A *Prompt* Response to the Demand for Automatic Gender-Neutral Translation

Beatrice Savoldi,[1] Andrea Piergentili,[1,2] Dennis Fucci,[1,2] Matteo Negri[1], Luisa Bentivogli[1]
[1] Fondazione Bruno Kessler
[2] University of Trento
{bsavoldi,apiergentili,dfucci,negri,bentivo}@fbk.eu

Abstract

Gender-neutral translation (GNT) that avoids biased and undue binary assumptions is a pivotal challenge for the creation of more inclusive translation technologies. Advancements for this task in Machine Translation (MT), however, are hindered by the lack of dedicated parallel data, which are necessary to adapt MT systems to satisfy neutral constraints. For such a scenario, large language models offer hitherto unforeseen possibilities, as they come with the distinct advantage of being versatile in various (sub)tasks when provided with explicit instructions. In this paper, we explore this potential to automate GNT by comparing MT with the popular GPT-4 model. Through extensive manual analyses, our study empirically reveals the inherent limitations of current MT systems in generating GNTs and provides valuable insights into the potential and challenges associated with prompting for neutrality.

1 Introduction

To foster greater inclusivity in our communication practices, there has been a rise in the adoption of gender-neutral language strategies (Hord, 2016; Papadimoulis, 2018), which challenge gender norms and embrace all identities by eschewing unnecessary gendered terms (e.g. *police officer* vs *policeman*). Such strategies are now widespread across various domains – including institutions (Höglund and Flinkfeldt, 2023), academia (APA, 2020), and industry (Langston, 2020), with their consequential investigation for various natural language processing (NLP) technologies (Cao and Daumé III, 2020; Brandl et al., 2022; Wagner and Zarrieß, 2022).

While recent advancements in NLP have seen the modeling of neutral language into monolingual applications (Vanmassenhove et al., 2021; Sun et al., 2021; Amrhein et al., 2023; Veloso et al., 2023), research in cross-lingual settings is relatively limited. Previous works in MT (Costa-jussà and de Jorge, 2020; Savoldi et al., 2021; Choubey et al., 2021; Alhafni et al., 2022; Piazzolla et al., 2023, *inter alia*) have been mostly confined within binary perspectives to improve the generation of masculine/feminine forms into grammatical gender languages (e.g. *doctors* → it: *dottor**i**/**esse***).[1] Under realistic scenarios though, systems often encounter ambiguous input sentences that do not convey gender distinctions (Saunders, 2023; Piergentili et al., 2023a), and for which GNT would be preferable to prevent undue gender assignments in the target language (e.g. en: ***doctors*** → *it:* ***personale medico***`[the medical staff]`).

Despite individual studies indicating that existing MT systems are ill-equipped to handle neutrality (Cho et al., 2019; Piergentili et al., 2023b; Savoldi et al., 2023), the automation of GNT remains an open challenge, hampered by the lack of dedicated resources. To the best of our knowledge, the work by Saunders et al. (2020) stands as the sole effort to create gender-neutral MT models, but their fine-tuning approach does not generalize from their small artificial adaptation set. Within this landscape, large language models (LLMs) can offer a solution to meet the demand for gender neutrality, thanks to their adaptability to perform new (sub)tasks based on explicit instructions and few examples (Brown et al., 2020). In fact, albeit LLMs still lag slightly behind traditional MT in overall translation quality (Robinson et al., 2023; Vilar et al., 2023; Zhang et al., 2023), their versatility for controlling specific aspects in the output translation was proven for several attributes (Moslem et al., 2023; Sarti et al., 2023; Garcia and Firat, 2022; Yamada, 2023).

In this paper, we thus seek to advance the automation of neutral translation by exploring the po-

[1] Although in grammatical gender languages also inanimate nous are formally assigned to a gender class (Corbett, 1991), we are hereby only concerned with (social) gender assignment for human referents.

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 2: Short Papers, pages 256–267
March 17-22, 2024

tential of instruction-following models. To this aim, we focus on English→Italian and systematically compare the neutral capabilities of traditional MT models with GPT-4 (OpenAI, 2023). By experimenting with different prompts and shot-exemplars, we conduct a fine-grained, manual evaluation showing that: ***i)*** used *as is* neither MT nor GPT are suitable for GNT, but prompting GPT shows surprising neutralization capabilities elicited with just a few examples; ***ii)*** while including test set terms as neutralization exemplars in the prompts leads to slightly better GNT performance, GPT can generalize well also when provided with unseen examples. Finally, extensive manual evaluations unveil that ***iii)*** judging the quality and acceptability of automatic GNT is a subjective task, with notable variations across annotators. To promote future research, we make all our manual output annotations freely available at: `https://mt.fbk.eu/gente/`.[2]

2 Methods and Settings

Test set. We run our experiments on GeNTE (Piergentili et al., 2023b), a recently released parallel test set designed to evaluate models' GNT capabilitites. Built on Europarl data (Koehn, 2005), it allows us to test MT on naturalistic instances for en-it, a language pair that is highly representative of the challenges of performing GNT into languages with extensive gendered morphology. For such languages, neutral strategies can range from simple word changes (e.g. omissions or synonyms) to complex reformulations that can alter the sentence structure (Gabriel et al., 2018). Hence, generating suitable GNTs is a delicate and difficult task, to be carefully weighted not to impact the acceptability of a translation. Here, we use a portion of GeNTE consisting of 750 English sentences that are gender-ambiguous,[3] and which are thus to be neutrally translated so as to avoid any undue gender inference in Italian (e.g. *I, with* ***all my colleagues*** *wish to...*, it-M: *Io, con* ***tutti i colleghi*** *desidero...* → it-GNT: *Io, con* ***ogni collega***`[each colleague]`, *desidero...*).[4]

Systems. As MT models, we select two state-of-the-art commercial systems: Amazon Translate[5] and DeepL.[6] For GNT-PROMPTING, we use GPT (`gpt-4-0613`), which achieved promising results in translation (Jiao et al., 2023), though especially for high-resource languages (Robinson et al., 2023; Stap and Araabi, 2023). As an *instruction-following* model (Chung et al., 2022; Ouyang et al., 2022), GPT is suited to keep adherence to provided guidance when performing a task, a valuable aspect to control the neutral translation of gendered terms.

	BLEU	CHRF	BLEURT	COMET
Amazon	31.04	57.54	82.84	84.07
DeepL	30.75	56.30	82.80	83.90
GPT-4	25.08	51.94	80.56	82.60

Table 1: Overall quality results for en-it.

Experiments. We explore models' neutralization abilities under two experimental settings: *i)* BASELINE, to compare if the MT models and GPT in zero-shot conditions[7] can perform GNT, without being explicitly instructed/adapted for the task; and *ii)* GNT-PROMPTING, to leverage GPT potential when prompted with dedicated instructions and examples. In both settings, for GPT we use temperature 0.0, since Peng et al. (2023) attested a progressive translation degradation with higher temperature values.

Before delving into their GNT capabilities, in Table 1 we report the performance of all models on the Europarl common test set.[8] Such results confirm that GPT exhibits good cross-lingual capabilities, but does not match traditional MT models.

3 GNT-PROMPTING

To elicit GPT's flexibility for neutral translations, in the GNT-PROMPTING condition we experiment with three few-shot templates inspired by existing literature on prompting (Liu et al., 2023; Dong et al., 2023). Our prompts, shown in Table 2, are:

(1) `Contr`: consisting of *contrastive* examples of gendered and neutral translations for each English sentence, without additional verbalized instructions. This simple template has shown promising results for controlling the generation of (binary) gender forms (Sánchez et al., 2023).

(2) `CoT-src`: based on *chain-of-thought* demonstrations that break complex tasks into intermediate reasoning steps (Wei et al., 2023). This prompt first guides the identification of *source* terms that cor-

[2]Released under a Creative Commons Attribution 4.0 International license (CC BY 4.0).
[3]`Set-N` in the original corpus.
[4]For more details, see Appendix A.
[5]`https://aws.amazon.com/it/translate/`.
[6]`https://www.deepl.com/en/translator`.
[7]We adopt the best performing prompt by Peng et al. (2023): "`Please provide the [TGT] translation of the following sentence:`".
[8]`https://www.statmt.org/europarl/`.

`Contr`	[English]: Secondly, how far does it increase transparency and accountability **of the writers**? [Italian, gendered]: Secondariamente, fino a che punto aumenta la trasparenza e la responsabilità **degli scrittori**? [Italian, neutral]: Secondariamente, fino a che punto aumenta la trasparenza e la responsabilità **di chi scrive**?
`CoT-src`	Q: Translate the following English sentence into Italian using a gender-neutral language to refer to human entities: [Secondly, how far does it increase transparency and accountability **of the writers**]. Think step by step. A: In the English sentence there is one expression which refers to human entities and could be translated in a non-neutral way: <of the writers>. A gender-neutral translation of <of the writers> is <di chi scrive>. The final gender-neutral translation is [Secondariamente, fino a che punto aumenta la trasparenza e la responsabilità **di chi scrive**?]
`CoT-tgt`	Q: Translate the following English sentence into Italian using a gender-neutral language to refer to human entities: [Secondly, how far does it increase transparency and accountability **of the writers**?]. Think step by step. A: The English sentence can be translated as [Secondariamente, fino a che punto aumenta la trasparenza e la responsabilità **degli scrittori**?]. There is one «expression with <non-neutral terms>» that refers to human entities: «<degli scrittori>». A gender-neutral alternative to «<degli scrittori>» is «di chi scrive». The final gender-neutral translation is [Secondariamente, fino a che punto aumenta la trasparenza e la responsabilità **di chi scrive**?].

Table 2: Examples of each prompt template. The source "***of the writers***" – corresponding to the gendered "***degli scrittori***" in Italian – is neutralized as "***di chi scrive***" `[of who writes]`. `CoT-tgt` and `CoT-src` templates are structured as Questions and Answers. The final gender-neutral translations are highlighted.

Seen		Not seen	
en	**it**	**en**	**it**
MEPs	parlamentari europei	writers	scrittori
President	Signora Presidente	manager	direttore
everyone	tutti	employees	impiegati
politicians	politici	musicians	musicisti
fishermen	pescatori	freshmen	studenti del primo anno

Table 3: Source English and target Italian pairs of *seen* and *not seen* terms used in the exemplar sentences.

		Examples	Neut.	Acc.
A	SRC OUT	I am **pleased** to make my contribution. Sono ***lieto*** di potere contribuire.	`G`	–
B	SRC OUT	Respect for standards lies with **the judges**. ... spetta ***all'autorità giudiziaria***. `[judicial authority]`	`N`	`Acc`
C	SRC OUT	May I quote three **actors** in this field. Posso citare tre ***persone*** `[people]`...	`N`	`Un`
D	SRC OUT	**Commissioner**, I would like to congratulate **the rapporteur**. ***Commissario***, vorrei congratularmi con ***chi ha redatto la relazione***. `[who wrote the report]`	`P`	`S-Acc`

Table 4: Output examples with annotations.

respond to a gendered expression in Italian, then elaborates on the neutralization of each term to provide the final target translation.

(3) `CoT-tgt`: similar to `CoT-src`, but with different steps, i.e. this prompt provides an (intermediate) gendered translation and identifies the *target* terms to be neutralized in the final translation.

Each prompt is used with 3 exemplar sentences taken from the institutional domain, a context where neutral language is increasingly employed, and which is also covered by GeNTE. To verify GPT's ability to generalize from the provided examples, we experiment with two sets of sentences, which only differ for the inclusion of terms to be neutralized that are either *i)* present in GeNTE – hence *seen* – or *ii)* terms that never occur in the test set – hence *not seen*. We list such terms in Table 3, whereas we refer to Appendix B for further details concerning our prompting experiments.

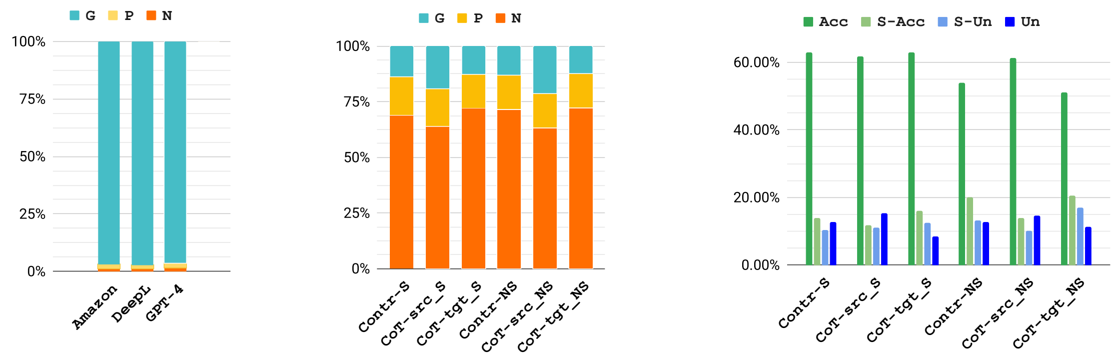
4 Manual Evaluation Results

In this section, we present the results obtained by all our models in BASELINE conditions, and by GPT in GNT-PROMPTING conditions. Although the assesment of GNT capabilites can be automated with the official GeNTE evaluation protocol, the approach would present two inherent limitations. Since the protocol simply classifies whether the whole output translation is gendered or neutral, it does not consider neutralization success/failure for multiple terms in the sentence individually, nor the correctness and acceptability of the corresponding translations.[10] To account for these aspects, we hence resort to a **two-layered manual evaluation** that first distinguishes *i)* fully Neutral (`N`) and *ii)* fully Gendered (`G`), from *iii)* Partially neutral (`P`) outputs where one or more gendered expressions in the sentence are not neutralized. Then, we judge whether the generated GNTs are acceptable (i.e. if they sound fluent and adequately represent the source meaning) on the Likert scale *i)* acceptable (`Acc`), *ii)* somewhat acceptable (`S-Acc`), *iii)* somewhat unacceptable (`S-Un`), *iv)* unacceptable (`Un`).[11] Example judgements are shown in Table 4.

For each model and prompt, we analyze the same 200 randomly selected and anonymized output sentences, equally distributed across three evaluators – all Italian native speakers, highly familiar with

[10] E.g., *I am* ***happy*** → *Sono* ***triste*** ("sad") counts as a – implicitly correct – neutralization, despite its inadequacy.

[11] More information on the manual analysis setup and guidelines is provided in Appendix C.



(a) BASELINE neutrality. (b) GNT-PROMPTING neutrality. (c) GNT-PROMPTING acceptability.

Figure 1: Manual Evaluation Results.[9]

neutral language.[12] While each annotator worked independently, for each system we ensured a 10% of output sentences judged by all raters to verify inter-annotator agreement (IAA).

For the first annotation layer (G,N,P), the Fleiss' kappa on label assignment (Fleiss, 1971) amounts to 0.89, which corresponds to "almost perfect agreement" (Landis and Koch, 1977). Disagreements were all oversights and thus reconciled.

For the acceptability annotations, instead, we measure IAA with the intraclass correlation coefficient (ICC)[13] (Fisher, 1925; Shrout and Fleiss, 1979). In this way, rather than solely focusing on label assignments (i.e. Acc, S-Acc, S-Un, Un) we can account for the actual distance in scores across raters on the 4-point acceptability Likert scale, and thus capture when annotators strongly disagree (e.g. Acc vs. Un) with respect to closer judgements (e.g. Acc vs. S-Acc). The resulting ICC amounts to 0.48. Thus, and as we further discuss in section §4.2, judging acceptability emerges as a more complex and variable task featuring moderate agreement. Notably, the generative nature of the GNT task does not entail a definitive 'correct' answer, and the diverse perspectives can contribute to a range of valid judgments (Popović, 2021; Plank, 2022). To acknowledge such a variability, we did not enforce reconciliation for disagreements.

4.1 BASELINE Results

In Figure 1a, the results achieved by Amazon, DeepL and GPT in the BASELINE condition empirically confirm that, **used *as is*, these models are unsuitable for GNT**. They indeed generate only a discouraging ~3% of neutral translations (both N and P), with a ~97% of the outputs comprising only (mostly masculine) gendered terms. Based on qualitative insights, such sporadic neutralizations largely correspond to (highly probable) literal translations, which incidentally avoid gendered expressions (e.g. src: *we have **addressed***, ref-it: *ci siamo **occupati*** [took care] → out-it: ***abbiamo affrontato*** [have addressed]). The few neutralizations were unsurprisingly considered acceptable by all evaluators, but their negligible amount and sporadic occurrence motivate testing GPT's versatility with dedicated prompts.

4.2 GNT-PROMPTING Results

Starting from the distribution of generated neutralizations, Figure 1b provides the results achieved by GPT *i)* for each prompt template, and *ii)* across the two sets of in-domain exemplars, respectively including gendered terms that occur in GeNTE (S, for *seen*) and terms that are not present in the test set (NS, for *not seen*), for a total of six configurations (§3). A bird's eye view of these scores reveals very promising results. **Across all configurations, GPT produces a notable amount of GNTs** (~65-70% N and ~15% P). Interestingly, despite slightly lower GNT performance for CoT-src,[14] we do not find notable differences across templates for S and NS examples, thus attesting GPT abilities to generalize to newly encountered gendered terms.

By turning to the results in Figure 1c,[15] instead,

[12] They are authors of the paper.

[13] We use the statistical analysis package Pingouin to compute the ICC3 score: https://pingouin-stats.org/build/html/generated/pingouin.intraclass_corr.html.

[14] We hypothesize that the lack of a contrastive gendered translation in the prompt negatively impacts the GNT task.

[15] For the 10% commonly annotated outputs, we include acceptability results by averaging the scores provided by the three evaluators.

[13] For automatic evaluation results, see Appendix D.

the use of NS exemplars seems to slightly reduce the acceptability degree of the generated GNTs. Still, the results are overall positive, with **the best configurations** that **produce over 60% of good quality neutralizations,** like the one in example B in Table 4, which ensures neutrality while fully preserving fluency and adequate source meaning. Notably, we attest a considerable number of somewhat acceptable (S-Acc) / unacceptable (S-Un) GNTs. Indeed, for several instances the raters found that GNT was complex to perform without compromising fluency, up to the point where in ~20-30% of the cases the neutral rephrasings generated by GPT were considered as borderline or not completely satisfactory – as in Table 4 example D, where a "*rapporteur*" is the person in charge of reporting, but not necessarily the one writing a report.

Indeed, the difficulty of judging GNTs is also reflected in the modest IAA measured for acceptability (§4). Examples such as the following one attest to the complexities of determining what makes a good – or *acceptable* – neutralization:

src: Paramilitary groups have stepped up the murders **journalists** and human rights **activists**...

out: I gruppi paramilitari hanno intensificato gli omicidi di **persone che lavorano nel giornalismo**[people working in journalism] e **persone attive nella difesa dei diritti umani**[people active in human right defence]

Two raters judged the GNT as S-ACC and S-UN due to the allegedly awkward repetition of "*people*". Instead, the third evaluator considered the GNT unacceptable due also to adequacy issues (i.e. *working in journalism* does not necessarily imply to be a *journalist*). Overall, we thus recognize different sensitivities with respect to the potential trade-off between adequacy, fluency and the satisfaction of neutral constraints. As such, the qualitative evaluation of **GNT emerges as a subjective task**, even across annotators with comparable expertise in neutral language. This holds implications not only from an evaluation perspective, but also for an effective modeling of future automatic GNT that accounts for such a variability (Kanclerz et al., 2022; Frenda et al., 2023).

5 Conclusions

In response to the rising demand for inclusive language (technologies), this study has focused on the possibilities of automating the generation of gender-neutral translations. In particular, given the limitations of general-purpose MT models due to the need for dedicated parallel data, we have explored the potential of GPT to produce gender-neutral outputs when translating from English into Italian. Through extensive, fine-grained manual analyses, we demonstrated that GPT offers promising avenues, as it can grapple with this complex task when given only a few examples and still generalizes beyond them. Importantly, our evaluations also show that determining the acceptability of what constitutes a good, acceptable neutral translation comes with notable subjectivity. To enable future research, all our manual output annotations are made available [16] to the community to explore the modeling and assessment of such variability.

6 Limitations

Naturally, this work comes with several limitations.

One language pair. Our experiments are carried out for en-it only, and we are thus cautious to indiscriminately generalize our findings. Nonetheless, Italian is a highly representative example of the challenges faced in cross-lingual transfer from English. Accordingly, we believe that our observations can broadly apply to other target grammatical gender languages for high-resource scenarios, too. Crucially, the decision to work on en-it was determined by the fact that – to the best of our knowledge – the bilingual GeNTE corpus (§2) is the only available resource for testing GNT.

Closed-source models. The study relies on different closed-source models. This has reproducibility consequences, since these systems are regularly updated, thus potentially yielding future results that differ from those reported in this paper. As a first attempt to a new, complex task with relevant societal impact such as GNT, we considered reasonable to *i)* focus on general-purpose models used at scale by millions of users *ii)* experiment GNT prompting on the strong GPT model, which as of October 2023 holds the first position on the AlpacaEval leaderboard.[17] In the future, we plan to test open-source models for this task and investigate how to weigh the strengths of MT (i.e. higher translation quality) with those of LLMs (i.e. adaptability to neutral constraints).

Prompts configurations. We tested the use gen-

[16] https://mt.fbk.eu/gente/.

[17] https://tatsu-lab.github.io/alpaca_eval/.

der terms occurring/not occurring in GeNTE for prompt exemplar sentences (§3), so as to investigate GPT's ability to generalize from the given examples. We recognize that a more comprehensive investigation of GPT's generalization ability would advocate for the use of sentence exemplars from varying domains, with more radical structural and stylistic differences. However, for this first exploration we followed existing studies advocating for the choice of demonstrations based on input stylistic and semantic similarity (Zhang et al., 2023; Vilar et al., 2023; Agrawal et al., 2023).

Evaluation. By relying on manual analyses (§4), we enabled a comprehensive GNT evaluation, and overcame the shortcomings of available automated protocols. To provide an alternative method was beyond the scope of this paper, though. Also, although we attest moderate agreement for the GNT acceptability judgments, it should not be regarded as a shortcoming of our evaluation procedure. Rather, on the one hand, it highlights the nuances of judging open-ended generations, for which multiple solutions and subjective perspective are valid (Basile et al., 2021; Rottger et al., 2022). On the other, as newly emerging forms, the perceived acceptability of neutral language is highly dependent on people's attitudes and exposure to such forms, and it is reasonable to expect that they will change over time (Koeser and Sczesny, 2014). Among other aspects, our annotated sentences could also allow to *i)* model this subjectivity, and *ii)* track the acceptability trajectory of GNT in time.

7 Ethics Statement

By investigating the automation of gender-neutral translation, this work has an inherent ethical component. In particular, it is concerned with the impact of translation technologies that reflect exclusionary language, which potentially reinforces stereotypes, masculine visibility, and preclude the representation of non-binary gender identities.[18] Specifically, here we focus on gender-neutralization techniques that rework existing forms and grammars to avoid using needless gendered terminology, and which are endorsed by several institutions (e.g. universities, the EU). These tactics can be viewed as an example of Indirect Non-binary Language (INL) (Attig and López, 2020), which prevent misgendering by eschewing gender assumptions and, as we do in this paper, *equally elicit* all gender identities in language (Strengers et al., 2020). Instead, to *enhance* the visibility of non-binary individuals, Direct Non-binary Language (Attig and López, 2020) resorts to the creation of neologisms, neopronouns, or even neomorphemes (Lauscher et al., 2022). Therefore, many concurring forms can fulfill the demand for inclusive language (Comandini, 2021; Knisely, 2020; Lardelli and Gromann, 2023). It is thus important to emphasize that the neutralizing techniques implemented in our work are not prescriptively intended. Instead, they are orthogonal to other approaches and non-binary expressions for inclusive language (technologies) (Lauscher et al., 2023; Ginel and Theroine, 2022).

[18]We use non-binary as an umbrella term to encompass all identities within and outside the masculine/feminine binary, and that are not represented by binary language expressions.

Acknowledgements

This work is part of the project "Bias Mitigation and Gender Neutralization Techniques for Automatic Translation", which is financially supported by an Amazon Research Award AWS AI grant. Moreover, we acknowledge the support of the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU.

References

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

Bashar Alhafni, Nizar Habash, and Houda Bouamor. 2022. User-centric gender rewriting. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 618–631, Seattle, United States. Association for Computational Linguistics.

Chantal Amrhein, Florian Schottmann, Rico Sennrich, and Samuel Läubli. 2023. Exploiting biased models to de-bias text: A gender-fair rewriting model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4486–4506, Toronto, Canada. Association for Computational Linguistics.

APA. 2020. *Publication Manual of the American Psychological Association*, 7th edition. American Psychological Association.

Remy Attig and Ártemis López. 2020. Queer Community Input in Gender-Inclusive Translations. *Linguistic Society of America [Blog]*.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.

Stephanie Brandl, Ruixiang Cui, and Anders Søgaard. 2022. How conservative are language models? adapting to the introduction of gender-neutral pronouns. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3624–3630, Seattle, United States. Association for Computational Linguistics.

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. Curran Associates, Inc.

Yang Trista Cao and Hal Daumé III. 2020. Toward gender-inclusive coreference resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595.

Won Ik Cho, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. 2019. On Measuring Gender bias in Translation of Gender-neutral Pronouns. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 173–181, Florence, IT. Association for Computational Linguistics.

Prafulla Kumar Choubey, Anna Currey, Prashant Mathur, and Georgiana Dinu. 2021. Improving gender translation accuracy with filtered self-training. *arXiv preprint arXiv:2104.07695*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Gloria Comandini. 2021. Salve a tutt, 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.

Greville G. Corbett. 1991. *Gender*. Cambridge University Press, Cambridge, UK.

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.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A Survey on In-context Learning. ArXiv:2301.00234 [cs].

R.A. Fisher. 1925. *Statistical Methods for Research Workers*. Biological monographs and manuals. Oliver and Boyd.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Simona Frenda, Alessandro Pedrani, Valerio Basile, Soda Marem Lo, Alessandra Teresa Cignarella, Raffaella Panizzon, Cristina Marco, Bianca Scarlini, Viviana Patti, Cristina Bosco, and Davide Bernardi. 2023. EPIC: Multi-perspective annotation of a corpus of irony. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13844–13857, Toronto, Canada. Association for Computational Linguistics.

Ute Gabriel, Pascal M. Gygax, and Elisabeth A. Kuhn. 2018. Neutralising linguistic sexism: Promising but cumbersome? *Group Processes & Intergroup Relations*, 21(5):844–858.

Xavier Garcia and Orhan Firat. 2022. Using natural language prompts for machine translation. ArXiv:2202.11822 [cs].

María Isabel Rivas Ginel and Sarah Theroine. 2022. Neutralising for equality: All-inclusive games machine translation. In *Proceedings of New Trends in Translation and Technology*, pages 125–133. NeTTT.

Frida Höglund and Marie Flinkfeldt. 2023. De-gendering parents: Gender inclusion and standardised language in screen-level bureaucracy. *International Journal of Social Welfare*.

Levi C. R. Hord. 2016. Bucking the linguistic binary: Gender neutral language in English, Swedish, French, and German. *Western Papers in Linguistics/Cahiers linguistiques de Western*, 3(1):4.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*.

Kamil Kanclerz, Marcin Gruza, Konrad Karanowski, Julita Bielaniewicz, Piotr Milkowski, Jan Kocon, and Przemyslaw Kazienko. 2022. What if ground truth is subjective? personalized deep neural hate speech detection. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 37–45, Marseille, France. European Language Resources Association.

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

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the tenth Machine Translation Summit*, pages 79–86, Phuket, TH. AAMT.

Sara Koeser and Sabine Sczesny. 2014. Promoting gender-fair language: The impact of arguments on language use, attitudes, and cognitions. *Journal of Language and Social Psychology*, 33(5):548–560.

J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174.

Jennifer Langston. 2020. New AI tools help writers be more clear, concise and inclusive in Office and across the Web. `https://blogs.microsoft.com/ai/microsoft-365-ai-tools/`. Accessed: 2021-02-25.

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.

Anne Lauscher, Archie Crowley, and Dirk Hovy. 2022. Welcome to the modern world of pronouns: Identity-inclusive natural language processing beyond gender. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1221–1232, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Anne Lauscher, Debora Nozza, Ehm Miltersen, Archie Crowley, and Dirk Hovy. 2023. What about "em"? how commercial machine translation fails to handle (neo-)pronouns. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 377–392, Toronto, Canada. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland. European Association for Machine Translation.

OpenAI. 2023. GPT-4 Technical Report.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. ArXiv:2203.02155 [cs].

Dimitrios Papadimoulis. 2018. *GENDER-NEUTRAL LANGUAGE in the European Parliament*. European Parliament 2018.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards Making the Most of ChatGPT for Machine Translation.

Silvia Alma Piazzolla, Beatrice Savoldi, and Luisa Bentivogli. 2023. Good, but not always fair: An evaluation of gender bias for three commercial machine translation systems. *HERMES - Journal of Language and Communication in Business*, (63):209–225.

Andrea Piergentili, Dennis Fucci, Beatrice Savoldi, Luisa Bentivogli, and Matteo Negri. 2023a. 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, Dennis Fucci, Matteo Negri, and Luisa Bentivogli. 2023b. Hi guys or hi folks? benchmarking gender-neutral machine translation with the GeNTE corpus. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14124–14140, Singapore. Association for Computational Linguistics.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Maja Popović. 2021. Agree to disagree: Analysis of inter-annotator disagreements in human evaluation of machine translation output. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 234–243, Online. Association for Computational Linguistics.

Nathaniel R. Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages.

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Gabriele Sarti, Phu Mon Htut, Xing Niu, Benjamin Hsu, Anna Currey, Georgiana Dinu, and Maria Nadejde. 2023. RAMP: Retrieval and attribute-marking enhanced prompting for attribute-controlled translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1476–1490, Toronto, Canada. Association for Computational Linguistics.

Danielle Saunders, Rosie Sallis, and Bill Byrne. 2020. Neural machine translation doesn't translate gender coreference right unless you make it. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 35–43, Barcelona, Spain (Online). Association for Computational Linguistics.

Jarem Saunders. 2023. Improving automated prediction of English lexical blends through the use of observable linguistic features. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 93–97, Toronto, Canada. Association for Computational Linguistics.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.

Beatrice Savoldi, Marco Gaido, Matteo Negri, and Luisa Bentivogli. 2023. Test suites task: Evaluation of gender fairness in MT with MuST-SHE and INES. In *Proceedings of the Eighth Conference on Machine Translation*, pages 252–262, Singapore. Association for Computational Linguistics.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. Language Models are Multilingual Chain-of-Thought Reasoners. ArXiv:2210.03057 [cs].

Patrick E. Shrout and Joseph L. Fleiss. 1979. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428.

Dagmar Stahlberg, Friederike Braun, Lisa Irmen, and Sabine Sczesny. 2007. Representation of the Sexes in Language. *Social Communication. A Volume in the Series Frontiers of Social Psychology, ed. K. Fiedler (New York, NY: Psychology Press)*, pages 163–187.

David Stap and Ali Araabi. 2023. ChatGPT is not a good indigenous translator. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 163–167, Toronto, Canada. Association for Computational Linguistics.

Yolande Strengers, Lizhen Qu, Qiongkai Xu, and Jarrod Knibbe. 2020. Adhering, steering, and queering: Treatment of gender in natural language generation. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–14, New York, NY, USA. Association for Computing Machinery.

Tony Sun, Kellie Webster, Apu Shah, William Yang Wang, and Melvin Johnson. 2021. They, Them, Theirs: Rewriting with Gender-Neutral English. *arXiv preprint arXiv:2102.06788*.

Eduardo Sánchez, Pierre Andrews, Pontus Stenetorp, Mikel Artetxe, and Marta R. Costa-jussà. 2023. Gender-specific Machine Translation with Large Language Models. ArXiv:2309.03175 [cs].

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.

Leonor Veloso, Luisa Coheur, and Rui Ribeiro. 2023. A rewriting approach for gender inclusivity in Portuguese. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8747–8759, Singapore. Association for Computational Linguistics.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for Translation: Assessing Strategies and Performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427, Toronto, Canada. Association for Computational Linguistics.

Jonas Wagner and Sina Zarrieß. 2022. Do gender neutral affixes naturally reduce gender bias in static word embeddings? In *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*, pages 88–97, Potsdam, Germany. KONVENS 2022 Organizers.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. ArXiv:2201.11903 [cs].

Masaru Yamada. 2023. Optimizing Machine Translation through Prompt Engineering: An Investigation into ChatGPT's Customizability. ArXiv:2308.01391 [cs].

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting Large Language Model for Machine Translation: A Case Study. ArXiv:2301.07069 [cs].

A Test set and GNT

The GeNTE corpus (Piergentili et al., 2023b) represents, to the best of our knowledge, the only available resource for neutral translation into grammatical gender languages and for a variety of gender phenomena. The only other resource being the synthetic dataset by Cho et al. (2019), which only focuses preserving *pronouns* neutrality for English→Korean, namely into a genderless target language (Stahlberg et al., 2007). The dataset INES (Savoldi et al., 2023), instead, focuses on inclusive translation from a grammatical gender language – namely German – into English.

For each of its entry sentences, GeNTE includes aligned *i)* source English, *ii)* gendered reference translation, and *iii)* gender-neutral references translation triplets. The 750 sentences which we are focusing on contain at least one – and potentially several more – source expressions corresponding to Italian gendered terms that require to be either neutralized. Their gendered translations corresponds to the original Europarl references (Koehn, 2005), which propagate the use of masculine generics to refer to generic referents (e.g., en: *It represents a threat to* ***man*** *and animals*→ ref-g: *Rappresenta una minaccia per* ***l'uomo*** *e gli animali*) or assign target masculine forms to unspecified referents (e.g., en: ***All the citizens***→ ref-g: ***Tutti i cittadini***). The neutral translations are created by replacing the gendered expressions and terms with neutral alternatives (e.g. ***essere umano***`[human beings]`, ***tutta la cittadinanza***`[the whole citizenship]`) with different degrees of interventions to ensure *i)* adherence to the source meaning, and *ii)* fluency in the target language, so to avoid perceiving the use of neutral language as intrusive and unsuitable. Accordingly, for each source gender-ambiguous human entity it is ensured that a gender-neutral translation in the target language is feasible.

B Prompts

This section discusses relevant aspects of the prompts used in the experiments and the interaction with GPT-4.

Language. As English emerged as the most effective language for prompting (Shi et al., 2022; Zhang et al., 2023), we use English instructions in our prompts, except for the Italian examples in the task demonstrations.

Task demonstrations. We use 3-shots prompts, which were shown to be a valid compromise between performance and prompt length (i.e. affecting costs and inference time) in our preliminary experiments. The creation of sentence exemplars proceeded as follows:

- The three initial parallel source sentences and the gendered references used in the demonstrations were selected from Europarl's en-it test set, excluding any entry that was already included in GeNTE.
- Source and reference translations were then modified to the include pre-selected *seen* gendered terms, which occur more than 20 times in the used GeNTE subset, and *ii)* the *unseen* terms, which never occur in the used GeNTE subset.
- For such parallel sentences, all gender-neutral translations were produced by one of the evaluators, a linguist experienced with neutral language strategies.
- Finally, the resulting 6 exemplar sentences (shown in Table 5) and their GNTs were approved by all evaluators before proceeding with the experiments.

Length. Table 6 reports the length of each prompt configuration (each template and set of sentence demonstrations) measured per number of tokens. The values were calculated via OpenAI's tokenizer.[19]

Model interaction. We interacted with GPT-4 via the chat completions API. Iterating over the test set, we included the complete content of the prompt and the input source sentence in a single message with the `user` role. The overall cost for the generation of 200 completions for each of the three prompts with both sets of shots was 29.15$.

Post-processing To perfom our manual analysis, we post-process GPT's output so to only extract the final neutral translations to be evaluated.

[19] `https://platform.openai.com/tokenizer.`

	Seen
SRC	Secondly, how far does it increase transparency and accountability **of the MEPs**?
GEND	Secondariamente, fino a che punto aumenta la trasparenza e la responsabilità **dei parlamentari europei**?
NEUT	Secondariamente, fino a che punto aumenta la trasparenza e la responsabilità **dei membri del Parlamento Europeo** [of the members of the European Parliament]?
SRC	**President**, **everyone** must continue to adopt an ambitious approach on these issues.
GEND	**Signora Presidente**, su tali questioni sarà necessario che **tutti** continuino a dare prova d'ambizione.
NEUT	**Presidente** [President], su tali questioni sarà necessario che **ogni persona** [every person] continui a dare prova d'ambizione.
SRC	**Several fishermen** have **joined** with **the politicians** in Belgrade.
GEND	A Belgrado, **molti pescatori** si sono **schierati** dalla parte **dei politici**.
NEUT	A Belgrado, **molte persone che lavorano nella pesca** [many people who work in fishery] hanno **preso le parti** [have taken the side of] di **chi fa politica** [of those who engage in politics].
	Not seen
SRC	Secondly, how far does it increase transparency and accountability **of the writers**?
GEND	Secondariamente, fino a che punto aumenta la trasparenza e la responsabilità **degli scrittori**?
NEUT	Secondariamente, fino a che punto aumenta la trasparenza e la responsabilità **di chi scrive** [of those who write]?
SRC	**HR manager**, **the employees** must continue to adopt an ambitious approach on these issues.
GEND	**Direttore delle risorse umane**, su tali questioni sarà necessario che **gli impiegati** continuino a dare prova d'ambizione.
NEUT	**Responsabile delle risorse umane** [HR manager], su tali questioni sarà necessario che **il personale** [the staff] continui a dare prova d'ambizione.
SRC	**Several freshmen** have **joined** with **the musicians** in Belgrade.
GEND	A Belgrado, **molti studenti del primo anno** si sono **schierati** dalla parte **dei musicisti**.
NEUT	A Belgrado, **molte matricole** [many first-years] hanno **preso le parti** [have taken the side of] **delle persone del mondo della musica** [of the people in the music business].

Table 5: All the <source sentence, gendered translations, and neutral translations> triplets used as demonstrations in both the S and NS sets of examples. Relevant terms for the gendered/neutral comparison are in bold. GNT glosses are available in square brackets.

Prompt	Tokens
Contr_S	294
Contr_NS	304
CoT-src_S	560
CoT-src_NS	568
CoT-tgt_S	743
CoT-tgt_NS	781

Table 6: Number of tokens of for each of the six prompt configurations.

C Manual Analysis

In our analysis, we evaluate the same set of 200 output translations for each models in the BASELINE condition (Amazon, DeepL, GPT) and for each of the six GNT-PROMPTING configurations of GPT (i.e. Contr/CoT-tgt/CoT-src, with both S and NS exemplares). Hence, for a total of 9 generations and 1,800 output sentences. The evaluations were carried based on detailed **guidelines** – created by the same evaluator that designed the prompt examples – which are available with the annotated data release.

Evaluation Design. To annotate the neutrality and acceptability of the outputs sentence, we provided all evaluators with the GeNTE *i)* source English sentences, and the *ii)* gendered reference translations, so to allow them to – respectively – identify which gendered terms had to be neutralized in the output as well as judge the adequacy of the translation with respect to the input sentence. By design, the annotators were tasked to only focus on and judge the portions of the sentence that had to be neutralized, thus disregarding the overall quality of rest of the sentence.[20] To ensure consistency and train the evaluators, we conducted a first round of trial annotations, which allowed to us to address liminal instances and identify blindspots. We have updated the final annotations guidelines accordingly.

[20]To facilitate this task, we *i)* automatically extracted all gendered terms in the Italian references, i.e. only words differing between the gendered and neutral reference in GeNTE, and *ii)* marked them in the sentences provided to the annotators.

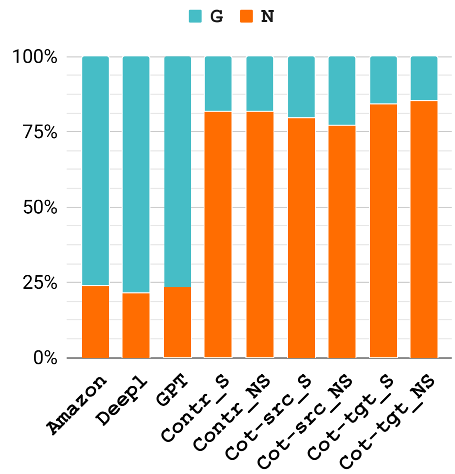


Figure 2: Neutrality for the BASELINE and the GNT-PROMPTING settings evaluated by the classifier.

	Overall	Neutral	Gendered
Amazon	85.35	7.84	86.53
DeepL	86.94	8.70	88.14
GPT-4	86.30	12.00	87.43
Contr_NS	74.65	84.69	49.46
Contr_S	79.30	87.42	61.22
CoT-src_NS	77.55	85.11	64.41
CoT-src_S	79.34	86.81	66.07
CoT-tgt_NS	75.50	87.08	47.62
CoT-tgt_S	79.07	87.90	55.81

Table 7: Percentage agreement (F1 scores) between classifier-based and manual annotation evaluations, with percentages presented for both the overall agreement (weighted F1) and individual class agreements.

D Automatic Evaluation

We report the automatic evaluations results for all models and GPT configurations using the GeNTE evaluation protocol.[21] As displayed in Figure 2, the classifier's scores contrast with the outcomes of our manual analysis. For example, there is a visible disparity in the number of output sentences of the MT systems automatically classified as GNTs. For this reason we exploit our manual analysis contribution to verify the reliability of such an evaluation by calculating *i)* Kendall's Tau (τ) on the GNT system rankings resulting from the classifier and manual analysis,[22] and *ii)* percentage agreement calculated as F1 scores of the classifier on the ground truth labels obtained with the manual evaluation (see Table 7). To ensure a fair assessment of the classifier – which offers a binary classification (Neutral vs Gendered) – we combined the G and P human labels. The τ coefficient yields a positive value of 0.91, indicating that the classifier correlates very well with humans in raking systems based on the amount of generated GNTs. In general, the F1 results vary depending on the system, showing varying levels of satisfaction. F1 scores range from 7.84 for Amazon, where the number of true neutral sentences is notably low (as reflected in the weighted global scores), to 87.90 in the CoT-tgt_S for the neutral class. This calls for future investigation on the performance of the classifier, which is however beyond the scope of this paper.

[21] Classifier v2.0: https://github.com/hlt-mt/fbk-NEUTR-evAL/blob/main/solutions/GeNTE.md.

[22] Calculated with SciPy (https://scipy.org/).