Fine-Tuning vs Prompting Techniques for Gender-Fair Rewriting of Machine Translations

Paolo Mainardi, Federico Garcea and Alberto Barrón-Cedeño
DIT, Università di Bologna, Forlì, Italy
paolo.mainardi3@studio.unibo.it, {federico.garcea2, a.barron}@unibo.it

Abstract

Increasing attention is being dedicated by the NLP community to gender-fair practices, including emerging forms of non-binary language. Given the shift to the prompting paradigm for multiple tasks, direct comparisons between prompted and fine-tuned models in this context are lacking. We aim to fill this gap by comparing prompt engineering and fine-tuning techniques for gender-fair rewriting in Italian. We do so by framing a rewriting task where Italian gender-marked translations from English gender-ambiguous sentences are adapted into a gender-neutral alternative using direct non-binary language. We augment existing datasets with gender-neutral translations and conduct experiments to determine the best architecture and approach to complete such task, by fine-tuning and prompting seq2seq encoder-decoder and autoregressive decoder-only models. We show that smaller seq2seq models can reach good performance when fine-tuned, even with relatively little data; when it comes to prompts, including task demonstrations is crucial, and we find that chat-tuned models reach the best results in a few-shot setting. We achieve promising results, especially in contexts of limited data and resources.

1 Introduction

Current practices in many languages involve the use of the masculine gender as a generic form (Sczesny et al., 2016), a norm —which we refer to as masculine generics (MG)— that may result in the erasure of other gender identities, including both women and non-binary (NB) people[1]. NLP models, based on dominant linguistic practices, reproduce this behavior (Costa-jussà et al., 2023). The reliance on MG implies the under-representation of people who do not identify as men (Dev et al., 2021), as well as an increased effort and a reduced quality of service for them (Savoldi et al., 2024a); examples of representational and allocational harms (Blodgett et al., 2020), respectively.

In this paper, we acknowledge the limited availability of NLP resources for NB language and, especially, the lack of a shared NB "grammar" for Italian, so we produce original, detailed guidelines for the use of an Italian NB language paradigm. They are included in Appendix A and are meant to serve as a basis for future works and for further discussions around the topic, with the ultimate aim of fostering the recognition of NB identities in Italian language technologies. Our guidelines were written by one of the authors of this paper, and they were validated by experienced Italian linguists.

We also define a rewriting task based on replacing masculine and feminine gender marks with NB endings, inspired by the Fair reformulation task described by Frenda et al. (2024). We focus on existing translations of gender-ambiguous English sentences: while the reference translations we collected use masculine or feminine gender marks, our goal is to obtain new translations that preserve the gender neutrality of the source sentences. The spans we aim to replace include examples of gendered language used in an *overextended* or *generic* way, as defined by Rosola et al. (2023).

To do so, we manually rewrite existing Italian translations that use gendered language so that they maintain the gender ambiguity of the corresponding English source sentences. We then use the original translations as inputs and our rewritten translations as labels to expand on recent related works on gender-fair NLP by comparing transfer and in-context learning on both encoder-decoder and decoder-only architectures.

While our approach is essentially monolingual, it is meant to be applied not only to Italian texts

[1] We use *non-binary* as an umbrella term to refer to individuals who do not recognize themselves in the gender binary typical of Western society, consisting of a clear distinction between the male and female genders, as intended for example by Kendall (2024).

Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 320–337
August 1, 2025 ©2025 Association for Computational Linguistics

which use gendered language and MG, but also to gender-marked translations provided by machine translation (MT) or human translators (in this case, it can be defined as a post-editing task). Sentences obtained this way could potentially be used to make texts inclusive of all gender identities, as well as to train future NLP models on more diverse datasets.

We release the data we used in our experiments, the outputs of our models, and the main scripts used to carry out this study.[2]

The rest of the paper is distributed as follows. Section 2 provides background on gender and language. We discuss our conceptualization of gender bias in Section 3. We present related work in Section 4. Section 5 describes our approach and Section 6 discusses our results. Section 7 draws conclusions and discusses future work.

2 Background

The relationship between gender and language is especially relevant in the context of translation, due to the need to resolve discrepancies between different gender systems (Nissen, 2002). We focus on English and Italian. English is a representative of notional gender languages, where only a few classes — mostly pronouns — are gender-marked, while nouns, verbs, and adjectives are usually gender-ambiguous; i.e., they can refer to people of any gender identity. Italian is a grammatical gender language, where most words are gender-marked and, usually, all components of a noun or verb phrase have to be inflected according to the same gender. Refer to Sczesny et al. (2016) for an overview of the grammatical gender systems of various languages. MG are also used in the context of translation; for example, referents whose gender is ambiguous in English are often translated as masculine in Italian.

From the 1970s, feminist movements initiated the debate around the social dynamics underlying the use of MG and gendered language (see Pusterla, 2019; Ludbrook, 2022). More recently, the need for alternative solutions has been reiterated by works on cognitive biases resulting from the extended use of masculine words (Gygax et al., 2008; Xiao et al., 2023). Interest in NB language increased starting in the 2010s (Pusterla, 2019; Ludbrook, 2022).

	Specific
Corpus ID	MT-GenEval geneval-test-954
Source	That led to a second career as a **writer**.
Gendered	Ciò **la** portò a intraprendere una carriera parallela come **scrittrice**.
Gender-neutral	Ciò **lə** portò a intraprendere una carriera parallela come **scrittorə**.
	Generic
Corpus ID	mGeNTE ep-en-it-3332
Source	[. . .] no audits were carried out by **the financial controller** [. . .]
Gendered	[. . .] **il controllore finanziario** non effettuava audit [. . .]
Gender-neutral	[. . .] **lə controllorə finanziariə** non effettuava audit [. . .]
	Group
Corpus ID	mGeNTE ep-en-it-13688
Source	**Citizens** must of course be **protected**.
Gendered	**I cittadini** devono essere **tutelati**.
Gender-neutral	**ə cittadinə** devono essere **tutelatə**.

Table 1: Examples of binary generics based on the type of human referent(s). Bolded expressions refer to human beings, and they identify the scope of our task.

NB language includes various sets of linguistic practices aimed at representing NB and gender-non-conforming identities. This is especially challenging in highly inflected languages with a binary grammatical gender system (e.g., Italian: Comandini, 2021; Scotto Di Carlo, 2020; French: Knisely, 2020, Ashley, 2019; Spanish: López, 2019), although challenges exist in all languages due to a lack of widespread recognition of such identities (for example, in Swedish: Gustafsson Sendén et al., 2015).

Such practices can be categorized into direct and indirect strategies (López, 2019). Indirect non-binary language (INL) mainly aims at avoiding gendered expressions by using synonyms and paraphrases, while direct non-binary language (DNL) introduces morpho(phono)logical changes to explicitly recognize NB identities. Both are used to avoid misgendering (i.e. addressing someone with a gender they do not identify as) and masculine generics. One of the main differences between INL and DNL is that the latter was born as a militant practice within queer communities (see Acanfora, 2022 and Gheno, 2022b for the case of Italian) and its use is still controversial (at least in Italy: Formato and Somma, 2023; Sulis and Gheno, 2022).

3 Bias statement

In the context of this paper, we consider as biased behavior the use of both masculine and feminine gender marks whenever referring to specific individuals whose gender identity is unknown or NB, to groups of people that may include individuals

[2] https://github.com/TinfFoil/dnl_rewriter

of various gender identities, or to generic referents that do not identify a specific individual. We define all these cases collectively as *binary generics*, as they all imply taking the gender binary as the general norm. Table 1 provides one example from our dataset for each of these cases.

A translation is biased according to our definition if it contains one or more masculine or feminine gender marks when the corresponding source text does not, as defined by Piergentili et al. (2023) in their desiderata for gender-neutral translation. In such cases, the translation should be rewritten, and that is our goal in this study. We consider binary generics to be harmful as they erase the existence of people whose gender identity does not adhere to the gender binary, and as they imply the risk of misgendering individuals who do not recognize themselves in a binary gender (Dev et al., 2021).

4 Related work

4.1 Gender-fair language in NLP

While most early works in the area mainly focused on binary gender (Dev et al., 2021), coverage of NB language has increased in recent years.

Earlier approaches to gender-fair NLP include reducing the association of certain words to the masculine or feminine gender in an embedding space (Bolukbasi et al., 2016), or making training data more balanced through counterfactual data augmentation (CDA; Lu et al., 2019), mainly by converting gender marks from masculine to feminine and vice versa. While the former was proven to be a superficial solution (Gonen and Goldberg, 2019), data balanced through some form of CDA has been extensively used to create evaluation benchmarks (Stanovsky et al., 2019; Bentivogli et al., 2020; Vanmassenhove et al., 2021; Currey et al., 2022) or to fine-tune models on downstream tasks or specific datasets (Saunders and Byrne, 2020; Costa-jussà and de Jorge, 2020).

Gender-fair post-editing has been proposed as a solution (Lardelli and Gromann, 2023) and it requires less data, since it relies on robust models that can provide high-quality translations, although biased. Crucially, such a task can be automated; e.g., Jain et al. (2021). Similarly to Sun et al. (2021), Vanmassenhove et al. (2021) obtain training data through a rule-based algorithm, then train a model on a rewriting task; Bartl and Leavy (2024) use their rewriting system to create a gender-fair dataset and fine-tune large language models (LLMs) on it. This task can generally be referred to as gender-fair rewriting or reformulation (Frenda et al., 2024).

Recently, the popularization of conversational LLMs has brought attention on prompting techniques for obtaining gender-fair texts. The novelty of many gender-fair communication strategies and the limited availability of task-specific datasets make prompting a very promising approach in this area. Sánchez et al. (2024), Vanmassenhove (2024), Savoldi et al. (2024b), and Piergentili et al. (2024) all compare different prompting strategies for gender-fair MT: specifically, the former two aim at obtaining all possible combinations of (binary) gender marks in the translations of gender-ambiguous source sentences, while the latter two focus on INL and DNL, respectively. Finally, Sant et al. (2024) use a similar approach to reduce gender stereotyping in generative models.

We adopt a NB perspective and focus on automatic gender-fair post-editing or rewriting in Italian. We use this approach to directly compare task-specific fine-tuning and zero- and few-shot prompting for gender-fair NLP. To the best of our knowledge, the only existing work partially comparable to our own is the one by Piergentili et al. (2024), who test their (prompted) models on the same test set and using the same metrics. However, their study is fundamentally different in that it focuses on translation rather than on rewriting.

4.2 Adaptation methods for transformers

Prompting is generally associated with causal decoder-only models since their emergence in the NLP landscape (Radford et al., 2019; Brown et al., 2020), while the main approach for typically smaller encoder-decoder architectures entails further training the model on an unseen task by updating its weights (see Wang et al., 2022).

The success of autoregressive LLMs and the prompting paradigm is related to in-context learning, an emergent ability (Wei et al., 2022b) that allows these models to reach state-of-the-art performance on unseen tasks when provided with a natural language description, which can be followed by a small set of examples. Few-shot prompting (Brown et al., 2020) is now an established method for adapting LLMs to specific tasks. Some authors propose methods that depart from this typical setting. While Lee et al. (2024) successfully leverage in-context learning for seq2seq models through curated prompting techniques, Zhang et al. (2023) compare (few-shot) prompting and fine-tuning on decoder-only models for MT, demonstrating the benefits of updating model weights by leveraging parameter-efficient fine-tuning for this task.

With this work, we aim to contribute to research on gender-fair NLP by directly comparing both approaches as applied to both architectures.

5 Method

We carry out experiments with various models, with the ultimate goal of obtaining a gender-neutral alternative for each translation in our dataset, as exemplified in Table 1. Our experiments are aimed at verifying which model architecture (seq2seq or autoregressive) and adaptation method (fine-tuning or in-context learning) is most suitable for our task. To create the labels for our experiments, we use the schwa DNL paradigm, currently one of the most commonly used NB language strategies in Italian (Comandini, 2021). To have a coherent basis for our own reformulations, we define guidelines on the use of the schwa, covering different parts of speech and noun classes (they appear in Appendix A). Our approach is based on informal interactions with the interested communities, as well as on sources that directly come from them or are involved with them, and that propose some systematization of this paradigm:[3]

- The *Italiano Inclusivo* (*Inclusive Italian*) project, which is among the early promoters of the schwa as an Italian NB neomorpheme and provides a guide for its use[4];

- The *Gender in Language* project, which includes an overview of current NB language strategies used in Italian and other languages (Papadopoulos et al., 2025).

Our guidelines are also similar to how Piergentili et al. (2024) use this paradigm in their dataset.

In this section, we describe the data we used (5.1), the models we included (5.2), and the experiments we carried out (5.4).

5.1 Data

We use the pairs of Italian gender-marked translations and corresponding schwa reformulations as a fine-tuning dataset and as the source of the examples used in our prompts.

The gender-marked reference translations were selected from the Italian sections of two datasets meant for the evaluation of gender-fair language: MT-GenEval (Currey et al., 2022)[5] and mGeNTE (Savoldi et al., 2025).[6] We choose these datasets as they contain gender-neutral source sentences and corresponding gender-marked translations, which fits our intended use case. Both allow for easy control over the rewriting task as each reference translation contains only one set of gender marks (masculine or feminine) for all human entities.

Specifically, the Contextual set in MT-GenEval contains 1,559 gender-ambiguous source sentences and for each of these, two alternative translations, one masculine and one feminine. We extract a gender-balanced subset by selecting only the feminine translations from the first half of the dataset and only the masculine translations from the second half. Eventually, 31 sentences were removed from the original set, because they could not be rewritten using the schwa (e.g., because they contain fixed gender nouns[7]), thus leaving us with 1,528 sentence pairs.

As for mGeNTE, we collect input sentences from the Set-N subset, which consists of 750 gender-ambiguous source sentences, each paired with two Italian translations: one gender-marked (either masculine or feminine) and one using INL. We use all gender-marked reference translations contained in this subset; we do not control the distribution of masculine and feminine gender marks in these sentences.

For each sentence pair in the combined dataset, we add a schwa reformulation —manually crafted based on our guidelines by one of the authors of this paper, supervised by a linguist experienced in inclusive communication— to serve as target sentences or labels in our experiments. The resulting dataset contains 2,278 pairs, each consisting of a masculine- or feminine-marked sentence and its schwa-based NB version. We leave out 10% of these for validation when fine-tuning, while for

[3] The schwa was also adopted by the Italian publisher effequ as a solution for the translation of texts using gender-fair language. Their choice was documented at https://www.effequ.it/schwa/ [Last accessed 26 February 2025], now archived at https://web.archive.org/web/20250214054644/https://www.effequ.it/schwa/ [Last accessed 8 June 2025]

[4] https://italianoinclusivo.it/scrittura/ [Last accessed 8 June 2025]

[5] We use the version hosted at https://huggingface.co/datasets/gsarti/mt_geneval.

[6] https://huggingface.co/datasets/FBK-MT/mGeNTE

[7] As defined in Appendix A.3.

each prompt, we select a random sample of pairs from the whole dataset.

5.2 Models

We choose models representative of both encoder-decoder and decoder-only architectures. Two of them (BLOOM and IT5) are base pretrained models, while the others underwent some kind of fine-tuning prior to this study.

IT5 (Sarti and Nissim, 2024) was the first encoder-decoder model specifically pre-trained on Italian.[8] It is based on the original T5 by Raffel et al. (2020), whose distinguishing feature is multi-task pretraining, which should give these models an advantage in the context of novel tasks like ours. We use the `base` and `large` versions of IT5, with 220 and 738 million parameters, respectively.

mT0 is based on mT5 —multilingual T5 (Xue et al., 2021).[9] It represents a special case as it has an encoder-decoder architecture, but it was further trained on zero-shot instructions (Muennighoff et al., 2023), on top of its multi-task and multi-lingual capabilities. We use the `base` and `large` versions, with 580 million and 1.2 billion parameters respectively.

BLOOM is an open and multilingual autoregressive model (BigScience Workshop, 2023).[10] As the base model is not suitable for inference without further training, for the prompting experiments we use its instruction-tuned (Wei et al., 2022a) version: **BLOOMZ**[11]; released by Muennighoff et al. (2023) alongside mT0. We use the 560-million version of both models, as well as the 7.1-billion version of BLOOMZ for some experiments.

We also include two chat-tuned models in our prompting experiments: **Llama 3.1** (Llama Team, 2024) and **Ministral**[12], multilingual decoder-only further trained on multi-turn conversation. We use the 8 B parameter `Instruct` version of each.[13]

5.3 Evaluation

We evaluate our models on Neo-GATE[14], a dataset meant to be easily adaptable to any Italian DNL strategy based on neomorphemes (Piergentili et al., 2024). We adapted it to a formalized version of our guidelines. The process also involved slightly altering some of the sentences in the original dataset to comply with our rules for the use of the schwa.[15] Moreover, when evaluating our models, we follow the same approach we adopted for MT-GenEval (Section 5.1) to obtain gender-balanced versions of the inputs. We only use the 841 sentences in the `test` split for the final evaluation of our models, but we add the 100 dev sentences to our validation data for fine-tuning and to our pool of examples for few-shot prompts.

Neo-GATE is accompanied by an evaluation protocol and dedicated metrics, which measure the ability of a model to appropriately produce the target neomorpheme compared to the gold standard. Specifically, COVerage refers to the number of spans that are found both in the model's output and in the dataset annotations for each sentence, i.e. spans that refer to human entities, regardless of the model producing a standard gendered form or a NB reformulation. ACCuracy only takes into account the correct NB forms generated by the model. Coverage-Weighted Accuracy combines the two aspects, thus reflecting the overall performance on the task; we consider this to be the main metric. Conversely, MIS-generation measures the ratio of NB forms used inappropriately, for example on words not referring to human beings, or generally in a different way compared to the reference.

Together with dataset-specific metrics, we also use standard MT metrics that capture the overlap between input and target sentence, both on the sentence level (BLEU by Papineni et al., 2002; and TER by Snover et al., 2006) and on the character level (chrF by Popović, 2015). We use the Hugging Face `evaluate`[16] implementation of sacreBLEU (Post, 2018) to compute them.

[8] https://huggingface.co/gsarti/it5-base

[9] https://huggingface.co/bigscience/mt0-base

[10] https://huggingface.co/bigscience/bloom

[11] https://huggingface.co/bigscience/bloomz

[12] At the time of writing, only a release blog post is available for Ministral: https://mistral.ai/news/ministraux.

[13] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct; https://huggingface.co/mistralai/Ministral-8B-Instruct-2410

[14] https://huggingface.co/datasets/FBK-MT/Neo-GATE

[15] The changes involve replacing occurrences of "studentə" (*student*$_{\mathrm{N}}$) in the raw (unadapted) Neo-GATE files, since it was considered as a masculine form in the original, while we treat it as an epicene noun, which does not need a schwa. To do so, we used the following regular expressions, which returned 74 matches in the test split, and 12 in the dev split:
Search: `([sS]tudent)<ENDS>` Replace: `$1e`
Search: `([sS]tudent)<ENDP>` Replace: `$1i`

[16] https://github.com/huggingface/evaluate

5.4 Experiments

Our experiments focus on comparing the performance of models with different architectures on the task defined above by a) sending them requests in the form of prompts, containing explicit instructions, a set of examples, or both; and b) further training them on our task.

5.4.1 Prompting experiments

Since prompting requires relatively low computational resources, we include larger models in these experiments, namely the `large` versions of IT5 and mT0, the 7.1 B version of BLOOMZ, Llama 3.1, and Ministral. We quantize all models —except IT5-base, IT5-large, and mT0-base— to fit our constraints.

For encoder-decoder models, our approach is largely based on Lee et al.'s (2024) method for in-context learning, based on providing these models with prompts containing task examples.

We first carry out a preliminary study in which we prompt all our models on a subset of 100 test sentences and experiment with the number of instances provided in the prompt, if any (0, 2, 4, 8, 16, or 32 per prompt); for quantized models, we also compare 4-bit and 8-bit quantization. For each model, we then select the configuration that guarantees the best results on this subset based on the CWA metric (or BLEU, in case of parity on CWA) to run a full test.

Our prompts are made up of one or more of three components: a task description or instructions, a set of examples including clear indicators to distinguish inputs and targets, and a final request, which follows the same structure as the examples, except that the target side is left open for completion. We use a slightly different template depending on the model or experimental configuration:

- zero-shot prompts only contain instructions and a request;

- following Lee et al.'s (2024) template, prompts for T5-based models always include examples, but no explicit instructions, and a sentinel token is added at the end of each request;

- for chat models we use chat templates which differentiate roles; most notably, examples are split into inputs, sent under the `user` role, and labels, sent under the `assistant` role.

When building prompts, we separate each component (instructions, task examples, request) from the following with a newline character.

As suggested by Lee et al. (2024), for seq2seq models we adopt early fusion. Each example is passed separately, together with the final request, to the encoder; then, the resulting set of encoder hidden states is concatenated and used for decoder cross-attention. The decoder is prompted to generate text by using as input the same token added at the end of the request passed to the encoder.

The templates we used for our prompts are presented in Tables 6 and 7 in Appendix B. For parity with each model's pretraining, we write prompts in Italian for IT5 and in English for all other models.

5.4.2 Fine-tuning experiments

Since updating model weights is more resource-intensive than prompting, we exclude the larger versions of BLOOMZ, IT5, and mT0, as well as Llama 3.1 and Ministral. Moreover, we use QLoRA for models with over 500 M parameters to reduce computational needs and to obtain satisfactory results with our small dataset (Dettmers et al., 2023).

For these experiments, we mainly follow Zhang et al.'s (2023) method for effectively fine-tuning larger decoder-only models. We mimic their setup for all of our models, except for the parameters listed in Appendix C. When fine-tuning T5-based models, in alignment with pretraining (Raffel et al., 2020), we prepend a task-specific prefix to each input sentence and we add a sentinel token both at the end of each input and at the beginning of each target sentence, as shown in Table 8.

After fine-tuning, we conduct a study on T5 models, comparing their performance when adding a task prefix, sentinel tokens, or both at inference time. This is meant to complement results reported by Lee et al. (2024) on the use of sentinel tokens when prompting these models.

6 Results and Discussion

As mentioned in Section 4, the closest work to our own is Piergentili et al. (2024). We thus use similar metrics to evaluate our models and will also refer to their results in this section; however, MT metrics are not comparable across the two studies, since they focus on translation rather than on rewriting.

Although we could not carry out a systematic qualitative evaluation of the models' outputs, we randomly extracted 10 sentences from the predictions of each prompted/fine-tuned model in the

Category	Overgeneration
Target	L’atleta colombianə ha deciso [. . .]
Output	L’atletə colombianə ha deciso [. . .]
Category	Partial rewriting
Target	Lə miə amicə tedescə è andatə [. . .]
Output	Lə miə amicə tedesca è andatə [. . .]

Table 2: Examples of common mistakes found in the models’ outputs.

final experiments, and we analyzed them to verify where the most common mistakes are concentrated. We find that, apart from hallucinations, our models struggle the most with long dependencies, for example rewriting some noun phrases only partially, and with overgeneration, for example extending the schwa on nouns that do not refer to human beings. Table 2 reports two typical examples of these issues. We also found that several of mT0’s predictions were (partially) in the wrong language.

6.1 Prompting

Table 3 shows the results of the final experiment (results of the preliminary evaluation are reported in Appendix D, Table 10). The numbers suggest that chat models such as Llama and Ministral are the only ones that can effectively perform our task.

MT metrics confirm that most models’ outputs are linguistically well-formed and adherent to the input sentences; COV is also an indirect indicator of translation quality, but it is consistently higher than BLEU since it only takes into account spans that involve gender-related phenomena. Other Neo-GATE metrics, however, clearly highlight the shortcomings of non-chat models with respect to the specific task at hand.

Preliminary results can help clarify this gap between chat and non-chat models: while Llama and Ministral perform better when examples are added, more examples often result in lower scores for the other models. This suggests that the conversational format of the prompts used with chat models is more adequate for including examples, which seem to mostly introduce noise in the other cases.

The best of the two chat models is Ministral, notably with +7 on CWA compared to Llama and substantially better results on all other metrics. A clear shortcoming of both these models is the high misgeneration rate, which is over 25 for both.

When it comes to the MIS metric, it is worth pointing out that a ratio close to 0 likely means that the model’s outputs contain virtually no schwa forms. That means that it is not fulfilling the task. The higher misgeneration rates for chat models are thus partly balanced by their higher accuracy, although our lowest MIS for models with non-near-zero accuracy is still higher than the best one obtained by Piergentili et al. (2024) (25.45 vs 10.17).

In general, model size does not emerge as a clear guarantee of better performance, nor does the number of task demonstrations. For example, in the final prompting experiment, the bigger BLOOMZ consistently performs worse than the smaller one, but the opposite is true for mT0 and IT5. However, the two best performing models overall are also the biggest ones and they achieve their best results when adding the most examples.

6.2 Fine-tuning

As shown in Table 4, IT5 guarantees the best results for almost all metrics (including CWA with almost 67) when fine-tuned, despite having less than half the parameters of other models in this experiment. However, it does suffer from a rather high misgeneration rate (over 16). mT0 apparently makes less mistakes (with MIS at around 8), although at the price of a much lower accuracy (slightly less than 22).

BLOOM and BLOOMZ are comparable on all metrics and do not reach the best performance in any, but they generally achieve better results than mT0, which has a similar number of parameters, and follow IT5 closely. This confirms that the approach can work well with both seq2seq and decoder-only models, and that model size is not the most important aspect.

Results of the ablation study on the best input format for inference with our fine-tuned T5-based models (Table 9) reveal that prepending a task prefix to the input and appending a sentinel token at the end, as done during fine-tuning, guarantees the best performance in most cases. This is thus the configuration that we selected for both models to compare them against the others in Table 4.

6.3 Comparison

Table 5 reports the best CWA score obtained by each model with zero- and few-shot prompts and when fine-tuned. When comparing models that were both prompted and fine-tuned (BLOOMZ, IT5, and mT0), all of them achieve better results on all metrics when fine-tuned, with the exception of the COV metric for BLOOMZ-560m. Moreover, the best overall figure for each metric is consis-

Model	Bits	Shots	BLEU	chrF	TER↓	COV	ACC	CWA	MIS↓
bloomz-560m	4	0	61.10	82.60	26.51	**91.33**	00.00	00.00	00.08
bloomz-7b1	8	0	46.44	68.29	45.74	69.83	00.00	00.00	00.16
it5-base	full	2	30.24	52.73	63.38	60.55	00.53	00.32	05.89
it5-large	full	2	46.66	67.66	46.08	75.80	00.53	00.40	01.29
Llama-3.1	8	16	51.56	79.92	37.70	71.60	32.96	23.60	28.32
Ministral	8	32	**67.18**	**87.77**	**19.54**	86.57	**35.60**	**30.82**	25.45
mt0-base	full	32	11.14	36.18	84.20	35.78	00.00	00.00	**00.00**
mt0-large*	4	8	30.29	57.05	59.20	63.70	00.00	00.00	**00.00**

Table 3: Results obtained by the prompted models on the full test set. Bold and underlined figures identify the best performance on that metric. *Due to memory constraints, we reduced the maximum length of both input and label in each example to 10 tokens for mT0-large in this experiment.

Model	BLEU	chrF	TER↓	COV	ACC	CWA	MIS↓
bloom-560m	77.68	92.01	12.51	85.44	55.67	47.56	17.39
bloomz-560m	76.62	91.53	13.05	85.20	55.40	47.20	17.79
it5-base	**85.39**	**94.31**	**07.75**	84.15	**79.58**	**66.96**	16.14
mt0-base	46.64	85.72	24.80	**91.49**	23.85	21.82	**07.91**

Table 4: Results obtained with fine-tuned models. For IT5 and mT0, we only report metrics for the best configuration based on the ablation study as shown in Table 9. Underlined and bold values identify the best result for that metric. For parity with the fine-tuning setup, in the inference stage we quantize all models to 4 bits except IT5.

Model	Zero-shot	Few-shot	Fine-tuning
IT5-base	N/A	00.32	66.96
IT5-large	N/A	00.40	N/A
mT0-base	N/A	00.00	21.82
mT0-large	N/A	00.00	N/A
BLOOM-560m	N/A	N/A	47.56
BLOOMZ-560m	00.00	00.00	47.20
BLOOMZ-7b1	00.00	00.00	N/A
Llama	08.81	23.60	N/A
Ministral	16.74	30.82	N/A

Table 5: Best CWA reached by each model in three settings. Each score is the best obtained by that model in that setting.

tently (much) better in the second case, despite using smaller models. This was partially expected, since fine-tuning acts on weights and thus influences model behavior at a deeper level, while also exposing the models to the full dataset; however, this also suggests that using very large models might not always turn out to be the best approach on an absolute level.

Fine-tuned IT5 achieves the best CWA (and most other metrics) overall, including the results of the prompting experiment, where the best model (Ministral) stops at around 30, or less than half. Looking at the results obtained by Piergentili et al. (2024) on the schwa paradigm with few-shot prompts, it improves on their best-performing model in terms of COV and CWA.

Overall, fine-tuning is the most effective approach when using smaller and less refined models. The increased cost of this approach compared to prompting is balanced by the smaller dimensions of the models and by using parameter-efficient techniques like QLoRA (Dettmers et al., 2023).

7 Conclusions and Future work

In this paper, we discussed the lack of recognition of non-binary identities in language and the implications of this on language technologies, with a focus on Italian and on the Western European and North American context. To address this problem from a technical point of view, we designed a rewriting task and evaluated models representing different architectures and NLP paradigms, while comparing the results obtained through prompting and fine-tuning methods.

As training data, we used previously released evaluation benchmarks where gender-ambiguous

English sentences are paired with gender-marked Italian translations; we manually added alternative translations using direct non-binary language according to our original guidelines, and we trained our models to rewrite each original translation into our reformulation.

We evaluated two seq2seq encoder-decoder models and three causal decoder-only models. We included different versions of these models with varying dimensions based on memory requirements for each experiment, and we conducted a preliminary evaluation and an ablation study to investigate the impact of a variety of parameters on performance.

We achieved promising results and suggest some possible directions for future developments. On one hand, fine-tuning benefits all models, and we demonstrate that it can guarantee better results even with smaller models. On the other hand, given the innovative nature of our task, prompting seems to only be effective when examples of such task are included in the prompt, and when the model is able to effectively learn from them and generalize. In our case, chat-tuned models were the only ones to yield satisfactory results in this setting.

An important aspect to consider is that we used a rather small dataset, and fine-tuning results would likely improve with more data. Despite the notoriously scarce availability of data in this domain, collecting more than this seems feasible, especially as prompting techniques to obtain annotated data from LLMs likely improve in the future. For example, future works could involve prompting strong models to obtain a basis from which to create more annotated data, and then fine-tuning cheaper models using the resulting, bigger dataset to obtain the final DNL sentences. An example of a similar approach is Raunak et al. (2024), who fine-tune an NMT model to follow instructions using translations generated by causal LLMs.

Moreover, annotations could be added to the input data, so as to explicitly identify specific spans holding gender information, both for prompting and for fine-tuning. Our results could also be improved by implementing more refined prompt design. For example, breaking the task into simpler, consecutive steps would likely prove beneficial to the rewriter: this could be achieved both with chain-of-thought prompts (Wei et al., 2022c) or thanks to the the improved "reasoning" capabilities of models such as DeepSeek-R1 (DeepSeek-AI, 2025).

Another direction that could be investigated in the future is the multilingual generalization of our approach, for example by fully leveraging mT0's multilingual capabilities through cross-lingual training. Finally, we plan to validate our approach by carrying out a more thorough manual investigation of the models' outputs and implementing human evaluation metrics.

Limitations

This work is limited in its way of dealing with its main subject, i.e. gender bias: superficial attempts aimed at adjusting existing models or adding more representative data are not sufficient to eliminate the negative effects and biases of language models on a general level. In order to be effective, research must foster a broader conversation about its sociocultural implications, and must therefore be interdisciplinary and community-based (Birhane, 2021; Gromann et al., 2023). This necessarily complex and collective effort was not carried out for this study. Nevertheless, we hope that this contribution can be useful in spreading and advancing the discussion about this and related issues.

Our experiments could be expanded. Specifically, for in-context learning with seq2seq models, we limited our experiments to the method proposed by Zhang et al. (2023), and thus did not test zero-shot prompting. Due to memory limitations, we also could not test some configurations in both the fine-tuning and the prompting experiments, and we could only fine-tune relatively small models. In addition, mT0 likely suffered from the maximum length of the examples being capped in the few-shot prompting setting.

As for our guidelines, since there is currently no shared standard for DNL in Italian, they contain some arbitrary choices and leave some questions open; moreover, we did not conduct any kind of survey, nor collect suggestions directly from the interested groups. As such, the guidelines are not meant to be prescriptive, nor representative of all the possible ways people who identify as non-binary can refer to themselves in Italian.

Our study focuses on only one strategy for non-binary language and considers only one language pair. Despite this, our approach could be easily extended to other neomorpheme-based strategies for Italian non-binary language (such as the asterisk *), as well as to other types of strategies (such as INL), although that would require additional work. With the due modifications, the approach could be adopted in similar languages or language pairs, but possibly with very different results. However, the resources we used might not be available or appropriate for different settings (e.g., with other languages or non-binary language strategies); in particular, our training dataset as it is can only be used for our Italian DNL paradigm.

Finally, the datasets we use, although consisting of natural examples, are not representative of the complexity of real-world data, as they are meant for controlled experiments. Specifically, each sentence in our training data contains only one set of gender marks, and our method might not extend to more complex texts.

Acknowledgements

We thank the anonymous reviewers for their valuable suggestions.

We also thank Prof. Rachele Raus for her advice during the early stages of this project, as well as Profs. Maja Miličević Petrović and Cristiana Cervini for reviewing our guidelines, Prof. Beatrice Spallaccia for providing bibliography pointers and for reviewing our guidelines, and Gabriele Aragno for helping with writing the target sentences we used in our experiments and included in our dataset.

References

Fabrizio Acanfora. 2022. Schwa: Una questione identitaria. In *Lingua, grammatica e società: Senza, con e oltre lo schwa*. Istituto della Enciclopedia Italiana. Available at `https://www.treccani.it/magazine/lingua_italiana/speciali/Schwa/1_Acanfora.html` [Last accessed 8 June 2025].

Florence Ashley. 2019. Les personnes non-binaires en français : une perspective concernée et militante. *H-France Salon*, 11(14).

Marion Bartl and Susan Leavy. 2024. From 'showgirls' to 'performers': Fine-tuning with gender-inclusive language for bias reduction in LLMs. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 280–294, Bangkok, Thailand. Association for Computational Linguistics.

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.

BigScience Workshop. 2023. BLOOM: A 176b-parameter open-access multilingual language model. *Preprint*, arXiv:2211.05100.

Abeba Birhane. 2021. Algorithmic injustice: a relational ethics approach. *Patterns*, 2(2):1–9.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4356–4364, Red Hook, NY, USA. Curran Associates Inc.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Gloria Comandini. 2021. Salve a tuttə, tutt*, tuttu, tuttx e tutt@: l'uso delle strategie di neutralizzazione di genere nella comunità queer online. Indagine su un corpus di italiano scritto informale sul web. *Testo e Senso*, 23:43–64.

Marta Costa-jussà, Pierre Andrews, Eric Smith, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Daniel Licht, and Carleigh Wood. 2023. Multilingual holistic bias: Extending descriptors and patterns to unveil demographic biases in languages at scale. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14141–14156, Singapore. Association for Computational Linguistics.

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.

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*, pages 4287–4299, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *Preprint*, arXiv:2501.12948.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Federica Formato and Anna Lisa Somma. 2023. Gender inclusive language in Italy: A sociolinguistic overview. *Journal of Mediterranean and European Linguistic Anthropology*, 5(1):22–40.

Simona Frenda, Andrea Piergentili, Beatrice Savoldi, Marco Madeddu, Martina Rosola, Silvia Casola, Chiara Ferrando, Viviana Patti, Matteo Negri, and Luisa Bentivogli. 2024. GFG - gender-fair generation: A CALAMITA challenge. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 1106–1115, Pisa, Italy. CEUR Workshop Proceedings.

Vera Gheno. 2020. Ministra, portiera, architetta: le ricadute sociali, politiche e culturali dei nomi professionali femminili (prima parte). Available at `https://www.linguisticamente.org/nomi-femminili/` [Last accessed 8 June 2025].

Vera Gheno. 2022a. Questione di privilegi: come il linguaggio ampio può contribuire ad ampliare gli orizzonti mentali. *About Gender*, 11(21).

Vera Gheno. 2022b. Schwa: storia, motivi e obiettivi di una proposta. In *Lingua, grammatica e società: Senza, con e oltre lo schwa*. Istituto della Enciclopedia Italiana. Available at `https://www.treccani.it/magazine/lingua_italiana/speciali/Schwa/4_Gheno.html` [Last accessed 8 June 2025].

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 60–63, Florence, Italy. Association for Computational Linguistics.

Dagmar Gromann, Manuel Lardelli, Katta Spiel, Sabrina Burtscher, Lukas Daniel Klausner, Arthur Mettinger, Igor Miladinovic, Sigrid Schefer-Wenzl, Daniela Duh, and Katharina Bühn. 2023. Participatory research as a path to community-informed, gender-fair machine translation. In *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 49–59, Tampere, Finland. European Association for Machine Translation.

Marie Gustafsson Sendén, Emma A. Bäck, and Anna Lindqvist. 2015. Introducing a gender-neutral pronoun in a natural gender language: the influence of time on attitudes and behavior. *Frontiers in Psychology*, 6.

Pascal Gygax, Ute Gabriel, Oriane Sarrasin, Jane Oakhill, and Alan Garnham. 2008. Generically intended, but specifically interpreted: When beauticians, musicians, and mechanics are all men. *Language and Cognitive Processes*, 23(3):464–485.

Nishtha Jain, Maja Popović, Declan Groves, and Eva Vanmassenhove. 2021. Generating gender augmented data for NLP. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 93–102, Online. Association for Computational Linguistics.

Emily Kendall. 2024. gender binary. In *Encyclopedia Britannica*. Available at `https://www.britannica.com/topic/gender-binary` [Last accessed 8 June 2025].

Kris Aric Knisely. 2020. *Le français non-binaire*: Linguistic forms used by non-binary speakers of French. *Foreign Language Annals*, 53(4):850–876.

Manuel Lardelli and Dagmar Gromann. 2023. Gender-fair post-editing: A case study beyond the binary. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 251–260, Tampere, Finland. European Association for Machine Translation.

Jihyeon Lee, Dain Kim, Doohae Jung, Boseop Kim, and Kyoung-Woon On. 2024. Exploiting the potential of seq2seq models as robust few-shot learners. In *First Conference on Language Modeling*.

Llama Team. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2019. Gender bias in neural natural language processing. *Preprint*, arXiv:1807.11714.

Geraldine Ludbrook. 2022. From gender-neutral to gender-inclusive English. The search for gender-fair language. *Deportate, esuli, profughe*, 48.

Ártemis López. 2019. Tú, yo, elle y el lenguaje no binario. *La linterna del traductor*, 19.

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)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Uwe Kjær Nissen. 2002. Aspects of translating gender. *Linguistik Online*, 11(2):25–37.

Ben Papadopoulos, Sol Cintrón, Clio Hartman, and Drew Rusignuolo. 2025. Italian. In *Gender in Language project*. Available at www.genderinlanguage.com/italian/ [Last accessed 8 June 2025].

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.

Andrea Piergentili, Dennis Fucci, Beatrice Savoldi, Luisa Bentivogli, and Matteo Negri. 2023. Gender neutralization for an inclusive machine translation: from theoretical foundations to open challenges. In *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 71–83, Tampere, Finland. European Association for Machine Translation.

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

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

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.

Michela Pusterla. 2019. Parlare femminista: la lingua di *Non una di meno*. In *Non esiste solo il maschile. Teorie e pratiche per un linguaggio non discriminatorio dal punto di vista del genere*. Edizioni Università di Trieste. Available at http://hdl.handle.net/10077/27152 [Last accessed 8 June 2025].

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. OpenAI blog. Available at https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf [Last accessed 8 June 2025].

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Vikas Raunak, Roman Grundkiewicz, and Marcin Junczys-Dowmunt. 2024. On instruction-finetuning neural machine translation models. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1155–1166, Miami, Florida, USA. Association for Computational Linguistics.

Martina Rosola, Simona Frenda, Alessandra Teresa Cignarella, Matteo Pellegrini, Andra Marra, and Mara Floris. 2023. Beyond obscuration and visibility: Thoughts on the different strategies of gender-fair language in Italian. In *Proceedings of the 9th Italian Conference on Computational Linguistics*.

Alma Sabatini. 1987. Raccomandazioni per un uso non sessista della lingua italiana. Availbe through the Internet archive at https://web.archive.org/web/20241206151026/https://www.funzionepubblica.gov.it/sites/funzionepubblica.gov.it/files/documenti/Normativa%20e%20Documentazione/Dossier%20Pari%20opportunit%C3%A0/linguaggio_non_sessista.pdf [Last accessed 12 June 2025].

Eduardo Sánchez, Pierre Andrews, Pontus Stenetorp, Mikel Artetxe, and Marta R. Costa-jussà. 2024. Gender-specific machine translation with large language models. In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pages 148–158, Miami, Florida, USA. Association for Computational Linguistics.

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

Gabriele Sarti and Malvina Nissim. 2024. IT5: Text-to-text pretraining for Italian language understanding and generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9422–9433, Torino, Italia. ELRA and ICCL.

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, Eleonora Cupin, Manjinder Thind, Anne Lauscher, Andrea Piergentili, Matteo Negri, and Luisa Bentivogli. 2025. mGeNTE: A multilingual resource for gender-neutral language and translation. *Preprint*, arXiv:2501.09409.

Beatrice Savoldi, Sara Papi, Matteo Negri, Ana Guerberof-Arenas, and Luisa Bentivogli. 2024a. What the harm? quantifying the tangible impact of gender bias in machine translation with a human-centered study. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18048–18076, Miami, Florida, USA. Association for Computational Linguistics.

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

Giuseppina Scotto Di Carlo. 2020. An analysis of the use of inclusive language among Italian non-binary individuals: A survey transcending binary thinking. *I-LanD Journal*, 2:69–89.

Sabine Sczesny, Magda Formanowicz, and Franziska Moser. 2016. Can gender-fair language reduce gender stereotyping and discrimination? *Frontiers in Psychology*, 7.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

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.

Gigliola Sulis and Vera Gheno. 2022. The debate on language and gender in Italy, from the visibility of women to inclusive language (1980s–2020s). *The Italianist*, 42(1):153–183.

Tony Sun, Kellie Webster, Apu Shah, William Yang Wang, and Melvin Johnson. 2021. They, them, theirs: Rewriting with gender-neutral English. *Preprint*, arXiv:2102.06788.

Stefano Telve. 2011. Accordo [Prontuario]. In *Enciclopedia dell'Italiano*. Available at `https://www.treccani.it/enciclopedia/accordo-prontuario_(Enciclopedia-dell'Italiano)/` [Last accessed 8 June 2025].

Eva Vanmassenhove. 2024. Gender bias in machine translation and the era of large language models. *Preprint*, arXiv:2401.10016.

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

Thomas Wang, Adam Roberts, Daniel Hesslow, Teven Le Scao, Hyung Won Chung, Iz Beltagy, Julien Launay, and Colin Raffel. 2022. What language model architecture and pretraining objective work best for zero-shot generalization? *Preprint*, arXiv:2204.05832.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022a. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022b. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

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

Hualin Xiao, Brent Strickland, and Sharon Peperkamp. 2023. How fair is gender-fair language? Insights from gender ratio estimations in French. *Journal of Language and Social Psychology*, 42(1):82–106.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Xuan Zhang, Navid Rajabi, Kevin Duh, and Philipp Koehn. 2023. 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.

A Guidelines for Italian DNL

Our approach for Italian direct non-binary language (DNL) is based on the schwa neomorpheme paradigm.

A.1 Articles and articles combined with prepositions

Definite articles are *lə* for the singular and *ə* for the plural; if the noun begins with a vowel, the schwa in the singular article is elided (*l'*), regardless of the grammatical gender of the word.

The singular indefinite article is *unə*; even if the noun begins with a vowel, the article is never elided or truncated. The plural form corresponds to the partitive *deə* (composed of preposition *di* + article *ə*).

Some contracted forms are created by combining the base prepositions *de-*, *a-*, *da-*, and *su-* with definite articles: *del/dello / della > dellə*, *dei/degli / delle > deə*; *sul/sullo / sulla > sullə*, *sui/sugli / sulle > suə*. In some cases, the ending can be elided and the resulting forms do not express any binary gender: *dell'*; *sull'*.

A.2 Pronouns

The third-person singular personal pronoun is *ləi* when it is a subject, *lə* when it is a direct or indirect object. In the plural, the direct object is *lə*, the indirect one is *loro*; the latter is already gender-ambiguous in its standard form, but any words that agree with it might be gendered.

A drawback of this solutions is that the distinction between singular and plural is lost for the direct object pronoun. In example (1), the rewritten translation introduces some ambiguity with respect to the number of the underlined referent, for lack of context:

(1) mGeNTE en-it - ep-en-it-2277
"[. . .] we are too dubious [. . .] not to refrain from putting them on their guard."
"[. . .] nutriamo troppi dubbi [. . .] per astenerci dal metterli in guardia." > "[. . .] nutriamo troppi dubbi [. . .] per astenerci dal metterlə in guardia."

It is also worth noting that the pronoun *ləi* is the only case where the schwa is in a stressed and intrasyllabic position, which could make its pronunciation more difficult, as Gheno (2022a) points out.

The formal third-person singular pronoun *lei* can refer independently to any gender (unless some other words are gender-marked). The pronouns *egli/ella*, *essi/esse* (when referring to people) become *ellə* in the singular and *essə* in the plural; they can also be replaced with the more informal *ləi* (singular) or *loro* (plural), or omitted in almost all contexts.

A.3 Nouns

We treat nouns differently according to their gender morphology, using the categories defined by Gheno (2020). In all cases, we do not make any explicit distinction between the singular and the plural in the noun itself; the distinction is given by other words that agree with the noun, usually articles.

A.3.1 Mobile gender nouns

For nouns in this class, gender is distinguished on the morphological level, through inflection of the endings. We identify the following two subcategories based on such endings:

1. Nouns ending in -o/-a
   For example: *il maestro / la maestra > lə maestrə, i maestri / le maestre > ə maestrə.*
   Special cases:
   - Nouns ending in -co/-ca, -ci/-che and -go/-ga, -gi/-ghe: these consonants have a hard sound in front of a schwa, both in the singular and in the plural, without the need to add an -h- in writing: *l'amico / l'amica > l'amicə, gli amici / le amiche > ə amicə* (IPA [amikə]); *lo psicologo / la psicologa > lə psicologə, gli psicologi / le psicologhe > ə psicologə* (IPA [psikologə]);
   - Nouns in -cio/-cia, -ci/-cie and -gio/-gia, -gi/-ge: these consonants are made soft in front of a schwa by keeping the -i- in writing in both singular and plural: *il saggio / la saggia > lə saggiə, i saggi / le sagge > ə saggiə* (IPA [saʤ:ə];
   - Nouns ending in -io/-ia, -i(i)/-ie: the gender-neutral form with schwa always ends in -iə: *il segretario / la segretaria > lə segretariə, i segretari / le segretarie > ə segretariə.*

2. Nouns ending in -e/-a
   For example: *il pompiere / la pompiera > lə pompierə.*
   Special cases:

- Nouns in -tore/-trice/-tora or -sore/-ditrice/-sora: for the gender-neutral form with schwa, the -torə, -sorə ending is preferred: *l'autore / l'autrice > l'autorə*, *gli autori / le autrici > ə autorə*; *il difensore / la difenditrice/difensora > lə difensorə*, *i difensori / le difenditrici/difensore > ə difensorə*; *l'assessore / l'assessora > l'assessorə*, *gli assessori / le assessore > ə assessorə*.
- Feminine forms in -essa: The use of these forms has been discouraged by Italian linguists since the foundational work by Sabatini (1987).
  - Nouns based on present participles in -ente/-enti or -ante/-anti are epicene (see below), so they are valid for any gender. For example: *il presidente / la presidentessa/presidente > lə presidente*, *i/le/ə presidenti*; *lo studente / la studentessa/studente > lə studente*, *i/le/ə studenti*.
  - If the feminine form in -essa corresponds to a masculine form in -sore, the same logic used for the feminine forms in -trice/-tora, -ditrice/-sora applies: *il professore / la professoressa > lə professorə*, *i professori / le professoresse > ə professorə*.

A.3.2 Epicene nouns

Epicene nouns are mostly based on present participles and have the same form for any gender, both in the singular and in the plural. For example: *il/la/lə parlante*, *i/le/ə parlanti* from the present participle of *parlare* ("to speak"). Other nouns behave in the same way although they are not based on present participles. For example: *il/la/lə giudice*, *i/le/ə giudici*.

Some of these nouns are epicene in the singular, but not in the plural (mostly nouns ending in -eta, -ista, -iatra). For example: singular *l'atleta*, but plural *gli atleti / le atlete > ə atletə*; singular *il/la/lə dentista*, but plural *i dentisti / le dentiste > ə dentistə*. Special cases:

- Nouns ending in -ga, -ghi/-ghe: the consonant has a hard sound in front of a schwa, both in the singular and in the plural, without the need to add an -h- in writing: singular *il/la/lə collega*, but plural *i colleghi / le colleghe > ə collegə* (IPA [kol:egə]).

A.3.3 Invariable gender nouns

A restricted group of nouns have a fixed grammatical gender, unrelated to the referent's gender. Some examples are: *la persona*$_{\text{[F]}}$, *il membro*$_{\text{[M]}}$, *la guida*$_{\text{[F]}}$, *la spia*$_{\text{[F]}}$, *l'individuo*$_{\text{[M]}}$.

A.3.4 Lexical gender nouns

As opposed to the other categories, for these nouns, gender is determined at the lexical level. Most of them identify family relationships, as far as human referents are concerned (e.g., *madre-padre* ("father-mother"), *fratello-sorella* ("brother-sister"), etc.). Given their morphology, the grammatical and referential gender of these nouns is tied to their semantic root; they would thus need to be replaced by different words altogether to avoid expressing any binary gender. We did not find any shared non-binary solution for nouns in this class (see also Rosola et al., 2023).

A.4 Adjectives and participles

Adjectives generally follow the same rules as nouns with a corresponding morphology. Some specific cases are discussed in the following paragraphs.

Demonstrative adjectives and pronouns follow mobile gender nouns ending in -o/-a: they are *questə* and *quellə* both in the singular and in the plural; the ending can be elided in the singular form if the following noun starts with a vowel (*quest'*, *quell'*).

The distinction between singular and plural is usually given by other elements of the sentence. In example (2), the underlined expression in Italian (corresponding to English *either*) contains a singular determiner (*uno*) and a plural pronoun (*questi*). In the rewritten translation, they have the same ending, but context makes the meaning unequivocable:

(2) MT-GenEval - context_en_it - test - 73
"If relations break down with either, the Assistant[...]'s usefulness is [...] impaired."
"Se le relazioni si guastano con uno di questi, l'utilità dell'assistente [...] è [...] compromessa". > "Se le relazioni si guastano con unə di questə, l'utilità dell'assistente [...] è [...] compromessa."

However, differently from nouns — which are usually accompanied by an article — in some cases the context might not be enough to distinguish between the singular and plural forms of adjectives.

In example (3), the rewritten translation is ambiguous with respect to the number of the underlined referent:

(3) mGeNTE en-it - ep-en-it-14384
“You feel like telling those old leaders to open the door and success will flood in.”
“Si è quasi tentati di invitare questi anziani leader ad aprire la porta e a lasciare entrare il successo.” > “Si è quasi tentatə di invitare questə anzianə leader ad aprire la porta e a lasciare entrare il successo.”

The same goes for **possessive adjectives and pronouns**: *(il) mio / (la) mia > (lə) miə*, *(i) miei / (le) mie > (ə) miə*; *(il) nostro / (la) nostra > (lə) nostrə*, *(i) nostri / (le) nostre > (ə) nostrə*. The third person plural possessive *loro* applies to possessors of any gender, but the grammatical gender of the possessed (which could be a person) can still be expressed through determiners, e.g.: *il/i / la/le / lə/ə loro*. Since possessives are usually accompanied by articles, the number distinction is less of a problem for this class.

Participles follow either epicene or mobile gender nouns. Many present participles are actually used as epicene nouns (e.g., *presidente*), while past participles can be conjugated as mobile gender nouns ending in -o/-a. In contemporary Italian, past participles systematically agree with the subject only if the verb is intransitive and has *essere* as its auxiliary, or with the object, if it is a third-person personal pronoun (Telve, 2011). For example (4):

(4) mGeNTE en-it - ep-en-it-5307
“No one has been able to explain to me yet [...]”
“Finora nessuno è riuscito a spiegarmi [...]” > “Finora nessunə è riuscitə a spiegarmi [...]”

B Templates

Tables 6 and 7 show the templates we used to prompt standard and chat-tuned models, respectively. For chat models, we use the `assistant` role to provide example completions, i.e. labels.

Table 8 shows the template for input-label pairs used when fine-tuning T5-based models. For Italian prompts, we use “Frase originale” and “Riformulazione” to introduce example inputs and labels, respectively.

C Fine-Tuning Settings

For the fine-tuning experiments we follow Zhang et al.’s (2023) settings. We only set the following parameters differently: batch size: 2, training steps for evaluation and checkpointing: 200, and patience for early stopping: 2 checkpoints.

D Additional experiments

Table 9 reports on the ablation study on the use of task prefix and sentinel tokens for T5-based models, while Table 10 contains the full results of the preliminary prompting experiment.

Component	**Example**
`instructions`	Rewrite the following Italian sentence by replacing masculine and feminine endings with a schwa (ə) for human entities.
`example set`	Original sentence: `<example input>` Rewritten sentence: `<example label></s>`
`request`	Original sentence: `<example input>` Rewritten sentence:

Table 6: Generic template for zero- or few-shot prompting. If any, examples are repeated k times, with a newline between each of them.

Role	**Template**
`user`	Rewrite the following Italian sentence by replacing masculine and feminine endings with a schwa (ə) for human entities based on the examples provided. Original sentence: `<Example input.>` Rewritten sentence:
`assistant`	`<Example label.>`
`user`	Original sentence: `<Input.>` Rewritten sentence:

Table 7: Template for few-shot prompts used with chat models.

Language	**Input template**	**Label template**
Italian	Riformula: `<Input sentence.><sentinel>`	`<sentinel><Target sentence.>`
English	Rewrite: `<Input sentence.><sentinel>`	`<sentinel><Target sentence.>`

Table 8: Template for inputs and labels, with task prefix and sentinel tokens, used for fine-tuning T5 models.

Model	**Prefix**	**Sentinel**	**BLEU**	**chrF**	**TER**$^{\downarrow}$	**COV**	**ACC**	**CWA**	**MIS**$^{\downarrow}$
`it5-base`	No	No	51.49	71.11	38.29	62.12	13.96	08.67	**02.10**
`it5-base`	Yes	No	82.50	92.82	10.01	82.33	73.84	60.79	16.54
`it5-base`	No	Yes	64.45	82.70	24.42	77.61	18.09	14.04	02.90
`it5-base`	Yes	Yes	**85.39**	**94.31**	**07.75**	**84.15**	**79.58**	**66.96**	16.14
`mt0-base`	No	No	45.39	84.62	26.95	**92.94**	17.80	16.54	**05.08**
`mt0-base`	Yes	No	46.44	85.58	25.17	92.17	21.97	20.25	07.18
`mt0-base`	No	Yes	46.23	85.43	25.28	92.34	21.71	20.05	06.86
`mt0-base`	Yes	Yes	**46.64**	**85.72**	**24.80**	91.49	**23.85**	**21.82**	07.91

Table 9: Ablation study on the impact of adding a task prefix and a sentinel token at inference time with T5 models. For parity with the fine-tuning setup, mT0 is quantized, while IT5 uses full-precision inference.

Model	Bits	Shots	BLEU	chrF	TER$^{\downarrow}$	COV	ACC	CWA	MIS$^{\downarrow}$
bloomz-560m	4	0	**66.41**	85.98	**21.28**	**93.83**	00.00	00.00	00.44
bloomz-560m	4	2	50.21	79.15	52.40	84.58	00.00	00.00	00.44
bloomz-560m	4	4	47.94	84.28	59.95	90.75	00.00	00.00	00.88
bloomz-560m	4	8	63.55	82.48	26.54	88.99	00.00	00.00	00.88
bloomz-560m	4	16	43.19	79.22	75.13	90.75	00.00	00.00	**00.00**
bloomz-560m	4	32	38.14	78.55	90.62	87.67	00.00	00.00	01.76
bloomz-560m	8	0	66.23	**86.18**	21.59	92.51	00.00	00.00	**00.00**
bloomz-560m	8	2	47.68	80.03	56.75	88.55	00.00	00.00	01.32
bloomz-560m	8	4	65.01	84.18	23.80	90.31	00.00	00.00	00.44
bloomz-560m	8	8	63.42	82.92	25.48	88.55	00.00	00.00	01.76
bloomz-560m	8	16	35.75	74.49	104.65	85.90	00.00	00.00	01.76
bloomz-560m	8	32	64.82	84.88	22.88	91.19	00.00	00.00	**00.00**
bloomz-7b1	4	0	52.07	71.33	41.19	67.84	00.00	00.00	00.00
bloomz-7b1	4	2	31.03	53.64	70.40	46.70	00.00	00.00	00.00
bloomz-7b1	4	4	39.65	58.03	58.96	54.19	00.00	00.00	00.00
bloomz-7b1	4	8	47.76	66.58	48.67	60.79	00.00	00.00	00.00
bloomz-7b1	4	16	43.45	64.32	52.86	63.00	00.00	00.00	00.00
bloomz-7b1	8	0	**54.12**	**73.10**	**38.52**	**72.25**	00.00	00.00	00.00
bloomz-7b1	8	2	38.11	56.87	60.64	48.46	00.00	00.00	00.00
bloomz-7b1	8	4	41.80	60.31	55.99	52.42	00.00	00.00	00.00
bloomz-7b1	8	8	40.18	59.84	58.28	53.30	00.00	00.00	00.00
bloomz-7b1	8	16	43.43	62.42	53.62	57.71	00.00	00.00	00.00
it5-base	full	2	34.85	56.21	60.64	70.48	**01.25**	**00.88**	09.25
it5-base	full	4	**38.08**	**59.41**	**56.60**	73.13	00.00	00.00	00.88
it5-base	full	8	36.46	59.01	58.05	**74.89**	00.00	00.00	**00.00**
it5-base	full	16	31.97	54.20	63.16	69.16	00.00	00.00	01.32
it5-base	full	32	35.15	57.04	60.11	72.25	00.00	00.00	**00.00**
it5-large	full	2	**50.08**	69.39	43.40	80.62	00.00	00.00	04.85
it5-large	full	4	50.07	69.45	**43.33**	**84.14**	00.00	00.00	**00.00**
it5-large	full	8	49.78	**69.47**	43.40	**84.14**	00.00	00.00	**00.00**
it5-large	full	16	46.01	66.26	48.36	79.74	00.00	00.00	**00.00**
it5-large	full	32	48.87	68.52	45.46	80.18	00.00	00.00	**00.00**
Llama-3.1-8B-Instruct	4	0	58.39	85.29	27.31	**83.70**	10.53	08.81	49.78
Llama-3.1-8B-Instruct	4	2	44.19	76.24	49.35	75.77	25.58	19.38	25.55
Llama-3.1-8B-Instruct	4	4	58.42	81.79	28.91	80.18	34.07	27.31	28.63
Llama-3.1-8B-Instruct	4	8	59.16	84.61	27.54	71.81	36.81	26.43	33.04
Llama-3.1-8B-Instruct	4	16	59.47	84.11	25.86	70.04	37.74	26.43	31.28
Llama-3.1-8B-Instruct	4	32	58.03	83.03	29.60	66.96	31.58	21.15	27.31
Llama-3.1-8B-Instruct	8	0	60.14	86.38	25.86	81.06	10.33	08.37	57.27
Llama-3.1-8B-Instruct	8	2	61.69	**86.84**	24.79	74.45	30.77	22.91	29.07
Llama-3.1-8B-Instruct	8	4	62.76	86.10	25.17	77.09	36.57	28.19	**22.47**
Llama-3.1-8B-Instruct	8	8	61.75	85.84	25.40	73.13	**43.37**	**31.72**	31.28
Llama-3.1-8B-Instruct	8	16	**63.96**	86.78	**21.82**	74.89	42.35	**31.72**	31.28
Llama-3.1-8B-Instruct	8	32	59.71	83.12	29.44	73.57	40.72	29.96	29.96
Ministral-8B-Instruct	4	0	59.11	80.68	27.46	66.96	22.37	14.98	59.47
Ministral-8B-Instruct	4	2	71.92	89.82	16.55	83.26	28.57	23.79	21.59
Ministral-8B-Instruct	4	4	72.50	90.40	15.26	90.75	33.98	30.84	18.50
Ministral-8B-Instruct	4	8	74.65	90.79	15.03	87.22	40.40	35.24	22.03
Ministral-8B-Instruct	4	16	**75.63**	**91.65**	**13.65**	**91.63**	42.79	39.21	**18.06**
Ministral-8B-Instruct	4	32	73.70	91.27	14.80	88.99	47.03	41.85	22.91
Ministral-8B-Instruct	8	0	54.04	77.53	31.81	59.47	28.15	16.74	94.71
Ministral-8B-Instruct	8	2	69.12	88.76	17.77	83.70	35.26	29.52	33.04
Ministral-8B-Instruct	8	4	71.20	89.58	16.70	85.02	34.20	29.07	24.23
Ministral-8B-Instruct	8	8	70.93	89.61	16.70	83.70	46.32	38.77	31.72
Ministral-8B-Instruct	8	16	70.96	87.57	16.70	85.02	48.19	40.97	29.52
Ministral-8B-Instruct	8	32	72.60	90.54	15.26	86.78	**53.30**	**46.26**	30.84
mt0-base	full	2	07.88	30.88	92.68	26.43	00.00	00.00	01.32
mt0-base	full	4	10.87	34.37	86.96	28.19	00.00	00.00	**00.00**
mt0-base	full	8	10.59	34.80	84.44	24.67	00.00	00.00	**00.00**
mt0-base	full	16	11.65	36.21	**82.53**	**31.28**	00.00	00.00	**00.00**
mt0-base	full	32	**12.14**	**38.02**	82.61	29.07	00.00	00.00	**00.00**
mt0-large	4	2	22.58	46.85	73.53	42.29	00.00	00.00	01.32
mt0-large	4	4	21.24	45.56	81.46	38.33	00.00	00.00	00.44
mt0-large	4	8	**27.53**	**54.08**	**66.44**	**50.66**	00.00	00.00	**00.00**
mt0-large	8	2	06.15	25.98	112.97	08.81	00.00	00.00	01.32
mt0-large	8	4	21.19	45.00	82.46	37.00	00.00	00.00	00.44

Table 10: Preliminary prompting experiment. Missing combinations are due to memory constraints. Bold figures identify the best result for a model on that metric.