Gender-Neutral Rewriting in Italian: Models, Approaches, and Trade-offs

Andrea Piergentili[1,2,*], Beatrice Savoldi[2], Matteo Negri[2] and Luisa Bentivogli[2]

[1] *University of Trento, via Sommarive 5, 38123, Povo (TN), Italy*
[2] *Fondazione Bruno Kessler, Via Sommarive 18, 38123, Povo (TN), Italy*

Abstract

Gender-neutral rewriting (GNR) aims to reformulate text to eliminate unnecessary gender specifications while preserving meaning, a particularly challenging task in grammatical-gender languages like Italian. In this work, we conduct the first systematic evaluation of state-of-the-art large language models (LLMs) for Italian GNR, introducing a two-dimensional framework that measures both neutrality and semantic fidelity to the input. We compare few-shot prompting across multiple LLMs, fine-tune selected models, and apply targeted cleaning to boost task relevance. Our findings show that open-weight LLMs outperform the only existing model dedicated to GNR in Italian, whereas our fine-tuned models match or exceed the best open-weight LLM's performance at a fraction of its size. Finally, we discuss the trade-off between optimizing the training data for neutrality and meaning preservation.

Keywords

Ethics, fairness, gender rewriting, large language models, fine-tuning

1. Introduction

Language technologies reinforce existing gender stereotypes and binary assumptions by disproportionately favoring masculine references or representations [1], especially when gender information is ambiguous or unspecified [2, 3, 4]. Such biases result in the underrepresentation or misrepresentation of certain gender groups, reinforcing existing societal stereotypes, and erasing non-binary identities [5, 6]. Addressing these biases through gender-inclusive approaches is increasingly important to ensure language technologies contribute to more inclusive and equitable communication [7, 8, 9].

Gender-neutral rewriting (GNR) has emerged as a natural language generation task aimed at producing texts free from unnecessary gender specifications [10, 11]. This task is particularly challenging in grammatical-gender languages, such as Italian, due to the pervasive encoding of gender in the morphology. Consider the sentence *'Tutti i senatori sono stati informati'* (equivalent to *All$_M$ the$_M$ senators$_M$ have been$_M$ informed$_M$*): almost every word is morphologically inflected for (masculine) gender. Rephrasing this sentence in a gender-neutral way may require significant changes, e.g. '*Ogni membro del Senato ha ricevuto l'informazione*' (*Every member of the Senate has received the information*). A further challenge in automatic GNR is preserving the meaning of the original sentence beyond gender expression, to avoid generating output sentences that are neutral but semantically divergent from the input.

So far, GNR system development has been mostly confined to English [10, 11, 12, inter alia], where gender is expressed through specific sets of words, such as pronouns (e.g., *he/she*, *him/her*) and lexically gendered terms (e.g., *policeman/policewoman*), and gender-neutral alternatives (e.g., the singular *they* or synonyms like *police officer*) are generally available and attested. GNR systems for grammatical-gender languages generally target specific gendered phenomena, such as member nouns [13], or use neologistic [14] inclusive devices such as neomorphemes and graphemic solutions [15, 16, 17] that convey neutrality, but are not necessarily acceptable in all contexts. Currently, the sole model dedicated to Italian GNR was developed by Greco et al. [18], which, however, was developed and tested on proprietary, not publicly available data, hindering reproducibility and progress.

Towards addressing this gap, this paper explores the potential of state-of-the-art (SOTA) large language models (LLMs) to perform GNR in Italian. Specifically, we explore both prompting and fine-tuning approaches and assess both neutrality and meaning preservation in the reformulated texts.

Our contributions are threefold: ***i)*** The first systematic evaluation of SOTA LLMs for Italian GNR under a two-dimensional framework measuring both neutrality and meaning preservation; ***ii)*** A set of experiments in fine-tuning LLMs for GNR, enabling compact models to rival significantly larger-sized models; ***iii)*** An investigation of the GNR performance trade-off between meaning preser-

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy

[*] Corresponding author.

apiergentili@fbk.eu (A. Piergentili); bsavoldi@fbk.eu (B. Savoldi); negri@fbk.eu (M. Negri); bentivo@fbk.eu (L. Bentivogli)

0000-0003-2117-1338 (A. Piergentili); 0000-0002-3061-8317 (B. Savoldi); 0000-0002-8811-4330 (M. Negri); 0000-0001-7480-2231 (L. Bentivogli)

vation and neutrality in the outputs of LLMs fine-tuned on sentence similarity-optimized data.[1]

2. Background

Gender-Inclusive Language Inclusive language aims to prevent expressions that reinforce gender hierarchies or render non-binary identities invisible, promoting fairness and inclusion in alignment with UN Sustainable Development Goals of gender equality.[2] In grammatical-gender languages like Italian, inclusive language is both particularly challenging and increasingly urgent due to their entrenched gender systems [19, 20, 21] and the widespread use of masculine forms as default to mark generic or mixed-gender referents [22].[3] To address this issue, two main strategies have emerged, as reviewed by Rosola et al. [24] within the Italian linguistic context. On the one hand, *innovative* forms using neomorphemes and symbols (e.g., tutt* or tutt@) are mostly used in informal contexts like social media and online LGBTQIA+ communities, and are generally not accepted in more formal contexts [25]. Instead, *conservative* gender-neutral language strategies retool existing forms and grammar to avoid unnecessary gendered expressions [26, 27], e.g. by replacing *i professori* with *la docenza* [9]. As attested by Piergentili et al. [28], such neutral solutions are increasingly accepted in communication and are endorsed by institutions and universities to embrace all gender identities [29].[4]

Gender-Inclusive Rewriting In recent years, sexism and gender-exclusionary practices have been increasingly addressed in NLP, focusing initially on binary gender bias and more recently expanding to non-binary inclusive language technologies [6, 4]. NLP work has explored the modeling of inclusive language across various tasks [30, 31], including inclusive language generation. For instance, Bartl and Leavy [12] explored stereotype reduction in English LLMs fine-tuned on inclusive seeds and lexicon.

Intralingual inclusive rewriting has primarily been explored in English [10, 11, 12], where gender marking is scarce. Similar efforts in languages with grammatical gender include research on German [15], Portuguese [16], and French [17, 13], either by using *innovative* forms or targeting specific instances of gendered languages—such as masculine generics in member nouns. In Italian, prior work has explored gender-neutral translation [32, 33], whereas intra-lingual rewriting remains mostly limited to benchmarking efforts. [34]. Attanasio et al. [35] compared several instruction-following models prompted across fairness-related tasks—including GNR—but these underperformed, achieving less than 50% success in neutralization. Frenda et al. [34] proposed the gender-fair generation (GFG) challenge, where for one of the tasks models are prompted to reformulate gendered Italian sentences in a neutral way. Closest to our work, Greco et al. [18] developed a rewriter by fine-tuning language models specifically for Italian gender-neutral language. However, the data used for testing and developing these models are not publicly available, hampering further research and comparability.

REF-G	Spero di essere stato chiaro su questo punto.
EN	I hope that I am clear in this.
REF-N	Spero di *avere espresso con chiarezza* questo punto.
EN	I hope that I have expressed this point clearly.

Table 1
Example of an Italian mGeNTE entry. The gendered words in the REF-G are underlined, the corresponding neutralization In REF-N is italicized.

3. Experimental settings

We define GNR as the task of reformulating a sentence to remove explicit gender markings referring to human entities, without altering the sentence beyond what is necessary for neutralization, ensuring semantic equivalence to the input. We run a set of experiments evaluating different systems and approaches to GNR. Here, we first discuss the evaluation data and metrics (§3.1) and the set of models we experiment with (§3.2). Then, we describe two approaches to GNR: few-shot prompting SOTA LLMs (§3.3) and fine-tuning a subset of those LLMs on repurposed Italian data (§3.4).

3.1. Evaluation

Test data Following Frenda et al. [34], we conduct our GNR experiments on mGeNTE [33], a benchmark for gender-neutral translation from English into several grammatical-gender languages, including Italian. mGeNTE provides 1,500 parallel gendered and gender-neutral references created by professionals (REF-G and REF-N respectively), differing only in gender expression (see Table 1 for an example of an Italian mGeNTE entry). It is organized into two subsets: Set-G, containing sentences that require neutralization, and Set-N, containing sentences that do not. For our GNR experiments, we use the 750 Italian gendered references from Set-N as

[1] We release models and data at https://huggingface.co/FBK-MT
[2] See https://sdgs.un.org/goals/goal5
[3] English presents fewer challenges as gender marking is primarily limited to pronouns, allowing focused solutions like the singular *they* [23].
[4] See for instance the EU Parliament guidelines for gender-neutral language: https://www.europarl.europa.eu/cmsdata/151780/GNL_Guidelines_EN.pdf

Group	Model	Size (B)	Prompting	Fine-tuning	Paper / Report	Link
"Italian" models	Minerva	7	✔	✘	Orlando et al. [36]	🤗
	LLaMAntino	8	✔	✔	Basile et al. [37]	🤗
	Velvet	14	✔	✔	Almawave [38]	🤗
Multilingual LLMs	Llama 3.1	8	✔	✔	Llama Team [39]	🤗
	Phi 4	14	✔	✔	Abdin et al. [40]	🤗
	Llama 3.3	70	✔	✘	Llama Team [39]	🤗
Qwen3 family	Qwen3	4	✔	✘	Qwen Team [41]	🤗
	Qwen3	8	✔	✔		🤗
	Qwen3	14	✔	✔		🤗
	Qwen3	32	✔	✘		🤗
Commercial system	GPT 4.1	?	✔	✘	OpenAI [42]	-
Dedicated model	Inclusively	0.78	✳	✘	Greco et al. [18]	🤗

Table 2
Summary of the models used in this work, including their size, usage in prompting and fine-tuning experiments, and documentation. Inclusively (✳) is a sequence-to-sequence model and was thus not compatible with few-shot prompting. We evaluated it by inputting gendered sentences directly and used it as a baseline in all generation experiments.

input. Such sentences are ideal input for our task, as they include unnecessary gender specifications by design.

Metrics To evaluate gender-neutrality, we use the LLM-as-a-Judge [43] approach proposed by Piergentili et al. [44], which provides sentence-level binary gendered/neutral assessments, and was shown to be highly accurate in both human- and model-generated texts. We use their optimal configuration for monolingual evaluation.[5] We compute the percentage of neutralized sentences over the whole test set (750 entries).

To evaluate meaning preservation in GNR, we use BERTScore [45], an attested BERT-based [46] metric measuring the semantic similarity of two texts (the higher the better, indicating close similarity). We use BERTScore rather than common string-matching metrics like BLEU [47] and TER [48] because gender-neutralization can have a notable impact on the lexicon, morphology, and structure of a sentence [9], which would be penalized by such metrics. By contrast, BERTScore was found to be rather insensitive to gender-neutralization [28]. Therefore, lower BERTScore values should be attributed to differences in the meaning of the sentences beyond gender, which we evaluate separately, as described above. To identify reference values to guide the interpretation of BERTScore in GNR, we compute the distribution of BERTScore of mGeNTE REF-N sentences against the respective REF-G.[6] As these neutral reformulations were produced by human experts, the BERTScore distribution provides an empirical estimate of human-level performance in meaning preservation in GNR. We take the mean BERTScore minus one standard deviation $(0.9334 - 0.0546 = 0.879)$ as a reference threshold to obtain a conservative estimate of human-level performance. This accounts for the natural variability in human reformulations while filtering out outliers with lower semantic similarity, thus ensuring that models evaluated above this threshold perform within the typical human range.

3.2. Models

We experiment with a diverse set of models spanning different families, architectures, scales, and language coverage. Table 2 summarizes our selection of models and how we use them in our experiments. Our selection includes:

- **'Italian' models**, specifically designed or adapted for Italian language tasks: Minerva 7B, LLaMAntino 8B, and Velvet 14B.
- **Multilingual LLMs**, trained on multiple languages including Italian, to evaluate general-purpose models: Llama 3.1 8B, Phi 4 14B, Llama 3.3 70B. Among the multilingual LLMs, we include four different-sized models from the **Qwen3 family**, to analyze consistency and scalability within a single architecture.
- One **commercial system**, included as a high-performance reference system: GPT-4.1.[7]
- Inclusively,[8] a fine-tuning of `it5-large` [49], as the only **dedicated model** for Italian GNR. We consider this system the baseline for our experiments.

[5] Prompt: 'Mono+P+L'; GPT model: `gpt-4o-2024-08-06`
[6] We only use Set-N entries in this computation.
[7] Model `gpt-4.1-2025-04-14`
[8] https://huggingface.co/E-MIMIC/inclusively-reformulation-it5

GFG, Italian	**GFG, English**
Riformula la seguente frase utilizzando un linguaggio neutro rispetto al genere dei referenti umani, evitando l'uso di forme maschili e femminili.	Rewrite the following Italian sentence using a gender-neutral language in reference to human beings, avoiding masculine or feminine forms.
REWRITE, Italian	**REWRITE, English**
Sei un riscrittore di frasi italiane con l'obiettivo di rendere i testi neutrali rispetto al genere dei referenti umani. Ti viene fornita una frase che contiene riferimenti a persone in forme marcate per genere, come il maschile sovraesteso o coppie binarie. Il tuo compito è riformulare la frase in modo da: • rimuovere riferimenti espliciti al genere quando non necessari; • mantenere inalterato il significato originale; • preservare lo stile e la leggibilità del testo. Per farlo, usa strategie come: • sostantivi collettivi ("la cittadinanza", "il personale", "l'utenza"); • perifrasi impersonali ("si dovrebbe", "si consiglia"); • forme passive ("l'accesso è consentito"); • forme imperative ("allega il documento"); • pronomi relativi e costruzioni subordinate ("chi ha svolto attività di pesca"); • termini epiceni ("ogni giudice", "gentile collega"); • termini neutri ("l'individuo", "la persona interessata", "il membro"). IMPORTANTE: • evita l'uso del maschile come forma generica e non usare forme grafiche non standard come asterischi o schwa; • evita doppie formulazioni come "il/a cittadino/a" oppure "il professore o la professoressa"; • non rimuovere parti della frase che non richiedono modifiche (ad esempio, i nomi propri); • fornisci solo la frase riformulata.	You are a rewriter of Italian sentences with the goal of making texts gender-neutral with respect to human referents. You are given a sentence that contains references to people using gender-marked forms (such as masculine generics or binary pairs). Your task is to rewrite the sentence to: • remove explicit gender references when they are not necessary; • preserve the original meaning; • maintain the style and readability of the text. To do this, use strategies such as: • collective nouns ("la cittadinanza", "il personale", "l'utenza"); • impersonal phrases ("si dovrebbe", "si consiglia"); • passive constructions ("l'accesso è consentito"); • imperative constructions ("allega il documento"); • relative pronouns and subordinate clauses ("chi ha svolto attività di pesca"); • epicene terms ("ogni giudice", "gentile collega"); • neutral terms ("l'individuo", "la persona interessata", "il membro"). IMPORTANT: • avoid using the masculine form as a generic and do not use non-standard spellings such as asterisks or schwa; • avoid binary formulations such as "il/a cittadino/a" or "il professore o la professoressa"; • do not remove any part of the sentence that does not need to be rewritten (e.g. proper names); • only return the reformulated sentence.

Table 3
The 'system' role messages for the two prompt formats used in the few-shot prompting experiments, in both Italian and English.

All models, except for Inclusively, are instruction-tuned autoregressive LLMs.

3.3. Few-Shot Prompting

We run few-shot prompting experiments with all models in the selection described above,[9] to investigate the performance of LLMs without any task-specific fine-tuning. We use two prompt formats:

- **GFG**: a concise rewriting instruction, originally used by Frenda et al. [34] in their gender-fair generation challenge for Italian LLMs.
- **REWRITE**: a more detailed and analytical prompt, also featuring essential guidelines for the task with neutralization examples following the strategies identified by Piergentili et al. [9].

These prompts allow us to explore the impact of more complex instruction on models' performance. Moreover, we experiment with these two prompt formats by formulating them in both Italian an English, to investigate whether the language used is a relevant factor as well. The content of the prompts is reported in Table 3. We include the same 8 task exemplars—or shots—with all prompts, to elicit the in-context learning ability of LLMs [50]. We use vLLM [51] as the inference engine.

3.4. Fine-tuning

We perform fine-tuning experiments to assess whether and to which extent smaller open-weight LLMs can be adapted to the GNR task and approach the performance

[9] Except for Inclusively, which does not support few-shot prompting. We instead test its off-the-shelf generation capabilities.

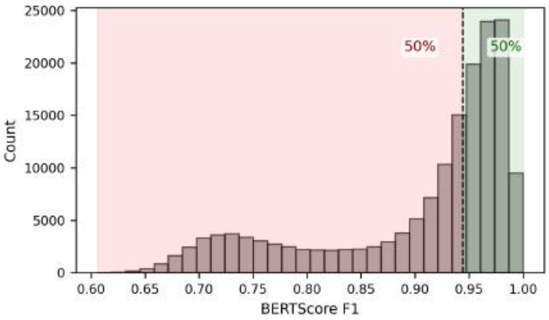

Figure 1: Distribution of BERTScore values over the FULL fine-tuning dataset. The CLEAN split is also visualized as the green portion starting at the median line (0.9443).

Set	Entries	Selection criterion	Avg. BERTScore
FULL	162,778	-	0.9044
CLEAN	81,389	BERTScore $\geq$ median	0.9697

Table 4
Training datasets statistics and summary.

of larger models or closed systems. Namely, we fine-tune LLaMAntino, Velvet, LLama 3.1, Phi 4, and the 8B and 14B Qwen3 models.

3.4.1. Data

The only openly available development data dedicated to Italian GNR is the synthetically generated training dataset used by Piergentili et al. [28] to train a gender-neutrality classifier.[10] This data consists in gendered Italian sentences and their gender-neutral counterparts, all generated starting from a dictionary of masculine, feminine, and neutral expressions, through a multi-step prompting pipeline. We repurpose this data to fine-tune autoregressive LLMs for GNR. We prepare the data as chat-formatted input, where each instance consists of a *user* role message containing a gendered sentence, and an *assistant* role message containing the corresponding neutral sentence. Consistent with the models' prior instruction-following fine-tuning, this method adopts a conversational prompt–response format while strictly adhering to a causal token-prediction objective [52].

As the sentences were partly LLM-generated, we note that the content of the gendered-neutral pairs may not always be aligned due to the unpredictability of LLMs in open-ended generation.[11] To investigate this aspect, we compare the gendered and neutral sentences in the dataset using BERTScore to identify dataset entries with semantically divergent gendered-neutral sentences. Figure 1 reports the BERTScore values for the entire dataset. We observe that while the score distribution is skewed towards almost-perfect values, there is a notable tail of gendered-neutral sentence pairs with a rather divergent semantic content. To investigate the impact of such data in GNR fine-tuning, we construct a subset to be used for training alongside the FULL dataset: a CLEAN subset obtained by filtering out the bottom 50% of sentence pairs based on the BERTScore values. Statistics about the fine-tuning data are reported in Table 4.

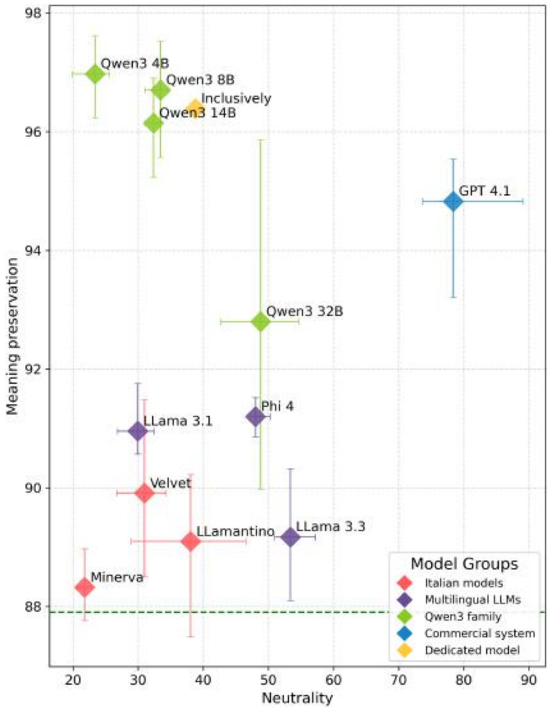

Figure 2: Results of the few-shot prompting experiments. The meaning preservation (vertical) axes report BERTScore values multiplied by 100 for easier visualization, whereas the neutrality (horizontal) axes report sentence-level neutralization accuracy. Each ◊ represents the average performance of a model across four prompts. The lines extending from each ◊ indicate the full range of values observed for that model on the respective axis. The dashed line indicates the reference value for human-level meaning preservation in GNR.

3.4.2. Method

We fine-tune the selected models using Low-Rank Adaptation (LoRA) [53]. Following common practices in LoRA fine-tuning [54] we set the `rank` and `alpha` at 32, and use the following hyperparameters to strike a

[10] More specifically, we use the cleaned version of the dataset later released by Savoldi et al. [32] at https://github.com/hlt-mt/fbk-NEUTR-evAL/blob/main/solutions/GeNTE.md

[11] While this is not necessarily an issue in the development of a classifier, where individual sentences are simply paired with neutrality labels, for a rewriting task the input-output sentences should be identical except for the attribute of interest, i.e., in this case, gender.

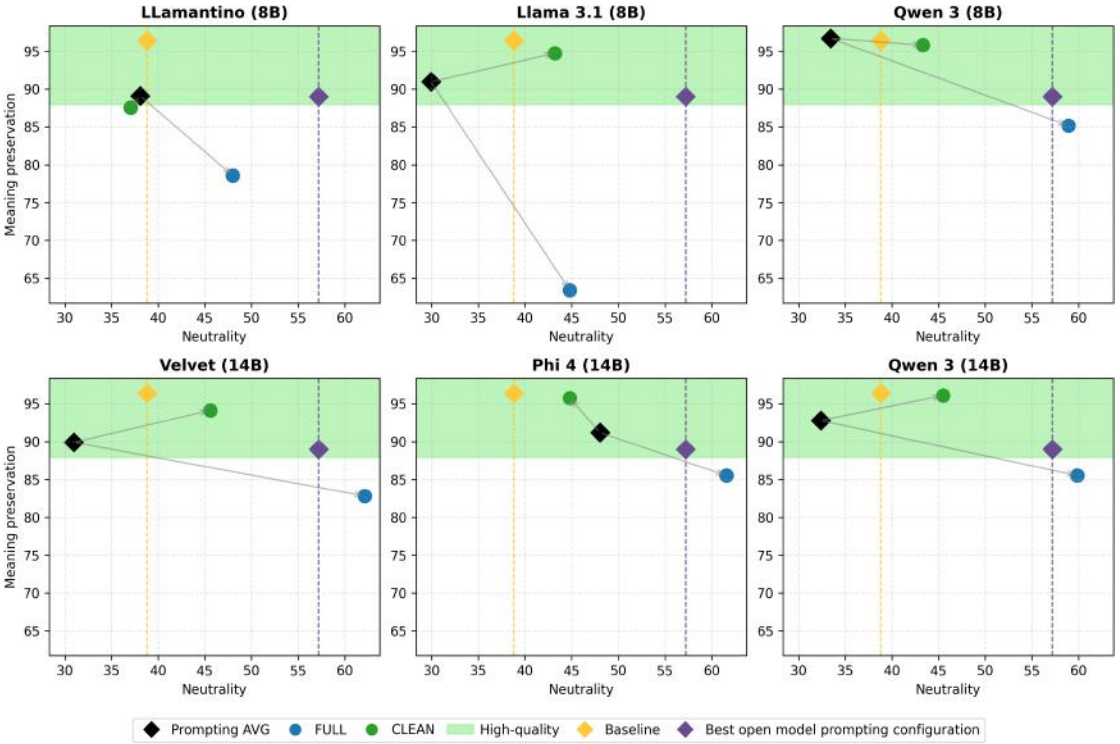

Figure 3: Results of the fine-tuning experiments. The meaning preservation (vertical) axes report BERTScore values multiplied by 100 for easier visualization, whereas the neutrality (horizontal) axes report sentence-level neutralization accuracy. The black diamond represents the average performance of the model in the prompting experiments. The blue and green points represent the performance of the model fine-tuned on the FULL and CLEAN datasets respectively. The green band at the top represents BERTScore values reaching human-level meaning preservation in GNR. The yellow and blue points and dashed vertical lines respectively represent the baseline (the dedicated model Inclusively) and the best configuration performance of an open-weight model (LLama 3.3 70B, GFG English prompt).

balance between hardware constraints[12] and consistency across model sizes and requirements: `learning rate`: 2×10^{-4}, `batch size`: 8 for the 8B models, 4 for the 14B models. We use early stopping with a patience of 20 steps for the 8B models and 40 steps for the 14B models.

4. Results

4.1. Few-Shot Prompting Results

Figure 2 summarizes the results of the few-shot prompting experiments showing all models' performance in neutrality and meaning preservation. Higher values on both axes indicate better performance; therefore, systems closer to the top-right corner perform best. As no consistent trend emerged across prompt formats (GFG vs. REWRITE, see Section 3.3) and languages (Italian vs. English), we report each model's average performance, along with the range of neutrality and BERTScore values observed across prompting conditions. In Appendix A we provide the complete and detailed results obtained with the two prompt formats, separately for Italian and English instructions.

Generally, and with rare exceptions, all models' BERTScore values are well above the quality threshold we identified in §3.1. This means that the models do not generate unrelated or additional text, confirming that their outputs remain adherent to the input and free of "hallucinations" [55].

Neutrality scores, on the contrary, vary significantly across models. Looking at our baseline, the GNR-dedicated model Inclusively, we observe that it performs rather poorly in neutrality. Across LLMs, we notice similar behavior within the groups. The "Italian" models, in the bottom left quarter of the chart, generally fail to neutralize, and alter the sentences the most. Within the multilingual LLMs group, only Phi 4, Qwen3 32B, and

[12] We run our experiments on nodes with 4 NVIDIA A100 GPUs with 64 GB VRAM each.

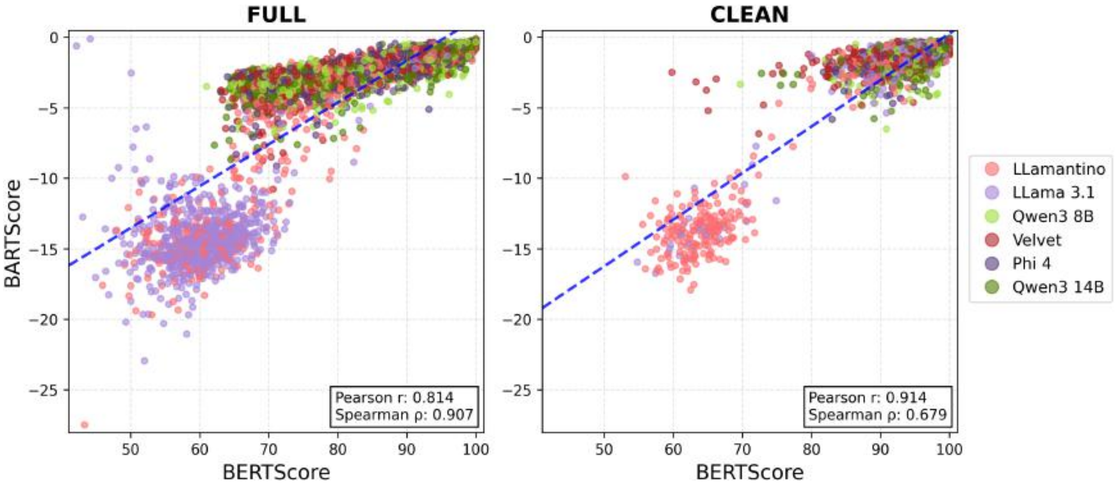

Figure 4: BERTScore and BARTScore for the outputs of the models fine-tuned on both FULL and CLEAN. The dashed lines are least-squares regression lines fitted to each set of points, modeling the relationship between the metrics. Points above the line have higher BARTScore than predicted by BERTScore (i.e. BERTScore underrates them), and vice versa for points below. We report Pearson r and Spearman ρ correlation coefficients for each split as well.

LLama 3.3 perform better than the Italian models. The rest of the Qwen3 family generally underperforms, with the high BERTScore suggesting that they make little to no change to the gendered sentences. The only model performing well on both axes is GPT 4.1, which tops at 89.07% neutralization accuracy and 93.21 BERTScore, indicating that it correctly alters the parts of the sentences expressing the gender of human beings while leaving the rest untouched.

Overall, we find that the LLMs we tested perform very differently in GNR in Italian, and that failure in this setting consists in overlooking the relevant (gendered) parts of the input to act upon, and/or unsuccessfully rendering them gender-neutral.

4.2. Fine-Tuning Results

Results of the fine-tuning experiments are reported in Figure 3. We first notice that on the neutrality axis all fine-tuned models outperform the baseline, except for LLamantino/CLEAN configuration. LLamantino shows the narrowest gains overall, and in one case even a drop in neutrality, echoing its weaker few-shot prompting results and suggesting it may be ill-suited to GNR. In four out of six instances, and always with the FULL dataset, the fine-tuned models also outperform the best performer among the open-weight models in the prompting experiments, i.e. LLama 3.3 70B with the GFG English prompt, though with a significant drop in BERTScore.

Such a drop indicates that these models fail by hallucinating unrelated content in their attempt to neutralize, rather than by leaving the input sentences untouched as observed in the prompting experiments (§4.1). This is possibly due to two factors: the significantly smaller size of the fine-tuned models with respect to LLama 3.3 70B (1/9 or 1/5, for the 8B and 14B models respectively), as larger LLMs have been shown to exhibit greater robustness and lower variance in downstream performance after fine-tuning compared to smaller counterparts [56], and/or the presence of many divergent gender-neutral sentence pairs in the fine-tuning dataset (see §3.4.1).

While FULL yields the highest improvements in neutrality, only CLEAN improves performance on both axes while keeping BERTScore within the human-level range. However, it yields significantly smaller gains in neutrality and even causes drops for two models (LLamantino, Phi 4). We hypothesize that CLEAN may be excessively conditioned by the data filtering method, i.e. a BERTScore based selection. In other words, by selecting only dataset entries with almost perfect BERTScore values we are optimizing the models to perform well on the sentence similarity dimension—as measured by BERTScore—rather than GNR.

The impact of metric-based data selection To investigate the hypothesis above, we evaluate the same outputs against the gendered inputs with another semantic similarity metric: BARTScore [57].[13] BERTScore and

[13] While similar in name and scope, BERTScore and BARTScore function differently. The first computes a sum of token-level cosine similarities between two sentences' embeddings encoded by a BERT (encoder-only) model; the latter is computed as the weighted sum of the log-probabilities that a pretrained BART (encoder-decoder) model assigns to each token in the generated text. In our experi-

BARTScore evaluations are visualized in Figure 4. To understand whether outputs of the models fine-tuned on CLEAN are actually very semantically similar to the corresponding input, and whether those models simply learned to game BERTScore, we compute[14] the Pearson r and Spearman ρ correlation coefficients between BERTScore and BARTScore assessments. The first captures linear correlations between the two metrics' raw scores, while the latter measures how well the relationship between the two variables can be described by a monotonic function, by comparing the rankings of the scores rather than their raw values. This combination allows us to assess both the alignment of the scores and the consistency in how the two metrics rank the outputs.

We find that in FULL, r equals 0.814 and ρ equals 0.907, whereas in CLEAN they are 0.914 and 0.679 respectively.[15] r is high in both cases, indicating a strong linear correlation between the two metrics—stronger in CLEAN, as in that case the data points are more tightly clustered, skewed towards higher values. This confirms that the metrics generally agree on the quality of the outputs. The substantial drop in ρ, instead, indicates that there are many instances in CLEAN where the monotonic trend is broken, i.e., higher BERTScore does not necessarily correspond to higher BARTScore. This suggests that the CLEAN models also learned to game BERTScore by reproducing features rewarded by that metric.

With respect to our hypothesis: by selecting high-similarity pairs for the CLEAN dataset, we effectively steered models toward preserving semantic alignment with the input; however, this emphasis on similarity appears to have hampered their improvement in neutralization. Indeed, the models learned to preserve the input to an excessive degree, as confirmed by the high r coefficient and high BARTScore values shown in Figure 4. We interpret our results as evidence of a broader trade-off between optimizing for neutrality and for sentence similarity. Our findings underscore the need for data curation strategies that strike a balance between neutrality and similarity, achieving the flexibility required for effective GNR.

5. Conclusions

We presented the first systematic investigation of state-of-the-art large language models for Italian gender-neutral rewriting under a two-dimensional evaluation of neutrality and meaning preservation. In our few-shot prompting experiments, open-weight models outperformed the only existing Italian-specific system but remained behind a closed commercial system.

Through fine-tuning experiments we showed that compact models can match or exceed the best open-weight LLM at a fraction of its size. Moreover, our BERTScore-based data cleaning highlighted a trade-off: models trained on cleaned data achieve human-level BERTScore but show smaller neutrality gains and exhibit ranking differences against another similarity metric, signaling over-fitting on BERTScore. Future work should take this trade-off into account and create dedicated, high-quality parallel data to aim at reaching the performance of the commercial system with open-weight models.

Acknowledgments

We acknowledge the support of the project InnovAction: Network Italiano dei Centri per l'Innovazione Tecnologica (CUP B47H2200437000), funded by MIMIT with NPRR - NextGenerationEU funds, in collaboration with Piazza Copernico S.r.l. We also received funding from the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by NextGenerationEU. Finally, we acknowledge the CINECA award under the ISCRA initiative (AGeNTE), for the availability of high-performance computing resources and support.

References

[1] B. Savoldi, S. Papi, M. Negri, A. Guerberof-Arenas, L. Bentivogli, 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, Association for Computational Linguistics, Miami, Florida, USA, 2024. URL: https://aclanthology.org/2024.emnlp-main.1002/. doi:10.18653/v1/2024.emnlp-main.1002.

[2] H. Kotek, R. Dockum, D. Sun, Gender bias and stereotypes in large language models, in: Proc. of The ACM Collective Intelligence Conference, CI '23, ACM, New York, NY, USA, 2023, p. 12–24. URL: https://doi.org/10.1145/3582269.3615599.

[3] R. Ostrow, A. Lopez, Llms reproduce stereotypes of sexual and gender minorities, 2025. arXiv:2501.05926.

[4] B. Savoldi, J. Bastings, L. Bentivogli, E. Vanmassenhove, A decade of gender bias in machine translation, Patterns (2025) 101257. URL: https://www.sciencedirect.com/science/article/pii/S2666389925001059.

[5] S. L. Blodgett, S. Barocas, H. Daumé III, H. Wallach, Language (technology) is power: A critical survey of "bias" in NLP, in: Proc. of the 58th Annual Meet-

ments, we use the BART model `facebook/bart-large` [58].

[14] We use the Python library `SciPy` [59].

[15] All p-values < 0.05.

ing of the ACL, ACL, Online, 2020, pp. 5454–5476. URL: https://aclanthology.org/2020.acl-main.485/.

[6] S. Dev, M. Monajatipoor, A. Ovalle, A. Subramonian, J. Phillips, K.-W. Chang, Harms of gender exclusivity and challenges in non-binary representation in language technologies, in: Proc. of the 2021 Conference on Empirical Methods in Natural Language Processing, ACL, Online and Punta Cana, Dominican Republic, 2021, pp. 1968–1994. URL: https://aclanthology.org/2021.emnlp-main.150/.

[7] U. Gabriel, P. M. Gygax, E. A. Kuhn, Neutralising linguistic sexism: Promising but cumbersome?, Group Processes & Intergroup Relations 21 (2018) 844–858.

[8] APA, Publication Manual of the APA, 7th ed., 2020.

[9] A. Piergentili, D. Fucci, B. Savoldi, L. Bentivogli, M. Negri, Gender neutralization for an inclusive machine translation: from theoretical foundations to open challenges, in: Proc. of the First Workshop on Gender-Inclusive Translation Technologies, EAMT, Tampere, Finland, 2023, pp. 71–83. URL: https://aclanthology.org/2023.gitt-1.7/.

[10] T. Sun, K. Webster, A. Shah, W. Y. Wang, M. Johnson, They, them, theirs: Rewriting with gender-neutral english, 2021. arXiv:2102.06788.

[11] E. Vanmassenhove, C. Emmery, D. Shterionov, NeuTral Rewriter: A rule-based and neural approach to automatic rewriting into gender neutral alternatives, in: Proc. of the 2021 Conference on Empirical Methods in Natural Language Processing, ACL, Online and Punta Cana, Dominican Republic, 2021, pp. 8940–8948. URL: https://aclanthology.org/2021.emnlp-main.704/.

[12] M. Bartl, S. Leavy, From 'showgirls' to 'performers': Fine-tuning with gender-inclusive language for bias reduction in LLMs, in: Proc. of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP), ACL, Bangkok, Thailand, 2024, pp. 280–294. URL: https://aclanthology.org/2024.gebnlp-1.18/.

[13] E. Doyen, A. Todirascu, Genre: A french gender-neutral rewriting system using collective nouns, 2025. arXiv:2505.23630.

[14] E. Rose, M. Winig, J. Nash, K. Roepke, K. Conrod, Variation in acceptability of neologistic English pronouns, Proc. of the Linguistic Society of America 8 (2023) 5526. URL: https://journals.linguisticsociety.org/proceedings/index.php/PLSA/article/view/5526.

[15] D. Pomerenke, Inclusify: A benchmark and a model for gender-inclusive german, 2022. arXiv:2212.02564.

[16] L. Veloso, L. Coheur, R. Ribeiro, A rewriting approach for gender inclusivity in Portuguese, in: Findings of the ACL: EMNLP 2023, ACL, Singapore, 2023, pp. 8747–8759. URL: https://aclanthology.org/2023.findings-emnlp.585/.

[17] P. Lerner, C. Grouin, INCLURE: a dataset and toolkit for inclusive French translation, in: Proc. of the 17th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC-COLING 2024, ELRA and ICCL, Torino, Italia, 2024, pp. 59–68. URL: https://aclanthology.org/2024.bucc-1.7/.

[18] S. Greco, M. La Quatra, L. Cagliero, T. Cerquitelli, Towards ai-assisted inclusive language writing in italian formal communications, ACM Trans. Intell. Syst. Technol. (2025). URL: https://doi.org/10.1145/3729237.

[19] B. Papadopoulos, Morphological Gender Innovations in Spanish of Gender queer Speakers, Department of Spanish and Portuguese, University of California, UC Berkeley, 2019. URL: https://escholarship.org/uc/item/6j73t666.

[20] G. S. di Carlo, Is italy ready for gender-inclusive language? an attitude and usage study among italian speakers, in: Inclusiveness Beyond the (Non)binary in Romance Languages, 1st edition ed., Routledge, 2024, p. 21. URL: https://doi.org/10.4324/9781003432906.

[21] G. V. Silva, C. Soares, Inclusiveness Beyond the (Non)binary in Romance Languages: Research and Classroom Implementation, 1st ed., Routledge, London, 2024. doi:10.4324/9781003432906.

[22] P. Gygax, S. Sato, A. Öttl, U. Gabriel, The masculine form in grammatically gendered languages and its multiple interpretations: a challenge for our cognitive system, Language Sciences 83 (2021) 101328. URL: https://www.sciencedirect.com/science/article/pii/S0388000120300619.

[23] L. Ackerman, Syntactic and cognitive issues in investigating gendered coreference, Glossa: a journal of general linguistics 4 (2019).

[24] M. Rosola, S. Frenda, A. T. Cignarella, M. Pellegrini, A. Marra, M. Floris, Beyond obscuration and visibility: Thoughts on the different strategies of gender-fair language in Italian, in: Proc. of the 9th Italian Conference on Computational Linguistics (CLiC-it 2023), CEUR Workshop Proc., Venice, Italy, 2023, pp. 369–378. URL: https://aclanthology.org/2023.clicit-1.44/.

[25] G. Comandini, 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 (2021) 43–64.

[26] J. Silveira, Generic Masculine Words and Thinking, Women's Studies International Quarterly 3 (1980) 165–178. URL: https://www.sciencedirect.com/science/article/pii/S0148068580921132.

[27] A. H. Bailey, A. Williams, A. Cimpian, Based on

billions of words on the internet, people= men, Science Advances 8 (2022) eabm2463.

[28] A. Piergentili, B. Savoldi, D. Fucci, M. Negri, L. Bentivogli, Hi guys or hi folks? benchmarking gender-neutral machine translation with the GeNTE corpus, in: Proc. of the 2023 Conference on Empirical Methods in Natural Language Processing, ACL, Singapore, 2023, pp. 14124–14140. URL: https://aclanthology.org/2023.emnlp-main.873/.

[29] F. Höglund, M. Flinkfeldt, De-gendering parents: Gender inclusion and standardised language in screen-level bureaucracy, International Journal of Social Welfare (2023).

[30] Y. T. Cao, H. Daumé III, Toward gender-inclusive coreference resolution, in: Proc. of the 58th Annual Meeting of the ACL, ACL, Online, 2020, pp. 4568–4595. URL: https://aclanthology.org/2020.acl-main.418/.

[31] A. Waldis, J. Birrer, A. Lauscher, I. Gurevych, The Lou dataset - exploring the impact of gender-fair language in German text classification, in: Proc. of the 2024 Conference on Empirical Methods in Natural Language Processing, ACL, Miami, Florida, USA, 2024, pp. 10604–10624. URL: https://aclanthology.org/2024.emnlp-main.592/.

[32] B. Savoldi, A. Piergentili, D. Fucci, M. Negri, L. Bentivogli, A prompt response to the demand for automatic gender-neutral translation, in: Proc. of the 18th Conference of the European Chapter of the ACL (Volume 2: Short Papers), ACL, St. Julian's, Malta, 2024, pp. 256–267. URL: https://aclanthology.org/2024.eacl-short.23/.

[33] B. Savoldi, G. Attanasio, E. Cupin, E. Gkovedarou, J. Hackenbuchner, A. Lauscher, M. Negri, A. Piergentili, M. Thind, L. Bentivogli, Mind the inclusivity gap: Multilingual gender-neutral translation evaluation with mGeNTE, 2025. URL: https://openreview.net/forum?id=dBUHC2QyBh.

[34] S. Frenda, A. Piergentili, B. Savoldi, M. Madeddu, M. Rosola, S. Casola, C. Ferrando, V. Patti, M. Negri, L. Bentivogli, GFG - gender-fair generation: A CALAMITA challenge, in: Proc. of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR Workshop Proc., Pisa, Italy, 2024, pp. 1106–1115. URL: https://aclanthology.org/2024.clicit-1.122/.

[35] G. Attanasio, P. Delobelle, M. La Quatra, A. Santilli, B. Savoldi, ItaEval and TweetyIta: A new extensive benchmark and efficiency-first language model for Italian, in: Proc. of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR Workshop Proc., Pisa, Italy, 2024, pp. 39–51. URL: https://aclanthology.org/2024.clicit-1.6/.

[36] R. Orlando, L. Moroni, P.-L. Huguet Cabot, S. Conia, E. Barba, S. Orlandini, G. Fiameni, R. Navigli, Minerva LLMs: The first family of large language models trained from scratch on Italian data, in: Proc. of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR Workshop Proc., Pisa, Italy, 2024, pp. 707–719. URL: https://aclanthology.org/2024.clicit-1.77/.

[37] P. Basile, E. Musacchio, M. Polignano, L. Siciliani, G. Fiameni, G. Semeraro, Llamantino: Llama 2 models for effective text generation in italian language, 2023. arXiv:2312.09993.

[38] Almawave, Velvet, 2025. URL: https://www.almawave.com/it/tecnologia/velvet/.

[39] M. Llama Team, The llama 3 herd of models, 2024. arXiv:2407.21783.

[40] M. Abdin, J. Aneja, H. Behl, S. Bubeck, R. Eldan, S. Gunasekar, M. Harrison, R. J. Hewett, M. Javaheripi, P. Kauffmann, J. R. Lee, Y. T. Lee, Y. Li, W. Liu, C. C. T. Mendes, A. Nguyen, E. Price, G. de Rosa, O. Saarikivi, A. Salim, S. Shah, X. Wang, R. Ward, Y. Wu, D. Yu, C. Zhang, Y. Zhang, Phi-4 technical report, 2024. arXiv:2412.08905.

[41] A. Qwen Team, Qwen3 technical report, 2025. arXiv:2505.09388.

[42] OpenAI, Introducing gpt-4.1 in the api, 2025. URL: https://openai.com/index/gpt-4-1/, accessed: 2025-05-15.

[43] D. Li, B. Jiang, L. Huang, A. Beigi, C. Zhao, Z. Tan, A. Bhattacharjee, Y. Jiang, C. Chen, T. Wu, K. Shu, L. Cheng, H. Liu, From generation to judgment: Opportunities and challenges of llm-as-a-judge, 2025. arXiv:2411.16594.

[44] A. Piergentili, B. Savoldi, M. Negri, L. Bentivogli, An LLM-as-a-judge approach for scalable gender-neutral translation evaluation, in: Proceedings of the 3rd Workshop on Gender-Inclusive Translation Technologies (GITT 2025), EAMT, Geneva, Switzerland, 2025, pp. 46–63. URL: https://aclanthology.org/2025.gitt-1.3/.

[45] T. Zhang*, V. Kishore*, F. Wu*, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, in: International Conference on Learning Representations, 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr.

[46] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proc. of the 2019 Conference of the North American Chapter of the ACL: Human Language Technologies, Volume 1 (Long and Short Papers), ACL, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423/.

[47] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proc. of the 40th Annual Meeting of the ACL, ACL, Philadelphia, Pennsylvania, USA,

2002, pp. 311–318. URL: https://aclanthology.org/P02-1040/.

[48] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, J. Makhoul, A study of translation edit rate with targeted human annotation, in: Proc. of the 7th Conference of the AMTA: Technical Papers, AMTA, Cambridge, Massachusetts, USA, 2006, pp. 223–231. URL: https://aclanthology.org/2006.amta-papers.25/.

[49] G. Sarti, M. Nissim, IT5: Text-to-text pretraining for Italian language understanding and generation, in: Proc. of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 9422–9433. URL: https://aclanthology.org/2024.lrec-main.823.

[50] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, Neelakantan, et al., Language models are few-shot learners, in: Advances in NeurIPS, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[51] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, I. Stoica, Efficient memory management for large language model serving with pagedattention, 2023. arXiv:2309.06180.

[52] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, R. Lowe, Training language models to follow instructions with human feedback, in: Proc. of the 36th International Conference on NeurIPS, NIPS '22, Curran Associates Inc., Red Hook, NY, USA, 2022.

[53] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, 2021. arXiv:2106.09685.

[54] Unsloth Documentation, Lora hyperparameters guide, 2025. URL: https://docs.unsloth.ai/get-started/fine-tuning-llms-guide/lora-hyperparameters-guide.

[55] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, T. Liu, A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, ACM Transactions on Information Systems 43 (2025) 1–55. URL: http://dx.doi.org/10.1145/3703155.

[56] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tai, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, A. Castro-Ros, M. Pellat, K. Robinson, D. Valter, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, J. Wei, Scaling instruction-finetuned language models, J. Mach. Learn. Res. 25 (2024).

[57] W. Yuan, G. Neubig, P. Liu, Bartscore: evaluating generated text as text generation, in: Proc. of the 35th International Conference on NeurIPS, NIPS '21, Curran Associates Inc., Red Hook, NY, USA, 2021.

[58] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, CoRR abs/1910.13461 (2019). URL: http://arxiv.org/abs/1910.13461. arXiv:1910.13461.

[59] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, SciPy 1.0 Contributors, SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python, Nature Methods 17 (2020) 261–272.

A. Detailed results

Tables 5 and 6 report the detailed results of our fine-tuning experiments.

Neutrality

Model		Size (B)	GFG Ita	GFG Eng	Rewrite Ita	Rewrite Eng	AVG
"Italian" models	Minerva	7	20.67	22.80	22.67	21.07	21.80
	LLaMAntino	8	28.93	31.07	46.53	45.73	38.07
	Velvet	14	32.40	34.27	30.53	26.67	30.97
Multilingual LLMs	Llama 3.1	8	26.80	28.27	32.27	32.40	29.93
	Phi 4	14	47.47	47.20	47.20	50.27	48.03
	Llama 3.3	70	52.93	57.20	52.40	50.93	53.37
Qwen3 family	Qwen3	4	23.87	19.87	25.60	24.27	23.40
	Qwen3	8	33.60	34.67	34.40	31.07	33.43
	Qwen3	14	32.27	31.07	33.47	32.67	32.37
	Qwen3	32	54.67	52.80	42.67	45.07	48.80
Commercial system	GPT 4.1	?	75.33	**89.07**	73.73	75.33	78.37
Dedicated model	Inclusively	0.78	38.80				38.80

Table 5

Neutrality results of the few-shot prompting experiments. The best model settings are underlined, the best settings across the categories are highlighted, and the best overall performer is in **bold**.

BERTScore

Model		Size (B)	GFG Ita	GFG Eng	Rewrite Ita	Rewrite Eng	AVG
"Italian" models	Minerva	7	87.78	88.78	87.76	88.97	88.32
	LLaMAntino	8	89.97	90.22	87.49	88.70	89.09
	Velvet	14	89.60	91.48	88.50	90.06	89.91
Multilingual LLMs	Llama 3.1	8	91.76	90.70	90.78	90.57	90.95
	Phi 4	14	90.86	90.95	91.52	91.46	91.20
	Llama 3.3	70	88.10	89.00	89.26	90.32	89.17
Qwen3 family	Qwen3	4	96.23	96.98	97.07	**97.62**	96.97
	Qwen3	8	96.49	95.57	97.23	97.52	96.70
	Qwen3	14	95.23	96.72	95.72	96.91	96.14
	Qwen3	32	89.98	91.31	94.04	95.86	92.80
Commercial system	GPT 4.1	?	95.12	93.21	95.54	95.44	94.83
Dedicated model	Inclusively	0.78	96.39				96.39

Table 6

Sentence-similarity results of the few-shot prompting experiments. The best model settings are underlined, the best settings across the categories are highlighted, and the best overall performer is in **bold**.