Mind the Inclusivity Gap:
Multilingual Gender-Neutral Translation Evaluation with mGeNTE

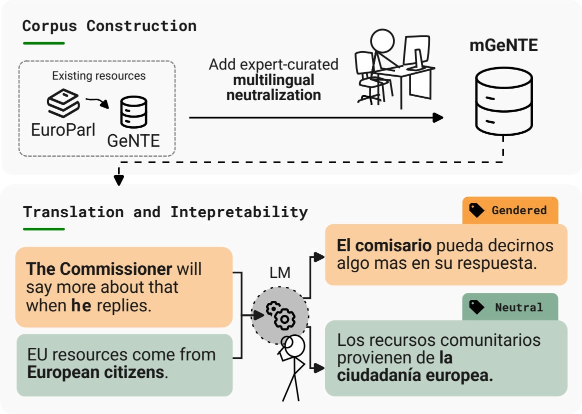
Beatrice Savoldi[◊], **Giuseppe Attanasio**[♠], **Eleonora Cupin**[°], **Eleni Gkovedarou**[•],
Janiça Hackenbuchner[•], **Anne Lauscher**[⋆], **Matteo Negri**[◊],
Andrea Piergentili[◊△], **Manjinder Thind**[°], **Luisa Bentivogli**[◊]

[◊]Fondazione Bruno Kessler, [♠]Instituto de Telecomunicações, [°]DIT Forlì - University of Bologna
[⋆]Data Science Group - University of Hamburg, [△]University of Trento, [•]LT^3 - Ghent University
bsavoldi@fbk.eu

Abstract

Avoiding the propagation of undue (binary) gender inferences and default masculine language remains a key challenge towards inclusive multilingual technologies, particularly when translating into languages with extensive gendered morphology. Gender-neutral translation (GNT) represents a linguistic strategy towards fairer communication across languages. However, research on GNT is limited to a few resources and language pairs. To address this gap, we introduce mGeNTE, an expert-curated resource, and use it to conduct the first systematic multilingual evaluation of inclusive translation with state-of-the-art instruction-following language models (LMs). Experiments on en-es/de/it/el reveal that while models can *recognize* when neutrality is appropriate, they cannot consistently *produce* neutral translations, limiting their usability. To probe this behavior, we enrich our evaluation with interpretability analyses that identify task-relevant features and offer initial insights into the internal dynamics of LM-based GNT.

🤗 datasets/FBK-MT/mGeNTE


Corpus Construction
Existing resources
EuroParl
GeNTE
Add expert-curated multilingual neutralization
mGeNTE
Translation and Intepretability
LM
Gendered
Neutral

The Commissioner will say more about that when **he** replies.

EU resources come from **European citizens**.

El comisario pueda decirnos algo mas en su respuesta.

Los recursos comunitarios provienen de **la ciudadanía europea.**

Figure 1: **Study overview** (English→Spanish example). We create a multilingual benchmark for inclusive MT in English→German, Greek, Italian, and Spanish. We test instruction-following language models on *recognizing* gendered (top) vs. ambiguous source sentences (bottom) and their ability to *produce* correctly gendered and neutral translations. Next, we explain models' behavior with interpretability tools.

1 Introduction

Amid societal and linguistic shifts, inclusive language practices that promote gender equality have gained relevance and use (APA, 2020; Ashwell et al., 2023; Silva and Soares, 2024). Gender-neutral language—as an inclusive strategy also endorsed by international institutions[1]—advances this goal by substituting unnecessary gendered terms with unmarked forms that embrace all gender identities (e.g. *spokesperson* instead of *spokes**man***) (Silveira, 1980; Höglund and Flinkfeldt, 2023).

Towards fairer technologies, prior work in natural language processing (NLP) has explored the use of inclusive language in many tasks (Sun et al., 2021; Hossain et al., 2023; Ovalle et al., 2023; Bartl et al., 2025, *among others*)—including machine translation (MT) (Saunders et al., 2020; Lardelli et al., 2024b). Indeed, translation into grammatical gender languages presents challenges due to extensive gender marking for human referents (e.g. *the citizens*→ es: ***los** ciudadan**os*** M. vs ***las** ciudadan**as*** F.) (Gygax et al., 2019). Ambiguous input often leads MT systems to default to masculine forms, reinforcing stereotypes and marginalizing minority gender groups (Savoldi et al., 2025).

Among other efforts to mitigate exclusionary language in cross-lingual settings (Lauscher et al., 2023; Daems, 2023), gender-neutral translation (GNT) represents a desirable direction, also to

[1] e.g., see the EU Parliament guidelines https://www.europarl.europa.eu/cmsdata/151780/GNL_Guidelines_EN.pdf

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 13698–13720
November 4-9, 2025 ©2025 Association for Computational Linguistics

Benchmark	Languages	Size	Data Type	Ref (Contrastive)	Metric
mGeNTE (ours)	en-it,es,de,el	6,000	natural	✓	✓
GeNTE (Piergentili et al., 2023b)	en-it	1,500	natural	✓	✓
Building Bridges (Lardelli et al., 2024a)	en-de	758	natural	✗	✗
Neo-GATE (Piergentili et al., 2024)	en-it	841	naturalistic	✓	✓
GenderQueer (Friidhriksdóttir, 2024)	en-is	331	naturalistic	✓	✓
INES (Savoldi et al., 2023)	de-en	162	naturalistic	✗	✓
Fair Translate (Jourdan et al., 2025)	en-fr	2,418	template	✓	✗

Table 1: Summary of multilingual resources for gender inclusive MT. We distinguish *natural* data—spontaneously occurring language in authentic contexts—and *naturalistic* data—human-written examples intentionally produced to represent gender-related phenomena–and *template*-based data.

avoid undue binary gender inferences (Piergentili et al., 2023a). In fact, as shown in Figure 1, while preserving gendered forms in the target is justified when they are explicitly marked in the source (orange boxes), GNT offers an *appropriate* alternative when gender in the source is unknown or irrelevant (green boxes). However, progress towards inclusive communication across languages is limited by few existing resources, and by evidence that MT models do not support GNT (Savoldi et al., 2024b).

We address these limitations by introducing **MGeNTE**, a multilingual GNT benchmark, created by extending the language coverage and annotation layers of the existing bilingual GeNTE test set (Piergentili et al., 2023b). With MGeNTE, we examine the potential of instruction-following language models (LMs) in this space and ask: ***Can multilingual LMs enable automatic neutral translation, and crucially, only when appropriate?*** We investigate what factors influence their ability to generate neutral outputs, evaluating five open-weight models of different sizes and families, across four configurations and four language pairs, totaling 80 conditions.

Contributions We lay new groundwork for research on neutral MT with two main contributions. **(1)** A high-quality, natural resource covering English→Italian/Spanish/German/Greek, freely available at https://huggingface.co/datasets/FBK-MT/mGeNTE. **(2)** The first systematic multilingual evaluation of open LMs for GNT. To this end, we pair downstream performance measures with interpretability analyses that shed light on the internal mechanisms supporting LM-based GNT. We release code and artifacts at https://github.com/g8a9/mgente-gap.

Findings We show that—while models *recognize* when neutrality is appropriate across all language pairs—they do not consistently *produce* neutral outputs (§5). Model size and prompt context impact translation outcomes, with larger models better at leveraging context information to improve GNT. The distinction between *recognizing* and *producing* is supported by an interpretability analysis (§6), showing how such tasks rely on different context signals. Our findings inform the current limits in usability of LMs for GNT and point to better integration of these two steps as a potential area for future improvement in inclusive MT.

2 Background

With a growing awareness that language encodes social inequalities and can negatively affect perceptions of gender and identity (Stahlberg et al., 2007; Sczesny et al., 2016), research on gender inclusive language technologies has been gaining traction. In monolingual NLP, this has led to a bulk of studies covering different tasks (Bunzeck and Zarrieß, 2024; Subramonian et al., 2025), with most work centered on English (Cao and Daumé III, 2020; Bartl and Leavy, 2024; Gautam et al., 2024, *inter alia*), and emerging efforts in grammatical gender languages like German (Amrhein et al., 2023; Waldis et al., 2024), Portuguese (Veloso et al., 2023), and Italian (Attanasio et al., 2021; Frenda et al., 2024; Greco et al., 2025).

In the multilingual and translation contexts, however, research remains comparatively scarce. Earlier work has shown that MT systems struggle to preserve neutrality (Cho et al., 2019), especially when translating into grammatical gender languages (Saunders and Byrne, 2020; Piergentili et al., 2023b; Lardelli et al., 2024a). Recently, LMs have been explored to meet the demands of automatic inclusive translation. Yet, the current landscape is fragmented. Individual studies focus on different language pairs, linguistic phenomena, and evaluation approaches, making it difficult to compare findings. Notably, Savoldi et al. (2024b) evaluate GNT for en-it with one *closed* model only, and exclusively on test instances that always require neutralizations—leaving unexplored whether LMs can discern when neutrality is desirable. Jourdan et al. (2025) investigate open LMs capabilities in translating both gendered forms and neomorphemes[2] into French—but their testbed consists of sentences with limited variability and relies on standard MT metrics like COMET (Rei et al., 2021), which are ill-equipped for dedicated analyses and can favor masculine forms (Zaranis et al., 2025).

Overall, a current limitation is the lack of a robust, multilingual benchmark to address cross-lingual inclusivity concerns. As illustrated in Table 1, existing resources are limited—in language coverage, size (Savoldi et al., 2023), naturalness (Jourdan et al., 2025), or lacking dedicated evaluation protocols (Lardelli et al., 2024a). With mGeNTE, we fill this gap. Rather than building our resource from scratch, we enrich the existing GeNTE resource, which is selected for being *i)* the sole natural dataset available, and *ii)* allowing consistent extension to multiple language pairs to enable comparable, multilingual evaluation of GNT. Additionally, we complete our benchmarking effort by pairing mGeNTE with a dedicated evaluation framework for assessing GNT (§4.2).

3 The mGeNTE Corpus

We create mGeNTE by extending the GeNTE benchmark (Piergentili et al., 2023b)—originally designed for English-Italian—to three additional language pairs: English-German, English-Spanish, and English-Greek.[3] Built from naturally occurring data in the Europarl corpus (Koehn, 2005), mGeNTE preserves the original design rationale and curation methodology to ensure comparability. Overall, each language pair comprises 1,500 <*source, target-gendered*> and newly created <*target-neutral*> triplets aligned at the sentence-level, resulting in a total of 6,000 entries.

As represented in Figure 1, mGeNTE entries are evenly distributed across two translation scenarios that allow benchmarking of models' ability to perform neutral translations, but only where appropriate: *i)* Set-N, featuring gender-ambiguous human referents in the source, for which a neutral translation is desirable in the target; and *ii)* Set-G, featuring explicit gender mentions in the source that should be correctly rendered with gendered (masculine or feminine) forms in the target.

	Segments		**Gendered Words**	
mGeNTE	**Set-G**	**Set-N**	**#**	**#Unique**
en-it	750	750	4115	802
en-es	750	750	4363	644
en-de	750	750	3977	613
en-el	750	750	3736	743
Parallel	578	409	–	–

Table 2: Distribution of mGeNTE segments by subset and language pair, sentences that are fully parallel across all pairs (Parallel), and total (#) and unique (#Unique) annotated gendered words per language.

To ensure *high-quality*, we entrust data extraction (§3.1) and annotation (§3.2) to experts and translation students specialized in GNT with native or C1 competence in their assigned target language.[4] Creating new neutral references in the target is tasked to professional translators (§3.3).

Final statistics of mGeNTE segments and annotations are given in Table 2.

3.1 Multilingual Expansion and Alignment

Parallel data extraction We prioritize selecting sentences aligned across en-it/es/de/el. To this aim, we started by retrieving the Europarl[5] sentences that are already contained in GeNTE en-it. Each automatically extracted sentence[6] was checked to confirm that it contained a gender-related phenomenon; if not, we discarded it.[7] This process yielded 987 sentences that are fully parallel across all four pairs, henceforth referred to as Parallel-Set.

Language-specific extraction To reach the target of 1,500 sentences per language pair, the remaining sentences were extracted *ex novo* using regular expressions. These were chosen to represent both subsets (Set-N or Set-G) as well as

[2] i.e., innovative linguistic solutions. See also §7.

[3] These target languages were selected because, as grammatical gender languages, they have rich gendered morphology, making them suitable for GNT. In addition to three widely studied Indo-European languages from distinct branches (Italian, Spanish, German), we deliberately included Greek—a lower-resource language with a distinct script—to broaden the linguistic diversity in mGeNTE and to ensure the coverage of underrepresented, challenging cases in inclusive MT research.

[4] They are all authors of this paper.

[5] Specifically, from the `common test set 2`, which contains data aligned across multiple languages: https://www.statmt.org/europarl/archives.html

[6] Using the original Europarl IDs available in GeNTE.

[7] This is possible due to cross-linguistic differences in how different languages encode and express genders. For instance, "child" is gender-marked in Italian (***il/la*** *bambin**o/a***) but remains invariant in German (*das Kind*).

SET-N	SRC	**Pensioners** are in favour of strengthening criminal law, [...]
en-it	REF-G REF-N$_1$	**I pensionati** sono favorevoli a un rafforzamento del diritto penale, [...] **Le persone pensionate**[pensioned people] sono favorevoli a un rafforzamento del diritto penale, [...]
en-es	REF-G REF-N$_1$	**Los pensionistas** están a favor de reforzar el Derecho penal no solo nacional, [...] **Hay pensionistas**[there are pensioners] que están a favor de reforzar el Derecho penal no solo nacional, [...]
en-de	REF-G REF-N$_1$	Die **Rentner** begrüßen den Ausbau nicht nur des einzelstaatlichen, [...] **Die Menschen in Rente**[people in retirement] begrüßen den Ausbau nicht nur des einzelstaatlichen, [...]
en-el	REF-G REF-N$_1$	Οι **συνταξιούχοι** είναι υπέρ της ενίσχυσης του ποινικού δικαίου, [...] **Τα συνταξιοδοτημένα άτομα**[the retired individuals] είναι υπέρ της ενίσχυσης του ποινικού δικαίου, [...]
SET-G	SRC	I trust the **Commissioner** will promise that he will exercise extra vigilance.
en-it	REF-G REF-N$_1$	Spero che **il Commissario** ora prometta di vigilare attentamente a tale riguardo. Spero che **il membro della Commissione**[the member of the board] ora prometta di vigilare attentamente a tale riguardo.
en-es	REF-G REF-N$_1$	Espero que **el Comisario** prometa controlar exhaustivamente esta situación. Espero que **la representación de la Comisión**[the representative of the board] prometa...
en-de	REF-G REF-N$_1$	Von **dem Herrn Kommissar** erwarte ich heute die Zusage, **er** werde mit Argusaugen darüber wachen. **Von dem Kommissionsmitglied**[From the board member] erwarte ich heute die Zusage, **es**[they] werde mit Argusaugen...
en-el	REF-G REF-N$_1$	Προσδοκώ από **τον** Επίτροπο να δεσμευτεί ότι θα επιβλέψει αυστηρά την κατάσταση. Προσδοκώ από **το μέλος της Επιτροπής**[the member of the Commission] να δεσμευτεί ότι θα επιβλέψει [...].

Table 3: MGENTE PARALLEL. Entries from SET-N and SET-G, with gendered references (REF-G) and REF-N with neutralizations. Words in **bold** mention human referents; underlined source words express the referent's gender.

language-specific gender patterns. Accordingly, for SET-G we targeted unambiguous English source sentences, i.e. containing explicit gender cues such as titles (*Mr, Mrs*) and pronouns (*him, her*). Vice versa, for SET-N we excluded segments with source gender information and matched expressions that—while gender ambiguous in English—correspond to either masculine or feminine forms in the target language (e.g., *deputy* → es: *deputa**do/a***, de: *Stellvertret**er/erin***, el: *αναπληρω**τής/τρια***).

Sentence Alignment To align with the original GeNTE corpus and streamline evaluation, source sentences containing multiple referents requiring different gender forms were edited to consistently require a single form in the target language.[8] Additionally, to address the under-representation of unambiguous feminine data in SET-G, the corpus was adjusted through gender-swapping interventions to achieve a balanced distribution of feminine and masculine forms.[9] Minor corrections (e.g., typos, translation errors) were also applied. Full details of the editing process and intervention statistics are provided in Appendix A. At the end of the editing process, all sentences were manually sorted into SET-N, or as either feminine (F) or masculine (M) instances from SET-G. Such a distinction is central to MGENTE design and explored in the following analyses.

3.2 Gendered Words Annotation

To further enrich MGENTE, all target sentences—including the original en-it ones—were manually annotated at the word level [10] to identify gendered words requiring neutralization, enabling the exploration of cross-linguistic variations in gender phenomena. As shown in Table 2, the corpus displays substantial lexical variability, with over 600 unique gendered words per language. Notably, Italian and Greek exhibit a higher unique count of gendered words, reflecting their morphosyntactic tendency to mark gender not only on nouns and adjectives but also on verbal forms. Also, based on qualitative observations, SET-N annotated words are vastly populated with masculine forms used generically to refer to mixed or unknown referents. These cases, though translated with gendered (masculine) forms in the original Europarl references, are exactly the target of language neutrality efforts (Gygax et al., 2021) and well-suited candidates for GNT. For more insights and details on the annotation process, see Appendix A.

3.3 Gender-Neutral Reference Creation

For each German, Spanish and Greek Europarl (gendered) reference translation— henceforth REF-G—we created an additional gender-neutral reference (i.e. REF-N), which differs from the original one only in that it refers to human entities with neutral expressions. Neutralization is an open-ended

[8] In this way, each entry can be treated as a coherent gendered or neutral unit.

[9] Masculine forms are over-represented in the original target references, reflecting a well-known under-representation of women and feminine forms in existing resources (Vanmassenhove et al., 2018; Gaido et al., 2020).

[10] We compute inter-annotator agreement (IAA) on the exact matches of the gendered words annotated by two annotators. IAA (Dice coefficient) was ≥ 0.92. For further details and a breakdown by language, see Appendix A.

task with a high degree of variability in its possible solutions. In our target languages with rich gendered morphology, this can involve a range of strategies (Papadopoulos, 2022; di Carlo, 2024; Müller-Spitzer et al., 2024)—from local lexical changes (e.g. epicene synonyms, collective nouns) to more extensive rephrasings (e.g. impersonal or passive constructions) to be selected contextually to preserve the adequacy of the neutral translations.

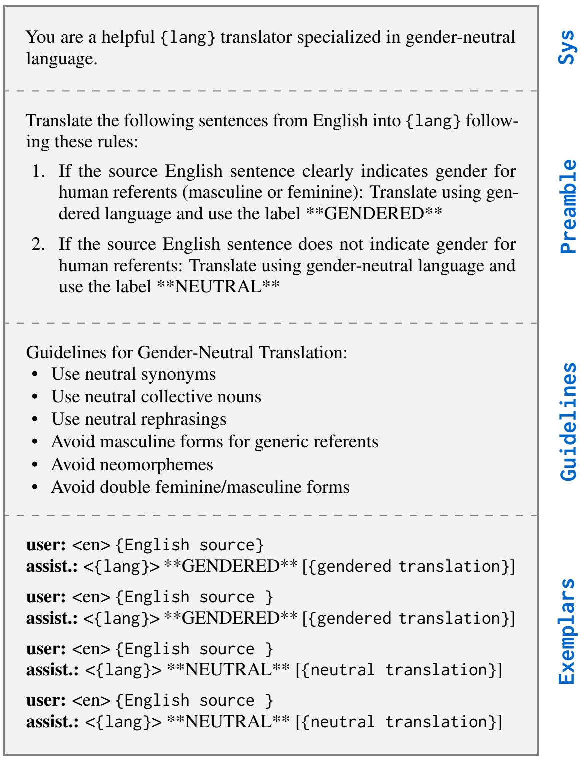
We ensured high-quality, diversified and contextually appropriate neutralizations by commissioning three professional translators per language pair—selected for their expertise in neutral language—and informed by detailed language-specific neutralization guidelines.[11]

Common Set Following the GeNTE design, we additionally created a common of 200 sentences (100 from each Set) to be neutralized by all translators. We thus obtain three Ref-Ns per source sentence. The common set, besides adding a richer dimension to the corpus, allows us to ask: *how much variability do translators yield when neutralizing the same sentence?* Our analysis reveals considerable variation across language pairs, with identical neutralizations occurring in only 11% (en-es), 9.3% (en-de), and 14.9% (en-el)[12] of cases. This variability is directly reflected in the corpus, attesting to its natural diversity, an asset for studying and evaluating open-ended tasks like GNT.

We show a full entry from the common set in Table 7 Appendix A, and refer to Table 3 for a Parallel entry example across all language pairs. Generally, in ambiguous Set-N entries (top), translators avoid masculine generics through rephrasing strategies, whereas neutralization of gender-unambiguous sentences Set-G can be more verbose and unnecessary.

4 Experimental Setup

We conceptualize the GNT task as a model's ability to produce correctly gendered translations when the source specifies gender (i.e. Set-G), and neutral translations when the source refers to unspecified referents (i.e. Set-N). In our setup—to disentangle source sentence categorization from generation performance—we enforce models to output both (*i*) a *label* indicating the source category, as well as (*ii*) the *translation* of the source sentence.


Sys
Preamble
Guidelines
Exemplars

You are a helpful {lang} translator specialized in gender-neutral language.

Translate the following sentences from English into {lang} following these rules:

1. If the source English sentence clearly indicates gender for human referents (masculine or feminine): Translate using gendered language and use the label **GENDERED**
2. If the source English sentence does not indicate gender for human referents: Translate using gender-neutral language and use the label **NEUTRAL**

Guidelines for Gender-Neutral Translation:

- Use neutral synonyms
- Use neutral collective nouns
- Use neutral rephrasings
- Avoid masculine forms for generic referents
- Avoid neomorphemes
- Avoid double feminine/masculine forms

user: <en> {English source}
assist.: <{lang}> **GENDERED** [{gendered translation}]

user: <en> {English source }
assist.: <{lang}> **GENDERED** [{gendered translation}]

user: <en> {English source }
assist.: <{lang}> **NEUTRAL** [{neutral translation}]

user: <en> {English source }
assist.: <{lang}> **NEUTRAL** [{neutral translation}]

Figure 2: GNT prompt overview with labeled sections. The prompt consists of system instructions, translation rules (preamble), gender-neutral guidelines, and exemplars provided as conversational turns.

4.1 Models setup

Models We experiment with open-weight multilingual models. Starting from an initial pool of 10 SOTA models[13] from 5 families (Qwen, LLama, Mistral, Gemma, Phi), we used translation quality and format adherence as thresholds for inclusion in the main experiments (details in Appendix B.3). Based on these criteria, we retain five final instruction models of varying sizes: Llama 3.1 8B and Llama 3.3 70B (Grattafiori et al., 2024), Qwen 2.5 72B (Qwen, 2024), Gemma 2 9B (Team et al., 2024), and Phi 4 14B (Abdin et al., 2024). Experimental details are in Appendix B.1.

Setup We test models' in-contex-learning capabilities (Brown et al., 2020) for the GNT task. Our prompt (shown in Figure 2) includes a system prompt (*Sys*), the task description (*Preamble*), language-specific GNT *guidelines*,[14] and four task demonstrations (2 gendered from Set-G, 2 neutral

[11] The guidelines are available with the data release. Translators were compensated at market rates: €25/hr for en-es/el and €35/hr for en-de.

[12] For the original en-it it amounts to 13.5%.

[13] As of March 2024 on Open LLM Leaderboard.

[14] Depending on the target, different linguistic examples are provided. Full prompts available in our repository.

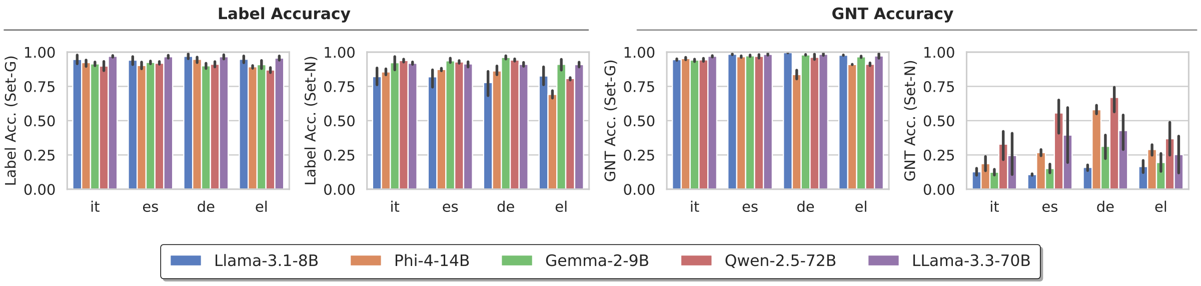


Figure 3: Source category (*left*) and GNT accuracy (*right*) results across MGENTE Sets (avg. across prompts).

from SET-N) randomly sampled from MGENTE PARALLEL and excluded from evaluation. To analyze robustness and the impact of context components, we test four different configurations: including both the system prompt and the guidelines, excluding one of them, or excluding both— totaling 4×5 model configurations per language pair.[15]

4.2 Evaluation

Overall quality We assess the selected models' translation quality using xCOMET (Guerreiro et al., 2024),[16] scoring outputs against their respective correct references (REF-G for SET-G, REF-N for SET-N). The models achieve high average scores (en-de 0.96, en-es 0.95, en-it 0.95), with Greek—being a less supported language—showing comparatively lower average performance (0.83).

Source sentence category As a first inclusivity-related metric, we measure the *accuracy* of label generation (either *Gendered* or *Neutral*) against the gold SET-G/N annotations in MGENTE.

Gender-Neutral Translation We evaluate the *accuracy* of models in producing correctly gendered and neutral translations by using an *LLM-as-a-judge* approach (Gu et al., 2025), which enables scalable GNT evaluation across multiple languages. In practice, we adapt the structured approach proposed by Piergentili et al. (2025), which provides sentence-level neutrality binary judgments and was tested on human-written gendered vs. neutral text.[17] We rely on their optimal prompt[18] by adapting it to all languages covered in MGENTE, and validate its effectiveness on automatic translations. We tested different LLMs on 1,000 manually annotated model outputs. Our best-performing evaluation setup relies on GPT-4O,[19] which achieves 0.87 macro F1 and 92% accuracy. Full results are in Table 5 (Appendix B.4), along with prompt and data annotation details.

5 Gender-Neutrality Results

We present GNT results on the MGENTE benchmark. Full multilingual results for both label and translation generation are in Figure 3.

LMs effectively distinguish ambiguous from gendered source sentences. Source category scores in Figure 3 show strong, consistent performance across languages, models, and with minor variance in prompt configurations (see error bars). Overall, *gendered* accuracy is only slightly higher.[20]

GNT is challenging, with variations across models and languages. Figure 3 shows that while SET-G sentences are consistently translated with the correct gendered forms, accuracy on SET-N is systematically lower and characterized by higher variance (see error bars), with difficulty in producing neutral translations. Results vary by language: en-el/it achieves the lowest rates of correct GNTs. Greek's results align with its lower overall generic performance (§4.2), but en-it's underperformance is more surprising given its otherwise solid overall translation quality. Broadly, larger models outperform smaller ones, with QWEN 2.5 72B leading overall, followed by LLAMA 3.3 70B (en-es, en-it) and PHI 4 14B (en-de, en-el).

Correct source categorization does not guarantee correct GNT. Given the mismatch between source categorization and GNT performance, in Figure 4 we measure how consistently the output

[15] We always provide the task definition and the four shots, though our preliminary experiments in Appendix B.1 also include zero-shot and 2-shot settings.

[16] https://huggingface.co/Unbabel/XCOMET-XL

[17] The original GeNTE evaluation method is limited to en - it, and is superseded by the LLM-as-a-judge approach by Piergentili et al. (2025).

[18] i.e. Cross+P+L in the original paper.

[19] gpt-4o-2024-08-06

[20] Worst overall scores are by PHI 4 14B for el Set-N < 75. Complete results are available in Table 6 in Appendix C.

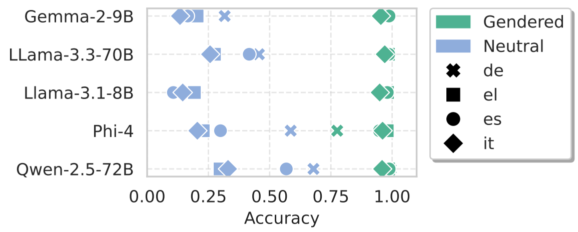


Figure 4: **Accuracy of label-translation coherence**. Measures the agreement of translation forms (gendered/neutral) with the generated label. Scores are averaged across prompt configurations.

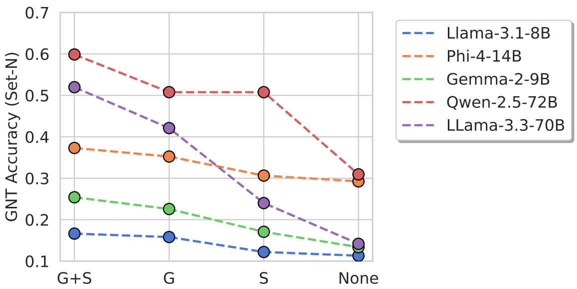


Figure 5: **GNT Accuracy (Set-N) across prompt configurations**, i.e. using both context parts *system* and *guidelines* (G+S), only one of them, or *None*. Markers show values averaged over all language pairs.

translation form (gendered/neutral) and the generated label are in agreement. These results highlight a gap between recognizing the source sentence category and generating the appropriate translation. Models systematically produce gendered translations when assigning a gendered label, but coherence drops sharply for neutral cases—often below random chance. Only QWEN 2.5 72B and PHI 4 14B maintain higher agreement in their strongest language pairs (en-es/de and en-de, respectively).

(Larger) LMs benefit from richer context. Figure 5 breaks down GNT results by prompt configuration to explore the impact of provided context.[21] Including both *guidelines* and *system* (G+S) yields higher GNT accuracy, while removing both (*None*) leads to the worst performance. Larger models show higher gains from rich prompts, better leveraging in-context information. Notably, while QWEN 2.5 72B maintains relatively higher GNT performance across setups, LLAMA 3.3 70B drops to small-model levels in the *None* setup. For individual language pairs results, see Figure 13 in Appendix C.1.

🔍 Overall, our findings highlight that (*i*) though LMs reliably detect when gender neutrality is needed, they do not consistently produce neutral translations, with GNT capabilities notably less robust. GNT (*ii*) varies across languages;[22] (*iii*) depends on model choice—with larger models generally performing better—and (*iv*) is sensitive to prompt variations—indicating that contextual information in prompts can affect real-world usability of LMs towards inclusive translation.

To better understand this behavior, we use for the first time context attribution and explainability techniques to shed light on how LMs handle gender-neutral translation.

6 Context Analysis

Section 5 established that source category detection and GNT translation differ in performance, and the prompt has a decisive impact on GNT. Hence, it is of utmost interest to measure *when* and *how* models use contextual information when carrying out GNT. This can help explain observed mismatches, guide future development, and support immediate use of LMs by enabling prompt steering for inclusivity.

Inspired by recent long-context attribution work (Sarti et al., 2024; Cohen-Wang et al., 2024; Liu et al., 2024), we hence resolve to post-hoc interpretability (Madsen et al., 2022) for a finer-grained analysis. We are interested in assessing the contribution of each input token to generating a specific output. This type of task is commonly known as *feature attribution*, and several methods have been proposed for NLP task (Mosca et al., 2022; Ferrando et al., 2022, among others), including MT (Zaranis et al., 2024). Prior work has leveraged interpretability for gender bias in MT (Sarti et al., 2023; Attanasio et al., 2023), focusing on binary gender and source token (i.e., pronouns) contributions. Our study extends them to multilingual GNT and context parts.

6.1 Attribution Setup

We use Attention-Aware Layer-Wise Relevance Propagation (Achtibat et al., 2024, AttnLRP), a leading method reporting strong faithfulness and plausibility. Given an input and model output, AttnLRP attributes a numerical score to each input

[21] Focusing on SET-N GNT. Full disaggregated results across Sets are in Appendix C.1.

[22] Potentially reflecting both data availability and sociolinguistic factors—the higher performance for en-de may relate to greater distribution and progress in language inclusivity for German; we leave this for future research.

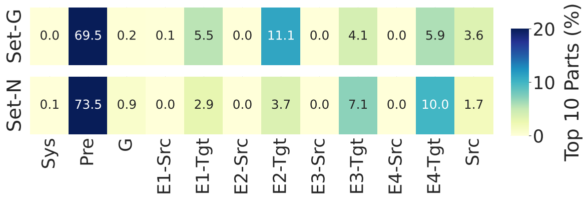


Figure 6: **Relevance of context parts to: Source Label**. Ratio of occurrence within the top 10 scores. Results for gendered (top) and neutral (bottom) source sentences.

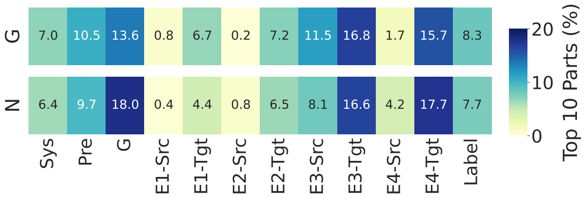


Figure 7: **Relevance of context parts to: Translation**. Ratio of occurrence within the top 10 scores. Results for ambiguous sentences with wrongly gendered (G) and correctly neutral (N) translations.

token defined as $\mathrm{s}(i, a : b) \in \mathbb{R}$ where i is the positional index of the input token and a, b are the (inclusive) boundaries of a span in the output (full derivation in Appendix B.5). This quantity indicates the *contribution of a given token for generating all the tokens within* $[a, b]$.

We compute two sets of token contributions to explain different parts of the output: 1) $\mathbf{S_L}$ collects the contributions for generating the source category label 2) $\mathbf{S_T}$ collects the scores explaining the actual translation—including the tokens of the label itself as it is always generated before the translation. We exclude all chat template special tokens (e.g., `<|im_start|>`) and the source tokens from $\mathrm{S_T}$—expectedly, their contribution is always the strongest, preventing a meaningful interpretation of the other context parts. For each part (the system prompt Sys, preamble Pre, guidelines G, four task exemplars E, the source sentence being translated (for $\mathrm{S_L}$) Src, and source label Label when explaining $\mathrm{S_T}$) we collect and aggregate token-level scores (see Appendix B.5 for full details and Appendix C.2 for complementary results).

Data We focus on QWEN 2.5 72B with a full prompt (G+S), the configuration that yielded the best GNT results (§5).[23] We compute contributions on 4,000 outputs across four language pairs and both MGENTE sets. To prioritize soundness and avoid potential errors from the automatic LLM-as-judge evaluation, $\mathbf{S_T}$ contributions are computed on output translations that are manually evaluated as gendered or neutral. All data and annotation details are provided in Appendix B.5. Given that the number of correct label predictions and GNT outputs differs across languages (see Figure 3), overall results in the following section are averaged across languages for equal representation.[24]

6.2 Findings

Does QWEN 2.5 72B use the context in a similar way to *detect* and *translate* gender neutrality? No, it does not. To detect the source category, it mostly relies on the preamble Pre (Figure 6). In contrast, contributions in $\mathrm{S_T}$ are more heterogeneous (Figure 7), with major contributions from the guidelines (G) and the assistant's neutral exemplars (E3/4-Tgt). Asymmetry in context use can be explained by the different nature of the task, i.e., binary classification vs. open-ended generation.

While *detecting* the source category, which context part contributes the most relevant tokens? The preamble (task definition) Pre, and by a large margin (Figure 6). The Figure shows how often the preamble is among the top 10 relevant tokens. Given that the preamble is present across all prompt configurations, this explains the consistently high performance on the source category task. When observing the other top contributors, we see that gendered exemplars (E1, E2) weigh more in SET-G, and neutral ones in SET-N (E3, E4). A closer look at the label generations for SET-N reveals a somewhat surprising result. Despite being accurate, QWEN 2.5 72B does not use the input Src nor the source of the shot examples (E3/4-Src) when predicting the label. Src contribution is marginally higher in SET-G, where explicit gender cues such as *he*, *she*, or *her* appear frequently among top contributors. This finding suggests a lexical overfitting phenomenon: the model detects a gendered category when source gender cues are present, and assumes neutrality when they are absent.[25]

What drives QWEN 2.5 72B's neutral translations of ambiguous source sentences? The assistant's part in neutral exemplars E3/4-Tgt, gender neutral guidelines G, and preamble Pre

[23] We focus on one model for resource availability; the strongest one to ensure a sufficient number of GNT outputs.

[24] For a breakdown by language, see Figure 18 and 19 in Appendix C.2.

[25] By design, SET-N has no explicit source gender cues since neutrality is defined precisely by their absence.

(Figure 7). The Figure shows how often these are among the top 10 relevant tokens, comparing wrongly gendered and correctly neutral translations from Set-N. Crucially, the source label (Label—always *Neutral*) has a similar contribution regardless of whether the output continued with a gendered or neutral translation. This finding suggests that, in our setup, the previously generated label *Neutral* (Label) is not discriminative of the produced translation, whether gendered or neutral. This opens up to future inquiries on conditioning mechanisms to link label predictions with translation behavior towards improving GNT.

7 Conclusion

We presented mGeNTE, the only existing multilingual benchmark for gender-neutral translation, covering English→Italian, Spanish, German, and Greek. This expert-curated resource provides rich annotations, contrastive references, and diverse neutralization strategies to support inclusive MT research. Using mGeNTE, we carried out the first systematic multilingual evaluation of open LMs for gender-neutral translation. Our findings reveal language gaps, the influence of model size, and prompt context, with interpretability analyses explaining the capabilities of LMs for this task as well as opportunities for improvement. Moving forward, we aim to leverage mGeNTE's richness to advance fairer, more inclusive NLP and we release it publicly to support the broader research community.

Limitations

While mGeNTE represents a significant step forward in evaluating gender-neutral machine translation, we acknowledge four limitations that can be addressed in future work.

First, the findings presented in this paper are based on state-of-the-art models using both a base prompt (i.e., configuration *None*, with only task definition and shots) that is incrementally made richer in information based on established prompting approaches (e.g. adding a *system prompt*). However, it is likely that our results will not generalize to *every* possible prompt formulation or model configuration for translation systems. Alternative model architectures or novel prompts may yield different performance patterns that are not captured in our current framework.

Second, mGeNTE employs a sentence-level design that, while effective and still quite widespread in the field, represents a simplification of real-world translation scenarios. This approach does not account for potential extra-sentential or long-range dependencies (e.g., discourse-level phenomena such as coreference resolution across multiple sentences) that may influence gender expression in translation.

Third, our best, LM-based GNT evaluation metric relies on a closed, commercial language model. While this model represents the current state-of-the-art, its proprietary nature may hinder reproducibility. Moreover, as the underlying model receives updates over time or is discontinued, our metric itself may evolve, potentially complicating longitudinal comparisons. Open-weight alternatives offer a promising, more transparent avenue for future research. We provide one of such in Appendix B.4. Namely, our second-best *LM evaluator* setup utilizes Qwen 2.5 72B as a judge, achieving competitive results (approximately 7 points behind GPT-4o).

Fourth, our interpretability analysis is conducted at a relatively coarse level, e.g., by computing attributions over the entire output translation. However, gender-translation phenomena typically manifest and regard only a portion of the overall translation output. As a result, computing attributions over the entire translation may dilute the precision of the interpretability analysis in relation to gender-specific effects. The reason for this approach, however, is that gender-neutralization often involves multiple, non-consecutive words within a sentence, making it conceptually challenging to define and isolate the relevant spans for targeted analysis. We intend to address this limitation in future work by leveraging the contrastive references provided by mGeNTE to enable more fine-grained, span-level attribution methods.

Ethics Statement

This paper concerns gender-neutral language and translation, inherently addressing ethical considerations. Specifically, it is intended to tackle language technologies that can generate service disparities and perpetuate exclusionary language (Savoldi et al., 2024a; Ungless et al., 2025), thereby reinforcing stereotypes, promoting masculine dominance, and neglecting the representation of non-binary gender identities.[26] Our focus is on gender-neutralization strategies that modify existing forms and grammatical structures to eliminate unnecessary gendered language. These methods aim to avoid assumptions about gender and *equally represent* all gender identities in language (Strengers et al., 2020). By contrast, Direct Non-binary Language (López, 2020) seeks to *increase* the visibility of non-binary individuals by introducing new linguistic elements such as neologisms, neopronouns, or even neomorphemes (Lauscher et al., 2022; Ginel and Theroine, 2022; Piergentili et al., 2024).

A range of strategies can be employed to meet the demand for inclusive language (Scharrón-del Río and Aja, 2020; Comandini, 2021; Knisely, 2020). It is crucial to note that the neutralization methods used in this research are not intended to be prescriptive. Instead, they complement other approaches and expressions for achieving inclusivity in language technologies.

Acknowledgements

We thank Beatriz Riva, Begoña Martínez, Jens Wargenau, Pedro Cano Navarro, Sylke Denfeld, Zografia Kefalopoulou, Eleni Vlachou, and Matthew Louis Papapoulios for their essential translation work in creating the neutral references of the mGeNTE corpus. The translation work for mGeNTE en-el was funded by the Department of Translation, Interpreting and Communication of Ghent University. We also thank Verdiana Crociani for their invaluable help in the annotation and IAA computation for mGeNTE en-de, as well as Maria Tsaousidou and Stella Tatalia for their contribution to the annotation of mGeNTE en-el.

Beatrice Savoldi is supported by the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU. Matteo Negri and Luisa Bentivogli are supported by the Horizon Europe research and innovation programme, under grant agreement No 101135798, project Meetween (My Personal AI Mediator for Virtual MEETtings BetWEEN People). Eleonora Cupin and Manjinder Thind's carried out their work during internships at FBK in the spring and summer of 2024, and Eleni Gkovedarou during internship at LT^3 - Ghent University in the summer and winter of 2024/5. Janiça Hackenbuchner was funded by The Research Foundation – Flanders (FWO), research project 1SH5V24N (from 01.11.2023 until 31.10.2027). Anne Lauscher's work is funded under the excellence strategy of the German federal government and the states. Giuseppe Attanasio is supported by the Portuguese Recovery and Resilience Plan through projects C645008882-00000055 (Center for Responsible AI) and UID/50008: Instituto de Telecomunicações.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, et al. 2024. Phi-4 technical report.

Reduan Achtibat, Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Aakriti Jain, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. 2024. AttnLRP: Attention-aware layer-wise relevance propagation for transformers. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 135–168. PMLR.

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.

Leila Arras, Bruno Puri, Patrick Kahardipraja, Sebastian Lapuschkin, and Wojciech Samek. 2025. A close look at decomposition-based xai-methods for transformer language models. *arXiv preprint arXiv:2502.15886*.

Sabrina J. Ashwell, Patricia K. Baskin, Stacy L. Christiansen, Sara A. DiBari, and Annette Flanagin. 2023. Three recommended inclusive language guidelines for scholarly publishing: Words matter. *Learned Publishing*, 36(1):94–99.

Giuseppe Attanasio, Salvatore Greco, Moreno La Quatra, Luca Cagliero, Michela Tonti, Tania Cerquitelli, and Rachele Raus. 2021. E-mimic: Empowering multilingual inclusive communication. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 4227–4234. IEEE.

Giuseppe Attanasio, Flor Miriam Plaza del Arco, Debora Nozza, and Anne Lauscher. 2023. A tale of pronouns: Interpretability informs gender bias mitigation for fairer instruction-tuned machine translation.

[26] The term "non-binary" is used here as an inclusive umbrella term to represent identities that exist both within and beyond the masculine/feminine binary, and are not conveyed through binary linguistic expressions.

In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3996–4014, Singapore. Association for Computational Linguistics.

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

Marion Bartl, Thomas Brendan Murphy, and Susan Leavy. 2025. Adapting psycholinguistic research for llms: Gender-inclusive language in a coreference context. *arXiv preprint arXiv:2502.13120*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Bastian Bunzeck and Sina Zarrieß. 2024. The SlayQA benchmark of social reasoning: testing gender-inclusive generalization with neopronouns. In *Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP*, pages 42–53, Miami, Florida, USA. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

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.

Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Madry. 2024. Contextcite: Attributing model generation to context. *Advances in Neural Information Processing Systems*, 37:95764–95807.

Gloria Comandini. 2021. Salve a tuttə, tutt*, tuttu, tuttx e tutt@: l'uso delle strategie di neutralizzazione di genere nella comunità queer online. *Testo e Senso*, 23:43–64.

Joke Daems. 2023. Gender-inclusive translation for a gender-inclusive sport: strategies and translator perceptions at the international quadball association. In *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 37–47, Tampere, Finland. European Association for Machine Translation.

Giuseppina Scotto di Carlo. 2024. 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 edition, page 21. Routledge.

Lee R. Dice. 1945. Measures of the Amount of Ecologic Association Between Species. *Ecology*, 26(3):297–302.

Javier Ferrando, Gerard I. Gállego, Belen Alastruey, Carlos Escolano, and Marta R. Costa-jussà. 2022. Towards opening the black box of neural machine translation: Source and target interpretations of the transformer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8756–8769, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Simona Frenda, Andrea Piergentili, Beatrice Savoldi, Marco Madeddu, Martina Rosola, Silvia Casola, Chiara Ferrando, Viviana Patti, Matteo Negri, Luisa Bentivogli, et al. 2024. Gfg-gender-fair generation: A calamita challenge. In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*.

Steinunn Rut Friidhriksdóttir. 2024. The GenderQueer test suite. In *Proceedings of the Ninth Conference on Machine Translation*, pages 327–340, Miami, Florida, USA. Association for Computational Linguistics.

Marco Gaido, Beatrice Savoldi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2020. Breeding gender-aware direct speech translation systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3951–3964, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Vagrant Gautam, Eileen Bingert, Dawei Zhu, Anne Lauscher, and Dietrich Klakow. 2024. Robust pronoun fidelity with english llms: Are they reasoning, repeating, or just biased? *Transactions of the Association for Computational Linguistics*, 12:1755–1779.

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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models.

Salvatore Greco, Moreno La Quatra, Luca Cagliero, and Tania Cerquitelli. 2025. Towards ai-assisted inclusive language writing in italian formal communications. *ACM Trans. Intell. Syst. Technol.*

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Pascal Gygax, Sayaka Sato, Anton Öttl, and Ute Gabriel. 2021. The masculine form in grammatically gendered languages and its multiple interpretations: A challenge for our cognitive system. *Language Sciences*, 83:101328.

Pascal M. Gygax, Daniel Elmiger, Sandrine Zufferey, Alan Garnham, Sabine Sczesny, Lisa von Stockhausen, Friederike Braun, and Jane Oakhill. 2019. A Language Index of Grammatical Gender Dimensions to Study the Impact of Grammatical Gender on the Way We Perceive Women and Men. *Frontiers in Psychology*, 10:1604.

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

Tamanna Hossain, Sunipa Dev, and Sameer Singh. 2023. MISGENDERED: Limits of large language models in understanding pronouns. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5352–5367, Toronto, Canada. Association for Computational Linguistics.

Fanny Jourdan, Yannick Chevalier, and Cécile Favre. 2025. Fairtranslate: An english-french dataset for gender bias evaluation in machine translation by overcoming gender binarity. *arXiv preprint arXiv:2504.15941*.

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.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Manuel Lardelli, Giuseppe Attanasio, and Anne Lauscher. 2024a. Building bridges: A dataset for evaluating gender-fair machine translation into German. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7542–7550, Bangkok, Thailand. Association for Computational Linguistics.

Manuel Lardelli, Timm Dill, Giuseppe Attanasio, and Anne Lauscher. 2024b. Sparks of fairness: Preliminary evidence of commercial machine translation as English-to-German gender-fair dictionaries. In *Proceedings of the 2nd International Workshop on Gender-Inclusive Translation Technologies*, pages 12–21, Sheffield, United Kingdom. European Association for Machine Translation (EAMT).

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.

Fengyuan Liu, Nikhil Kandpal, and Colin Raffel. 2024. Attribot: A bag of tricks for efficiently approximating leave-one-out context attribution. *arXiv preprint arXiv:2411.15102*.

Ártemis López. 2020. Cuando el lenguaje excluye: Consideraciones sobre el lenguaje no binario indirecto. *Cuarenta Naipes*, 3:295–312.

Andreas Madsen, Siva Reddy, and Sarath Chandar. 2022. Post-hoc interpretability for neural nlp: A survey. *ACM Computing Surveys*, 55(8):1–42.

Edoardo Mosca, Ferenc Szigeti, Stella Tragianni, Daniel Gallagher, and Georg Groh. 2022. SHAP-based explanation methods: A review for NLP interpretability. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4593–4603, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Carolin Müller-Spitzer, Samira Ochs, Alexander Koplenig, Jan Oliver Rüdiger, and Sascha Wolfer. 2024. Less than one percent of words would be affected by gender-inclusive language in german press texts. *Humanities and Social Sciences Communications*, 11(1):1–13.

Anaelia Ovalle, Palash Goyal, Jwala Dhamala, Zachary Jaggers, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2023. "i'm fully who i am": Towards centering transgender and non-binary voices to measure biases in open language generation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1246–1266, New York, NY, USA. Association for Computing Machinery.

Ben Papadopoulos. 2022. A brief history of gender-inclusive Spanish. *Deportate, esuli, profughe*, 48(1):31–48.

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.

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.

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

Andrea Piergentili, Beatrice Savoldi, Matteo Negri, and Luisa Bentivogli. 2025. 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)*, pages 46–63, Geneva, Switzerland. European Association for Machine Translation.

Qwen. 2024. Qwen2.5: A party of foundation models! https://qwenlm.github.io/blog/qwen2.5/. Accessed: 2025-05-03.

Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are references really needed? unbabel-IST 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online. Association for Computational Linguistics.

Gabriele Sarti, Grzegorz Chrupała, Malvina Nissim, and Arianna Bisazza. 2024. Quantifying the plausibility of context reliance in neural machine translation. In *The Twelfth International Conference on Learning Representations*.

Gabriele Sarti, Nils Feldhus, Ludwig Sickert, and Oskar van der Wal. 2023. Inseq: An interpretability toolkit for sequence generation models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 421–435, Toronto, Canada. Association for Computational Linguistics.

Danielle Saunders. 2022. Domain adaptation for neural machine translation. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 9–10, Ghent, Belgium. European Association for Machine Translation.

Danielle Saunders and Bill Byrne. 2020. Addressing exposure bias with document minimum risk training: Cambridge at the WMT20 biomedical translation task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 862–869, Online. 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.

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

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.

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

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

María R Scharrón-del Río and Alan A Aja. 2020. Latinx: Inclusive language as liberation praxis. *Journal of Latinx Psychology*, 8(1):7.

Sabine Sczesny, Magda Formanowicz, and Franziska Moser. 2016. Can gender-fair language reduce gender stereotyping and discrimination? *Frontiers in psychology*, 7:154379.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3145–3153. PMLR.

Gláucia V. Silva and Cristiane Soares, editors. 2024. *Inclusiveness Beyond the (Non)binary in Romance Languages: Research and Classroom Implementation*, 1st edition edition. Routledge.

Jeanette Silveira. 1980. Generic Masculine Words and Thinking. *Women's Studies International Quarterly*, 3(2-3):165–178.

Dagmar Stahlberg, Friederike Braun, Lisa Irmen, and Sabine Sczesny. 2007. Representation of the Sexes in Language. *Social communication*, pages 163–187.

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.

Arjun Subramonian, Vagrant Gautam, Preethi Seshadri, Dietrich Klakow, Kai-Wei Chang, and Yizhou Sun. 2025. Agree to disagree? a meta-evaluation of llm misgendering. *arXiv preprint arXiv:2504.17075*.

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*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size.

Eddie L. Ungless, Sunipa Dev, Cynthia L. Bennett, Rebecca Gulotta, Jasmijn Bastings, and Remi Denton. 2025. Amplifying trans and nonbinary voices: A community-centred harm taxonomy for LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20503–20535, Vienna, Austria. Association for Computational Linguistics.

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

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.

Andreas Waldis, Joel Birrer, Anne Lauscher, and Iryna Gurevych. 2024. The Lou dataset - exploring the impact of gender-fair language in German text classification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10604–10624, Miami, Florida, USA. Association for Computational Linguistics.

Brandon T Willard and Rémi Louf. 2023. Efficient guided generation for llms. *arXiv preprint arXiv:2307.09702*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Emmanouil Zaranis, Giuseppe Attanasio, Sweta Agrawal, and André F. T. Martins. 2025. Watching the watchers: Exposing gender disparities in machine translation quality estimation.

Emmanouil Zaranis, Nuno M Guerreiro, and Andre Martins. 2024. Analyzing context contributions in LLM-based machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14899–14924, Miami, Florida, USA. Association for Computational Linguistics.

A mGeNTE Details

Sentence Editing We apply three key editing interventions. **(1)** Some intricate source sentences containing mentions of multiple referents—and which required the combination of different forms in translation (i.e. neut/masc/fem)—were edited to allow handling them as a coherent unit and in alignment with GeNTE sentences that had been previously edited in the original dataset. **(2)** We compensated for the under-representation of unambiguous feminine data from SET-G. To this end, we adjusted the corpus to achieve a balanced representation of feminine and masculine forms through lexical gender-swapping—statistics on the original gender distribution in Europarl sentences are provided in Figure 8. To ensure alignment with en-it segments, we performed 652 (en-de), 621 (en-es), and 702 (en-el) interventions, ~60% of which served gender balancing. Finally, **(3)** minor edits were made to improve the quality of the corpus (e.g. fixing typos or inaccurate translations). Such changes were applied to 16, 59 and 40 sentences for en-es, en-de and en-el, respectively.

Gendered words annotation To further enrich MGENTE, the linguists in charge of creating the corpus annotated all gendered (masculine/feminine) words in the target sentences. Then, to ensure data quality, a second annotator—with either

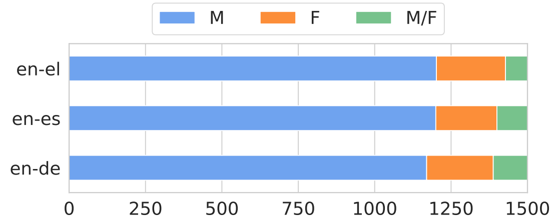


Figure 8: Gender distribution in the original Europarl target sentences. We distinguish between instances containing masculine forms, feminine, or both within the same sentence.

a native or C1 competence of the assigned target language—re-annotated 200 sentences for each language pair. We calculate inter-annotator agreement (IAA) on the exact matches of the gendered words annotated on these subsets. The resulting Dice coefficients (Dice, 1945) are of 0.95 (en-de), 0.94 (en-es), 0.92 (en-el) and 0.95 (en-it), which are considered highly satisfactory. All disagreements were double-checked and reconciled.

		Set-G	Set-N	All
en-it	Tokens	1974	2141	4115
	Types	391	543	802
en-es	Tokens	2389	1974	4363
	Types	306	429	644
en-de	Tokens	2646	1331	3977
	Types	303	403	613
en-el	Tokens	2045	1691	3736
	Types	327	546	743

Table 4: Counts of all and unique MGENTE gendered words annotated by language pair.

The total number of gendered words annotated in MGENTE is shown in Table 4, whereas a qualitative overview of the most frequent words across MGENTE subsets is provided in Figures 20. Also, we note a higher incidence of gendered words in the ambiguous SET-N, consistent with findings by Saunders (2022) that Europarl contains numerous gender-ambiguous cases. As shown in Figure 20, SET-N annotated words are vastly populated with masculine, plural lexical items (e.g. *citizens*, *everyone*, *colleagues*): that is, masculine forms used generically and indiscriminately to refer to mixed or unspecified groups of referents.

B Additional Experimental Details

B.1 Model Inference Details

For all experiments, we used code and model implementations from `transformers` (Wolf et al., 2020) and `vLLM` (Kwon et al., 2023) as the inference engine. For the translation experiments, we prompted the instruct version of Llama 3.1 8B (https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct), Llama 3.3 70B (https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct), Qwen 2.5 72B (https://huggingface.co/Qwen/Qwen2.5-72B-Instruct) and 7B (https://huggingface.co/Qwen/Qwen2.5-7B), Gemma 2 9B (https://huggingface.co/google/gemma-2-9b-it), Phi 4 14B (https://huggingface.co/microsoft/phi-4), Falcon 3 7B (https://huggingface.co/tiiuae/Falcon3-7B-Instruct), Mistral v0.3 7B (https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3), EuroLLM 9B (https://huggingface.co/utter-project/EuroLLM-9B-Instruct), and TowerInstruct Mistral v0.2 7B (https://huggingface.co/Unbabel/TowerInstruct-Mistral-7B-v0.2). We formatted the input using each model's chat template. We provided four in-context exemplar shots, two gendered and two neutral cases in a fixed order, and formatted as the first eight conversation turns. We set the temperature to 0, used `bfloat16` precision, prefix caching, and disabled the attention sliding window for all models but Gemma 2. We guided the decoding via `outlines` (Willard and Louf, 2023), forcing the output to match the following regex:

```
<{lang}>\s \*\*( GENDERED|NEUTRAL)\*\*\s
    \[[^\]]+\]
```

where `lang` is replaced with the standard ISO 639-2 language code of the target language (i.e., it, es, de, or el). For the evaluations, we post-process the output by extracting the labels (**LABEL**) and the translations ([translation]).

B.2 Computational Details

We conducted our experiments on in-house computing infrastructures using nodes with 4x NVIDIA A6000 GPU accelerators. Based on our estimates, translation runs required 10 minutes per configuration on average, which totals to 80 hours (10' x 10 (models) x 4 (langs) x 12 (prompt configura-

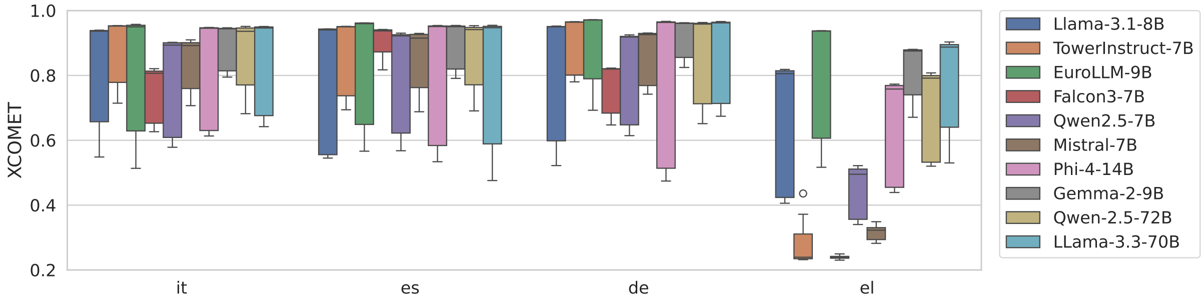


Figure 9: Overall translation quality results for each model across language pairs. Scores reported across all prompt configurations for both zero-shot and few-shots (2,4).

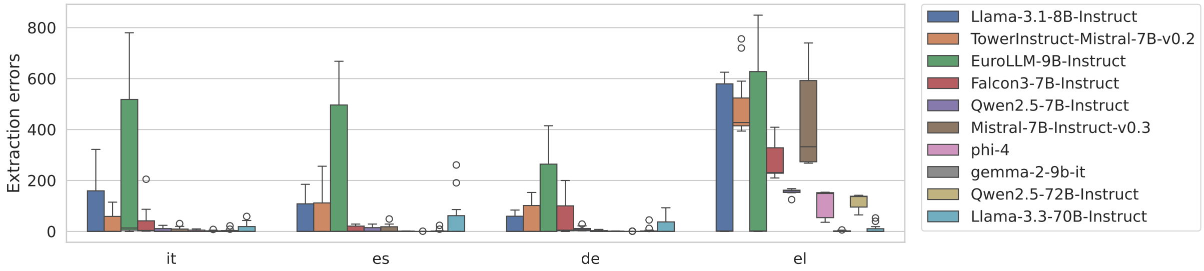


Figure 10: Extraction error count for each model across language pairs. Scores reported across all prompt configurations for both zero-shot and few-shots (2,4).

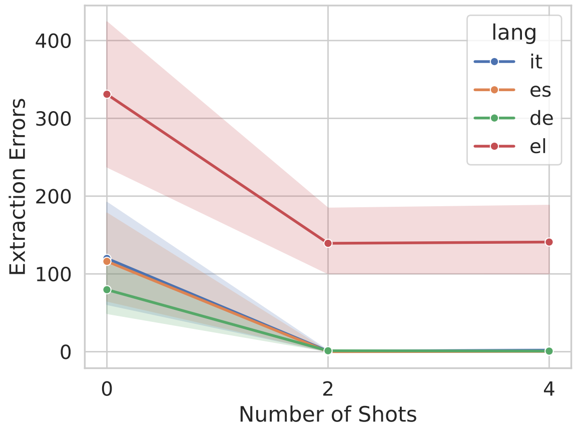


Figure 11: Extraction error count for all models averaged across across language pairs. Trends reported across zero-shot, 2-shot, and 4-shot setups.

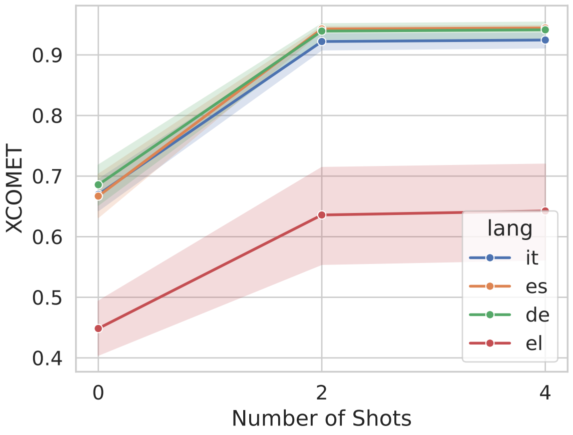


Figure 12: xCOMET scores for all models averaged across across language pairs. Trends reported across zero-shot, 2-shot, and 4-shot setups.

tions)).[27] Running AttnLRP on QWEN 2.5 72B required on average 20 minutes on the same infrastructure, totaling 80 minutes.

B.3 Model and Settings Selection Details

Model selection We conduct extensive preliminary experiments on 10 multilingual LLMs (listed in Appendix B.1). We exclude 5 of them, i.e. Falcon 3 7B, Mistral 7B, Qwen2.5 7B, EuroLLM 9B, and Tower Instruct 7B. This decision is guided by overall translation quality scores and the number of extraction errors, i.e. models' output that do not adhere to the structured expected output described in Appendix B.1. As shown in Figure 9, Falcon 3 7B, Mistral 7B and Qwen2.5 7B achieve comparatively lower scores, especially on Greek. Tower Instruct 7B—on Greek—obtains xCOMET scores below 0.4. In some scenarios, low-quality results can stem from certain languages being underrepresented in the models' training data. Considering the extrac-

[27] i.e. the four prompt variants described in §4.1 in zero-shot, 2-shot and 4-shot, as described in Appendix B.3.

tion errors in Figure 10, EuroLLM 9B largely fails to adhere to the output format requirements.

Shot selection We experiment with zero-shot, 2-shot, and 4-shot configurations. We decided on the 4-shot settings given that it increases translation quality (Figure 12) by reducing extraction errors (Figure 11). Performance across 2-shot and 4-shot do not exhibit notable differences, but we opt for the 4-shot setting for richer context analyses with interpretability tools.

B.4 LLM-as-a-Judge Details

For GNT evaluation, we use LLMs as evaluators, providing source-output pairs as input and constraining the models to generate (intermediate) span-level annotations and (final) sentence-level labels, either *Gendered* or *Neutral*. By design, the translation is classified as neutral *only* if all human referents are neutralized, otherwise it is considered Gendered. To enforce the above-mentioned constraints, we perform structured generation using `outlines` (Willard and Louf, 2023) and the JSON schemas used in Piergentili et al. (2025). Further details about the prompting settings are described below. We compare QWEN 2.5 72B and GPT-4O as *evaluator models* on a gold standard of 1,000 annotated model translations. Results are shown in Table 5.

	Lang	Acc.	Macro-F1
Qwen2.5-72B Instruct	*all*	0.85	0.80
	es	0.90	0.86
	it	0.87	0.78
	el	0.88	0.78
	de	0.84	0.79
gpt-4o-2024-08-06	*all*	**0.92**	**0.87**
	es	0.96	0.94
	it	0.92	0.87
	el	0.89	0.80
	de	0.89	0.85

Table 5: Over(all) evaluation results for Qwen2.5-72B and GPT-4o and across language pairs.

Evaluation Prompt (en-el)

You are an expert language annotator and evaluator of gender-neutral translation for English-Greek. Your task is to extract target Greek phrases that refer to human beings, determine whether each phrase is masculine, feminine, or neutral, and assess if the gender expressed in each phrase is correct with respect to the source. Based on the phrases, **determine whether the sentence was translated in a correctly gendered, wrongly gendered, or neutral way**.

Guidelines:

1. Identify relevant phrases: Carefully read the Greek sentence and extract all phrases that refer to human beings or groups of human beings, including:
- Noun phrases (e.g., “μία άριστη ομιλήτρια”, “η πολιτεία”, “ένας πρόεδρος”),
- Adjective phrases (e.g., “εξαιρετικά κουρασμένος”, “το παντρεμένο”, “ικανοποιημένη”).

2. Evaluate gender information: Consider only the social gender conveyed by the phrases, not grammatical gender, and assign a label to each phrase [M/F/N]. For example:

- Phrases like “ο ομιλητής”, “είναι πολύ χαρούμενος”, “όλοι οι συνάδελφοι”, and “οι εργαζόμενοι” are masculine [M];
- Phrases like “η ομιλήτρια”, “είναι πολύ χαρούμενη”, “όλες οι συναδέλφισσες”, and “οι εργαζόμενες” are feminine [F];
- Phrases like “ένα άτομο που μιλάει στο κοινό”, “είναι πολύ χαρούμενο”, “όλα τα άτομα με τα οποία δουλεύω”, and “η πολιτεία” do not express social gender, therefore they must be considered neutral [N].

3. Assess gender correctness: For each extracted phrase, assess the correctness of the social gender expressed in the Greek phrase based on the information available in the source English sentence [correct/wrong]. Consider that:

- If a phrase is masculine, the English source must contain masculine gender cues (e.g., *he*, *him*, *Mr*, *man*) for it to be correct.
- If a phrase is feminine, the English source must contain feminine gender cues (e.g., *she*, *her*, *Ms*, *woman*) for it to be correct.
- If a phrase is neutral, it is always correct, regardless of gender cues in the source. Note that proper names do not count as valid gender cues — ignore them.

4. Assign a sentence-level label to the translation:

- If there are masculine or feminine phrases in the Greek text and the source contains matching gender cues, label the sentence as “CORRECTLY **GENDERED**”.
- If there are masculine or feminine phrases in the Greek text and the source does not contain matching gender cues, label the sentence as “WRONGLY **GENDERED**.
- If there are only neutral phrases in the Greek text, label the sentence as **“NEUTRAL”**.

Data Annotation For data annotation, we randomly sampled 50 Set-G and 50 Set-N outputs from each of our five models (250 sentences per language pair). Native speakers provided binary annotations (Gendered/Neutral) following comprehensive guidelines available in our project repository at https://github.com/g8a9/mgente-gap.

Evaluation Prompt We use the best prompts and settings identified by Piergentili et al. (2025), with prompts for en-es/it/de from their original data release. We created a new prompt for en-el (see en-el box). Each prompt includes 8 annotated exemplars randomly sampled from the MGENTE PARALLEL-SET, which are excluded at test time.

B.5 Context Attribution Details

B.5.1 Method

Computing Contribution Scores We used the official implementation of the Attention-Aware

Layer-Wise Relevance Propagation algorithm (Achtibat et al., 2024).[28] In particular, we used the efficient implementation where the computed gradients are multiplied by the input embeddings, as introduced by Arras et al. (2025). The library, based on top of `transformers`, patches specific modules in the model code, and allows convenient and efficient computation of logit gradients with respect to input embeddings.

Formally, we compute our contribution scores as follows. Given an input instance, defined by a set of input embeddings $\mathbf{e}$ and a set of output logits ℓ, we first compute the gradient of the *j*-th logit with respect to the *i*-th input as

$$\nabla_{\mathbf{e}_i} \ell_j = \frac{\partial \ell_j}{\partial \mathbf{e}_i} \in \mathbb{R}^N \tag{1}$$

where $\mathbf{e}_i \in \mathbb{R}^N$ is the input embedding vector for token i and N the embedding dimension. This gradient quantifies how a change in each dimension of the input embedding $\mathbf{e}_i$ affects the value of the logit ℓ_j. Then, to attribute a set of adjacent logits (e.g., those corresponding to the translation label or the entire translation), we compute the sum of logits within a span ($\mathcal{L}_{a:b} = \sum_{j=a}^{b} \ell_j$), where a and b are the (inclusive) limits, and compute

$$\nabla_{\mathbf{e}_i} \mathcal{L}_{a:b} = \frac{\partial}{\partial \mathbf{e}_i} \sum_{j=a}^{b} \ell_j = \sum_{j=a}^{b} \frac{\partial \ell_j}{\partial \mathbf{e}_i} = \sum_{j=a}^{b} \nabla_{\mathbf{e}_i} \ell_j \tag{2}$$

For each input embedding $\mathbf{e}_i$, we can compute its contribution to a specific logit span by calculating $\nabla_{\mathbf{e}_i} \mathcal{L}_{a:b}$. Finally, to obtain the contribution of the input embedding $\mathbf{e}_i$ to the logit span $\mathcal{L}_{a:b}$, we take the absolute value of the dot product between the input embedding and its gradient:

$$\mathrm{s}(i, a:b) = |\mathbf{e}_i \cdot \nabla_{\mathbf{e}_i} \mathcal{L}_{a:b}| \tag{3}$$

The dot product is commonly used to weigh each gradient component by its corresponding embedding value (Shrikumar et al., 2017; Achtibat et al., 2024) while the absolute value allows to focus on the magnitude of the contribution rather than its direction.

Score Normalization Once we computed $\mathrm{s}(i, a:b)$ for all input tokens, we apply a max-normalization such that the scores are all scaled within $[0, 1]$. Note that we never consider special tokens from each model's chat template.

B.5.2 Data and annotations

We describe the rationale for the data used in the context analysis. All manually annotated data, corresponding annotation guidelines, and contributions scores are publicly available at https://github.com/g8a9/mgente-gap.

For computing $\mathbf{S_L}$ (contributions to source label), we retain Qwen 2.5 72B outputs with *correct* label predictions for both Set-G (gendered label) and Set-N (neutral label) Given the high performance of the model on this task, we retain ∼4,000 outputs balanced across language pairs and sets. We do not focus on correct vs wrong label predictions given the low number of outputs with wrongly predicted labels, thus hindering reliable analyses.

For computing $\mathbf{S_T}$ (contributions to translation), we focus on ambiguous source sentences from Set-N, where neutral translations are expected as correct outputs. To ensure reliable interpretability analyses and to avoid noise from automatic evaluations of gendered vs. neutral translations (conducted with GPT-4o as a judge), we asked the original language experts who created the corpus to manually annotate translations produced by Qwen 2.5 72B for each language pair, starting from ∼2,000 sentences from Set-N. Our goal was to maximize the number of GNTs per language pair while maintaining a balanced distribution of gendered and neutral outputs both within and across languages. Annotators labeled each translation as either fully gendered or fully neutral, discarding cases where (i) the translation was unintelligible or too low-quality for reliable evaluation, or (ii) mixed scenarios occurred, with both neutral and gendered mentions in the same sentence.

Following this process, we obtained between 400–600 manually validated gendered/neutral translations per language pair to be included in the context attribution analyses. German and Spanish contained higher proportions of neutral translations compared to Italian and French, reflecting the lower GNT performance of the model on these languages.

C Complementary Results

C.1 Gender Neutrality and Label Results

In Figures 13 and 14, we show disaggregated translation results across each model and prompt configuration for Set-N and Set-G respectively. For label macro-f1 trends across models and configurations are calculated over both Sets, see Figure 15.

[28] https://github.com/rachtibat/LRP-eXplains-Transformers

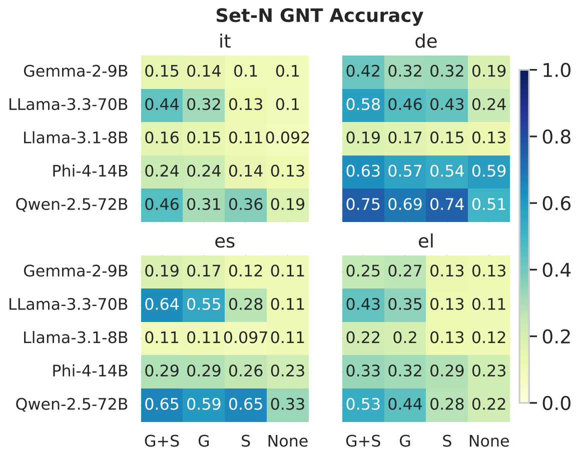


Figure 13: **GNT Accuracy (Set-N) across prompt configurations**, i.e. using both system and guidelines (G+S), only one of them, or None.

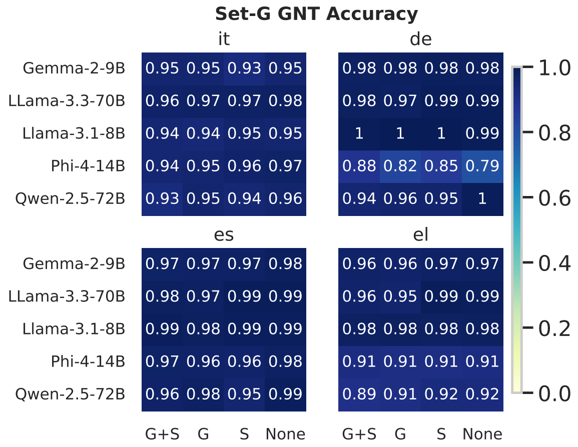


Figure 14: **GNT Accuracy (Set-G) across prompt configurations**, i.e. using both system and guidelines (G+S), only one of them, or None.

C.2 Context attribution results

As complementary results, we provide *i)* Figure 17, which reports context relevance (top 10 contributors) for Translation and Label (All Sets). Also, we report context relevance for Label (18) and Translation (19) across language pairs.

For a complementary view, we also show the mean average contribution of each context part to Label and Translation in Figure 16.

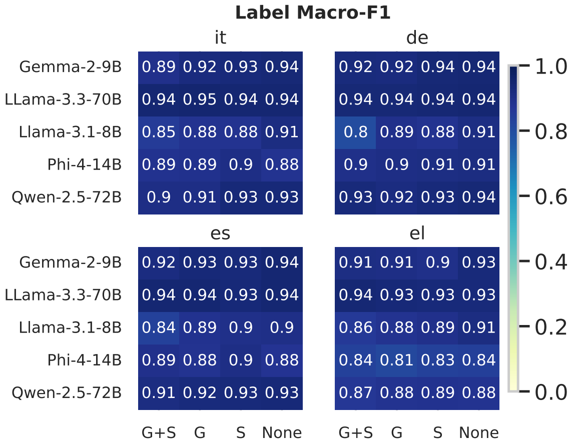


Figure 15: **Label macro-f1 across prompt configurations**, i.e. using both system and guidelines (G+S), only one of them, or None.

	GEMMA9B	LLAMA70B	LLAMA8B	PHI4	QWEN72B
de	$0.93^{\pm0}$	**$0.94^{\pm0}$**	$0.87^{\pm0}$	$0.91^{\pm0}$	$0.93^{\pm0}$
el	$0.91^{\pm0}$	**$0.93^{\pm0}$**	$0.89^{\pm0}$	$0.83^{\pm0}$	$0.88^{\pm0}$
es	$0.93^{\pm0}$	**$0.94^{\pm0}$**	$0.88^{\pm0}$	$0.89^{\pm0}$	$0.92^{\pm0}$
it	$0.92^{\pm0}$	**$0.94^{\pm0}$**	$0.88^{\pm0}$	$0.89^{\pm0}$	$0.92^{\pm0}$
	$0.92^{\pm0}$	**$0.94^{\pm0}$**	$0.88^{\pm0}$	$0.88^{\pm0}$	$0.91^{\pm0}$

Table 6: Macro-F1 Label scores calculated over both Sets, by language and model (mean across configurations, $^{\pm\text{ std}}$).

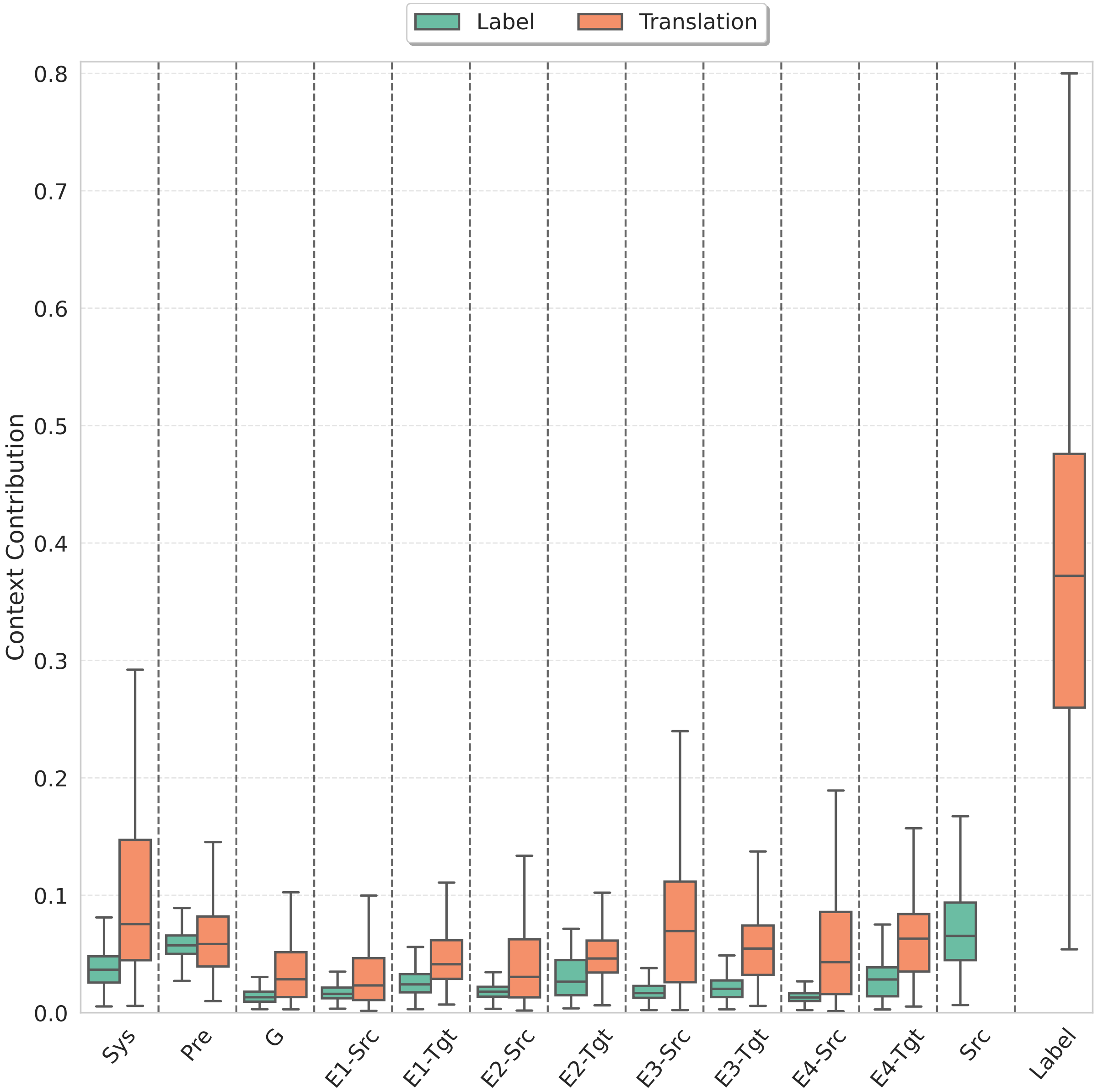


Figure 16: **Average contribution of each input context part** to the generated output sequence, calculated for both the final label and the full generated translation. Set-N and Set-G together.

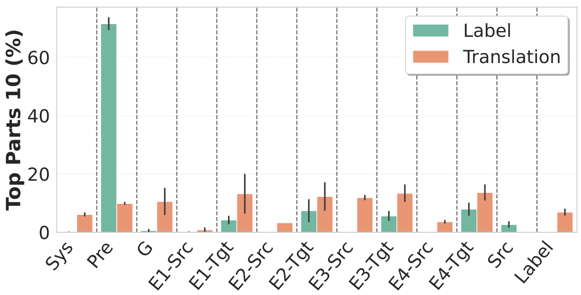


Figure 17: **Overall Relevance of contexts part to Source Label and Translation**. Ratio of occurrence within the top 10 scores for both Set-N and Set-G together.

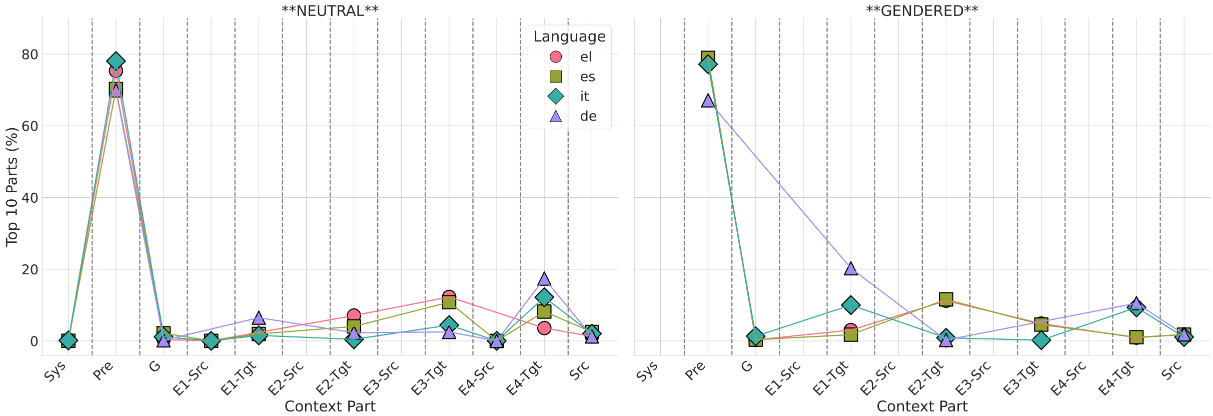


Figure 18: **Relevance of contexts part to Source Label by Language**. Ratio of occurrence within the top 10 scores. Results for gendered (left) and neutral (right) source sentences, respectively from Set-G and Set-N.

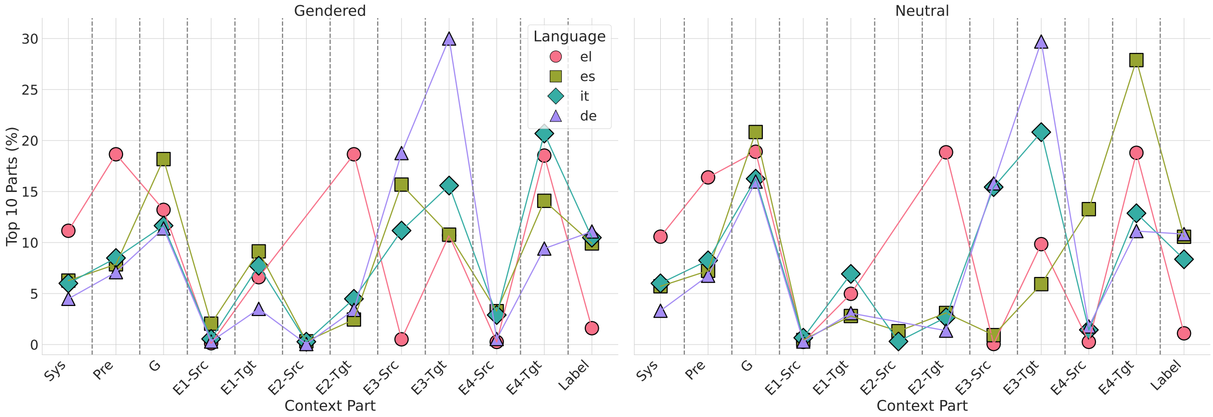


Figure 19: **Relevance of context parts to Translation by Language.** Ratio of occurrence within the top scores. Results for ambiguous sentences from Set-N with wrongly gendered (left) and correctly neutral (right) translations.

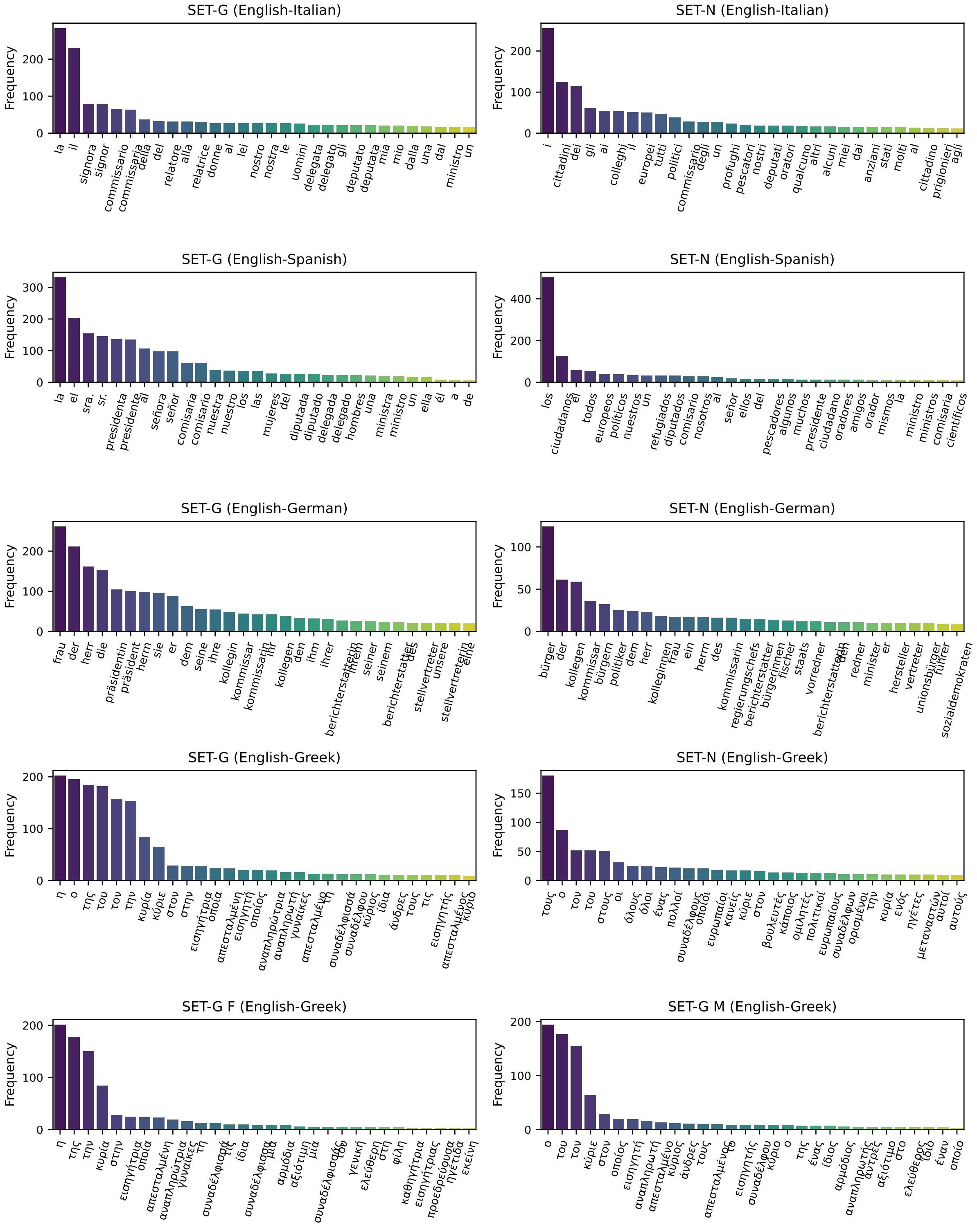


Figure 20: Top 30 most frequent gendered words annotated in MGENTE.

SET-N	SRC	**Pensioners** are in favour of strengthening criminal law, [...]	
en-it	REF-G	**I pensionati** sono favorevoli a un rafforzamento del diritto penale, [...]	
	REF-N$_1$	**Le persone pensionate**[pensioned people] sono favorevoli a un rafforzamento del diritto penale, [...]	
	REF-N$_2$	**Chi percepisce una pensione**[those receiving a pension] è favorevole a un rafforzamento del diritto penale, [...]	
	REF-N$_3$	**Le persone in pensione**[retired people] sono favorevoli a un rafforzamento del diritto penale, [...]	
en-es	REF-G	**Los** pensionistas están a favor de reforzar el Derecho penal no solo nacional, [...]	
	REF-N$_1$	**Hay pensionistas**[there are pensioners] que están a favor de reforzar el Derecho penal no solo nacional, [...]	
	REF-N$_2$	**Quienes reciben pensiones**[Those receiving a pension] están a favor de reforzar el Derecho penal no solo nacional, [...]	
	REF-N$_3$	**Las personas pensionistas**[pensioned people] están a favor de reforzar el Derecho penal no solo nacional, [...]	
en-de	REF-G	Die **Rentner** begrüßen den Ausbau nicht nur des einzelstaatlichen, [...]	
	REF-N$_1$	**Die Menschen in Rente**[people in retirement] begrüßen den Ausbau nicht nur des einzelstaatlichen, [...]	
	REF-N$_2$	**Die Personen im Ruhestand**[persons in retirement] begrüßen den Ausbau nicht nur des einzelstaatlichen, [...]	
	REF-N$_3$	**Pensionierte Menschen**[Pensioned people] begrüßen den Ausbau nicht nur des einzelstaatlichen, [...]	
en-el	REF-G	Οι **συνταξιούχοι** είναι υπέρ της ενίσχυσης του ποινικού δικαίου, [...]	
	REF-N$_1$	**Τα συνταξιοδοτημένα άτομα**[the retired individuals] είναι υπέρ της ενίσχυσης του ποινικού δικαίου, [...]	
	REF-N$_2$	**Τα συνταξιοδοτημένα άτομα**[the retired individuals] είναι υπέρ της ενίσχυσης του ποινικού δικαίου, [...]	
	REF-N$_3$	**Ο συνταξιοδοτημένος πληθυσμός**[the retired population] είναι υπέρ της ενίσχυσης του ποινικού δικαίου, [...]	
SET-G	SRC	I trust the **Commissioner** will promise that he will exercise extra vigilance.	M.
en-it	REF-G	Spero che **il Commissario** ora prometta di vigilare attentamente a tale riguardo.	
	REF-N$_1$	Spero che **il membro della Commissione**[the member of the board] ora prometta di vigilare attentamente a tale riguardo.	
	REF-N$_2$	Spero che **l'esponente della Commissione**[the representative of the board] ora prometta di vigilare attentamente a tale riguardo.	
	REF-N$_3$	Spero che **il membro della Commissione**[the member of the board] ora prometta di vigilare attentamente a tale riguardo.	
en-es	REF-G	Espero que **el Comisario** prometa controlar exhaustivamente esta situación.	
	REF-N$_1$	Espero que **la representación de la Comisión**[the representative of the board] prometa...	
	REF-N$_2$	Espero que **la persona de la Comisión que vaya a ocuparse de ello**[the person of the board in charge of this] prometa...	
	REF-N$_3$	Espero **que quien está a la cabeza de la Comisión**[who is in charge of the board] prometa...	
en-de	REF-G	Von **dem Herrn Kommissar** erwarte ich heute die Zusage, **er** werde mit Argusaugen darüber wachen.	
	REF-N$_1$	**Von dem Kommissionsmitglied**[From the board member] erwarte ich heute die Zusage, **es**[they] werde mit Argusaugen...	
	REF-N$_2$	**Von dem Kommissionsmitglied**[From the board member] erwarte ich heute die Zusage, **es**[they] werde mit Argusaugen...	
	REF-N$_3$	**Von dem Kommissionsmitglied**[From the board member] erwarte ich heute die Zusage, **es**[they] werde mit Argusaugen...	
en-el	REF-G	Προσδοκώ από **τον** Επίτροπο να δεσμευτεί ότι θα επιβλέψει αυστηρά την κατάσταση.	
	REF-N$_1$	Προσδοκώ από **το μέλος της Επιτροπής**[the member of the Commission] να δεσμευτεί ότι θα επιβλέψει αυστηρά την κατάσταση.	
	REF-N$_2$	Προσδοκώ από **το μέλος της Επιτροπής**[the member of the Commission] να δεσμευτεί ότι θα επιβλέψει αυστηρά την κατάσταση.	
	REF-N$_3$	Προσδοκώ από **το μέλος της Επιτροπής**[the member of the Commission] να δεσμευτεί ότι θα επιβλέψει αυστηρά την κατάσταση.	
SET-G	SRC	It is true that we women are those who suffer most in war zones but we are **the bearers** of alternatives to war.	F.
en-it	REF-G	...noi **donne** siamo **le** più **colpite** nei luoghi di guerra ma siamo **portatrici** di alternative alla guerra.	
	REF-N$_1$	...**le persone come me sono le più colpite**[people like me] [...] ma siamo **portatrici**[people bringing] di alternative...	
	REF-N$_2$	...**noi esseri umani più colpiti**[we human beings] [...] ma **portiamo** [we bring] alternative...	
	REF-N$_3$	...noi siamo **la tipologia di persone più colpita**[the type of people] [...] ma **portiamo con noi**[bringing with us] alternative...	
en-es	REF-G	...**nosotras las mujeres** somos **las** más **afectadas** en los lugares donde hay guerra, pero somos **portadoras** de alternativas...	
	REF-N$_1$	...**las personas de género femenino somos las más afectadas**[people of feminine gender] [...] **portadoras**[people bringing]...	
	REF-N$_2$	...**las personas de género femenino somos las más afectadas**[people of feminine gender] [...] **portadoras**[people bringing]...	
	REF-N$_3$	...**las personas más afectadas**[the people] [...] **aportamos**[we bring] alternativas...	
en-de	REF-G	...wir **Frauen** am stärksten in den Kriegsgebieten zu leiden haben, sondern sind wir auch **Trägerinnen** von Alternativen...	
	REF-N$_1$	....**wir**[we_] am stärksten in den Kriegsgebieten zu leiden haben, sondern **tragen wir**[we bring] auch Alternativen...	
	REF-N$_2$	...**wir**[we_] am stärksten in den Kriegsgebieten zu leiden haben, sondern sind wir auch **Anbietende**[suppliers] von Alternativen...	
	REF-N$_3$	...wir **als weibliches Geschlecht**[as feminine gender] am stärksten [...], **tragen wir**[we bring] zu Alternativen...	
en-el	REF-G	...εμείς οι **γυναίκες** είμαστε **αυτές** που πλήττονται περισσότερο στις εμπόλεμες περιοχές, αλλά είμαστε **φορείς**...	
	REF-N$_1$	...εμείς **τα άτομα θηλυκού γένους**[as individuals of female gender] [...], **φέρνουμε**[we bring]...	
	REF-N$_2$	...εμείς **τα άτομα θηλυκού φύλου**[as individuals of female gender] [...], είμαστε και **φορείς**[people bringing]...	
	REF-N$_3$	...εμείς **τα άτομα γυναικείου γένους**[as individuals of female gender] [...], είμαστε και **φορείς**[people bringing]...	

Table 7: Parallel, multilingual MGENTE entries from the COMMON-SET. We provide an example entry from SET-N, and two examples (masculine and feminine) from SET-G. REF-G indicates the gendered references, REF-N$_{1,2,3}$ highlight the neutralized expressions produced by Translator 1, 2, and 3 respectively. Within each language, identical neutralizations are shown with the same color highlight. Words in **bold** are mentions of human referents; underlined source words are linguistic cues informing about the referents's gender.