Coreference Resolution* task. Ideally, our goal is to narrow the gap in gender scores with respect to NMT models, as this would represent a significant breakthrough.

The procedure goes as follows: Initially, we develop a range of prompts based on strategies outlined in the literature, including few-shot prompting, context-supplying, and chain-of-thought instructions. To assess the impact of these prompts, we test them on WinoMT and obtain gender-bias scores for each prompt. Thereafter, the prompt that demonstrates the most considerable reduction in bias on WinoMT, as indicated by numerical gender-bias scores, is evaluated on the remaining test sets (Gold BUG, MuST-SHE, and FLoRes-200). By doing so, we want to determine: firstly, the prompt's generalizability across the remaining gender-bias test sets, and secondly, if it affects the overall machine translation capabilities of the LLM.

		English → Catalan				English → Spanish			
		DEV		DEVTEST		DEV		DEVTEST	
		BLEU	**COMET**	**BLEU**	**COMET**	**BLEU**	**COMET**	**BLEU**	**COMET**
NMT	Google Translate	**45.1**	**0.8838**	**46.0**	**0.8811**	**29.6**	**0.8737**	**30.1**	**0.8724**
	NLLB-200-1.3B	38.7	0.8645	38.6	0.8626	27.2	0.8591	27.7	0.8578
	M2M-100-1.2B	40.1	0.8687	40.4	0.8623	25.6	0.8450	25.4	0.8422
	Mt-aina-en-ca	43.0	0.8735	43.9	0.8730	-	-	-	-
LLM	Ăguila-7B	29.1	0.8359	30.3	0.8368	18.2	0.8212	19.5	0.8198
	Flor-6.3B	**37.9**	**0.8641**	**39.6**	**0.8680**	**23.8**	**0.8498**	**25.5**	**0.8528**
	Llama-2-7B	31.6	0.8443	32.9	0.8458	23.3	0.8486	23.5	0.8454
	Llama-2-7B-chat	**30.1**	**0.8284**	**29.9**	**0.8250**	**22.6**	**0.8427**	**22.9**	**0.8423**
	Llama-2-7B-chat (GB prompt)	27.6	0.8176	28.4	0.8140	22.4	0.8251	21.8	0.8277

Table 1: BLEU and COMET scores for FLoRes-200

6.1 Baseline

Before embarking on the search for the prompt, it is essential to establish a baseline for Llama-2-7B-chat. Therefore, we use the same prompt employed in the benchmarking, with minor adaptations necessary for Llama-2-7B-chat, such as the use of special tags («SYS», [INST]...). The resulting prompt after the adjustments is detailed in Section F of the Appendices. With this prompt, we obtain MT and gender-bias scores across the four test sets. Refer to Tables 1, 3, 4, and 5 to observe the results. These initial results offer valuable insights, revealing that the instructed version (Llama-2-7B-chat) achieves lower MT scores compared to its base model (Llama-2-7B) for both directions.

6.2 Crafting and testing prompts on WinoMT

After conducting several experiments using Llama-2-7B-chat, we proceeded to curate and test multiple prompts on the WinoMT test set. The curated prompts in detail can be found in section G of the Appendices. We recommend consulting them for a comprehensive understanding of this section.

In the design of all our prompts, we incorporated the 5-shot strategy already used in the benchmarking and the baseline. However, we substituted the FLoRes-200 examples and introduced additional modifications to the curated prompts.

A significant aspect of our crafted prompts involves the inclusion of translation examples that encompass more gender-related phenomena compared to the ones from the FLoRes-200 dataset, which comprises mainly gender-neutral or impersonal sentences. Specifically, one of our curated prompts included examples from the MuST-SHE dataset, while in another prompt, we intentionally created five sentences (or rather, translations) adhering to a Winograd structure, wherein each sentence comprises two human entities and one pronoun used to disambiguate one of them. These crafted translations were deliberately designed to contain more female representation and anti-stereotypical content. These invented translations are provided in section H of the Appendices.

For another prompt, in addition to including 5 shots from MuST-SHE, we also adopted an approach that involved providing more contextual information to the model. We explicitly stated the objective of translating while simultaneously reducing gender bias. By offering this additional context, the model should gain a clearer understanding of the goal to mitigate gender bias and the factors it should consider to do so effectively.

Afterwards, we adopted a chain-of-thought strategy for the remaining curated prompts, each following again a 5-shot structure. We integrated the previously crafted Winograd examples into these prompts. Two of them resulted in complex and detailed chain-of-thought prompts, incorporating all the necessary steps and reasoning that the model should do to carefully solve the *Gender Coreference Task* and provide a correct translation. The only distinction between these two complex

		English → Catalan				English → Spanish			
Model	**Examples from:**	**G Acc**	**F1-male**	**F1-female**	**ΔG**	**G Acc**	**F1-male**	**F1-female**	**ΔG**
Llama-2-7B	FLoRes-200	48.0	62.6	36.4	26.2	53.1	64.9	41.3	23.6
Llama-2-7B-chat	FLoRes-200	46.4	61.4	35.5	25.9	53.3	65.3	41.6	23.7
	MuST-SHE	46.9	60.6	39.4	21.2	49.9	62.7	35.4	27.3
	Invented Winograd examples	46.6	58.8	43.2	15.6	49.8	61.8	37.5	24.3
	MuST-SHE + context on Gender Bias issue	46.9	60.0	40.7	19.3	50.6	62.6	38.7	23.9
	Invented Winograd examples + chain-of-thought ("agent")	55.2	65.8	56.7	9.1	60.6	69.7	56.8	12.9
	Invented Winograd examples + chain-of-thought ("human entity")	54.5	65.2	55.3	9.9	59.5	68.9	54.8	14.1
	Invented Winograd examples + SHORT chain-of-thought	**58.8**	**68.9**	**60.5**	**8.4**	**65.0**	**73.3**	**63.6**	**9.7**

Table 2: WinoMT scores using different prompting techniques for En → Ca and En → Es

prompts was the terminology used to refer to the human entities in the examples, either as "human entity" or "agent".

Finally, we constructed another chain-of-thought prompt that yielded the best results. In this prompt, the steps were significantly simplified compared to the previous two prompts. Here, explicit instructions of the steps were not included, and instead, schematic steps accompanied by arrows were provided in the shots.

For a comprehensive summary of the results obtained on WinoMT for all these prompts, please consult Table 2.

6.3 Top-performing prompt

The resulting top-performing prompt on WinoMT is the one named *Invented Winograd examples + SHORT chain-of-thought* from Table 2. With this prompt, we have achieved remarkable increases of 12.4 (En → Ca) and 11.7 (En → Es) on WinoMT compared to the baseline. In short, this prompt follows a simplified chain-of-thought approach with 5-shots on anti-stereotypical content and increased female representation. The examples in the prompt were invented following the Winograd sentence structure, designed to address gender coreference.

The phrase "Proceed step by step" is also included before the shots. In the initial experiments, we observed that incorporating this sentence led to the model providing a more structured response. Based on this observation, we replicated the same pattern generated by the LLM in our crafted shots.

7 Results

After testing our top-performing prompt on the remaining gender-bias test sets, Gold BUG and MuST-SHE, we observe a significant reduction in gender bias within those test sets too. These results are detailed in sections A and B of the Appendices. Subsequently, all the three Tables 3, 4, and 5 demonstrate a remarkable improvement in gender-bias scores, significantly reducing the upper bound in each test set compared to the best NMT model. This places the LLM on par with NMT models in terms of gender bias manifestation. For example, on the WinoMT test set, the model achieves the second-best position in En → Ca and the third-best position in En → Es. In MuST-SHE, the mitigation is less pronounced as this test set also encompasses other gender-related tasks, unlike WinoMT and Gold BUG, which focus solely on *Gender Coreference Resolution*.

Regarding the MT metrics, we observe a small loss compared to the baseline when testing on

		English → Catalan					English → Spanish				
		G Acc	**F1-male**	**F1-female**	**ΔG**	**ΔS**	**G Acc**	**F1-male**	**F1-female**	**ΔG**	**ΔS**
NMT	Google Translate	57.1	67.5	55.6	11.9	**23.9**	**70.9**	**76.6**	**74.4**	**2.2**	**24.3**
	NLLB-200-1.3B	**60.9**	**70.1**	**64.0**	**6.1**	28.1	67.2	74.0	68.9	5.1	33.9
	M2M-100-1.2B	51.5	64.2	44.6	19.6	24.6	57.9	68.6	50.4	18.2	26.5
	Mt-aina-en-ca	48.9	63.1	37.9	25.2	27.3	-	-	-	-	-
LLM	Ăguila-7B	46.1	60.4	34.5	**25.9**	36.1	49.3	63.3	32.5	30.8	**28.4**
	Flor-6.3B	47.7	62.2	35.2	27.0	33.1	**53.4**	**65.1**	**42.5**	**22.6**	30.1
	Llama-2-7B	**48.0**	**62.6**	**36.4**	26.2	**32.8**	53.1	64.9	41.3	23.6	33.1
	Llama-2-7B-chat	46.4	61.4	35.5	25.9	33.1	53.3	65.3	41.6	23.7	32.0
	Llama-2-7B-chat (GB prompt)	**58.8**	**68.9**	**60.5**	**8.4**	**27.8**	**65.0**	**73.3**	**63.6**	**9.7**	**22.1**

Table 3: WinoMT gender scores

FLoRes-200 (Table 1).

8 Discussion

Initially, we believed that reducing gender bias through prompting would possibly be straightforward. However, it was surprising to find that the model only began effectively mitigating the bias after implementing the chain-of-thought approach. In fact, the results presented in Table 2 demonstrate that without the chain-of-thought approach and relying solely on the same invented Winograd examples from the top-performing prompt, no improvement was observed. Furthermore, we noticed that describing the problem of gender bias or including MuST-SHE examples did not lead to any improvement. Additionally, we observed that the Llama-2-7B-chat model comprehends and responds better to schematic chain-of-thought prompts compared to highly detailed and elaborate prompts, resulting in higher gender scores in the former case. Besides, the inclusion of the phrase "Proceed step by step" seems to be beneficial.

Fortunately, after identifying our successful prompt, we can confidently affirm that leveraging prompting can indeed serve as an effective method to mitigate gender bias in an instructed LLM (at least, for *Gender Coreference Resolution*).

9 Conclusions

This work investigates gender bias in the translation outputs generated by various LLMs through two distinct approaches. Firstly, by benchmarking three base models (Ăguila-7B, Flor-6.3B and Llama-2-7B) using different gender-bias test sets and comparing the results with state-of-the-art NMT models (M2M-100-1.2B, NLLB-200-1.3B, Mt-aina-en-ca, and Google Translate). Secondly, by experimenting with the prompting mechanism of an instruction-tuned LLM (Llama-2-7B-chat) and trying to mitigate its gender bias in the output. This study is done in the En → Ca and En → Es directions.

Results reveal the presence of gender bias across all models, with base LLMs exhibiting more gender bias than NMT models. Moreover, the performance of all models correlates with gender stereotypes. In the absence of gender cues in the source sentence, they tend to generate predominantly male terms, while female terms are generated primarily when encountering female-stereotypical content. To mitigate this bias, prompting engineering techniques have been implemented in an instruction-tuned LLM. After curating and testing several prompts, one prompt was identified that resulted in a significant reduction in gender bias, achieving impressive gender scores. The prompt follows a simplified chain-of-thought approach with 5-shots relying on anti-stereotypical content and increased female representation. This prompt enables the instructed LLM to perform competitively in terms of gender scores, achieving results comparable to NMT models and even surpassing some of them. However, it is observed that using this prompt leads to a slight loss in the translation quality.

10 Ethical statement

In this evaluation, we have only focused on the binary male and female genders, without considering other gender identities. Additional experiments on new datasets would be required to assess the performance of these methods on non-binary scenarios.

About the proposed definition of gender bias, we tried to characterize different aspects of the problem. Even though we recognize that it is a complex problem and our metrics and experiments focus only on some specific manifestations.

11 Acknowledgments

This research has been promoted and financed by the Government of Catalonia through the Aina project, by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia (Funded by EU – NextGenerationEU within the framework of the project ILENIA with reference 2022/TL22/00215337, 2022/TL22/00215336, 2022/TL22/00215335, 2022/TL22/00215334). It has also been supported by the Horizon Europe program [Grant Number 101135916] and by DeepR3 (TED2021-130295B-C32) (Funded by MCIN/AEI/10.13039/501100011033 and European Union NextGeneration EU/PRTR).

References

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

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Rachel Bawden and François Yvon. 2023. Investigating the translation performance of a large multilingual language model: the case of BLOOM. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation, EAMT 2023, Tampere, Finland, 12-15 June 2023*, pages 157–170. European Association for Machine Translation.

Luisa Bentivogli, Beatrice Savoldi, Matteo Negri, Mattia A. Di Gangi, Roldano Cattoni, and Marco Turchi. 2020. Gender in danger? evaluating speech translation technology on the MuST-SHE corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6923–6933, Online. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Yuchen Cai, Ding Cao, Rongxi Guo, Yaqin Wen, Guiquan Liu, and Enhong Chen. 2024. Locating and mitigating gender bias in large language models. *CoRR*, abs/2403.14409.

Roldano Cattoni, Mattia Antonino Di Gangi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Must-c: A multilingual corpus for end-to-end speech translation. *Computer Speech and Language*, 66:101155.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings of SSST@EMNLP 2014, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, Doha, Qatar, 25 October 2014*, pages 103–111. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang,

Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

Marta R. Costa-jussà, Carlos Escolano, Christine Basta, Javier Ferrando, Roser Batlle, and Ksenia Kharitonova. 2020. Gender bias in multilingual neural machine translation: The architecture matters. *CoRR*, abs/2012.13176.

Kate Crawford. 2017. The trouble with bias. in conference on neural information processing systems (nips) – keynote, long beach, usa.

Lisa Irmen Dagmar Stahlberg, Friederike Braun and Sabine Sczesny. 2007. Representation of the sexes in language. social communication, pages 163–187.

Jasmina S. Ernst, Sascha Marton, Jannik Brinkmann, Eduardo Vellasques, Damien Foucard, Martin Kraemer, and Marian Lambert. 2023. Bias mitigation for large language models using adversarial learning. In *Proceedings of the 1st Workshop on Fairness and Bias in AI co-located with 26th European Conference on Artificial Intelligence (ECAI 2023), Kraków, Poland, October 1st, 2023*, volume 3523 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. *CoRR*, abs/2010.11125.

Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Trans. Inf. Syst.*, 14(3):330–347.

Jun Gao, Huan Zhao, Changlong Yu, and Ruifeng Xu. 2023a. Exploring the feasibility of chatgpt for event extraction. *CoRR*, abs/2303.03836.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023b. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 6465–6488. Association for Computational Linguistics.

Xavier Garcia, Yamini Bansal, Colin Cherry, George F. Foster, Maxim Krikun, Melvin Johnson, and Orhan Firat. 2023. The unreasonable effectiveness of few-shot learning for machine translation. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 10867–10878. PMLR.

Sourojit Ghosh and Aylin Caliskan. 2023. Chatgpt perpetuates gender bias in machine translation and ignores non-gendered pronouns: Findings across bengali and five other low-resource languages. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2023, Montréal, QC, Canada, August 8-10, 2023*, pages 901–912. ACM.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2021. The FLORES-101 evaluation benchmark for low-resource and multilingual machine translation. *CoRR*, abs/2106.03193.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are GPT models at machine translation? A comprehensive evaluation. *CoRR*, abs/2302.09210.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023a. Is chatgpt A

good translator? A preliminary study. *CoRR*, abs/2301.08745.

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

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Trans. Assoc. Comput. Linguistics*, 5:339–351.

Evan Hazenberg Lal Zimman and Miriam Meyerhoff. 2017. Trans people's linguistic self-determination and the dialogic nature of identity. representing trans: Linguistic, legal and everyday perspectives, pages 226–248.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Thirteenth International Conference, KR 2012, Rome, Italy, June 10-14, 2012*. AAAI Press.

Roger J. R. Levesque. 2011. *Sex Roles and Gender Roles*, pages 2622–2623. Springer New York, New York, NY.

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

Bo Li, Gexiang Fang, Yang Yang, Quansen Wang, Wei Ye, Wen Zhao, and Shikun Zhang. 2023. Evaluating chatgpt's information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness. *CoRR*, abs/2304.11633.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 22631–22648. PMLR.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1412–1421. The Association for Computational Linguistics.

Audrey Mash, Carlos Escolano, Aleix Sant, Maite Melero, and Francesca de Luca Fornaciari. 2024. Unmasking biases: Exploring gender bias in English-Catalan machine translation through tokenization analysis and novel dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17144–17153, Torino, Italia. ELRA and ICCL.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3470–3487. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M. Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15991–16111. Association for Computational Linguistics.

Munan Ning, Yujia Xie, Dongdong Chen, Zeyin Song, Lu Yuan, Yonghong Tian, Qixiang Ye, and Li Yuan. 2023. Album storytelling with iterative story-aware captioning and large language models. *CoRR*, abs/2305.12943.

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

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

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

Parmy Olson. 2018. The algorithm that helped google translate become sexist. https://www.forbes.com/sites/parmyolson/.

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

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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

Isabelle Régner, Catherine Thinus-Blanc, Agnès Netter, Toni Schmader, and Pascal Huguet. 2019. Committees with implicit biases promote fewer women when they do not believe gender bias exists. *Nature Human Behaviour*, 3.

Eduardo Sánchez, Pierre Andrews, Pontus Stenetorp, Mikel Artetxe, and Marta R. Costa-jussà. 2023. Gender-specific machine translation with large language models. *CoRR*, abs/2309.03175.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

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. 2021. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.

Beatrice Savoldi, Andrea Piergentili, Dennis Fucci, Matteo Negri, and Luisa Bentivogli. 2024. 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.

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

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Hsuan Su, Cheng-Chu Cheng, Hua Farn, Shachi H. Kumar, Saurav Sahay, Shang-Tse Chen, and Hung-yi Lee. 2023. Learning from red teaming: Gender bias provocation and mitigation in large language models. *CoRR*, abs/2310.11079.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text classification via large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 8990–9005. Association for Computational Linguistics.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Kathleen S. Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Agüera y Arcas, Claire Cui, Marian Croak, Ed H. Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications. *CoRR*, abs/2201.08239.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura,

Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George F. Foster. 2023. Prompting palm for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15406–15427. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, Yong Jiang, and Wenjuan Han. 2023. Zero-shot information extraction via chatting with chatgpt. *CoRR*, abs/2302.10205.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. 2023. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *CoRR*, abs/2304.13712.

Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. 2022. Re3: Generating longer stories with recursive reprompting and revision. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 4393–4479. Association for Computational Linguistics.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023a. Prompting large language model for machine translation: A case study. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 41092–41110. PMLR.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2023b. Instruction tuning for large language models: A survey. *CoRR*, abs/2308.10792.

Xuan Zhang, Navid Rajabi, Kevin Duh, and Philipp Koehn. 2023c. Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with QLoRA. In *Proceedings of the Eighth Conference on Machine Translation*, pages 468–481, Singapore. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *CoRR*, abs/2304.04675.

Appendices

A Gender Scores on Gold BUG

		English → Catalan					English → Spanish				
		G Acc	**F1-male**	**F1-female**	**ΔG**	**ΔS**	**G Acc**	**F1-male**	**F1-female**	**ΔG**	**ΔS**
NMT	Google Translate	**62.3**	**77.5**	56.9	20.6	26.7	47.5	62.8	55.0	**7.8**	**14.9**
	NLLB-200-1.3B	62.1	77.4	**57.9**	**19.5**	**13.6**	**65.2**	**79.4**	**61.9**	17.5	20.4
	M2M-100-1.2B	60.4	76.3	49.8	26.5	23.9	63.8	78.5	56.5	22.0	22.7
	Mt-aina-en-ca	60.3	76.4	51.2	25.2	20.5	-	-	-	-	
LLM	Ăguila-7B	54.5	71.7	**43.9**	**27.8**	22.5	58.8	75.2	**46.5**	**28.7**	18.8
	Flor-6.3B	**57.8**	74.5	43.0	31.5	**18.6**	**61.2**	**77.1**	46.2	30.9	**14.9**
	Llama-2-7B	57.7	**74.9**	37.1	37.8	18.7	60.2	76.9	37.4	39.5	16.1
	Llama-2-7B-chat	57.8	74.5	39.3	35.2	25.3	58.9	75.6	37.0	38.6	16.9
	Llama-2-7B-chat (GB prompt)	**59.8**	**75.0**	**58.7**	**16.3**	**15.4**	**63.7**	**78.5**	**58.9**	**19.6**	**18.1**

Table 4: Gold BUG gender scores

B Gender Scores on MuST-SHE

		English → Catalan				English → Spanish			
		G Acc	**F1-male**	**F1-female**	**ΔG**	**G Acc**	**F1-male**	**F1-female**	**ΔG**
NMT	Google Translate	89.5	90.6	88.0	2.6	95.1	95.5	94.7	0.8
	NLLB-200-1.3B	**93.3**	**93.7**	**92.7**	**1.0**	**96.0**	**96.2**	**95.8**	**0.5**
	M2M-100-1.2B	84.4	86.6	81.4	5.2	87.4	89.2	84.8	4.3
	Mt-aina-en-ca	87.1	88.5	85.4	3.1	-	-	-	-
LLM	Ăguila-7B	87.1	88.5	85.4	3.1	92.2	93.0	91.0	2.0
	Flor-6.3B	89.6	90.7	88.2	2.5	93.3	93.9	92.4	1.5
	Llama-2-7B	**91.1**	**91.9**	**90.0**	**1.8**	**95.1**	**94.5**	**93.2**	**1.3**
	Llama-2-7B-chat	88.1	89.7	86.0	3.7	91.0	92.0	89.6	2.4
	Llama-2-7B-chat (GB prompt)	**88.4**	**89.8**	**86.5**	**3.3**	**92.0**	**92.6**	**91.4**	**1.2**

Table 5: MuST-SHE gender scores

C Prompts employed in the Benchmarking

The prompts employed with Ăguila-7B when testing FLoRes-200 devtest set for En → Ca and En → Es respectively:

```
Translate the following sentence from English to Catalan:
English: <s>Hangeul is the only purposely invented alphabet in popular daily
use. The alphabet was invented in 1444 during the reign of King Sejong
(1418-1450).</s>
Catalan: <s>El hangul és l'únic alfabet creat arbitràriament que té un ús
estès en la vida diària. L'alfabet es va inventar l'any 1444 durant el regnat
de King Sejong (1418-1450).</s>
English: <s>They also said in a statement, "The crew is currently working to
determine the best method of safely extracting the ship".</s>
Catalan: <s>També han dit en un comunicat, "La tripulació treballa ara
mateix per a determinar la millor tècnica per a extreure la nau de manera
segura".</s>
English: <s>This is becoming less of an issue as lens manufacturers achieve
higher standards in lens production.</s>
Catalan: <s>Això és cada vegada menys important perquè els fabricants de lents
estan assolint estàndards més elevats en la producció de lents.</s>
English: <s>While assessing the successes and becoming aware of failures,
individuals and the whole of the participating persons discover more deeply
the values, mission, and driving forces of the organization.</s>
Catalan: <s>Mentre confirmen els èxits i prenen consciència dels fracassos,
els individus i el grup de participants descobreixen més profundament els
valors, la missió i les forces motrius de l'organització.</s>
English: <s>Entering Southern Africa by car is an amazing way to see all the
region's beauty as well as to get to places off the normal tourist routes.</s>
Catalan: <s>Entrar a l'Àfrica del Sud en cotxe és una forma impressionant
de veure tota la bellesa de la regió i d'arribar a llocs fora de les rutes
turístiques més habituals.</s>
English: <s>____sentence_to_translate____</s>
Catalan: <s>
```

```
Translate the following sentence from English to Spanish:
English: <s>Hangeul is the only purposely invented alphabet in popular daily
use. The alphabet was invented in 1444 during the reign of King Sejong
(1418–1450).</s>
Spanish: <s>El alfabeto coreano es el único diseñado en forma deliberada que
aún se utiliza a diario popularmente. Se inventó en 1444, durante el reinado
de Sejong (1418 a 1450).</s>
English: <s>They also said in a statement, "The crew is currently working to
determine the best method of safely extracting the ship".</s>
Spanish: <s>También se dijo en un comunicado que: «La tripulación se encuentra
actualmente trabajando para decidir cuál es el método más seguro para extraer
el barco».</s>
English: <s>This is becoming less of an issue as lens manufacturers achieve
higher standards in lens production.</s>
Spanish: <s>Este problema cada vez es menos importante gracias a que los
fabricantes de lentes logran estándares más altos en su producción.</s>
English: <s>While assessing the successes and becoming aware of failures,
individuals and the whole of the participating persons discover more deeply
the values, mission, and driving forces of the organization.</s>
Spanish: <s>Durante el proceso de análisis de los éxitos y toma de conciencia
de los fracasos, los individuos y grupos de personas involucrados descubren
con mayor profundidad los valores, el objetivo y las fuerzas que impulsan a la
organización.</s>
English: <s>Entering Southern Africa by car is an amazing way to see all the
region’s beauty as well as to get to places off the normal tourist routes.</s>
Spanish: <s>Una fantástica forma de contemplar todo el encanto de la región
del sur África es ingresar en automóvil, lo que, a su vez, le permitirá
acceder a lugares fuera de las rutas turísticas habituales.</s>
English: <s>____sentence_to_translate____</s>
Spanish: <s>
```

The prompts employed with Ăguila-7B when testing FLoRes-200 dev set, WinoMT, Gold BUG and MuST-SHE for En → Ca and En → Es were:

```
Translate the following sentence from English to Catalan:
English: <s>The feathers' structure suggests that they were not used in flight
but rather for temperature regulation or display. The researchers suggested
that, even though this is the tail of a young dinosaur, the sample shows adult
plumage and not a chick's down.</s>
Catalan: <s>L'estructura de les plomes fa pensar que no s'usaven per a volar
sinó per a regular la temperatura o per a exhibir-se. Els investigadors han
suggerit que, tot i que es tracta de la cua d'un dinosaure jove, la mostra
presenta el plomatge d'un adult i no d'un pollet.</s>
English: <s>They found the Sun operated on the same basic principles as other
stars: The activity of all stars in the system was found to be driven by their
luminosity, their rotation, and nothing else.</s>
Catalan: <s>Han descobert que el Sol funcionava sota els mateixos principis
bàsics que altres estrelles: s'ha vist que l'activitat de totes les estrelles
del sistema depèn de llur brillantor, llur rotació i res més.</s>
English: <s>The speeds of 802.11n are substantially faster than that of its
predecessors with a maximum theoretical throughput of 600Mbit/s.</s>
Catalan: <s>Les velocitats de 802.11n són substancialment més ràpides que les
dels seus predecessors amb un rendiment teòric màxim de 600Mbit/s.</s>
English: <s>Over four million people went to Rome to attend the funeral.</s>
Catalan: <s>Més de quatre milions de persones van anar a Roma per a assistir
al funeral.</s>
English: <s>Mrs. Kirchner announced her intention to run for president at the
Argentine Theatre, the same location she used to start her 2005 campaign for
the Senate as member of the Buenos Aires province delegation.</s>
Catalan: <s>La Sra. Kirchner va anunciar la seva intenció de presentar-se a la
presidència al Teatre de l'Argentina, el mateix lloc on va engegar la campanya
al Senat de 2005 com a membre de la delegació provincial de Buenos Aires.</s>
English: <s>____sentence_to_translate____</s>
Catalan: <s>
```

```
Translate the following sentence from English to Spanish:
English: <s>The feathers' structure suggests that they were not used in flight
but rather for temperature regulation or display. The researchers suggested
that, even though this is the tail of a young dinosaur, the sample shows adult
plumage and not a chick's down.</s>
Spanish: <s>La estructura que presenta el plumaje sugiere que su función
no estaba relacionada con el vuelo, sino que las usaban para regular la
temperatura o como indicador de la misma. Los investigadores sostienen que,
aunque se trata de la cola de un dinosaurio joven, la muestra analizada
presenta rasgos del plumaje de un adulto y no de un polluelo.</s>
English: <s>They found the Sun operated on the same basic principles as other
stars: The activity of all stars in the system was found to be driven by their
luminosity, their rotation, and nothing else.</s>
Spanish: <s>Se descubrió que el sol se regía por los mismos principios básicos
que otras estrellas: los únicos factores que impulsaban su actividad dentro
del sistema eran su luminosidad y su rotación.</s>
English: <s>The speeds of 802.11n are substantially faster than that of its
predecessors with a maximum theoretical throughput of 600Mbit/s.</s>
Spanish: <s>Las velocidades del estándar 802.11n son mucho más altas que las
alcanzadas por los que lo precedieron, con un rendimiento teórico máximo de
600 Mbps.</s>
English: <s>Over four million people went to Rome to attend the funeral.</s>
Spanish: <s>Más de cuatro millones de individuos se concentraron en Roma para
presenciar el funeral.</s>
English: <s>Mrs. Kirchner announced her intention to run for president at the
Argentine Theatre, the same location she used to start her 2005 campaign for
the Senate as member of the Buenos Aires province delegation.</s>
Spanish: <s>El Teatro Argentino fue el lugar donde la señora Kirchner anunció
su intención de candidatearse como presidenta; este es el mismo sitio donde
inició su campaña para el senado en el año 2005, en representación de la
provincia de Buenos Aires.</s>
English: <s>____sentence_to_translate____</s>
Spanish: <s>
```

The prompts employed with Flor-6.3B and Llama-2-7B when testing FLoRes-200 devtest set for En → Ca and En → Es respectively:

```
Translate the following sentence from English to Catalan:
English: <BOS>Hangeul is the only purposely invented alphabet in popular
daily use. The alphabet was invented in 1444 during the reign of King Sejong
(1418-1450).<EOS>
Catalan: <BOS>El hangul és l'únic alfabet creat arbitràriament que té un ús
estès en la vida diària. L'alfabet es va inventar l'any 1444 durant el regnat
de King Sejong (1418-1450).<EOS>
English: <BOS>They also said in a statement, "The crew is currently working to
determine the best method of safely extracting the ship".<EOS>
Catalan: <BOS>També han dit en un comunicat, "La tripulació treballa ara
mateix per a determinar la millor tècnica per a extreure la nau de manera
segura".<EOS>
English: <BOS>This is becoming less of an issue as lens manufacturers achieve
higher standards in lens production.<EOS>
Catalan: <BOS>Això és cada vegada menys important perquè els fabricants de
lents estan assolint estàndards més elevats en la producció de lents.<EOS>
English: <BOS>While assessing the successes and becoming aware of failures,
individuals and the whole of the participating persons discover more deeply
the values, mission, and driving forces of the organization.<EOS>
Catalan: <BOS>Mentre confirmen els èxits i prenen consciència dels fracassos,
els individus i el grup de participants descobreixen més profundament els
valors, la missió i les forces motrius de l'organització.<EOS>
English: <BOS>Entering Southern Africa by car is an amazing way to see
all the region's beauty as well as to get to places off the normal tourist
routes.<EOS>
Catalan: <BOS>Entrar a l'Àfrica del Sud en cotxe és una forma impressionant
de veure tota la bellesa de la regió i d'arribar a llocs fora de les rutes
turístiques més habituals.<EOS>
English: <BOS>____sentence_to_translate____<EOS>
Catalan: <BOS>
```

```
Translate the following sentence from English to Spanish:
English: <BOS>Hangeul is the only purposely invented alphabet in popular
daily use. The alphabet was invented in 1444 during the reign of King Sejong
(1418–1450).<EOS>
Spanish: <BOS>El alfabeto coreano es el único diseñado en forma deliberada que
aún se utiliza a diario popularmente. Se inventó en 1444, durante el reinado
de Sejong (1418 a 1450).<EOS>
English: <BOS>They also said in a statement, "The crew is currently working to
determine the best method of safely extracting the ship".<EOS>
Spanish: <BOS>También se dijo en un comunicado que: «La tripulación se
encuentra actualmente trabajando para decidir cuál es el método más seguro
para extraer el barco».<EOS>
English: <BOS>This is becoming less of an issue as lens manufacturers achieve
higher standards in lens production.<EOS>
Spanish: <BOS>Este problema cada vez es menos importante gracias a que los
fabricantes de lentes logran estándares más altos en su producción.<EOS>
English: <BOS>While assessing the successes and becoming aware of failures,
individuals and the whole of the participating persons discover more deeply
the values, mission, and driving forces of the organization.<EOS>
Spanish: <BOS>Durante el proceso de análisis de los éxitos y toma de
conciencia de los fracasos, los individuos y grupos de personas involucrados
descubren con mayor profundidad los valores, el objetivo y las fuerzas que
impulsan a la organización.<EOS>
English: <BOS>Entering Southern Africa by car is an amazing way to see
all the region’s beauty as well as to get to places off the normal tourist
routes.<EOS>
Spanish: <EOS>Una fantástica forma de contemplar todo el encanto de la región
del sur África es ingresar en automóvil, lo que, a su vez, le permitirá
acceder a lugares fuera de las rutas turísticas habituales.<BOS>
English: <BOS>____sentence_to_translate____<EOS>
Spanish: <BOS>
```

The prompts employed with Flor-6.3B and Llama-2-7B when testing FLoRes-200 dev set, WinoMT, Gold BUG and MuST-SHE for En → Ca and En → Es were:

```
Translate the following sentence from English to Catalan:
English: <BOS>The feathers' structure suggests that they were not used in
flight but rather for temperature regulation or display. The researchers
suggested that, even though this is the tail of a young dinosaur, the sample
shows adult plumage and not a chick's down.<EOS>
Catalan: <BOS>L'estructura de les plomes fa pensar que no s'usaven per a volar
sinó per a regular la temperatura o per a exhibir-se. Els investigadors han
suggerit que, tot i que es tracta de la cua d'un dinosaure jove, la mostra
presenta el plomatge d'un adult i no d'un pollet.<EOS>
English: <BOS>They found the Sun operated on the same basic principles as
other stars: The activity of all stars in the system was found to be driven by
their luminosity, their rotation, and nothing else.<EOS>
Catalan: <BOS>Han descobert que el Sol funcionava sota els mateixos principis
bàsics que altres estrelles: s'ha vist que l'activitat de totes les estrelles
del sistema depèn de llur brillantor, llur rotació i res més.<EOS>
English: <BOS>The speeds of 802.11n are substantially faster than that of its
predecessors with a maximum theoretical throughput of 600Mbit/s.<EOS>
Catalan: <BOS>Les velocitats de 802.11n són substancialment més ràpides que
les dels seus predecessors amb un rendiment teòric màxim de 600Mbit/s.<EOS>
English: <BOS>Over four million people went to Rome to attend the
funeral.<EOS>
Catalan: <BOS>Més de quatre milions de persones van anar a Roma per a assistir
al funeral.<EOS>
English: <BOS>Mrs. Kirchner announced her intention to run for president at
the Argentine Theatre, the same location she used to start her 2005 campaign
for the Senate as member of the Buenos Aires province delegation.<EOS>
Catalan: <BOS>La Sra. Kirchner va anunciar la seva intenció de presentar-se
a la presidència al Teatre de l'Argentina, el mateix lloc on va engegar la
campanya al Senat de 2005 com a membre de la delegació provincial de Buenos
Aires.<EOS>
English: <BOS>____sentence_to_translate____<EOS>
Catalan: <BOS>
```

```
Translate the following sentence from English to Spanish:
English: <BOS>The feathers’ structure suggests that they were not used in
flight but rather for temperature regulation or display. The researchers
suggested that, even though this is the tail of a young dinosaur, the sample
shows adult plumage and not a chick’s down.<EOS>
Spanish: <BOS>La estructura que presenta el plumaje sugiere que su función
no estaba relacionada con el vuelo, sino que las usaban para regular la
temperatura o como indicador de la misma. Los investigadores sostienen que,
aunque se trata de la cola de un dinosaurio joven, la muestra analizada
presenta rasgos del plumaje de un adulto y no de un polluelo.<EOS>
English: <BOS>They found the Sun operated on the same basic principles as
other stars: The activity of all stars in the system was found to be driven by
their luminosity, their rotation, and nothing else.<EOS>
Spanish: <BOS>Se descubrió que el sol se regía por los mismos principios
básicos que otras estrellas: los únicos factores que impulsaban su actividad
dentro del sistema eran su luminosidad y su rotación.<EOS>
English: <BOS>The speeds of 802.11n are substantially faster than that of its
predecessors with a maximum theoretical throughput of 600Mbit/s.<EOS>
Spanish: <BOS>Las velocidades del estándar 802.11n son mucho más altas que las
alcanzadas por los que lo precedieron, con un rendimiento teórico máximo de
600 Mbps.<EOS>
English: <BOS>Over four million people went to Rome to attend the
funeral.<EOS>
Spanish: <BOS>Más de cuatro millones de individuos se concentraron en Roma
para presenciar el funeral.<EOS>
English: <BOS>Mrs. Kirchner announced her intention to run for president at
the Argentine Theatre, the same location she used to start her 2005 campaign
for the Senate as member of the Buenos Aires province delegation.<EOS>
Spanish: <BOS>El Teatro Argentino fue el lugar donde la señora Kirchner
anunció su intención de candidatearse como presidenta; este es el mismo sitio
donde inició su campaña para el senado en el año 2005, en representación de la
provincia de Buenos Aires.<EOS>
English: <BOS>____sentence_to_translate____<EOS>
Spanish: <BOS>
```

D Gender Scores on the Pro- and Anti-Stereotypical sets from WinoMT and Gold BUG

Below you can see the results for the WinoMT:

		English → Catalan				English → Spanish			
		G Acc	**F1-male**	**F1-female**	**ΔG**	**G Acc**	**F1-male**	**F1-female**	**ΔG**
NMT	Google Translate	74.1	80.9	71.1	9.8	**89.8**	**91.1**	**90.4**	**0.7**
	NLLB-200-1.3B	**79.7**	**85.1**	**80.7**	**4.4**	89.3	90.8	89.8	1.0
	M2M-100-1.2B	67.7	76.4	61.1	15.3	73.7	79.2	68.7	10.5
	Mt-aina-en-ca	65.7	76.0	56.9	19.1	-	-	-	-
LLM	Ăguila-7B	66.0	76.1	**57.8**	**18.3**	65.7	74.1	53.8	20.3
	Flor-6.3B	66.4	76.3	57.6	18.7	71.9	77.9	63.3	14.6
	Llama-2-7B	**66.5**	**78.0**	57.7	20.3	**73.1**	**78.5**	**66.4**	**12.1**

Table 6: WinoMT pro-stereotypical set gender scores

		English → Catalan				English → Spanish			
		G Acc	**F1-male**	**F1-female**	**ΔG**	**G Acc**	**F1-male**	**F1-female**	**ΔG**
NMT	Google Translate	51.8	60.6	43.7	16.9	**66.9**	**72.4**	**60.6**	**11.8**
	NLLB-200-1.3B	**53.0**	**62.3**	**47.3**	**15.0**	58.1	66.7	46.2	20.5
	M2M-100-1.2B	45.9	58.3	30.0	28.3	52.5	65.3	29.6	35.7
	Mt-aina-en-ca	42.5	56.8	21.5	35.3	-	-	-	-
LLM	Ăguila-7B	34.5	49.8	12.0	37.8	43.1	58.5	12.7	45.8
	Flor-6.3B	38.7	**54.0**	13.8	40.2	45.2	58.2	**22.8**	**35.4**
	Llama-2-7B	**38.8**	53.6	**16.6**	**37.0**	**45.3**	**59.0**	19.7	39.3

Table 7: WinoMT anti-stereotypical set gender scores

Below you can see the results for the Gold BUG:

		English → Catalan				English → Spanish			
		G Acc	**F1-male**	**F1-female**	**ΔG**	**G Acc**	**F1-male**	**F1-female**	**ΔG**
NMT	Google Translate	**69.6**	**82.6**	**67.5**	**15.1**	70.8	83.9	60.5	23.4
	NLLB-200-1.3B	66.9	81.3	52.7	28.6	**71.5**	**84.1**	**66.6**	**17.5**
	M2M-100-1.2B	67.4	81.8	54.7	27.1	70.3	83.6	61.3	22.3
	Mt-aina-en-ca	68.0	81.8	64.4	17.4	-	-	-	-
LLM	Ăguila-7B	60.7	76.3	54.8	**21.5**	64.4	79.1	**56.0**	**23.1**
	Flor-6.3B	65.0	80.1	54.7	25.4	69.8	83.2	54.6	28.6
	Llama-2-7B	**66.5**	**81.3**	**56.1**	25.2	**69.9**	**83.4**	53.1	30.3

Table 8: Gold BUG pro-stereotypical set gender scores

		English → Catalan				English → Spanish			
		G Acc	**F1-male**	**F1-female**	**ΔG**	**G Acc**	**F1-male**	**F1-female**	**ΔG**
NMT	Google Translate	43.6	61.0	35.7	25.3	**51.4**	**67.1**	**47.6**	**19.5**
	NLLB-200-1.3B	**46.9**	**62.9**	**44.0**	18.9	48.8	65.5	44.4	21.1
	M2M-100-1.2B	41.9	59.5	29.2	30.3	46.2	62.8	36.8	26.0
	Mt-aina-en-ca	46.0	61.3	43.9	**17.4**	-	-	-	-
LLM	Ăguila-7B	40.0	59.1	27.0	32.1	46.0	64.8	32.8	32.0
	Flor-6.3B	44.5	62.5	35.1	**27.4**	49.0	66.2	**41.9**	**24.3**
	Llama-2-7B	**46.7**	**64.4**	**35.7**	28.7	**49.8**	**69.0**	35.4	33.6

Table 9: Gold BUG anti-stereotypical set gender scores

E Proportion of Predicted Male and Female terms in Absence of Gender Cues

The following Figures 3 and 4 depict a range of pie diagrams illustrating the proportion of predicted male and female terms in the translations per model when testing on instances of MuST-SHE without gender cues for disambiguation.

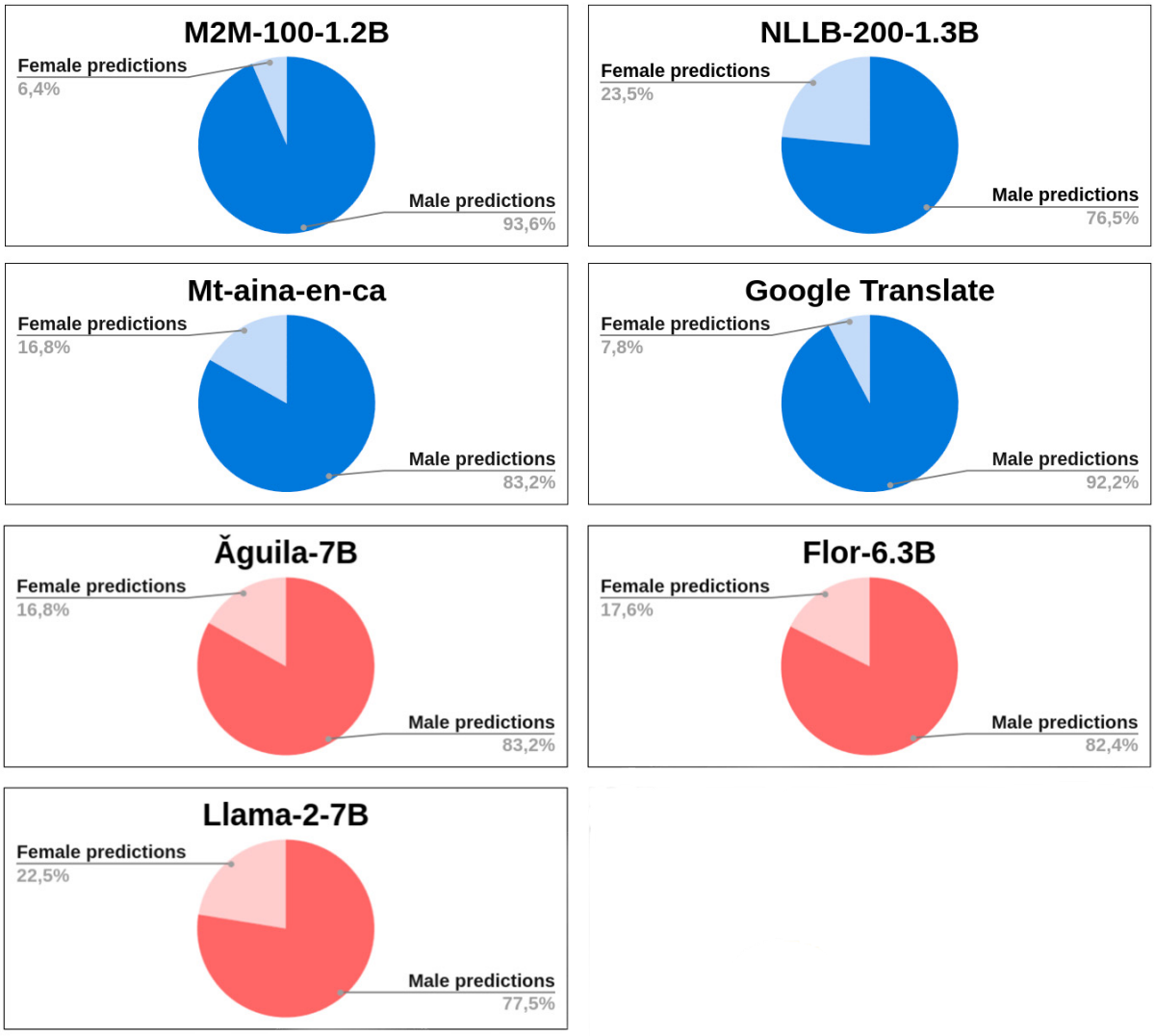

Figure 3: Male and female predicted terms across models for En → Ca in absence of gender cues

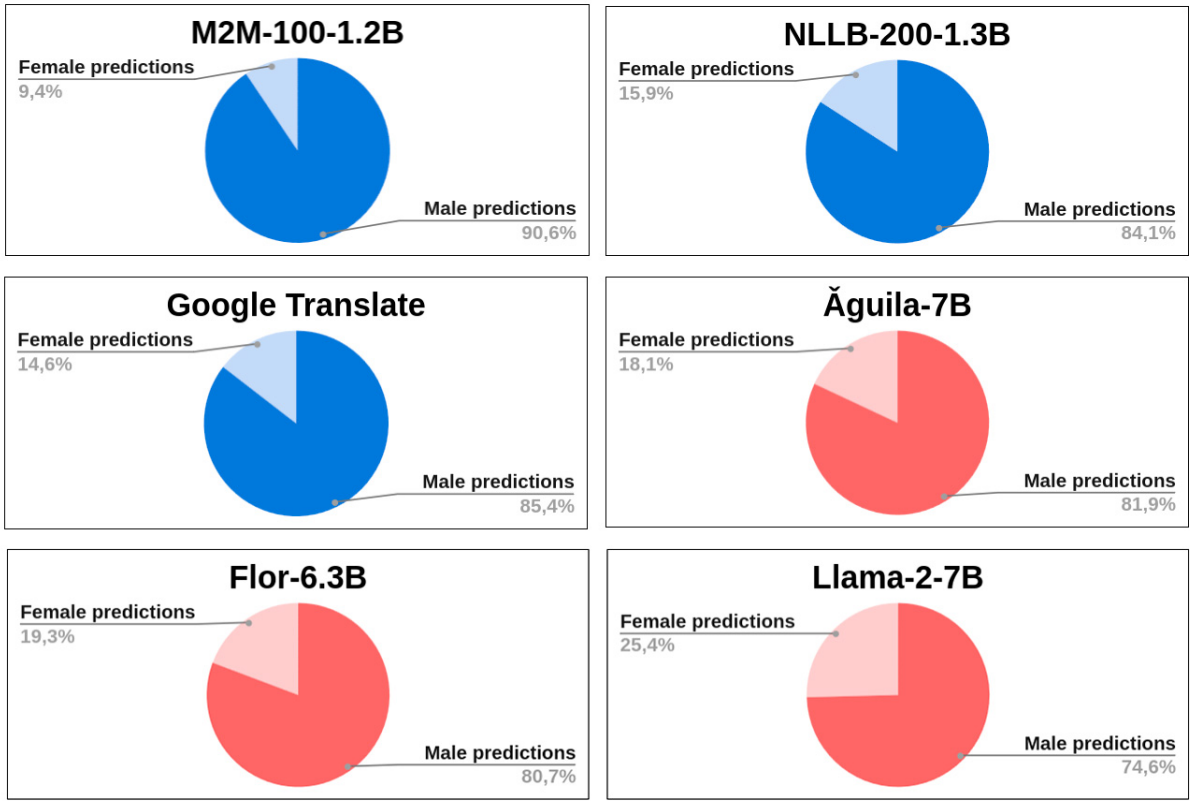

Figure 4: Male and female predicted terms across models for En → Es in absence of gender cues

F Prompt used for the Baseline in the Investigation into Prompting

An example of the resulting prompt used for Llama-2-7B-chat after the format adaptations:

```
«SYS» Translate the following sentence from English to Catalan: «/SYS»
[INST] English: <BOS>The feathers' structure suggests that they were not used
in flight but rather for temperature regulation or display. The researchers
suggested that, even though this is the tail of a young dinosaur, the sample
shows adult plumage and not a chick's down.<EOS> [/INST]
Catalan: <BOS>L'estructura de les plomes fa pensar que no s'usaven per a volar
sinó per a regular la temperatura o per a exhibir-se. Els investigadors han
suggerit que, tot i que es tracta de la cua d'un dinosaure jove, la mostra
presenta el plomatge d'un adult i no d'un pollet.<EOS>
[INST] English: <BOS>They found the Sun operated on the same basic principles
as other stars: The activity of all stars in the system was found to be driven
by their luminosity, their rotation, and nothing else.<EOS> [/INST]
Catalan: <BOS>Han descobert que el Sol funcionava sota els mateixos principis
bàsics que altres estrelles: s'ha vist que l'activitat de totes les estrelles
del sistema depèn de llur brillantor, llur rotació i res més.<EOS>
[INST] English: <BOS>The speeds of 802.11n are substantially faster than that
of its predecessors with a maximum theoretical throughput of 600Mbit/s.<EOS>
[/INST]
Catalan: <BOS>Les velocitats de 802.11n són substancialment més ràpides que
les dels seus predecessors amb un rendiment teòric màxim de 600Mbit/s.<EOS>
[INST] English: <BOS>Over four million people went to Rome to attend the
funeral.<EOS> [/INST]
Catalan: <BOS>Més de quatre milions de persones van anar a Roma per a assistir
al funeral.<EOS>
[INST] English: <BOS>Mrs. Kirchner announced her intention to run for
president at the Argentine Theatre, the same location she used to start
her 2005 campaign for the Senate as member of the Buenos Aires province
delegation.<EOS> [/INST]
Catalan: <BOS>La Sra. Kirchner va anunciar la seva intenció de presentar-se
a la presidència al Teatre de l'Argentina, el mateix lloc on va engegar la
campanya al Senat de 2005 com a membre de la delegació provincial de Buenos
Aires.<EOS>
[INST] English: <BOS>____sentence_to_translate____<EOS> [/INST]
Catalan: <BOS>
```

G Curated Prompts for the Investigation into Prompting

Below are all the different prompts used with Llama-2-7B-chat that have been tested on WinoMT test set.

Prompt with 5-shot MuST-SHE examples:

```
«SYS» Translate the following sentence from English to Catalan: «/SYS»
[INST] English: <BOS>Early on, Laura Hughes could see that I was a little lost
in this habitat, so she often sat right next to me in meetings so she could be
my tech translator, and I could write her notes and she could tell me, "That's
what that means." Laura was 27 years old, she'd worked for Google for four
years and then for a year and a half at Airbnb when I met her.<EOS> [/INST]
Catalan: <BOS>Al principi, la Laura Hughes va poder veure que estava una
mica perdut en aquest hàbitat, així que sovint s'asseia al meu costat a
les reunions per poder ser la meva traductora de tecnologia, i jo podia
escriure-li notes i ella em podria dir, "Això és el que això significa." La
Laura tenia 27 anys, havia treballat a Google durant quatre anys i després
durant un any i mig a Airbnb quan la vaig conèixer.<EOS>

[INST] English: <BOS>When I found the captain, he was having a very engaging
conversation with the homeowner, who was surely having one of the worst days
of her life.<EOS> [/INST]
Catalan: <BOS>Quan vaig trobar el capità, estava mantenint una conversa molt
atractiva amb la propietària, que segurament vivia un dels pitjors dies de la
seva vida.<EOS>

[INST] English: <BOS>And in this program, girls who have been studying
computer skills and the STEM program have a chance to work side by side with
young professionals, so that they can learn firsthand what it's like to be an
architect, a designer or a scientist.<EOS> [/INST]
Catalan: <BOS>I en aquest programa, les noies que han estudiat informàtica
i el programa STEM tenen l'oportunitat de treballar colze a colze amb joves
professionals, per tal que puguin conèixer de primera mà com és ser una
arquitecta, una dissenyadora o una científica.<EOS>

[INST] English: <BOS>One government scientist, a friend of mine, we'll call
him McPherson, was concerned about the impact government policies were having
on his research and the state of science deteriorating in Canada.<EOS> [/INST]
Catalan: <BOS>Un científic del govern, un amic meu, l'anomenarem McPherson,
estava preocupat per l'impacte que tenien les polítiques governamentals en la
seva investigació i el deteriorament de l'estat de la ciència al Canadà.<EOS>

[INST] English: <BOS>The architect Emmanuelle Moureaux uses this idea in her
work a lot.<EOS> [/INST]
Catalan: <BOS>L'arquitecta Emmanuelle Moureaux utilitza molt aquesta idea en
la seva obra.<EOS>

[INST] English: <BOS>____sentence_to_translate____<EOS> [/INST]
Catalan: <BOS>
```

```
«SYS» Translate the following sentence from English to Spanish: «/SYS»
[INST] English: <BOS>Early on, Laura Hughes could see that I was a little lost
in this habitat, so she often sat right next to me in meetings so she could be
my tech translator, and I could write her notes and she could tell me, "That's
what that means." Laura was 27 years old, she'd worked for Google for four
years and then for a year and a half at Airbnb when I met her.<EOS> [/INST]
Spanish: <BOS>Al principio, Laura Hughes se dio cuenta de que estaba perdido
en este hábitat, así que solía sentarse a mi lado en las reuniones para ser mi
traductora de tecnología, y yo le escribía notas y ella me decía, "Esto es lo
que significa". Laura tenía 27 años, trabajó en Google durante 4 años, y luego
por un año y medio en Airbnb cuando la conocí.<EOS>

[INST] English: <BOS>When I found the captain, he was having a very engaging
conversation with the homeowner, who was surely having one of the worst days
of her life.<EOS> [/INST]
Spanish: <BOS>Cuando encontré al capitán, estaba enfrascado en una
conversación con la propietaria que sin duda atravesaba uno de los peores
días de su vida.<EOS>

[INST] English: <BOS>And in this program, girls who have been studying
computer skills and the STEM program have a chance to work side by side with
young professionals, so that they can learn firsthand what it's like to be an
architect, a designer or a scientist.<EOS> [/INST]
Spanish: <BOS>En este programa, las niñas que estudian informática y el
programa CTIM tienen la oportunidad de trabajar junto a jóvenes profesionales,
para que puedan aprender de primera mano qué es ser una arquitecta,
diseñadora, o científica.<EOS>

[INST] English: <BOS>One government scientist, a friend of mine, we'll call
him McPherson, was concerned about the impact government policies were having
on his research and the state of science deteriorating in Canada.<EOS> [/INST
Spanish: <BOS>Un científico del gobierno, un amigo mío, lo llamaremos
McPherson, estaba preocupado por el impacto que las políticas gubernamentales
estaban teniendo en su investigación y el estado de la ciencia deteriorándose
en Canadá.<EOS>

[INST] English: <BOS>The architect Emmanuelle Moureaux uses this idea in her
work a lot.<EOS> [/INST
Spanish: <BOS>La arquitecta Emmanuelle Moureaux usa mucho esta idea en su
trabajo.<EOS>

[INST] English: <BOS>____sentence_to_translate____<EOS> [/INST]
Spanish: <BOS>
```

Prompt with 5-shot invented Winograd examples:

```
«SYS» Translate the following sentence from English to Catalan: «/SYS»
[INST] English: <BOS>The grower delivered tomatoes to the housekeeper, because
he needed them for dinner.<EOS> [/INST]
Catalan: <BOS>La productora va lliurar tomàquets al mestre de casa, perquè els
necessitava per sopar.<EOS>

[INST] English: <BOS>The developer was upset with the designer because she
was disappointed with the final design.<EOS> [/INST]
Catalan: <BOS>La desenvolupadora estava molesta amb el dissenyador perquè
estava decebuda amb el disseny final.<EOS>

[INST] English: <BOS>The doctor congratulated the nurse saying his assistance
was great.<EOS> [/INST]
Catalan: <BOS>La doctora va felicitar a l'infermer dient que la seva
assistència va ser fantàstica.<EOS>

[INST] English: <BOS>The construction worker asked the chief for an increase
in her salary.<EOS> [/INST]
Catalan: <BOS>La treballadora de la construcció va demanar al cap un augment
del seu sou.<EOS>

[INST] English: <BOS>The carpenter asked the CEO if she liked the new
furniture.<EOS> [/INST]
Catalan: <BOS>La fustera li va preguntar a la directora si li agradaven els
mobles nous.<EOS>

[INST] English: <BOS>____sentence_to_translate____<EOS> [/INST]
Catalan: <BOS>
```

```
«SYS» Translate the following sentence from English to Spanish: «/SYS»
[INST] English: <BOS>The grower delivered tomatoes to the housekeeper, because
he needed them for dinner.<EOS> [/INST]
Spanish: <BOS>La productora entregó tomates al amo de casa, porque los
necesitaba para la cena.<EOS>

[INST] English: <BOS>The developer was upset with the designer because she
was disappointed with the final design.<EOS> [/INST]
Spanish: <BOS>La desarrolladora estaba enojada con el diseñador porque estaba
decepcionada con el diseño final.<EOS>

[INST] English: <BOS>The doctor congratulated the nurse saying his assistance
was great.<EOS> [/INST]
Spanish: <BOS>La doctora felicitó al infermero diciendo que su asistencia fue
fantástica.<EOS>

[INST] English: <BOS>The construction worker asked the chief for an increase
in her salary.<EOS> [/INST]
Spanish: <BOS>La trabajadora de la construcción pidió al jefe un aumento de su
salario.<EOS>

[INST] English: <BOS>The carpenter asked the CEO if she liked the new
furniture.<EOS> [/INST]
Spanish: <BOS>La carpintera preguntó a la directora general si le gustaban los
muebles nuevos.<EOS>

[INST] English: <BOS>____sentence_to_translate____<EOS> [/INST]
Spanish: <BOS>
```

Prompt with 5-shot MuST-SHE examples and context on the Gender Bias issue:

```
«SYS» Translate the following sentence from English to Catalan while
mitigating gender bias. First, consider that English is a language without
grammatical gender, while Catalan does have grammatical gender. Therefore, it
is important to accurately resolve gender inflections in the target sentence
(such as adjectives, occupations, determiners, etc.) based on the gender
information provided in the source sentence. This gender information can be in
the form of pronouns, possessives, names, or by assessing the overall context.
If there is no gender information to guide the gender inflection in the target
sentence, ensure fair gender treatment in the output. This means using random
gender inflections in the translation. «/SYS»

[INST] English: <BOS>Early on, Laura Hughes could see that I was a little
lost in this habitat, so she often sat right next to me in meetings so she
could be my tech translator, and I could write her notes and she could tell
me, "That's what that means." Laura was 27 years old, she'd worked for Google
for four years and then for a year and a half at Airbnb when I met her.<EOS>
[/INST]
Catalan: <BOS>Al principi, la Laura Hughes va poder veure que estava una
mica perdut en aquest hàbitat, així que sovint s'asseia al meu costat a
les reunions per poder ser la meva traductora de tecnologia, i jo podia
escriure-li notes i ella em podria dir, "Això és el que això significa." La
Laura tenia 27 anys, havia treballat a Google durant quatre anys i després
durant un any i mig a Airbnb quan la vaig conèixer.<EOS>

[INST] English: <BOS>When I found the captain, he was having a very engaging
conversation with the homeowner, who was surely having one of the worst days
of her life.<EOS> [/INST]
Catalan: <BOS>Quan vaig trobar el capità, estava mantenint una conversa molt
atractiva amb la propietària, que segurament vivia un dels pitjors dies de la
seva vida.<EOS>

[INST] English: <BOS>And in this program, girls who have been studying
computer skills and the STEM program have a chance to work side by side with
young professionals, so that they can learn firsthand what it's like to be an
architect, a designer or a scientist.<EOS> [/INST]
Catalan: <BOS>I en aquest programa, les noies que han estudiat informàtica
i el programa STEM tenen l'oportunitat de treballar colze a colze amb joves
professionals, per tal que puguin conèixer de primera mà com és ser una
arquitecta, una dissenyadora o una científica.<EOS>

[INST] English: <BOS>One government scientist, a friend of mine, we'll call
him McPherson, was concerned about the impact government policies were having
on his research and the state of science deteriorating in Canada.<EOS> [/INST]
Catalan: <BOS>Un científic del govern, un amic meu, l'anomenarem McPherson,
estava preocupat per l'impacte que tenien les polítiques governamentals en la
seva investigació i el deteriorament de l'estat de la ciència al Canadà.<EOS>

[INST] English: <BOS>The architect Emmanuelle Moureaux uses this idea in her
work a lot.<EOS> [/INST]
Catalan: <BOS>L'arquitecta Emmanuelle Moureaux utilitza molt aquesta idea en
la seva obra.<EOS>

[INST] English: <BOS>____sentence_to_translate____<EOS> [/INST]
Catalan: <BOS>
```

«SYS» Translate the following sentence from English to Spanish while mitigating gender bias. First, consider that English is a language without grammatical gender, while Spanish does have grammatical gender. Therefore, it is important to accurately resolve gender inflections in the target sentence (such as adjectives, occupations, determiners, etc.) based on the gender information provided in the source sentence. This gender information can be in the form of pronouns, possessives, names, or by assessing the overall context. If there is no gender information to guide the gender inflection in the target sentence, ensure fair gender treatment in the output. This means using random gender inflections in the translation. «/SYS»

[INST] English: <BOS>Early on, Laura Hughes could see that I was a little lost in this habitat, so she often sat right next to me in meetings so she could be my tech translator, and I could write her notes and she could tell me, "That's what that means." Laura was 27 years old, she'd worked for Google for four years and then for a year and a half at Airbnb when I met her.<EOS> [/INST]
Spanish: <BOS>Al principio, Laura Hughes se dio cuenta de que estaba perdido en este hábitat, así que solía sentarse a mi lado en las reuniones para ser mi traductora de tecnología, y yo le escribía notas y ella me decía, "Esto es lo que significa". Laura tenía 27 años, trabajó en Google durante 4 años, y luego por un año y medio en Airbnb cuando la conocí.<EOS>

[INST] English: <BOS>When I found the captain, he was having a very engaging conversation with the homeowner, who was surely having one of the worst days of her life.<EOS> [/INST]
Spanish: <BOS>Cuando encontré al capitán, estaba enfrascado en una conversación con la propietaria que sin duda atravesaba uno de los peores días de su vida.<EOS>

[INST] English: <BOS>And in this program, girls who have been studying computer skills and the STEM program have a chance to work side by side with young professionals, so that they can learn firsthand what it's like to be an architect, a designer or a scientist.<EOS> [/INST]
Spanish: <BOS>En este programa, las niñas que estudian informática y el programa CTIM tienen la oportunidad de trabajar junto a jóvenes profesionales, para que puedan aprender de primera mano qué es ser una arquitecta, diseñadora, o científica.<EOS>

[INST] English: <BOS>One government scientist, a friend of mine, we'll call him McPherson, was concerned about the impact government policies were having on his research and the state of science deteriorating in Canada.<EOS> [/INST]
Spanish: <BOS>Un científico del gobierno, un amigo mío, lo llamaremos McPherson, estaba preocupado por el impacto que las políticas gubernamentales estaban teniendo en su investigación y el estado de la ciencia deteriorándose en Canadá.<EOS>

[INST] English: <BOS>The architect Emmanuelle Moureaux uses this idea in her work a lot.<EOS> [/INST]
Spanish: <BOS>La arquitecta Emmanuelle Moureaux usa mucho esta idea en su trabajo.<EOS>

[INST] English: <BOS>____sentence_to_translate____<EOS> [/INST]
Spanish:

Prompt with 5-shot invented Winograd examples and chain-of-thought prompting (“agent” version):

```
«SYS» Identify the agents in the sentence. Identify the pronouns referring
to the agents in the sentence. Relate each pronoun to the correct agent.
Establish the gender of each agent according to the pronoun. Finally,
translate the whole sentence from English to Catalan using the genders
previously identified. Proceed step by step. «/SYS»

[INST] English: <BOS>The grower delivered tomatoes to the housekeeper, because
he needed them for dinner.<EOS> [/INST]
1. Identify the agents in the sentence:
     * grower (agent)
     * housekeeper (agent)
2. Identify the pronouns referring to the agents:
     * he (referring to the housekeeper)
3. Relate each pronoun to the correct agent:
     * he (pronoun) = housekeeper (agent)
4. Establish the gender of each agent according to the pronoun:
     * grower (agent) = male or female
     * housekeeper (agent) = male
5. Translate the sentence from English to Catalan:
Catalan: <BOS>La productora va lliurar tomàquets al mestre de casa, perquè els
necessitava per sopar.<EOS>

[INST] English: <BOS>The developer was upset with the designer because she
was disappointed with the final design.<EOS> [/INST]
1. Identify the agents in the sentence:
     * developer (agent)
     * designer (agent)
2. Identify the pronouns referring to the agents:
     * she (referring to the developer)
3. Relate each pronoun to the correct agent:
     * she (pronoun) = developer (agent)
4. Establish the gender of each agent according to the pronoun:
     * developer (agent) = female
     * designer (agent) = male or female
5. Translate the sentence from English to Catalan:
Catalan: <BOS>La desenvolupadora estava molesta amb el dissenyador perquè
estava decebuda amb el disseny final.<EOS>

[INST] English: <BOS>The doctor congratulated the nurse saying his assistance
was great.<EOS> [/INST]
1. Identify the agents in the sentence:
     * doctor (agent)
     * nurse (agent)
2. Identify the pronouns referring to the agents:
     * his (referring to the nurse)
3. Relate each pronoun to the correct agent:
     * his (pronoun) = nurse (agent)
4. Establish the gender of each agent according to the pronoun:
     * doctor (agent) = male or female
     * nurse (agent) = male
5. Translate the sentence from English to Catalan:
Catalan: <BOS>La doctora va felicitar a l’infermer dient que la seva
assistència va ser fantàstica.<EOS>
```

```
[INST] English: <BOS>The construction worker asked the chief for an increase
in her salary.<EOS> [/INST]
1. Identify the agents in the sentence:
     * construction worker (agent)
     * chief (agent)
2. Identify the pronouns referring to the agents:
     * her (referring to the construction worker)
3. Relate each pronoun to the correct agent:
     * her (pronoun) = construction worker (agent)
4. Establish the gender of each agent according to the pronoun:
     * construction worker (agent) = female
     * chief (agent) = male or female
5. Translate the sentence from English to Catalan:
Catalan: <BOS>La treballadora de la construcció va demanar al cap un augment
del seu sou.<EOS>

[INST] English: <BOS>The carpenter asked the CEO if she liked the new
furniture.<EOS> [/INST]
1. Identify the agents in the sentence:
     * carpenter (agent)
     * CEO (agent)
2. Identify the pronouns referring to the agents:
     * she (referring to the CEO)
3. Relate each pronoun to the correct agent:
     * she (pronoun) = CEO (agent)
4. Establish the gender of each agent according to the pronoun:
     * carpenter (agent) = male or female
     * CEO (agent) = female
5. Translate the sentence from English to Catalan:
Catalan: <BOS>La fustera li va preguntar a la directora general si li
agradaven els mobles nous.<EOS>

[INST] English: <BOS>____sentence_to_translate____<EOS> [/INST]
1.
```

```
«SYS» Identify the agents in the sentence. Identify the pronouns referring
to the agents in the sentence. Relate each pronoun to the correct agent.
Establish the gender of each agent according to the pronoun. Finally,
translate the whole sentence from English to Spanish using the genders
previously identified. Proceed step by step. «/SYS»

[INST] English: <BOS>The grower delivered tomatoes to the housekeeper, because
he needed them for dinner.<EOS> [/INST]
1. Identify the agents in the sentence:
     * grower (agent)
     * housekeeper (agent)
2. Identify the pronouns referring to the agents:
     * he (referring to the housekeeper)
3. Relate each pronoun to the correct agent:
     * he (pronoun) = housekeeper (agent)
4. Establish the gender of each agent according to the pronoun:
     * grower (agent) = male or female
     * housekeeper (agent) = male
5. Translate the sentence from English to Spanish:
Spanish: <BOS>La productora entregó tomates al amo de casa, porque los
necesitaba para la cena.<EOS>

[INST] English: <BOS>The developer was upset with the designer because she
was disappointed with the final design.<EOS> [/INST]
1. Identify the agents in the sentence:
     * developer (agent)
     * designer (agent)
2. Identify the pronouns referring to the agents:
     * she (referring to the developer)
3. Relate each pronoun to the correct agent:
     * she (pronoun) = developer (agent)
4. Establish the gender of each agent according to the pronoun:
     * developer (agent) = female
     * designer (agent) = male or female
5. Translate the sentence from English to Spanish:
Spanish: <BOS>La desarrolladora estaba enojada con el diseñador porque estaba
decepcionada con el diseño final.<EOS>

[INST] English: <BOS>The doctor congratulated the nurse saying his assistance
was great.<EOS> [/INST]
1. Identify the agents in the sentence:
     * doctor (agent)
     * nurse (agent)
2. Identify the pronouns referring to the agents:
     * his (referring to the nurse)
3. Relate each pronoun to the correct agent:
     * his (pronoun) = nurse (agent)
4. Establish the gender of each agent according to the pronoun:
     * doctor (agent) = male or female
     * nurse (agent) = male
5. Translate the sentence from English to Spanish:
Spanish: <BOS>La doctora felicitó al infermero diciendo que su asistencia fue
fantástica.<EOS>
```

```
[INST] English: <BOS>The construction worker asked the chief for an increase
in her salary.<EOS> [/INST]
1. Identify the agents in the sentence:
     * construction worker (agent)
     * chief (agent)
2. Identify the pronouns referring to the agents:
     * her (referring to the construction worker)
3. Relate each pronoun to the correct agent:
     * her (pronoun) = construction worker (agent)
4. Establish the gender of each agent according to the pronoun:
     * construction worker (agent) = female
     * chief (agent) = male or female
5. Translate the sentence from English to Spanish:
Spanish: <BOS>La trabajadora de la construcción pidió al jefe un aumento de su
salario.<EOS>

[INST] English: <BOS>The carpenter asked the CEO if she liked the new
furniture.<EOS> [/INST]
1. Identify the agents in the sentence:
     * carpenter (agent)
     * CEO (agent)
2. Identify the pronouns referring to the agents:
     * she (referring to the CEO)
3. Relate each pronoun to the correct agent:
     * she (pronoun) = CEO (agent)
4. Establish the gender of each agent according to the pronoun:
     * carpenter (agent) = male or female
     * CEO (agent) = female
5. Translate the sentence from English to Spanish:
Spanish: <BOS>La carpintera preguntó a la directora general si le gustaban los
muebles nuevos.<EOS>

[INST] English: <BOS>____sentence_to_translate____<EOS> [/INST]
1.
```

Prompt with 5-shot on invented Winograd examples and chain-of-thought prompting ("human entity" version):

```
«SYS» Identify the human entities in the sentence. Identify the pronouns
referring to the human entities in the sentence. Relate each pronoun to the
correct human entity. Establish the gender of each human entity according to
the pronoun. Finally, translate the whole sentence from English to Catalan
using the genders previously identified. Proceed step by step. «/SYS»

[INST] English: <BOS>The grower delivered tomatoes to the housekeeper, because
he needed them for dinner.<EOS> [/INST]
1. Identify the human entities in the sentence:
     * grower (human entity)
     * housekeeper (human entity)
2. Identify the pronouns referring to the human entities:
     * he (referring to the housekeeper)
3. Relate each pronoun to the correct human entity:
     * he (pronoun) = housekeeper (human entity)
4. Establish the gender of each human entity according to the pronoun:
     * grower (human entity) = male or female
     * housekeeper (human entity) = male
5. Translate the sentence from English to Catalan:
Catalan: <BOS>La productora va lliurar tomàquets al mestre de casa, perquè els
necessitava per sopar.<EOS>

[INST] English: <BOS>The developer was upset with the designer because she was
disappointed with the final design.<EOS> [/INST]
1. Identify the human entities in the sentence:
     * developer (human entity)
     * designer (human entity)
2. Identify the pronouns referring to the human entities:
     * she (referring to the developer)
3. Relate each pronoun to the correct human entity:
     * she (pronoun) = developer (human entity)
4. Establish the gender of each human entity according to the pronoun:
     * developer (human entity) = female
     * designer (human entity) = male or female
5. Translate the sentence from English to Catalan:
Catalan: <BOS>La desenvolupadora estava molesta amb el dissenyador perquè
estava decebuda amb el disseny final.<EOS>

[INST] English: <BOS>The doctor congratulated the nurse saying his assistance
was great.<EOS> [/INST]
1. Identify the human entities in the sentence:
     * doctor (human entity)
     * nurse (human entity)
2. Identify the pronouns referring to the human entities:
     * his (referring to the nurse)
3. Relate each pronoun to the correct human entity:
     * his (pronoun) = nurse (human entity)
4. Establish the gender of each human entity according to the pronoun:
     * doctor (human entity) = male or female
     * nurse (human entity) = male
5. Translate the sentence from English to Catalan:
Catalan: <BOS>La doctora va felicitar a l'infermer dient que la seva
assistència va ser fantàstica.<EOS>
```

```
[INST] English: <BOS>The construction worker asked the chief for an increase
in her salary.<EOS> [/INST]
1. Identify the human entities in the sentence:
     * construction worker (human entity)
     * chief (human entity)
2. Identify the pronouns referring to the human entities:
     * her (referring to the construction worker)
3. Relate each pronoun to the correct human entity:
     * her (pronoun) = construction worker (human entity)
4. Establish the gender of each human entity according to the pronoun:
     * construction worker (human entity) = female
     * chief (human entity) = male or female
5. Translate the sentence from English to Catalan:
Catalan: <BOS>La treballadora de la construcció va demanar al cap un augment
del seu sou.<EOS>

[INST] English: <BOS>The carpenter asked the CEO if she liked the new
furniture.<EOS> [/INST]
1. Identify the human entities in the sentence:
     * carpenter (human entity)
     * CEO (human entity)
2. Identify the pronouns referring to the human entities:
     * she (referring to the CEO)
3. Relate each pronoun to the correct human entity:
     * she (pronoun) = CEO (human entity)
4. Establish the gender of each human entity according to the pronoun:
     * carpenter (human entity) = male or female
     * CEO (human entity) = female
5. Translate the sentence from English to Catalan:
Catalan: <BOS>La fustera li va preguntar a la directora general si li
agradaven els mobles nous.<EOS>

[INST] English: <BOS>____sentence_to_translate____<EOS> [/INST]
1.
```

```
«SYS» Identify the human entities in the sentence. Identify the pronouns
referring to the human entities in the sentence. Relate each pronoun to the
correct human entity. Establish the gender of each human entity according to
the pronoun. Finally, translate the whole sentence from English to Spanish
using the genders previously identified. Proceed step by step. «/SYS»

[INST] English: <BOS>The grower delivered tomatoes to the housekeeper, because
he needed them for dinner.<EOS> [/INST]
1. Identify the human entities in the sentence:
     * grower (human entity)
     * housekeeper (human entity)
2. Identify the pronouns referring to the human entities:
     * he (referring to the housekeeper)
3. Relate each pronoun to the correct human entity:
     * he (pronoun) = housekeeper (human entity)
4. Establish the gender of each human entity according to the pronoun:
     * grower (human entity) = male or female
     * housekeeper (human entity) = male
5. Translate the sentence from English to Spanish:
Spanish: <BOS>La productora entregó tomates al amo de casa, porque los
necesitaba para la cena.<EOS>

[INST] English: <BOS>The developer was upset with the designer because she
was disappointed with the final design.<EOS> [/INST]
1. Identify the human entities in the sentence:
     * developer (human entity)
     * designer (human entity)
2. Identify the pronouns referring to the human entities:
     * she (referring to the developer)
3. Relate each pronoun to the correct human entity:
     * she (pronoun) = developer (human entity)
4. Establish the gender of each human entity according to the pronoun:
     * developer (human entity) = female
     * designer (human entity) = male or female
5. Translate the sentence from English to Spanish:
Spanish: <BOS>La desarrolladora estaba enojada con el diseñador porque estaba
decepcionada con el diseño final.<EOS>

[INST] English: <BOS>The doctor congratulated the nurse saying his assistance
was great.<EOS> [/INST]
1. Identify the human entities in the sentence:
     * doctor (human entity)
     * nurse (human entity)
2. Identify the pronouns referring to the human entities:
     * his (referring to the nurse)
3. Relate each pronoun to the correct human entity:
     * his (pronoun) = nurse (human entity)
4. Establish the gender of each human entity according to the pronoun:
     * doctor (human entity) = male or female
     * nurse (human entity) = male
5. Translate the sentence from English to Spanish:
Spanish: <BOS>La doctora felicitó al infermero diciendo que su asistencia fue
fantástica.<EOS>
```

```
[INST] English: <BOS>The construction worker asked the chief for an increase
in her salary.<EOS> [/INST]
1. Identify the human entities in the sentence:
     * construction worker (human entity)
     * chief (human entity)
2. Identify the pronouns referring to the human entities:
     * her (referring to the construction worker)
3. Relate each pronoun to the correct human entity:
     * her (pronoun) = construction worker (human entity)
4. Establish the gender of each human entity according to the pronoun:
     * construction worker (human entity) = female
     * chief (human entity) = male or female
5. Translate the sentence from English to Spanish:
Spanish: <BOS>La trabajadora de la construcción pidió al jefe un aumento de su
salario.<EOS>

[INST] English: <BOS>The carpenter asked the CEO if she liked the new
furniture.<EOS> [/INST]
1. Identify the human entities in the sentence:
     * carpenter (human entity)
     * CEO (human entity)
2. Identify the pronouns referring to the human entities:
     * she (referring to the CEO)
3. Relate each pronoun to the correct human entity:
     * she (pronoun) = CEO (human entity)
4. Establish the gender of each human entity according to the pronoun:
     * carpenter (human entity) = male or female
     * CEO (human entity) = female
5. Translate the sentence from English to Spanish:
Spanish: <BOS>La carpintera preguntó a la directora general si le gustaban los
muebles nuevos.<EOS>

[INST] English: <BOS>____sentence_to_translate____<EOS> [/INST]
1.
```

Prompt with 5-shot on invented Winograd examples and SHORT chain-of-thought prompting:

```
«SYS» Translate the following sentence from English to Catalan. Proceed step
by step. «/SYS»

[INST] English: <BOS>The grower delivered tomatoes to the housekeeper, because
he needed them for dinner.<EOS> [/INST]
"he" (M) → "the housekeeper" (Male) → "mestre de casa"
Catalan: <BOS>La productora va lliurar tomàquets al mestre de casa, perquè els
necessitava per sopar.<EOS>

[INST] English: <BOS>The developer was upset with the designer because she
was disappointed with the final design.<EOS> [/INST]
"she" (F) → "the developer" (Female) → "la desenvolupadora"
Catalan: <BOS>La desenvolupadora estava molesta amb el dissenyador perquè
estava decebuda amb el disseny final.<EOS>

[INST] English: <BOS>The doctor congratulated the nurse saying his assistance
was great.<EOS> [/INST]
"his" (M) → "the nurse" (Male) → "l'infermer"
Catalan: <BOS>La doctora va felicitar a l'infermer dient que la seva
assistència va ser fantàstica.<EOS>

[INST] English: <BOS>The construction worker asked the chief for an increase
in her salary.<EOS> [/INST]
"her" (F) → "the construction worker" (Female) → "la treballadora de la
construcció"
Catalan: <BOS>La treballadora de la construcció va demanar al cap un augment
del seu sou.<EOS>

[INST] English: <BOS>The carpenter asked the CEO if she liked the new
furniture.<EOS> [/INST]
"she" (F) → "the CEO" (Female) → "la directora general"
Catalan: <BOS>La fustera li va preguntar a la directora general si li
agradaven els mobles nous.<EOS>

[INST] English: <BOS>____sentence_to_translate____<EOS> [/INST]
Catalan: <BOS>
```

```
«SYS» Translate the following sentence from English to Spanish. Proceed step
by step. «/SYS»

[INST] English: <BOS>The grower delivered tomatoes to the housekeeper, because
he needed them for dinner.<EOS> [/INST]
"he" (M) → "the housekeeper" (Male) → "amo de casa"
Spanish: <BOS>La productora entregó tomates al amo de casa, porque los
necesitaba para la cena.<EOS>

[INST] English: <BOS>The developer was upset with the designer because she
was disappointed with the final design.<EOS> [/INST]
"she" (F) → "the developer" (Female) → "la desarrolladora"
Spanish: <BOS>La desarrolladora estaba enojada con el diseñador porque estaba
decepcionada con el diseño final.<EOS>

[INST] English: <BOS>The doctor congratulated the nurse saying his assistance
was great.<EOS> [/INST]
"his" (M) → "the nurse" (Male) → "el infermero"
Spanish: <BOS>La doctora felicitó al infermero diciendo que su asistencia fue
fantástica.<EOS>

[INST] English: <BOS>The construction worker asked the chief for an increase
in her salary.<EOS> [/INST]
"her" (F) → "the construction worker" (Female) → "la trabajadora de la
construcción"
Spanish: <BOS>La trabajadora de la construcción pidió al jefe un aumento de su
salario.<EOS>

[INST] English: <BOS>The carpenter asked the CEO if she liked the new
furniture.<EOS> [/INST]
"she" (F) → "the CEO" (Female) → "la directora general"
Spanish: <BOS>La carpintera preguntó a la directora general si le gustaban los
muebles nuevos.<EOS>

[INST] English: <BOS>____sentence_to_translate____<EOS> [/INST]
Spanish: <BOS>
```

H Invented Examples following Winograd structure

The subsequent sentences (with their respective translations) are the ones created during the crafting of prompts. As you can see, they are characterized by containing more female representation and anti-stereotypical content.

EXAMPLE 1:

- English: The grower delivered tomatoes to the *housekeeper*, because he needed them for dinner.
- Catalan: La productora va lliurar tomàquets al *mestre de casa*, perquè els necessitava per sopar.
- Spanish: La productora entregó tomates al *amo de casa*, porque los necesitaba para la cena.

EXAMPLE 2:

- English: The *developer* was upset with the designer because she was disappointed with the final design.
- Catalan: La *desenvolupadora* estava molesta amb el dissenyador perquè estava decebuda amb el disseny final.
- Spanish: La *desarrolladora* estaba enojada con el diseñador porque estaba decepcionada con el diseño final.

EXAMPLE 3:

- English: The doctor congratulated the *nurse* saying his assistance was great.
- Catalan: La doctora va felicitar a *l'infermer* dient que la seva assistència va ser fantàstica.
- Spanish: La doctora felicitó al *infermero* diciendo que su asistencia fue fantástica.

EXAMPLE 4:

- English: The *construction worker* asked the chief for an increase in her salary.
- Catalan: La *treballadora de la construcció* va demanar al cap un augment del seu sou.
- Spanish: La *trabajadora de la construcción* pidió al jefe un aumento de su salario.

EXAMPLE 5:

- English: The carpenter asked the *CEO* if she liked the new furniture.
- Catalan: La fustera li va preguntar a la *directora general* si li agradaven els mobles nous.
- Spanish: La carpintera preguntó a la *directora general* si le gustaban los muebles nuevos.