Enhancing Gender-Inclusive Machine Translation with Neomorphemes and Large Language Models

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

Abstract

Machine translation (MT) models are known to suffer from gender bias, especially when translating into languages with extensive gendered morphology. Accordingly, they still fall short in using gender-inclusive language, also representative of non-binary identities. In this paper, we look at gender-inclusive neomorphemes, neologistic[1] elements that avoid binary gender markings as an approach towards fairer MT. In this direction, we explore prompting techniques with large language models (LLMs) to translate from English into Italian using neomorphemes. So far, this area has been under-explored due to its novelty and the lack of publicly available evaluation resources. We fill this gap by releasing NEO-GATE,[2] a resource designed to evaluate gender-inclusive en→it translation with neomorphemes. With NEO-GATE, we assess four LLMs of different families and sizes and different prompt formats, identifying strengths and weaknesses of each on this novel task for MT.

1 Introduction

Machine translation (MT) has been found to be susceptible to gender bias, i.e. the tendency to produce default masculine outputs or stereotypical gender associations (Saunders et al., 2020; Savoldi et al., 2021; Piazzolla et al., 2023) when gender information about human referents is absent. This is especially relevant when translating from notional gender languages like English, which express gender only through a limited set of elements (e.g., *he/she* pronouns), into grammatical gender target languages, such as Italian, German, and Spanish, which mark gender extensively in their morphology (e.g., en: *"My friends are rich"* → it: *"I miei amici sono ricchi"* [M] vs *"Le mie amiche sono ricche"* [F]). The consequences of this behavior are systematically harmful (Blodgett et al., 2020) to women, who risk being under-represented and stereotypically defined, and non-binary individuals, who are erased from representation or misgendered within binary gender linguistic frameworks (Misiek, 2020; Dev et al., 2021).

In light of this, in this paper we look at neologistic solutions – which are emerging from grassroots efforts to make language more inclusive – as a path towards gender-inclusive MT. Linguistic innovations such as neopronouns (e.g., en *ze* instead of *he/she*, sw *hen* instead of *han/hon*) and neomorphemes (e.g., it *-ə/-ɜ*, es *-e/-es* in place of gendered inflectional morphemes) add new elements to the grammar and morphology to allow for the expression of non-binary gender identities or to convey gender neutrality (Bradley et al., 2019). To date, the use of neologistic solutions is not systematized yet, with alternative paradigms coexisting and new ones continuously emerging. The choice and use of one paradigm of neologistic devices (e.g., the neopronouns *xe/xem/xyr/xyrs/xemself* vs *ze/zir/zir/zirs/zirself*, etc.) depends on individuals' identity and preferences in gender expression.

The use of neologistic devices in MT is still a largely unexplored research area, due to the novelty of this approach and to the lack of dedicated

[1]Following (Rose et al., 2023), we refer to them as *neologistic* because of their linguistically innovative nature.

[2]Available at https://huggingface.co/datasets/FBK-MT/Neo-GATE under the CC-BY-4.0 licence.

Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1), pages 300–314
June 24-27, 2024 ©2024 European Association for Machine Translation

EN	I like being surrounded by my friends.
A	Mi piace essere circondat**o** da**i** mie**i** amic**i**.
B	Mi piace avere persone amiche intorno a me. [en: I like to have people who are friends around me]
C	Mi piace che intorno a me siano presenti persone che considero mie amiche. [en: I like that there are people around me who I consider my friends]
D	Mi piace essere circondat* da* mie* amic*.
E	Mi piace essere circondatə daɜ mieɜ amicɜ.

Table 1: Examples of en→it translations with no gender information in the source. Example A uses generic masculine formulations to refer to human beings (in bold), while the rest employ different gender-inclusive strategies (underlined). B and C use periphrases of different verbosity, while D and E employ different neomorpheme paradigms.

resources, which in turn is complicated by the unfixed nature of these solutions. Ideally, gender-inclusive MT research should factor in the multiplicity of paradigms that make up the landscape of neologistic devices (Lauscher et al., 2022). However, the unavailability of evaluation and training resources which can be adjusted to any paradigm is a bottleneck for the investigation of gender-inclusive MT. Also, neural MT has been proven to fail in handling neologistic gender-inclusive language (Lauscher et al., 2023). Looking at other options, LLMs' ability to adapt to unseen tasks through in-context learning (Brown et al., 2020; Min et al., 2022) offers a viable path toward gender-inclusive MT without the need for extensive training data. Thus, in this work we investigate multilingual LLMs' ability to adapt to new, inclusive morphological paradigms in translation.

To this aim, we *i)* release NEO-GATE, a benchmark to evaluate gender-inclusive en→it translation with any of the ever-emerging neomorpheme paradigms; *ii)* explore different prompting strategies for neologistic gender-inclusive MT, across three open and one commercial LLMs, and the two most popular Italian neomorpheme paradigms.

2 Background

Following evolving social and linguistic phenomena (Sendén et al., 2021; Waldendorf, 2023), there has been a rising demand for the integration of gender-inclusive language in natural language processing (NLP) technologies to make them inclusive of all gender identities (Dev et al., 2021). In MT, gender-neutral translation (GNT) was recently proposed as a gender-fair approach to make translation technologies less biased and more inclusive (Piergentili et al., 2023a). GNT consists in using gender-neutralization strategies, such as epicene formulations, (e.g., 'persone' – en: *people* – in examples B and C in Table 1), to avoid expressing the gender of human beings in the target language. However, this approach has considerable limitations: *i)* it can result in verbose phrasings, as in example C, which are only acceptable in specific contexts and textual domains, namely formal and institutional communication (Piergentili et al., 2023b); *ii)* it is arguably impossible to translate some terms by applying these strategies in grammatical gender languages like Italian (e.g., kinship terms, such as *parent* → it *genitore/genitrice*) (Motschenbacher, 2014). Moreover, the use of circumlocutory language to avoid expressing gender is regarded as a form of *indirect* non-binary language (Attig and López, 2020), in that it conceals gender, while other, *direct* solutions emphasize it.

Indeed, innovative alternatives have been proposed by queer communities as well. Neologistic elements, such as neopronouns and neomorphemes, have emerged in notional gender languages, such as Swedish (Gustafsson Sendén et al., 2015) and English (McGaughey, 2020), as well as in grammatical gender languages, such as Spanish (Sarmiento, 2015), French (Kaplan, 2022), and German (Paolucci et al., 2023). These devices aim to enrich the language with extra resources, which act as gender-neutral alternatives to gendered linguistic elements, and allow for a manifest inclusion of gender identities beyond the masculine-feminine binary (Bradley et al., 2019). Individuals choose to use neologistic devices for themselves as they best fit their gender identity and as an open statement of it, rather than using gender-neutralization strategies, which would instead circumvent it (Gautam, 2021). Such innovative solutions are mostly used within LGBT+ communities, over informal channels. However, their use and acceptance are on the rise (Waldendorf, 2023; Rose et al., 2023). While there is no *one-fits-all* approach to gender-inclusive language (Lardelli and Gromann, 2023), neologistic devices have naturally emerged as a response to the demand for a direct solution that deserves attention.

In this work, we focus on the use of neomorphemes in en→it translation, a scenario in which gender-related ambiguities – and, consequently, the need for gender-inclusive solutions – are crucial. Indeed, Italian is characterized by a pervasive gender-marking system, which assigns a gender to each noun and every word syntactically linked to it, including some verbal forms. Coherently, there have been several proposals of neomorpheme paradigms, which currently coexist and are not yet ultimately codified (Sulis and Gheno, 2022). Such proposals promote the use of specific characters in place of gendered morphemes (e.g., masculine -o and feminine -a, as in "uno scienziato" [M], "una scienziata" [F] – en: a scientist). The proposed neomorphemes range from letters of the Latin alphabet (e.g., 'u'→unu scienziatu), to typographical symbols (e.g., '*'→un* scienziat*), to letters of the international phonetic alphabet (IPA), like the Schwa neomorpheme paradigm, which uses the IPA letter 'ə' for the singular number (unə scienziatə) and 'ɜ' for the plural (alcunɜ scienziatɜ – en: a few scientists) (Baiocco et al., 2023).

Gender Inclusivity in NLP So far, research on gender-inclusive neologistic solutions in NLP has been mainly limited to first explorations in monolingual settings, and mostly confined to English neopronouns. In a pioneering effort, Lauscher et al. (2022) discussed the adoption of neopronouns and formulated a list of desiderata to model the use of pronouns in language technologies. They redefined pronouns as an *open class*, i.e., a class which is not fixed and allows for the inclusion of emerging neopronoun paradigms each user may identify with. This is crucial when dealing with such novel and constantly evolving devices.

In the context of generative tasks, several studies highlight the difficulty of LLMs in handling neopronouns in zero-shot settings (Brandl et al., 2022; Hossain et al., 2023; Ovalle et al., 2023a). Ovalle et al. (2023b) identify byte pair encoding tokenization (Sennrich et al., 2016) as a major cause of LLMs' shortcomings, coherently with Gaido et al. (2021), which observed the same phenomenon in gender bias investigation. Tokenization, paired with the un-fixed nature of innovative gender-inclusive solutions, may represent a crucial problem for LLMs in correctly generating neomorphemes as well. Indeed, as mentioned above, the range of characters used as neomorphemes is wide and *i)* not all characters are necessarily represented in the training data of LLMs; *ii)* the use of different characters in place of more common gendered morphemes may result in different tokenizations for otherwise identical terms, which in turn could interfere with LLMs' ability to generate fluent text.

In MT research, the sole experiment in developing systems partially compatible with neologistic devices is a proof-of-concept built by Saunders et al. (2020) in a gender bias mitigation experiment. They fine-tuned en→de and en→es MT models which use placeholder tags in place of determiners and inflectional morphemes, to be replaced with non-binary forms post-inference. In a broader analysis of gender bias in LLMs, Vanmassenhove (2024) reports that ChatGPT never produces gender-inclusive neomorphemes when translating ambiguous English sentences into Italian, although without specifically prompting the model to do so. The sole analysis dedicated to the use of neologistic devices in MT is the one by Lauscher et al. (2023), which shows how commercial systems fail to deal with English neopronouns, resulting in either misgendering or low-quality outputs.

A major bottleneck hindering the exploration of gender-inclusive neologistic devices in MT is the lack of publicly available evaluation resources. To bridge this gap, in the next section we introduce a dedicated resource: NEO-GATE.

3 The NEO-GATE benchmark

NEO-GATE is designed to evaluate the use of neomorpheme paradigms in en→it translation. Following Lauscher et al. (2022), and extending their desiderata to gender-inclusive translation, we treat neomorphemes as an open class embracing all possible neomorpheme paradigms. To this aim, we design NEO-GATE to be adjustable to any neomorpheme paradigm in Italian, thanks to a set of adaptable references and annotations.

NEO-GATE is built upon GATE (Rarrick et al., 2023), a benchmark for the evaluation of gender bias in MT. In GATE, the gender of human entities is unknown, i.e. there are no elements providing gender information about human referents in the (English) source sentences. GATE also provides target language references which only differ in the feminine/masculine gendered words that refer to human entities (see Table 2). Since in our gender-inclusive translation task we envision the use of neomorphemes for human referents whose gender is unknown, GATE is an ideal candidate corpus as a basis for the creation of our resource. NEO-GATE includes GATE's test set entries,[3] with the addition of references and (word-level) annotations based on a set of placeholder tags, which can be automatically replaced with the de-

[3]Except for two of GATE's entries, which do not feature gender-marked terms in the references.

GATE	**Source**	The department chair said they might hire new professors
	Ref. Masc.	Il direttore del dipartimento ha detto che potrebbero assumere nuovi professori
	Ref. Fem.	La direttrice del dipartimento ha detto che potrebbero assumere nuove professoresse
NEO-GATE	**Ref. tagged**	<DARTS> direttor<ENDS> del dipartimento ha detto che potrebbero assumere nuov<ENDP> professor<ENDP>
	Annotation	il la <DARTS>; direttore direttrice direttor<ENDS>; nuovi nuove nuov<ENDP>; professori professoresse professor<ENDP>;
NEO-GATE ADAPTED *	**Reference**	L* direttor* del dipartimento ha detto che potrebbero assumere nuov* professor*
	Annotation	il la l*; direttore direttrice direttor*; nuovi nuove nuov*; professori professoresse professor*;
NEO-GATE ADAPTED ə/ɜ	**Reference**	Lə direttorə del dipartimento ha detto che potrebbero assumere nuovɜ professorɜ
	Annotation	il la lə; direttore direttrice direttorə; nuovi nuove nuovɜ; professori professoresse professorɜ;

Table 2: Examples of a single entry in GATE, NEO-GATE, and adapted to the two neomorpheme paradigms used in our experiments (§4.2). The terms of interest for our evaluation are highlighted.

sired forms. The tagged references and the annotations are discussed in §3.1, while NEO-GATE's evaluation metrics are described in §3.2.

3.1 Tagged references and annotations

For each entry in GATE's test set (see Table 2 for an example), we want to create an additional reference translation featuring neomorphemes. To this aim, for each gendered target word we replace gendered morphemes and function words (articles, prepositions, etc.) with placeholder tags. The placeholders serve to identify words of interest for our task and make this reference adjustable to any neomorpheme paradigm by automatically replacing them with the desired forms. The tagset was designed to cover all parts of the grammar which express grammatical gender, and accounts for distinct singular and plural forms (e.g., the tags <DARTS> and <DARTP> for the singular and plural definite articles respectively). This enables the evaluation of neomorpheme paradigms that use different characters for the singular and the plural case, e.g., the 'Schwa' paradigm mentioned in §2. While for content words we only replace the inflectional morpheme with a tag (either <ENDS> or <ENDP>), for function words we use placeholders that cover the whole word. We do so because: *i)* in Italian, some function words are not morphologically derived but paradigmatically opposed (e.g., the definite article singular masculine forms 'il' and 'lo' vs the feminine form 'la'); *ii)* as neomorpheme use is not yet settled, there are instances where competing forms exist for a single word and differ in the root part (e.g., the forms 'lɜ' and 'ə' have been proposed[4] for the plural definite article). Since it would be impossible to account for all existing forms with the sole inflectional placeholders, we replace all function words entirely with dedicated tags (see Appendix B for further details).

We performed the same annotation on a subset of GATE's dev set as well, so as to have a pool of exemplar sentences for our experiments (see §4). Table 8 in Appendix A describes all the tags used in NEO-GATE, as well as the forms we used to replace them in our experiments. NEO-GATE statistics are reported in Table 3.

	Entries	Tags	Content	Function	Singular	Plural
Test	841	2,479	1,539	940	1,316	1,163
Dev	100	345	211	134	184	161

Table 3: Statistics of NEO-GATE's test and dev sets.

To ensure the quality of our resource, the references were manually annotated by a linguist following dedicated guidelines.[5] Using the same guidelines, a second linguist[6] independently reannotated a 15% randomly selected subset of target language sentences. Inter-annotator agreement computed with Cohen's kappa (Cohen, 1960)[7] on label assignment for the placeholder tags amounts to 0.94, indicating almost perfect agreement (Landis and Koch, 1977). The few disagreements were overlooks and were thus reconciled.

NEO-GATE's set of annotated words is automatically extracted by comparing the masculine, feminine, and tagged references. It serves to define the words upon which the evaluation is based. It includes the three forms required for the evaluation, i.e the masculine and feminine forms, and the forms with the placeholder tags, which are to be

[4]The first form was proposed in https://italianoinclusivo.it/scrittura/, and the second in https://effequ.it/schwa/

[5]The guidelines are available in NEO-GATE's release page.
[6]Both linguists are authors of the paper.
[7]We use scikit-learn (Pedregosa et al., 2011).

replaced with a neomorpheme (e.g., *direttore*, *direttrice*, and *direttor** in 'NEO-GATE ADAPTED *', in Table 2).

3.2 Evaluation metrics

While holistic metrics like BLEU (Papineni et al., 2002) have been previously explored to inform the evaluation of gender bias in MT (Bentivogli et al., 2020; Currey et al., 2022), these metrics are not designed to provide fine-grained assessments for specific linguistic phenomena. Rather, they offer a coarse-grained indication of overall translation quality, thus motivating the use of dedicated metrics that allow for pinpointed evaluations, isolating gender from other factors that could impact generic performance. To this aim, we rely on NEO-GATE's annotations associated with each source sentence (e.g. "*the department chair* said they might hire *new professors*" in Table 2). Every annotation comprises three forms for each gender-related word: masculine, feminine, and the form with neomorphemes (e.g. "il la l*", "direttore direttrice direttor*", "nuovi nuove nuov*", "professori professoresse professor*"). In the description of our metrics, we refer to the total number of annotated triplets as '$annotations$' (4 triplets in our example). Scores computation is carried out by scanning the models' output translations word by word, and checking whether such words match any of the three forms in the annotated triplets. Each matched word increases the '$matched$' count. If the matched form is the one with neomorphemes (e.g. "direttor*"), we count it as '$correct$'. To further monitor models' behavior we also count the generated words that include a neomorpheme, regardless of their presence in the annotations, in the $found\ neomorphemes$ tally. With these parameters, we compute the following metrics.

Coverage (COV) and accuracy (ACC). As our primary evaluation method, we draw from the metrics defined by Gaido et al. (2020) in the context of binary gender translation. Such a method first computes *coverage* as the ratio of annotated words $matched$ in the outputs over NEO-GATE's $annotations$: $COV = \frac{matched}{annotations}$. This score serves two purposes: *i)* it is indicative of the informativeness of the accuracy evaluation, as a low coverage indicates that the accuracy score described below is calculated over a relatively low number of annotations; *ii)* it can function as an indirect indicator of translation quality (Savoldi et al., 2022), i.e. a higher coverage suggests that the model generates the expected target words.

On this basis, we then compute *accuracy* as the proportion of $correct$ neomorphemes generated by the model over the total number of annotations $matched$ in the outputs: $ACC = \frac{correct}{matched}$. This score measures models' ability to correctly produce neomorphemes.

The combination of these two metrics allows to distinguish between the generation of an annotated word (regardless of its gender) from its gender realization (fem./mas./neom.), thus ensuring pinpointed analyses.

Coverage-weighted accuracy (CWA). For a comprehensive view of models' overall performance, CWA takes into account both how accurately a model generates neomorphemes and the proportion of *annotations* covered by the evaluation: $CWA = \frac{correct}{matched} * \frac{matched}{annotations}$. This score allows for the comparison of different systems, for which both coverage and accuracy should be taken into account. Indeed, a system's high accuracy may be the result of an evaluation based on a particularly small set of matched annotations, impairing the comparison with other systems' performance evaluated on bigger portions of the corpus. While the other metrics serve to investigate each model's behavior, coverage-weighted accuracy allows for a fairer comparison of different systems.

Mis-generation (MIS). We also consider the case where models generate neomorphemes inappropriately, for instance by applying the use of neomorphemes to words that do not refer to human entities (e.g., table→tavol* instead of 'tavolo'). Such scenario is crucial, as overgeneralizing the use of neomorphemes compromises the intelligibility of the translation. Thus, to we quantify *mis-generations*, i.e. the number of output words – which are not annotated in NEO-GATE – that feature neomorphemes. Accordingly, $MIS = \frac{found\ neomorphemes - correct}{annotations}$. This score complements the evaluation, as it can signal undesired behaviors even despite good accuracy and coverage.

4 Experimental settings

4.1 Models

We experiment with three open, decoder only LLMs. TowerInstruct-7B-v0.1,[8] is fine-tuned

[8] https://huggingface.co/Unbabel/TowerInstruct-7B-v0.1

	BLEU	chrF	TER ↓	BERTSc.	COMET
OPUS-MT	27.53	57.61	58.95	87.42	82.68
GPT-4	**32.34**	**61.11**	**54.87**	**88.76**	**87.05**
Tower	30.88	59.41	56.96	88.17	86.21
Mixtral	29.63	58.68	59.35	87.81	86.11
LLama 2	26.28	55.92	61.98	87.02	84.23

Table 4: Translation quality results. Best scores are in **bold**. Cases where LLMs outperform the MT model are underlined.

for MT, whereas the other two – Mixtral-8x7B-Instruct-v0.1 (Jiang et al., 2024) and LLama 2 70B chat (Touvron et al., 2023) – are not specialized for MT. We also include the commercial model GPT-4 (Achiam et al., 2023),[9] which proved to perform well in gender-inclusive MT experiments (Savoldi et al., 2024). We use the models' default settings, except for the temperature parameter, which we set to 0 following Peng et al. (2023). We do not include neural MT models as no model currently supports neomorphemes and no dedicated training or fine-tuning data is available.

To ensure the suitability of the selected LLMs for translation-related tasks, we preemptively test their generic en→it translation performance on the FLORES 101 benchmark (Goyal et al., 2022). We prompt the models to translate with a few-shot prompt (see Appendix C). In these experiments, we include `opus-mt-en-it`,[10] a state-of-the-art neural MT model, as a reference system for translation quality. For general MT evaluation, we use BLEU, chrF (Popović, 2015), TER (Snover et al., 2006), BERTScore (Zhang et al., 2019), and COMET (Rei et al., 2020). Using these metrics allows for a comparative evaluation of translation performance based on different aspects, namely the surface similarity to human-made reference translations (BLEU, chrF, TER), and the semantic adherence to those references (BERTScore) and to the source (COMET). The results in Table 4 show that the LLMs perform very well in MT, often outperforming the SOTA MT system in this setting.

4.2 Neomorphemes

We focus on the two most popular Italian neomorpheme paradigms (Comandini, 2021): *i)* the *Asterisk*, which uses the symbol '*' as a graphemic device in place of regular inflectional morphemes (Haralambous and Dichy, 2018); *ii)* the *Schwa*, which features both a singular form, for which the character 'ə' is used, and a plural form, represented with the character 'ɜ' (Baiocco et al., 2023).

For each paradigm, we create a tagset mapping (see Appendix A) that associates the tags used in the tagged references with the desired form for that specific paradigm. As no complete codification of the use and the orthography of neomorphemes in Italian is available (Thornton, 2020), we referenced established resources such as the website *Italiano Inclusivo*,[11] and examples found in scientific literature, such as Rosola et al. (2023). As these sources do not cover the whole set of possibly gendered elements in the grammar, we derived the missing forms by analogy from elements of the same class. For example, since none of these sources describes the full set of articulated prepositions, which express gender in Italian, we used the given examples as a model for the rest of the class.

4.3 Prompts

We experiment with one zero-shot and three few-shot formats, illustrated by the examples in Table 5. The few-shot prompts follow the format used in Sánchez et al. (2023), which was found to be useful for controlling gender expression in translation. We instantiate different conceptualizations of the task, ranging from a simple pairing of source sentences directly with gender-inclusive translations, to a ternary opposition of masculine, feminine, and gender-inclusive translations:

◇**Zero-Shot:** a verbalized description of the task is provided without any demonstration.

◇**Direct:** the same verbalized instruction is provided along with demonstrations that include the English source sentence and the gender-inclusive Italian translation.

◇**Binary:** an intermediate gendered (masculine) Italian translation is also included in the format. This format follows the one used in Savoldi et al. (2024), which frames the task as a double output translation. The models are asked to produce a gendered translation first and then a second one with neomorphemes, which should be identical to the first except for the words expressing gender.

◇**Ternary:** two intermediate gendered translations (one masculine, one feminine) are included. The rationale for this format is that, by instantiating a ternary opposition, the models may better identify parts of the target language sentences that should be identical among the three transla-

[9] Model `gpt-4-0125-preview`

[10] `https://huggingface.co/Helsinki-NLP/opus-mt-en-it`

[11] `https://italianoinclusivo.it/scrittura/`

PROMPT	ROLE	EXAMPLE
Zero-shot	`user`	Translate the following English sentence into Italian using the neomorpheme '*'. To do so, the neomorpheme '*' should be used as a substitute for masculine and feminine morphemes in words that refer to human beings. [English] <{`input sentence`}> [Italian]
Direct	`user`	[English] <I never buy flowers for my friends.> [Italian]
	`assistant`	<Non compro mai fiori per l* mi* amic*.>
Binary	`user`	[English] <I never buy flowers for my friends.> [Italian, gendered]
	`assistant`	<Non compro mai fiori per i miei amici.> [Italian, neomorpheme] <Non compro mai fiori per l* mi* amic*.>
Ternary	`user`	[English] <I never buy flowers for my friends.> [Italian, masculine]
	`assistant`	<Non compro mai fiori per i miei amici.> [Italian, feminine] <Non compro mai fiori per le mie amiche.> [Italian, neomorpheme] <Non compro mai fiori per l* mi* amic*.>

Table 5: Examples of all the prompts used in our experiments. The few-shots prompt examples include the Asterisk neomorpheme. Words expressing gender are highlighted.

tions and the parts that should differ, i.e. those expressing gender. Framing the task as a triple output translation could help the models infer that the gender expressed in the third translation should be something other than masculine or feminine.

We enclose the exemplar sentences in angle brackets <>. Models are expected to reproduce this structure, thus facilitating the extraction of the final translation from the output in postprocessing.

All four models expect prompts in a 'chat' format, with `user` messages providing input and `assistant` messages representing the model's desired output.[12] For the few-shot prompts we adhere to this structure, whereas for the zero-shot prompts, we only provide a single `user` message.

Demonstrations In the few-shot settings (i.e. Direct, Binary, Ternary) we included **1**, **4**, and **8** task demonstrations in the prompts. The extremes were chosen as the minimum necessary to elicit in-context learning (1) and a compromise between a high number of demonstrations and the computational cost of inference (8). The exemplar sentences were selected from NEO-GATE's dev set (§3.1). The exemplars were chosen so as to represent the average tag *density* of the dev set, i.e., the number of tags in each reference, and to offer a balanced mix of singular and plural tags. The prompts were then formatted using each paradigm's tagset mapping before presenting them to the model.

[12]`https://huggingface.co/docs/transformers/main/en/chat_templating`

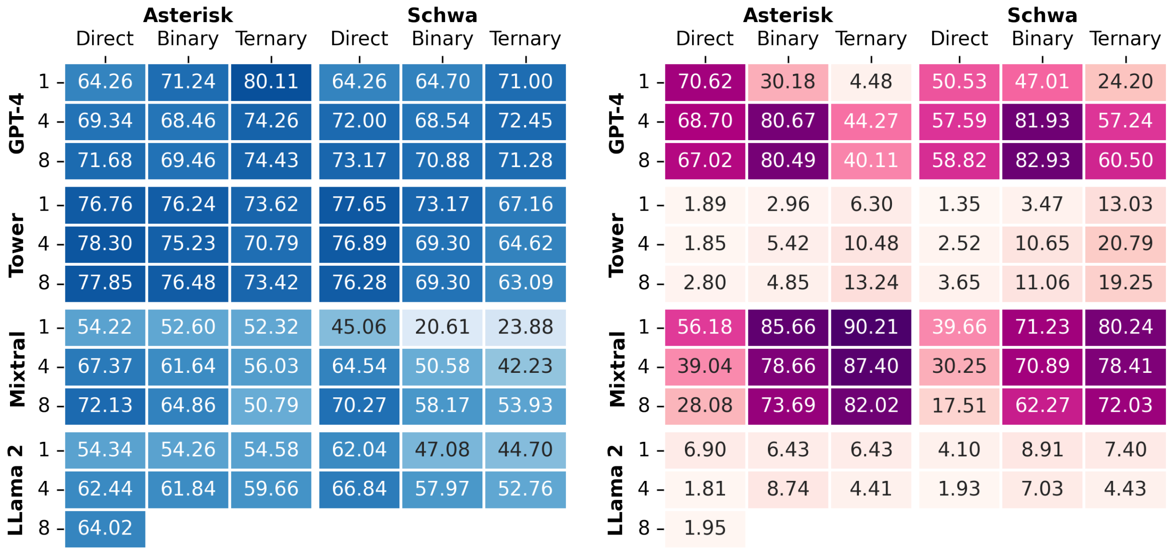
5 Results and discussion

ASTERISK	COV ↑	ACC ↑	CWA ↑	MIS ↓
GPT-4	57.08	**74.63**	**42.60**	45.78
Tower	**77.57**	0.00	0.00	**0.00**
Mixtral	35.22	37.92	13.35	52.20
LLama 2	56.72	0.57	0.32	16.70
SCHWA	COV ↑	ACC ↑	CWA ↑	MIS ↓
GPT-4	46.91	**60.19**	**28.24**	72.77
Tower	**77.25**	0.00	0.00	**0.00**
Mixtral	30.05	27.79	8.35	61.44
LLama 2	57.60	0.35	0.20	12.79

Table 6: Zero-shot setting results. We report the coverage (COV), accuracy (ACC), coverage-weighted accuracy (CWA), and mis-generation (MIS) scores.

5.1 Zero-shot results

The results of our zero-shot experiments are reported in Table 6, which unveils very different model behaviors. On the one hand, GPT-4 and Mixtral achieve significantly higher accuracy scores compared to the other two models, with GPT doubling Mixtral's performance. The accuracy scores indicate that, out of the matched terms, GPT correctly generated 74.63% Asterisk neomorphemes and 60.19% Schwa neomorphemes, with Mixtral reaching 37.92% and 27.79% respectively. Accounting for coverage, the gap widens further, with GPT's coverage-weighted accuracy amounting to more than three times that of Mixtral (42.60 and 28.24, vs 13.35 and 8.35). Both models tend to



(a) Coverage percentage scores.

(b) Accuracy percentage scores.

Figure 1: Coverage and accuracy results in the few-shot settings. Darker shades indicate better performance.

produce considerable amounts of mis-generations, which are most often higher than the respective coverage scores. This implies that **GPT and Mixtral generate plenty of neomorphemes, but they use them incorrectly for the most part**. Regardless, according to all the metrics, both models perform better with the Asterisk paradigm rather than the Schwa, possibly due to the latter's use of distinct singular (ə) and plural (ɜ) forms, adding a further challenge to the task.

On the other hand, **LLama 2 and Tower severely under-generate neomorphemes**, regardless of the paradigm. More specifically, LLama's near zero accuracy scores (0.57 and 0.35) paired with its low mis-generation scores (16.70 and 12.79) indicate that LLama 2 rarely generates neomorphemes and, when it does, it uses them inaccurately. Finally, Tower's high coverage scores (77.57 and 77.25) combined with the rest of the metrics, all of which report 0 scores, indicate that the model produces fluent, gendered outputs and never generates neomorphemes in the zero-shot setting. This can be due to the fact that in Tower-instruct's fine-tuning data set, TowerBlocks,[13] our neomorpheme characters are practically absent (3 occurrences of 'ə' in English segments, and no occurrences at all of 'ɜ' and '*'). However, since the development data of the other two models is not publicly available, we cannot further investigate this hypothesis and draw definitive conclusions.

5.2 Few-shots experiments

For the few-shot experiments we report each of the four metrics separately. We do not report all LLama 2 scores because in some cases, namely all the 8-shots settings, the model struggled to reproduce the format described in §4. In such instances, LLama 2 failed to insert the angle brackets or the labels we included in our prompts, and its outputs contained too many hallucinations to be automatically post-processed and evaluated. As the model did not seem to yield better performance or exhibit interesting phenomena in 8-shot settings, the additional effort required to process its unpredictable outputs was unjustified. Therefore, we only report the scores of one of the 8-shots settings outputs, namely the Asterisk, Direct prompt setting.

5.2.1 Coverage and accuracy

The coverage and accuracy scores are reported in Figure 1. **Looking at coverage (1a), we observe that few-shot prompting generally leads to improvements compared to the zero-shot results. Also, for Mixtral and LLama, the scores increase at higher numbers of demonstrations.** As for the prompts, the Direct format generally produces higher coverage scores, with only GPT performing better with the Ternary format. Interestingly, the neomorpheme paradigm has an impact on coverage, as we see generally higher scores with the Asterisk paradigm compared to the

[13]https://huggingface.co/datasets/Unbabel/TowerBlocks-v0.1/

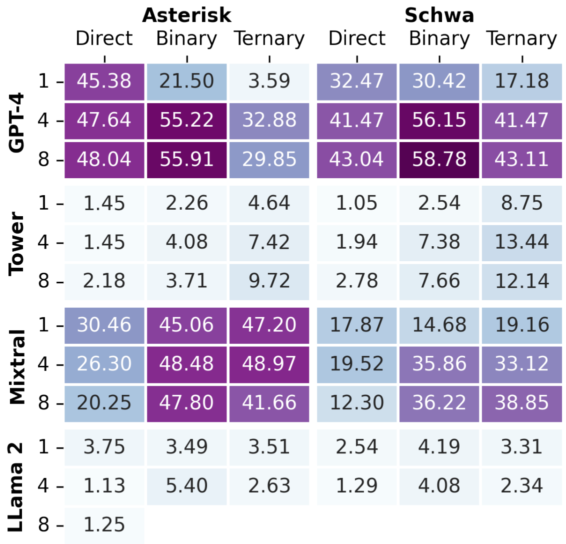


Figure 2: Coverage-weighted accuracy percentage scores for the few-shot settings. Darker shades indicate better performance.

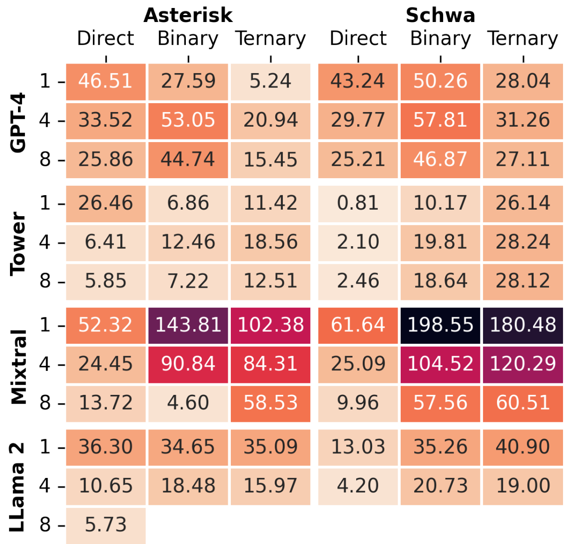


Figure 3: Mis-generation percentage scores for the few-shot settings. Higher scores (darker shades) indicate worse performance.

Schwa. As discussed in §5.2.3, this can be ascribed to the models' tendency to produce more mis-generations with the latter.

Coverage, however, is only informative of the proportion of annotated terms the models generated and disregards how many of those words include neomorphemes. **Looking at accuracy (1b) we find that all models improve their performance in at least one setting** compared to the zero-shot experiments, confirming the benefits of in-context learning for generative tasks involving neologistic expressions (Hossain et al., 2023). Mixtral and GPT are confirmed as the models which produce the highest rates of correct neomorphemes, with the first topping at 90.21 and the latter at 82.93 accuracy. On the contrary, Tower and LLama 2 are unfit for this task despite their improvements, as their scores remain low.

Surprisingly, **a greater number of demonstrations does not necessarily lead to higher accuracy.** While coverage generally increased with more demonstrations, this trend generally holds true for accuracy only for GPT and Tower, indicating that they generate more neomorphemes and do so more accurately. On the contrary, the accuracy of LLama 2 and Mixtral significantly decreases with more demonstrations. Paired with their rising coverage, this indicates that they produce fewer neomorphemes and more gendered terms. Both behaviors may result from systems better modeling the task with more demonstrations, as LLama's and Mixtral's higher accuracy and lower coverage in the 1 shot settings may be due to fortuitous correct generations of neomorphemes in a context where they over-generate them. We discuss this aspect below, looking at mis-generation (§5.2.3).

As for the neomorpheme paradigms, Mixtral and Tower perform better with the Asterisk and the Schwa respectively, as in the zero-shot experiments. GPT does not seem to be consistently affected by the neomorpheme paradigm. LLama presents negligible differences between the paradigms as well. The ability to more correctly generate one neomorpheme over another is possibly due to models' robustness to likely unseen grammatical paradigms and to the representations of the specific characters used as neomorphemes in each model's training data. Unfortunately, we cannot investigate this aspect as such data is not publicly available, with the exception of Tower's fine-tuning data set, as mentioned in §5.1.

5.2.2 Coverage-weighted accuracy

To compare models' overall performance in gender-inclusive MT, we look at coverage-weighted accuracy in Figure 2. This metric offers a comprehensive view of model performance in each setting, allowing for a fair comparison of different systems in light of both coverage and accuracy.

We first find that **all models improve their performance in the few-shots experiments**. The upside offered by in-context learning is notable, and there is arguably room for improvement at higher numbers of demonstrations. GPT and Mixtral are confirmed as the best models, and the gap between them narrows significantly with respect to the zero-shot experiments. In the best configurations, GPT scores 58.78 (Schwa, Binary, 8 shots) and Mixtral scores 48.97 (Asterisk, Ternary, 4 shots). Generally, GPT performs better with the Binary prompt and 4 or 8 shots, whereas Mixtral achieves its best results in the Binary/Ternary, 4/8 shots region, especially with the Asterisk paradigm. As for the other two models, despite the very low scores Tower generally outperforms the ten times bigger LLama 2, but both come across as unfit for this task.

5.2.3 Mis-generation

So far, the discussion of our few-shots experiments has focused on the correct generation of neomorphemes when referring to human entities, thus only on relevant phenomena we annotated in our test set. To better investigate models' behavior, we look at mis-generation, i.e. inappropriate neomorphemes generations, as well (Figure 3).

We first note that **Mixtral stands out as the model producing the most mis-generations**, especially in the 1 and 4 shots, Binary and Ternary region. Table 7 reports examples of mis-generation from Mixtral's outputs.

Source	I hope the shaman can help us.
Annotation	lo la lə; sciamano sciamana sciamanə;
Output	**Sperə** che lə sciamanə possa aiutarci.
Source	They asked everyone to remain silent.
Annotation	tutti tutte tuttз;
Output	Hanno chiesto a **tз** di **rimanerə** in **silenziə**.

Table 7: Examples of mis-generation found in Mixtral's Schwa, 1 shot, Binary prompt outputs. Words containing neomorphemes are underlined, mis-generations are in bold.

As hypothesized in §5.2.1, in these settings Mixtral over-generates neomorphemes, resulting in both correct generations and mis-generations. This behavior is reflected in the high accuracy and low coverage: by over-generating neomorphemes, Mixtral produces fewer gendered words – which would contribute to coverage – and many words that are either *a)* correct or *b)* mis-generations. With more task demonstrations, Mixtral generates significantly fewer mis-generations, and while its accuracy decreases the coverage improves, meaning that it produces better formed outputs. Mixtral's example testifies to how the mis-generation metric complements the analysis of models' behavior, as it sheds light on unwanted phenomena related to neomorphemes usage, which coverage and accuracy alone (or combined) cannot signal.

Similarly to Mixtral, LLama 2 produces more mis-generations given fewer demonstrations and progressively mitigates this behavior when given more. With an opposite trend, GPT generates fewer mis-generations, and Tower even less. However, the best performing settings of both models are also the ones in which they produce the most mis-generations. Hence, future work should focus on improving the ratio of correctly generated neomorphemes over the total neomorphemes generated by these models.

6 Conclusions

We discussed a neologistic approach to gender-inclusive machine translation, an underexplored area constrained by the lack of publicly available dedicated data. Our first contribution, the release of the NEO-GATE benchmark, allowed us to give a first fundamental impulse to research in this direction. As a second contribution, we explored the possibility of performing gender-inclusive translation from English to Italian with four popular Large Language Models: three open models – Mixtral, Tower, and LLama 2 – and a commercial one – GPT-4. Our comparisons across different prompting settings reveal that GPT-4 and Mixtral generally exhibit promising results when properly prompted, while LLama 2 and Tower are unfit for the task. More specifically, models' understanding of the task is significantly influenced by prompt complexity, the number of demonstrations, and the specific characters employed as neomorphemes (possibly depending on the representation of those characters in each model's training data).

While our investigation suggests LLMs' potential for neologistic gender-inclusive MT, there remains room for improving their accuracy. NEO-GATE and the analyses presented herein lay the groundwork for rising to the challenge and for future research on gender-inclusive MT tailored to existing neologistic paradigms, and those that may emerge in this new and evolving landscape.

7 Acknowledgements

We acknowledge the support of the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU.

References

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

Attig, Remy and Ártemis López. 2020. Queer Community Input in Gender-Inclusive Translations. *Linguistic Society of America [Blog]*, June 23.

Baiocco, Roberto, Fau Rosati, and Jessica Pistella. 2023. Italian proposal for non-binary and inclusive language: The schwa as a non-gender–specific ending. *Journal of Gay & Lesbian Mental Health*, 27(3):248–253, July. Publisher: Routledge _eprint: https://doi.org/10.1080/19359705.2023.2183537.

Bentivogli, Luisa, Beatrice Savoldi, et al. 2020. Gender in Danger? Evaluating Speech Translation Technology on the MuST-SHE Corpus. In *58th ACL*, pages 6923–6933.

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

Bradley, Evan D., Julia Salkind, Ally Moore, and Sofi Teitsort. 2019. Singular 'they' and novel pronouns: Gender-neutral, nonbinary, or both? *Proceedings of the Linguistic Society of America*, 4:36:1–7, March.

Brandl, Stephanie, Ruixiang Cui, and Anders Søgaard. 2022. How conservative are language models? adapting to the introduction of gender-neutral pronouns. In Carpuat, Marine, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3624–3630, Seattle, United States, July. Association for Computational Linguistics.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

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

Currey, Anna, Maria Nadejde, Raghavendra Reddy Pappagari, Mia Mayer, Stanislas Lauly, Xing Niu, Benjamin Hsu, and Georgiana Dinu. 2022. MT-GenEval: A counterfactual and contextual dataset for evaluating gender accuracy in machine translation. In Goldberg, Yoav, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4287–4299, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

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

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

Gaido, Marco, Beatrice Savoldi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. How to split: the effect of word segmentation on gender bias in speech translation. In Zong, Chengqing, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3576–3589, Online, August. Association for Computational Linguistics.

Gautam, Vasundara. 2021. Guest lecture in pronouns: Vagrant. `https://link.medium.com/viFawWyPVHb`. Accessed: Feb 20, 2024.

Goyal, Naman, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

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

Haralambous, Yannis and Joseph Dichy. 2018. Graphemic Methods for Gender-Neutral Writing. In *Graphemics in the 21st Century, Brest 2018*, volume Graphemics in the 21st Century: 2018 Conference, pages 41 – 89, Brest, France, Jun. Fluxus Editions.

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

Jiang, Albert Q, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Kaplan, Jennifer Marisa. 2022. Pluri-grammars for pluri-genders: Competing gender systems in the nominal morphology of non-binary french. *Languages*, 7(4).

Landis, J. Richard and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Lardelli, Manuel and Dagmar Gromann. 2023. Gender-fair post-editing: A case study beyond the binary. In Nurminen, Mary, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, Eva Vanmassenhove, Sergi Alvarez Vidal, Nora Aranberri, Mara Nunziatini, Carla Parra Escartín, Mikel Forcada, Maja Popovic, Carolina Scarton, and Helena Moniz, editors, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 251–260, Tampere, Finland, June. European Association for Machine Translation.

Lauscher, Anne, Archie Crowley, and Dirk Hovy. 2022. Welcome to the Modern World of Pronouns: Identity-Inclusive Natural Language Processing beyond Gender. In *29th COLING*, pages 1221–1232.

Lauscher, Anne, 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, July. Association for Computational Linguistics.

McGaughey, Sebastian. 2020. Understanding neopronouns. *The Gay & Lesbian Review Worldwide*, 27:27+.

Min, Sewon, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Misiek, Szymon. 2020. Misgendered in translation?: Genderqueerness in polish translations of english-language television series. *Anglica. An International Journal of English Studies*, pages 165–185.

Motschenbacher, Heiko. 2014. Grammatical gender as a challenge for language policy: The (im)possibility of non-heteronormative language use in German versus English. *Language policy*, 13(3):243–261.

Ovalle, Anaelia, Palash Goyal, Jwala Dhamala, Zachary Jaggers, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2023a. "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.

Ovalle, Anaelia, Ninareh Mehrabi, Palash Goyal, Jwala Dhamala, Kai-Wei Chang, Richard Zemel, Aram Galstyan, Yuval Pinter, and Rahul Gupta. 2023b. Are you talking to ['xem'] or ['x','em']? on tokenization and addressing misgendering in llms with pronoun tokenization parity. *arXiv preprint arXiv:2312.11779*.

Paolucci, Angela Balducci, Manuel Lardelli, and Dagmar Gromann. 2023. Gender-fair language in translation: A case study. In Vanmassenhove, Eva, Beatrice Savoldi, Luisa Bentivogli, Joke Daems, and Janiça Hackenbuchner, editors, *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 13–23, Tampere, Finland, June. European Association for Machine Translation.

Papineni, Kishore, Salim Roukos, et al. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *40th ACL*, pages 311–318.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peng, Keqin, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and

Dacheng Tao. 2023. Towards making the most of ChatGPT for machine translation. In Bouamor, Houda, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5622–5633, Singapore, December. Association for Computational Linguistics.

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

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

Piergentili, Andrea, 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 Bouamor, Houda, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14124–14140, Singapore, December. Association for Computational Linguistics.

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

Rarrick, Spencer, Ranjita Naik, Varun Mathur, Sundar Poudel, and Vishal Chowdhary. 2023. Gate: A challenge set for gender-ambiguous translation examples. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, page 845–854, New York, NY, USA. Association for Computing Machinery.

Rei, Ricardo, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702.

Rose, Ell, Max Winig, Jasper Nash, Kyra Roepke, and Kirby Conrod. 2023. Variation in acceptability of neologistic English pronouns. *Proceedings of the Linguistic Society of America*, 8(1):5526, April.

Rosola, Martina, Simona Frenda, Alessandra Teresa Cignarella, Matteo Pellegrini, Andrea Marra, Mara Floris, et al. 2023. Beyond obscuration and visibility: Thoughts on the different strategies of gender-fair language in italian. *CLiC-it 2023. Proceedings of the 9th Italian Conference on Computational Linguistics. Venice, Italy, November 30-December 2, 2023.*, 3596:1–10.

Sánchez, Eduardo, Pierre Andrews, Pontus Stenetorp, Mikel Artetxe, and Marta R Costa-jussà. 2023. Gender-specific machine translation with large language models. *arXiv preprint arXiv:2309.03175*.

Sarmiento, Miguel Angel. 2015. La e para la desexualización del género en beneficio de la motivación de ele en suecia : Revitalizando la propuesta de Álvaro garcía meseguer. In *La enseñanza de ELE centrada en el alumno :*, pages 863–872. ASELE.

Saunders, Danielle, Rosie Sallis, and Bill Byrne. 2020. Neural Machine Translation Doesn't Translate Gender Coreference Right Unless You Make It. In *2nd Workshop on Gender Bias in Natural Language Processing*, pages 35–43.

Savoldi, Beatrice, Marco Gaido, et al. 2021. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.

Savoldi, Beatrice, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2022. Under the morphosyntactic lens: A multifaceted evaluation of gender bias in speech translation. In Muresan, Smaranda, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1807–1824, Dublin, Ireland, May. Association for Computational Linguistics.

Savoldi, Beatrice, Andrea Piergentili, Dennis Fucci, Matteo Negri, and Luisa Bentivogli. 2024. A prompt response to the demand for automatic gender-neutral translation. In Graham, Yvette and Matthew Purver, editors, *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, March. Association for Computational Linguistics.

Sendén, Marie Gustafsson, Emma Renström, and Anna Lindqvist. 2021. Pronouns beyond the binary: The change of attitudes and use over time. *Gender & Society*, 35(4):588–615.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In Erk, Katrin and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August. Association for Computational Linguistics.

Snover, Matthew, Bonnie Dorr, et al. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *7th AMTA*, pages 223–231.

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

Thornton, Anna Maria. 2020. Genere e igiene verbale: l'uso di forme con əin italiano | annali del dipartimento di studi letterari, linguistici e comparati. sezione linguistica. *Annali Del Dipartimento Di Studi Letterari, Linguistici E Comparati. Sezione Linguistica*, 11:11–54.

Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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

Waldendorf, Anica. 2023. Words of change: The increase of gender-inclusive language in German media. *European Sociological Review*, September.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

A Tagset and annotation

Table 8 reports the complete tagset used in NEO-GATE, as well as the tagset mappings for the Schwa and the Asterisk paradigms.

B Function words anchoring

We include an additional information for function words which maps them to an anchor, in respect to which they are expected to be correctly positioned. This check allows for a more precise evaluation of function words, as it ensures that the evaluation is performed on the appropriate function word, and not on other ones which may occur in the sentence.

An anchor consists in the longest possible subword common to the masculine, feminine, and tagged content word which the function word is syntactically linked to. Looking at Table 9, the first Annotation reports an example of anchor for a function word: 'student=1'. It indicates that the sub-word sequence 'student' is the anchor for the function word forms 'il la l*', meaning that if one of the three forms is found it will only be evaluated if the anchor is found immediately after it (i.e., at a distance of 1 word). Similarly, the second annotation of the table reports anchor annotations for two function words. The first, 'amic=2' indicates that if one of the three forms 'i le l*' is found, it will only be evaluated if the anchor 'amic' is found at a distance of two words. The second anchor annotation 'amic=1' maps the function word forms 'tuoi tue tu*' to the same anchor 'amic', which should be positioned one word after.

We did not include anchor annotations in the main body to simplify the examples. However, all function words annotated in NEO-GATE are assigned with anchors, including the ones reported in the examples throughout the paper.

C Translation experiments prompt

Table 10 reports the prompt we used to assess the general translation quality of the systems, as discussed in §4.1. We include three demonstrations taken from FLORES' dev set, so as to provide the LLMs with an interaction structure to reproduce. This facilitates the process of filtering out extra comments and hallucination produced by the models, and extract the output translation.

TAG	Description	Masculine	Feminine	Asterisk	Schwa
<ENDS>	inflectional morpheme (word ending), singular	o, e, tore	a, essa, trice	*	ə
<ENDP>	inflectional morpheme (word ending), plural	i, tori	e, esse, trici	*	ɜ
<DARTS>	definite article, singular	il, lo, l’	la, l’	l*	lə
<DARTP>	definite article, plural	i, gli	le	l*	lɜ
<IART>	indefinite article	uno, un	una, un’	un*	unə
<PARTP>	partitive article, plural	dei, degli	delle	de*	deɜ
<PREPdiS>	articulated preposition with root ‘di’, singular	del, dello, dell’	della, dell’	dell*	dellə
<PREPdiP>	articulated preposition with root ‘di’, plural	dei, degli	delle	dell*	dellɜ
<PREPaS>	articulated preposition with root ‘a’, singular	al, allo, all’	alla, all’	all*	allə
<PREPaP>	articulated preposition with root ‘a’, plural	agli, ai	alle	all*	allɜ
<PREPdaS>	articulated preposition with root ‘da’, singular	dal, dallo, dall’	dalla, dall’	dall*	dallə
<PREPdaP>	articulated preposition with root ‘da’, plural	dagli	dalle	dall*	dallɜ
<PREPinP>	articulated preposition with root ‘in’, plural	negli	nelle	nell*	nellɜ
<PREPsuS>	articulated preposition with root ‘su’, singular	sul, sullo, sull’	sulla, sull’	sull*	sullə
<PREPsuP>	articulated preposition with root ‘su’, plural	sugli	sulle	sull*	sullɜ
<DADJquelS>	demonstrative adjective (far), singular	quel, quello, quell’	quella, quell’	quell*	quellə
<DADJquelP>	demonstrative adjective (far), plural	quegli	quelle	quell*	quellɜ
<DADJquestS>	demonstrative adjective (near), singular	questo, quest’	questa, quest’	quest*	questə
<DADJquestP>	demonstrative adjective (near), plural	questi	queste	quest*	questɜ
<POSS1S>	possessive adjective, 1st person singular, singular	mio	mia	mi*	miə
<POSS1P>	possessive adjective, 1st person singular, plural	miei	mie	mi*	miɜ
<POSS2S>	possessive adjective, 2nd person singular, singular	tuo	tua	tu*	tuə
<POSS2P>	possessive adjective, 2nd person singular, plural	tuoi	tue	tu*	tuɜ
<POSS3S>	possessive adjective, 3rd person singular, singular	suo	sua	su*	suə
<POSS3P>	possessive adjective, 3rd person singular, plural	suoi	sue	su*	suɜ
<POSS4S>	possessive adjective, 1st person plural, singular	nostro	nostra	nostr*	nostrə
<POSS4P>	possessive adjective, 1st person plural, plural	nostri	nostre	nostr*	nostrɜ
<PRONDOBJS>	direct object pronoun, singular	lo	la	l*	lə
<PRONDOBJP>	direct object pronoun, plural	li	le	l*	lɜ

Table 8: The full tagset used in NEO-GATE and the tagset mappings to the Italian gendered forms and the desired forms in the Asterisk and Schwa nomorpheme paradigms.

Source	The student was worried about going off topic.
Tagged reference	L* student* era preoccupat* di andare fuori tema.
Annotation	lo la l* **student=1**; studente studentessa student*; preoccupato preoccupata preoccupat*;
Source	Come out to the balcony and let your friends see you.
Tagged reference	Esci sul balcone e lascia che l* tu* amic* ti vedano.
Annotation	i le l* **amic=2**; tuoi tue tu* **amic=1**; amici amiche amic*;

Table 9: Examples of NEO-GATE annotations including anchors. The Tagged references and the Annotations are adapted to the Asterisk paradigm, the anchors are in bold.

Role	Content
`user`	Translate the following English sentence into Italian: <Construction is ongoing for five new skyscrapers at the site, with a transportation center and memorial park in the middle.>
`assistant`	<Nel sito sono in corso i lavori di costruzione di cinque nuovi grattacieli, con un centro trasporti e un parco alla memoria nel mezzo.>
`user`	<The game is based on the Second Battle of Fallujah, a vicious battle between American and Iraqi forces.>
`assistant`	<Il videogioco simula la seconda battaglia di Fallujah, un conflitto cruento tra le forze armate americane e quelle irachene.>
`user`	<Born in the Croatian capital, Zagreb, Bobek gained fame while playing for Partizan Belgrade.>
`assistant`	<Nato nella capitale croata Zagabria, Bobek è diventato famoso giocando nel Partizan Belgrado.>
`user`	`<{input sentence}>`

Table 10: The 3 shots prompt used in the general translation preliminary experiments.