Linguistic Knowledge-Infused Fine-Tuning for Mitigating Gender Bias in Machine Translation

Ernesto Garcia-Estrada[1], **Audrey Mash**[2], **Carlos Escolano**[1], **Maite Melero**[2], **Christine Basta**[3,4]

[1]Universitat Politècnica de Catalunya, Spain, [2]Barcelona Supercomputing Center, Barcelona, Spain,
[3]HiTZ Center, University of the Basque Country, Spain
[4]Faculty of Computers and Data Science, Alexandria University, Egypt
{luis.ernesto.garcia, carlos.escolano}@upc.edu,
{audrey.mash,maite.melero}@bsc.es
christine.basta@alexu.edu.eg

Abstract

Large Language Models (LLMs) achieve strong performance in machine translation (MT) but often encode gender bias, particularly when translating from non-gendered into gendered languages. This paper introduces a fine-tuning strategy to mitigate such bias in English→Spanish and English→Catalan translation. Using parameter-efficient LoRA fine-tuning, we apply *linguistic knowledge infusion*—a reasoning-based method that trains models to identify gendered referents and syntactic cues before generating translations. Experiments with **Mistral–7B** and **Salamandrata–7B** on MT-GenEval show that linguistically infused models (T2) improve gender accuracy by 15 percentage points and reduce gender gaps by 27 points in English→Spanish translation, with comparable trends for Catalan. Gains are strongest for Mistral, suggesting that explicit linguistic reasoning particularly benefits general-purpose LLMs. Overall, these results demonstrate that structured linguistic priors can enhance fairness and referential consistency in multilingual machine translation.

Keywords: machine translation, gender bias, linguistic reasoning, large language models, fine-tuning

1. Introduction

Large Language Models (LLMs) have demonstrated remarkable success across various Natural Language Processing (NLP) tasks, including text generation, question answering, and complex reasoning. This performance surge is largely attributable to training on massive datasets, which has instilled rich linguistic representations and facilitated high efficacy in zero-shot and few-shot learning paradigms (Brown et al., 2020; Chowdhery et al., 2023; Achiam et al., 2023; Touvron et al., 2023; Team et al., 2024). However, this expansive data reliance introduces a critical challenge: LLMs inherit and often amplify biases encoded within the training corpora. A pervasive manifestation of this issue is gender bias. Gender bias, in this context, refers to the systematic and disproportionate association of certain professions, adjectives, or titles with a specific gender, frequently overriding or neglecting contextual linguistic information (Kotek et al., 2023; Navigli et al., 2023).

The problem of gender bias is particularly challenging in the domain of Neural Machine Translation (NMT). A common scenario arises when translating from non-gendered (or less grammatically gendered) source languages to grammatically gendered target languages. In such cases, the NMT model lacks explicit gender-marking cues in the source text and tends to default to the statistically dominant, and often stereotypical, gender representation learned from the pre-training data when translating professions, adjectives, or titles (Stanovsky et al., 2019; Savoldi et al., 2021). Furthermore, models occasionally disregard subtle or explicit gender cues provided in the source text, indicating that the deeply ingrained, statistically driven tendency to produce a stereotypical translation overshadows the immediate contextual evidence (Basta et al., 2020; Savoldi et al., 2025).

The study of gender bias in NMT has focused extensively on certain language pairs (Costa-Jussà et al., 2022), while others remain underexplored—often due to smaller speaker populations or a scarcity of suitable corpora for comprehensive analysis (Joshi et al., 2020). This research aims to investigate a less-resourced language pair (English→Catalan) added to a high-resourced one (English→Spanish). Both Spanish and Catalan are highly grammatical gendered languages, making them excellent candidates for the explicit detection and quantification of bias.

Existing approaches to gender bias mitigation primarily in LLMs utilize inference-time strategies, such as Chain-of-Thought (CoT) prompting, which explicitly instructs the model to reason about gender markers before translation (Kaneko et al., 2024; Sant et al., 2024). While effective at runtime, a deeper solution requires interventions at the parameter level. Therefore, our work builds on the need for data- and training-based methods designed to

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 8699–8709
11–16 May 2026.

systematically correct the biases embedded in the model's core representations.

In this paper, we propose a technique to reduce gender bias in English→Spanish and English→Catalan translation using LLMs. We apply parameter-efficient fine-tuning (PEFT) to **Mistral-7B** (Jiang et al., 2023), a general purpose model, and **Salamandrata-7B** (Gilabert et al., 2025), an MT-oriented model based on LLaMA–7B, under two methodologies. The first method fine-tunes on gender-balanced parallel data containing both masculine and feminine sentence pairs. The second incorporates linguistically inspired reasoning prompts, encouraging the model to identify gendered reference markers within sentence context before producing a translation. We evaluate our systems following the methodology of (Currey et al., 2022) for gender accuracy.

In summary, this paper contributes: (1) a linguistically informed fine-tuning strategy for bias mitigation, (2) A new evaluation of gender bias in English→Catalan MT using MT-GenEval, and (3) evidence that reasoning-based fine-tuning improves fairness without harming translation quality.

2. Gender Bias Statement

In general, gender bias refers to a preference for, or prejudice against, one gender over another. In the context of translation, this bias often manifests as the default use of masculine forms or as the omission of gender markers when they are required by the target language. In NLP systems, and Neural Machine Translation (NMT) and Large Language Model (LLM) translation in particular, we define gender bias as outputs that reinforce stereotypical associations, such as communal roles for women (e.g., person-oriented, supportive) and agentic roles for men (e.g., assertive, task-focused) (Currey et al., 2022).

Building on this, we distinguish between *gender bias* in a broad sense and *gender accuracy* in translation. We define **gender accuracy** as the extent to which a machine translation output correctly reflects the gender of human referents explicitly marked or linguistically disambiguated in the source input (Currey et al., 2022). Gender accuracy can be treated as the binary task where the model is evaluated on whether the output gender matches the source gender. Beyond this, gender bias may also surface as differences in translation quality between masculine and feminine inputs, a phenomenon we refer to as the gender accuracy gap.

This distinction is particularly important in morphologically rich languages such as Spanish and Catalan, where gender is not optional but systematically encoded across nouns, adjectives, and verb forms. In these languages, even small mismatches in gender agreement can distort meaning, reduce fluency, and perpetuate bias, making accurate evaluation and mitigation strategies essential.

3. Related Work

Gender bias in MT has been widely explored in previous studies, showing the prevailing gender biases in the standard NMT systems (Stanovsky et al., 2019; Savoldi et al., 2021; Costa-Jussà et al., 2022).Multiple approaches worked in evaluating gender bias in MT and multiple benchmarks were introduced for the sake of this evaluation (Stanovsky et al., 2019; Bentivogli et al., 2020; Renduchintala et al., 2021; Levy et al., 2021), others worked on mitigating by adding context (Basta et al., 2020; Vanmassenhove et al., 2018), domain adaptation (Saunders and Byrne, 2020) and finetuning on balanced data set (Costa-jussà and de Jorge, 2020).

When Large Language Models (LLMs) began achieving strong results in machine translation (MT), research attention shifted toward evaluating gender bias in their outputs — specifically, whether LLMs continue to encode and reproduce gender stereotypes during translation. (Vanmassenhove, 2024) conducted an evaluation of ChatGPT's gender handling in Italian→English translation, assessing whether its outputs exhibit reduced bias. Their findings indicated that the model still tends to favor masculine forms, revealing persistent gender asymmetries. Subsequent studies have explored various evaluation methodologies (Sant et al., 2024; Aly et al., 2025; Kaneko et al., 2024), including adaptations of standard benchmarks, the creation of novel translation templates, and the use of prompting-based evaluation frameworks to systematically measure bias in LLM translations.

Recent work has shown that well-crafted prompting strategies can play a key role in reducing gender bias in LLM-based translation. Several studies highlight that encouraging models to explicitly reason about gender cues can lead to more balanced outputs. For instance, multi-step prompting frameworks that first identify gender markers before producing translations (Tran et al., 2025; Qiu et al., 2025), leveraging chain-of-thought reasoning (Wei et al., 2022) have been found to outperform standard zero-shot baselines (Sant et al., 2024). Collectively, these findings illustrate how LLMs' strengths in instruction-following and multi-step reasoning can be leveraged to counteract biases.

While advanced prompt engineering techniques offer an immediate fix for biased outputs, the research explored a wide array of techniques that adjust the model's parameters or intervene during inference. At the core, many successful strategies involve fine-tuning the model's internal weights us-

ing specialized data, such as corpora that are either explicitly gender-balanced or counterfactually augmented. This approach has been consistently shown to significantly reduce the problematic gender asymmetries found in generated text and machine translation (Raza et al., 2024; Zhang et al., 2024). Crucially, we can not depend on expensive full model updates generally, therefore, the rise of parameter-efficient methods like Low-Rank Adaptation (LoRA) (Hu et al., 2022) now allows for lightweight fine-tuning to mitigate gender bias. This efficiency is vital because it ensures we can adjust a model's fairness without the risk of catastrophic forgetting—thereby preserving the LLM's vast general language knowledge and capabilities (Zheng et al., 2024; Ding et al., 2024).

4. Methodology

4.1. Linguistic-Knowledge Fine-Tuning

Unlike previous fine-tuning approaches, our focus is to reduce gender bias by explicitly teaching the model the linguistic rules required to perform the task. Our method is designed to prompt models to perform a pseudo-syntactic analysis of English source sentences before producing translations. To achieve this goal, we first create a training dataset that includes each instance with step-by-step instructions guiding the model's generation. Figure 1 shows the process over the example *Maria called her brother and said she would arrive late.*:

- **Step 0: Sentence Scope.** Count the number of sentences to detect whether contextual information extends beyond the current segment.
- **Step 1: Coordination** Identify coordinating (*and*, *or*) particle.
- **Step 2: Subordination.** Identify subordinating particles (*that*, *whose*) particles, which signal shared or dependent clause structures.
- **Step 3: Actor Identification.** Locate the subjects and objects performing verbal actions—entities most likely to carry gender information (e.g., *Maria* and *brother*).
- **Step 4: Gendered Pronouns.** Detect gendered pronouns (e.g., *her*, *she*) that substitute for previously mentioned actors.
- **Step 5: Coreference Resolution.** Link each pronoun to its antecedent to determine referential gender. In the example, *her* refers to *Maria*, not *brother*.

The objective of this reasoning step was to encourage models to explicitly track gendered referents and maintain accurate gender agreement in translation.

```
Instruction:
Analyze the input sentence step-by-
step by identifying coordination and
subordination, actors, and gendered
pronouns.  Use coreference reasoning
to identify antecedents, then out-
put the translation.  Translate into
Spanish.

<source_text>
Maria called her brother and said she
would arrive late.
</source_text>

<reasoning>
0.  Count sentences:  1
1.  Count coordinate particles:  1
("and")
2.  Count subordinate particles:  0
3.  Locate the actors of the sen-
tence:  Maria (F), her brother (M)
4.  Locate the gendered pronouns:
she (F)
5.  Use coreference to determine each
pronoun's antecedent:  "she" refers
to Maria, not to her brother.
</reasoning>

<translation>
María llamó a su hermano y dijo que
ella llegaría tarde.
</translation>
```

Figure 1: Example from the Instructional Linguistic-Infused Dataset

From a linguistic perspective, gender assignment in Spanish and Catalan requires correct **coreferential mapping** between pronouns and their antecedents, often across variable syntactic distances. When these dependencies span multiple clauses or appear after intervening constituents, surface cues in the source language (English) may be insufficient for automatic alignment. By introducing intermediate reasoning steps that identify actors, pronouns, and dependency relations, the dataset trains the model to perform a form of **referential grounding**—a prerequisite for accurate gender agreement.

Moreover, coordination and subordination particles were explicitly included because they frequently introduce syntactic structures that obscure referential dependencies. For instance, in sentences such as *"The doctor called the nurse because she was late"* the pronoun *she* is ambiguous—it could refer either to the doctor or the nurse. In English, such ambiguity is tolerable, but in Spanish or Catalan the translator must commit to a gendered form (*la doctora* or *la enfermera*). By

prompting the model to analyze clause structure and resolve pronoun–antecedent relations before translation, we encourage internal representations that treat gender as a function of syntactic and discourse context rather than as a stereotypical lexical default.

In this case, a conventional model would rely on stereotypical priors (*"doctor → he", "nurse → she"*), while our linguistically infused system explicitly identifies the clause structure (*because she was late*), enumerates potential antecedents, and performs a reasoning step to determine which entity the pronoun refers to before generating the translation. This enables the model to disambiguate referential gender through syntactic reasoning rather than frequency bias.

4.2. Model Selection

For this study, we selected two open-source 7B parameter models. The choice of 7B models was motivated by their balance between computational feasibility and representational power, which makes them suitable for controlled fine-tuning experiments without requiring prohibitively large resources.

We deliberately included two different model types. **Mistral-7B**, a general-purpose decoder-only language model, serves as a baseline for evaluating whether our approach improves translation capabilities in models not originally designed for machine translation. In contrast, **Salamandrata-7B** is a decoder-only model oriented toward machine translation tasks, allowing us to examine the effect of our method on a model already adapted to translation. Comparing these two tracks provides insight into whether linguistic knowledge infusion is primarily beneficial as a bias-mitigation strategy for general-purpose LLMs, or whether it also enhances fairness in MT-specialized models.

We trained and evaluated three methodologies for each model:

- **T0 (Zero-Shot Prompting):** The original model without fine-tuning, prompted for English→Spanish and English→Catalan translation.
- **T1 (Counterfactual Fine-Tuning):** The model fine-tuned on parallel masculine and feminine sentence pairs drawn from the development split of the MT-GenEval counterfactual dataset.
- **T2 (Linguistic-Knowledge Infusion):** The model fine-tuned on an instruction dataset designed to encourage reasoning about gendered reference markers in context before generating translations. These are the same segments utilized for T1 with the added reasoning steps.

This setup enables a systematic comparison between (i) zero-shot LLM translation, (ii) data-driven fine-tuning with balanced gender examples, and (iii) linguistically informed fine-tuning designed to integrate explicit reasoning about gender.

Preliminary experiments with **Mistral-7B** on English→Catalan translation yielded non-functional outputs: the model frequently failed to generate Catalan text or produced degenerate, multilingual outputs mixing Spanish and English. This behavior indicates that Mistral lacks official Catalan support in its pretraining corpus and is therefore unsuitable for fine-tuning or evaluation in this language without prior adaptation. These results highlight that while our setting reports positive results on gender bias translation, it seems to be insufficient for language adaptation.

To ensure fair comparison and meaningful evaluation, we therefore restricted Mistral experiments to English→Spanish, where its zero-shot translation capabilities were stable and reproducible.

4.3. Fine-Tuning Methodology

We applied **parameter-efficient fine-tuning (PEFT)** using the **LoRA framework** (Hu et al., 2022). LoRA adapts pretrained models by introducing low-rank trainable matrices into existing weight projections, enabling efficient adaptation without modifying the full parameter set. This choice was motivated by the resource efficiency of LoRA and its demonstrated effectiveness in adapting large decoder-only LLMs for downstream tasks.

For both **Mistral-7B** and **Salamandrata-7B**, we fine-tuned the following modules: query, key, value, output, gate, up, and down projections. We set the rank to 16, the scaling factor (alpha) to 32, and applied no LoRA dropout. The task type was specified as causal language modeling.

Salamandrata-7B reported 36,831,232 trainable parameters of 4,969,598,976 (0.74%), while **Mistral-7B** reported 41,943,040 trainable parameters of 3,800,305,664 (1.10%) on all methods.

All fine-tuning experiments were performed using 4-bit quantization via the BitsAndBytes library (Dettmers et al., 2023). Each model was fine-tuned for one epoch with a learning rate of 2e-4, a per-device batch size of 2, and gradient accumulation steps of 16, resulting in an effective batch size of 32. All experiments were executed on a single NVIDIA A100 (80GB) GPU with mixed-precision (FP16) training enabled. Checkpoints were saved every 10 steps.

This setup ensured that each method (T1, T2) was adapted under consistent fine-tuning conditions, making performance differences attributable to the training data rather than architectural or optimization changes.

4.4. Datasets

4.4.1. Instructional Linguistic-Infused Dataset

To implement our linguistically informed fine-tuning (T2), we constructed an **instructional dataset** derived from the **MT-GenEval counterfactual and contextual development subsets** (Currey et al., 2022).

We initially created **100 seed examples manually**, which were then **augmented to 949 entries** using GPT-5 reasoning completions. 15% (145) of the generated entries were randomly sampled and manually reviewed to ensure consistency and validity of both the instructional reasoning and the final translations. Both T1 and T2 models have been trained with the same set of sentences, with the only exception of not including the linguistic steps.

For the Catalan experiments, we employed the same dataset, with the Spanish target-side sentences translated into Catalan. Since the linguistic preprocessing steps were applied exclusively to the English source sentences, no additional modifications were required.

4.4.2. MT-GenEval

We used the **MT-GenEval** benchmark (Currey et al., 2022) as the main dataset for evaluating gender bias. It contains English source sentences paired with masculine and feminine reference translations for several target languages, including Spanish. The benchmark comprises two subsets: (i) **Contextual**, where gender must be inferred from linguistic cues, and (ii) **Counterfactual**, which includes minimal masculine–feminine pairs to test consistency.

In our experiments, the English–Spanish MT-GenEval data were used for fine-tuning and evaluation under all tiers (T0–T2). We followed the official scoring procedure to compute **Gender Accuracy**—the proportion of translations preserving the correct referent gender—and **Gender Gap**, the absolute difference between masculine and feminine accuracies, complemented by COMET scores (Rei et al., 2020) to monitor translation quality.

4.4.3. MT-GenEval Catalan

To extend coverage beyond Spanish, we created a **Catalan version** of the MT-GenEval benchmark.[1] Starting from the original English–Spanish data, professional translators produced English–Catalan translations by translating the Spanish side. Ambiguous or non-gendered sentences in Catalan were flagged for review, and a second translator carried out a full revision. In a final pass, all remaining ungendered Catalan sentences were removed.

This process was applied to both the development and test sets of the contextual subset, and to the test set of the counterfactual subset. For the counterfactual development set, we used automatic translation.

In total, the Catalan extension contains 764 sentences in the contextual development set, 397 in the counterfactual development set, and 600 in the counterfactual test set. The counterfactual development set remains 1164 sentences, equal to the original English-Spanish dataset.

4.4.4. WinoMT

To evaluate the generalization and robustness of our proposed fine-tuning approach across diverse datasets and domains, we also conducted an evaluation using the WinoMT benchmark (Stanovsky et al., 2019). This dataset is specifically designed to assess gender bias in machine translation.

The WinoMT corpus comprises 3,888 English sentences such as “The nurse helped the patient while he was recovering.”. These sentences are structured such that 1,584 examples align with common gender stereotypes concerning professions (pro-stereotypical), and 1,584 examples intentionally contradict these stereotypes (anti-stereotypical). The sentences adhere to a consistent template, pairing a profession with a pronoun. The core task requires the model to correctly resolve the co-reference between the profession and the pronoun to generate the appropriate gendered term in the translation. No reference translations are provided. Instead, accuracy is determined through an automated process. First, Alignment is computed between the source and target sentences. Then, a Part-of-Speech (POS) tagger is subsequently employed to detect the translated gender.

4.5. Experimental Framework

Our experiments were designed to evaluate the effect of linguistic knowledge infusion on gender bias mitigation across two target languages (Spanish and Catalan) and two model families (**Mistral-7B** and **Salamandrata-7B**). Each model was evaluated under the three methods.

We evaluate each model on two MT-GenEval subsets:

- **Contextual:** Sentences where gender must be inferred from surrounding linguistic context.
- **Counterfactual:** Paired sentences differing only in gender, used to assess consistency across masculine and feminine forms.

For translation quality, we report **COMET** (Rei et al., 2020) scores. COMET offers a reference-based semantic evaluation of translation quality.

[1] available at https://huggingface.co/datasets/LangTech-MT/geneval_catalan

Model	Lang.	Task	Tier	F Acc. ↑	M Acc. ↑	Gender Gap ↓	Gender Acc. ↑
Mistral-7B	ES	CTX	T0	0.451	0.936	0.485	0.694
			T1	0.698	0.931	0.233	0.815
			T2	**0.733**	**0.949**	**0.216**	**0.841**
Mistral-7B	ES	CFT	T0	0.690	0.857	0.167	0.773
			T1	0.773	**0.867**	0.094	**0.820**
			T2	**0.787**	0.840	**0.053**	0.814
Salamandrata-7B	ES	CTX	T0	0.874	**0.962**	0.088	0.918
			T1	0.789	0.955	0.166	0.872
			T2	**0.901**	0.953	**0.052**	**0.927**
Salamandrata-7B	ES	CFT	T0	0.830	0.870	0.040	0.850
			T1	0.793	0.857	0.064	0.825
			T2	**0.850**	**0.860**	**0.010**	**0.855**
Salamandrata-7B	CA	CTX	T0	0.919	**0.968**	0.049	0.944
			T1	0.839	0.966	0.127	0.902
			T2	**0.935**	0.955	**0.020**	**0.945**
Salamandrata-7B	CA	CFT	T0	0.797	0.867	0.070	0.832
			T1	0.727	**0.877**	0.150	0.802
			T2	**0.830**	0.857	**0.027**	**0.844**

Table 1: Unified results across **Contextual (CTX)** and **Counterfactual (CFT)** tasks. F Acc. and M Acc. refer to feminine and masculine accuracy respectively. Gender Gap is computed as |F Acc. – M Acc.|, and Gender Acc. represents overall accuracy across both genders. ↑ indicates higher is better; ↓ lower is better.

To assess fairness, we compute two gender-related metrics:

- **Gender Accuracy:** The proportion of translations that correctly preserve the referent's gender.
- **Gender Gap:** The absolute difference between masculine and feminine accuracy scores.

All metrics are reported for both the contextual and counterfactual subsets, enabling direct comparison of accuracy, balance, and translation quality under each fine-tuning regime.

For consistent comparison across all experiments, WinoMT results are reported using the same set of metrics over all models studied.

5. Results

We report results for all models on the **Contextual** and **Counterfactual** subsets of the MT-GenEval benchmark for both Spanish (ES) and Catalan (CA) in table 1.

Across all settings, the linguistically infused fine-tuning (**T2**) consistently reduces Gender Gap. The effects, however, vary in across models and languages, reflecting different inductive biases and pretraining backgrounds.

For **Mistral–7B**, the T2 configuration yields the strongest improvements. On the **Contextual (CTX)** task, feminine accuracy increases from 0.45 to 0.73, masculine accuracy remains high (0.93→0.95), and the gender gap narrows sharply (0.49→0.22). Overall gender accuracy rises from 0.69 to 0.84 (0.15 improvement), confirming that reasoning-based fine-tuning enhances referential gender tracking without sacrificing performance. Similarly, in the **Counterfactual (CFT)** task, the model achieves a steady improvement in feminine accuracy (0.69→0.79) and a reduction in gender gap (0.17→0.05), indicating more balanced method of explicitly gendered pairs.

For **Salamandrata–7B**, changes are more moderate. On the CTX task, feminine accuracy rises slightly (0.87→0.90), and the gender gap narrows (0.09→0.05), though masculine accuracy remains nearly saturated at 0.95. In the CFT task, feminine accuracy increases modestly (0.83→0.85), and the gap decreases from 0.04 to 0.01. These stable yet consistent gains suggest that Salamandrata -7B, trained for machine translation tasks, benefits from the structured reasoning step mainly as a form of balancing gender prediction rather than learning to perform coreference between the elements in the sentence..

For **Salamandrata–7B** for Catalan, the pattern mirrors Spanish. The CTX task shows small improvements in gender accuracy (0.94→0.95) and a marked reduction in gap (0.05→0.02). In the CFT task, the feminine accuracy increases from 0.73 to 0.83, and the gap is minimized from 0.15 to 0.03, confirming that linguistic cues generalize effectively across typologically related Romance languages.

COMET scores remain broadly stable across fine-tuning tiers (+0.02 / 0.01), confirming that lin-

Model	Lang.	Tier	F Acc. ↑	M Acc. ↑	Gender Gap ↓	Gender Acc. ↑
Mistral-7B	ES	T0	0.400	0.638	0.238	0.515
		T1	0.422	0.650	0.228	0.531
		T2	**0.537**	**0.677**	**0.14**	**0.580**
Salamandrata-7B	ES	T0	0.796	**0.804**	0.008	**0.745**
		T1	0.498	0.675	0.177	0.568
		T2	**0.799**	0.798	**-0.001**	0.740
Salamandrata-7B	CA	T0	**0.731**	**0.735**	0.004	**0.637**
		T1	0.465	0.640	0.175	0.496
		T2	0.713	0.713	**0.000**	0.619

Table 2: WinoMT results for Mistral–7B and Salamandrata–7B on English→Spanish (ES) and English→Catalan (CA). Acc. represents overall accuracy, F Acc. and M Acc. correspond to feminine and masculine accuracy respectively, Gender Gap is |F Acc. – M Acc.|.

guistic reasoning does not compromise adequacy or fluency. For **Mistral–7B**, COMET improves from 0.85 (T0) to 0.87 (T2) on the CTX task, indicating that reasoning-based fine-tuning enhances referential capabilities while maintaining translation quality. In the CFT setting, COMET remains steady (0.84 → 0.83), showing that fairness improvements do not come at the expense of translation adequacy. For **Salamandrata–7B** in Spanish, COMET values decrease minimally (0.88 → 0.86) across CTX and CFT tasks, while Salamandrata–7B for Catalan exhibits a modest decline (0.87 → 0.83) alongside gains in gender accuracy (0.94 → 0.95).

To evaluate cross-benchmark robustness, we additionally report results on the **WinoMT** dataset (Table 2). Overall trends mirror those observed in MT-GenEval: the linguistically infused fine-tuning (T2) consistently improves gender balance, increasing accuracy and reducing gender gaps without degrading translation quality. Notably, Mistral–7B again benefits most from T2, improving overall WinoMT accuracy from 0.52 (T0) to 0.58, while narrowing the gender gap from 0.24 to 0.14. Salamandrata–7B shows smaller yet consistent gains across both Spanish and Catalan, confirming that linguistic reasoning generalizes to out-of-domain, profession-centered data.

6. Discussion

Our results show that linguistically infused fine-tuning (T2) consistently narrows the gender gap across both models and languages, while maintaining competitive translation quality. These findings support the hypothesis that explicit linguistic reasoning helps LLMs better track referential gender information throughout the translation process. We hypothesise that these gains arise because T2 encourages a lightweight pseudo-syntactic analysis prior to generation: the model is prompted to detect actors, pronouns, and coordinating or subordinate structures, then use those cues to resolve coreference and propagate referential gender across clauses and sentences.

The effect is particularly evident for Mistral–ES, which achieved the largest overall gains. The increase in accuracy across both genders together with the reduction in the gender gap suggests that linguistic infusion serves as an inductive bias, injecting structural priors that promote more controlled, context-aware generation. In this way, feminine forms become more reliably realized, masculine dominance diminishes, and overall accuracy improves. This pattern aligns with prior research showing that translation systems systematically over-generate masculine forms. This bias is rooted in the asymmetries of training data and is thus counteracted in T2 by the enforcement of syntactic reasoning.

In contrast, Salamandrata's smaller gains reflect its translation specialization and lower initial bias. Its translation-focused pretraining likely encodes stronger co-reference resolution and gender-handling behaviors, leaving less headroom for T2 to change surface outputs; nonetheless, the steady gains validate that the reasoning-based approach is broadly applicable even to models with robust pretraining.

When comparing tiers, T0 (zero-shot) often yields competitive translation quality but larger gender asymmetries. T1 (balanced fine-tuning) partially mitigates bias but tends to overfit to surface-level gender patterns in limited data, leading to noisier and sometimes degraded performance. T2 achieves a better equilibrium, introducing explicit linguistic cues that regularize gender representation without eroding fluency or adequacy. In this sense, reasoning-based fine-tuning acts as a form of structural regularization, stabilizing outputs through explicit referential awareness.

To illustrate the practical impact of linguistic reasoning, Table 3 presents representative examples from both the Spanish Counterfactual subset and the Spanish WinoMT output across models. These examples were selected to show typical improve-

Dataset	System	Gender	Source Sentence	T0 (Baseline)	T2 (Linguistic Reasoning)
CTX	Mistral	F	The new colonel-in-chief took up her appointment at a ceremony on March 17, 2007, at the Regimental Headquarters in Edmonton.	***El nuevo coronel-en-jefe*** *asumió su cargo en una ceremonia el 17 de marzo de 2007, en la Cabecera de Regimiento en Edmonton.*	***La nueva coronel-jefa*** *asumió su cargo en una ceremonia el 17 de marzo de 2007 en el Cuartel General Regimental de Edmonton.*
	Salamandrata	M	The priests of Vesta, known as Vestal Virgins, administered his temple and watched the eternal fire.	***Las sacerdotisas*** *de Vesta, conocidas como vírgenes vestales, administraban su templo y vigilaban el fuego eterno.*	***Los sacerdotes*** *de Vesta, conocidos como vírgenes vestales, administraban su templo y vigilaban el fuego eterno.*
WinoMT	Mistral	M	The developer dropped by the nurse and discussed with him about the surgery.	*El desarrollador se detuvo con* ***la enfermera*** *y discutió con él sobre la cirugía.*	*El desarrollador se detuvo con* ***el enfermero*** *y se discutió con él sobre la cirugía.*
	Salamandrata	F	The mechanic called the editor and told her to get the car back.	***La mecánica*** *llamó* ***al editor*** *y le dijo que recuperara el auto.*	***El mecánico*** *llamó a* ***la editora*** *y le dijo que recuperara el coche.*

Table 3: Examples from the CTX and WinoMT Spanish tasks. Bold highlights the forms changed between T0 and T2.

ments in referential gender agreement after T2 fine-tuning. As can be seen in these examples, mistakes at the T0 level often relate to stereotypically masculine or feminine contexts or occupations, while the linguistic reasoning introduced in T2 allows the model's context to override biases learnt from pretraining data.

Taken together, these findings demonstrate that linguistic reasoning provides a viable mechanism for embedding structured grammatical knowledge into LLMs. Beyond mitigating gender bias, this approach may generalize to other aspects of linguistic control such as politeness, formality, and discourse coherence in which interpretability and fairness converge. Future work will be needed to test whether reasoning-based fine-tuning can transfer effectively to such domains.

7. Conclusions

Our study explored how linguistic knowledge infusion can mitigate gender bias in large language model based translation. By embedding reasoning prompts that explicitly target referential and morphosyntactic gender cues, we demonstrated consistent improvements in gender accuracy and fairness without heavily compromising translation quality.

The two models exhibited distinct behaviors under our methods. For the general-purpose Mistral, linguistically infused fine-tuning (T2) led to the largest gains, confirming that explicit linguistic supervision can guide models lacking translation-specific pretraining toward more controlled and context-sensitive generation. In contrast, the translation-oriented Salamandrata, which already encoded strong gender agreement tendencies, displayed smaller yet consistent improvements. This indicates that the benefits of linguistic reasoning are most pronounced when the model's inductive biases are not already shaped by translation objectives.

However, our findings also show that the method's effectiveness depends on the interaction between model pretraining and task type. In the contextual task, where gender cues must be inferred across longer or more syntactically complex spans, Salamandrata underperformed under T1 and showed limited benefit from T2. This suggests that fine-tuning on templatic or reasoning-augmented short sentences may not sufficiently transfer to discourse-level gender resolution. Future iterations could expand the dataset to include multi-sentence reasoning and broader discourse contexts to strengthen generalization. These findings correlate with (Zaranis et al., 2024), showing that general LLMs without translation training show stronger positional bias for source sentence contributions.

Overall, this work provides evidence that explicit linguistic reasoning can enhance fairness and referential consistency in LLM-based translation. Extending this approach to phenomena such as politeness, formality, or non-binary gender represen-

tation represents a promising direction for building more interpretable and equitable multilingual systems.

8. Limitations

While our approach demonstrates consistent improvements in gender accuracy, several limitations should be acknowledged.

First, the linguistic-infused fine-tuning dataset was derived and augmented from the MT-GenEval development set, which primarily contains short, single-sentence examples. As a result, the training material offers limited syntactic and discourse variety, constraining the model's exposure to gender dependencies that operate across clauses or sentence boundaries. This may partially explain the reduced generalization observed in complex or multi-sentence contexts.

Second, we observe that longer or syntactically intricate segments tend to yield less stable gender agreement. This degradation likely arises from accumulated contextual uncertainty in decoder-only LLMs, which can obscure gender cues and affect both translation fluency and referential consistency.

Finally, our evaluation framework is restricted to binary gender distinctions, reflecting the structure of existing benchmarks. Future research should expand this analysis to include non-binary and neutral gender representations, which are increasingly relevant in both Spanish and Catalan.

9. Ethical Statement

This research adheres to the LREC 2026 Code of Ethics. All datasets used or derived in this study are publicly available or were created synthetically for research purposes. No personally identifiable information or sensitive data were processed. The Catalan extension of MT-GenEval was produced by professional translators who provided informed consent and were compensated under standard contractual conditions.

The linguistically infused dataset was partially generated with GPT-based reasoning completions. We conducted manual review and correction to ensure quality and remove potentially biased or inappropriate outputs. Our work focuses on mitigating gender bias in translation; however, residual forms of bias may persist, and we encourage responsible reuse of our models and data.

All resources released from this study are intended for non-commercial research use under the same licensing conditions as their source datasets.

10. Acknowledgements

This work has been promoted and financed by the Government of Catalonia through the Aina project. It has also been supported by ALIA Models Development Project, Resolution of the Secretary of State for Digitalization and Artificial Intelligence (SEDIA) 19/08/2024, within the framework of the National Language Technologies Plan – ENIA 2024, funded by the Ministry for Digital Transformation and the Civil Service (MTDFP), by the Recovery, Transformation and Resilience Plan (PRTR), and by the European Union – NextGenerationEU. Carlos Escolano has been funded by the Ministerio de Ciencia, Innovación y Universidades and the Agencia Estatal de Investigación through the project LLM4LS (PID2024-157855OA-C33).

11. Bibliographical References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ranwa Aly, Yara Allam, Rana Gaber, and Christine Basta. 2025. ArGAN: Arabic gender, ability, and nationality dataset for evaluating biases in large language models. In *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 256–267, Vienna, Austria. Association for Computational Linguistics.

Christine Basta, Marta R Costa-jussà, and José AR Fonollosa. 2020. Towards mitigating gender bias in a decoder-based neural machine translation model by adding contextual information. In *Proceedings of the Fourth Widening Natural Language Processing Workshop*, pages 99–102.

Luisa Bentivogli, Beatrice Savoldi, Matteo Negri, Mattia Antonino Di Gangi, Roldano Cattoni, and Marco Turchi. 2020. Gender in danger? evaluating speech translation technology on the must-she corpus. *arXiv preprint arXiv:2006.05754*.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Marta R. Costa-jussà and Adrià de Jorge. 2020. Fine-tuning neural machine translation on gender-balanced datasets. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 26–34, Barcelona, Spain (Online). Association for Computational Linguistics.

Marta R Costa-Jussà, Carlos Escolano, Christine Basta, Javier Ferrando, Roser Batlle, and Ksenia Kharitonova. 2022. Interpreting gender bias in neural machine translation: Multilingual architecture matters. In *Proceedings of the aaai conference on artificial intelligence*, volume 36, pages 11855–11863.

Anna Currey, Maria Nadejde, Raghavendra Reddy Pappagari, Mia Mayer, Stanislas Lauly, Xing Niu, Benjamin Hsu, and Georgiana Dinu. 2022. Mt-geneval: A counterfactual and contextual dataset for evaluating gender accuracy in machine translation. 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 4287–4299. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. 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*.

Zhoujie Ding, Ken Ziyu Liu, Pura Peetathawatchai, Berivan Isik, and Sanmi Koyejo. 2024. On fairness of low-rank adaptation of large models. *arXiv preprint arXiv:2405.17512*.

Javier García Gilabert, Xixian Liao, Severino Da Dalt, Ella Bohman, Audrey Mash, Francesca de Luca Fornaciari, Irene Baucells, Joan Llop, Miguel Claramunt Argote, Carlos Escolano, and Maite Melero. 2025. From SALAMANDRA to SALAMANDRATA: BSC submission for WMT25 general machine translation shared task. *CoRR*, abs/2508.12774.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Masahiro Kaneko, Danushka Bollegala, Naoaki Okazaki, and Timothy Baldwin. 2024. Evaluating gender bias in large language models via chain-of-thought prompting. *arXiv preprint arXiv:2401.15585*.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference*, pages 12–24.

Shahar Levy, Koren Lazar, and Gabriel Stanovsky. 2021. Collecting a large-scale gender bias dataset for coreference resolution and machine translation. *arXiv preprint arXiv:2109.03858*.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: origins, inventory, and discussion. *ACM Journal of Data and Information Quality*, 15(2):1–21.

Hongye Qiu, Yue Xu, Meikang Qiu, and Wenjie Wang. 2025. Dr.gap: Mitigating bias in large language models using gender-aware prompting with demonstration and reasoning.

Shaina Raza, Ananya Raval, and Veronica Chatrath. 2024. Mbias: Mitigating bias in large language models while retaining context. *arXiv preprint arXiv:2405.11290*.

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.

Adithya Renduchintala, Denise Diaz, Kenneth Heafield, Xian Li, and Mona Diab. 2021. Gender

bias amplification during speed-quality optimization in neural machine translation. *arXiv preprint arXiv:2106.00169*.

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

Danielle Saunders and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. In *Proceedings of ACL*.

Beatrice Savoldi, Jasmijn Bastings, Luisa Bentivogli, and Eva Vanmassenhove. 2025. A decade of gender bias in machine translation. *Patterns*, 6(6):101257.

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

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of ACL*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Sharan Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Van-Hien Tran, Huy-Hien Vu, Hideki Tanaka, and Masao Utiyama. 2025. Can explicit gender information improve zero-shot machine translation? In *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 171–181.

Eva Vanmassenhove. 2024. Gender bias in machine translation and the era of large language models. In *Gendered Technology in Translation and Interpreting*, pages 225–252. Routledge.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22.

Emmanouil Zaranis, Nuno Miguel Guerreiro, and André F. T. Martins. 2024. Analyzing context contributions in llm-based machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 14899–14924. Association for Computational Linguistics.

Tao Zhang, Ziqian Zeng, Yuxiang Xiao, Huiping Zhuang, Cen Chen, James Foulds, and Shimei Pan. 2024. Genderalign: An alignment dataset for mitigating gender bias in large language models. *arXiv preprint arXiv:2406.13925*.

Jiawei Zheng, Hanghai Hong, Feiyan Liu, Xiaoli Wang, Jingsong Su, Yonggui Liang, and Shikai Wu. 2024. Fine-tuning large language models for domain-specific machine translation. *arXiv preprint arXiv:2402.15061*.