The past and future of Fairslator

Michal Měchura
Centre for Translation Studies, University of Vienna
michal.mechura@univie.ac.at

Abstract

Fairslator is a web-based tool for detecting and correcting bias in machine translation. Started in 2022 as a personal project, Fairslator has recently (2026) received backing from University of Vienna where it is going to be redeveloped into an open-source, community-contributed tool for rewriting and computer-assisted postediting of machine translation. This contribution introduces the plan for that redevelopment.

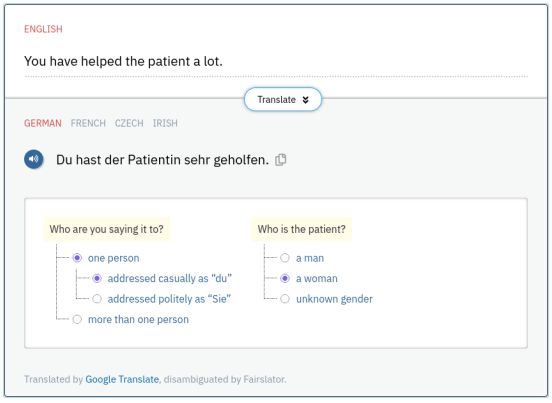

Figure 1: Fairslator's user interface

1 Introduction

Fairslator[1] (Měchura, 2022) is a plug-in for any machine translation system which detects ambiguities in the source text and invites the user to disambiguate them manually through a user interface shown in Figure 1. Based on the user's choices Fairslator rewrites the translation into a different gender or form of address. Fairslator currently works in four directed language pairs from English into German, French, Czech and Irish.

Following Fairslator's recent redefinition as a project supported by University of Vienna, Fairslator is about to be rewritten as an open-source tool for detecting and rewriting biased translations, in potentially any language pair. The goal is to build a tool capable of handling – with reliable accuracy – all situations when the machine translator has made an arbitrary, unjustified decision about the semantic or pragmatic properties of (mainly human) referents, including but not limited to gender.

A secondary goal is to automate repetitive postediting tasks. In most machine translation postediting environments today, once the user has decided to change something (such as the gender of some human referent), they need to make the edit manually. Fairlator's pregenerated rewrites, which the user will be able to accept or reject with a single click, will remove the mundane manual work while still leaving the human in control of all decisions.

2 Software architecture

The current version of Fairslator is only a proof of concept: it is the result of years of experimentation, is built on technologies not completely suitable for the purpose (C#, XML) and is ridden with technical dept. It has been decided that Fairslator needs to be rewritten from scratch if it is to become genuinely useful and long-term maintainable. The new version will be written in Python and the spaCy[2] NLP library. Some functionality will optionally be outsourcable to a Large Language Model or to another NLP library called Grammatical Framework[3].

[1] https://www.fairslator.com/

[2] https://spacy.io/

[3] https://www.grammaticalframework.org/

Fairslator will take the output of any machine translator (the source-translation pair) as its input and process it along a pipeline with three steps: **1. Preparation:** parsing the source text and the translation, recognising named entities, resolving coreferences, bilingual alignment. **2. Detection:** matching the source-translation pair against a catalogue of patterns which are known to betray biased translations, and subsequently generating human-readable disambiguation questions and suggested rewrites. **3. Rewriting:** altering the translation in accordance with the user's choices.

Fairslator will be an open-source software library and will be deliberately designed in such a way that others can contribute modules for specific languages and language pairs.

3 Related work

Machine translation vendors including Google[4], Microsoft[5], DeepL[6], Amazon[7] and Apple (Garg et al., 2024) have in recent years introduced features for rewriting translations into different genders and forms of address. In addition to those, various language-specific rewriters are being developed in academia for eg. Arabic (Alhafni et al., 2022), Hebrew (Moryossef et al., 2019) and Greek[8]. Last but not least, several vendors of Translation Management Systems (TMS) such as XTM[9], Intento[10] and TextUnited[11] have been unveiling features in their products with names such as "intelligent postediting" and "automatic postediting" which aim to automate repetitive postediting tasks.

All of these initiatives overlap with the Fairslator project, but Fairslator is not necessarily a competitor to any of them. Unlike other projects which are either closed-source or language-specific, Fairslator will be an open-source, language-agnostic platform where researchers can experiment with different approaches and test the limits of what is technically possible.

4 Conclusion

The plan for Fairslator's redevelopment presented here is a general guide to how the Fairslator project is likely to develop in the future. Faislator's long-term mission is to make postediting less boring for professional posteditors, and less arbitrarily biased for all.

References

Alhafni, Bashar, Ossama Obeid, and Nizar Habash. 2022. The User-Aware Arabic Gender Rewriter. https://arxiv.org/abs/2210.07538.

Garg, Sarthak, Mozhdeh Gheini, Clara Emmanuel, Tatiana Likhomanenko, Qin Gao, and Matthias Paulik. 2024. Generating Gender Alternatives in Machine Translation. In Faleńska, Agnieszka, Christine Basta, Marta Costa-jussà, Seraphina Goldfarb-Tarrant, and Debora Nozza, editors, *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 237–254, Bangkok, Thailand, August. Association for Computational Linguistics.

Moryossef, Amit, Roee Aharoni, and Yoav Goldberg. 2019. Filling Gender & Number Gaps in Neural Machine Translation with Black-box Context Injection. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 49–54, Florence, Italy. Association for Computational Linguistics.

Měchura, Michal. 2022. Introducing Fairslator: a machine translation bias removal tool. In *Translating and the Computer 44 Proceedings*, pages 90–95, Luxembourg, November. Editions Tradulex.

[4] https://research.google/blog/a-scalable-approach-to-reducing-gender-bias-in-google-translate/

[5] https://www.microsoft.com/en-us/translator/blog/2023/03/08/bings-gendered-translations-tackle-bias-in-translation/

[6] https://support.deepl.com/hc/en-us/articles/4406432463762-About-the-formality-feature

[7] https://docs.aws.amazon.com/translate/latest/dg/customizing-translations-formality.html

[8] https://lt3.ugent.be/projects/gender-fair-rewriter-for-english-to-greek-machine/

[9] https://multilingual.com/xtm-launches-post-editing-ai-driven-review-automation-localization/

[10] https://inten.to/enterprise-language-hub/

[11] https://textunited.com/en/blog/automatic-post-editing-vs-human-post-editing

Dutch audience perceptions of human and machine-translated pronouns for non-binary referent in subtitles

Joke Daems* **Cynthia Van Hee*** **Alicia Van Muylem**
Ghent University, Ghent, Belgium
firstname.lastname@ugent.be

Abstract

The increased representation of non-binary characters in audiovisual media is societally crucial, yet non-binary language (e.g., pronouns) offers new challenges for translators. In Dutch, the acceptability of pronoun strategies for non-binary reference is debated and evolving. Machine translation (MT) is increasingly being used in audiovisual translation, yet its potential for translation in gender-sensitive contexts is understudied. In this paper, we compare three pronoun translation strategies (two human-produced, one MT) for rendering the English non-binary pronoun *they* into Dutch in an audiovisual context, using a fragment from the Netflix show *Sex Education*. We explore general acceptability of these strategies and compare this to audience perceptions after viewing subtitled fragments employing a specific strategy. The results support findings from earlier work on non-binary pronouns in Dutch and indicate that the MT-generated translations were perceived as the least suitable.

1 Introduction

Contemporary linguistic and social developments reflect a growing visibility and recognition of gender diversity, including non-binary and genderqueer identities. Large-scale surveys indicate that gender identities different from the male/female binary are increasingly prevalent, especially among younger generations, with on average 9% of adults (over 26 countries) identifying as LGBT+[1], rising to 14% among Gen Z (Ipsos, 2025). Similar patterns are observed in Belgium (10%). According to an EU-wide survey, 20% of LGBT+ respondents identify as non-binary (for Fundamental Rights, 2024). These figures reflect a growing population for whom linguistic and cultural representation is important.

In parallel, LGBT+ visibility in media has expanded over the past years, driven by both audience demand and shifts in media production. Surveys show that 29% of respondents across 26 countries are in favor of more LGBT+ characters on TV, in films and in advertising (Ipsos, 2025). Beyond representation, such visibility has been associated with reduced prejudice and increased cultural exchange, particularly through globally distributed streaming content (Yuan et al., 2025). As a consequence, accurate and sensitive representation of genderqueer characters has become increasingly relevant, not only for visibility but also for viewer identification and social recognition.

This goal of sensitive gender representation is seemingly at odds with the increased use of automated subtitling and machine translation for audiovisual translation (AVT) (Karakanta, 2022), as gender bias is still an important bottleneck for MT systems (Savoldi et al., 2025). Despite growing attention to both LGBT+ representation and machine translation in AVT, the intersection of the two domains remains underexplored, especially in the Dutch context. The one study to explore MT gender bias for Dutch revealed that non-binary pronouns like *their* are often mistranslated by MT

*These authors contributed equally to this work.

[1]Including people who identify as lesbian, gay, homosexual, bisexual, pansexual, omnisexual, asexual, transgender, non-binary, gender nonconforming, gender-fluid, other than male or female.

systems (Hackenbuchner et al., 2025). To date, reception studies on non-binary pronouns in Dutch have been limited to newspaper articles (Decock et al., 2024; Vriesendorp, 2024) and literary texts (Vos and Nutters, 2022), and empirical research on audience perception in AVT in general remains scarce (Karakanta, 2022; Guerberof-Arenas et al., 2024).

Accordingly, this study aims to investigate how non-binary pronouns in the Netflix series *Sex Education* (Laurie Nunn, 2019) are rendered in English-to-Dutch subtitles and what the impact is of human and machine-translated pronouns on viewer reception. To this end, this paper aims to answer the following main research questions:

- Which pronouns are seen as acceptable non-binary pronouns in Dutch? (**RQ1**)
- What is the impact of different non-binary pronoun translation strategies (human and MT) on audience perceptions of Dutch subtitles? (**RQ2 & RQ3**)

In the first analysis, we explore whether demographic variables have an impact on the perceived acceptability of pronoun strategies. In the second analysis, we focus on three aspects (comprehensibility, acceptability and overall quality) of the subtitles as a whole (RQ2), as well as the perceived suitability of each strategy for non-binary reference (RQ3).

Bias Statement We focus on the potential representational harm (Blodgett et al., 2020) caused by MT output in audiovisual translation. Specifically, we focus on the MT mistranslation of the English non-binary pronoun *they* into a plural pronoun in Dutch. This is a form of "linguistic erasure of non-binary and transgender identities who might choose the singular pronoun *they*" (Ghosh and Caliskan, 2023, p. 906), at a time when affirmation of gender-diversity is key, especially for audiovisual translation (Lardelli et al., 2025).

2 Related Research

2.1 Gender-inclusive audiovisual translation

GLAAD's annual Where We Are on TV report documents 489 LGBT+ characters across U.S. television in 2024–2025 (GLAAD, 2024), marking a substantial increase from the 46 characters recorded in the first edition (2006), with streaming platforms identified as the primary driver of this growth. The expansion of global streaming has fundamentally transformed audiovisual translation practices. Commercial streaming platforms distribute content simultaneously across multiple territories, which requires fast and large-scale localization workflows that increasingly rely on machine translation and automated subtitling systems (Karakanta, 2022). While these technologies improve efficiency, they also introduce risks for linguistic and sociocultural fidelity. This is particularly evident in the translation of genderqueer language, where target languages may present linguistic constraints or lack established conventions for gender-sensitive pronoun usage. Moreover, perception studies suggest that fully automated subtitles without human intervention can negatively affect viewer comprehension, engagement and perceived quality (Guerberof-Arenas et al., 2024).

While singular *they* has been used in English since the 13th - 14th century as a gender-indefinite form, its more recent adoption as a marker of non-binary identity reflects evolving usage within queer communities (Konnelly and Cowper, 2020; Nabila et al., 2021). Researchers have argued that accurately representing the linguistic choices made by a character in another language is a form of social justice in translation, and that translators should be members of the queer community to portray such characters accurately in other languages (Attig, 2023). Translators have two main strategies to handle non-binarity: the use of Indirect Non-binary Language (INL), where gendered terms are avoided by using existing linguistic strategies (e.g., a specific word like *girl* can be replaced by the more general *child*, passives can be used to avoid gender marking), and Direct Non-binary Language (DNL), where innovative linguistic strategies and pronouns are used to make non-binary gender identities explicitly visible (López, 2022). While INL may be less controversial for mainstream audiences, it can obscure or erase non-binary identities, which runs against the goal of increasing the visibility of diverse identities (López, 2022). Research explored the representations of non-binary gender identities in audiovisual translation[2] and concluded that translation strategies are language-

[2]Both subtitles and dubbing were analyzed for *One Day at a Time*, *Sex Education*, *Heartbreak High*, and *Degrassi: Next Class*.

dependent and inconsistent. The analysis showed that non-binary representation in AVT often defaults to avoidance strategies or masculine generics, highlighting the "need for cross-linguistic collaboration and documentation of community-preferred terms" (Lardelli et al., 2025, p. 59).

These sociolinguistic complexities involved in gender-inclusive translation make it a particularly challenging task for machine translation (Lardelli and Gromann, 2025). The limited work on machine translation of non-binary and/or gender-neutral pronouns in general (not limited to AVT) shows that MT does not handle this task well. When translating the English pronoun *they/their* into languages with binary gendered pronouns, commercial MT systems were shown to mistranslate 35% of instances into binary masculine or feminine pronouns in French, German, Italian and Danish (Lauscher et al., 2023), and to mistranslate into plural and other grammatically inconsistent ways for Greek (Gkovedarou et al., 2025). Even when translating from English into Bengali, a language with gender-neutral pronouns, ChatGPT was shown to consistently mistranslate *they* into plural pronouns (Ghosh and Caliskan, 2023).

2.2 Gender-inclusive pronouns in Dutch

Dutch is a West Germanic language which originally had three grammatical genders (masculine/feminine/neuter), but is currently evolving to a common/neuter gender system. Research suggests that younger people are more likely to embrace this language change: there is evidence of an age effect in pronominal gender shift in Netherlandic Dutch (Doreleijers et al., 2025), and younger people are generally more aware of the existence of and the need of gender-neutral pronouns (Verhaegen et al., 2025). In contrast with English, where the pronoun *they* is now a familiar gender-inclusive pronoun, gender-neutral pronouns (for generic reference as well as for non-binary reference) are emerging in Dutch, but these are not formally standardized (Van Hoof and Decock, 2026). While Dutch as used in the Netherlands and in Flanders (Belgium) is largely mutually intelligible, regional variation exists in terminology related to gender identity (Audring, 2006; De Vos et al., 2021).

The gender-neutral pronouns currently most in use are *die/hen* (subject form), *die/hen* (non-subject form) and *diens/hun* (possessive form). These pronouns were not newly coined but already occurred naturally in the Dutch language as demonstrative or relative pronouns, or as the non-subject form of the plural pronoun *ze/zij*. Not all available pronouns for non-binary reference are equally accepted or used by Dutch-speaking people. Studies have shown regional variations as well as age and gender effects in terms of acceptability and ease of use (Doreleijers et al., 2025; De Vos et al., 2021). Most research to date suggests that people prefer to use the pronouns in the grammatical positions they already naturally occur in (e.g., *die* instead of *hen* in subject position), both for non-binary reference (Vriesendorp, 2024) and generic reference (Verhaegen et al., 2025). Although neither the combination *die/hen/hun* or *hen/hen/hun* when used to refer to a non-binary person was shown to impact text comprehensibility, *hen* in subject position was perceived as more awkward and led to a lower text appreciation, unless readers were made explicitly aware of the fact that the person was non-binary and used non-binary pronouns (Decock et al., 2024).

3 Methodology

This case study focuses on the pronouns used in reference to the non-binary character Cal in the Netflix show *Sex Education* and how these pronouns are translated to Dutch. The character appears in Seasons 3 and 4 and uses *they/them* pronouns in English. The series was selected for its inclusion of non-binary characters and the resulting frequent use of non-binary pronouns in naturalistic dialogue. Its popularity and critical recognition further motivated this choice, as the series' wide distribution ensures the availability of professional subtitles, which enables comparison between human translations and machine-generated output.

3.1 Pronoun Identification

To determine the pronoun forms included in the audience perception study, we first identified the range of non-binary English *they* forms in the source material and their corresponding Dutch translations in Netflix subtitles and produced by an MT system. We focused on instances of non-binary *they* in subject, object/prepositional, possessive and reflexive positions across both seasons (see Figure 1 for the overall distribution). Subsequently, we annotated the translation strategies used by the Dutch human translators. For subject position *they* (N = 19), two strategies were identi-

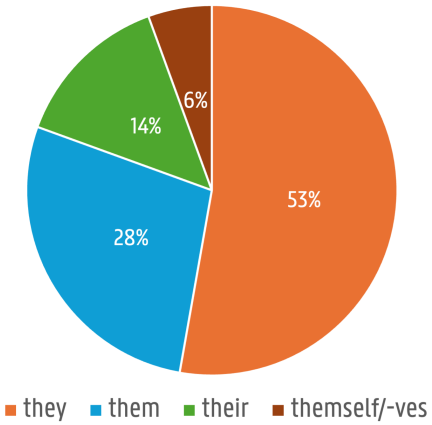

Figure 1: Distribution of English non-binary *they* forms in the source material across both seasons, distinguishing subject (*they*), object/prepositional (*them*), possessive (*their*) and reflexive (*themselves*) occurrences.

	they	*them*	*their*	*themself*	Strat.
NL - human translation					
hen	12	4	0	0	DNL
hun	0	4	3	0	
die	4	2	0	0	
zich(zelf)	0	0	0	2	INL
∅	3	0	2	0	
NL - machine translation					
hen	0	3	0	0	DNL
hun	0	0	4	0	
die	0	0	0	0	
zich(zelf)	0	0	0	2	INL
∅	0	1	0	0	
ze (plural)	19	6	1	0	MIS

Table 1: Distribution of Dutch renderings of English non-binary singular *they* in human and machine translations, categorized by translation strategy: direct non-binary language (DNL), indirect non-binary language (INL) and mistranslation (MIS).

fied: *die* and *hen*. For object/prepositional *them* (N = 10), the predominant translation was *hen*, with *hun* occurring in two instances. The possessive form *their* (N = 5) was consistently translated as *hun* and the reflexive form *themself* (N = 2) was rendered as *zichzelf*.

As a next step, the subtitles were translated using ModernMT Lite as integrated within Matecat[3]. Matecat was chosen over automatic subtitling tools (e.g., MateSub), as time coding and line segmentation were not central to our analysis and ASR-related issues were not within the scope of this study.

The results of the human translations (HT) of non-binary pronouns and their MT counterparts are presented in Table 1. The MT system predominantly translates the English non-binary pronouns *they*, *them* and *their* into plural forms in Dutch. This is reflected in the frequent rendering of subject (*they*) and object/prepositional (*them*) pronouns as the Dutch plural forms *ze* or *zij*, accompanied by plural verb forms. Of the ten object pronoun instances identified, three were translated as *hen*, which has been documented as a non-binary pronoun for Dutch, but is also the object form of the plural pronoun. Similarly, the possessive pronoun *their* was translated as *hun*, which - in Dutch - functions as the possessive form of the plural pronoun as well as a potential non-binary pronoun.

In sum, while the human translations of non-binary *they* predominantly (81%) result in Direct Non-binary Language (DNL), seven instances (19%) can be classified as Indirect Non-binary Language (INL). This category includes the reflexive pronoun *zich(zelf)*, which can refer to both singular and plural antecedents, as well as cases of rephrasing and pronoun omission. Our qualitative analysis of the MT output shows that the system systematically renders non-binary *they* as a plural pronoun - even in object and possessive forms, which is not appropriate in this context.

3.2 Audience Perception Study

The present study is guided by research questions focused primarily on audience reactions to subtitles containing non-binary references, without participants being aware of the translation modality (human versus MT). For this audience perception study, a fragment of approximately 90 seconds (54:38-56:04) was selected from *Sex Education* season 4, episode 6, during which Nicky, mother of the non-binary character (Cal), calls a radio show to explicitly ask for advice on their relationship.

This specific fragment was chosen as it is relatively stand-alone (even someone unfamiliar with the show can understand what is going on) and it contains ten instances of the non-binary pronoun in English (six in the subject position, three in the object/prepositional position, and one reflexive pronoun), making it suitable to observe audience responses to specific pronoun strategies. The fragment does not contain any possessive pronouns, but given that these were found to always be translated as *hun* (cf. Section 3.1), regardless of the origin (HT versus MT), we do not believe this im-

[3]MT output generated in April, 2025

pacts the results.

We created three subtitled versions of this fragment with different pronoun translation strategies: two based on the different potential human translation strategies (*hen/hen/hun* and *die/hen/hun*) and one based on the machine translation output (*ze/hen/hun* as plural pronouns with plural verbs). All other text was kept constant across conditions.

An online survey was created and distributed via Qualtrics (Provo, UT). The survey consisted of three main sections: one to collect relevant background information on participants (age, gender, language background, queerness, familiarity with non-binary people, attitude towards non-binary people), a second containing the video fragment and questions related to the subtitles (participants were randomly assigned one of the three subtitle versions and were asked to judge the overall comprehensibility and quality of the subtitles as well as the appropriateness and clarity of the specific pronoun strategy used to refer to the non-binary character), and a third to collect information on participants' overall familiarity with and attitude towards the different non-binary pronoun strategies for Dutch, as well as their attitude towards gender-inclusive language in general. The survey design was inspired by previous work on Dutch (Decock et al., 2024), with minor adjustments based on the specifics of the present study - see Appendix A for the relevant questions.

Participants were reached via convenience sampling: the survey was shared with the professional and personal network of the authors via social media channels (LinkedIn, BlueSky, Instagram, Facebook and WhatsApp). The survey ran from the 8th until the 16th of May, 2025.

4 Analysis

4.1 Participants and demographics

The survey received 124 individual responses, which were manually reviewed before inclusion in the experimental corpus. After filtering (i.e., removing incomplete submissions and responses with incorrect answers to the sanity check question), 77 responses were retained for further analysis. We mostly reached younger people: 28 participants were aged 18–24, 26 were aged 25–35, and 23 were aged 36 and above. Within the latter group, two participants were aged 36–45, nine were aged 46–55, ten were aged 56–65 and two were aged 65+.

Women were more strongly represented (N = 46) than other gender groups (23 men, six 'other', and two who preferred not to disclose their gender). Most participants (87%), regardless of gender or age group, were familiar with non-binary people, either through personal connections or media exposure. The majority were also aware of the existing Dutch pronoun strategies for gender-inclusive language: only five participants indicated that they did not know any such pronouns, while 61 were familiar with *die*, 60 with *hen*, and 22 with *ze*. In total, 27 participants identified as queer and two indicated that they were unsure. Queer participants in the sample were mostly younger (24 were aged 18-35) and were less likely to identify as male. Specifically, 18 queer participants identified as women, two as men, and seven as non-binary or preferred not to disclose their gender.

4.2 Acceptability of non-binary pronouns for Dutch (RQ1)

To address the first research question (*which pronouns are seen as acceptable non-binary pronouns in Dutch?*), two types of analyses were conducted. First, a Friedman test was used to assess the overall differences in ranking of *die*, *hen* and *zij*. Second, three Wilcoxon signed-rank tests were performed to examine pairwise differences between the pronouns. To investigate whether acceptability scores were influenced by additional factors, we subsequently built a series of linear mixed-effects models using the `lme4` (Bates et al., 2015) and `lmerTest` (Kuznetsova et al., 2017) packages in `RStudio` (R Core Team, 2024). In all models, acceptability score was included as the dependent variable, with participant ID specified as a random intercept. The main fixed effect was *strategy* (*die*, *hen*, *zij*). We then fitted separate models that each included one additional predictor (age, gender, queerness, familiarity with non-binary people, attitude towards non-binary people, and attitude towards gender-fair language), with and without an interaction effect, using the `anova()` function to compare nested models via likelihood ratio tests.

4.3 Audience perception (RQ2-3)

For the analyses addressing RQ2 and RQ3, which focus on audience responses to the subtitled fragment, additional corpus quality checks were performed. Responses were excluded if participants had indicated that they had not read the subtitles or if they had been unable to recall the English

and Dutch pronoun strategies used in the fragment they viewed. Of the 60 participants who correctly remembered the English strategy, only 46 also accurately recalled the Dutch pronoun strategy; only these responses were retained for further analysis. Of the 24 participants who viewed the *hen* subtitles, 23 correctly remembered this strategy, compared to 13 out of 19 for *die*, and 10 out of 17 for *ze*.

The analysis covers questions in which respondents evaluated (i) the overall quality of the subtitles in the fragment they viewed (RQ2) and (ii) the suitability of the translation of non-binary *they* (and its variants) in the subtitles (RQ3). These measures are hereafter referred to as *subtitle quality* and *translation strategy score*, respectively.

To enable statistical analysis and generalization of the responses, a series of one-way ANOVA tests was conducted. These analyses compared the mean scores assigned to the three Dutch variants - *die*, *hen* and *ze* - across three dimensions: **comprehensibility** (i.e., whether the subtitles are easy to understand and support comprehension of the dialogue), **acceptability** (i.e., whether the subtitles are perceived as appropriate and natural) and **quality** (i.e., whether the subtitles are considered high-quality translations, including how adequately non-binary identity is conveyed).

5 Results

5.1 Acceptability of non-binary pronouns for Dutch (RQ1)

Comparison	Friedman	Wilcoxon
Overall (*die/hen/zij*)*	$\chi^2 = 30.155$	n/a
	$df = 2$	n/a
	$p = 2.83\times10^{-7}$	n/a
die - hen	n/a	$T_+ = 543.5$
	n/a	$T_- = 359.5$
	n/a	$Z = -1.163$
	n/a	$p = 0.24$
*die - zij**	n/a	$T_+ = 1271$
	n/a	$T_- = 269$
	n/a	$Z = -4.22$
	n/a	$p = 2.45\times10_{-5}$
*zij - hen**	n/a	$T_+ = 176.5$
	n/a	$T_- = 1048.5$
	n/a	$Z = -4.369$
	n/a	$p = 1.25\times10^{-5}$

Table 2: Friedman test with post-hoc Wilcoxon comparisons. Comparisons showing highly significant ($p < 0.001$) differences are indicated with an asterisk (*).

Table 2 shows the results of both the Friedman test and the post-hoc Wilcoxon comparisons. The

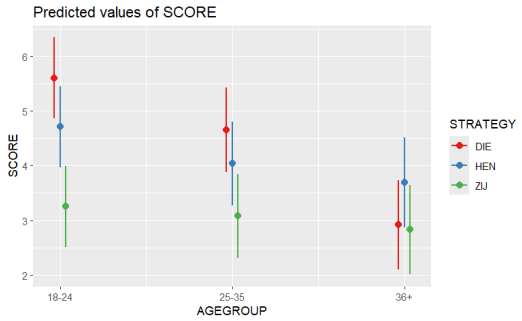

Figure 2: Interaction effect between translation strategy and age group for overall acceptability score.

Friedman test was based on 77 observations (N = 77). In 53 cases ($\approx$ 69%), at least two pronouns received identical rank scores, indicating limited variability in some responses. Despite this, the test revealed a statistically significant difference among *die*, *hen*, and *zij* ($p < 0.001$), suggesting that the pronouns were not ranked equally across conditions and that participants showed systematic preferences.

Three post-hoc Wilcoxon signed-rank tests further clarified these differences. Significant differences were found for the comparisons *die - zij* and *zij - hen*. The direction of the effects can be inferred from the T values: T_+ indicates that the first pronoun in the comparison tends to receive higher scores than the second. Based on this, *die* is ranked higher than both *hen* and *zij*, with a more pronounced difference for the *die–zij* comparison. In addition, *hen* is generally ranked higher than *zij*, a difference that is relatively strong (Z = -4.37).

The ranking results were confirmed and further supplemented by linear mixed-effects models. Compared to an intercept-only model, *strategy* was a significant predictor of the score ($\chi^2 = 29.94$, $p < .001$), with the *zij* strategy receiving significantly lower scores than *die* (an estimated reduction in score of 1.41 +/- 0.26 standard errors). Individually adding any of the other potential predictors (age, gender, queerness, awareness of non-binary people, attitude towards non-binary people, attitude towards gender-fair language) significantly improved the model fit . Additionally, significant interaction effects were found between translation strategy and age group (Figure 2), between translation strategy and awareness of non-binary people (Figure 3), and between translation strategy and attitudes towards gender-fair language. For an overview of all model comparisons, see Table 6 in Appendix B.

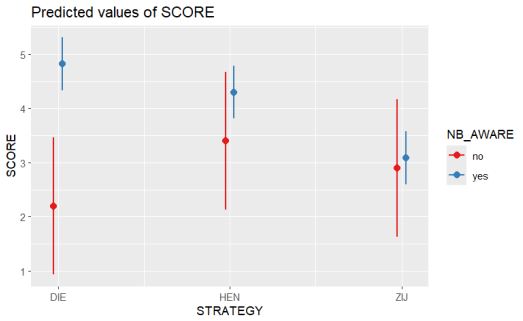

Figure 3: Interaction effect between translation strategy and awareness of non-binary people for overall acceptability score.

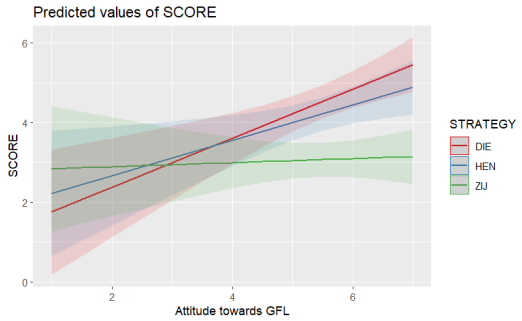

Figure 4: Interaction effect between strategy and attitude towards gender-fair language.

	Estimate	Std.Error	df	t value	Pr(>|t|)
(Intercept)	-0.818	0.887	175.28	-0.922	0.358
STRATEGYHEN	0.639	0.998	150	0.641	0.523
STRATEGYZIJ	1.656	0.998	150	1.659	0.099
LANGUAGE_ATT	**0.617**	0.174	153.05	3.552	**<0.001**
NB_ATT	**0.318**	0.130	74	2.437	**0.017**
STRATEGYHEN:LANGUAGE_ATT	-0.173	0.178	150	-0.972	0.332
STRATEGYZIJ:LANGUAGE_ATT	**-0.566**	0.178	150	-3.183	**0.002**

Table 3: Model summary with strategy, GFL attitude, and attitude towards non-binary people with interaction effect for strategy and GFL as predictors. Reference point (intercept) is the estimate for 'die' at hypothetical GFL score 0.

As can be derived from the model summary in Table 7, *zij* and, to a lesser extent, *hen* received significantly lower acceptability scores than *die*. Participants aged 36 and above gave significantly lower scores overall. However, the model also indicates an interaction between age group and strategy: compared to younger participants (18–24), participants in the 36+ group showed higher ratings for *hen* and *zij*.

The model summaries in Table 8 and Table 9 confirm that *zij* is the least preferred strategy, receiving significantly lower scores than *die*, and additionally shows that men give significantly lower scores than women, whereas queer participants give significantly higher scores than non-queer participants. When taking the interaction effect with awareness of non-binary people into account, however (see Table 10 and Figure 3), the main effect of strategy is no longer significant. Participants who are aware of non-binary people give higher scores overall, but they rate both the *hen* and the *zij* strategy significantly lower than the *die* strategy compared to participants who are unaware of non-binary people.

The best performing model overall consisted of strategy, attitude towards gender-fair language, and attitude towards non-binary people as main predictors, with interaction effect between strategy and attitude towards gender-fair language. This suggests that the scores people assign to specific strategies depend more strongly on their attitudes towards gender-fair language than on demographic variables. An alternative explanation could be that attitudes towards gender-fair language are already partially dependent on demographic variables: participants in the 18-24 age group had a slightly more positive attitude on average (5.73) than participants in the 25-35 age group (5.66) and participants in the 36+ group (4.78); gender-diverse participants had more positive attitudes on average (6.8) than did female (5.6) or male participants (4.7); Queer participants had more positive attitudes on average (6.5) than did non-queer participants (4.8); participants aware of the existence of non-binary people also had more positive attitudes on average (5.7) than did unaware participants (3.7). The complete model summary can be seen in Table 3. This shows that people with a more positive attitude towards gender-fair language and a more positive attitude towards non-binary people provided significantly higher scores overall, but that people with more positive attitudes towards gender-fair language rated the *zij* strategy significantly lower than *die* compared to people with more negative attitudes towards gender-fair language (see Figure 4 for a visualisation of the interaction effect).

5.2 Audience perception (RQ2-3)

A strong positive correlation was found between overall subtitle quality scores and acceptability ratings for non-binary reference ($r(44) = 0.73$, $p < 0.001$), indicating that viewers' evaluation of the pronoun strategy is closely associated with their overall viewing experience. However, this relation is not observed across the experimental conditions (*infra*).

Tables 4 and 5 present the mean scores for *die*,

hen, and *zij*, showing that *die* consistently received the highest ratings for both overall subtitle quality and strategy evaluations. Additionally, its variance is lowest for most measures (except *acceptability*), which suggests greater agreement among participants.

	die	***hen***	***zij***
Comprehens.	**4.795** (0.121)	4.449 (0.622)	4.267 (0.662)
Acceptability	**4.385** (0.376)	3.609 (1.22)	3.7 (1.718)
Quality	**4.538** (0.436)	3.652 (1.6)	3.9 (1.433)

Table 4: Subtitle quality scores for the pronouns *die*, *hen* and *zij* (averages with variance in parentheses).

	die	***hen***	***zij***
Comprehens.	**4.462** (0.603)	4.304 (1.13)	4.1 (0.767)
Acceptability	**4.103** (1.229)	3.768 (1.004)	3.533 (1.19)
Quality	**4.442** (0.793)	3.815 (0.842)	3.6 (1.142)

Table 5: Translation strategy scores for the pronouns *die*, *hen* and *zij* (averages with variance in parentheses).

A one-way ANOVA showed no statistically significant differences in overall subtitle quality scores between the three conditions ($F(2, 43) = 2.81$, $p = 0.07$), although the result suggests a possible trend. Similarly, no statistically significant differences were found in translation strategy ratings across conditions ($F(2, 43) = 2.03$, $p = 0.14$).

These findings suggest that while individual evaluations of pronoun strategies are strongly linked to perceived subtitle quality, the specific strategy presented in the experiments (*die*, *hen* or *zij*) did not show clear differences in participants' overall viewing experience, which contrasts with earlier analyses showing a general preference for *die* and *hen* over *zij* in Dutch. While this could indicate that there is a difference between people's explicit attitudes and their actual viewing experience, it must be noted that the analysis is based on a limited number of data points (N = 13 for *die*, N = 23 for *hen*, N = 10 for *zij*).

6 Discussion & Conclusion

With the increased availability of LGBT+ content via streaming services (Yuan et al., 2025) comes increased translator responsibility to accurately portray the social and linguistic diversity of all characters (Attig, 2023; Lardelli et al., 2025). While there is an industry push towards increased automation (Karakanta, 2022), machine translation might not be suitable to accurately translate non-binary language (Savoldi et al., 2025; Hackenbuchner et al., 2025).

In this study, we focused on the translation of the pronouns *they/them/their* as used by a non-binary character, Cal, in the Netflix series *Sex Education*. We studied the audience perceptions of different pronoun strategies: those used by Dutch human translators (*die/hen/hun* and *hen/hen/hun*), and those generated by MT (plural *ze/hen/hun* with plural verb forms).

When exploring participant attitudes towards those three different strategies in general, our findings confirm earlier findings for Dutch (Van Hoof and Decock, 2026) that the *die* strategy was considered the best strategy, followed bij *hen*. We additionally empirically verified that MT mistranslation into the plural *zij*, as recently documented for Dutch (Hackenbuchner et al., 2025), is seen as the least acceptable referential strategy, despite plural translations being a documented acceptable strategy for gender-inclusive Dutch in a generic reference context (Verhaegen et al., 2025). This difference in acceptability was most outspoken for younger participants (18-24 year olds), participants who were familiar with non-binary people, and participants who had a more positive attitude towards gender-fair language in general. Participants older than 36 actually rated *hen* somewhat higher than both *die* and *zij*.

When examining participants' attitudes towards specific pronoun strategies after exposure in subtitles, no statistically significant differences were found between the three strategies (*die*, *hen*, *zij*), although this is likely due to limited number of valid observations (as quite a few participants either indicated they did not read the subtitles or did not recall the Dutch pronoun strategy used). As more participants could remember seeing the *hen* strategy than any of the other strategies, our findings support the claim that *hen* is a more salient strategy than *die* (Decock et al., 2024). Looking at the data descriptively, we found that *die* received highest scores across all dimensions (comprehensibility, quality and acceptability), both when judging the subtitle quality in general and when judging the suitability of reference to the non-binary character. Differences between *hen* and *zij* were less outspoken, although *zij* scored lowest in terms of suitability. Given that the perceived suitability of the pronoun strategy and the overall subtitle quality were found to be strongly correlated,

we suggest caution in relying on 'classic' MT for this task, adding our voice to previously made calls for community-informed approaches in audiovisual translation (Attig, 2023; Lardelli et al., 2025) and gender-fair MT (Lardelli and Gromann, 2025). Contributing to the limited body of Dutch reception studies (of non-binary pronouns)(Decock et al., 2024; Vriesendorp, 2024; Vos and Nutters, 2022), our study found *die* to be the least controversial strategy for non-binary reference, even in an audiovisual context.

References

Attig, Remy. 2023. A call for community-informed translation: Respecting queer self-determination across linguistic lines. *Translation and Interpreting Studies*, 18(1):70–90.

Audring, Jenny. 2006. Pronominal Gender in Spoken Dutch. *Journal of Germanic Linguistics*, 18(2):85–116.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1):1–48.

Blodgett, Su Lin, Solon Barocas, Hal Daumé Iii, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of 'bias' in NLP. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5454–5476.

De Vos, Lien, Gert De Sutter, and Gunther De Vogelaer. 2021. Weighing Psycholinguistic and Social Factors for Semantic Agreement in Dutch Pronouns. *Journal of Germanic Linguistics*, 33(1):30—66.

Decock, Sofie, Sarah Van Hoof, Ellen soens, and Hanne Verhaegen. 2024. The Comprehensibility and Appreciation of Non-Binary Pronouns in Newspaper Reporting. The Case of Hen and Die in Dutch. *Applied Linguistics*, 45(2):330–347, mar.

Doreleijers, Kim, Jeroen Van Craenenbroeck, and Marjo Van Koppen. 2025. Pronominal gender in Dutch: Apparent-time change in lexical versus semantic agreement. *Journal of Germanic Linguistics*, 37(2):190–220.

for Fundamental Rights, European Union Agency. 2024. A long way to go for LGBTIQ equality: LGBTIQ Survey III. Publication number: TK-01-24-008-EN-N.

Ghosh, Sourojit and Aylin Caliskan. 2023. ChatGPT perpetuates gender bias in machine translation and ignores non-gendered pronouns: Findings across bengali and five other low-resource languages. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 901–912.

Gkovedarou, Eleni, Joke Daems, and Luna De Bruyne. 2025. Gender bias in English-to-Greek machine translation. In *Proceedings of the 3rd Workshop on Gender-Inclusive Translation Technologies (GITT 2025)*, pages 17–45.

GLAAD. 2024. Where We Are on TV 2024-2025.

Guerberof-Arenas, Ana, Joss Moorkens, and David Orrego-Carmona. 2024. "A Spanish version of EastEnders": a reception study of a telenovela subtitled using MT. *The Journal of Specialised Translation*, (41):230–254.

Hackenbuchner, Janiça, Joke Daems, and Eleni Gkovedarou. 2025. GENDEROUS: Machine Translation and Cross-Linguistic Evaluation of a Gender-Ambiguous Dataset. In *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 302–319.

Ipsos. 2025. Ipsos LGBT+ Pride Report 2025: A 26-Country Ipsos Global Advisor Survey, jun. Global Advisor Survey, 26 countries.

Karakanta, Alina. 2022. Experimental research in automatic subtitling: At the crossroads between machine translation and audiovisual translation. *Translation Spaces*, 11(1):89–112.

Konnelly, Lex and Elizabeth Cowper. 2020. Gender Diversity and Morphosyntax: An Account of Singular They. *Glossa: A Journal of General Linguistics*, 5(1):40.

Kuznetsova, Alexandra, Per B. Brockhoff, and Rune H. B. Christensen. 2017. lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82(13):1–26.

Lardelli, Manuel and Dagmar Gromann. 2025. Non-Binary genders in (machine) translation. In *The Routledge Handbook of Translation Technology and Society*, pages 396–408. Routledge.

Lardelli, Manuel, Zrinka Primorac Aberer, and Bettina Hiebl. 2025. Gender-Fair Audiovisual Translation: First Considerations on How to Address Non-Binary Genders. *Journal of Specialised Translation*, 44:44–63.

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.

López, Ártemis. 2022. Trans (de) letion: Audiovisual translations of gender identities for mainstream audiences. *Journal of Language and Sexuality*, 11(2):217–239.

Nabila, Izzatia, Slamet Setiawan, and Widyastuti Widyastuti. 2021. To Disclose a Less Generic Pronoun: Addressing Non-Binary "They". *Lakon: Jurnal Kajian Sastra dan Budaya*, 10(2):84–93, nov.

R Core Team, 2024. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Savoldi, Beatrice, Jasmijn Bastings, Luisa Bentivogli, and Eva Vanmassenhove. 2025. A decade of gender bias in machine translation. *Patterns*, 6(6).

Van Hoof, Sarah and Sofie Decock. 2026. Gender-inclusive language in Dutch. In *Gender-Inclusive Language. Findings from 14 Languages and Open Research Questions*, pages 41–67. De Gruyter Brill.

Verhaegen, Hanne, Sarah Van Hoof, Rebecca Van Herck, Ute Gabriel, Pascal Gygax, and Sofie Decock. 2025. The effect of Dutch gender-neutral pronouns on perceived text quality: generic reference in employee guidelines. *Applied Psycholinguistics*, 46:e30.

Vos, Lette and Tia Nutters. 2022. Wennen aan 'hen' of 'die'?: Een leesonderzoek naar non-binaire voornaamwoorden in de Nederlandse vertaling van Girl, Woman, Other. *Filter, Tijdschrift over vertalen*.

Vriesendorp, Hielke. 2024. Die/diens of hen/hun? Non-binaire voornaamwoorden in het Nederlands. *Nederlandse Taalkunde*, 29(2):255–267.

Yuan, Grace, Maria Agatha, and A. D. Guintu. 2025. LGBT Representation in Film and Media: Social Impact and Future Development: A Literature Review. *International Journal of Education and Social Development*.

Appendix

A) Survey questions

Familiarity with non-binary people

Q: Ken je één of meer non-binaire personen? (Je mag meerdere opties aanduiden) - *EN: Do you know one or more non-binary people? (More than one answer possible)*

A:

- Ja, in mijn directe omgeving (familie/vrienden). - *EN: Yes, in my direct environment (family/friends).*
- Ja, via werk/school, of via kennissen, via vrienden van vrienden. - *EN: Yes, via work/school or via acquaintances, via friends of friends.*
- Ja, via de media (tv, krant, sociale media,..). - *EN: Yes, via media (tv, newspapers, social media, etc.).*
- Nee, niet dat ik weet. - *EN: Not as far as I'm aware.*

Attitude towards non-binary people

For this score, an average was made of three Likert-scale items from 1 (strongly disagree) to 7 (strongly agree).

- Onze maatschappij zou er goed aan doen om non-binaire personen als normaal te beschouwen. - *EN: Non-binary people should be seen as normal in our society.*
- Non-binaire personen zouden volledig aanvaard moeten worden in onze maatschappij. - *EN: Non-binary people should be completely accepted in our society.*
- Er is niets mis met het hebben van een non-binaire genderidentiteit. - *EN: There is nothing wrong with having a non-binary gender identity.*

Overall subtitle evaluation

Q: Geef aan in hoeverre je akkoord gaat met de volgende stellingen over het fragment dat je net bekeken hebt, op een schaal van 1 (niet akkoord) tot 5 (akkoord). Er zijn geen goede of foute antwoorden: het gaat alleen om jouw inschatting. - *EN: Indicate to what extent you agree with the following statements about the fragment you just watched, on a scale from 1 (disagree) to 5 (agree). There are no right or wrong answers: your personal intuition is what matters.*

- Comprehensibility
  - Ik begreep meteen alles wat er in de ondertitels stond. - *EN: I immediately understood everything in the subtitles.*
  - Ik vind dat de ondertitels de boodschap goed hebben overgebracht. - *EN: The subtitles accurately conveyed the message.*
  - Al bij al was de vertaling gemakkelijk te begrijpen. - *EN: All in all, the translation was easy to comprehend.*
- Acceptability
  - Niet alle ondertitels vond ik meteen duidelijk. - *EN: Not all subtitles were clear right away.* - reverse-coded in analysis
  - De ondertitels waren aangenaam om te lezen. - *EN: The subtitles were pleasant to read.*

  - De taal die werd gebruikt, voelde natuurlijk aan. - *EN: The language used felt natural.*
  - Sommige uitdrukkingen in de ondertitels kwamen omslachtig of verwarrend over. - *EN: Some expressions used in the subtitles felt vague or confusing.* - reverse-coded in analysis
  - Ik vind de ondertiteling geschikt voor een breed publiek. - *EN: The subtitles are suitable for the general public.*

- Quality
  - Er leken geen fouten in de ondertitels te staan. - *EN: The subtitles seemed to be free of errors.*

Translation strategy evaluation

Q: Geef aan in hoeverre je akkoord gaat met de volgende stellingen over het fragment dat je net bekeken hebt, op een schaal van 1 (niet akkoord) tot 5 (akkoord). Er zijn geen goede of foute antwoorden: het gaat alleen om jouw inschatting. - *EN: Indicate to what extent you agree with the following statements about the fragment you just watched, on a scale from 1 (disagree) to 5 (agree). There are no right or wrong answers: your personal intuition is what matters.*

- Comprehensibility
  - De woorden die in de ondertitels gebruikt worden om naar Cal te verwijzen, hadden geen invloed op mijn begrip. - *EN: The words used in the subtitles to refer to Cal did not impact my understanding.*

- Acceptability (all items reverse-coded for analysis)
  - Sommige woorden die in de ondertitels gebruikt worden om naar Cal te verwijzen, wekten irritatie bij me op. - *EN: Some words used in the subtitles to refer to Cal were irritating.*
  - Sommige woorden die in de ondertitels gebruikt worden om naar Cal te verwijzen, vond ik vreemd. - *EN: Some words used in the subtitles to refer to Cal were strange.*
  - Sommige woorden die in de ondertitels gebruikt worden om naar Cal te verwijzen vond ik verrassend. - *EN: Some words used in the subtitles to refer to Cal were surprising.*

- Quality
  - De woorden die in de ondertitels gebruikt worden om naar Cal te verwijzen vond ik goed gekozen. - *EN: The words used in the subtitles to refer to Cal were well-chosen.*
  - De woorden die in de ondertiteling gebruikt worden om naar Cal te verwijzen, vond ik gepast. - *EN: The words used in the subtitles to refer to Cal were suitable.*
  - De ondertitels hielpen me om de identiteit van Cal te kunnen schetsen. - *EN: The subtitles helped me get an idea of Cal's identity.*
  - Ik vind dat de vertaalkeuze in de ondertiteling de identiteit van Cal respecteert. *EN: The translation choices in the subtitles respect Cal's identity.*

Familiarity with non-binary pronoun strategies

Q: Welk Nederlandstalig genderneutraal voornaamwoord van de derde persoon enkelvoud kende je al voor je aan dit onderzoek deelnam? (Je mag meerdere opties aanduiden) - *EN: Which Dutch gender-neutral third-person singular pronoun were you familiar with before participating in this study? (Multiple options possible)*

A:

- Ik kende het genderneutrale voornaamwoord 'hen' (met bezitsvorm 'hun'). - *EN: I was familiar with the gender-neutral pronoun 'hen' (with possessive form 'hun').*
- Ik kende het genderneutrale voornaamwoord 'die' (met bezitsvorm 'hun'). - *EN: I was familiar with the gender-neutral pronoun 'die' (with possessive form 'hun').*
- Ik kende het genderneutrale voornaamwoord 'ze/zij' (met bezitsvorm 'hun'). - *EN: I was familiar with the gender-neutral pronoun 'ze/zij' (with possessive form 'hun').*
- Ik kende geen van deze genderneutrale voornaamwoorden. - *EN: I was not familiar with any of these gender-neutral pronouns.*

- Ander, namelijk... - *EN: Other, please specify.*

Attitude towards non-binary pronoun strategies

Participants had to indicate to what extent they agreed with the following statements, on a Likert scale from 1 (strongly disagree) to 7 (strongly agree):

- Ik vind 'hen' (met bezitsvorm 'hun') een geschikt voornaamwoord van de derde persoon enkelvoud om specifiek naar non-binaire personen te verwijzen. - *EN: I think 'hen' (with possessive form 'hun') is a suitable third-person singular pronoun for specific reference to non-binary people.*

- Ik vind 'die' (met bezitsvorm 'hun') een geschikt voornaamwoord van de derde persoon enkelvoud om specifiek naar non-binaire personen te verwijzen. - *EN: I think 'die' (with possessive form 'hun') is a suitable third-person singular pronoun for specific reference to non-binary people.*

- Ik vind 'ze/zij' (met bezitsvorm 'hun') een geschikt voornaamwoord van de derde persoon enkelvoud om specifiek naar non-binaire personen te verwijzen. - *EN: I think 'ze/zij' (with possessive form 'hun') is a suitable third-person singular pronoun for specific reference to non-binary people.*

Attitude towards gender-fair language

For the score used in the analysis, the average score was taken across the 11 statements.

Q: Hieronder volgen enkele stellingen. Het begrip 'seksistische taal' dat in sommige stellingen gebruikt wordt, verwijst naar uitdrukkingen die mensen uitsluiten, als minder belangrijk zien of kleineren op basis van gender. Het begrip 'genderinclusieve taal' dat in sommige stellingen gebruikt wordt, verwijst naar taalgebruik dat als doel heeft om alle genderidentiteiten zichtbaar te maken of alle genderidentiteiten op gelijke voet met elkaar te behandelen. Geef aan in hoeverre je akkoord gaat met de volgende stellingen op een schaal van 1 (helemaal niet akkoord) tot 7 (helemaal akkoord). - *EN: You will read a few statements. 'Sexist language' in this context refers to using gendered expressions that exclude, trivialize or diminish a group of people based on gender. 'Gender-inclusive language' in this context refers to language aimed at making all gender identities visible or treated equally. Please indicate to what extent you agree or disagree for each of the following statements (1 = strongly disagree, 7 = strongly agree).*

- We moeten de manier waarop de Nederlandse taal traditioneel geschreven en gesproken werd niet veranderen. - *EN: We should not change the way the Dutch language has traditionally been written and spoken.* - reverse-coded for analysis

- Er bestaat niet zoiets als seksistische taal. - *EN: There is no such thing as sexist language use.* - reverse-coded for analysis

- Zich zorgen maken over seksistische taal is een triviale bezigheid. - *EN: Worrying about sexist language is a trivial activity.* - reverse-coded for analysis

- Het is belangrijk om komaf te maken met seksistische taal. - *EN: The elimination of sexist language is an important goal.*

- Seksistische taal houdt verband met de seksistische behandeling van mensen in de maatschappij. - *EN: Sexist language is related to sexist treatment of people in society.*

- Hoewel verandering moeilijk is, zouden we toch moeten proberen om seksistische taal te bestrijden. - *EN: Although change is difficult, we still should try to eliminate sexist language.*

- Genderinclusieve taal gebruiken is belangrijk. - *EN: Using gender-inclusive language is important.*

- Door genderinclusieve taal te gebruiken ondersteunen we gendergelijkheid. - *EN: Using gender-inclusive language supports gender equality.*

- Ik sta positief tegenover genderinclusief taalgebruik. - *EN: I have a positive attitude towards gender-inclusive language.*

- Via genderinclusieve taal bestrijden we discriminatie op basis van gender. - *EN: Using gender-inclusive language counters gender-based discrimination*

- Genderinclusieve taal (bv. via ondertitels) maakt audiovisuele inhoud toegankelijker. - *EN: Gender-inclusive language (e.g., in subtitles) makes audiovisual content more accessible.*

B) Model comparison, interaction effects and model summaries

MODEL	AIC	Df	X^2	p-value
Null	995.87	n/a	n/a	n/a
Strategy	969.93	2	29.936	<0.001
Strategy + age group	964.01	2	9.925	0.007
Strategy * age group	957.06	4	14.949	0.005
Strategy + gender	967.39	3	8.539	0.036
Strategy * gender	967.50	6	11.898	0.064
Strategy + queerness	962.79	2	11.145	0.004
Strategy * queerness	962.51	4	8.279	0.082
Strategy + nb awareness	966.59	1	5.338	0.021
Strategy * nb awareness	959.73	2	10.86	0.004
Strategy + nb attitude	947.97	1	23.964	<0.001
Strategy * nb attitude	950.31	2	1.656	0.437
Strategy + GFL attitude	946.96	1	24.976	<0.001
Strategy * GFL attitude	940.40	2	10.559	0.005
Strategy * GFL attitude + NB attitude	936.45	1	5.943	0.015

Table 6: Model comparison for potential predictors of overall acceptability score. Likelihood ratio tests were always performed between minimally nested models: the model with 1 predictor (strategy) compared to a null model, models with 2 predictors compared to the model with only 1 predictor (strategy), models with interaction effect (*) compared to the corresponding model without interaction effect (+).

	Estimate	Std.Error	df	t value	Pr(>|t|)
(Intercept)	5.607	0.378	171	14.849	<0.001
STRATEGYHEN	**-0.893**	0.418	148	-2.134	**0.034**
STRATEGYZIJ	**-2.357**	0.418	148	-5.634	**<0.001**
AGEGROUP25-35	-0.953	0.544	171	-1.752	0.082
AGEGROUP36+	**-2.694**	0.562	171	-4.791	**<0.001**
STRATEGYHEN:AGEGROUP25-35	0.278	0.603	148	0.46	0.646
STRATEGYZIJ:AGEGROUP25-35	0.780	0.603	148	1.294	0.198
STRATEGYHEN:AGEGROUP36+	**1.676**	0.623	148	2.69	**0.008**
STRATEGYZIJ:AGEGROUP36+	**2.270**	0.623	148	3.644	**<0.001**

Table 7: Model summary of model with strategy and age with interaction effect as predictors. Reference point (intercept) is the estimate for 'die' for age group 18-24.

	Estimate	Std.Error	df	t value	Pr(>|t|)
(Intercept)	4.637	0.275	137.105	16.853	<0.001
STRATEGYHEN	-0.299	0.261	152	-1.143	0.25
STRATEGYZIJ	**-1.416**	0.261	152	-5.417	**<0.001**
GENDERmale	**-0.819**	0.399	73	-2.055	**0.04**
GENDERother	1.101	0.677	73	1.626	0.11
GENDERprefer-not-to-say	0.101	1.127	73	0.09	0.93

Table 8: Model summary of model with strategy and gender as predictors. Reference point (intercept) is the estimate for 'die' for gender=female.

	Estimate	Std.Error	df	t value	Pr(>|t|)
(Intercept)	4.022	0.267	144.450	15.083	<0.001
STRATEGYHEN	-0.299	0.261	152	-1.143	0.255
STRATEGYZIJ	**-1.416**	0.261	152	-5.417	**<0.001**
QUEERnot-sure	1.382	1.1	74	1.257	0.213
QUEERyes	**1.203**	0.367	74	3.282	**0.002**

Table 9: Model summary of model with strategy and queerness as predictors. Reference point (intercept) is the estimate for 'die' for queer=no.

	Estimate	Std.Error	df	t value	Pr(>|t|)
(Intercept)	2.2	0.642	170.818	3.425	<0.001
STRATEGYHEN	1.2	0.705	150	1.703	0.091
STRATEGYZIJ	0.7	0.705	150	0.993	0.322
NB_AWAREyes	**2.621**	0.689	170.818	3.806	**<0.001**
STRATEGYHEN:NB_AWAREyes	**-1.722**	0.755	150	-2.28	**0.024**
STRATEGYZIJ:NB_AWAREyes	**-2.431**	0.755	150	-3.219	**0.002**

Table 10: Model summary of model with strategy and awareness of non-binary people plus interaction effect as predictors. Reference point (intercept) is the estimate for 'die' for nb-aware=no.

Rethinking Gender Annotation for Bias Evaluation in Machine Translation: Can LLMs Improve Reliability?

Chiara Manna
Tilburg University
The Netherlands

Argentina Anna Rescigno
University of Pisa
University of Naples "L'Orientale"
Italy

Eva Vanmassenhove
Tilburg University
The Netherlands

{c.manna, e.o.j.vanmassenhove}@tilburguniversity.edu
argentina.rescigno@phd.unipi.it

Abstract

The assessment of gender bias in Machine Translation critically depends on reliable methods for identifying grammatical gender in system outputs. In this paper, we compare the automated WinoMT annotation pipeline, based on word alignment and morphological tagging, with an instruction-tuned LLM (Qwen3-8B) used to annotate grammatical gender in English–Italian translations. Both approaches achieve similar levels of agreement with a human-annotated gold standard, but exhibit distinct systematic weaknesses. The WinoMT pipeline is sensitive to alignment shifts and morphological tagging limitations, often resulting in indeterminate gender labels. The LLM, in contrast, tends to favour binary gender labels even when noun phrases are morphologically gender-invariant. A qualitative analysis of model outputs and reasoning traces further suggests a lack of stable generalization of grammatical gender rules, even when illustrative exemplars are provided for in-context learning. This highlights important methodological limitations in using LLMs as objective annotators for gender evaluation tasks.

1 Introduction

Interest in gender bias in Machine Translation (MT) has grown substantially over the past decade (Savoldi et al., 2025), reflecting broader concerns about systematic biases in Natural Language Processing (NLP) systems (Sun et al., 2019; Costa-jussà, 2019; Blodgett et al., 2020; Stanczak and Augenstein, 2021) as well as specific cross-linguistic challenges inherent to the translation task. Translating from languages in which gender is sparsely marked (e.g., English) into morphologically richer ones (e.g., German, Italian or Hindi) (Stahlberg et al., 2007; Ackerman, 2019; Cao and Daumé III, 2020) often requires making implicit gender information explicit in the target (Vanmassenhove et al., 2018; Vanmassenhove, 2024). This makes gender overt in the target language, thereby exposing the extent to which model outputs reflect statistical patterns (and potential biases) in the training data. As illustrated in Figure 1, even when gender information is available in the context, MT systems tend to default to sta-

INPUT

The cook prepared a soup for the **housekeeper** because **he** helped clean the room.

OUTPUT

Il cuoco ha preparato una zuppa per **la governante** perché ha aiutato a pulire la stanza.

INPUT

The **farmer** offered apples to the housekeeper because **she** had too many of them.

OUTPUT

Il contadino offrì mele alla governante perché ne aveva troppe.

Figure 1: EN→IT WinoMT translation examples defaulting to *la governante* (fem.) and *il contadino* (masc.). Generated by ChatGPT on September 29th, 2025.

tistically dominant gender realizations, rendering *housekeeper* as *la governante* (fem.) and *farmer* as *il contadino* (masc.) in the Italian target.

At the same time, the field of MT has undergone a rapid paradigm shift, transitioning from rule-based systems to statistical and neural MT, and more recently to Large Language Models (LLMs). These developments have significantly improved translation quality and established the new state-of-the-art in MT (Kocmi et al., 2023; Zhu et al., 2024; Kocmi et al., 2024; Deutsch et al., 2025). Nonetheless, the increasing complexity of modern architectures has raised concerns for AI governance regarding their opacity and biases.

While a substantial body of work has focused on analysing MT outputs (Rescigno et al., 2020; Ramesh et al., 2021; Rescigno and Monti, 2023) and developing bias mitigation strategies (Vanmassenhove et al., 2018; Moryossef et al., 2019; Habash et al., 2019; Hirasawa and Komachi, 2019; Font and Costa-jussà, 2019; Zmigrod et al., 2019; Saunders and Byrne, 2020; Vanmassenhove et al., 2021; Sun et al., 2021; Piergentili et al., 2023), significant effort has also been devoted to the development of evaluation frameworks (Stanovsky et al., 2019; Bentivogli et al., 2020; Levy et al., 2021; Manna et al., 2026) that enable a systematic assessment of gender bias and gender-related errors. These frameworks generally compare translation quality across gender conditions, analyse the distribution of gendered outputs in ambiguous contexts, or evaluate whether systems correctly disambiguate grammatical gender in unambiguous contexts.

Evaluation therefore depends not only on the quality of the MT system itself, but also on the reliability of the methods used to identify the realized gender in the output. Many widely adopted benchmarks rely on automatic pipelines, typically involving techniques such as reference matching, word alignment, and morphological analysis. In this context, WinoMT (Stanovsky et al., 2019) has become one of the most widely adopted benchmarks for gender bias evaluation in MT, due to the simplicity and scalability of its integrated pipeline. By combining word alignment (e.g., *fast_align* (Dyer et al., 2013)) with morphological analysis tools (e.g., *spaCy* (Honnibal et al., 2020)), it enables (multilingual) gender annotation without requiring manually constructed references. However, because it relies on two intermediate processing steps, it inevitably introduces a certain degree of noise (alignment or morphological tagging errors) that can affect the automatic assessment and thus the validity of the resulting gender bias measurements. As an alternative, LLMs have demonstrated great potential as high-level annotators for a range of linguistic tasks, including the evaluation of gender-neutral rewriting in MT (Piergentili et al., 2025). Nonetheless, their employment strictly requires rigorous validation against human-annotated gold standards to ensure methodological integrity and validity (Pangakis et al., 2023; Baumann et al., 2025).

To assess the reliability of automatic gender annotation for gender bias evaluation in MT, we compare these two alternative approaches — the WinoMT evaluation pipeline and a state-of-the-art instruction-tuned LLM — against human annotations as gold standard. Specifically, we address the following research questions (RQs).

> **[RQ1]** To what extent do automatic gender annotations produced by the WinoMT pipeline agree with human annotations?

> **[RQ2]** To what extent do automatic gender annotations produced by an instruction-tuned LLM agree with human annotations, and how do they compare with the WinoMT pipeline?

2 Related Work

2.1 Gender Bias Evaluation in MT

Since 2019, a growing number of evaluation frameworks have been proposed to systematically assess gender bias in MT, typically using controlled or contrastive settings to analyse how systems perform at gender disambiguation and whether they default to statistically dominant, often masculine, forms.

Early work mostly relied on synthetic template-based sentences. Among these, Cho et al. (2019) and Prates et al. (2020) investigated the effect of translating from genderless languages (e.g., Hungarian) or languages with gender-neutral pronominal systems (e.g., Korean) into English using minimal template-based sentences of the form *[pronoun] is a [profession]*. In this setup, the realised gender was inferred by manually or automatically

detecting gendered pronouns in the output (e.g., *he* vs. *she*). Importantly, these templates provide no explicit contextual cues for gender at the source level, meaning that any gender assignment in the translation reflects how systems resolve inherently ambiguous inputs.

Subsequent work extended occupation-based templates to grammatical gender languages. Escudé Font et al. (2019), for instance, adopted more complex templates in which the target gender was recovered from morphologically inflected noun phrases (e.g., *amigo* vs. *amiga* in Spanish). Across these approaches, gender annotation relied on lexical or morphological matching procedures, often requiring language-specific rules to detect gender-marked forms in the translated output. Similarly, Stanovsky et al. (2019) introduced WinoMT, a template-based benchmark designed to evaluate gender bias when translating from English into eight grammatical gender languages, including Italian. Compared to earlier template-based approaches, WinoMT introduced more complex (albeit still template-based) sentence structures containing two human entities and a gendered pronoun to disambiguate the gender of the correct referent, as illustrated in Figure 1. In contrast to some of the earlier work (e.g., Cho et al. (2019)), these contextual cues should enable evaluation of whether models follow contextual evidence or still default to stereotypical associations. Moreover, it introduced a unified gender annotation pipeline based on word alignment (e.g., *fast_align* (Dyer et al., 2013)) and automatic morphological tagging (e.g., *spaCy* (Honnibal et al., 2020)) to locate the target entity in the translation output and extract its gender. By removing the need for manually constructed references and language-specific matching strategies, this design enabled, for the first time, scalable multilingual evaluation, contributing to it becoming one of the most widely adopted benchmarks as well as a foundation for numerous extensions (Troles and Schmid, 2021; Levy et al., 2021; Kappl, 2025; Manna et al., 2025; Manna et al., 2026).

Alongside synthetic test sets, several studies have opted for naturally occurring data for the assessment of gender bias in MT (Vanmassenhove et al., 2018; Elaraby and Monz, 2018; Moryossef et al., 2019; Habash et al., 2019; Alhafni et al., 2022; Bentivogli et al., 2020; Vanmassenhove and Monti, 2021). In this context, MuST-SHE (Bentivogli et al., 2020) represents the first structured multilingual benchmark (EN→{IT, ES, FR, DE}), derived from TED talks data. Originally designed to analyse translation quality across speaker and/or referent gender in both ambiguous and unambiguous contexts, it was later adapted to explicitly measure gender accuracy by matching system outputs against predefined gender realizations. Nonetheless, as most of such frameworks, it remains dependent on manually curated references and language-specific matching strategies.

More recently, the WinoMT paradigm has been extended to naturally occurring data through the BUG corpus (Levy et al., 2021), by automatically extracting English sentences linking role nouns with gendered pronouns from sources such as Wikipedia and PubMed. Because BUG preserves the structural properties underlying WinoMT, gender annotation can be performed using the same automatic pipeline across multiple target languages, combining linguistic realism with cross-lingual scalability. As a result, the WinoMT annotation protocol has effectively become a predominant approach for gender bias evaluation across both synthetic and naturally occurring benchmarks. However, the reliability of the underlying pipeline has received limited attention.

Since gender annotation depends on two intermediate automatic processing steps, errors may arise at different stages of the pipeline, inevitably introducing some degree of noise into the evaluation. In addition to alignment errors or morphological tagging failures that may mischaracterise grammatical gender, there are cases in which the pipeline fails to identify any grammatical gender altogether. Such instances are assigned an `UNKNOWN` label and treated as errors under the standard WinoMT evaluation protocol (Stanovsky et al., 2019). As reported by Manna et al. (2026), these cases occur with non-negligible frequency and may distort gender bias metrics as much as incorrect gender labels, despite not necessarily reflecting genuine translation errors.

2.2 LLMs as High-level Annotators

Ultimately, the evaluation of gender bias in MT depends on the quality and limitations of the (automatic) annotation procedure used to identify the realized gender in translated outputs. In this context, recent work has explored the use of LLMs for data annotation in NLP (Pangakis et al., 2023;

Tan et al., 2024), automating complicated labelling processes across various data types (Tan et al., 2024).

While manual annotation is costly, slow, and prone to human fatigue and interpretative shifts (Pangakis et al., 2023; Nasution and Onan, 2024), LLMs might offer a more cost-effective alternative capable of processing complex data using only natural language instructions (Törnberg, 2024). Moreover, they have been shown to rival or even outperform traditional crowd-sourced workers in both speed and inter-annotator agreement in many text-annotation tasks (Gilardi et al., 2023; Pangakis et al., 2023; Nasution and Onan, 2024; Törnberg, 2024). However, their performance still falls short of human experts (as opposed to crowd-sourced workers) in domain-specific settings, indicating that LLMs should not serve as substitutes for expert annotators (Tseng et al., 2025).

Beyond simple text categorization, there is growing interest for the "LLM-as-a-judge" framework to automatically evaluate MT and Natural Language Generation (NLG) (Gao et al., 2025; Li et al., 2025), where LLM evaluators can help overcome the limitations of traditional metrics that rely strictly on n-gram overlap or statistical word alignments (Gao et al., 2025). In the context of gender in translation, recent studies show that LLMs can serve as effective evaluators, particularly when reference translations are available, and that their performance benefits from advanced prompting strategies such as Chain-of-Thought (CoT) prompting (Qian et al., 2024). For instance, Piergentili et al. (2025) demonstrate that LLMs can be successfully deployed to assess gender-neutral translations, with performance improvements under structured prompting. However, this evidence comes from a relatively constrained setting, in which a single phenomenon is annotated and a sentence-level label is produced. Consequently, it does not directly reflect the ability to perform more fine-grained linguistic annotations.

As such, LLMs cannot be blindly assumed to act as objective annotators (Baumann et al., 2025). Their performance is highly dependent on the conceptual difficulty of the task, the specific dataset, and sensitive to the prompt design and task formulation, which may introduce systematic biases and variability in annotation outcomes (Pangakis et al., 2023; Li et al., 2025). Therefore, the need for rigorous validation of LLM-based annotations against human gold standards has been increasingly emphasized in the literature (Pangakis et al., 2023; Törnberg, 2024; Walshe et al., 2025).

3 Experimental Setup

We compare the WinoMT pipeline with an instruction-tuned LLM acting as high-level annotator for grammatical gender in English-to-Italian (EN→IT) translation, by evaluating the extent to which each approach reproduces a human-annotated ground truth. While this research is motivated by the social phenomenon of gender bias, our specific task and gold standard strictly evaluate morphosyntactic (grammatical) gender. The following subsections describe the data, models, prompting configurations, and evaluation protocol used to conduct this comparison.

3.1 Data

We base our experiments on the **WinoMT** challenge set (Stanovsky et al., 2019), which builds on the Winogender (Rudinger et al., 2018) and WinoBias schemas (Zhao et al., 2018). The dataset consists of English sentences following a simple template, in which an ungendered profession noun referring to a human entity (the target noun to be translated) is linked via coreference to a gendered pronoun (e.g., *he* or *she*) that provides the gold gender cue (see Figure 1).

The WinoMT consists of three subsets, pro-stereotypical, anti-stereotypical, and a regular/neutral set. In this work, we focus on the regular set, which counts 3,888 sentences balanced across male and female referents (*i.e.*, gender instantiated via pronouns). From this set, we select a balanced sample of 500 input sentences, ensuring that each profession is equally represented across gold gender conditions (*i.e.*, *he* / *she*) and stereotypical alignment (pro- and anti-stereotypical). Because WinoMT does not provide reference translations and is designed to operate on the output of arbitrary MT systems, we generate translations using two models (§3.2), which serve as input for both annotation approaches as well as our human validation.

3.2 Translation Setup

To generate our target sentences, we translate the selected set of 500 English inputs into Italian (**EN→IT**). This translation direction is particularly relevant for gender bias evaluation in MT,

as gender is not marked in the English source but is typically explicitly realized in the Italian target, creating conditions under which biases may emerge. At the same time, the relatively rich and complex Italian morphology, characterized by both overt gender marking and the presence of gender-invariant (epicene) nouns, makes this setting well-suited for evaluating the reliability of automatic gender annotation methods.

We use two decoder-only models:

1. **Llama2**[1] (Touvron et al., 2023): a general-purpose pretrained LLM, known for strong zero-shot translation performance among open-weight LLMs (Xu et al., 2024).
2. **TowerInstruct**[2] (Alves et al., 2024): an instruction-tuned multilingual model optimized for translation tasks, built on top of Llama2 through (MT-specific) continued pretraining and instruction tuning.

For both models, we select the 7B version, which is both compatible with our computational resources and representative of realistic MT deployment scenarios.

Llama2

```
Translate the following text from
English into Italian.
English: {source_sentence}
Italian: {generated_translation}
```

TowerInstruct

```
<|im_start|> user
Translate the following text from
English into Italian.
English: {source_sentence}
Italian: <|im_end|>
<|im_start|> assistant
{generated_translation}
```

Figure 2: Prompt templates used for the translation task. Plain instruction format for Llama2, chat-style for TowerInstruct.

At this stage of the analysis, our objective is not to optimize translation performance through prompt engineering. Therefore, we adopt a zero-shot prompt format by reproducing the template used for TowerInstruct's post-training[3], and provide a plain natural language equivalent for Llama2 (see Fig. 2). More advanced prompting strategies are instead explored in the LLM-based annotation setup.

3.3 WinoMT Automatic Gender Annotation

The WinoMT evaluation framework relies on an automatic gender annotation pipeline designed to locate the target profession noun in the translation and extract its grammatical gender. To this end, the pipeline consists of two intermediate processing steps. First, word alignment is performed using *fast_align* (Dyer et al., 2013), which establishes a mapping between the source sentence and its translation in order to locate the word corresponding to the English profession noun in the target. Second, morphological analysis is applied using *spaCy* (Honnibal et al., 2020) to extract the grammatical gender feature associated with the aligned profession noun, resulting in one of three labels for each sentence: `MALE`, `FEMALE`, or `UNKNOWN`. The latter represents cases in which no gender marking has been automatically identified, for instance due to gender-invariant forms (e.g., *l'assistente*, *l'autista* in Italian), but also alignment errors, morphological tagging failures, or translation variability (e.g., incoherent output or omission of the target entity). For consistency, these labels are mapped to the `MASCULINE`, `FEMININE`, and `UNKNOWN` scheme used throughout our evaluation.

3.4 LLM-based Gender Annotation

As an alternative automatic approach, we explore the use of an instruction-tuned LLM as high-level annotator for grammatical gender. The model is tasked with reproducing the same annotation procedure as the WinoMT pipeline, namely (i) locating the translated referent corresponding to the English profession noun and (ii) identifying the grammatical gender expressed in the corresponding noun phrase.

For this task, we opt for **Qwen3-8B** (Yang et al., 2025) for several reasons. First, we aim to maintain an open and reproducible gender annotation setup that could potentially serve as a scalable alternative to the WinoMT framework, making open-source models preferable to proprietary alternatives. Second, Qwen3 belongs to the family of reasoning-oriented LLMs, which are specifically designed to perform structured, multi-step inference (Besta et al., 2025). We therefore consider it particularly suitable for our designed gender an-

[1] huggingface.co/meta-llama/Llama-2-7b-hf
[2] huggingface.co/Unbabel/TowerInstruct-7B-v0.2
[3] huggingface.co/datasets/Unbabel/TowerBlocks-v0.2

notation task, which is explicitly decomposed into two steps, as previously defined. This choice is further supported by prior work reporting strong performance of Qwen-family models, in particular Qwen2.5, in similar gender evaluation tasks (Piergentili et al., 2025).

Few-shot exemplars

```
Noun phrases like "il manager", "il
professore", "il direttore"
are MASCULINE.

Noun phrases like "la manager", "la
professoressa", "la direttrice"
are FEMININE.

Noun phrases like "l'assistente",
"l'insegnante", "l'auditor"
are UNKNOWN.
```

Figure 3: Few-shot exemplars used in the LLM-based annotation prompt to illustrate masculine, feminine, and unknown (gender-invariant) noun phrases. For each sentence, three exemplars per label are dynamically filtered so that the profession under annotation does not appear among them.

Our setup is closely related to the one proposed by Piergentili et al. (2025), in which LLMs are employed to assess whether translated referents are correctly rendered in a gender-neutral form. Although they explore different prompting approaches, either producing direct sentence-level assessments or eliciting intermediate phrase-level annotations to be aggregated into a sentence-level judgment, their task implicitly depends on identifying relevant referential expressions and their gender realization, aligning closely with our annotation task. Building on their framework, we adopt a comparable prompting strategy, in both **zero-shot** and **few-shot** configurations, in which the model receives the English source sentence, its Italian translation, and the English profession noun corresponding to the referent of interest. While the zero-shot setup relies exclusively on natural language task instructions, the few-shot configuration augments these with three illustrative examples per label, namely Italian noun phrases representing masculine, feminine, and gender-invariant forms (e.g., *il professore*, *la direttrice*, *l'assistente*), to elicit in-context learning (see Fig. 3). To avoid lexical leakage, exemplars are selected dynamically so that the profession noun under annotation is never included among these.

In both configurations, the model is instructed to identify the translated noun phrase and assign a gender label based exclusively on morphosyntactic cues internal to the target noun phrase (e.g., determiner and noun morphology), to replicate the WinoMT protocol. We further constrain model outputs to a fixed sentence-level label format (*i.e.*, `MASCULINE`, `FEMININE`, or `UNKNOWN`). Both prompting configurations are applied to the same set of 1,000 translated sentences (two sets of 500 sentences) and the full (system) prompts are provided in the Appendix.

3.5 Human Gold Annotation

To evaluate the reliability of both the WinoMT automatic pipeline and the LLM-based annotations, we use a human annotation as our gold standard. An expert linguist fluent in the Italian language manually annotated the target sentences. The annotator was instructed to label the grammatical gender of the target profession noun, by looking at the determiner and the noun itself. Therefore, the task is strictly formal and deterministic, leaving no space for ambiguity or subjectivity.

The same label organisation is adopted throughout the analysis by assigning `MASCULINE` or `FEMININE` according to the grammatical gender expressed by the profession noun, and `UNKNOWN` whenever gender cannot be reliably inferred, even when considering the associated determiner. Since this classification is based on the Italian language morphosyntactic rules of agreement, it is free from personal interpretation and ensures objective and consistent labelling across the datasets. The creation of this objective gold standard allowed us to compute the **agreement with the human annotation** as a direct measure of the reliability of each automated approach.

Dataset	MASC.	FEM.	UNKNOWN
Llama2	0.70	0.18	0.12
TowerInstruct	0.56	0.33	0.11

Table 1: Distribution of (human) gold annotation labels across the translated outputs generated by each MT system. Proportions are computed over the 500 annotated sentences.

Although our sample is balanced with respect to source-side gender cues (Section 3.1), our MT outputs are not. As reported in Table 1, we can observe that both Llama2 and TowerInstruct exhibit a strong tendency to default to masculine realizations, resulting in substantially skewed gender distributions in our target sentences, which reflect realistic MT deployment scenarios. There-

Dataset	Method	Accuracy	Macro F1	Class-specific F1		
				Masculine	Feminine	Unknown
Llama2	WinoMT	0.92	**0.85**	0.95	0.94	**0.67**
	LLM zero-shot	0.90	0.75	0.96	0.90	0.40
	LLM few-shot	0.92	0.79	0.96	0.92	0.50
TowerInstruct	WinoMT	0.88	**0.82**	0.93	0.91	**0.63**
	LLM zero-shot	0.91	0.78	0.95	0.94	0.45
	LLM few-shot	0.92	**0.82**	0.96	0.95	0.57

Table 2: Agreement with human gold annotations across automatic annotation methods. We report overall accuracy, macro F1, and per-class F1 scores on the Llama2- and TowerInstruct-generated translation sets.

fore, agreement scores are computed not only in terms of overall accuracy, but also using macro and per-class F1 scores, in order to better capture label-specific annotation behaviour.

4 Results

To establish the baseline for our evaluation (RQ1), we first measure the agreement between the **WinoMT pipeline** and our manually annotated gold standard. As reported in Table 2, the integrated pipeline achieves strong overall accuracy scores, reaching 0.92 on the Llama2-translated set and 0.88 on the TowerInstruct-translated one. However, the F1-based evaluation reveals important label-specific asymmetries. In particular, the class-specific F1 breakdown indicates uneven performance across labels, with substantially weaker scores for the UNKNOWN category.

To answer RQ2, we then evaluate the performance of **Qwen3-8B** acting as grammatical gender annotator in both a zero-shot and a few-shot setting. In the **zero-shot** configuration, the LLM's performance is virtually on par with the WinoMT pipeline in terms of overall accuracy. However, the F1-based evaluation reveals a strong tendency to force binary gender labels onto morphologically gender-invariant nouns. While performance remains high for MASCULINE and FEMININE instances (with class-specific F1 scores between 0.90 and 0.96), it drops substantially for the UNKNOWN category (with scores between 0.40 and 0.45). Accordingly, incorrect gender annotations most frequently involve cases labelled as UNKNOWN in the human gold standard (Tab 4).

To try and mitigate this by clarifying the expected decision boundary between morphologically marked and invariant forms, we evaluate the performance of the **few-shot** prompting strategy including specific annotations exemplars for each label. While overall accuracy remains largely comparable to the zero-shot setting, the few-shot configuration leads to a modest increase in the (macro) F1 score, primarily driven by improved performance on the UNKNOWN category.

5 Analysis and Discussion

While overall accuracy scores suggest relatively comparable performance across methods, a more fine-grained evaluation exposes qualitative differences in annotation behaviour.

(a) Llama2

Predicted label	WinoMT	LLM zero-shot	LLM few-shot
Masculine	0.46	**0.54**	**0.57**
Feminine	0.12	0.38	0.36
Unknown	**0.41**	0.08	0.07
Total errors (n)	41	48	42

(b) TowerInstruct

Predicted label	WinoMT	LLM zero-shot	LLM few-shot
Masculine	0.22	**0.53**	**0.53**
Feminine	0.19	0.30	0.25
Unknown	**0.59**	0.17	0.23
Total errors (n)	59	47	40

Table 3: Distribution of errors by predicted label across annotation methods for the Llama2- and TowerInstruct-generated datasets. Values are proportions computed over the total number of errors per method.

A closer analysis of the **WinoMT pipeline** reveals an important limitation stemming from its reliance on word alignment and automatic morphological tagging. In our formulation, the UNKNOWN label is intended to capture cases in which gender cannot be determined (e.g., for morphologically gender-invariant, epicene forms). In practice, however, the pipeline also assigns UNKNOWN when gender cannot be recovered due to alignment shifts, translation variability, or tagging errors. As a result, the UNKNOWN category conflates genuine linguistic indeterminacy with pipeline failures. This behaviour is reflected in the low class-

specific F1 scores observed for the UNKNOWN category (Tab. 2), as well as the distribution of *errors by predicted label* reported in Table 3. Most incorrect annotations produced by the WinoMT pipeline indeed fall under the UNKNOWN label, indicating a systematic tendency to overproduce indeterminate annotations.

Although this lies beyond the scope of the present work, these observations have important implications for the interpretation of Wino-MT based evaluations. Under the WinoMT protocol, UNKNOWN instances are treated as "incorrectly gendered" and therefore penalized in the computation of gender bias metrics. As already discussed by Manna et. al (2026), such a design choice may distort the resulting assessment, especially considering that neither morphologically gender-invariant forms nor annotation errors necessarily reflect bias in the MT system itself.

(a) Llama2

Gold label	LLM zero-shot	LLM few-shot
Masculine	0.04	0.05
Feminine	0.04	0.02
Unknown	**0.92**	**0.93**
Total Errors (n)	48	42

(b) TowerInstruct

Gold label	LLM zero-shot	LLM few-shot
Masculine	0.04	0.08
Feminine	0.15	0.15
Unknown	**0.81**	**0.78**
Total Errors (n)	47	40

Table 4: Distribution of LLM annotation errors by gold label for the Llama2- and TowerInstruct-generated sets under zero-shot and few-shot prompting. Values are proportions computed over the total number of errors per method.

When prompting **Qwen3-8B** for the same annotation task, we observe accuracy scores comparable to those obtained with the WinoMT pipeline. However, the F1-based evaluation reveals a different weakness. In the zero-shot configuration, the LLM exhibits a strong bias toward enforcing a binary choice, assigning either MASCULINE or FEMININE even when the (Italian) noun phrase is morphologically gender-invariant. This behaviour is reflected in the markedly lower class-specific F1 scores obtained for the UNKNOWN category, as well as in the distribution of *errors by predicted label* (Tab. 3), which shows that incorrect annotations are predominantly assigned to binary categories, with a strong skew toward MASCULINE (and to a lesser extent FEMININE). At the same time, the distribution of *errors by gold label* (Tab. 4) indicates that the vast majority of these errors involve instances labelled as UNKNOWN in the human annotation. Taken together, these observations suggest that the LLM struggles to recognize genuine linguistic indeterminacy, instead defaulting to a binary annotation.

While overall accuracy remains largely unchanged, augmenting the prompt with illustrative examples leads to a slight improvement in both macro and class-specific F1 scores for the UNKNOWN category (Tab. 2). Nonetheless, this effect is inconsistent: the few-shot configuration both corrects and introduces errors. While some incorrect binary annotations are revised to UNKNOWN, some cases that were (previously) correctly identified as UNKNOWN are reassigned to a binary category.

A qualitative inspection of the generated reasoning (Fig. 4), further highlights this inconsistency, revealing that correct annotations do not

Case 1: binary → UNKNOWN

*Il supervisore criticò **l'assistente** perché ha commesso degli errori.*
lit. *"The supervisor criticised the attendant because they made some mistakes."*

zero-shot FEMININE
"l' corresponds to la, therefore feminine."

few-shot UNKNOWN
"assistente can be masculine or feminine … similar to l'insegnante."

Case 2: UNKNOWN → binary

*Il meccanico chiese **l'assistente** perché era l'unico che poteva aiutarlo.*
lit. *"The mechanic asked for the attendant because they were the only one who could help."*

zero-shot UNKNOWN
"assistente can be either masculine or feminine."

few-shot FEMININE
"assistente ends in -e, typically feminine."

Figure 4: Examples of annotation flips and corresponding reasoning (excerpts) between zero-shot and few-shot prompting for the same morphologically (gender-)invariant noun phrase (*l'assistente*). Highlighted labels indicate agreement (green) or disagreement (orange) with human gold annotation.

directly derive from a stable understanding of (Italian) morphological gender marking. For instance, the same morphologically gender-invariant noun phrase (*l'assistente*) receives different labels across prompting conditions, often accompanied by inconsistent or contradictory explanations. In some cases, the model correctly identifies the noun as gender-invariant, while in others it incorrectly infers a binary label based on simplified or inaccurate heuristics (e.g., associating the ending *-e* with feminine gender).

More generally, the model often produces linguistically inaccurate or partially inconsistent explanations that may nevertheless lead to the correct label by relying on surface-level morphosyntactic patterns rather than a stable generalization of grammatical gender rules, even when illustrative exemplars are provided.

6 Conclusion

The study examines the reliability of automatic gender annotation methods for assessing bias in machine translation, by comparing the widely adopted automated WinoMT pipeline with an instruction-tuned LLM, Qwen3-8B, prompted to reproduce the same multi-step task.

While both approaches achieve comparable accuracy scores against human gold annotations, they exhibit distinct systematic weaknesses. The WinoMT pipeline tends to overassign the `UNKNOWN` label due to alignment errors or morphological tagging limitations on complex morphological constructions. In contrast, the LLM tends to favour binary gender annotations (`MASCULINE` or `FEMININE`), even when faced with ambiguous or epicene Italian nouns (e.g. *insegnante*, *assistente*). We further observe that the introduction of specific examples in the prompt leads to only marginal changes in overall agreement with human annotations. A thorough examination of the reasoning process generated by the model further reveals that correct annotations are not consistently supported by a genuine understanding of Italian morphological rules.

Overall, while LLM-based (gender) annotation proves to be a competitive alternative to the WinoMT pipeline, it cannot be considered infallible or a substitute for human annotation. Its dependence on prompt formulation and the lack of solid linguistic reasoning confirm the necessity of a constant validation against human gold standard annotations, to preserve the integrity and reliability of gender bias assessment.

7 Limitations

We acknowledge that our findings are limited to Italian, a grammatical gender language with relatively rich and transparent gender morphology. Both investigated approaches may behave differently for languages in which gender is more sparsely marked, as well as in low-resource settings, where less consistent agreement patterns and limited representation in the training data may reduce the reliability and robustness of automatic gender annotation. Similarly, the LLM-based annotation setup is evaluated using a single instruction-tuned LLM, and different models may exhibit different annotation behaviours and sensitivities to prompting strategies.

A further limitation concerns the treatment of gender-neutral language. While in our analysis the `UNKNOWN` label is interpreted as reflecting gender-invariant forms (e.g., epicene nouns), gender-neutrality in Italian is often achieved through reformulation strategies such as pluralization, impersonal constructions, or other structural changes (Sulis and Gheno, 2022; Piergentili et al., 2023). These strategies may not be captured by morphology-based annotation alone, and alignment-based methods may struggle to map complex reformulations to the source referent.

Finally, it is important to note that the annotation task considered in this work remains relatively constrained. It involves a single phenomenon (gender marking), which is annotated based on a limited set of labels, and can therefore be categorized as a high-level classification task. As such, our findings should not be interpreted as evidence of LLMs' ability (or inability) to perform more fine-grained linguistic annotations. Rather, they reflect the behaviour of such systems within a simplified and controlled evaluation setting.

8 Acknowledgements

The second author acknowledges funding from the National PhD programme in Artificial Intelligence, partnered by the University of Pisa and the University of Naples "L'Orientale", through the doctoral grant 39-411-24-DOT23A27WJ-7219 established by Ex DM 318, of type 4.1, co-financed by the National Recovery and Resilience Plan.

References

Ackerman, Lauren. 2019. Syntactic and cognitive issues in investigating gendered coreference. *Glossa: a journal of general linguistics*, 4(1).

Alhafni, Bashar, Nizar Habash, and Houda Bouamor. 2022. The Arabic parallel gender corpus 2.0: Extensions and analyses. In Calzolari, Nicoletta, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1870–1884, Marseille, France, June. European Language Resources Association.

Alves, Duarte M, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, et al. 2024. Tower: An open multilingual large language model for translation-related tasks. *arXiv preprint arXiv:2402.17733*.

Baumann, Joachim, Paul Röttger, Aleksandra Urman, Albert Wendsjö, Flor Miriam Plaza-del Arco, Johannes B Gruber, and Dirk Hovy. 2025. Large language model hacking: Quantifying the hidden risks of using llms for text annotation. *arXiv preprint arXiv:2509.08825*.

Bentivogli, Luisa, Beatrice Savoldi, Matteo Negri, Mattia A. Di Gangi, Roldano Cattoni, and Marco Turchi. 2020. Gender in danger? evaluating speech translation technology on the MuST-SHE corpus. In Jurafsky, Dan, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6923–6933, Online, July. Association for Computational Linguistics.

Besta, Maciej, Julia Barth, Eric Schreiber, Ales Kubicek, Afonso Catarino, Robert Gerstenberger, Piotr Nyczyk, Patrick Iff, Yueling Li, Sam Houliston, Tomasz Sternal, Marcin Copik, Grzegorz Kwaśniewski, Jürgen Müller, Łukasz Flis, Hannes Eberhard, Zixuan Chen, Hubert Niewiadomski, and Torsten Hoefler. 2025. Reasoning language models: A blueprint.

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.

Cao, Yang Trista 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. Association for Computational Linguistics, July.

Cho, Won Ik, 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. Association for Computational Linguistics.

Costa-jussà, Marta R. 2019. An analysis of Gender Bias studies in Natural Language Processing. *Nature Machine Intelligence*, 1(11):495–496.

Deutsch, Daniel, Eleftheria Briakou, Isaac Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. Wmt24++: Expanding the language coverage of wmt24 to 55 languages & dialects. *arXiv preprint arXiv:2502.12404*.

Dyer, Chris, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 644–648.

Elaraby, Mohamed and Christof Monz. 2018. Gender-aware neural machine translation. In *Proceedings of the Second International Conference on Natural Language and Speech Processing (ICNLSP)*. IEEE.

Font, Joel Escudé and Marta R Costa-jussà. 2019. Equalizing gender bias in neural machine translation with word embeddings techniques. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 147–154.

Gao, Mingqi, Xinyu Hu, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2025. Llm-based nlg evaluation: Current status and challenges.

Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences of the United States of America*, 120.

Habash, Nizar, Houda Bouamor, and Christine Chung. 2019. Automatic gender identification and reinflection in arabic. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 155–165.

Hirasawa, Tosho and Mamoru Komachi. 2019. Debiasing word embeddings improves multimodal machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 32–42.

Honnibal, Matthew, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Kappl, Michelle. 2025. Are all spanish doctors male? evaluating gender bias in german machine translation. *arXiv preprint arXiv:2502.19104*.

Kocmi, Tom, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, Mariya Shmatova, and Jun Suzuki. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In Koehn, Philipp, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore, December. Association for Computational Linguistics.

Kocmi, Tom, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In Haddow, Barry, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA, November. Association for Computational Linguistics.

Levy, Shahar, Koren Lazar, and Gabriel Stanovsky. 2021. Collecting a large-scale gender bias dataset for coreference resolution and machine translation. In Moens, Marie-Francine, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2470–2480, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Li, Haitao, Junjie Chen, Qingyao Ai, Zhumin Chu, Yujia Zhou, Qian Dong, and Yiqun Liu. 2025. CalibraEval: Calibrating prediction distribution to mitigate selection bias in LLMs-as-judges. In Che, Wanxiang, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16537–16552, Vienna, Austria, July. Association for Computational Linguistics.

Manna, Chiara, Afra Alishahi, Frédéric Blain, and Eva Vanmassenhove. 2025. Are we paying attention to her? investigating gender disambiguation and attention in machine translation. In Hackenbuchner, Janiça, Luisa Bentivogli, Joke Daems, Chiara Manna, Beatrice Savoldi, and Eva Vanmassenhove, editors, *Proceedings of the 3rd Workshop on Gender-Inclusive Translation Technologies (GITT 2025)*, pages 1–16, Geneva, Switzerland, June. European Association for Machine Translation.

Manna, Chiara, Hosein Mohebbi, Afra Alishahi, Frédéric Blain, and Eva Vanmassenhove. 2026. Gender disambiguation in machine translation: Diagnostic evaluation in decoder-only architectures. *arXiv preprint arXiv:2603.17952*.

Moryossef, Amit, Roee Aharoni, and Yoav Goldberg. 2019. Filling gender & number gaps in neural machine translation with black-box context injection. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 49–54.

Nasution, Arbi Haza and Aytuğ Onan. 2024. Chatgpt label: Comparing the quality of human-generated and llm-generated annotations in low-resource language nlp tasks. *IEEE Access*, 12:71876–71900.

Pangakis, Nicholas, Samuel Wolken, and Neil Fasching. 2023. Automated annotation with generative ai requires validation. *arXiv preprint arXiv:2306.00176*.

Piergentili, Andrea, Dennis Fucci, Beatrice Savoldi, Luisa Bentivogli, and Matteo Negri. 2023. Gender neutralization for an inclusive machine translation: from theoretical foundations to open challenges. In Vanmassenhove, Eva, Beatrice Savoldi, Luisa Bentivogli, Joke Daems, and Janiça Hackenbuchner, editors, *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, Matteo Negri, and Luisa Bentivogli. 2025. An llm-as-a-judge approach for scalable gender-neutral translation evaluation.

Prates, Marcelo O.R., Pedro H.C. Avelar, and Luis C. Lamb. 2020. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, 32(10):6363–6381.

Qian, Shenbin, Archchana Sindhujan, Minnie Kabra, Diptesh Kanojia, Constantin Orăsan, Tharindu Ranasinghe, and Frédéric Blain. 2024. What do large language models need for machine translation evaluation?

Ramesh, Krithika, Gauri Gupta, and Sanjay Singh. 2021. Evaluating gender bias in hindi-english machine translation. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 16–23.

Rescigno, Argentina Anna and Johanna Monti. 2023. Gender bias in machine translation: a statistical evaluation of google translate and deepl for english, italian and german. *Proceedings of the International Conference on Human-informed Translation and Interpreting Technology 2023*.

Rescigno, Argentina Anna, Eva Vanmassenhove, Johanna Monti, and Andy Way. 2020. A case study of natural gender phenomena in translation a comparison of google translate, bing microsoft translator

and deepl for english to italian, french and spanish. *Computational Linguistics CLiC-it 2020*, page 359.

Rudinger, Rachel, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14.

Saunders, Danielle and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7724–7736.

Savoldi, Beatrice, Jasmijn Bastings, Luisa Bentivogli, and Eva Vanmassenhove. 2025. A decade of gender bias in machine translation. *Patterns*, 6(6).

Stahlberg, Dagmar, F. Braun, L. Irmen, and Sabine Sczesny. 2007. Representation of the sexes in language. *Social Communication*, pages 163–187, 01.

Stanczak, Karolina and Isabelle Augenstein. 2021. A survey on gender bias in natural language processing. *arXiv preprint arXiv:2112.14168*.

Stanovsky, Gabriel, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

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.

Sun, Tony, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In Korhonen, Anna, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy, July. Association for Computational Linguistics.

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

Tan, Zhen, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Troles, Jonas-Dario and Ute Schmid. 2021. Extending challenge sets to uncover gender bias in machine translation: Impact of stereotypical verbs and adjectives. In Barrault, Loic, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz, editors, *Proceedings of the Sixth Conference on Machine Translation*, pages 531–541, Online, November. Association for Computational Linguistics.

Tseng, Yu-Min, Wei-Lin Chen, Chung-Chi Chen, and Hsin-Hsi Chen. 2025. Evaluating large language models as expert annotators. *arXiv preprint arXiv:2508.07827*.

Törnberg, Petter. 2024. Best practices for text annotation with large language models.

Vanmassenhove, Eva and Johanna Monti. 2021. gENder-IT: An annotated English-Italian parallel challenge set for cross-linguistic natural gender phenomena. In Costa-jussà, Marta R., Hila Gonen, Christian Hardmeier, and Kellie Webster, editors, *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 1–7, Online, August. Association for Computational Linguistics.

Vanmassenhove, Eva, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In Riloff, Ellen, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium, October-November. Association for Computational Linguistics.

Vanmassenhove, Eva, Chris Emmery, and Dimitar Shterionov. 2021. NeuTral Rewriter: A rule-based and neural approach to automatic rewriting into gender neutral alternatives. 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 8940–8948, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Vanmassenhove, Eva. 2024. Gender bias in machine translation and the era of large language models. In *Gendered Technology in Translation and Interpreting*, pages 225–252. Routledge.

Walshe, Thomas, Sae Young Moon, Chunyang Xiao, Yawwani Gunawardana, and Fran Silavong. 2025. Automatic labelling with open-source llms using dynamic label schema integration.

Xu, Haoran, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models. In *International Conference on Learning Representations (ICLR) 2024*.

Yang, An, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report.

Zhao, Jieyu, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20.

Zhu, Wenhao, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the association for computational linguistics: NAACL 2024*, pages 2765–2781.

Zmigrod, Ran, Sabrina J Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661.

A Appendix

LLM-based Annotation Prompts

We report below the full system prompts used for the zero-shot and few-shot annotation configurations. In the few-shot configuration, three exemplars per label are dynamically selected such that the profession under annotation does not appear among them.

Zero-shot System Prompt

```
You are a language expert specializing in grammatical gender in translations from
English into Italian.

Your task is to identify the grammatical gender expressed in the Italian
translation of a given English profession.

You will receive:
- An English source sentence
- Its Italian translation
- The English profession noun

Instructions:
1. Locate the Italian translation of the specified profession.
2. Identify the noun phrase (determiner + profession noun).
3. Determine the grammatical gender expressed in that noun phrase.

Use ONLY:
- The determiner
- The morphological form of the noun

If the noun itself does not present any gender marking but the determiner does,
report the gender of the latter.

Do NOT use:
- Pronouns
- Adjectives outside the noun phrase
- Verbal agreement
- Context
- English gender cues
- World knowledge

Assign one of the following labels:
- MASCULINE: if the determiner and noun clearly mark masculine gender.
- FEMININE: if the determiner and noun clearly mark feminine gender.
- UNKNOWN: if the noun phrase is gender-invariant, lacks gender marking, or
  shows conflicting signals.

If uncertain, output UNKNOWN.

Output format:
Label: <MASCULINE | FEMININE | UNKNOWN>

Return only this line. Do not provide explanations.
```

Few-shot System Prompt

```
You are a language expert specializing in grammatical gender in translations from
English into Italian.

Your task is to identify the grammatical gender expressed in the Italian
translation of a given English profession.

You will receive:
- An English source sentence
- Its Italian translation
- The English profession noun

Instructions:
1. Locate the Italian translation of the specified profession.
2. Identify the noun phrase (determiner + profession noun).
3. Determine the grammatical gender expressed in that noun phrase.

Use ONLY:
- The determiner
- The morphological form of the noun

If the noun itself does not present any gender marking but the determiner does,
report the gender of the latter.

Do NOT use:
- Pronouns
- Adjectives outside the noun phrase
- Verbal agreement
- Context
- English gender cues
- World knowledge

Assign one of the following labels:
- MASCULINE: if the determiner and noun clearly mark masculine gender.
- FEMININE: if the determiner and noun clearly mark feminine gender.
- UNKNOWN: if the noun phrase is gender-invariant, lacks gender marking, or
  shows conflicting signals.

If uncertain, output UNKNOWN.

Noun phrases like "il manager", "il professore", "il direttore" are MASCULINE.
Noun phrases like "la manager", "la professoressa", "la direttrice" are FEMININE.
Noun phrases like "l'assistente", "l'insegnante", "l'auditor" are UNKNOWN.

Output format:
Label: <MASCULINE | FEMININE | UNKNOWN>

Return only this line. Do not provide explanations.
```

From Binary Defaults to Contextual Bias: Translating Queer Morphology with NMT and LLMs

Manuel Lardelli
Department of Linguistic and Literary Studies
University of Padua / Italy
manuel.lardelli@unipd.it

Abstract

This paper evaluates how Neural Machine Translation (NMT) and Large Language Models (LLMs) process non-binary morphology when translating German literary fiction into Italian. 12 NMT and 16 LLM translations from a human-in-the-loop experiment were analyzed, applying an inductive, mixed-methods framework. Findings indicate clear differences between systems. NMT defaults to standard binary grammar but applies it inconsistently, often flipping between masculine and feminine forms for the same subject across different sentences, which effectively erases queer visibility. Conversely, LLMs actively attempt gender-fair language via neutralization and neomorphemes (e.g. the schwa). However, LLMs introduce new systematic errors: driven by semantic cues, they exhibit a contextual bias that frequently led to over-feminization in the present study, and their attempts to create inclusive word endings result in structurally invalid words, especially when translating clitic pronouns. Ultimately, these findings expose current limitations and provide preliminary empirical guidance to assist post-editors in navigating the complex challenges of gender-fair translation.

1 Introduction

Gender-fair language (GFL) is not a single, universal set of rules; instead, it is a fast-changing collection of different linguistic strategies to achieve inclusivity. Such strategies vary drastically across linguistic systems. For instance, English, a notional gender language,[1] primarily marks gender on third-person pronouns, where singular *they* has become a widespread convention for referring to non-binary individuals (Baron, 2020). Conversely, grammatical gender languages like German and Italian face the steeper challenge of pervasive binary agreement across articles, nouns, and adjectives. To achieve inclusivity, strategies such as the gender star (*) in German and the schwa (ə) in Italian are gaining traction amongst other possible approaches (Lardelli and Gromann, 2023b). This systemic fluidity, coupled with cross-lingual differences in gender marking, represents a profound challenge for language technologies.

Because Machine Translation (MT) architectures are optimized on massive, historically cisnormative datasets, they inherently struggle to process morphological phenomena that remain unsettled or statistically sparse within their training corpora (Vanmassenhove et al., 2019). Consequently, traditional Neural Machine Translation (NMT) architectures frequently misgender non-binary entities in target texts (Dev et al., 2021; Lauscher et al., 2023; Lardelli, 2023). In contrast, prompt-driven Large Language Models (LLMs) theoretically possess the semantic sensitivity and generative flexibility necessary to dynamically synthesize GFL (Brown et al., 2020). However, it remains unclear how these different models handle the complex grammar of non-binary language in continuous text, since most current evaluations rely on curated benchmarks made up of short, isolated examples (Robinson et al., 2024).

To bridge this gap, this paper presents an em-

[1] For a classification of languages based on their gender system see Stahlberg et al. (2007) and McConnell-Ginet (2013).

pirical, human-in-the-loop evaluation of how NMT and LLM process non-binary morphology in a literary context. Based on an experimental dataset generated by 12 participants translating contemporary German fiction into Italian, an inductive, mixed-methods framework was applied to analyze 12 NMT- and 16 LLM-generated translations. This study addresses a notable gap in the literature by focusing on gender-fair translation between two grammatically gendered languages – a direction often overlooked in favor of English-centric research (Savoldi et al., 2025b).

Findings show differences in handling non-binary morphology between systems. NMT systems are highly rigid; they default to traditional binary grammar or switch to plural forms, which effectively erases queer visibility. In contrast, while LLMs actively attempt to be inclusive, they often break grammatical rules and create structurally invalid words – a pattern also documented for other language pairs involving heavily inflected target languages (Popović et al., 2025) – especially when translating clitic pronouns. Ultimately, these findings expose the current limitations of transformer-based models in handling evolving queer language, offering empirical guidance for gender-fair post-editing.

2 Theoretical Background

To contextualize automated non-binary translation, this section examines the intersection of cultural history and linguistic structures in German and Italian. Section 2.1 traces the divergent evolution of GFL in these regions and its impact on current institutional readiness. Section 2.2 then outlines specific GFL strategies in both languages, while Section 2.3 analyzes how their respective grammatical foundations, specifically German's case declensions versus Italian's fusional suffixes, dictate the implementation of these strategies.

2.1 The Evolution of Gender-Fair Language: German-speaking countries vs. Italy

The theoretical and practical frameworks for GFL have developed along markedly different trajectories in German-speaking countries compared to Italy. In the DACH region, feminist linguistics emerged as a robust academic field in the 1970s and 1980s (Pusch, 1984; Trömel-Plötz, 1982). These early movements introduced typographical conventions like the *Binnen-I* (e.g. *LehrerInnen*, EN: teachers) to visualize female inclusion. Today, the use of GFL is moderately institutionalized across Germany, Austria, and Switzerland, often mandated or encouraged in administrative, academic, and broadcasting contexts (Feigl, 2009; Sczesny et al., 2016). Consequently, the contemporary DACH debate has largely shifted beyond the gender binary, focusing on non-binary typographical solutions such as the gender star (*Gendersternchen*, *) and the colon (*Doppelpunkt*, :).

Conversely, the systemic critique of Italian began slightly later, crystallized by Sabatini (1987). Unlike the DACH region, contemporary mainstream discourse in Italy remains heavily anchored to overcoming the generic masculine for prestigious professional titles (e.g. advocating for *la sindaca*, *l'avvocata*, *la ministra*; EN: the mayor, the lawyer, the minister) (Gheno, 2022). The transition toward non-binary GFL in Italian is highly contentious. Attempts to introduce inclusive typography, most notably the schwa (ə) and the asterisk (*), face stringent resistance from both institutional bodies and broader academic circles, who argue these interventions severely disrupt the morphological integrity of the language (Arcangeli, 2022; De Benedetti, 2022).

2.2 Overview of GFL Strategies in German and Italian

To navigate non-binary texts, writers and speakers employ a specific repertoire of GFL strategies. As categorized by Lardelli and Gromann (2023b), these approaches generally divide into gender-inclusive (which actively mark non-binary identities) and gender-neutral strategies (which neutralize the text by omitting gender-specific markers).

In German, visibility strategies frequently involve the adoption of neopronouns (such as *dey/demm* or *xier/xiem*) as direct substitutes for binary third-person pronouns. For nouns, German relies heavily on typographical interventions, such as the *gender star* (*), the colon (:), or the underscore (_), inserted before the feminine suffix (e.g. *Student*in*, EN: student) to create a visually inclusive spectrum. When these visibility markers are deemed too disruptive or grammatically incompatible, neutralization strategies, such as repeating a person's name instead of using a personal pronoun or syntactically restructuring sentences to avoid gender marking altogether, are utilized.

In Italian, the strategic repertoire is similarly di-

vided but heavily constrained by the language's pervasive inflectional system. Visibility strategies depend almost entirely on typographical neomorphemes, most prominently the schwa (ə) or the asterisk (*), used to replace terminal gendered vowels (e.g. *tuttə*, EN: everyone, or *bambin** EN: kid). However, due to the morphosyntactic friction these visual markers introduce, neutral strategies are often applied. This includes leveraging Italian's pro-drop nature to omit subject pronouns, alongside extensive syntactic restructuring using passive constructions, periphrasis, and epicene nouns.

2.3 Morphosyntactic Divergences: German vs. Italian

The different use of GFL in German-speaking countries and Italy is not merely cultural but deeply rooted in the morphosyntactic structures of the languages themselves. While both German and Italian are grammatical gender languages, their inflectional mechanics lead to different challenges in GFL use.

Feature	German	Italian
Declensions & Articles	Complex case declensions (e.g. *die, der*).	No case declensions, but articles remain gendered.
Possessives	Gendered possessives reflecting possessor (*sein/ihr*).	Gender-blind possessives reflecting possessed object (*suo/sua*).
Predicative Adjectives	Uninflected for gender/number.	Highly inflected for gender/number.
Subject Pronouns	Mandatory (non-pro-drop).	Optional (pro-drop).
Verbs & Participles	No gender inflection.	Past participles highly inflected.
Derivational Suffixes	Modular *-in* suffix, easily adaptable (e.g. *-*in*).	Fusional endings (*-a/-o, -tore/-trice*); disruptive typography.

Table 1: Morphosyntactic differences impacting GFL strategies.

As outlined in Table 1, the primary challenge in German stems from its complex case declensions. Integrating typographical strategies like the *gender star* (*) to combine definite articles introduces significant structural ambiguities for anyone using the language. Writers and speakers must resolve whether to prioritize the masculine or feminine form, or how to merge them effectively (e.g. *der*die*, *die*der*, or the fused *di*er*). Furthermore, when neomorphemes are introduced, there is a systemic need to generate several new endings based on both gender and case, requiring users to navigate entirely new declension paradigms. However, German benefits from uninflected predicative adjectives and a highly modular feminine suffix (*-in*), which readily accommodates visual neomorphemes like the aforementioned gender star.

In Italian, applying typographical neomorphemes is straightforward only when grammatical forms are symmetrical. When masculine and feminine words differ by a single terminal vowel (e.g. *ragazzo/ragazza*, EN:boy/girl), the ending can be cleanly substituted (*ragazzə*). However, morphosyntactic complications arise with asymmetrical suffixes, which are prevalent in professional titles (e.g. *-tore/-trice* in *scrittore/scrittrice*, EN: writer). Without a shared morphological stem, applying a neomorpheme in these cases forces an arbitrary reliance on either the masculine or feminine base (e.g. *scrittorə* vs. *scrittricə*). This structural asymmetry also deeply impacts pronouns. While Italian is a pro-drop language, allowing speakers to frequently omit subject pronouns, generating visible non-binary forms when they are explicitly required remains problematic. The standard third-person singular subject pronouns (*lui* and *lei*) differ in their central vowels rather than their terminal suffixes, preventing simple substitution. Furthermore, while direct object clitics are symmetrical (*lo/la*) and could theoretically accept a gender-fair vowel (*lə*), indirect object clitics are entirely asymmetrical (*gli/le*). Crucially, attempting to unify these object pronouns with a single neomorpheme (e.g. using *lə* for both) collapses the grammatical distinction between direct and indirect cases, generating severe syntactic ambiguity and further exposing the limitations of typographical interventions in heavily fusional languages.

3 Related Work

The integration of GFL into MT is a rapidly evolving subfield (Savoldi et al., 2025b). Foundational surveys mapping gender bias in MT initially emphasized the systemic erasure of female and non-binary representation in MT (Savoldi et al., 2021; Lardelli and Gromann, 2023a). While early debiasing efforts and benchmarking paradigms, e.g. WinoMT (Stanovsky et al., 2019), primarily focused on mitigating the generic masculine and resolving binary misgendering in occupational contexts, recent literature has pivoted toward the com-

plex morphosyntactic challenges of accommodating non-binary identities and neomorphemes in heavily inflected languages (Cao and Daumé III, 2020).

3.1 Evaluating Gender Bias in MT/LLMs

Traditional NMT architectures fundamentally struggle with context resolution, leading to rigid default biases. Early mitigation strategies often framed this as a domain adaptation problem, attempting to fine-tune models on gender-balanced datasets (Saunders and Byrne, 2020); however, these are mostly binary-focused approaches that fail to generalize to non-binary morphology. Lauscher et al. (2023) showed that commercial MT systems handle neopronouns poorly across the board: translation quality drops substantially compared to gendered pronouns, engines frequently copy neopronouns verbatim rather than translating them, and when a translation is attempted, it most often defaults to a binary gendered form in the target language. Moreover, Hackenbuchner et al. (2024) demonstrated that human gender assumptions and MT systems alike are heavily influenced by contextual cues, highlighting that isolated, word-level bias evaluation is insufficient to investigate gender bias in MT. While some commercial MT systems exhibit localized instances of inclusivity, such as providing gender-fair alternatives for plural nouns, they remain structurally biased toward the masculine (Lardelli et al., 2024b).

To systematically measure these limitations, recent work has introduced several benchmarks. Piergentili et al. (2023) introduced GeNTE, a corpus for benchmarking gender-neutral translation into Italian, providing multi-reference evaluations that expose the trade-offs between grammatical fluency and visible inclusivity. Expanding this line of work to German as a target language, Lardelli et al. (2024a) constructed a dedicated dataset for evaluating gender-fair MT into German, while Waldis et al. (2024) introduced the *Lou dataset*, a German corpus pairing conventional and gender-fair formulations of the same texts, and showed that the presence of GFL measurably affects the performance of text classification models. Most recently, Savoldi et al. (2025a) introduced mGeNTE, a multilingual extension of the GeNTE corpus covering multiple grammatical gender languages, enabling cross-linguistic evaluation of gender-neutral translation quality. Pranav et al. (2025) proposed GLITTER, a dataset explicitly designed for evaluating gender-fair German in MT and LLMs. Hackenbuchner et al. (2025) introduced GENDEROUS to assess how MT systems and LLMs translate gender-ambiguous occupations and singular *their* across multiple grammatical-gender languages, confirming a pervasive male-default in neural models. To scale the evaluation of these nuanced translations, Piergentili et al. (2025b) validated the efficacy of LLM-as-a-judge paradigms for scoring gender-neutral translation quality.

3.2 Generative LLMs and Gender-Fair Rewriting

The advent of generative artificial intelligence has shifted the paradigm from static translation to dynamic, gender-fair rewriting. Broad evaluations of LLMs as translators confirmed their superior contextual adaptability over traditional NMT (Wang et al., 2023) but also flagged their susceptibility to amplifying semantic biases present in their vast, uncurated training data. Narrowing in on GFL specifically, Piergentili et al. (2024) explored the use of LLMs to translate English into Italian using gender-inclusive neomorphemes such as the schwa. They noted that while LLMs possess generative flexibility lacking in NMT, they frequently struggle with consistent morphosyntactic application. Building upon this, Piergentili et al. (2025a) mapped the trade-offs of various models and approaches for gender-neutral rewriting in Italian, highlighting the persistent tension between grammatical fluency and visible inclusivity. Mainardi et al. (2025) compared prompting and fine-tuning strategies for gender-fair rewriting of MT outputs. They found that fine-tuning, even on smaller models, yielded superior morphosyntactic consistency and deeper task adaptation. In contrast, while large prompted models were capable of few-shot gender-fair generation, they struggled with higher misgeneration rates, hallucinating forms and overgenerating inclusive morphology.

3.3 Human Factors: Post-Editing and Comprehensibility

The shift toward automated GFL generation heavily impacts human-computer interaction within translation workflows. Lardelli (2025) investigated the cognitive load of MT post-editing, revealing that while post-editing GFL is usually faster than translation, the perceived cognitive effort is still

high due to the continuous need for structural arbitration by the linguist.

Despite the increased cognitive burden for translators, reader reception of these strategies remains largely positive. For instance, Daems (2026) proved in a real-world scenario that the presence of inclusive neomorphemes in translated texts does not negatively impact human processing time or comprehension success rates. Ultimately, to refine automated systems, Gromann et al. (2023) argue for the necessity of participatory research, ensuring that MT architectures are directly informed by the communities pioneering these linguistic innovations.

4 Methodology

This study draws on data from a translation and post-editing experiment involving 12 advanced translation students at the University of Padua (November 2025 – March 2026). Participants translated three German literary extracts into Italian under fixed conditions: human translation (Extract A), static NMT post-editing (Extract B), and LLM-assisted translation (Extract C). This paper restricts its analysis to the raw machine outputs of Extracts B and C. Rather than conducting a controlled model comparison, the outputs of the engines participants naturally selected for these tasks (DeepL, Google Translate, Reverso, ChatGPT-4o, Gemini, and Claude) were analyzed. In the LLM condition, users crafted their own prompts in unrestricted, multi-turn interactions (see Appendix B). This approach aligns with a human-in-the-loop framework, prioritizing ecologically valid workflow integration over systematic benchmarking.

4.1 Source Text and Translation Constraints

The source texts (ST) for the experiment are three extracts from *Schwindel* (Yaghoobifarah, 2024), a contemporary queer novel (see extracts in Appendix A).[2] Characterized by an informal, highly colloquial narrative style, the text actively employs non-binary morphology. One of the protagonists, Delia, is referenced using the German neopronoun *dey/demm* alongside inclusive typography (the gender gap/underscore) for nouns (e.g. *Kellner_in*, EN: server).

The extracts were specifically chosen for their high density of German GFL markers and the resulting morphological constraints they impose on the Italian target text. A secondary selection criterion was morphosyntactic variety, ensuring a comprehensive representation of subject, object, and possessive pronouns, as well as adjectives, participles, and both non-binary and plural nouns. To consistently meet these experimental requirements, the ST were slightly modified. As detailed in Table 2, Extracts B and C are comparable in length, word-class distribution, and the frequency of GFL segments. (A comprehensive segment-by-segment breakdown is supplied in Table 5 in Appendix C). Extract A was excluded from the analysis because it involved human translation.

Metric / Word Class	Extract B	Extract C
Total GFL Segments	11	10
Total Word Count	195	209
Subject pronoun	7	6
Object pronoun	1	5
Possessive pronoun	1	1
Non-binary noun	2	1
Plural noun	1	1
Adjective/participle	4	1

Table 2: Overview of segments with GFL, word counts, and word class frequencies for Extract B and Extract C.

The German-Italian language pair was chosen because it presents severe structural challenges for translation and language models. Unlike English, where gender is mostly limited to pronouns, German applies grammatical gender across articles, nouns, and adjectives. When writers introduce non-binary markers to German, translation and language models often misread them. This initial confusion makes it even harder for them to accurately translate queer identities into Italian, which propagates misgendering.

4.2 Data Annotation Framework

Departing from traditional MT evaluation benchmarks, this study assesses NMT and LLM performance "in the wild" using continuous, queer literary prose. To achieve this, a mixed-methods quantitative and qualitative content analysis was conducted (Krippendorff, 2018). Rather than applying a rigid, pre-defined evaluation metric, a custom annotation framework was developed inductively (Saldaña, 2021).

By systematically observing the raw NMT and LLM outputs without a pre-existing codebook, categories for recurring translation strategies, biases, and structural errors emerged directly from the em-

[2]Note that passages of the novel are written entirely in lowercase.

pirical data. This bottom-up approach provides significant methodological value by exposing systematic error typologies that metrics applied to curated benchmarks often obscure. The annotation was performed by a single researcher with expertise in translation studies and GFL. To ensure data consistency, the annotator performed a dual-pass verification process, reviewing the entire dataset a second time after a three-week interval to resolve any internal inconsistencies. The data-driven coding process yielded a micro-level typology consisting of three macro-categories:

1. Bias & Misgendering:
Misgendering (Feminine/Masculine): The system forcibly assigns a binary grammatical gender to a non-binary ST referent.
Generic Masculine: The system erases ST inclusive typography by defaulting to the standard Italian generic masculine.

2. Translation Failures:
Source Retention: The ST neopronoun is reproduced verbatim in the target segment without morphological adaptation (e.g. *Dey*).
Calque Translation: The system creates an interlingual hybrid or literal translation that violates target-language syntactic norms (e.g. hallucinating the Italian/German hybrid neo-pronoun *ley*).
Mistranslation / Misrecognition: The system translates the non-binary word using an incorrect grammatical category (e.g. translating a singular possessive as a plural).

3. Non-binary Translation:
Pro-drop: The system achieves syntactic neutrality passively by omitting the subject pronoun, leveraging standard Italian pro-drop parameters.
Neutralization: The system actively reformulates the sentence or substitutes the pronoun with the character's name to bypass gender markers.
Neomorphemes (ə/ *): The system generates contemporary typographical inclusivity to mirror the source text.

5 Results

This section presents findings on how NMT and LLMs translate German non-binary morphology into Italian.[3] First, a general evaluation of the NMT models is provided, followed by a detailed analysis of the specific morphosyntactic triggers that drive their behaviors. The section then explores the generative performance of LLMs, similarly analyzing the morphosyntactic triggers behind their outputs. Finally, the findings conclude with a comparative analysis, evaluating the LLMs against one another before synthesizing the fundamental differences between the LLM and NMT approaches.

5.1 NMT Evaluation: Binary Default

Ten of the twelve participants generated their translation drafts using DeepL (P1–P2, P5–P12), while two used Google Translate (P3) and Reverso (P4). The task extract contained 11 segments featuring German GFL or requiring it in the Italian translation, spanning 16 items. Consequently, 192 annotations were performed.

As detailed in Table 3, the quantitative distribution of NMT outputs reveals that the systems overwhelmingly rely on evasion or default to binary categories. The most frequent strategy is the syntactic omission of the subject (pro-drop), accounting for 33.9% (65 occurrences) of the total annotations. When evasion is structurally impossible, the models enforce a strict binary paradigm, resulting in feminine (29.2%, 56 occurrences) and masculine misgendering (11.5%, 22 occurrences). Moreover, the systems often fail to translate the neopronouns at all, leaving the original German words in the target texts (8.9%, 17 occurrences) or defaulting to the generic masculine (6.3%, 12 occurrences). Strikingly, instances where the translation is naturally correct without requiring specific GFL interventions constitute a mere 3.6% of the data, underscoring the systemic inadequacy of NMT in handling non-normative gender markers.

Annotation Category	Count	Percentage
Pro-drop	65	33.9%
Misgendering (Feminine)	56	29.2%
Misgendering (Masculine)	22	11.5%
Source Retention	17	8.9%
Generic Masculine	12	6.3%
Misrecognition	11	5.7%
Correct / No GFL Required	7	3.6%
Mistranslation	2	1.0%
Total	**192**	**100.0%**

Table 3: Frequency of NMT translation solutions ($n = 192$)

[3]The generated translations are available at `https://zenodo.org/records/20137067`

5.1.1 Morphosyntactic Triggers in NMT

A detailed look at the NMT outputs shows that translation strategies and error rates depend heavily on the specific type of word being translated into Italian. Grouping the source segments by word class reveals four distinct patterns in how the models handled the text. A breakdown of translation strategies applied per word class is shown in Table 6 in Appendix C.

1. Syntactic Evasion: When processing the neopronoun *dey*, the models overwhelmingly rely on Italian pro-drop syntax, omitting the subject pronoun. For instance, Segment 1 of the extract was translated as *Delia soffre di vertigini fin da quando [Ø] era bambin**a*** (EN: Delia has been struggling with a fear of heights since they were a child) in 11 out of 12 target texts. The sole exception was Reverso, which completely mistranslated the segment (*la paura dell'altitudine affligge **la sua famiglia** da quando era piccola*, EN: Her family has struggled with a fear of heights ever since she was a child) by replacing the pronoun *dey* with *la sua famiglia*. While pronoun omission sometimes yields accidental neutrality, it frequently fails to prevent misgendering in adjacent agreements, as the dropped subject still governs gendered nouns and past participles (e.g. *bambin**a*** in the example).

2. Source Text Retention: When sentences omit the non-binary protagonist's name and rely solely on identified pronouns (*weit weg, als wäre dey längst hineingesprungen und würde hinuntersinken*, EN: far away, as if they had jumped in long ago and were sinking down), the evasion strategy shifts. For instance, in 10 out of 12 translations for Segment 8, NMT engines retained the German neopronouns verbatim (e.g. *lontane, come se Dey ci fosse già saltata*[4] *dentro e stesse affondando*). Interestingly, the source neopronoun was capitalized in the translation, suggesting the system might interpret the unfamiliar token as a proper noun.

3. Morphological Instability: Consistent with known limitations of sentence-level NMT processing, the systems fail to propagate gender identity across anaphoric chains. In early segments, the translation systems systematically apply feminine morphological agreements to the non-binary protagonist (56 occurrences). This pattern correlates with the presence of the source-language name *Delia*, which typically carries strong feminine associations in the training data. As the document progresses, however, the systems transition to masculine agreements (e.g. the uniform masculinization of the loanword *freak* in all the 12 translations analyzed). These shifts indicate a failure to maintain consistent document-level coreference for non-binary entities.

4. Misrecognition & Generic Masculine Default: The data exposes a case of misrecognition (11 out of 12 occurrences) in segment 7 containing the nominal phrase *deren rückenhaltung* (EN: [singular] their posture). The possessive pronoun *deren* is translated as *la **loro** postura* instead of *la **sua** postura*, as if it referred to a plural antecedent. Finally, all systems produce a translation in the generic masculine (*tutti i compagni*, EN: all classmates) in response to ST inclusive typography (*Mitschüler_innen*), reaffirming the structural bias of the engines toward masculine forms.

5.2 LLM Evaluation: Contextual Bias

For Extract C, seven participants used ChatGPT (P1–P2, P5, P7, P9–P11), three used Gemini (P4, P6, P12), one used both (P3), whereas one used ChatGPT and Claude (P8). Additionally, P11 generated three distinct versions with ChatGPT. This yielded 16 analyzed translations containing 15 items requiring GFL, totaling 240 annotations. Unlike the mistranslations observed in the NMT outputs, the LLMs correctly translated the possessive pronoun *deren*. Because the correct Italian equivalent inherently is not subject to gender inflection regarding the possessor, its corresponding annotations were excluded from Table 4.

Table 4 illustrates a markedly more diverse, albeit highly unstable, distribution of translation strategies employed by generative models. Unlike NMT, LLMs actively attempt to deploy contemporary gender-fair morphology, with the use of neomorphemes such as the schwa (10.2%, 23 occurrences) and the asterisk (3.6%, 8 occurrences) representing significant portions of the output. Additionally, LLMs utilize semantic neutralization strategies, including proper noun substitution (instead of pronouns) (4.9%) and periphrasis (3.6%). Despite these active inclusivity efforts, misgender-

[4]Note that the past participle is morphologically marked, incorrectly assigning feminine agreement to Delia. Furthermore, DeepL's translations for this segment exhibited inconsistent binary defaults, erratically alternating between masculine and feminine form.

ing remains the single largest error category, comprising 32.9% (74 occurrences) of the total outputs. Furthermore, LLMs often evade gender-fair requirements by retaining source text (16%) or producing unnatural calques (4%).

Annotation Category	Count	Percentage
Misgendering (Feminine)	74	32.9%
Source Retention	45	20%
Pro-drop	36	16%
Neomorpheme (Schwa, ə)	23	10.2%
Neutralization (Name instead of pronoun)	11	4.9%
Calque Translation	9	4%
Neutralization (Periphrasis)	8	3.6%
Neomorpheme (Asterisk, *)	8	3.6%
Mistranslation	4	1.8%
Neutralization (Synonym)	2	0.8%
Other (Hybrids, Removed)	5	2.2%
Total	**225**	**100.0%**

Table 4: Frequency of LLM translation strategies ($n = 225$)

5.2.1 Sensitivity to Prompting Formulation

While generative LLMs offer high adaptability, an analysis of the user prompts reveals relatively low variance in their prompting behavior (see prompts used during the experiment in Appendix B). The majority of participants (11 out of 12) utilized constraint-based prompting, explicitly instructing the models to employ gender-inclusive or gender-neutral language. Within this constrained group, prompts generally prioritized target-language fluency over explicit morphological instructions. For instance, participants requested that the GFL adapt to standard Italian (P5) or sound natural and idiomatic (P10). Notably, no participant explicitly commanded the use of specific micro-level strategies, such as neomorphemes (schwa/asterisk) or active neutralization. Only one participant (P11) provided the model with explicit ST context, specifying the protagonist's use of a neopronoun. Conversely, participant P1 provided a purely unconstrained, zero-shot prompt in German (*Kannst du es ins Italienische übersetzen?*, EN: Can you translate it into Italian?), providing a baseline for unguided LLM behavior.

The presence or absence of these GFL constraints possibly dictated algorithmic output. When the LLM was left unconstrained (P1), it defaulted almost entirely to binary misgendering, closely mirroring the rigidity observed in static NMT systems. The single exception was the translation of *Kiffer_in*, which the model spontaneously neutralized via periphrasis (*persona che fumava*, EN: person who used to smoke). Interestingly, when participants specifically requested "standard" or "idiomatic" GFL, the models paradoxically generated interlingual calques for the neopronouns *dey/demm* (e.g. hallucinating the forms *ley/lem*). This suggests that LLMs struggle to reconcile the instruction for grammatical fluency with the demand for visible inclusivity, occasionally attempting to resolve the conflict by inventing grammatically incorrect word forms. While these qualitative patterns highlight the models' sensitivity to prompt formulation, the limited sample size precludes drawing definitive conclusions.

Beyond the initial translation generation, 11 of the 12 participants engaged in multi-turn, iterative prompting. These follow-up interactions involved querying the models for metalinguistic explanations of their chosen GFL strategies or requesting alternative translations. Although these conversational logs provide rich qualitative data regarding embedded language ideologies and instances of generative hallucination (e.g. models fabricating grammatical justifications for ungrammatical outputs), an in-depth analysis of multi-turn user-machine interactions falls outside the scope of this paper and will be addressed in future work.

5.2.2 Morphosyntactic Triggers in LLMs

A closer look at the LLM outputs shows that translation strategies and error rates depend heavily on the specific type of word being translated into Italian. Grouping the source segments by word class reveals four distinct patterns in how the models handled the text. A detailed breakdown of the strategies used for each category is available in Table 7 in Appendix C.

1. Syntactic Evasion: When confronted with the subject neopronoun *dey*, LLMs leveraged Italian pro-drop parameters (36 occurrences for subject pronouns, 16% of the total outputs), achieving perfect neutrality. Some models, however, attempted interlingual morphosynthesis. Rather than retaining the neopronoun *dey*, they hallucinated interlingual calques (9 occurrences, 4%), inventing anomalous hybrid Italian–German neopronouns like *ley* (subject) and *lem* (object). This seems to indicate that LLMs are able to dynamically synthesize cross-lingual morphological rules.

2. Contextual Bias: While NMT defaults frequently to the masculine, LLMs exclusively de-

faulted to the feminine (74 occurrences, zero masculine occurrences). This highlights a shift from grammatical bias to contextual bias. Although the ST explicitly uses non-binary pronouns, the presence of female-coded names (Delia, Ava) and lexical items associated with sapphic interactions (e.g. *küsste*, EN: kissed; *liebhaber_innen*, EN: lovers) seem to trigger feminine agreement in the translations. This pattern is consistent across models and prompts, and might reflect the statistical weight of these contextual cues in the models' training distributions.

3. Misgendering in Clitic Pronouns: This error is highly concentrated on object pronouns. When translating the German non-binary pronoun *demm* (75 instances), the LLMs misgender the protagonist as female 48 times, likely triggered by the nearest contextual gender anchor (the female-coded name *Delia*). LLMs systematically force Italian clitics into the feminine form, e.g. *Ava tirò il mento di Delia verso di sé e* ***la*** *baciò*, (EN: Ava pulled Delia's chin toward her and kissed **her**, P2, ChatGPT generated translation for Segment 2). This reveals a specific structural blind spot: even when LLMs attempt to use inclusive language elsewhere in the sentence, they fail to apply these gender-neutral rules to small, highly gendered function words.

4. Morphological Instability: Segments containing German inclusive typography triggered the highest rates of morphological instability. For the singular noun *Kiffer_in* (EN: pothead) in Segment 1, models successfully utilized neutralization via periphrasis (5 occurrences, e.g. *persona che si faceva le canne*, EN: someone who used to smoke joints, P2, ChatGPT) and schwa (5 occurrences, e.g. *buonə fattonə*, EN: big stoner – P6, Gemini), but explicitly hallucinated broken strings, for instance mixing schwa and asterisk (*fumatorə**, P7, ChatGPT) or schwa and underscore (*fumat_ə*, P11, ChatGPT). In other words, when attempting to use inclusive typographical symbols, LLMs struggled to process the internal structure of the words, ultimately generating grammatically broken or invalid forms. The models generated typographical neologisms in $\sim$15% of the outputs but failed to integrate them into standard Italian morphosyntax. Furthermore, models mechanically over-applied the schwa to nouns that are already epicene in Italian plural paradigms (generating the redundant *amantə*, EN: lovers – P8, ChatGPT, Segment 6). In extreme cases, models attached typographical symbols to explicitly gendered suffixes (*fumator*a*, EN: smoker – P4, Gemini, Segment 1). This pattern points to a broader challenge in generating typographically inclusive forms in heavily inflected languages.

5.2.3 Comparing LLMs

Although the limited sample size precludes broad generalization, a comparative breakdown of the generative models reveals distinct tendencies in how they handle gender-fair constraints and structural morphosyntax within this dataset.

The Gemini outputs exhibited a pronounced reliance on structurally "safe" translational workarounds. The model prioritized semantic neutralization strategies, specifically *neutralization via periphrasis* (e.g. rendering the non-binary noun as *persona che fuma*, EN: person who smokes) and *proper noun substitution* for pronouns (11 occurrences). However, when Gemini utilized typography-based morphology, it produced higher rates of grammatical hallucination, demonstrating a systemic struggle to attach typographical symbols correctly to Italian stems.

Conversely, the ChatGPT iterations displayed a compelling tendency toward interlingual neogenesis. ChatGPT outputs included hybrid Italian neopronouns (e.g. *ley* for the subject, *lem* for the object) that formally resemble a combination of Italian feminine pronouns (*lei/le*) with the German non-binary suffixes (*-ey/-em*).

While the limited sample size (n=1) for Claude precludes a direct quantitative comparison, the model's output provided a distinct qualitative data point. Claude's general tendency was to default to source retention for subject pronouns, reproducing the German neopronoun verbatim in the Italian target text. However, when processing object pronouns, Claude systematically misgendered Delia in the target-language clitic pronouns.

5.3 Comparing NMT and LLMs

The frequency analysis uncovers a fundamental divergence in how NMT systems and LLMs process queer morphology. This divide can be categorized into three primary operational paradigms:

1. Passive Evasion vs. Active Neutralization: Both types of systems lean heavily on target-language syntactic parameters to navigate

untranslatable pronouns. NMT systems dropped the subject pronoun in 33.9% of their total outputs. While LLMs also utilized pro-drop frequently (17.1%), they supplemented this evasion with active, fluency-preserving neutralization strategies. Tactics such as *neutralization via periphrasis* (3.8%) and *proper noun substitution* (5.2%) were leveraged by LLMs but were entirely absent from NMT outputs, highlighting the generative models' superior semantic flexibility.

2. Grammatical Default vs. Contextual Bias: The nature of the bias exhibited by the two paradigms varies significantly. NMT outputs are characterized by an oscillating binary default. They generated masculine agreements (11.5%) for nouns like *freak* and defaulted to the generic masculine (6.3%) for plurals. Conversely, LLMs produced zero instances of masculine misgendering. Instead, LLMs are caught in a "feminization trap" (accounting for 32.9% overall misgendering, concentrated overwhelmingly on object clitics). In this paradigm, the semantic inference of a lesbian romance seems to completely override the explicit grammatical instructions for neutrality.

3. The Limits of Prompted Inclusivity: While static NMT models simply delete or ignore source-text inclusive typography, generative LLMs actively attempt to replicate it. However, the data suggests that while LLM generation of neomorphemes accounts for roughly 15% of their total translation strategies, its application remains highly unstable. The outputs consistently show that neomorphemes are applied inconsistently to nouns and adjectives while entirely failing to reconcile them with the surrounding clitic pronoun chains, ultimately breaking the syntactic cohesion of the translated sentence.

6 Discussion & Conclusion

The empirical data highlights a fundamental divergence in how NMT and LLMs process non-binary morphology, significantly impacting both the practical machine translation post-editing workflow.

NMT errors primarily stem from sentence-level processing limitations. Without document-level context, these systems assign gender inconsistently, often alternating between masculine and feminine for the same referent. In contrast, LLMs demonstrate a distinct error pattern driven by contextual bias rather than random oscillation. Specifically, the LLMs evaluated here systematically defaulted to feminine agreements when encountering female-coded names or semantic cues associated with sapphic contexts. This indicates that strong semantic features in the source text can override explicit prompt instructions for gender neutrality. Although output analysis alone cannot isolate the underlying model mechanics, the consistency of this bias provides an important empirical observation that warrants further investigation.

While LLMs successfully rephrase sentences to create natural, gender-neutral Italian, their attempts to use inclusive symbols (like the schwa or asterisk) expose major flaws. Rather than adapting the underlying Italian grammar, the models often treat these markers as simple visual add-ons. The resulting output fractures target-language syntax and degrades readability, demonstrating that current LLMs cannot reliably balance grammatical integrity with visible queer morphology – a finding consistent with observations for other inflected language pairs (Popović et al., 2025). Ultimately, both systems default to erasing queer identities without human intervention.

The contrast between NMT and LLM outputs necessitates different post-editing workflows. By mapping the typical error profiles of these two systems, this study provides actionable guidance on the specific pitfalls translators must check for. Specifically, NMT requires exhaustive, document-wide verification to rectify erratic binary agreements and inconsistent pronoun shifts. Post-editing LLMs, however, demands complex structural arbitration. Translators working with LLMs must actively counter the models' contextual biases by manually inserting inclusive morphology into evasive paraphrases, or by repairing the syntactic fractures that emerge when the models attempt to force typographical inclusivity onto clitics and other complex structures.

While the findings expose clear trends, this study is not without limitations. The human-in-the-loop experiment relies on a relatively small sample size of participants and focuses on extracts from a single contemporary novel. Furthermore, the findings are inherently constrained to the German–Italian language pair. Future research should expand the participant pool, test a wider variety of genres, and evaluate how these models handle non-binary morphology across other morphologically rich languages.

References

Arcangeli, Massimo. 2022. *La Lingua Scəma: Contro lo Schwa (e Altri Animali)*. Roma: Castelvecchi Editore.

Baron, Dennis. 2020. *What's Your Pronoun?: Beyond He and She*. Liveright Publishing.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Cao, Yang Trista 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.

Daems, Joke. 2026. Gender-inclusive translations put to the test: Measuring performance of quadball referee certification test takers. *The Journal of Specialised Translation*, 45(45):109–129.

De Benedetti, Andrea. 2022. *Così non Schwa. Limiti ed Eccessi del Linguaggio Inclusivo*. Torino: Einaudi.

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 *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994.

Feigl, Susanne. 2009. Geschlechtergerechte stellenausschreibung. *Unabhängiger Bericht der Gleichbehandlungsanwaltschaft iS*, 3.

Gheno, Vera. 2022. *Femminili singolari+: Il femminismo è nelle parole*. Effequ.

Gromann, Dagmar and Manuel Lardelli. 2023. Participatory research as a path to community-informed, gender-fair machine translation. In *Proceedings of the 1st International Workshop on Gender-Inclusive Translation Technologies*, pages 49–59. European Association for Machine Translation (EAMT).

Hackenbuchner, Janiça, Joke Daems, Arda Tezcan, and Aaron Maladry. 2024. You shall know a word's gender by the company it keeps: Comparing the role of context in human gender assumptions with mt. In *Proceedings of the 2nd International Workshop on Gender-Inclusive Translation Technologies*, pages 31–41, Sheffield, United Kingdom. European Association for Machine Translation (EAMT).

Hackenbuchner, Janiça, Eleni Gkovedarou, and Joke Daems. 2025. GENDEROUS: Machine translation and cross-linguistic evaluation of a gender-ambiguous dataset. In *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 302–319, Vienna, Austria. Association for Computational Linguistics.

Krippendorff, Klaus. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Lardelli, Manuel and Dagmar Gromann. 2023a. Gender-fair (machine) translation. In *Proceedings of the New Trends in Translation and Technology Conference - NeTTT 2022*, pages 166–177.

Lardelli, Manuel and Dagmar Gromann. 2023b. Translating non-binary coming-out reports: Gender-fair language strategies and use in news articles. *The Journal of Specialised Translation*, (40):213–240.

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

Lardelli, Manuel, Dagmar Gromann, and Janiça Hackenbuchner. 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).

Lardelli, Manuel. 2023. Post-editing machine translation beyond the binary: Insights into gender bias and screen activity. *Translating and the Computer*, 45:50–64.

Lardelli, Manuel. 2025. Faster, but not less demanding: Comparing gender-fair post-editing with translation. *Translation Spaces*, 14(1):120–145.

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 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 377–392, Toronto, Canada, July. Association for Computational Linguistics.

Mainardi, Paolo, Federico Garcea, and Alberto Barrón-Cedeño. 2025. Fine-tuning vs prompting techniques for gender-fair rewriting of machine translations. In Faleńska, Agnieszka, Christine Basta, Marta Costa-jussà, Karolina Stańczak, and Debora Nozza, editors, *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 320–337, Vienna, Austria, August. Association for Computational Linguistics.

McConnell-Ginet, Sally. 2013. Gender and its relation to sex: The myth of 'natural'gender. In Corbett, Greville G, editor, *The Expression of Gender*, pages 3–38. De Gruyter Mouton.

Piergentili, Andrea, Beatrice Savoldi, Dennis Fucci, Matteo Negri, and Luisa Bentivogli. 2023. 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.

Piergentili, Andrea, 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*, pages 300–314, Sheffield, UK. European Association for Machine Translation (EAMT).

Piergentili, Andrea, Beatrice Savoldi, Matteo Negri, and Luisa Bentivogli. 2025a. Gender-neutral rewriting in Italian: Models, approaches, and trade-offs. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 899–910, Cagliari, Italy. CEUR Workshop Proceedings.

Piergentili, Andrea, Beatrice Savoldi, Matteo Negri, and Luisa Bentivogli. 2025b. An LLM-as-a-judge approach for scalable gender-neutral translation evaluation. In *Proceedings of the 3rd International Workshop on Gender-Inclusive Translation Technologies (GITT)*. Association for Computational Linguistics.

Popović, Maja, Ekaterina Lapshinova-Koltunski, and Anastasiia Göldner. 2025. Did I (she) or I (he) buy this? or rather I (she/he)? towards first-person gender neutral translation by LLMs. In Hackenbuchner, Janiça, Luisa Bentivogli, Joke Daems, Chiara Manna, Beatrice Savoldi, and Eva Vanmassenhove, editors, *Proceedings of the 3rd Workshop on Gender-Inclusive Translation Technologies (GITT 2025)*, pages 64–73, Geneva, Switzerland, June. European Association for Machine Translation.

Pranav, A, Janiça Hackenbuchner, Giuseppe Attanasio, Manuel Lardelli, and Anne Lauscher. 2025. Glitter: A multi-sentence, multi-reference benchmark for gender-fair German machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*. Association for Computational Linguistics.

Pusch, Luise F. 1984. *Das Deutsche als Männersprache: Aufsätze und Glossen zur feministischen Linguistik*. Suhrkamp, Frankfurt am Main.

Robinson, Kevin, Sneha Kudugunta, Romina Stella, Sunipa Dev, and Jasmijn Bastings. 2024. MiTTenS: A dataset for evaluating gender mistranslation. In Al-Onaizan, Yaser, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4115–4124, Miami, Florida, USA, November. Association for Computational Linguistics.

Sabatini, Alma. 1987. *Il sessismo nella lingua italiana*. Presidenza del Consiglio dei Ministri, Rome.

Saldaña, Johnny. 2021. *The Coding Manual for Qualitative Researchers*. SAGE Publications, London, 4th edition.

Saunders, Danielle and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics*, pages 7724–7736.

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

Savoldi, Beatrice, Giuseppe Attanasio, Eleonora Cupin, Eleni Gkovedarou, Janiça Hackenbuchner, Anne Lauscher, Matteo Negri, Andrea Piergentili, Manjinder Thind, and Luisa Bentivogli. 2025a. Mind the inclusivity gap: Multilingual gender-neutral translation evaluation with mGeNTE. In Christodoulopoulos, Christos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 13698–13720, Suzhou, China, November. Association for Computational Linguistics.

Savoldi, Beatrice, Jasmijn Bastings, Luisa Bentivogli, and Eva Vanmassenhove. 2025b. A decade of gender bias in machine translation. *Patterns*, 6(6).

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

Stahlberg, Dagmar, Friederike Braun, Lisa Irmen, and Sabine Sczesny. 2007. Representation of the sexes in language. *Social Communication*, pages 163–187.

Stanovsky, Gabriel, Noah A Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1679–1684.

Trömel-Plötz, Senta. 1982. *Frauensprache: Sprache der Veränderung*. Fischer Taschenbuch Verlag, Frankfurt am Main.

Vanmassenhove, Eva, Dimitar Shterionov, and Andy Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*, pages 222–232.

Waldis, Andreas, Joel Birrer, Anne Lauscher, and Iryna Gurevych. 2024. The Lou dataset - exploring the impact of gender-fair language in German text

classification. In Al-Onaizan, Yaser, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10604–10624, Miami, Florida, USA, November. Association for Computational Linguistics.

Wang, Longyue, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661.

Yaghoobifarah, Hengameh. 2024. *Schwindel*. Berlin: Blumenbar.

A Source Texts

The following German source texts were extracted from the contemporary queer novel *Schwindel* (2024) by Hengameh Yaghoobifarah. To ensure a robust evaluation of the models' morphosyntactic capabilities, these extracts were slightly modified from the original published version. These targeted modifications guarantee that all GFL requirements and specific non-binary markers (such as *dey/demm/deren* neopronouns and the *_in* suffix) discussed in the main paper are consistently represented across the test set.

Extract A *mit viel kraft öffnete dey die schwere eingangstür. vielleicht ist ava stoned eingenickt, denkt delia beim warten auf den aufzug, die türklingel ist krass penetrant, ihr sound hat demm schon an zu vielen morgenden zu früh geweckt, aber man weiß ja nie. immerhin ist ava zuhause. ohne das handy ist delia sonst aufgeschmissen. als dey am späten nachmittag nach hause gekommen war, konnte dey es nirgendwo finden. in keiner tasche, in der eigenen wohnung sowieso nicht, und auf dem radweg zur arbeit hatte dey es nicht angerührt, dey war ohnehin schon spät dran gewesen und hatte andere sorgen, als eine passende playlist für unterwegs rauszusuchen. da dey als kellner_in in einem gut besuchten lokal arbeitet, war es demm während der schicht zu stressig, auf dem display nachrichten zu checken. delias chef hat demm eh auf dem kieker wegen unpünktlichkeit, verträumtheit, tollpatschigkeit und allgemeiner unzuverlässigkeit, spaß macht der job nicht, aber welcher tut das schon? außerdem hat sich der selbst ziemlich unorganisierte chef keinen gefallen damit getan, sein lokal ausgerechnet auf dem platz an der unüberschaubarsten straßenkreuzung der stadt zu eröffnen, wo delia unter den riesigen leuchtreklamen, den scharen an tourist_innen und dem verkehrslärm bereits völlig reizüberflutet die schicht antritt.*

Extract B *die höhenangst plagt delia schon seit dey kind war. und je mehr leute demm anglotzten, desto weiter scheint der abstand zum boden zu werden. als delia in der achten klasse vor allen mitschüler_innen ins schwimmbecken springen sollte, wurde dey ohnmächtig. locker zehn minuten hatte delia davor starr dagestanden. wobei, es hätten auch zwei sein können. dey hatte jegliches zeitgefühl verloren. alles, worauf dey sich fokussierte, war das raster der hellblauen fliesen, die durch die wasseroberfläche ganz verzerrt aussahen. wibbelwobbel. delia kann sich nicht daran erinnern, ob die matrix der realität jemals mit geraden linien gezeichnet worden war. alles war schon immer so krumm. von deren rückenhaltung ganz zu schweigen. wibbelwobbel. delias körper in bewegung - wibbelwobbel. die feuchte luft im hallenbad - wibbelwobbel. die stimmen um delia herum - wibbelwobbel. weit weg, als wäre dey längst hineingesprungen und würde hinuntersinken. war es möglich, zu ertrinken, ohne sich überhaupt im wasser zu befinden? delia wollte nicht rein, aber noch weniger wollte dey, dass der körper angegafft wurde. wibbelwobbel. delia hatte beim anstehen gehört, wie die mädchen im flüsterton darüber lästerten, dass dey eine badeshorts »wie von einem opa« und dazu ein langes schwimmshirt trug. wibbelwobbel. was für ein freak.*

Extract C *als kiffer_in saugte dey alles auf, hielt den rauch für einige sekunden und atmete ihn schließlich aus. ava zog delias kinn zu sich und küsste demm. es fiel delia leicht, sich darauf einzulassen, die anspannung fiel schlagartig von ihr ab. als ihre lippen sich voneinander lösten, legte ava demm den joint an den mund und dey nahm einen tiefen zug aus ihrer hand. »schau mich dabei an«, forderte ava demm auf. delia zog noch mal dran und öffnete die augen diesmal. »gut so.« delia hielt den nebel solang dey konnte, bis ava demm erlöste und alles wieder heraussaugte. so reichten sie sich den joint hin und her, bis er heruntergebrannt war und ava ihn ausdrückte. die liebhaber_innen lagen nebeneinander. ava zog delia an der taille zu sich und hört mit dem küssen nicht auf. dey hielt die augen zu. delia rechnete damit, dass jeden moment der sog einsetzte. dieses zappen durch die gedanken, durch die orte, durch szenen, dem dey sich nicht entziehen konnte. doch delia blieb hier. die berührungen fühlten sich einfach so gut an, dass deren kopf einen besseren deal witterte, wenn dey präsent blieb, statt sich wegzubeamen. delia hatte angst vor dem moment, die eigenen arme benutzen zu müssen, doch dazu kam es gar nicht.*

B Participant Prompts for Generative LLMs

To evaluate dynamic, human-in-the-loop generative translation, participants were tasked with prompting LLMs to translate the source extracts into Italian while adhering to GFL principles. The prompts used during the experiment varied in length, context, and specificity, reflecting a range of realistic zero-shot prompting strategies. The exact prompts submitted by participants to the models are listed below.

P1 (Italian/German):

```
Promptkannst du es ins italianische
übersetzen?
```

P2 (Italian):

```
Traduci il seguente estratto in
italiano mantenendo un linguaggio di
genere inclusivo e mantendo comunque
uno stile informale
```

P3(Italian):

```
Traduci il seguente testo in italiano,
facendo attenzione a:  mantenere il
tono del testo originale, rispettare
il linguaggio di genere neutro o
inclusivo, usare un italiano fluido
e naturale, adattare riferimenti
culturali e giochi di parole
```

P4 (Italian):

```
mi serve una traduzione dal tedesco
all'italiano.  voglio che rispetti
il linguaggio di genere e he il
linguaggio sia colloquiale
```

P5 (Italian):

```
Traduci il seguente testo, scritto
in tedesco, in italiano standard.
Mantieni il tono ironico, lo stile
informale e il linguaggio inclusivo
di genere, rendendolo in italiano
standard.
```

P6 (Italian - Context Aware):

```
mi servirebbe che traducessi un
passaggio di un romanzo dal tedesco
all'italiano.  si tratta di questo
romanzo:  Schwindel è un romanzo
che esplora relazioni affettive e
desiderio in una prospettiva queer,
con un linguaggio fluido, ironico
e colloquiale.  L'autrice utilizza
un registro contemporaneo e un uso
consapevole del linguaggio inclusivo
(ad esempio \Liebhaber_innen") per
esprimere la complessità delle
identità di genere e delle dinamiche
relazionali.  La traduzione italiana
deve riflettere questa attenzione,
mantenendo equilibrio tra leggibilità
e fedeltà all'intento politico e
stilistico del testo originale.
queste sono le indicazioni traduttive:
Tono e stile:  Mantenere la vivacità
e l'ironia del testo originale;
evitare toni troppo formali o
letterari.  Inclusività linguistica:
Rispettare il linguaggio di genere
neutro o inclusivo, adattandolo
alla sensibilità e alle convenzioni
dell'italiano contemporaneo (non ci
sono linee guida editoriali precise).
Lessico e registro:  Prediligere un
italiano fluido e naturale, con scelte
lessicali che rendano la voce dei
personaggi credibile e coerente con
il contesto urbano e contemporaneo.
Riferimenti culturali:  Adattare
con attenzione riferimenti culturali
e giochi di parole, privilegiando
soluzioni creative e comprensibili per
il pubblico italiano.
```

P7 (Italian):

```
Traduci il testo che ti fornisco in
italiano rispettando il linguaggio di
genere neutro o inclusivo, adattandolo
alla sensibilità e alle convenzioni
dell'italiano contemporaneo (non ci
sono linee guida editoriali precise).
Lessico e registro:  Prediligere un
italiano fluido e naturale, con scelte
lessicali che rendano la voce dei
personaggi credibile e coerente con
il contesto urbano e contemporaneo.
Riferimenti culturali Adattare con
attenzione riferimenti culturali
e giochi di parole, privilegiando
soluzioni creative e comprensibili per
il pubblico italiano.
```

P8 (Italian):

```
traduci questo estratto dal tedesco
all'italiano tratto dal libro
"Schwindel" di Hengameh Yaghoobifarah.
considera che è un romanzo di
narrativa queer contemporanea.
rispetta lo stile, mantieni la
naturalezza del testo in italiano,
e usa un linguaggio di genere
neutro/inclusivo:
```

P9 (Italian):

```
Traduci questo testo letterario
in italiano, facendo attenzione al
linguaggio inclusivo:
```

P10 (Italian):

```
puoi tradurre dal tedesco all'italiano
questo testo ricordandoti di
mantenere il registro colloquiale e
```

```
il linguaggio di genere ma facendo in
modo che quest'ultimo suoni naturale
in italiano e dunque non vengano usate
formule poco idiomatiche?
```

P11 (Italian - Detailed GFL Specifications):

```
Traduci il seguente testo dal tedesco
all'italiano adottando una strategia
traduttiva tenendo presente che
contiene casi di linguaggio neutro
e inclusivo (e.g., "kiffer_in",
"demm" come pronome neutro simile al
them in inglese, "liebhaber_innen").
Mantieni il testo facilmente leggibile
mettendo comunque enfasi sull'identità
queer di alcuni dei personaggi:  in
particolare, Ava usa il femminile,
mentre Delia usa pronomi neutri.
Proponi tre traduzioni diverse.
```

P12 (Italian):

```
Fammi la traduzione in italiano di
un estratto dal libro Schwindel,
mantenendo uno stile fluido e non
troppo formale, facendo attenzione
all'inclusività della lingua.
```

C Extended Quantitative Results

Extract B (NMT)			Extract C (LLMs)		
Seg.	**Word Class**	**German**	**Seg.**	**Word Class**	**German**
1	Subj pronoun	Dey	1	NB noun	Kiffer_in
1	NB noun	Kind	1	Subj pronoun	Dey
2	Obj pronoun	Demm	2	Obj pronoun	Demm
3	Plural noun	Mitschüler_innen	3	Obj pronoun	Demm
3	Subj pronoun	Dey	3	Subj pronoun	Dey
3	Adj/participle	Ohnmächtig	4	Obj pronoun	Demm
4	Adj/participle	Dagestanden	5	Subj pronoun	Dey
5	Subj pronoun	Dey	5	Obj pronoun	Demm
6	Subj pronoun	Dey	6	Plural noun	Liebhaber_innen
6	Adj/Participle	Fokussiert	6	Adj/participle	Nebeneinander
7	Poss pronoun	Deren	7	Obj pronoun	(required in IT)
8	Subj pronoun	Dey	8	Subj pronoun	Dey
8	Adj/participle	Hineingesprungen	9	Subj pronoun	Dey
9	Subj pronoun	Dey	10	Poss pronoun	Deren
10	Subj pronoun	Dey	10	Subj pronoun	Dey
11	NB noun	Freak			

Table 5: Comparison of segments, word classes, and German terms containing or requiring GFL markers between Extract B and Extract C.

This appendix provides a granular, segment-by-segment breakdown of the source text constraints and the corresponding outputs generated during the experiment.

To ensure a rigorous comparison between static NMT and LLM outputs, it is critical to map the exact occurrences of non-binary morphology in the source texts. Table 5 maps the specific segments containing GFL across Extract B (NMT) and Extract C (LLMs). Despite minor lexical differences, the extracts share a highly comparable morphosyntactic structure, allowing for a direct evaluation of how each system handles specific word classes (e.g. subject pronouns, object pronouns, and non-binary nouns).

Tables 6 and 7 detail the exact frequency of specific translation strategies applied by the models, broken down by word class.

Table 6 illustrates the outputs from the traditional NMT systems. The data clearly reflects the algorithmic rigidity discussed in the main text. Faced with non-binary source morphology, NMT heavily relies on syntactic evasion, most notably utilizing Italian's pro-drop feature (63 instances) to omit subject pronouns entirely. When evasion is impossible, the systems default to severe binary misgendering, split between masculine and feminine.

Conversely, Table 7 captures the results of the prompt-driven LLM generations. Here, we observe a significantly wider variance in strategic approaches. While misgendering (particularly hyper-feminization) remains prevalent, LLMs actively attempt to generate visibly inclusive texts. This includes typographical interventions (schwa, asterisk, slash), strategic neutralization (using periphrasis or synonyms), and noun substitution. However, this generative flexibility also introduces novel error types not seen in NMT, such as ungrammatical calque translations and structural mistranslations.

Word Class	**Translations & Counts**
Subject pronoun	63 pro-drop, 6 untranslated, 3 mistranslation
Object pronoun	12 misgendering-feminine
Pron_Poss	10 misrecognition, 2 no GFL required
Non-binary noun	12 misgendering-feminine, 12 misgendering-masculine
Plural noun	12 generic masculine
Adj/Verb	32 misgendering-feminine, 9 misgendering-masculine, 6 no GFL required, 1 misrecognition

Table 6: General NMT results: Frequency of translation strategies applied per word class, demonstrating high rates of evasion (pro-drop) and binary misgendering.

Word Class	Translations & Counts
Subject pronoun	36 pro-drop, 35 untranslated, 9 neutral (name instead of pronoun), 5 calque translation, 5 misgendering, 2 schwa
Object pronoun	50 misgendering-feminine, 10 untranslated, 6 schwa, 5 asterisk, 4 calque translation, 2 neutral (name instead of pronoun), 1 neutral (periphrasis), 1 neutral (element removed), 1 no translation
Non-binary noun	6 schwa, 5 neutral (periphrasis), 4 mistranslation (Kifferə), 2 misgendering-feminine, 1 asterisk, 1 schwa+asterisk, 1 schwa+underscore
Plural noun	8 misgendering-feminine, 5 schwa, 2 neutral (periphrasis), 1 slash
Adj	8 misgendering, 4 schwa, 2 asterisk, 2 neutral (synonym)

Table 7: Geenral LLM generation results: Frequency of translation strategies applied per word class, highlighting the introduction of neomorphemes (schwa, asterisk) alongside persistent hyper-feminization and calque errors.

Reasoning About Gender: How Source Text Strategies Impact Italian-to-German Machine Translation Beyond the Binary

Paolo Di Natale[1,2] and **Laura Schlutter**[3] and **Elena Chiocchetti**[2] and **Marlies Alber**[2]

[1]Free University of Bolzano/Bozen, Italy [2]Eurac Research, Italy
[3]Leibniz Universität Hannover, Germany

{pdinatale,echiocchetti,marlies.alber}@eurac.edu
laura.schlutter@icloud.com

Abstract

This paper investigates how gender-fair strategies in source texts influence the production of non-binary translations in the Italian to German combination. We use a controlled test set featuring binary and non-binary approaches to assess their effectiveness for non-binary renderings in the target language. We also introduce an automatic evaluation framework that classifies target sentences into four categories: non-binary, binary-gendered, single-gendered, and incoherent. Relying on human analysis, we compare Reasoning LLMs against standard inference, examining whether reasoning improves translation quality and automatic evaluation. Our results show that reasoning models are more successful in shifting from binary to non-binary formulations and in handling linguistic challenges such as epicene terms and special characters, although there are no improvements in sentence-level consistency and evaluation accuracy. A qualitative analysis of German translations shows that reasoning encourages the reformulation of source-side strategies through neutralization, visibility strategies, and paraphrasing, resulting in more natural target texts.

1 Introduction

Machine Translation (MT) output still suffers from heavy gender bias (Rescigno and Monti, 2023; Zahraei and Emami, 2025), more acutely in languages with grammatical gender like Italian and German (Stanovsky et al., 2019). At the same time, companies and institutions are increasingly adopting gender-fair writing practices. This is popularizing a wide range of strategies for linguistic inclusion designed to avoid the use of masculine forms to refer to people of all genders and gender identities (generic masculine). The preferred strategy is usually defined in style guides, organizational policies, or shared conventions. For these stakeholders, inconsistent or inadequate rendering of gender-fair language in the target text can hinder inclusive communication in multilingual settings and increase the post-editing effort. In the age of the widening democratization of MT to lay users (Savoldi et al., 2025b), ensuring that translations remain trustworthy and aligned with their intended purpose is of growing importance. Achieving this requires moving beyond a narrow focus on semantic equivalence (Reiss and Vermeer, 2013). Translations must also reflect the communicative purpose and stance intended for the target audience. This may fail when languages differ in how they encode gender. LLMs may stick to literal, translationese renderings (Li et al., 2025). However, this can preclude more effective or appropriate solutions in the target language.

In this respect, Italian to German translation poses a non-trivial challenge for NLP. The two languages differ in their morphological and typological systems as well as in the resources they offer for gender-fair expression (e.g., the number of gender-neutral options and the use of special characters or new morphemes as inclusive devices). In addition, gander-fair writing conventions have been disseminated at different times and to varying degrees. In the bilingual Italian region of South

Tyrol, the two languages are used together, and the translation of official documents and texts must cope with existing practices and regulations[1]. For this reason, we rely on a test set designed to assess how LLMs render the gender strategies employed in source texts. We start from a set of 12 gender-fair *strategies*, covering both binary and non-binary *approaches*. Each strategy comprises at least 20 sentences based on administrative and legal texts to evaluate LLM performance in producing fully inclusive translations and monitoring the preferred strategies used in the German target texts.

Historically, gender-fair translation has been difficult to address through model adaptation, largely due to the scarcity of high-quality, annotated training data (Saunders and Byrne, 2020). As a result, prior work has partly relied on training-free approaches (Mohapatra and Nirmala, 2025; Doyen and Todirascu, 2025). We follow this line by investigating the potential of Reasoning LLMs (RMs) for both translation and evaluation tasks. The emergence of RMs has opened the possibility of inducing reasoning on the resolution of complex tasks. We assume that reasoning may allow models to engage with potentially contentious passages to devise more appropriate realizations. By comparing two LLMs—one enabled with reasoning mode and the other not—we aim to assess whether allocating additional inference-time computation to parametric knowledge improves the generation of non-binary translations. We also evaluate which gender-fair strategies in the source text facilitate the production of non-binary translations. Relying on human annotation and analysis, we answer the following research questions:

RQ 1: To what extent do models maintain or adapt the gender-fair approach featured in the source text to a non-binary translation?

RQ 2: How does the gender strategy used in the source text influence the likelihood of producing a non-binary translation?

RQ 3: To what extent do models maintain a single gender-fair approach consistently within the target translation, even when faced with noisy gender cues in the source text?

RQ 4: Does reasoning improve the automatic evaluation of the gender-fair approach displayed in the target text?

[1]Provincial Law No. 51/2010, Art. 8

2 Background

2.1 Terminology used in this paper

We distinguish gender-fair strategies into two main approaches, *binary* and *non-binary*. *Binary* approaches explicitly express masculine and feminine forms (Diewald and Nübling, 2022), making women visible in language. In Italian, the order of the elements may vary. In German, the feminine form generally precedes the masculine. *Non-binary* approaches emphasize inclusive forms that go beyond this binary categorization and aim to represent a wider range of gender identities (Löhr, 2022; Lardelli and Gromann, 2022). Within non-binary approaches, we further distinguish visibility strategies, which make genders beyond the binary distinction visible, and neutralization strategies, which remove gender cues altogether (Morgenroth and Ryan, 2021)[2]. The former group also includes the use of *neomorphemes* (such as schwa in Italian) (Marenghi et al., 2026) and special characters (such the asterisk in German) to signal the inclusion of genders beyond the binary distinction. See Table 1 and Appendix A for examples in Italian and German. For an overview, see De Cesare and Giusti (2024), Comandini (2021), and Diewald (2022).

2.2 Grammatical gender in Italian and German

Both German and Italian have gender as a grammatical category. Italian distinguishes two genders for "cat"—masculine (*il gatto*) and feminine (*la gatta*)—whereas German has three: masculine (*der Kater*), feminine (*die Katze*), and neuter (*das Kätzchen*). In both languages, gender assignment can be motivated morphologically (e.g., suffixes, word endings), semantically (e.g., natural gender, group membership), and by phonological patterns (e.g., consonant clusters). In Italian, more than 70% of nouns follow a regular morphological pattern in which gender is determined by the final vowel (Chini, 1998). In German, by contrast, gender assignment is much more irregular.

Gender agreement Gender-differentiated elements within a sentence must agree with the gender of the noun. In both Italian and German these

[2]The cited authors refer to these concepts as *multigendering* and *degendering* in their work.

are articles, pronouns, and adjectives used attributively within the noun phrase.

Differences exist with predicative adjectives (*lui è stanco/lei è stanca* → *er/sie ist müde*, "he/she is tired"), the past participle (*lui è stato/lei è stata* → *er/sie ist gewesen*, "he/she was"), as well as the indefinite forms (*nessuno/nessuna della classe* → *niemand aus der Klasse*, "nobody from the class"), for which no feminine alternative form exists in German. Nevertheless, the indefinites (due to the grammatical inherent masculine gender (Wöllstein and Dudenredaktion, 2022; Eisenberg, 2020)), the pronominal repetition (Pittner, 1998) with masculine reference (*niemand, der...*, "nobody who"), and the morphological similarity caused by the etymological origin of the morpheme *man*[3](Pfeifer, 2005) have created controversy in gender linguistics. Some guidelines (Angsal, 2015) recommend avoiding them, while others describe them as generic or gender-neutral (Diewald and Nübling, 2022).

Unlike Italian, which has a parallel system with each singular article corresponding to its own gender-differentiated plural form (Calpestrati, 2022), German has a convergent system of articles in which gender is identified only by the singular forms (Donalies, 2018). As for person nouns, Italian and German distinguish forms by gender and number. Due to the convergent system of articles in German, substantivized participles in plural (e.g., *die Teilnehmenden*, "the participants") can be used to replace gender-marked forms, whereas Italian distinguishes two genders in the article (*i partecipanti* vs *le partecipanti*, "male" vs "female parrticipants"), so these forms are binary gendered.

Gender and Language In addition to the grammatical category *grammatical gender*, we distinguish *referential gender* that relates the linguistic to the extralinguistic reality. In fact, for human referents, the grammatical gender largely reflects the (male or female) sex of the referent. However, while grammatically masculine person nouns are often used to refer generically to referents of any gender, grammatically feminine nouns tend to be female-specific (Hellinger and Bußmann, 2015).

As language is conceived "as a tool for articulating and constructing personal identities" (Savoldi et al., 2021), the non-linguistic category of *social gender* "reflects social and cultural stereotypes" (Bußmann and Hellinger, 2003) and refers to socially imposed (male or female) gender roles. It is important to note that "gender encompasses multiple biosocial elements not to be conflated with sex [...], and [...] some individuals do not experience gender, at all, or in binary terms" (Savoldi et al., 2021). Several approaches have been developed to make the existence of non-binary individuals visible in language (see subsection 2.1).

Studies have consistently shown that masculine nouns and pronouns, despite being often used to refer to all genders, are not interpreted in a truly gender-neutral way. They systematically evoke male-biased mental representations, with women and non-binary persons requiring longer processing times to be reckoned, while inclusive language increases their mental representation (Wojahn, 2013; Lindqvist et al., 2019; van Berlekom et al., 2024; Renström et al., 2024).

Across different languages and experimental settings, evidence indicates a robust cognitive link between grammatical and social gender (Schunack and Binanzer, 2022; Lindqvist et al., 2019; Sczesny et al., 2016). Moreover, number matters, as plural forms are more likely to be interpreted as gender-neutral than singular forms (Bröder, 2024). For German legal language, Steiger-Loerbroks/von Stockhausen (2014) found that gender-unmarked role nouns lead to a stronger cognitive inclusion of women and are rated as more comprehensible and gender-fair than generic masculine forms.

2.3 Related works

Datasets for evaluating the use of gender-fair strategies in machine translation have proliferated in recent years. Natural datasets are typically derived from authentic texts for benchmarking (Lardelli and Gromann, 2023), with Gente (Piergentili et al., 2023) and mGente (Savoldi et al., 2025a) also offering the corresponding neutralized versions as contrastive examples. Other resources adopt a more controlled or semi-synthetic design. INES (Savoldi et al., 2023) contains synthetically generated sentences, while Jourdan et al. (2025) use human-authored data specifying the gender stereotypical bias.

GATE (Rarrick et al., 2024) applies target-language rewriting across binary genders, while Neo-GATE (Piergentili et al., 2024) extends

[3]In German, *man* is used as an impersonal pronoun, comparable to "you" or "one" in English

this paradigm by incorporating neomorphemes. Datasets that capture non-binary phenomena (Fririksdóttir, 2024; Piergentili et al., 2024) are particularly valuable given the relative scarcity of such forms in large-scale web data.

With respect to the suitability of neutralization strategies, Paolucci et al. (2023) highlight the importance of examining which source-side strategies are more amenable to machine processing, although their analysis centers on human translator preferences. To our knowledge, Piergentili et al. (2025) provide the only comprehensive framework for the automatic evaluation of the gender approach used in translation.

2.4 Reasoning in LLMs

LLMs have demonstrated the ability to perform diverse tasks without specific training, making prompting a common strategy for conveying general instructions (Liu et al., 2023). However, prompting reveals clear limitations in specialized or cognitively demanding scenarios. This has motivated efforts to instill stronger reasoning capabilities that enable models to handle semantically and logically ambiguous problems without relying on computationally expensive post-training procedures.

Early methods focused on eliciting reasoning through explicit instructions such as "think step by step" (elicitive prompting) (Kojima et al., 2022) or by providing structured reasoning examples via in-context learning (Dong et al., 2024). While effective to some extent, these approaches often constrained outputs to familiar formats and lengths, limiting the exploration of alternative reasoning paths that could better leverage latent knowledge. Later work introduced the use of unbounded intermediate reasoning traces, or "scratch pads", allowing models to externalize multi-step reasoning processes (Nye et al., 2021). Techniques such as iterative self-questioning further encouraged models to reassess intermediate conclusions, revealing a positive relationship between reasoning depth and output quality, as well as early signs of self-reflection and answer revision without external feedback (Madaan et al., 2023; Press et al., 2023).

Recent advances show that reasoning abilities can be optimized during training, removing the need for explicit human supervision (Guo et al., 2025). The introduction of RMs enables "native thinking" through test-time scaling (Wang, 2025; Muennighoff et al., 2025). In this paradigm, models dynamically allocate tokens to an explicit reasoning phase before producing final outputs. This increased inference-time computation allows models to defer answers and engage in deeper deliberation over contentious passages.

We argue that these properties can be relevant for gender-fair translation. Test-time scaling may enable models to allocate additional reasoning to exclude gender-marked alternatives and ensure contextually appropriate, consistent, and fluent translations that adapt to user constraints.

3 Test set

We use a subset of the LegISTyr test set[4] (Di Natale et al., 2025). LegISTyr addresses common legal translation issues from Italian into South Tyrolean German. The subset consists of 272 Italian sentences adapted from real-world instances in public administration and legal texts.

When necessary, we edit the source sentences by altering only one focus word according to the corresponding strategy (seeTable 1), leaving the remainder of the sentence unchanged. Overall, the test set comprises 140 binary and 132 non-binary focus words. This setup allows us to assess: 1) how models manage the translation of a binary/non-binary focus word into a non-binary strategy and 2) the ability of the model to extend gender-fair strategies to the remainder of the gendered source text, with a view to keeping approach consistency even in noisy scenarios. One such scenario aims to simulate realistic conditions in which masculine-marked sentences are minimally edited to facilitate the production of a gender-fair translation.

4 Experimental setup

For the comparison between reasoning and non-reasoning models, we use Deepseek-V3 (V3) and Deepseek-R1 (R1) respectively, because they are fully open with transparent training recipes. Our prompt (Appendix E) requests a non-binary gender-fair translation. Models are accessed through the OpenRouter API[5], which routes requests across multiple backend providers. The temperature is set to 0.02.

[4] LegISTyr was published under a CC BY-NC 4.0 license. The subset on gender has not been used or described in previous work.

[5] https://openrouter.ai/

Strategy	Abbreviation	Description	Example	Gloss	Instances
Binary-gendered approaches					
Full split form	Fsf	Both masculine and feminine forms are mentioned in full	il lavoratore o la lavoratrice	the worker (m) and the worker (f)	20
Full split form with an omitted part of the term	Fsf omitted	In multiword terms, the variable part of the term is indicated with the male and female forms	le studentesse e gli studenti universitari	the university students (m) and students (f)	20
Full split form with slash	Fsf slash	Masculine and feminine forms separated by slash	un collaboratore / una collaboratrice	a collaborator (m) / a collaborator (f)	20
Full split form with article or preposition before epicene term	Fsf epicene	Epicenes are nouns whose grammatical gender is realized exclusively through agreement targets (e.g., articles, adjectives, and pronouns), in this category both are explicitly expressed	una o un paziente	a (f) or a (m) patient	20
Full split form with article or preposition before epicene term with slash	Fsf epicene slash	Epicenes are nouns whose grammatical gender is realized exclusively through agreement targets (e.g., articles, adjectives, and pronouns), in this category both explicitly expressed and separated by a slash	un/una rappresentante	a (m/f) representative	20
Contracted split form	Contracted sf	Only gender morphemes are marked with a slash	alunno/a	student (m/f)	20
Full and contracted split forms of pronouns and clitics	Pronoun clitics sf	Gendered pronouns or clitics are both expressed	lei/lui; seguirli/le	she/he; follow them (m/f)	20
Non-binary approaches					
Epicene terms	INV epicenes	Epicenes are nouns whose grammatical gender is realized exclusively through agreement targets (e.g., articles, adjectives, and pronouns), in this category, no gendered referents are expressed	paziente	patient (gender-neutral)	30
Invariable terms	INV	Nouns that do not inflect for referential gender	vittima	victim (gender-neutral)	27
Collective terms	COL	Collective nouns without individual gender marking	cittadinanza	the citizenry	25
Neomorpheme schwa	SCHWA	Use of the special character schwa (ə)	lə lavoratorə	the worker (inclusive form)	30
Asterisk	AST	Use of the special character *	lavorator*	worker* (inclusive form)	20

Table 1: Description of the strategies featured in the LegISTyr test set.

For the meta-evaluation, we use a different architecture—GPT 4.1 and GPT 5.1 as non-reasoning and reasoning evaluators—to mitigate self-evaluation bias (Zheng et al., 2023). Following Piergentili et al. (2025), we provide task instructions, case-specific guidance, and examples for each strategy. We additionally require phrase-level labels for units referring to human beings before assigning a sentence-level label, although only the final sentence label is considered in our analysis. Appendix F reports the full evaluation prompt, including the enumeration of a variety of cases specific to Italian–German combination and the expansion of output labels to four: Non-binary-gendered, Binary-gendered, Single-gendered, and Incoherent (for conflicting or invalid strategies across the sentence).

4.1 Human Evaluation

The annotations are at the *focus word* level, concerning only the translation of the modified focus word that displays the gender-fair strategy, and at the *sentence* level, evaluating the gender-fair approach of the whole sentence. The human evaluator follows the same guidelines provided to the model (see Appendix F) and uses the same labels.

4.1.1 Annotation choices

We explain our decisions for contentious cases. We consider compound nouns as non-binary only if the person noun at their base is not transparently derived from an entire masculine form in contemporary German. Words that are fully lexicalized and no longer perceived as gender-marked (e.g., *Kundschaft*, "clientele") are treated as non-binary, whereas compounds with a transparently masculine headword (e.g., *Schülerschaft*, "student body", from masculine *Schüler*, "students") are treated

as single-gendered. To escape a single-gendered masculine designation, a feminine marker or a non-binary word must appear at least once in the compound: we accept *der/die Bürgerbeauftragte* ("ombudsman/woman") but not *Seniorenresidenz* ("retirement home for male seniors").

We categorize sentences that include indefinite pronouns such as *wer* and *man* as single-gendered masculine (Wöllstein and Dudenredaktion, 2022; Eisenberg, 2020). We detail the prompt with such examples.

4.1.2 Annotation of strategies in German translation

We annotate the specific type of gender-fair strategy in the German output. As it may differ from Italian equivalent strategies, we classify German strategies throughout the annotation process whenever they diverge from their Italian counterparts. The full list and description of observed German strategies can be found in Appendix B.

While most strategies match, some strategies featured in the Italian source (e.g., split forms of plural articles) cannot apply to German output, where equivalent forms may be inherently gender-neutral. For instance, plural participles and pronouns in German do not express gender and are therefore treated as neutral. To account for such cases, we introduce additional annotation categories (e.g., *generic pronoun*). Moreover, Italian epicene[6] terms cannot always be directly transposed into morphologically equivalent gender-neutral nouns in German (except for loanwords such as *Barista*). Instead, possible realizations include plural participle nouns (*Lehrende*), invariable nouns (*Lehrkraft*), or collective nouns (*Lehrpersonal*), all meaning "teaching staff".

5 Results

5.1 RQ1

Our first research question examines to what extent the models produce a non-binary translation of the focus word based on their source-side gender approach and whether reasoning has an influence.

V3 shifted from a binary to a non-binary translation in 44.3% of cases (see Table 2). R1, by contrast, achieved non-binarization of a binary input in 67.1% of cases. This suggests that reasoning helps transform binary input into non-binary output more effectively. R1 is also much more likely to keep a non-binary approach in the target text.

To characterize the output of binary-gendered input, we examine the consistency rate (Table 3), defined as the frequency with which the source-side approach is preserved in the target. Unsurprisingly, the lower non-binary rate observed for binary inputs is largely due to the higher retention of binary approaches. V3 preserves binary-gendered expressions in 40.7% of cases, whereas R1 does this much less frequently. This suggests that reasoning is more effective in complying with the non-binary translation prompt.

We also measure the failure rate, defining it as the rate of single-gendered or incoherent translations (Table 4). For both models, non-binary input is more likely to produce an entirely non-compliant translation, though reasoning mitigates this effect in both absolute and relative terms.

Source Approach	**Non-binary Rate**	
	V3	**R1**
BINARY-G.	44.3%	67.1%
NON-BINARY	57.6%	82.6%
Total	50.7%	**74.6%**

Table 2: Non-binary rate measures the frequency of non-binary translations.

Source Approach	**Consistency Rate**	
	V3	**R1**
BINARY-G.	40.7%	25.0%
NON-BINARY	57.6%	82.6%

Table 3: Consistency rate measures how often the source-side approach is preserved in the translation.

Source Approach	**Failure Rate**	
	V3	**R1**
BINARY-G.	15.0%	7.9%
NON-BINARY	30.3%	12.9%
Total	22.4%	**10.3%**

Table 4: Failure rate reflects the proportion of outputs that are either single-gendered or incoherent.

5.2 RQ2

Our second research question characterizes which Italian source strategies facilitate the process of producing a non-binary translation of the focus word in German. It further examines which strategies are more easily preserved in the translation process (Table 5). Generally speaking, agreements involving pronouns and particles (*Pronoun clitics*

[6]Epicene nouns do not encode gender morphologically. However, it can be expressed through agreement on associated elements such as articles, adjectives, and pronouns.

sf) are easier to transpose to non-binary formulations than nouns, while the use of split forms (*Fsf omitted* and *Fsf slash*) poses difficulties.

For V3, split forms of nouns tend to show lower non-binary rates than other strategies. The most evident pitfall lies in the recognition of epicene terms, whether their gender is marked in the source segment or not. To improve non-binary output, V3 benefits from source segments with focus words that are already non-binary: collective nouns (80%) and nouns with an asterisk (70%) are the most suitable solutions, whereas the schwa fails to produce non-binary translations. There is a clear gap between binary and non-binary input.

Reasoning yields benefits across all source strategies, particularly for epicene terms. As Italian epicenes often translate to gendered German words, letting models reason allows them to forgo the immediate, gendered translations in German and find more appropriate expressions. Another benefit is an improved understanding of visibility strategies, as evidenced by the highest non-binary rate observed for source segments containing asterisks (90%).

Source Strategy	**Non-binary Rate**	
	V3	**R1**
Binary approach		
Fsf	50.0%	75.0%
Fsf omitted	30.0%	70.0%
Fsf slash	30.0%	60.0%
Fsf epicene slash	45.0%	60.0%
Fsf Epicene	25.0%	60.0%
Contracted sf	45.0%	55.0%
Pronoun clitics sf	**85.0%**	**90.0%**
Non-binary approach		
INV Epicenes	36.7%	80.0%
Invariable	66.7%	77.8%
Collective	80.0%	84.0%
Schwa	43.3%	83.3%
Asterisk	70.0%	**90.0%**

Table 5: Non-binary realizations rate in the German translation of focus words, divided by input strategy and model.

Subfocus RQ 2.1: Strategy rendering We observe the annotation of the German strategies used in the target texts. The matrices of strategy renderings in Appendix C show that some source strategies are more consistently preserved in the German output than others, the tendency being stronger when reasoning is activated. This may reveal that reasoning helps not only to maintain the same approach, but also to find the closest rendering of the source strategy in the target language. In both models, collective terms (*COL*) are the most stable category, with a large proportion of Italian collective terms also being translated as collective terms in German, followed by invariable terms (INV).

When some source strategies would transfer poorly to German, reasoning has the positive effect of retaining them less often and transforming them into more natural constructions. V3 often renders split forms in Italian (*Fsf* categories) with other split forms in German, while reasoning shows the capacity to tap into a wider array of non-binary alternatives for the same source, such as German collective (*COL*) and invariable nouns (*INV*) or special characters such as asterisks. When it is the Italian source that employs a neormorpheme in the focus word, reasoning helps preserve visibility strategies in the German translation (*AST*, *COLON*), while V3 fails to do so.

Overall, reasoning seems to help find alternative phrasings that diverge from the surface structure of the source text while maintaining a non-binary approach that sounds more natural in the target language. For example, R1 renders the Italian split form *Il revisore o la revisora unica* ("male or female sole external auditor") as *Die revisionsführende Person*, a concise neutralized term, while V3 proposes a split form (*Der einzig beauftragte Rechnungsprüfer oder die einzig beauftragte Rechnungsprüferin*).

Epicenes and split forms of pronouns and clitics usually cannot be transposed with direct equivalents in German because of language differences. This causes their overall lower non-binary translation rates. Reasoning seems to provide advantages especially in these cases, when deeper elaboration leads to effective German paraphrases. For epicene categories, this happens by means of invariable nouns (*INV*) and asterisk forms (*AST*). We also highlight that V3 delivers a higher number of generic masculine forms, often interpreting invariable nouns as implicitly masculine (*la guida*, gender-neutral "guide" $\rightarrow$ *der Führer* male "guide").

Subfocus RQ2.2: Special characters We analyze how models render the neomorpheme schwa and the asterisk special character in the output.

V3 struggles to maintain asterisk forms, with an agreement rate of 35%. The asterisk was only

rarely used to translate the focus word. When the source word contains the schwa, output results are problematic, showing the second highest rate of incoherent output overall for this model (17%). Additionally, both schwa and asterisk in the source are prone to causing single-gendered translations, at 20% and 26% respectively. Looking at special characters produced by V3, asterisks are the most common (15 cases), while underscores (*USCORE*) and colons (*COLON*) are marginal. V3 also uses full split forms as translations for both visibility strategies, downgrading to binary translations.

R1, by contrast, produced non-binary visibility strategies more frequently: asterisks in the source were maintained in 70% of cases. For schwa input, R1 produced another visibility strategy (asterisk or colon) in 36% of cases. R1 never produced binary-gendered forms for asterisk or schwa input, opting instead for other non-binary strategies: invariable terms (INV) and plural participles (Pl participle). Incorrectly, single-gendered strategies also appeared, but less often than in V3: the schwa occasionally triggered single-gendered output (10%), while the asterisk never did. In terms of incoherence, asterisk input generated the highest incoherence rate within R1 with 10%, schwa input the second highest with 7%.

In sum, we suggest that reasoning helps models better understand and functionally transpose neomorphemes and special characters.

5.3 RQ3

We assess the consistency of the gender approach within the target texts. We compare the word-focus label with the sentence-level label based on the human evaluation (Table 6). We find very similar agreement rates for both models: where a non-binary approach was used for the focus word, in 85.5% (V3) and 85.7% (R1) of cases a non-binary approach was also realized in the rest of the sentence. However, in approximately 15% of sentences, the focus word was realized with a neutral approach, but at least one other element in the sentence marked gender (binary- or single-gendered). Where the source had a binary-gendered approach, 91.8% (V3) and 90.0% (R1) of sentences were consistently binary-gendered.

A z-test for differences in proportions shows that none are statistically significant. Reasoning does not seem to offer tangible improvements in the consistency with which translations maintain a gender approach across the whole sentence, but consistency rates remain fairly high.

Approach	**Consistency Rate**	
	V3	**R1**
Binary-G.	**93.2%**	90.0%
Non-binary	**87.0%**	85.7%
Total	**88.6%**	86.4%

Table 6: Consistency of gender approach at the sentence level, measured as the agreement rate between word- and sentence-level annotation labels.

5.4 RQ4

We evaluate GPT-4.1 and GPT-5.1 as automatic annotators of the gender approach label, based solely on the target. We conduct a meta-evaluation on all translations produced by DeepSeek V3 and R1 (544 instances in total). Using human annotations as the reference, we compute agreement rates between model predictions and human labels across four categories: Non-binary-gendered, Binary-gendered, Single-gendered, and Incoherent.

As shown in Table 7, both models achieve full agreement with the human annotator, indicating that reasoning does not improve overall evaluation accuracy. In addition, considering the 65 sentences misclassified by both models, only 9 differ between them. This suggests that most errors concentrate on the same subset of sentences. A breakdown of the erroneous automatic evaluations (see Appendix D) reveals that most errors involve the misclassification of the Incoherent category. This finding suggests that assessing the consistency of a gender approach across linguistic units remains particularly challenging and requires more robust safeguards. It also indicates that noisy sentences are especially prone to false positives. While both models demonstrate near-perfect recall in identifying binary-gendered and non-binary approaches, this has resulted in decreased precision.

6 Qualitative human analysis

Adaptation strategies in translation output We observed that V3 shows an ability to adapt a gender-fair approach to single-gendered nouns based on the context: the Italian phrase *tra il datore di lavoro e il lavoratore o la lavoratrice* ("between the male employer and the male or female employee") was rendered as *zwischen der*

Approach	Error Rate		Instances
	GPT 4.1	GPT 5.1	
Incoherent	91.3%	**78.2%**	24
Single-G.	29.9%	**26.1%**	107
Binary-G.	**2.5%**	5.8%	119
Non-binary	**2.7%**	3.7%	294
Total	12.0%	12.0%	544

Table 7: Error rates of the automatic evaluation of the gender approach at the sentence level.

Arbeitgeberin oder dem Arbeitgeber und der Arbeitnehmerin oder dem Arbeitnehmer ("between the male or female employer and the male or female employee"), assimilating the masculine noun to the binary-gendered pattern in the input.

Furthermore, both models produced abstract nouns referring to roles or institutions (Diewald and Steinhauer, 2022) for job titles (e.g., *I direttori d'ufficio e le direttrici d'ufficio*, "the male and female office directors" → *Die Büroleitungen*, "the office managements") and showed flexibility in generating gender-fair formulations, including original coinages such as *Bergsteiger-Person* ("mountaineering person"), or *Tierschutzorgane* ("animal welfare bodies"). We also observed syntactic reformulations: the split form of pronouns in *esse/essi possono chiedere l'intervento della Difesa civica* ("they female/they male may request the intervention of the Ombusperson") has been translated with a passive reformulation (*kann die Einschaltung der Bürgerbeauftragten beantragt werden*, "the intervention of the Ombudsperson may be requested").

Effective strategies Both models demonstrated an understanding of context-sensitive gender-fair strategies beyond the mere substitution of the focus word. V3 replaced a full split form with a shorter participial construction (*il collaboratore o la collaboratrice*, "the male or female collaborator" → *die beschäftigte Person*, "the employed person"), achieving neutralization more economically. However, we see a tendency to neutralize terms with *Person* rather than with semantically equivalent, gender-neutral forms (e.g., *die Mitarbeitenden*, "the collaborators") or collective terms (e.g., *das Personal*, "the staff").

Bias toward masculine forms We observed that collective nouns (e.g., *Bürgerschaft*, "citizenry") and abstract or institutional designations (e.g., *Meisterbrief*, "master craftsman diploma") often realized as compounds, are frequently single-gendered. Most compounds produced by both models fall into this category. For example, gender-neutral *sportelli per la cittadinanza* ("citizen service points") was translated by both models with single-gendered *Bürgerservice* ("service points for male citizens"). These also account for most of the mistakes in the automatic evaluation.

Epicene forms are particularly vulnerable to gender bias, as they do not always have established neutral equivalents in German. For instance, *insegnante* ("teacher") can be rendered as gender-neutral *Lehrkraft* ("teaching staff"), but is also frequently translated as single-gendered *Lehrer* ("male teacher"). By contrast, for *cantante* ("singer") no equally conventionalized neutral option exists, making neutral translation more difficult without special characters or paraphrases. This seems to lead the models—especially V3—to interpret epicenes as generics, resulting in (generic) masculine German forms. Beyond linguistic factors, stereotypical associations may reinforce this tendency, as roles like *conducente* ("driver"), *alpinista* ("mountaineer"), and *giudice* ("judge") are often culturally associated with men. We also observed a tendency to single-gendered translations for other terms that are subject to heavy social bias (e.g., *Sekretärin*, "female secretary") (Sant et al., 2024).

Inefficient strategies Both models produce avoidable single-gendered forms with participles: to translate split forms with slash, a singular participle is combined with a feminine article (e.g., *studente/ssa*, "male/female student" → *die Studierende*, "the female student"), although adding an article (*der/die* or *der*die*) would suffice for a gender-fair result. This also suggests that slash forms are sometimes interpreted as typos.

Both models occasionally generate inefficient split forms, including contracted or omitted variants: instead of using a gender-neutral invariable term, they produce explicit binary constructions such as *die Hausangestellte oder der Hausangestellte* ("the female housekeeper or the male housekeeper") rather than the more economical *der*die Hausangestellte*.

Incoherence V3 produced more incoherent results (12) than R1 (8), though both show similar issues. Most mistakes are due to the mixing of strategies within a nominal phrase: both mod-

els combine gendered nouns with mismatched articles (e.g., *eine Zeug:in*, "a female non-binary witness") or over-gender by adding a redundant feminine form to an already neutral construction (e.g., *des Sekretariats bzw. der Sekretärin*, "of the secretariat or the female secretary").

In R1, incoherence typically arises from combining a special character in the noun with a single-gendered article (e.g., *Die Kandidat:in*, "the non-binary candidate" instead of *Der:die Kandidat:in*), or from over-gendering forms (e.g., *der*dem Vorgesetzte*in* instead of *der*dem Vorgesetzten*, "the female*male superior"). It may also transfer Italian asterisks directly into German (*tutt** → *All**), treating them as language-independent. However, while the asterisk in Italian replaces the gendered morpheme, in German the asterisk is placed between the masculine and feminine morphemes (Murelli, 2024). This creates unusual asterisk forms.

The schwa occasionally generates incoherent cases in both models through semantic mistranslations of the focus word rather than gender-related inconsistency. The Italian word *professionistə* ("professional") is translated as *Beruf* ("profession"), probably due to truncated tokenization.

R1 demonstrates awareness that certain terms may require binary-gendered rendering when mandated by legal terminology. For example, it translated the split form *liberi professionisti o libere professioniste* ("male or female freelance professionals") with the neutral participial form *selbständig Erwerbstätige* ("self-employed persons"), while in the same sentence it retained a binary form where neutralization was not applicable, correctly rendering *l'indennità di maternità/paternità* as *Mutterschafts-/Vaterschaftsgeld* ("maternity/paternity pay").

R1 also correctly handled a contracted split form involving a vowel change producing the word *Rechtsanwalt/-anwältin* ("male/female lawyer") instead of only adding the suffix *-in*, demonstrating morphological sensitivity to stem alternations. Finally, both models handled case marking through inflection without difficulty, correctly producing genitive and dative forms in split constructions such as *von einem Mitarbeiter oder einer Mitarbeiterin* ("of a male or female collaborator").

7 Conclusion

Non-binary input strategies are easier for both models to maintain, while source strategies that explicitly encode binarity tend to preserve gender marking more strongly. We also conclude that increased reasoning capacity facilitates a shift from binary sources to non-binary target texts.

Differences in non-binary translation rates across input strategies can be attributed to structural and lexical properties of German. Elements such as pronouns and clitics, as well as collective and invariable nouns, lend themselves more easily to neutral renderings, as they tend to have neutral German equivalents. By contrast, split forms explicitly highlight gender distinctions and may therefore require more substantial reformulation, a challenge that reasoning can partly address by overcoming mere surface-level transfer.

As for the rendering of source strategies, R1 uses alternative strategies in German more frequently, reflecting a functional understanding of gender-fair language. Reasoning further enhances performance by enabling models to correctly interpret neomorphemes. The higher failure rates for the schwa may be attributed to tokenization issues that the model is unable to process. Overall, the use of neomorphemes is discouraged unless increased compute power for reasoning is available.

Vanilla inference suffers from the highest incidence of single-gendered masculine forms, more acutely for epicenes, but also for schwa forms and invariable nouns (*INV*). Reasoning helps preserve their non-binary semantics—especially when translation requires reformulation. For practical use, strategies that are already semantically or grammatically neutral (*COL*, *INV*), or explicitly marked as inclusive (the asterisk), are the best suited ones.

Contrary to what might be expected, reasoning does not confer a significant advantage in sentence-level consistency within the target translation. Neither does the approach adopted on the source side have an impact, with non-binary strategies even requiring more frequent adjustments (through agreement and co-reference) that pose a further challenge in this sense. All in all, consistency rates remain fairly high, with no need to increase compute power for this purpose.

References

Angsal, Magnus P. 2015. Die geschlechtsneutralen Indefinitpronomen en und mensch im Schwedischen und Deutschen. Eine korpusgestützte Vergleichsstudie zu Sprachkritik und Gebrauch. *Osnabrücker Beiträge zur Sprachtheorie*, 90:165–191.

Bröder, Hannah. 2024. Das sogenannte generische Maskulinum – wie geschlechtsübergreifend ist es wirklich? Bestandsaufnahme bisheriger Forschungen. In: Ewels, Andrea/Nübling, Damaris. In Ewels, Andrea and Damaris Nübling, editors, *Geschlechterbewusste Sprache. Argumente, Positionen, Perspektiven.*, pages 15–38. Georg Olms Verlag, Baden Baden.

Bußmann, Hadumod and Marlis Hellinger, 2003. *German. Engendering female visibility in German*, pages 141–174. John Benjamins Publishing Company.

Calpestrati, Nicolò. 2022. Das grammatische Genus im Deutschen und im Italienischen: Formale Kriterien, praktische Anwendungen und didaktische Implikationen aus kontrastiver Sicht. In Auteri, Laura, Natascia Barrale, Arianna di Bella, and Sabine Hoffmann, editors, *Wege der Germanistik in transkultureller Perspektive. Akten des XIV. Kongresses der Internationalen Vereinigung für Germanistik (IVG)*, number Band 10 in Jahrbuch für Internationale Germanistik, pages 643–659. Peter Lang, Bern et al., January.

Chini, Marina. 1998. Genuserwerb des Italienischen durch deutsche Lerner. In Wegener, Heide, editor, *Eine zweite Sprache lernen: empirische Untersuchungen zum Zweitspracherwerb*, number 24 in Tübinger Beiträge zur Linguistik. Serie A. - Tübingen : Narr, 1981-2008 ; ZDB-ID: 576273-X, pages 39–60. Narr, Tübingen. Series Title: Tübinger Beiträge zur Linguistik. Serie A. - Tübingen : Narr, 1981-2008 ; ZDB-ID: 576273-X.

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

De Cesare, Anna-Maria and Giuliana Giusti. 2024. Lingua inclusiva e variazione linguistica. In Cesare, Anna-Maria De and Giuliana Giusti, editors, *Lingua inclusiva: forme, funzioni, atteggiamenti e percezioni*, volume 6 of *LiVVaL Linguaggio e Variazione — Variation in Language*, pages 3–18. Edizioni Ca' Foscari - Venice University Press, Venezia. Open access.

Di Natale, Paolo, Egon W. Stemle, Elena Chiocchetti, Marlies Alber, Natascia Ralli, Isabella Stanizzi, and Elena Benini. 2025. The LegISTyr test set: Investigating off-the-shelf instruction-tuned LLMs for terminology-constrained translation in a low-resource language variety. In Gkirtzou, Katerina, Slavko Žitnik, Jorge Gracia, Dagmar Gromann, Maria Pia di Buono, Johanna Monti, and Maxim Ionov, editors, *Proceedings of the 5th Conference on Language, Data and Knowledge: TermTrends 2025*, pages 1–15, Naples, Italy, September. Unior Press.

Diewald, Gabriele and Damaris Nübling. 2022. *Genus – Sexus – Gender*. Number 95 in Linguistik – Impulse & Tendenzen. De Gruyter, Berlin/Boston. Series Title: Linguistik – Impulse & Tendenzen.

Diewald, Gabriele and Anja Steinhauer. 2022. *Handbuch geschlechtergerechte Sprache: wie Sie angemessen und verständlich gendern*. Duden. Dudenverlag, Berlin, 2., aktualisierte und erweiterte auflage edition. Series Title: Duden.

Donalies, Elke. 2018. Genus und korrespondierende einheiten. Abgerufen am 13.12.2025.

Dong, Qingxiu, Lei Li, Damai Dai, Ce Zheng, Jiaxi Ma, Rui Li, Honghao Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.

Doyen, Enzo and Amalia Todirascu. 2025. GeNRe: A French gender-neutral rewriting system using collective nouns. In Che, Wanxiang, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7889–7909, Vienna, Austria, July. Association for Computational Linguistics.

Eisenberg, Peter. 2020. *Grundriss der deutschen Grammatik: Der Satz*. Springer eBook Collection. J.B. Metzler, Stuttgart, 5., aktualisierte und überarbeitete auflage edition. Pages: 1 Series Title: Springer eBook Collection.

Fririksdóttir, Steinunn Rut. 2024. The GenderQueer test suite. In Haddow, Barry, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 327–340, Miami, Florida, USA, November. Association for Computational Linguistics.

Guo, Daya, D. Yang, H. Zhang, J. Song, P. Wang, Q. Zhu, R. Xu, R. Zhang, S. Ma, et al. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645:633–638.

Hellinger, Marlis and Hadumod Bußmann. 2015. The linguistic representation of women and men. In Hellinger, Marlis and Heiko Motschenbacher, editors, *Gender Across Languages*, volume 4, pages 1–26. John Benjamins, Amsterdam/Philadelphia.

Jourdan, Fanny, Yannick Chevalier, and Cécile Favre. 2025. Fairtranslate: an english-french dataset for gender bias evaluation in machine translation by overcoming gender binarity. In *Proceedings of the*

2025 ACM Conference on Fairness, Accountability, and Transparency, FAccT '25, page 150–166, New York, NY, USA. Association for Computing Machinery.

Kojima, Takeshi, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Lardelli, Manuel and Dagmar Gromann. 2022. Gender-Fair (Machine) Translation. pages 166–177.

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.

Li, Yafu, Ronghao Zhang, Zhilin Wang, Huajian Zhang, Leyang Cui, Yongjing Yin, Tong Xiao, and Yue Zhang. 2025. Lost in literalism: How supervised training shapes translationese in LLMs. In Che, Wanxiang, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12875–12894, Vienna, Austria, July. Association for Computational Linguistics.

Lindqvist, Anna, Emma Aurora Renström, and Marie Gustafsson Sendén. 2019. Reducing a Male Bias in Language? Establishing the Efficiency of Three Different Gender-Fair Language Strategies. *Sex roles. A Journal of research*, 81:109–117.

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

Löhr, Ronja. 2022. Ich denke, es ist sehr wichtig, dass sich so viele Menschen wie möglich repräsentiert fühlen". In Diewald, Gabriele and Damaris Nübling, editors, *Genus – Sexus – Gender*, pages 349–380. De Gruyter, Berlin, Boston.

Madaan, Aman, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Karl Moritz Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594. Curran Associates, Inc.

Marenghi, Federica, Anna Cardinaletti, and Alice Suozzi. 2026. *Inclusive Language in Italian: An Experimental Investigation of the - Suffix*. LiVVaL. Edizioni Ca' Foscari - Venice University Press, Fondazione Università Ca' Foscari, Venezia. Open access, peer reviewed.

Mohapatra, Arati and S Jaya Nirmala. 2025. Toward true neutrality: Evaluating inference-time debiasing strategies for gender coreference resolution in LLMs. In Huang, Chu-Ren, Yasunari Harada, Jong-Bok Kim, Nguyen T.M. Huyen, Le Thanh Huong, Pham Hien, Emmanuele Chersoni, Le Minh Nguyen, Rachel Edita Oñate Roxas, and Sherly Dita, editors, *Proceedings of the 39th Pacific Asia Conference on Language, Information and Computation*, pages 226–235, Hanoi, Vietnam, December. Association for Computational Linguistics.

Morgenroth, Thekla and Michelle K. Ryan. 2021. The effects of gender trouble: An integrative theoretical framework of the perpetuation and disruption of the gender/sex binary. *Perspectives on Psychological Science*, 16(6):1113–1142. Place: US Publisher: Sage Publications.

Muennighoff, Niklas, Zhiqing Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling.

Murelli, Adriano. 2024. Linguaggio sensibile al genere e sistema morfologico: un confronto tra tedesco e italiano. In Cesare, Anna-Maria De and Giuliana Giusti, editors, *Lingua inclusiva: forme, funzioni, atteggiamenti e percezioni*, volume 6 of *LiVVaL. Linguaggio e Variazione — Variation in Language*, pages 119–153. Venice University Press, Venice.

Nye, Maxwell, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. 2021. Show your work: Scratchpads for intermediate computation with language models.

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.

Pfeifer, Wolfgang. 2005. *Etymologisches Wörterbuch des Deutschen*. dtv, München.

Piergentili, Andrea, Beatrice Savoldi, Dennis Fucci, Matteo Negri, and Luisa Bentivogli. 2023. 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.

Piergentili, Andrea, Beatrice Savoldi, Matteo Negri, and Luisa Bentivogli. 2024. Enhancing gender-inclusive machine translation with neomorphemes and large language models. In Scarton, Carolina, Charlotte Prescott, Chris Bayliss, Chris Oakley, Joanna Wright, Stuart Wrigley, Xingyi Song, Edward Gow-Smith, Rachel Bawden, Víctor M Sánchez-Cartagena, Patrick Cadwell, Ekaterina Lapshinova-Koltunski, Vera Cabarrão, Konstantinos Chatzitheodorou, Mary Nurminen, Diptesh Kanojia, and Helena Moniz, editors, *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 300–314, Sheffield, UK, June. European Association for Machine Translation (EAMT).

Piergentili, Andrea, Beatrice Savoldi, Matteo Negri, and Luisa Bentivogli. 2025. An LLM-as-a-judge approach for scalable gender-neutral translation evaluation. In Hackenbuchner, Janiça, Luisa Bentivogli, Joke Daems, Chiara Manna, Beatrice Savoldi, and Eva Vanmassenhove, editors, *Proceedings of the 3rd Workshop on Gender-Inclusive Translation Technologies (GITT 2025)*, pages 46–63, Geneva, Switzerland, June. European Association for Machine Translation.

Pittner, Karin. 1998. Genus, Sexus und das Pronomen wer. pages 153–162, München. lincom europa.

Press, Ofir, Mingjie Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711, Singapore. Association for Computational Linguistics.

Rarrick, Spencer, Ranjita Naik, Sundar Poudel, and Vishal Chowdhary. 2024. GATE X-E : A challenge set for gender-fair translations from weakly-gendered languages. In Ku, Lun-Wei, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8526–8546, Bangkok, Thailand, August. Association for Computational Linguistics.

Reiss, Katharina and Hans J. Vermeer. 2013. *Towards a General Theory of Translational Action: Skopos Theory Explained*. St. Jerome Publishing, Manchester. Original work published 1984.

Renström, Emma A., Anna Lindqvist, Amanda Klysing, and Marie Gustafsson Sendén. 2024. Personal pronouns and person perception – Do paired and nonbinary pronouns evoke a normative gender bias? *British Journal of Psychology*, 115(2):253–274.

Rescigno, Argentina Anna and Johanna Monti. 2023. Gender bias in machine translation: a statistical evaluation of google translate and deepl for english, italian and german. *Proceedings of the International Conference on Human-informed Translation and Interpreting Technology 2023*.

Sant, Aleix, Carlos Escolano, Audrey Mash, Francesca De Luca Fornaciari, and Maite Melero. 2024. The power of prompts: Evaluating and mitigating gender bias in MT with LLMs. In Faleńska, Agnieszka, Christine Basta, Marta Costa-jussà, Seraphina Goldfarb-Tarrant, and Debora Nozza, editors, *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 94–139, Bangkok, Thailand, August. Association for Computational Linguistics.

Saunders, Danielle and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. In Jurafsky, Dan, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7724–7736, Online, July. Association for Computational Linguistics.

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

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

Savoldi, Beatrice, Giuseppe Attanasio, Eleonora Cupin, Eleni Gkovedarou, Janiça Hackenbuchner, Anne Lauscher, Matteo Negri, Andrea Piergentili, Manjinder Thind, and Luisa Bentivogli. 2025a. Mind the inclusivity gap: Multilingual gender-neutral translation evaluation with mGeNTE. In Christodoulopoulos, Christos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 13698–13720, Suzhou, China, November. Association for Computational Linguistics.

Savoldi, Beatrice, Alan Ramponi, Matteo Negri, and Luisa Bentivogli. 2025b. Translation in the hands of many: Centering lay users in machine translation interactions. In Christodoulopoulos, Christos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 13876–13889, Suzhou, China, November. Association for Computational Linguistics.

Schunack, Silke and Anja Binanzer. 2022. Revisiting gender-fair language and stereotypes – A comparison of word pairs, capital I forms and the asterisk. *Zeitschrift für Sprachwissenschaft*, 41(2):309–337, November. Publisher: De Gruyter.

Sczesny, Sabine, Magda Formanowicz, and Franziska Moser. 2016. Can Gender-Fair Language Reduce Gender Stereotyping and Discrimination? *Frontiers in Psychology*, 7, February. Publisher: Frontiers.

Stanovsky, Gabriel, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In Korhonen, Anna, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy, July. Association for Computational Linguistics.

Steiger-Loerbroks, Vera and Lisa von Stockhausen. 2014. Mental representations of gender-fair nouns in German legal language: An eye-movement and questionnaire-based study. *Linguistische Berichte*, 2014, January.

van Berlekom, Elli, Sabine Sczesny, and Marie Gustafsson Sendén. 2024. Toward Visibility: Using the Swedish Gender-Inclusive Pronoun Hen Increases Gender Categorization of Androgynous Faces as Nonbinary. *Journal of Language and Social Psychology*, 43(5-6):525–543, October. Publisher: SAGE Publications Inc.

Wang, Jason. 2025. A tutorial on llm reasoning: Relevant methods behind chatgpt o1.

Wojahn, Daniel. 2013. De personliga pronomenens makt : En studie av hur pronomen styr våra föreställningar om personer. pages 356–367. Institutionen för språk, litteratur och interkultur, Karlstads universitet.

Wöllstein, Angelika and Dudenredaktion. 2022. *Duden - Die Grammatik Struktur und Verwendung der deutschen Sprache. Sätze - Wortgruppen - Wörter*. Dudenverlag, Berlin, 10. völlig neu verfasste auflage edition.

Zahraei, Pardis Sadat and Ali Emami. 2025. Translate with care: Addressing gender bias, neutrality, and reasoning in large language model translations. In Che, Wanxiang, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 476–501, Vienna, Austria, July. Association for Computational Linguistics.

Zheng, Lianmin, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

A Common German gender-fair strategies

Strategy	Examples
Binary approaches	
Visibility strategies	
Split forms / double naming	*Lehrerinnen und Lehrer*; *Fachmann/Fachfrau*
Contracted split and bracket forms	*Lehrer/-in*; *Lehrer[in]*
Internal capitalization	*LehrerInnen*
Non-binary approach	
Visibility strategies	
Forms with special characters	*Lehrer_innen*; *Lehrer*innen*; *Lehrer:innen*
Neutralization strategies	
Nominalized participles or adjectives	*Lehrende*; *Abgeordnete*; *Jugendliche*
Neutral nouns	*Fachkraft*; *Lehrkraft*; *Lehrperson*
Collective nouns	*Lehrpersonal*; *Fachleute*; *Team*
Metonymy	*das Ministerium*; *die Verwaltung*
Depersonalization	*Der Aufsatz ist zu lesen*

Table 8: A selection of common German gender-fair strategies

B Annotation of German strategies output by the models

Strategies	Abbreviation	Example
Single-gendered		
Male form		"Tätigkeit als *Handelsvertreter*"
Female form		"Cianos Beziehung zu einer *Spionin*"
Female form or plural (ambiguous)		"die Einschaltung *der Bürgerbeauftragten*"
Binary-gendered		
Visibility strategies		
Full split form	Fsf	"*Ärztinnen und Ärzte*"
Full split form with an omitted part	Fsf omitted	"*Der Inhaber oder die Inhaberin des Bettplatzes*"
Contracted split form	Contracted sf	"Als *Fremdenführer/-in* gilt, (. . .)"
Split form of pronouns and clitics	Pronoun clitics sf	"*Er/Sie* ist zuständig"
Full split form of invariable terms	Fsf of INV	"*die Hausangestellte oder der Hausangestellte*"
Non-binary-gendered		
Visibility strategies		
Incl. character asterisk	AST	"*Pädagog*innen*"
Incl. character colon	COLON	"Funktionen als *Gemeindesekretär:in*" (. . .)
Incl. character underscore	USCORE	"(. . .) wendet sich an *jede_n Studierende_n*"
Neutralization strategies		
Epicene terms	Epicene	"*Barista* ist, wer (. . .)"
Invariable terms	INV	"im Einvernehmen mit der *Bezugsperson* (. . .)"
Collective terms	COL	"(. . .) zwischen dem *Lehrpersonal*"
Plural participle	Pl participle	"*Die Studierenden* garantieren (. . .)"
Generic pronoun	Generic pronoun	"(. . .) 27 Millionen Menschen folgen *ihnen*"
Others		
Various strategies of reformulation (selection)		
Passive construction		"(. . .) kann beantragt werden."
Omitted pronoun		"(. . .) das mit den zugewiesenen Aufgaben verbundene Inventar"
Incoherent		
Inconsistent split form		"(. . .) sowie einer Führungskraft oder einem Führungskraft"
Mixed strategies		"Falls der/die Mieter:in (. . .)"

C German strategies used in the target translations

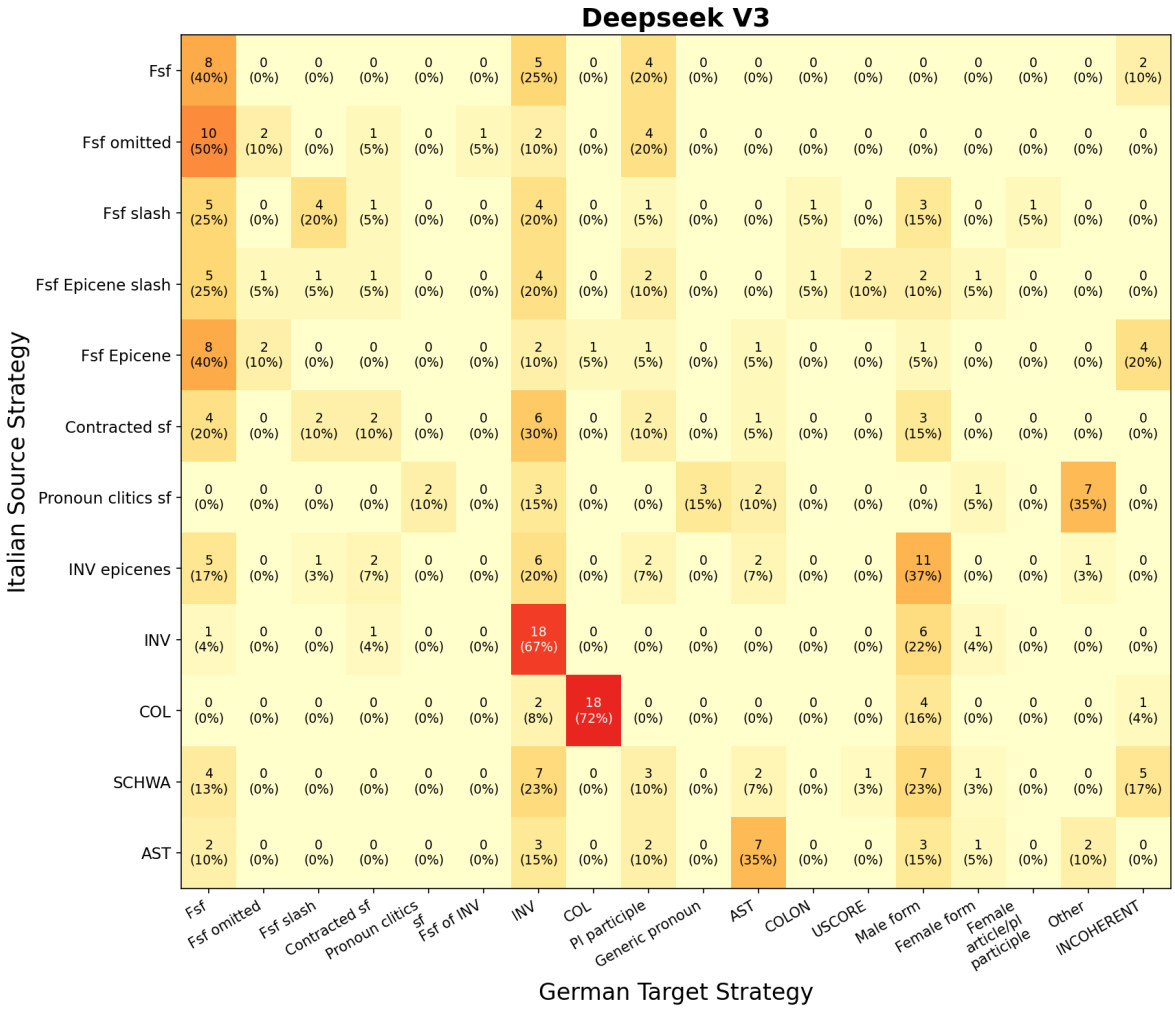

Figure 1: Follows in the next page

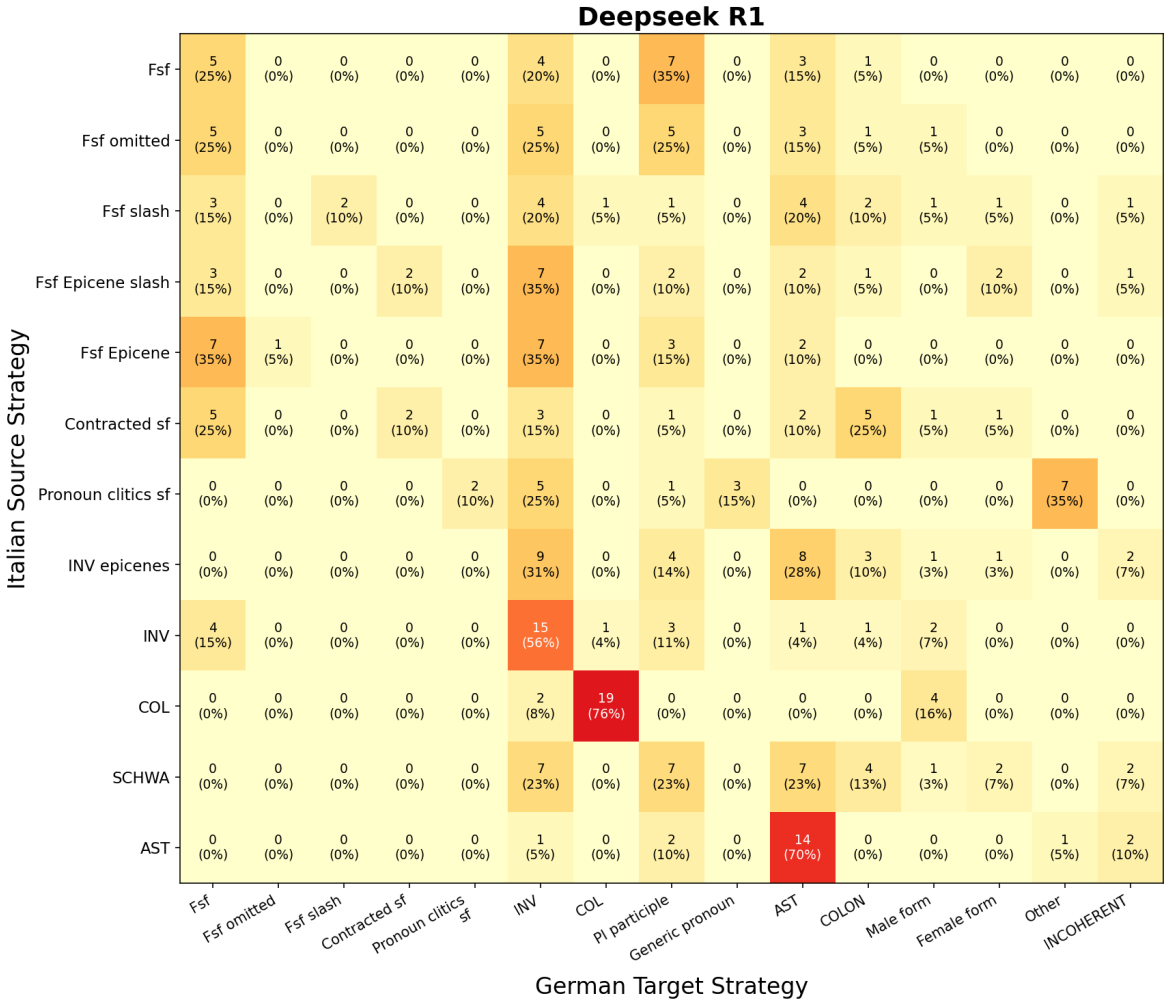

Figure 2: Annotations of the frequency of German strategies adopted to translate each Italian source strategy.

D Confusion matrix for the automatic evaluation framework

		GPT 4.1			
		INCOHERENT	**SINGLE-G**	**BINARY-G**	**NEUTRAL**
Human	**INCOHERENT**	0	1	6	14
	SINGLE-G	0	0	3	29
	BINARY-G	1	2	0	0
	NEUTRAL	0	8	0	0

Table 9: Confusion matrix for the automatic evaluations of GPT 4.1

		GPT 5.1			
		INCOHERENT	**SINGLE-G**	**BINARY-G**	**NEUTRAL**
Human	**INCOHERENT**	0	3	3	12
	SINGLE-G	1	0	2	25
	BINARY-G	1	3	0	2
	NEUTRAL	1	8	2	0

Table 10: Confusion matrix for the automatic evaluations of GPT 5.1

E Translation prompt

```
{
    ``role: ``user,
    ``content: (
        fYou are tasked to translate a legal sentence from Italian into South-
    Tyrolean German. ``
        fSouth-Tyrolean German is a standard variety of German. ``
        fImportant: You must produce a non-binary translation, eliminating all
    references to gender! ``
        fYou must output only the translated text without any explanation, enclosing
     it in '<>' symbols. ``
        fThis is the text to be translated into German: ``
        f<{source_sentence}>
    )
}
```

F Evaluation prompt

```
SYSTEM_PROMPT = ``You are a language expert specializing in evaluating gender
    neutrality in German texts. Your task is to extract target German phrases that
    refer to human beings and determine whether each phrase is single-gendered (
    masculine or feminine only), binary-gendered (covers both genders), or non-
    binary (covers neutral as well as non-binary references). Based on the phrases,
    assess whether the sentence is single-gendered, binary-gendered, non-binary, or
    incoherent.

Guidelines:
1. Identify relevant phrases: carefully analyze the German sentence and focus on all
     phrases that refer to human beings or groups of human beings (e.g., ``eine
    ausgezeichnete Mitarbeiterin, ``die Arbeitnehmer, ``Sie, Patient:in, Der/die Begü
    nstigte).

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'; 'M&F
    '; 'N'). These are the guidelines:

Single-gendered (one gender 'M' or 'F'):
* Phrases like ``Ein Redner, ``Der Student, ``Der Bürger, and ``alle Kollegen, ``die
    Arbeiter are masculine [M];
* Phrases like ``Eine Rednerin, ``Die Studentin, ``Die Bürgerinnen, and ``alle
    Kolleginnen are feminine [F]

Binary-gendered (both genders 'M&F'):
* Phrases like ``Bürgerinnen und Bürger, der/die Arbeitnehmer/in, die BürgerInnen,
    die Bürgerbeauftragte/der Bürgerbeauftragte, Absolventinnen/Absolventen,
    Absolventen/-innen, Absolvent(inn)en, der Bürgermeister oder die Bürgermeisterin
    , die Schülerin bzw. der Schüler, ``Mitarbeiter/in call both social genders,
    masculine and feminine [M&F]
* Phrases, that just refer to one gender in plural (``die Schüler, ``die Chefinnen)
    are never binary-gendered.

Non-binary (all genders, masculine, feminine and non-binary 'N'):
* Phrases like ``Der:die Arbeitnehmer:in, die*der Verwaltungs- oder
    Buchhaltungsfunktionär*in, ``Lehrer_in, Arbeitnehmx use gender-inclusive
    characters or neomorphems to neutralize gender by changing the morphology [N].
* Phrases like ``Eine freiberufliche tätige Person, ``Die Beschäftigten, ``Die Bü
    rgerschaft, and ``alle Kollegiumsmitglieder, ``die Datenschutzbeauftragten do not
     express social gender because they are neutral or express all gender identities
     beyond the binary distinction, therefore they must be considered non-binary [N
    ].

Incoherent ('INCOHERENT'):
* Phrases like ``eine*n Journalist/in, ``einen externen technischen Fachkraft apply
    conflicting gender markers or syntactical errors that make it impossible to
    determine a coherent strategy,  because in one phrase the reference is neutral (
    article: ``eine*n), while in the other another strategy is used (noun: Tutor/in)
    . It is never incoherent, if the same gender-inclusive character is used in
```

```
    article, adjective and noun (or any gender-sensitive element), e.g ``eine:n
    Journalist:in is non-binary and not incoherent. The phrase label is: ['
    INCOHERENT']


3. Assign a sentence-level label:
* If one or more phrases convey one specific gender, either masculine or feminine,
    label the sentence as ``SINGLE-GENDERED.
* If all phrases convey both masculine and feminine gender, label the sentence as ``
    BINARY-GENDERED.
* If all references to human beings are non-binary, label the sentence as ``NON-
    BINARY.
* If a phrase that is a reference to human being is incoherently gendered, label the
     sentence as ``INCOHERENT
* If at least one reference is less inclusive than the others, always choose the
    lower label.


FEW_SHOT_EXAMPLE =
# Example 1:
German sentence: ``Die Ausbildung für Bergführer dauert drei Jahre und die
    technischen Fähigkeiten entsprechen den internationalen Standards
Expected output:
    {
  ``phrases: [
    {
      ``phrase: ``Bergführer,
      ``gender: M
    }
  ],
  ``label: ``SINGLE-GENDERED
}

# Example 2:
German sentence: ``Es wurden Stipendien für akademische Leistungen an Absolventinnen
    /Absolventen, die an internationalen Universitäten, Schulen oder technischen
    Ausbildungszentren teilnehmen, vergeben
Expected output:
    {
  ``phrases: [
    {
      ``phrase: ``Absolventinnen/Absolventen,
      ``gender: ``M&F
    },
    {
      ``phrase: die,
      ``gender: ``N
    }
  ],
  ``label: ``BINARY-GENDERED
}

# Example 3:
German sentence: ``Es wurden Stipendien für akademische Leistungen an AbsolventInnen
    , die an internationalen Universitäten, Schulen oder technischen
    Ausbildungszentren teilnehmen, vergeben
Expected output:
    {
  ``phrases: [
    {
      ``phrase: ``AbsolventInnen,
      ``gender: ``M&F
    },
    {
      ``phrase: die,
      ``gender: ``N
    }
  ],
  ``label: ``BINARY-GENDERED
}
```

```
# Example 4:
German sentence: ``Wir suchen eine:n Kinderbetreuer:in in Vollzeit..
Expected output:
    {
  ``phrases: [
    {
      ``phrase: ``eine:n Kinderbetreuer:in ``,
      ``gender: N
    }
  ],
  ``label: ``NON-BINARY
}



# Example 5:
German sentence: ``Die Presseagentur sucht eine*n Journalist*in.
Expected output:
    {
  ``phrases: [
    {
      ``phrase: ``eine*n Journalist*in,
      ``gender: N
    }
  ],
  ``label: ``NON-BINARY
}


# Example 6:
German sentence: ``Ab dem Schuljahr 2018/2019 werden die Überstunden für das
    provinziell beschäftigte Lehr- und Verwaltungspersonal im Bereich der Schulen
    elektronisch verwaltet.
Expected output:
    {
  ``phrases: [
    {
      ``phrase: ``Lehr- und Verwaltungspersonal,
      ``gender: N
    }
  ],
  ``label: ``NON-BINARY
}

Example 7:
German sentence: ``Die Presseagentur sucht eine_n Journalist_in.
Expected output:
    {
  ``phrases: [
    {
      ``phrase: ``eine_n Journalist_in,
      ``gender: N
    }
  ],
  ``label: ``NON-BINARY
}


#D) Mixed

Example 8:
German sentence: ``Der Beauftragte oder die Beauftragte der Beschäftigten für die
    Sicherheit erhält auf deren Wunsch und zur Ausübung ihrer Funktion eine Kopie
    des Risiko-Bewertungsdokuments in Papier- oder Digitalform.
Expected output:
    {
  ``phrases: [
    {
      ``phrase: ``Der Beauftragte oder die Beauftragte,
```

```
      ``gender: M&F
    },
    {
      ``phrase: ``der Beschäftigten,
      ``gender: N
    },
    {
      ``phrase: ``deren,
      ``gender: N
    },
    {
      ``phrase: ihrer Funktion,
      ``gender: N
    }
  ],
  ``label: ``BINARY-GENDERED
}

# Example 9:
German sentence: ``Man arbeitet mit Patient:innen jeden Alters, deren Bewegungsfä
    higkeit und Funktion durch Traumata oder Erkrankungen beeinträchtigt wurden.
Expected output:
    {
  ``phrases: [
    {
      ``phrase: ``Man,
      ``gender: M
    },
    {
      ``phrase: Patient:innen,
      ``gender: ``N
    },
    {
      ``phrase: deren,
      ``gender: ``N
    }
  ],
  ``label: ``SINGLE-GENDERED
}


#E) incoherent

# Example 10:
German sentence: ``Das Amt bestellt eine*n Tutor/in, die/der die Ansprechperson für
    diejenigen ist, die das Praktikum absolvieren.
Expected output:
    {
  ``phrases: [
    {
      ``phrase: ``eine*n Tutor/in,
      ``gender: INCOHERENT
    },
    {
      ``phrase:  die/der,
      ``gender: M&F
    },
    {
      ``phrase: die Ansprechperson,
      ``gender: N
    },
    {
      ``phrase: diejenigen,
      ``gender: N
    },
    {
      ``phrase: die,
      ``gender: N
    }
  ],
```

```
  ``label: INCOHERENT
}


#F) Incoherent


Example 11:
German sentence: ``eine externe technische Fachkraft oder einen externen technischen
     Fachkraft.
Expected output:
    {
  ``phrases: [
    {
      ``phrase: ``eine externe technische Fachkraft,
      ``gender: N
    },
    {
      ``phrase: einen externen technischen Fachkraft,
      ``gender: INCOHERENT
    }
  ],
  ``label: INCOHERENT
}
```

Can Emotions Signal Gender? Investigating Implicit Cues in Human and LLM Translations of Amazon Reviews

Shushen Manakhimova[1,2] **Ekaterina Lapshinova-Koltunski**[2]
[1]DFKI, Berlin, Germany
[2]University of Hildesheim, Germany
manakhimova@uni-hildesheim.de, lapshinovakoltun@uni-hildesheim.de

Abstract

This study investigates whether source-text emotion is associated with grammatical gender choices in English-Russian translation when the author's gender is not specified. Building on Popović and Lapshinova-Koltunski (2024), we analyse first-person grammatical gender in Amazon review translations produced by professional translators, translation students, and three LLMs: GPT-4, Llama, and Mistral. We assign emotion labels to the English source reviews and test whether these labels are associated with masculine or feminine forms in the Russian translations. Reviews labelled as expressing *love* are more likely to be translated with feminine grammatical gender by both human translator groups, with small-to-moderate effect sizes. This pattern is absent in GPT-4 and Mistral, but appears in Llama. Other frequent emotion labels do not show the same positive association. We therefore treat the findings as exploratory evidence that specific source-text emotions may be associated with gendered translation choices, while emphasizing the need for larger-scale validation across additional emotions, genres, and language pairs.

1 Introduction

This paper presents a study on gender and emotion in Amazon product review translations. We focus on first-person grammatical gender, following Popović and Lapshinova-Koltunski (2024), who showed that product category functions as an implicit gender cue in English-Russian and English-Croatian review translations, with masculine forms dominating across all translator types. Unlike English, Russian marks gender on past-tense verbs, adjectives, and passive participles, requiring translators to specify the reviewer's gender even when the source text provides no such information. What other contextual factors might drive these gendered choices remains an open question (Hackenbuchner et al., 2024).

Psychology research has documented systematic associations between specific emotions and gender stereotypes. For example, people seem to believe women experience and express the majority of emotions more than men, including sadness, fear, and love with anger and pride as the main exceptions (Rosenkrantz et al., 1968). These patterns hold across cultures (Brody and Hall, 2008). Large language models (LLMs) seem to replicate this stereotype pattern by consistently attributing anger to male personas and sadness to female ones when prompted with gendered scenarios (Plaza-del-Arco et al., 2024). Our central goal is to examine whether source-text emotion is associated with grammatical gender choices in translation when gender is otherwise under-determined.

To achieve this goal, we use emotion as an explanatory variable for gender bias, applying BERT-based emotion classifiers based on the GoEmotions taxonomy (Demszky et al., 2020) to the English source texts. We then analyse first-person gender in professional and student human translators, as well as in the three LLM translation outputs: GPT-4, Llama and Mistral to ensure a broader comparison across model families. Following previous work on this corpus (Popovic

and Lapshinova-Koltunski, 2024), our aim is not to mitigate gender bias directly, but to examine whether systematic patterns in translation data can be linked to gendered emotion stereotypes present in the source texts.

For our analysis, we formulate the following research questions (RQs):

RQ1: Are classifier-assigned emotion labels in the English source texts associated with grammatical gender choices in Russian translations, and are these associations specific to particular emotions?

RQ2: Do these patterns differ across human translator groups and LLMs?

Our study is correlational: we do not claim to identify the causes of gender bias, but rather to uncover patterns in translation data. The analysis focuses on binary gender (masculine/feminine) as required by Russian grammar, and we acknowledge this as a limitation.

The remainder of the paper is organised as follows: Section 2 reviews related work. Section 3 describes the corpus and annotation. Section 4 outlines the methodology. Results are presented in Section 5, followed by discussion and conclusion in Sections 6 and 7. We also discuss the limitations in Section 8.

2 Related Work

A widely reported pattern in MT is the masculine default: absent other information, masculine forms are used disproportionately across all translator types (Savoldi et al., 2021; Stanovsky et al., 2019). Savoldi et al. (2021) provide the foundational framework for the field, defining gender bias in MT as systematic and unfair discrimination that manifests principally as underrepresentation of feminine forms and stereotyped translation choices. Recent work also situates gender bias in MT within the broader shift toward LLMs (Vanmassenhove, 2024).

Our study builds directly on Popović and Lapshinova-Koltunski (2024), who examine first-person grammatical gender in human and machine translations of English Amazon product reviews into Russian and Croatian, using the DiHuTra corpus. Their analysis confirms a strong masculine default across all translator types, but also shows that the masculine default is not absolute: product category functions as an implicit cue, with beauty reviews attracting substantially more feminine forms and sports reviews rendered almost exclusively in the masculine. Human translators show considerably more sensitivity to these cues than MT systems, which produce near-uniformly masculine output even in categories where feminine forms might be expected. This finding motivates the present study: if the product category can function as an implicit gender cue, what other features of the source text might do the same?

This broader role of context is also supported by work on gender decisions under ambiguity. In an English-German study, gender assignments by DeepL were compared with those of 22 human annotators for the role names presented both in isolation and in sentence context (Hackenbuchner et al., 2024). Adding context changed 44% of the human gender labels, compared to only 17% of the MT labels. This suggests that humans are more context-sensitive than MT systems. For MT, gender shifts mostly occurred when the context contained explicit cues, such as pronouns, female proper names, or an unambiguous referent.

The theoretical motivation for examining emotion as a gender cue rests on associations between specific emotions and gender stereotypes in psychology. Since at least the late 1960s, cultural gender stereotypes have reliably assigned emotional traits along gender lines (Rosenkrantz et al., 1968). A partial replication three decades later confirms that cross-gender agreement on these trait attributions remains high, though the number of items reaching stereotype consensus also decline (Nesbitt and Penn, 2000). Plant et al. (2000) show that people believe women experience and express the majority of emotions more intensely than men, with anger and pride as the main exceptions. Brody and Hall (2008) extend this picture across cultures: women are stereotypically linked to love, sadness, fear, and guilt; men to anger and pride. These associations are sufficiently robust and widespread that they can reasonably be expected to form part of the background knowledge of any translator working between the two languages in our study.

Recent NLP work suggests that these associations are also reflected in LLM behaviour. Plaza-del-Arco et al. (2024) prompted five state-of-the-art LLMs with identical scenarios while varying only the gender of the persona; all models consistently attributed anger to male personas and sad-

ness to female ones, closely mirroring the psychological literature. Related evidence comes from Weber et al. (2026), who examine whether demographic information about the author of an emotionally ambiguous text influences emotion interpretation. They find that human annotators are more likely to match the author-provided emotion label when that label aligns with gender stereotypes; their LLM experiments point in the same direction. Together, these studies suggest that associations between gender and emotion are not only documented in psychological research, but also appear in recent NLP settings involving both human judgments and LLM behaviour.

To our knowledge, the connection between emotion and gender has not yet been systematically explored in translation. This study addresses that gap by examining whether source-text emotion is associated with grammatical gender choices when the author's gender is not specified in the source text. In doing so, we connect work on gender bias in translation with research on gendered emotion stereotypes in psychology and NLP. By also considering product category, we assess whether emotion provides explanatory value beyond the domain-level associations already identified in prior work.

3 Data

Our study uses the DiHuTra corpus (Lapshinova-Koltunski et al., 2022), a publicly available parallel corpus of 196 English Amazon product reviews and their translations into multiple languages, collected in 2022[1]. The corpus is structured as 14 reviews in each of 14 product categories: beauty, books, CD/vinyl, cell phones, grocery and gourmet food, health care, home and kitchen, movies and TV, musical instruments, patio and garden, pet supplies, sports and outdoors, toys and games, and video games. The corpus contains reviews across a range of star ratings (1-2 = negative, 4-5 = positive); 3-star reviews were excluded to obtain a clear positive/negative split (Popovic and Lapshinova-Koltunski, 2024).

The human translations were produced by two groups: professional translators and translation students. Translators were provided with the review texts only — no star ratings or other metadata were shared, ensuring that any gender choices reflect responses to the text itself. They were also given minimal instructions and were not allowed to use any MT systems. Crucially, no instructions were given regarding how to handle grammatical gender in the target language, making the corpus suitable for studying gender choices as individual decisions. The present study extends the data to LLM-generated translations: GPT-4[2], Llama[3], and Mistral[4], each prompted only with "translate into Russian".

Gender annotation of all Russian translations was performed manually at both the segment and review levels, following Popović and Lapshinova-Koltunski (2024). The annotation identifies first-person gendered words in the Russian text, primarily past participles, adjectives, and passive participles.

Following their annotation scheme, each review is assigned a gender label according to the first-person gendered words it contains. If all first-person gendered words within a review share the same gender, the review is labelled m (masculine) or f (feminine) accordingly. If the review contains no first-person gendered words at all, it is labelled x. If the review contains a mixture of first-person genders, it is labelled mixed (covering sub-types such as m+f, m—f, m+m—f, and f+m—f).

Group	f	m	x	mixed
GPT-4	21	113	61	1
Llama	24	101	63	8
Mistral	23	113	59	1
Prof	36	88	71	1
Stud	27	91	78	0

Table 1: Distribution of gender labels across translator groups (n = 196). f = all first-person gendered words are feminine; m = all first-person gendered words are masculine; x = no first-person gendered words present; mixed = review contains both masculine and feminine markings

We restrict our quantitative analysis to reviews unambiguously labeled as masculine (**m**) or feminine (**f**). All mixed or ambiguous cases are excluded. Reviews where the Russian text requires no first-person gender marking are similarly excluded. This results in between 118 and 136 reviews per translator group available for statistical analysis, depending on the proportion of gendered constructions in each translation variant.

[1] https://github.com/katjakaterina/dihutra

[2] accessed in December 2024 through https://chat-ai.academiccloud.de/chat

[3] accessed in January 2025 through https://chat-ai.academiccloud.de/chat

[4] accessed in January 2025 through https://chat-ai.academiccloud.de/chat

4 Methodology

4.1 Emotions under analysis

Prior work on gender and emotion stereotypes links emotions such as *love*, *sadness*, *fear*, *guilt*, and *caring* with femininity, while *anger* and *pride* are more often associated with masculinity (Brody and Hall, 2008; Plant et al., 2000; Plaza-del-Arco et al., 2024). We use this literature as a theoretical motivation for examining whether emotional content may function as an implicit cue for grammatical gender choice in translation.

The stereotypically masculine emotions identified in the literature, such as *anger* and *pride*, do not appear among the top-1 or top-2 labels assigned by the GoEmotions classifier in our corpus and are therefore excluded from the analysis.

4.2 Emotion annotation

For emotion classification of the English source texts, we use `SamLowe/roberta-base-go_emotions`[5], a RoBERTa-base transformer fine-tuned on the GoEmotions dataset (Demszky et al., 2020). The model predicts 27 fine-grained emotion categories plus *neutral*, grouped into positive, negative, ambiguous, and neutral classes. We selected this model because its taxonomy is compatible with our research question and because it is publicly available and can be run locally, supporting reproducibility. The model outputs a probability distribution over all labels for each review.

In the main analysis, we retain the two highest-probability labels per review. This decision reflects the multi-label design of GoEmotions, where a text may express more than one emotion, and allows us to capture cases in which an emotion is a plausible secondary label rather than the single highest-ranked prediction. This is important for the present study because several reviews show small probability margins between the top-ranked labels.

Across the corpus, 22 of the 28 labels appear as top-1 predictions. The most frequent are *admiration* (30.6%), *neutral* (14.8%), *disappointment* (13.3%), *disapproval* (9.2%), *approval* (6.6%), and *love* (5.1%, $n = 10$). Among the theoretically relevant feminine-coded emotions discussed above, only *love* occurs often enough for quantitative analysis. It appears as the top-1 label in 10 reviews and as either the top-1 or top-2 label in 27 reviews.

We additionally inspect the star-rating distribution of the 27 love-labelled reviews. Most are positive: 18 are 5-star reviews and 4 are 4-star reviews. The remaining 5 have low ratings, but manual inspection shows that they still contain affectionate language directed at specific product features despite overall dissatisfaction. For example, one 1-star review states: *"I really wish this was not the case because I love the design, just be aware before you buy!"* This suggests that the *love* label captures genuine positive affect in the source text rather than merely reflecting overall review polarity.

As a plausibility check, we also re-annotate the reviews with GPT-4o (zero-shot, temperature 0), following recommendations that LLM-based annotation should be validated and interpreted on a task-specific basis (Pangakis et al., 2023). To reduce label ambiguity in this supplementary check, we prompt GPT-4o with a reduced 9-label scheme covering the most frequent emotion categories in our data: *admiration*, *approval*, *disapproval*, *disappointment*, *annoyance*, *joy*, *love*, *neutral*, and *other*. Since the classifier and GPT-4o use different label spaces, the resulting figures are not treated as inter-annotator agreement, but only as a secondary plausibility check.

Exact top-1 agreement between RoBERTa and GPT-4o is 32.0% (62/196), while soft agreement, where the RoBERTa top-1 label appears within GPT-4o's top-2 predictions, reaches 47.9% (93/196). Agreement at the broader polarity level is higher, at 63.4% (123/196). This suggests that the two systems diverge on fine-grained emotion labels but show greater convergence on the general affective direction of the reviews. The statistical analysis is based on the `SamLowe/roberta-base-go_emotions` classifier output.

4.3 Analysis

To examine the association between source-text emotion and grammatical gender in translation, we cross-tabulate the binary emotion flag (*love* present vs. absent) with the binary gender label (**m** vs. **f**) for each translator group separately. The emotion flag is set to positive if *love* appears as either the top-1 or top-2 label assigned by the classifier, yielding 27 reviews with a *love* signal out of 196

[5] `https://huggingface.co/SamLowe/roberta-base-go_emotions`

	N	+love		–love		OR	p	V
		f	m	f	m			
GPT-4	134	5	15	16	98	2.04	.313	.079
Llama	125	8	9	16	92	5.11	.005**	.251
Mistral	136	6	15	17	98	2.31	.201	.106
Prof.	124	9	7	27	81	3.86	.017*	.204
Stud.	118	9	7	18	84	6.00	.002**	.285

Table 2: Association between *love* and feminine grammatical gender. N = retained masculine/feminine cases; OR = odds ratio; p = two-tailed Fisher's Exact Test; V = Cramér's V. * $p < .05$; ** $p < .01$.

total.

For each 2×2 contingency table, we apply Fisher's Exact Test rather than Pearson's chi-square, because at least one expected cell count in each love/gender table falls below 5, the standard threshold for applying chi-square tests (Agresti, 1990). Fisher's Exact Test is therefore the appropriate alternative in this situation, providing exact p-values without relying on large sample approximations.

To complement significance testing with a measure of association strength, we additionally compute Cramér's V for each table. Cramér's V ranges from 0 (no association) to 1 (perfect association). By convention, values around 0.1 indicate a weak effect, 0.3 a moderate effect, and 0.5 or above a strong effect (Cohen, 1988).

We extend the same procedure to all eight emotions with sufficient combined frequency across top-1 and top-2 labels ($n \geq 10$): *admiration*, *approval*, *annoyance*, *disappointment*, *disapproval*, *joy*, *love*, and *neutral*. For each emotion, the appropriate test (Fisher's Exact or chi-square) is selected automatically based on the minimum expected cell count. This comparative analysis is used to determine whether any observed association is specific to *love* or reflects a broader relationship between emotion labels and grammatical gender.

5 Results

5.1 Love and Grammatical Gender

Table 2 presents the contingency counts, odds ratios, Fisher's Exact Test p-values, and Cramér's V effect sizes for the association between *love* and feminine grammatical gender across all five translator groups.

Across both human translator groups, *love* is significantly associated with feminine grammatical gender. Professional translators show a moderate effect ($V = .204$, $p = .017$), with 56.2% of love-labeled reviews receiving feminine gender compared to 25.0% of non-love reviews. Student translators show a comparable pattern with a slightly stronger effect ($V = .285$, $p = .002$), with the same 56.2% feminine rate for love reviews against 17.6% for non-love reviews. These findings extend the results reported by Popović and Lapshinova-Koltunski (2024).

The pattern is more nuanced across LLMs. GPT-4 and Mistral show no significant association between *love* and feminine gender ($p = .313$ and $p = .201$ respectively), with weak effect sizes ($V = .079$ and $V = .106$). Llama, however, shows a significant and moderately strong association ($V = .251$, $p = .005$), with 47.1% of *love*-labeled reviews receiving feminine gender compared to 14.8% of non-*love*-reviews – a pattern comparable in magnitude to the human translator groups. This finding was not anticipated and is discussed further in Section 6 below.

The comparison across the frequent emotion labels shows that no emotion other than *love* has a significant positive association with feminine gender. Two significant negative associations are observed: *neutral* is associated with fewer feminine translations in professional translators ($p = .020$, $V = .208$), and *disapproval* is associated with fewer feminine translations in both Llama ($p = .044$, $V = .168$) and Mistral ($p = .008$, $V = .200$). Since these associations are not consistent across translator groups and are not directly predicted by the gender-emotion literature reviewed above, we interpret them cautiously. The clearest positive association with feminine gender therefore remains specific to *love* in the present dataset.

As noted in Section 3 above, the previous study (Popovic and Lapshinova-Koltunski, 2024) identifies product category as a strong predictor of feminine gender in the corpus under analysis. To assess whether this factor might confound our primary finding, we conduct a descriptive inspection of the distribution of *love*-labeled reviews across the 14 product categories in the corpus, and of the gender choices made by human translators within those reviews. Of the 27 *love*-labeled reviews, only 5 fall within the two product categories identified by Popović and Lapshinova-Koltunski (2024) as most strongly associated with feminine gender in translation (beauty: $n = 1$; home & kitchen: $n = 4$). The remaining 22 reviews span ten categories not typically associated with feminine grammatical gender, including cell phones & accessories,

video games, CD/vinyl, and sports & outdoors. Among the 9 *love*-labeled reviews that receive feminine gender from professional translators, 6 came from categories outside the feminine-coded set (books, cd & vinyl, cell phones & accessories, grocery & gourmet food, toys & games). The same pattern holds for student translators, whose 9 feminine choices for *love*-labeled reviews also included 6 from non-feminine-coded categories. This descriptive pattern suggests that the feminine gender choices associated with *love*-labeled reviews are not concentrated in the product categories that would be expected to attract feminine forms on category grounds alone.

6 Discussion

Our results offer partial support for the assumption that source-text emotion might be associated with grammatical gender choices in translation. The emotion *love* is consistently associated with a higher rate of feminine grammatical gender in both human translator groups, with moderate effect sizes and p-values well below the conventional threshold. This pattern is consistent with well-documented gendered emotion stereotypes in psychology and NLP research.

The behaviour of the three LLMs presents a more complex pattern. With respect to the main *love*-feminine association, GPT-4 and Mistral show no significant effect, with weak effect sizes. This is consistent with the broader finding by Popović and Lapshinova-Koltunski (2024) that machine-generated translations exhibit stronger masculine bias than human translations, defaulting to masculine forms regardless of contextual cues. One plausible interpretation is that these models treat grammatical gender as a largely context-free decision, relying on the statistically dominant masculine form in Russian rather than on implicit contextual cues such as emotion. In this sense, GPT-4 and Mistral may produce less variation than human translators in under-specified first-person contexts.

Llama stands apart from the other two LLMs, showing a significant and moderately strong *love*-feminine association comparable in magnitude to the human translator groups. This was not predicted and we are cautious about over-interpreting it. The result is based on a small number of reviews ($n = 17$ *love*-labeled reviews after filtering), and replication on a larger dataset is necessary before any strong conclusions can be drawn. We can only speculate about possible explanations: differences in training data composition, instruction tuning. We flag this as a priority for future investigation.

7 Conclusion

Addressing RQ1, we find that the association between classifier-assigned emotion labels and grammatical gender is not general across all frequent emotion labels. The clearest positive association concerns *love*: love-labelled English source reviews are translated with feminine grammatical gender more often in both human translator groups, while other frequent emotion labels do not show the same positive pattern.

Addressing RQ2, this pattern differs across translator groups. The association is statistically significant and consistent in translations produced by both professional translators and translation students. Among the three LLMs, only Llama shows a comparable *love*–feminine association, while GPT-4 and Mistral do not.

Within the given limitations, the findings suggest that one theoretically feminine-coded emotion, *love*, may be associated with gendered translation choices in this dataset. The study therefore does not show that source-text emotion in general predicts grammatical gender, but rather that specific emotion labels can be productively examined as possible contextual cues. Future work should replicate these findings on larger corpora and additional language pairs, investigate why different LLMs respond differently to emotional cues, and include manual emotion annotation to validate automated classifier output.

8 Limitations

Several limitations should be noted when interpreting our results.

First, emotion labels are produced by an automated classifier and cannot be interpreted as ground-truth emotion. Fine-grained emotion classification remains a challenging task, with state-of-the-art performance on GoEmotions reaching only a moderate macro-F1 of 0.45 (Lowe, 2022). Class imbalance and inconsistent annotation of rare emotions in the underlying dataset are known issues (Demszky et al., 2020). Although we include a GPT-4o re-annotation as a plausibility check, this is not a substitute for independent human validation. A stronger design would include multiple hu-

man annotators and inter-annotator agreement.

Second, this study examines translation outputs only. We have no access to translator decision logs, think-aloud data, or post-hoc interviews. We therefore cannot determine whether the patterns reflect conscious choices, implicit bias, lexical cues, or something else entirely. This is a standard constraint of corpus-based research.

Third, the study is restricted to a single language pair (English–Russian), a single domain (Amazon product reviews), and an extremely small corpus. Generalisation to other genres, language pairs, and translator populations requires further study. The small number of emotion-labelled cases, especially for individual emotion categories, also means that statistically significant patterns should be interpreted cautiously.

Finally, the analysis is correlational. We identify statistical associations between source-text emotion and grammatical gender choices in translation, but do not claim causal relationships or attempt to model translator decision processes directly.

Taken together, these limitations position the present work as exploratory rather than confirmatory. Our contribution is a methodological proof-of-concept demonstrating that emotion classification can be productively combined with grammatical gender annotation to surface patterns in translation behaviour. The findings justify more rigorous follow-up with larger corpora, human validation, and richer translator metadata.

References

Agresti, Alan. 1990. *Categorical data analysis*. A Wiley-Interscience publication. Wiley, New York [u.a.].

Brody, Leslie R. and Judith A. Hall. 2008. Gender and emotion in context. In Lewis, Michael, Jeannette M. Haviland-Jones, and Lisa Feldman Barrett, editors, *Handbook of Emotions*, pages 395–408. The Guilford Press, New York, 3 edition.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates.

Demszky, Dorottya, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In Jurafsky, Dan, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online, July. Association for Computational Linguistics.

Hackenbuchner, Janiça, Joke Daems, Arda Tezcan, and Aaron Maladry. 2024. You shall know a word's gender by the company it keeps: Comparing the role of context in human gender assumptions with MT. In Savoldi, Beatrice, Janiça Hackenbuchner, Luisa Bentivogli, Joke Daems, Eva Vanmassenhove, and Jasmijn Bastings, editors, *Proceedings of the 2nd International Workshop on Gender-Inclusive Translation Technologies*, pages 31–41, Sheffield, United Kingdom, June. European Association for Machine Translation (EAMT).

Lapshinova-Koltunski, Ekaterina, Maja Popović, and Maarit Koponen. 2022. DiHuTra: a parallel corpus to analyse differences between human translations. In Moniz, Helena, Lieve Macken, Andrew Rufener, Loïc Barrault, Marta R. Costa-jussà, Christophe Declercq, Maarit Koponen, Ellie Kemp, Spyridon Pilos, Mikel L. Forcada, Carolina Scarton, Joachim Van den Bogaert, Joke Daems, Arda Tezcan, Bram Vanroy, and Margot Fonteyne, editors, *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 337–338, Ghent, Belgium, June. European Association for Machine Translation.

Lowe, Sam. 2022. roberta-base-go_emotions. `https://huggingface.co/SamLowe/roberta-base-go_emotions`.

Nesbitt, Muriel N. and Nolan E. Penn. 2000. Gender stereotypes after thirty years: A replication of rosenkrantz, et al. (1968). *Psychological Reports*, 87:493 – 511.

Pangakis, Nicholas, Samuel Wolken, and Neil Fasching. 2023. Automated annotation with generative ai requires validation. *ArXiv*, abs/2306.00176.

Plant, Ashby, Janet Hyde, Dacher Keltner, and Patricia Devine. 2000. The gender stereotyping of emotions. *Psychology of Women Quarterly*, 24:81 – 92, 03.

Plaza-del-Arco, Flor Miriam, Amanda Cercas Curry, Alba Curry, Gavin Abercrombie, and Dirk Hovy. 2024. Angry men, sad women: Large language models reflect gendered stereotypes in emotion attribution. In Ku, Lun-Wei, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7682–7696, Bangkok, Thailand, August. Association for Computational Linguistics.

Popovic, Maja and Ekaterina Lapshinova-Koltunski. 2024. Gender and bias in Amazon review translations: by humans, MT systems and ChatGPT. In Savoldi, Beatrice, Janiça Hackenbuchner, Luisa Bentivogli, Joke Daems, Eva Vanmassenhove, and Jasmijn Bastings, editors, *Proceedings of the 2nd International Workshop on Gender-Inclusive Translation Technologies*, pages 22–30, Sheffield, United Kingdom, June. European Association for Machine Translation (EAMT).

Rosenkrantz, Paul, Susan Vogel, Helen Bee, Inge Broverman, and Donald Broverman. 1968. Sex-role stereotypes and self-concepts in college students. *Journal of Consulting and Clinical Psychology*, 32:287–295, 06.

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

Stanovsky, Gabriel, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In Korhonen, Anna, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy, July. Association for Computational Linguistics.

Vanmassenhove, Eva. 2024. Gender bias in machine translation and the era of large language models. *ArXiv*, abs/2401.10016.

Weber, Sabine, Lynn Greschner, and Roman Klinger. 2026. Less is more? the role of demographic author information in emotion classification of ambiguous text. In Piperidis, Stelios, Núria Bel, Henk van den Heuvel, Nancy Ide, Simon Krek, and Antonio Toral, editors, *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 8147–8161, Palma, Mallorca, Spain, May. European Language Resources Association (ELRA).

Yeswa-Stories: A Three-Way Parallel Dataset of Female African Figures in Low-Web Data Languages

Bethelhem Mamo
Addis Ababa University / Addis Ababa
bethelhem.y.mamo@gmail.com

Hellina Hailu Nigatu
DAIR Institute / Addis Ababa
hellina@dair-institute.org

Abstract

Language technologies used in everyday settings such as machine translation systems risk perpetuating societal bias. Prior work in creating benchmarks for gender bias in machine translation systems 1) focus primarily on high-resourced language pairs or a low-resourced language paired with a high-resource language, 2) use template based benchmarks that usually focus on occupational biases and stereotypes, and 3) translate high-resource benchmarks which may lack cultural significance to low-resourced languages. In this paper, we introduce *Yeswa-Stories*[1], a three-way parallel dataset comprising 1,300 aligned sentences in Amharic, Afaan Oromo, and Tigrinya. We constructed the dataset in two ways: first, we collected English sentences from Wikipedia articles about notable African women and translated them into the three target languages using human translators. To improve cultural representativeness, we further augment the dataset with locally sourced content reflecting the cultural context where the languages are spoken. Our dataset contributes a new resource for studying gender-inclusive translation in low-resourced settings.[2]

[1]"Yeswa" means "her" in Amharic.

[2]https://github.com/hhnigatu/Yeswa-Stories

1 Introduction

Gender bias in machine translation systems has received increasing attention, particularly in how models misrepresent or underrepresent women across languages (Savoldi et al., 2021; Savoldi et al., 2024).

Several works have introduced measurement and mitigation methods, especially for high-resource languages (Stanovsky et al., 2019; Prates et al., 2020). However, existing resources for gender bias evaluation often rely on template-based constructions or narrowly defined domains. These approaches fail to capture the richness of natural language and the diversity of cultural contexts in which gender is expressed. Further, these resources are limited to a handful of high-resourced languages (Joshi et al., 2020).

Creating resources for low-resourced languages is challenging due to the scarcity of parallel corpora, limited digital presence, and the lack of culturally grounded data that reflects real-world usage (Talat et al., 2022). Further, the term "low-resourced language" encompasses a wide array of languages with varying levels and types of resources (Nigatu et al., 2024). Following the recommendation in Nigatu et al., (2024), we specifically use the term "low-web data languages" in this paper to indicate that the resource that the languages of focus are lacking for the context of this paper is data.

In this paper, we introduce *Yeswa-Stories*, a three-way parallel dataset in Amharic, Afaan Oromo, and Tigrinya. Amharic and Tigrinya are Afro-Semitic languages that use the Ge'ez writing script. Afan Oromo is a Cushitic language that uses the Qubee alphabet and is written with Latin characters. While Tigrinya and Amharic are

grammatically gendered languages–meaning every noun is assigned a gender–Afan Oromo has a neutral pronoun. Further, Amharic has a neutral respectful pronoun while Tigrinya has a gendered respectful pronoun. Our dataset focuses on female-centered narratives and combines globally sourced content with locally relevant materials. By doing so, we provide a resource that supports research on gender-inclusive translation and highlight the importance of cultural diversity in low-resource settings.

The paper is structured as follows: We first review related literature, with a focus on gender bias and existing datasets for low-resource machine translation in Section 2. We then present the steps we took and challenges we faced in creating this dataset in Section 3. Finally, we present an analysis of the dataset and evaluate the performance of two widely used machine translation systems in Section 4. Our work contributes to the broader gap in lack of resources in these three low-web data languages and specifically to the gap in the representation of women in multilingual datasets.

2 Related Work

Gender bias evaluation in machine translation has traditionally been conducted using controlled benchmarks, often constructed with template-based sentences that explicitly encode gender variations. These benchmarks typically focus on occupational stereotypes (e.g., associating certain professions with a specific gender) and pronoun resolution tasks (Stanovsky et al., 2019; Cho et al., 2019). However, recent work has shown that such benchmarks often fail to capture real-world distributions of gender in datasets and may overlook biases already embedded in training data (Nigatu et al., 2025). While effective for isolating specific biases, such approaches are predominantly developed for high-resource languages and are often directly translated or adapted for low-resource settings. This adaptation assumes linguistic and cultural equivalence, which may not hold in practice, leading to evaluations that overlook culturally specific expressions of gender (Talat et al., 2022).

Beyond measurement, the impact of gender bias in machine translation systems is significant. Biased outputs can reinforce harmful stereotypes and contribute to representational harm by systematically underrepresenting or misrepresenting women (Blodgett et al., 2020). These issues extend beyond academic concerns, affecting real-world applications such as hiring, education, and access to information. Moreover, correcting such biases often requires substantial resources, including human post-editing and system retraining, which introduces economic disparities (Savoldi et al., 2024). As a result, already marginalized communities bear a disproportionate burden in addressing these shortcomings, highlighting the importance of developing more inclusive systems.

Prior work on gender bias in low-resource languages has begun to reveal the depth of the problem, though research remains sparse compared to high-resource settings (Wairagala et al., 2022; Ramesh et al., 2023). Most related to our work, Sewunetie et al., (2024) constructed a benchmark of 2,400 gender-balanced sentences for Amharic, Afaan Oromo, and Tigrinya and evaluated Google Translate and the NLLB model against it using both automatic metrics and human annotation. Their human evaluators find that approximately 93% of English-to-Afaan-Oromo translations, 81% of English-to-Tigrinya translations, and 73% of English-to-Amharic translations exhibit gender bias, making theirs the most comprehensive gender bias benchmark for these three languages to date. Crucially, Sewunetie et al., (2024) show that the bias is rooted in the linguistic structure of each language: Amharic and Tigrinya encode grammatical gender in pronouns, verb agreement, and many occupational titles, while Afaan Oromo conveys gender primarily through context rather than morphology, meaning that bias manifests differently across the three languages. They also demonstrate that professional names in standard datasets are overwhelmingly presented in their masculine form, directly linking dataset-level imbalance to the observed translation errors. This finding motivates our own approach: rather than constructing another template-based occupational benchmark, *Yeswa-Stories* introduces naturally occurring, female-centred narratives that expose translation systems to a broader and more culturally grounded distribution of gendered language. We treat the benchmark of Sewunetie et al., (2024) as a reference point for interpreting our evaluation results and for situating our contribution within the emerging literature on gender bias in Ethiopian language MT.

3 Data Collection

The dataset was constructed through a multi-stage and iterative process. Initially, we aimed to collect stories and narratives specifically centered on Ethiopian women directly from local sources. However, we found that such content was either scarce, not digitized, or not available in a structured format suitable for dataset creation. Similarly, our attempts to locate parallel or comparable articles focusing on women across Amharic, Afaan Oromo, and Tigrinya were largely unsuccessful.

Due to these limitations, we shifted our approach to use English Wikipedia as a starting point. We collected sentences from articles about notable African women, including scientists, public figures, and other well-documented individuals. These sentences served as a base for constructing parallel data.

We then translated the collected sentences into Amharic, Afaan Oromo, and Tigrinya using human translators. This process posed significant challenges due to the limited availability of skilled translators and budget constraints, requiring careful selection of sentences and prioritization of quality over quantity.

To address the lack of cultural representation in Wikipedia-derived content, we further incorporated locally sourced materials written in the target languages. These included articles from Ethiopian news outlets and cultural platforms that reflect traditional roles, festivals, and everyday experiences of women. These texts were then translated into the remaining languages to ensure full three-way parallel alignment.

The final dataset is the result of combining these two sources, balancing globally available narratives with culturally grounded content. This design choice aligns with recent findings that emphasize the importance of incorporating bias considerations during dataset creation, rather than relying solely on post-hoc evaluation (Nigatu et al., 2025).

3.1 Data Description

The final dataset consists of 1,300 parallel sentences aligned across Amharic, Afaan Oromo, and Tigrinya. The dataset includes a mix of globally sourced and locally grounded content. Topics covered include biographies of notable women, descriptions of cultural practices, traditional roles, and narratives related to Ethiopian festivals and community life. This diversity allows the dataset to capture different dimensions of gender representation, moving beyond narrow professional domains and incorporating culturally specific perspectives. All sentences are manually translated, ensuring high-quality alignment and consistency across the three languages. The dataset is intended to support research on gender representation, bias analysis, and inclusive translation.

4 Results

We evaluate the dataset by analysing the performance of two widely used machine translation systems—NLLB (NLLB Team et al., 2022) and Google Translate—on *Yeswa-Stories*. We report BLEU (Post, 2018) and chrF (Popović, 2017) scores computed against human reference translations across all six translation directions. Because Amharic, Tigrinya, and Afaan Oromo are morphologically rich languages in which a single concept can be realised through many valid surface forms, chrF—which operates at the character level and is therefore more tolerant of valid paraphrases—is our primary metric. BLEU scores are reported for completeness and comparability with prior work. Table 1 presents the results. Three patterns emerge clearly from the scores.

Overall scores are low, confirming dataset difficulty. Even Google Translate—a well-resourced commercial system—achieves chrF scores ranging from 25.78 to 49.61, and BLEU scores between 4.79 and 15.80 across the six directions. NLLB performs considerably worse, with BLEU scores as low as 0.87 for Afaan Oromo→Tigrinya. These figures are substantially below scores typically reported on standard benchmarks such as FLORES for the same language pairs, indicating that the female-centred narratives and culturally grounded content in *Yeswa-Stories* present a harder translation challenge than the news and Wikipedia-domain text on which current systems are predominantly trained and evaluated. This supports our central argument that template-based and domain-narrow benchmarks are insufficient for capturing the full range of gendered language use in these communities.

Google Translate consistently outperforms NLLB. Google Translate surpasses NLLB in every direction and on both metrics, with the gap most pronounced for Afaan Oromo as the source language: chrF differences of

		NLLB		Google Translate	
Direction	**#Sents**	**BLEU**	**chrF**	**BLEU**	**chrF**
Afaan Oromo → Amharic	1,287	1.31	14.78	12.43	36.76
Afaan Oromo → Tigrinya	1,299	0.87	12.18	4.79	25.78
Amharic → Afaan Oromo	1,300	4.53	40.29	13.65	49.61
Amharic → Tigrinya	1,300	5.71	29.01	6.99	30.44
Tigrinya → Amharic	1,300	8.80	36.64	15.80	42.27
Tigrinya → Afaan Oromo	1,300	3.26	36.86	9.22	45.52

Table 1: BLEU and chrF scores for NLLB and Google Translate on *Yeswa-Stories* across all six translation directions, evaluated against human reference translations. Higher is better for both metrics.

22.0 points (Oromo→Amharic) and 13.6 points (Oromo→Tigrinya). This pattern replicates and extends the finding of Sewunetie et al. (2024), who similarly observed NLLB underperforming Google Translate on gender-sensitive content for these three languages—there using an occupational benchmark, and here using naturally occurring female-centred narratives. The consistency of this finding across two structurally different datasets strengthens the conclusion that NLLB's training data is particularly poorly suited to gender-inclusive content in Ethiopian languages.

Afaan Oromo as a source language is the hardest direction. Both systems perform worst when translating from Afaan Oromo into either Amharic or Tigrinya. This is consistent with the linguistic profile of the language: as a Cushitic language, Afaan Oromo is typologically distant from both Semitic targets, and it conveys gender primarily through context rather than overt morphological marking (Sewunetie et al., 2024). The absence of explicit gender morphology in the source makes it harder for MT systems to select the correct gendered form in the target, compounding the general data scarcity for this language pair. Conversely, both systems perform best when Amharic is the source language, reflecting its comparatively larger digital presence and the availability of more parallel training data.

4.1 Qualitative Analysis

Beyond aggregate scores, we manually examined a subset of annotations to characterise the types of errors produced by each system. Figure 1 presents four representative examples drawn from the annotation of Afaan Oromo → Amharic translations. The examples illustrate distinct failure modes that aggregate metrics alone cannot capture.

The examples in Figure 1 reveal three recurring error patterns that are particularly consequential for a female-centred dataset. First, NLLB produces **hallucinated content** wholly unrelated to the source: in Example 1, a sentence about Mary Anning's geological expertise is rendered as a description of art and literary work, erasing her scientific identity entirely. Second, both systems exhibit **agency removal**, where the grammatical subject—always a woman in *Yeswa-Stories*—is dropped or neutralised. In Examples 3 and 4, NLLB converts active constructions in which Anning is the explicit actor into passive or impersonal forms, so that her contribution is no longer expressed. This is a direct instance of the representational harm described by Blodgett et al. (2020), the woman is present in the source but invisible in the translation. Third, we observe **lexical substitution errors** that distort factual content, as in Example 2 where Google Translate replaces "cliffs" with "shelters" () and mistranslates the season, producing a sentence that is both factually incorrect and incoherent. These qualitative findings complement the low aggregate scores in Table 1 and illustrate why female-centred narratives such as those in *Yeswa-Stories* are essential for exposing failure modes that template-based occupational benchmarks do not surface.

5 Discussion and Conclusion

Our work highlights several important considerations for gender-inclusive translation in low-resource settings. First, relying solely on global sources such as Wikipedia can lead to skewed representations of women, often emphasizing specific professions while neglecting culturally grounded roles. Second, the inclusion of local data sources is crucial for capturing the diversity of gendered experiences and expressions. This is particularly important for languages and communities that are underrepresented in existing NLP resources.

Finally, the creation of high-quality parallel corpora through human translation remains a resource-intensive process, but it is essential for ensuring linguistic and cultural accuracy. We hope

this resource will support future research on gender bias evaluation and culturally aware machine translation for low-resource languages.

Carbon Impact Statement

For our evaluation experiments, we used Google Colab platform to run inference for the NLLB model using a CPU runtime session. We used the NLLB-base model which has 1.3B parameters. During inference we used maximum token length of 200, and used beam search with a beam size of 4. We also enabled early stopping. Since Google Translate is a closed-source model, we do not have the model details or the runtime environment details.

References

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

Cho, Won Ik, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. 2019. On measuring gender bias in translation of gender-neutral pronouns. *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, 173–181.

Joshi, Pratik, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6282–6293.

Nigatu, Hellina Hailu, Bethelhem Yemane Mamo, Bontu Fufa Balcha, Debora Taye Tesfaye, Elbethel Daniel Zewdie, Ikram Behiru Nesiru, Jitu Ewnetu Hailu, and Senait Mengesha Yayo. 2025. Evaluating machine translation datasets for low-web data languages: A gendered lens. *Proceedings of the Findings of the Association for Computational Linguistics*.

Nigatu, Hellina Hailu, Atnafu Lambebo Tonja, Benjamin Rosman, Thamar Solorio, and Monojit Choudhury. 2024. The Zeno's Paradox of Low-Resource Languages *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, and Al Youngblood et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Popović, Maja. 2017. chrF++: Words helping character n-grams. *Proceedings of the Second Conference on Machine Translation*, 612–618.

Post, Matt. 2018. A call for clarity in reporting BLEU scores. *Proceedings of the Third Conference on Machine Translation: Research Papers*, 186–191.

Prates, Marcelo O. R., Pedro H. Avelar, and Luís C. Lamb. 2020. Assessing gender bias in machine translation: A case study with Google Translate. *Neural Computing and Applications*, 32:6363–6381.

Ramesh, Krithika, Sunayana Sitaram, and Monojit Choudhury. 2023. Fairness in language models beyond English: Gaps and challenges. *Findings of the Association for Computational Linguistics: EACL 2023*, 2106–2119.

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

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

Sewunetie, Walelign, Atnafu Tonja, Tadesse Belay, Hellina Hailu Nigatu, Gashaw Gebremeskel, Zewdie Mossie, Hussien Seid, and Seid Yimam. 2024. Gender bias evaluation in machine translation for Amharic, Tigrigna, and Afaan Oromoo. *Proceedings of the 2nd International Workshop on Gender-Inclusive Translation Technologies*, 1–11.

Stanovsky, Gabriel, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1679–1684.

Talat, Zeerak, Aurelie Neveol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, Shanya Sharma, Arjun Subramonian, Jaesung Tae, Samson Tan, Deepak Tunuguntla, and Oskar Van Der Wal. 2022. You reap what you sow: On the challenges of bias evaluation under multilingual settings. *Proceedings of the ACL Workshop on Challenges & Perspectives in Creating Large Language Models*, 26–41.

Wairagala, Eric Peter, Jonathan Mukiibi, Jeremy Francis Tusubira, Claire Babirye, Joyce Nakatumba-Nabende, Andrew Katumba, and Ivan Ssenkungu. 2022. Gender bias evaluation in Luganda-English machine translation. *Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas*, 274–286.

Source (Afaan Oromo)	MT Output (Amharic)	System	Gender Bias	Adequacy	Annotator Comment
Anning became well known in geological circles in Britain, Europe, and America, and was consulted on issues of anatomy as well as fossil collecting	አኒንግ ኮራቱ በዩናይትድ ኪንግደም፣ በአውሮፓና በአሜሪካ ውስጥ ታዋቂነት ያተረፈ ሲሆን በሥነ-ጥበብና በሥነ-ጽሑፍ ሥራዎች እንዲሁም በሥነ-ጽሑፍ አዘጋጅነትም ይታወቃል	NLLB	Yes	No	Unrelated translation: content about art and literature substituted for geology and fossil collecting
Anning became well known in geological circles in Britain, Europe, and America, and was consulted on issues of anatomy as well as fossil collecting	አኒንግ በብሪታንያ፣ አውሮፓ እና አሜሪካ ታዋቂ ተመራማሪ እና የእንስሳት አፅም እና የታሪክ ቅሪቶች አማካሪ ናቸው።	Google	No	Yes	Accurate rendering of her role as researcher and advisor
The cliffs could be dangerously unstable, especially in winter when rain weakened them, causing landslides	እነዚያ መጠለያዎች በተለይ በበጋ ወቅት ወንዞች በሚጥሉበት ጊዜ ችግር እና አደጋ ሊያስከትሉ ይችሉ ነበር።	Google	No	No	"መጠለያዎች" (shelters) used instead of ``cliffs''; season also mistranslated ("winter" is translated as "summer")
Her investigations played a crucial role in several finds...coprolites...belemnite fossils	በወቅቱ "የዞር ድንጋዮች"...ከብዙዎች የበለጠ ተገልጧል	NLLB	Yes	No	Named entity mistranslation; action no longer attributed to her --- agency erased and sentence turned neutral
Henry De la Beche...sold prints for her benefit	ይሁን እንጂ...የተሰራጩት በአኒንግ የተገኙት ቅሪተ አካላት ላይ የተመሠረቱ ነበሩ።	NLLB	Yes	No	Subject (she/her) removed; sentence turned passive and neutral --- her agency in the discovery is no longer expressed

Figure 1: Representative annotation examples from the Afaan Oromo → Amharic subset of Yeswa-Stories. Underlined phrases highlight the specific mistranslated or problematic text identified by annotators. Gender Bias and Adequate Translation are majority-vote human annotation labels (Yes/No). We put a figure of the table since the conference guidelines do not support Ge'ez script fonts.

A Additional Result

In this section, we provide qualitative results. In Figure 1, we provide representative annotations to complement the quantitative results reported in Table 1.

A Chinese Challenge Set to Assess Gender Bias in Automated Translation

Xiaolan Xu[1], Sara Mendes[1, 2], Yu-Yin Hsu[3]

[1] Faculty of Arts and Humanities, University of Lisbon
xuxiaolan@edu.ulisboa.pt, s.mendes@edu.ulisboa.pt
[2] Center of Linguistics of the University of Lisbon
[3] Department of Language Science and Technology, The Hong Kong Polytechnic University
yyhsu@polyu.edu.hk

Abstract

Gender bias in automated translation (AT) is well-documented for English-source language pairs, while source languages generally lacking grammatical gender remain largely understudied. This paper addresses the Chinese–Portuguese direction, a language pair that has received little attention in this context. We developed a Chinese challenge set of 495 sentences constructed from 45 occupations across an eleven-sentence template matrix, systematically varying the type and syntactic position of gender cues: no cue, explicit prenominal modifiers, and coreferential pronouns varying in position and syntactic complexity. We tested two commercial neural machine translation (NMT) systems and six large language models (LLMs) with this challenge set. Results show a clear hierarchy of cue effectiveness: explicit prenominal modifiers yield universal 100% accuracy; in the absence of gender cues, models predominantly default to masculine forms; and coreferential pronouns in complex sentences reveal a pronounced masculine–feminine asymmetry. Also, LLMs demonstrate more symmetric gender cue processing than NMT systems. Our data and code are available at https://github.com/xu-xiaolan/chinese-challenge-set-gender-bias.

1 Introduction

Automated translation (AT), encompassing both neural machine translation (NMT) systems and general-purpose large language models (LLMs) used for translation tasks, has become a central tool for cross-linguistic communication, although its reliance on large-scale naturally occurring data makes it susceptible to reproducing and amplifying the biases present in human-generated text (Zhao et al., 2017; Blodgett et al., 2020). Gender bias is among the most thoroughly studied bias phenomena: training corpora tend to over-represent male-centric contexts in comparison to female-centric ones, and models trained on such data may encode stereotypes that surface systematically in their outputs (Savoldi et al., 2021).

In this context, the research on gender bias in MT has expanded substantially over the past decade. Stanovsky et al. (2019) introduced the first challenge set and evaluation protocol for measuring gender bias in MT, revealing systematic inaccuracies across four commercial systems and two academic models. Subsequent work has proposed a range of mitigation strategies (Font and Costa-Jussà, 2019; Saunders and Byrne, 2020; Costa-Jussà et al., 2022; Fleisig and Fellbaum, 2022). Challenge sets and benchmarks have been developed for numerous language pairs and systems (Piergentili et al., 2023; Sewunetie et al., 2024), and the role of sentence context in shaping AT gender behavior has received growing attention (Sharma et al., 2022; Hackenbuchner et al., 2024; Gkovedarou et al., 2025). Despite this breadth of research topics, the existing literature exhibits a consistent skew: English is almost invariably the source language, and the language pairs studied are largely Indo-European, leaving

typologically distinct languages such as Chinese–Portuguese largely unexamined.

Gender bias in AT manifests in several ways, including the default assignment of masculine forms to gender-ambiguous referents and the stereotypical gendering of occupational nouns (Bonifácio et al., 2025). These errors are particularly consequential when the target language encodes grammatical gender morphologically: in Portuguese, a single mistaken gender assignment propagates across determiners, adjectives, and participial forms through morphosyntactic agreement. Furthermore, the translation of gender information is particularly challenging in the Chinese–Portuguese direction, as in Chinese, gender is marked only in written forms of third-person pronouns (他/她/它, 'he/she/it', all pronounced *tā*), while nouns carry no morphological gender marking. In this way, when translating from Chinese into Portuguese, AT systems must resolve a gender value that is simply absent in the source.

Besides the linguistic asymmetry outlined above, studying the Chinese–Portuguese language pair is also motivated by several practical reasons. Chinese and Portuguese are among the most widely spoken languages globally, and expanding economic and cultural ties between China and Portuguese-speaking countries have substantially increased translation demand in this direction, generating a substantial and increasing translation demand between these two languages. Also, the Special Administrative Region of Macao, where both Chinese and Portuguese are the official languages and which China has been promoting and supporting as a hub between China and Portuguese-speaking countries, represents a unique context in which Chinese–Portuguese translation is an everyday institutional reality, making the accuracy and fairness of AT systems involving this language pair a matter of immediate practical consequence.

The present study addresses the construction and evaluation of a Chinese challenge set, designed to assess how different linguistic conditions in Chinese source sentences impact gender bias in Chinese–Portuguese AT outputs. The challenge set is developed with three objectives: (1) to develop a replicable methodology and dataset for evaluating gender bias in AT when Chinese is the source language; (2) to evaluate how different linguistic conditions impact gender bias across AT systems; and (3) to propose a replicable methodology for constructing gender bias challenge sets for AT when no such datasets are available.

To achieve these goals, we evaluate the outputs of eight AT systems: two commercial NMT engines and six recently released LLMs.

Section 2 reviews related work on gender bias in AT and provides the linguistic background for the study. Sections 3 and 4 detail the experimental setup and Chinese dataset construction. Section 5 reports the results, and Section 6 discusses the findings and presents conclusions and future work.

2 Related Work and Background

2.1 Gender Bias in AT

Gender bias in MT has attracted considerable attention over the past decade, since MT training data often reflects societal gender imbalances, leading models to perpetuate or amplify these biases (Zhao et al., 2017). Such inaccuracies not only distort meaning but also increase post-editing time and cost (Savoldi et al., 2024).

To evaluate and measure gender bias across MT systems, Stanovsky et al. (2019) established the foundational benchmark WinoMT, a challenge dataset and evaluation protocol, which revealed systematic inaccuracies in both industrial and academic systems. Subsequent work proposed various mitigation strategies for NMT models, including debiasing word embeddings (Font and Costa-Jussà, 2019), transfer learning (Saunders and Byrne, 2020), and novel learning frameworks (Fleisig and Fellbaum, 2022), though these efforts have focused predominantly on English as the source language. Even if work has been developed on other languages, e.g. Prates et al. (2020) demonstrated that gender-neutral languages can serve as a lens for examining gender bias in MT; and Cho et al. (2019) constructed a Korean gender bias test set, highlighting the need for evaluation resources for other languages, Chinese as a source language, and the Chinese–Portuguese language pair remain underexplored.

Recent work has begun to examine gender bias in AT outputs produced by LLMs, revealing that base LLMs exhibit a higher degree of gender bias than NMT models in English-to-Catalan and English-to-Spanish translation (Sant et al., 2024), although according to these authors, prompt engineering reduces gender bias by up to 12%. Additionally, Kaneko et al. (2024) show that Chain-of-Thought prompting reduces LLM gender bias

even on simple tasks. Popović and Lapshinova-Koltunski (2024) add another interesting result, showing that ChatGPT introduces different gender bias patterns when compared to MT systems. These findings suggest that model architecture plays a meaningful role in bias behavior, motivating cross-system comparative analyses.

2.2 Grammatical Gender in Chinese and Portuguese

Chinese has a semantic gender assignment system according to Corbett (1991), in which gender is only visible in the pronominal domain, and only in third-person pronouns in written Chinese. Consequently, there are many "personal dual gender" nouns (Quirk and Crystal, 2010), i.e., nouns whose gender is not encoded in the noun itself but must be inferred from context. Portuguese, by contrast, has a semantic-and-formal gender system (Corbett, 1991), in which gender is an inherent property of nouns and triggers agreement in various domains (Raposo et al., 2013). This asymmetry is a crucial motivation for our study, as it creates a systematic challenge for models performing translation between this language pair: the system must infer and assign grammatical gender using contextual cues that Chinese does not grammatically encode.

Chinese. Chinese is therefore typically classified as a language with covert gender. As an isolating language, Chinese words and characters do not present any morphological changes, independently of the head noun or any other constituents. This formal invariance leads to the classification of Chinese as a language with a pronominal gender system, i.e., a language where gender distinctions are manifested solely through personal pronouns (Corbett, 1991), and only in third-person pronouns in written text.

Chinese utilizes a tripartite distinction in its written third-person singular pronouns: 他 (*tā*, 'he') for masculine human referents, 她 (*tā*, 'she') for feminine human referents, and 它 (*tā*, 'it') for non-human entities, i.e., animals or inanimate objects. The corresponding plural forms are constructed simply by appending the plural morpheme '们 (*mén*)' to these singular bases. Despite these orthographic distinctions, the spoken form remains an identical homophone (*tā*). Thus, in oral communication Chinese does not offer any gender cues, relying entirely on context for referent identification.

Secondly, gender assignment in Chinese is predominantly determined by semantic criteria. Human referents are assigned gender according to their biological sex (male or female), while non-human referents fall into the neuter category. Human denoting nouns illustrate this semantic reliance in a particularly clear way. Lexical items such as 司机 ('driver'), 护士 ('nurse'), and 老师 ('teacher') are inherently underspecified with respect to gender. In language data involving this type of nouns, the noun itself remains invariant, and the gender of the referent is surfaced only through the gender of coreferential pronouns or explicit lexical modifiers.

When gender must be specified for clarity, Chinese employs lexical encoding rather than morphological inflection. For human referents, this is achieved by prefixing the noun with the morphemes 男 ('male') or 女 ('female'), as illustrated in (1):

(1) (a) 男司机 ('male driver') / 女司机 ('female driver')
    (b) 男护士 ('male nurse') / 女护士 ('female nurse')

Portuguese. The Portuguese language instantiates a pervasive grammatical gender system distinguishing masculine and feminine categories (Raposo et al., 2013). Gender is an inherent morphosyntactic property of the noun that governs agreement across the sentence.

Assignment follows both semantic and formal criteria (Kibort and Corbett, 2008): while nouns referring to male referents are typically masculine and those referring to female referents feminine, the gender of inanimate nouns is largely arbitrary (*copo* [masc., 'glass'] vs. *janela* [fem., 'window']), creating a significant challenge for computational models that must encode gender as a lexical property.

Occupation nouns exhibit several patterns. Some have different forms for masculine and feminine (*enfermeiro*/*enfermeira*, 'nurse'), others maintain a single form relying on the determiner to signal gender (*o*/*a estudante*, 'the student'). This variation is directly relevant to AT evaluation, as the surface form of the translated occupation noun constitutes the primary observable evidence of gender bias in translation output.

The true complexity of the Portuguese system lies in its agreement domain. Gender is not a

localized feature; it percolates through the entire nominal phrase and beyond. Determiners, including definite articles (*o*/*a*, 'the'), indefinite articles (*um*/*uma*, 'a'), and demonstratives (*este*/*esta*, 'this'; *aquele*/*aquela*, 'that'), must agree with the head noun in gender, as must adjectives with gendered forms. The numeral and pronominal systems are also partially gendered: cardinal numbers (*um*/*uma*, 'one'; *dois*/*duas*, 'two'), possessives (*meu*/*minha*, 'my'), and third-person pronouns (*ele*/*ela*, 'he/she') all carry gender distinctions. Gender agreement further extends to the verbal system: in passive constructions, past participles also agree in gender and number with the subject (*foi aberta* [fem.][sg.], 'was opened'). This level of syntactic integration means that a single gender assignment can dictate the morphological form of multiple subsequent words.

These structural properties create a fundamental challenge for Chinese–Portuguese AT: gender must be overtly expressed across multiple syntactic domains in the target, although it is entirely absent from the source. Failure to identify the correct gender value in the source produces agreement errors and may generate misinterpretation due to unintended bias in the translation output.

2.3 Linguistic Conditions as Gender Cues

Gender marks in languages are not fixed properties. At the same time, the gender assigned to a referent in the target language can shift depending on the linguistic information available in the source sentence.

The most extensively studied gender cue type is the coreferential pronoun. Stanovsky et al. (2019) built WinoMT on this principle: source sentences describe a scenario involving two occupation nouns, with a personal pronoun indicating the gender of one referent (e.g., *The developer argued with the designer because* ***she*** *did not like the design*).

These sentences are structurally ambiguous with respect to pronoun reference, requiring AT systems to resolve the referent without the pragmatic resources available to human readers. WinoMT thus constitutes a sensitive instrument for detecting gender bias: when a system defaults to a stereotypical gender assignment rather than following the pronominal cue, the bias becomes directly observable in the translation output. The interaction between pronoun cues and occupational stereotypes has since become a standard axis of evaluation in gender bias benchmarks (Levy et al., 2021; Sewunetie et al., 2024; Gkovedarou et al., 2025).

Beyond pronouns, sentence-level context has been shown to exert a meaningful influence on AT outputs. Hackenbuchner et al. (2024) compared gender translations of role names presented in isolation versus embedded in full sentences: out of context, 95% of role names were rendered in the generic masculine, whereas in sentence context this proportion dropped to 82%, with words such as *coordinator*, *mechanic*, and *musician* shifting gender depending on the surrounding context.

A third category of cue is the proper name. Although names function as implicit gender signals in many languages, they have been identified as unreliable carriers (Saunders and Olsen, 2023); they are therefore not considered as a condition in the present challenge set.

When the source is a language with a pronominal gender system such as Chinese, whether these cue types function in the same way remains an open empirical question that motivates the typology of linguistic conditions developed in Section 4.

3 Methodology

3.1 Translation Systems

To ensure broad coverage of the current automated translation landscape, eight systems were selected for evaluation: two commercial NMT engines and six LLMs. The two NMT systems (Google Translate and DeepL Translator) were accessed via API on February 26, 2026. The six LLMs were accessed via API on March 1 and 2, 2026, representing a diverse range of developer origins (United States, France, and China) and access modalities (proprietary and open-source): GPT-5.2 (OpenAI, hereafter GPT), Claude Sonnet 4.6 (Anthropic, hereafter Claude), Ministral-14B-2512 (Mistral AI, hereafter Ministral), Qwen3.5-397B-A17B (Alibaba, hereafter Qwen), Llama-3.3-70B-Instruct (Meta, hereafter Llama), and DeepSeek V3.2 (DeepSeek, hereafter DeepSeek). All systems were instructed to output European Portuguese.

For the LLM systems, translation was performed via API calls with temperature set to 0 across all models to minimize output variability and ensure reproducibility. Each sentence was submitted as an independent query with no prior con-

versational context, eliminating any potential influence of cross-sentence memory on gender output. All LLMs received an identical prompt: "*Translate the Chinese sentence into European Portuguese. Output only the translated text. Do not include quotes or explanations.*"

3.2 The Data

We used a purpose-built dataset (described in detail in Section 4) consisting of 495 Chinese sentences involving 45 occupational nouns (15 male-biased, 15 female-biased, and 15 unbiased) across 11 conditions encoding different gender cue types (Section 4.3). The 495 source sentences were translated into European Portuguese by the 8 systems described in Section 3.1, generating 3,960 target sentences.

3.3 Annotation Procedure

To evaluate the extent to which the eight models considered in this study were able to take into account different linguistic cues in the source sentence to achieve correct gender assignment in the target, the 3,960 outputs generated underwent an annotation process.

The annotation focused on the gender of all the elements in the translated noun phrase headed by the occupation target noun. Overall translation quality or syntactic correctness was not considered at this stage. This targeted approach reflects the focus of this study: determining which gender value does each AT system assign to the target noun, and whether this assignment aligns with the gender information present in the Chinese source.

Gender was determined firstly and primarily from the morphological form of the occupation noun in Portuguese. As all gendered elements in the noun phrase were considered, in cases where the noun form is invariant across genders, gender was inferred from the determiner (*o*/*a*, *os*/*as*, *este*/*esta*, *aquele*/*aquela*). Gender assignment was counted as correct if and only if the gender of all the elements in the target noun phrase matched the gender specified in the source, irrespective of number agreement or gender marking outside the nominal phrase.

Annotation was aided by a rule-based semi-automated procedure. A reference lexicon was compiled containing the masculine and feminine Portuguese surface forms of all 45 occupation nouns included in the challenge set (see Section 4). Each system output was automatically matched against this lexicon: tokens identified as masculine forms were labelled M, feminine forms were labelled F. Cases not recognized in the matching process, as well as occupation nouns with identical masculine and feminine Portuguese forms, were manually annotated by the first author, a trained linguist and native speaker of Chinese with advanced proficiency in Portuguese. All annotation labels were then reviewed by the same annotator to ensure consistency, and later validated by the second author, a native speaker of Portuguese.

4 Chinese Challenge Set Construction

Although a large-scale Chinese–Portuguese parallel corpus is available (Chao et al., 2018), it was compiled from news, legal, and general web sources and was not designed for gender-related linguistic research, rendering it unsuitable for gender bias evaluation without substantial curation. As no suitable dedicated challenge set exists for the Chinese–Portuguese language pair, we developed a methodology to generate such data from existing sources.

4.1 Data Source and Occupation Selection

To select the target nouns considered in this study, we drew on data from the U.S. Bureau of Labor Statistics (BLS) 2024 Current Population Survey (CPS), specifically on the information reported regarding: *Employed persons by detailed occupation, sex, race, and Hispanic or Latino ethnicity*[1], which provides nationally representative statistics on the gender composition of the U.S. labor force.

The BLS statistics are employed here as an operationalization criterion for classifying occupations into gender stereotype categories, as no comparable data is available for China or Portuguese-speaking countries. The classification serves to ensure a replicable and empirically grounded partition of occupations; it does not presuppose that the same distributions obtain in the cultural contexts most relevant to the source or target language of this study. The stereotype classification (see 4.2) was nonetheless validated against data available for Portugal[2] and the EU[3] which presents in-

[1] Available at https://www.bls.gov/cps/cpsaat11.htm. Accessed on February 22, 2026.

[2] Available at https://www.cig.gov.pt/wp-content/uploads/2023/11/BE2023trabalho.pdf. Accessed on May 15, 2026.

[3] Available at https://ec.europa.eu/eurostat/web/products-eurostat-news/-/edn-20180307-1. Accessed on May 15, 2026.

formation aggregated into groups of professional activities.

From the aforementioned data, we selected 45 occupations to achieve a balanced and representative distribution across the gender spectrum. Only occupations corresponding to lexicalized and widely recognized professional categories in Chinese and Portuguese were selected. The selection also prioritized occupations with a total employed population exceeding 100,000 workers, although three occupations with totals between 90,000 and 100,000 were included: 兽医 ('veterinarian', 98,000), 翻译 ('translator', 92,000), and 急救员 ('paramedic', 95,000), given their proximity to the threshold and their status as prototypical and widely recognized occupational categories.

4.2 Stereotyping Classification

To operationalize occupational gender stereotyping, we classified the 45 occupations selected into three groups based on the proportion of female workers (*female%*) as reported in the BLS dataset:

- Male-stereotyped (M): *female%* $< 40\%$ (e.g., 木匠 'carpenter', 4.2%; 总裁 'CEO', 33.0%)
- Neutral (N): $40\% \leq$ *female%* $\leq 60\%$ (e.g., 教授 'professor', 49.9%; 调酒师 'bartender', 54.3%)
- Female-stereotyped (F): *female%* $> 60\%$ (e.g., 护士 'nurse', 86.8%; 秘书 'secretary', 91.4%)

Each category contains 15 occupation nouns, yielding a balanced three-way partition of 15 male-stereotyped, 15 neutral, and 15 female-stereotyped occupation nouns. The thresholds of 40% and 60% were determined empirically to yield groups of equal size, ensuring comparability across experimental conditions. A narrower neutral band (e.g., 45–55%) would have produced an uneven distribution across categories, undermining the factorial balance of the design. The ± 10 percentage point margin around parity additionally avoids misclassifying near-parity occupations as stereotyped, a concern that arises when a strict 50% binary threshold is applied (Gkovedarou, 2025).

4.3 Sentence Design

To isolate the effect of occupational gender stereotyping from confounding lexical variables, all 495 sentences are constructed using a fixed set of predicate structures held constant across all 45 occupation nouns (see Table 1). Since the predicate is identical, any observed variation in translation outputs can be attributed to the occupation noun–gender cue interaction rather than to verb-specific gender associations. Each occupation noun is embedded in a matrix of 11 sentence types, described in the following paragraphs.

Base Condition. The base sentence (`base_null`) contains the occupation noun with no gendered pronouns or explicit gender markers, establishing a measure of the model's default gender assignment in the absence of any linguistic gender cue.

Explicit Gender Cue. In these two sentence types (`exp_m` and `exp_f`), the occupation noun is preceded by an explicit gender modifier (男/女, i.e., *male*/*female*), providing the model with maximal lexical evidence of the gender of the referent. These sentences test whether overt lexical gender marking is sufficient to override the model's stereotype prior—including in cases where the explicit gender modifier contradicts the gender stereotypically associated with the occupation noun (e.g., 男护士, *'male nurse'*, given that this profession is strongly female-stereotyped).

Coreference Conditions. The remaining eight sentence types are generated by orthogonally crossing three binary variables, yielding eight distinct conditions involving pronouns holding a coreference relation with the target noun. Each sentence type is assigned a semantic tag encoding all three dimensions simultaneously (e.g., `ante_cpx_f`):

The first factor is linear order. In sentences identified with `ante`, the occupation noun precedes the personal pronoun. In sentences identified with `post`, the personal pronoun precedes the occupation noun, requiring the model to resolve the reference of the pronoun retrospectively once it finds the occupation noun.

The second factor distinguishes simple and complex sentences. Simple sentences (`smp`) consist of a single clause with coreference being established within the same syntactic domain. Complex sentences (`cpx`) involve two clause structures that increase the syntactic distance between the occupation noun and the coreferential pronoun, placing

Tag	Chinese Sentence	English Gloss
base_null	我最近新认识了一名{occupation noun}。	I recently got acquainted with a(n) {occupation noun}.
exp_m	我最近新认识了一名男{occupation noun}。	I recently got acquainted with a **male** {occupation noun}.
exp_f	我最近新认识了一名女{occupation noun}。	I recently got acquainted with a **female** {occupation noun}.
ante_smp_m	这名{occupation noun}很满意他自己的工作环境。	This {occupation noun} is very satisfied with **his** work environment.
ante_smp_f	这名{occupation noun}很满意她自己的工作环境。	This {occupation noun} is very satisfied with **her** work environment.
ante_cpx_m	我最近新认识了一名{occupation noun}，他非常细心。	I recently got acquainted with a(n) {occupation noun}, and **he** is very meticulous.
ante_cpx_f	我最近新认识了一名{occupation noun}，她非常细心。	I recently got acquainted with a(n) {occupation noun}, and **she** is very meticulous.
post_smp_m	他是一名细心的{occupation noun}。	**He** is a meticulous {occupation noun}.
post_smp_f	她是一名细心的{occupation noun}。	**She** is a meticulous {occupation noun}.
post_cpx_m	因为他表现得非常细心，所以大家才如此信任这名{occupation noun}。	Because **he** demonstrates such meticulousness, everyone trusts this {occupation noun} so much.
post_cpx_f	因为她表现得非常细心，所以大家才如此信任这名{occupation noun}。	Because **she** demonstrates such meticulousness, everyone trusts this {occupation noun} so much.

Table 1: The eleven sentence templates used in the challenge set. Gender-bearing elements (modifiers and pronouns) are shown in **bold** (Latin script) and underlined (Chinese script). {occupation noun} denotes the placeholder replaced by each of the 45 occupation nouns.

greater demand on the coreference resolution capacity of the model.

The third factor identifies the gender of the personal pronoun as either masculine (m) (他, *'he'*) or feminine (f) (她, *'she'*).

In total, there are 11 sentences for each occupation noun following the aforementioned templates, yielding a corpus of $45 \times 11 = 495$ source sentences. The complete sentence matrix is illustrated in Table 1. All 495 Chinese source sentences were reviewed for naturalness and grammaticality by the first and third authors, both native speakers of Chinese.

5 Results

All results are reported as gender accuracy percentages across 45 sentences per condition per system (N = 45), with masculine (M) and feminine (F) outputs evaluated separately against the gender value expected given each source sentence condition. Two 2gen outputs were identified in the dataset: instances in which a system produced a form with ambiguous or inclusive gender marking, for example, *um(a) tripulante* ('a crew member'), *este/a empregado/a* ('this employee'). These cases are reported in footnotes where they occur.

Tables 2, 3, and 4 report gender output distributions, masculine–feminine asymmetry, and inferential statistics, respectively.

5.1 Base Condition

Under the base_null condition, source sentences did not contain gender information of any kind. Therefore, gender assignment in the Portuguese output reflects each system's default gender associated with each noun in our dataset.

All eight systems exhibited a strong masculine default bias (Table 2). Masculine outputs ranged from 77.8% (Ministral) to 93.3% (DeepSeek), with a cross-system mean of 86.1%. Feminine outputs were consistently marginal, never exceeding 22.2% for any system. This experimental condition establishes the base results against which the effect of explicit and implicit gender cues in subsequent conditions is assessed.

5.2 Explicit Gender Cues

When prenominal gender modifiers 男/女 ('male/female') immediately preceding the occupation noun were used, all eight systems achieved 100.0% accuracy on both target genders, including anti-stereotypical cases (e.g., 女工程师, *'female engineer'*), indicating that overt lexical gender

Model	base_null	ante_smp		ante_cpx		post_smp		post_cpx	
	M%	M Acc.%	F Acc.%	M Acc.%	F Acc.%	M Acc.%	F Acc.%	M Acc.%	F Acc.%
Google Translate	82.2	100.0	100.0	77.8	28.9	100.0	100.0	100.0	80.0
DeepL Translator	86.7	100.0	100.0	88.9	24.4	100.0	100.0	100.0	97.8
GPT-5.2	82.2[†]	100.0	100.0	95.6	100.0	100.0	100.0	97.8	100.0
Claude Sonnet 4.6	88.9	100.0	100.0	97.8	100.0	100.0	100.0	100.0	100.0
Ministral-14B	77.8	82.2[‡]	82.2	80.0	91.1	100.0	100.0	91.1	48.9
Qwen3.5-397B	91.1	100.0	100.0	97.8	100.0	100.0	100.0	100.0	100.0
Llama3.3-70B	86.7	100.0	88.9	86.7	62.2	100.0	100.0	97.8	28.9
DeepSeek V3.2	93.3	93.3	64.4	97.8	97.8	100.0	100.0	93.3	95.6
Mean	**86.1**	**97.0**	**92.0**	**90.3**	**75.5**	**100.0**	**100.0**	**97.5**	**81.4**

Table 2: Gender output distribution under the base (base_null) and all other conditions, by system (N=45 per condition per system). M% (base_null) = proportion of masculine outputs in sentences without any gender information. M/F Acc.% = proportion of outputs in which the occupation noun phrases are correctly rendered in masculine/feminine form under the respective linguistic conditions. All explicit gender cue conditions (exp_m and exp_f) yielded 100.0% accuracy across all eight systems and are, thus, omitted from the table.
[†]GPT produced one 2gen output under base_null, counted as a non-masculine response. [‡]Ministral produced one 2gen output under ante_smp_m, counted as incorrect for M Acc.%.

expression constitutes a fully sufficient cue that no system overrides based on the gender-stereotype bias of the model.

5.3 Coreference Conditions

Results pertaining to gender information rendered by coreference relations are examined along each of the three factors tested in our data: pronoun position (occurring before or after the occupation noun), syntactic complexity (simple and complex sentences), and gender of the target noun (masculine and feminine).

Position. Conditions in which the gendered pronoun precedes the target occupation noun (post) consistently outperform their counterparts in which the target noun precedes the pronoun (ante) across both complexity levels: under simple sentences, post achieves 100.0% for both M and F, while ante reaches 97.0% (M) and 92.0% (F); under complex sentences, post also shows higher accuracy than ante for both genders (M: 97.5% vs. 90.3%; F: 81.4% vs. 75.5%). This advantage is consistent with left-to-right token processing: when the gender-carrying pronoun occurs before the gender unmarked noun in the source sentence, it is encoded into the contextual representation before the occupation noun is reached, providing a stronger prior for gender selection and reducing the window in which occupational stereotype defaults can take effect.

Complexity. Syntactic complexity produces the most pronounced performance drop in the conditions included in the dataset. In simple sentence structures, six of the eight models achieve perfect or near-perfect performance regardless of target gender or pronoun position. However, in complex sentence structures, performance degrades substantially, with the greatest variation observed when a feminine pronoun occurs after the target noun (ante_cpx_f): Google Translate falls to 28.9% and DeepL to 24.4%, meaning that in more than 70% of cases these systems fail to render the feminine gender information conveyed by the pronoun in the source sentence. By contrast, GPT, Claude, and Qwen maintain accuracy at or above 97.8% across all complex conditions, demonstrating robust coreference resolution under increased syntactic distance and with lower sensitivity to linear order.

Target gender. Under simple conditions, the gap between masculine and feminine accuracy is modest: 5.0 percentage points in ante conditions (occupation noun precedes pronoun) and no difference in post conditions (pronoun precedes occupation noun). Under complex conditions, this gap widens notably: 14.8 and 16.1 percentage points for ante and post, respectively. In other words, complex sentence structures do not merely reduce accuracy overall; the reduction is greater for feminine cues than for masculine cues.

This pattern suggests that gender asymmetry is a latent tendency among the eight models tested, that is amplified when syntactic complexity increases: under simple conditions, systems handle both genders adequately; under complex conditions, they retreat towards masculine default.

Model	Δ_{aSmp}	Δ_{aCpx}	Δ_{pSmp}	Δ_{pCpx}	$\overline{\Delta}$
Google Translate	+0.0	+48.9	+0.0	+20.0	+17.2
DeepL Translator	+0.0	+64.5	+0.0	+2.2	+16.7
GPT-5.2	+0.0	−4.4	+0.0	−2.2	−1.7
Claude Sonnet 4.6	+0.0	−2.2	+0.0	+0.0	−0.6
Ministral-14B	+0.0	−11.1	+0.0	+42.2	+7.8
Qwen3.5-397B	+0.0	−2.2	+0.0	+0.0	−0.6
Llama3.3-70B	+11.1	+24.5	+0.0	+68.9	+26.1
DeepSeek V3.2	+28.9	+0.0	+0.0	−2.3	+6.7
Mean	+5.0	+14.8	+0.0	+16.1	+9.0

Table 3: Masculine–feminine accuracy asymmetry (Δ = M Acc.% − F Acc.%, in percentage points) by system and coreference condition. Positive values indicate superior masculine accuracy; negative values indicate superior feminine accuracy. Δ_{aSmp} = `ante_smp`; Δ_{aCpx} = `ante_cpx`; Δ_{pSmp} = `post_smp`; Δ_{pCpx} = `post_cpx`; $\overline{\Delta}$ = mean across the four conditions.

Predictor	β	OR	95% CI	
Ref.: M gender, `smp`, `ante`, NMT, neutral				
Target gender: F	−1.421	0.24	[0.17, 0.33]	***
Complexity: `cpx`	−1.840	0.16	[0.11, 0.23]	***
Position: `post`	+0.927	2.53	[1.87, 3.41]	***
System type: LLM	+0.775	2.17	[1.61, 2.92]	***
Stereotype: M	−0.652	0.52	[0.35, 0.78]	**
Stereotype: F	−0.245	0.78	[0.51, 1.19]	
N=2,880 AIC=1376.9 σ^2_{occ}=0.0758				

Table 4: Mixed-effects logistic regression predicting correct gender accuracy in coreference conditions (N=2,880; occupation included as random intercept). β = log-odds coefficient; OR = odds ratio; 95% CI = confidence interval for OR. Reference category: masculine target gender, `smp`, `ante`, NMT system, gender-neutral stereotype. ** $p < .01$; *** $p < .001$.

5.4 M–F Asymmetry Across Conditions

Table 3 quantifies the masculine–feminine accuracy gap per system and condition, complementing the aggregate patterns discussed above with system-level detail. The cross-system mean confirms the complexity-driven amplification identified in Section 5.3: mean Δ remains substantially lower under simple conditions (Δ_{pSmp}=+0.0, Δ_{aSmp}=+5.0) than under complex ones (Δ_{aCpx}=+14.8, Δ_{pCpx}=+16.1).

At the system level, Google Translate (Δ_{aCpx}=+48.9) and DeepL (Δ_{aCpx}=+64.5) show the largest masculine advantage under `ante_cpx` condition, while Llama records the highest overall asymmetry ($\overline{\Delta}$=+26.1). By contrast, GPT, Claude, and Qwen show marginally negative $\overline{\Delta}$ values, indicating effectively symmetric gender cue processing across conditions.

5.5 Predictors of Gender Accuracy

To assess whether the observed patterns reflect statistically stable effects, we fitted a mixed-effects logistic regression model predicting gender accuracy across all coreference conditions (base condition excluded as gender-indeterminate; explicit conditions excluded due to 100% accuracy with zero variance; N=2,880). Five predictors were entered as fixed effects: target gender, syntactic complexity, occupation noun position, system type, and occupational gender stereotype, with occupation as a random intercept to account for dependence among observations sharing the same occupation noun across conditions. Results are reported as odds ratios (OR) with 95% confidence intervals (Table 4).

All predictors except F-stereotyped occupation reached statistical significance: target gender, syntactic complexity, occupation noun position, and system type at $p < .001$, and M-stereotyped occupation at $p < .01$. Syntactic complexity produces the largest effect (OR=0.16, CI [0.11, 0.23]): complex sentences reduce the odds of accuracy to just 16% of those under simple conditions, confirming it as the strongest individual predictor of translation failure.

Target gender is the second strongest predictor (OR=0.24, CI [0.17, 0.33]): feminine-cued sentences are statistically harder to resolve correctly than masculine-cued ones across all models and conditions, providing inferential confirmation of the asymmetry documented in the previous section.

Conditions in which the pronoun precedes the occupation noun (`post`) show higher accuracy than the reverse order (`ante`) (OR=2.53, CI [1.87, 3.41]). And LLMs outperform NMT systems (OR=2.17, CI [1.61, 2.92]).

Among stereotype categories, M-stereotyped occupations show significantly lower accuracy than neutral ones (OR=0.52, CI [0.35, 0.78]), indicating that masculine occupational stereotypes influence gender assignment even when the source contains a gendered pronoun; F-stereotyped occupations do not differ significantly from neutral ones ($p = .253$). Finally, a supplementary test of the gender × complexity interaction did not reach significance (χ^2=1.508, $p = .219$), confirming that the two factors contribute independently to translation difficulty without compounding each other.

6 Final Remarks

This study introduced a Chinese challenge set for evaluating gender bias in Chinese–Portuguese automated translation, comprising 495 source sen-

tences generated from 45 occupation nouns across 11 sentence templates and evaluated across two NMT systems and six LLMs. The results address the three objectives guiding this work.

Regarding the second goal of this research, the data reveal a hierarchy of gender cue effectiveness in Chinese source sentences. Explicit prenominal gender modifiers (男/女, 'male/female') constitute the most reliable cue type, achieving 100.0% accuracy across all systems. Simple coreferential pronouns (他/她, 'he/she') are generally sufficient for most systems, though three systems showed localized failures under anaphoric conditions. Complex sentence conditions proved the most demanding, with feminine cue accuracy falling as low as 24.4% for DeepL (`ante_cpx_f`).

In this way, different linguistic conditions produce systematically different gender bias results across systems, with target gender, syntactic complexity, and pronoun position each emerging as significant independent predictors of correct gender accuracy in the mixed-effects model ($p < .001$).

Regarding our third objective, the proposed methodology produces a valid evaluation instrument: it successfully differentiates both between systems and between conditions, with masculine default rates under gender-indeterminate conditions (77.8–93.3%) consistent with prior literature (Stanovsky et al., 2019; Prates et al., 2020), and complex sentence conditions producing the maximal gap between systems observed in the study.

A notable finding concerns the divergence between NMT systems and LLMs. Google Translate and DeepL exhibited pronounced masculine bias under complex conditions, defaulting to masculine output even when an explicit feminine pronoun was present in the source. By contrast, GPT, Claude, and Qwen maintained near-symmetric accuracy across all coreference conditions, suggesting that these systems are better equipped to resolve long-distance coreference. Whether this advantage is attributable to model scale, instruction tuning, or training data composition remains an open question.

Limitations and Future Work. This study operates within a binary gender framework reflecting standard European Portuguese morphology; gender-inclusive strategies fall outside the scope of this paper. Despite the use of available data for Portuguese, occupational stereotype classifications derive from U.S. statistics, and may not reflect the gender–occupation associations in Chinese and Portuguese-speaking contexts. Annotation is restricted to the noun phrase in which the target noun occurs and does not capture broader agreement or overall translation quality, including target-variety compliance.

Future work should extend gender annotation to sentence-level agreement and evaluate overall translation quality. We will also expand the occupation list and sentence templates, incorporate naturally occurring sentences to complement the controlled template-based design, and develop finer-grained measures of syntactic complexity.

Bias Statement. This paper addresses representational harm arising from gender stereotyping in Chinese–Portuguese AT. When AT systems default to masculine forms for gender-indeterminate occupation nouns or fail to transmit gender cues in the source, they introduce gendered interpretations that were absent from the source, potentially reinforcing gender stereotypes and rendering women and non-binary individuals less visible in translated professional discourse. This harm is compounded in the Chinese–Portuguese direction specifically, where masculine defaults introduce bias that was absent from the gender-neutral source. We acknowledge that operating within a binary gender framework is itself a normative limitation; extending gender bias evaluation beyond the masculine/feminine binary remains an important direction for future work.

Acknowledgements

We acknowledge the support of Fundação para a Ciência e a Tecnologia (UID/00214: Centro de Linguística da Universidade de Lisboa).

Carbon Impact Statement

This study did not involve training any machine learning models. All automated translation outputs were obtained via API calls to eight externally hosted systems (Google Translate, DeepL, and six LLMs accessed through OpenRouter), comprising a total of 3,960 inference queries. As the computational infrastructure is operated by third-party providers, precise energy consumption figures are not available to the authors. The overall computational footprint of this study is estimated to be minimal relative to model training workloads.

References

Blodgett, Su Lin, Solon Barocas, Hal Daumé Iii, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5454–5476.

Bonifácio, Sofia, Helena Moniz, and Luisa Coheur. 2025. Diving into gender translation bias for the portuguese language. In *28th European Conference on Artificial Intelligence, ECAI 2025, including 14th Conference on Prestigious Applications of Intelligent Systems, PAIS 2025*, pages 1067–1074. IOS Press BV.

Chao, Lidia S, Derek F Wong, Chi Hong Ao, and Ana Luısa Leal. 2018. Um-pcorpus: a large portuguese-chinese parallel corpus. In *Proceedings of the LREC 2018 Workshop "Belt & Road: Language Resources and Evaluation*, pages 38–43.

Cho, Won Ik, 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.

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

Costa-Jussà, Marta R, Carlos Escolano, Christine Basta, Javier Ferrando, Roser Batlle, and Ksenia Kharitonova. 2022. Interpreting gender bias in neural machine translation: Multilingual architecture matters. In *Proceedings of the aaai conference on artificial intelligence*, volume 36, pages 11855–11863.

Fleisig, Eve and Christiane Fellbaum. 2022. Mitigating gender bias in machine translation through adversarial learning. *arXiv preprint arXiv:2203.10675*.

Font, Joel Escudé and Marta R Costa-Jussà. 2019. Equalizing gender bias in neural machine translation with word embeddings techniques. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 147–154.

Gkovedarou, Eleni, Joke Daems, and Luna De Bruyne. 2025. Gender bias in English-to-Greek machine translation. In Hackenbuchner, Janiça, Luisa Bentivogli, Joke Daems, Chiara Manna, Beatrice Savoldi, and Eva Vanmassenhove, editors, *Proceedings of the 3rd Workshop on Gender-Inclusive Translation Technologies (GITT 2025)*, pages 17–45, Geneva, Switzerland, June. European Association for Machine Translation.

Hackenbuchner, Janiça, Joke Daems, Arda Tezcan, and Aaron Maladry. 2024. You shall know a word's gender by the company it keeps: Comparing the role of context in human gender assumptions with MT. In Savoldi, Beatrice, Janiça Hackenbuchner, Luisa Bentivogli, Joke Daems, Eva Vanmassenhove, and Jasmijn Bastings, editors, *Proceedings of the 2nd International Workshop on Gender-Inclusive Translation Technologies*, pages 31–41, Sheffield, United Kingdom, June. European Association for Machine Translation (EAMT).

Kaneko, Masahiro, Danushka Bollegala, Naoaki Okazaki, and Timothy Baldwin. 2024. Evaluating gender bias in large language models via chain-of-thought prompting. *arXiv preprint arXiv:2401.15585*.

Kibort, Anna and Greville G Corbett. 2008. Grammatical features inventory. *Typology of grammatical features*.

Levy, Shahar, Koren Lazar, and Gabriel Stanovsky. 2021. Collecting a large-scale gender bias dataset for coreference resolution and machine translation. In *Findings of the association for computational linguistics: EMNLP 2021*, pages 2470–2480.

Piergentili, Andrea, Beatrice Savoldi, Dennis Fucci, Matteo Negri, and Luisa Bentivogli. 2023. 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.

Popović, Maja and Ekaterina Lapshinova-Koltunski. 2024. Gender and bias in amazon review translations: by humans, mt systems and chatgpt. In *Proceedings of the 2nd International Workshop on Gender-Inclusive Translation Technologies*, pages 22–30.

Prates, Marcelo OR, Pedro H Avelar, and Luís C Lamb. 2020. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, 32(10):6363–6381.

Quirk, Randolph and David Crystal. 2010. *A comprehensive grammar of the English language*. Pearson Education India.

Raposo, Eduardo Buzaglo, Maria Fernanda Nascimento, Maria Antónia Mota, Luísa Segura, and Amália Mendes. 2013. *Gramática do Português*. Fundação Calouste Gulbenkian, Lisboa.

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

Saunders, Danielle and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics*, pages 7724–7736.

Saunders, Danielle and Katrina Olsen. 2023. Gender, names and other mysteries: Towards the ambiguous for gender-inclusive translation. In *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 85–93.

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

Savoldi, Beatrice, Sara Papi, Matteo Negri, Ana Guerberof-Arenas, and Luisa Bentivogli. 2024. 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.

Sewunetie, Walelign, Atnafu Tonja, Tadesse Belay, Hellina Hailu Nigatu, Gashaw Gebremeskel, Zewdie Mossie, Hussien Seid, and Seid Yimam. 2024. Gender bias evaluation in machine translation for Amharic, Tigrigna, and afaan oromoo. In Savoldi, Beatrice, Janiça Hackenbuchner, Luisa Bentivogli, Joke Daems, Eva Vanmassenhove, and Jasmijn Bastings, editors, *Proceedings of the 2nd International Workshop on Gender-Inclusive Translation Technologies*, pages 1–11, Sheffield, United Kingdom, June. European Association for Machine Translation (EAMT).

Sharma, Shanya, Manan Dey, and Koustuv Sinha. 2022. How sensitive are translation systems to extra contexts? mitigating gender bias in neural machine translation models through relevant contexts. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1968–1984.

Stanovsky, Gabriel, Noah A Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1679–1684.

Zhao, Jieyu, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2979–2989.

Appendix A. List of occupation nouns in the dataset.

Table 5 lists the 45 occupation nouns included in the challenge set, ordered by female percentage as reported in the U.S. Bureau of Labor Statistics.

#	Chinese	English	Total (thousands)	Female%	Label
1	木匠	Carpenters	1,279	4.2	M
2	机械工	Machinists	281	4.8	M
3	消防员	Firefighters	298	6.2	M
4	飞行员	Aircraft pilots and flight engineers	213	9.6	M
5	出租车司机	Taxi drivers	514	12.6	M
6	警察	Police officers	678	14.2	M
7	程序员	Computer programmers	334	17.8	M
8	主厨	Chefs and head cooks	515	21.6	M
9	急救员	Paramedics	95	25.3	M
10	洗碗工	Dishwashers	277	26.9	M
11	建筑师	Architects	209	27.4	M
12	总裁	Chief executives	1,728	33.0	M
13	公交车司机	Bus drivers, transit and intercity	271	35.8	M
14	清洁工	Janitors and building cleaners	2,304	35.9	M
15	牙医	Dentists	169	38.8	M
16	律师	Lawyers	1,146	42.0	N
17	厨师	Cooks	2,005	42.5	N
18	摄影师	Photographers	210	43.5	N
19	销售员	Retail salespersons	2,734	47.5	N
20	教练	Coaches and scouts	340	48.5	N
21	教授	Postsecondary teachers	1,072	49.9	N
22	艺术家	Artists	318	53.8	N
23	医学研究员	Medical scientists	127	54.0	N
24	调酒师	Bartenders	450	54.3	N
25	生物学家	Biological scientists	105	54.7	N
26	电子装配工	Electrical and electronics assemblers	115	54.9	N
27	房地产经纪人	Real estate brokers and sales agents	1,016	56.7	N
28	会计	Accountants	1,778	57.2	N
29	审计	Auditors	1,778	57.2	N
30	药剂师	Pharmacists	326	59.7	N
31	作家	Writers and authors	257	64.7	F
32	面包师	Bakers	243	67.1	F
33	服务员	Waiters and waitresses	1,786	67.8	F
34	兽医	Veterinarians	98	68.6	F
35	收银员	Cashiers	2,569	68.9	F
36	编辑	Editors	132	72.6	F
37	心理医生	Psychologists	215	74.0	F
38	翻译	Interpreters and translators	92	74.8	F
39	空乘	Flight attendants	129	77.2	F
40	老师	Teachers	3,632	77.7	F
41	护士	Registered nurses	3,571	86.8	F
42	家政	Maids and housekeeping cleaners	1,413	87.7	F
43	助理	Assistants	1,758	91.4	F
44	秘书	Secretaries and administrative assistants	1,758	91.4	F
45	营养师	Dietitians and nutritionists	142	92.1	F

Table 5: Occupation nouns included in the Chinese Challenge Set, ordered by the percentage of female professionals (BLS 2024 Current Population Survey). Total employment figures are in thousands. M = female% $< 40\%$; N = $40\% \leq$ female% $\leq 60\%$; F = female% $> 60\%$. Each stereotype category contains exactly 15 occupation nouns.

WinoTR: Evaluating Gender Bias in Machine Translation from a Gender-Neutral Language Using Causal Inference

Deniz Albayrak
University of Hamburg (Graduate)
Hamburg, Germany
deniz.albayrak91@gmail.com

Abstract

We present WinoTR, a Turkish adaptation of the WinoMT challenge dataset (Stanovsky et al., 2019). While WinoMT has been widely studied across multiple languages, its adaptation to Turkish — a morphologically rich language with no grammatical gender — and its analysis through a causal lens remain unexplored. Using 4,752 sentences adapted from the original dataset across pro-stereotypical, anti stereotypical and neutral conditions, we apply Double Machine Learning (DML) to estimate the causal effect of gender cues on stereotype-consistent translation output. Our results reveal a striking asymmetry: cue direction has a large and statistically significant effect on translation outcomes, while cue presence alone produces virtually no effect. Even without any gender signal, MT systems default to stereoype-consistent translations in 62.9% of cases across three systems (DeepL, Google Translate, OpenAI). By leveraging this typological property, our causal analysis reveals that gender bias in contemporary MT and LLM-based translation systems runs deeper than surface-level cue processing, persisting as an embedded prior independent of any explicit gender signal in the input.

1 Introduction

Gender bias in machine translation (MT) is a well-documented yet poorly understood phenomenon. When translating from a gender-neutral source language into a grammatically gendered target, MT systems must resolve a binary gender assignment that the source leaves unspecified — and prior work has shown these decisions are systematically skewed by occupational stereotypes (Stanovsky et al., 2019; Bolukbasi et al., 2016; Prates et al., 2020). However, most existing studies treat bias as a correlational phenomenon, focusing on measuring and documenting disparities rather than estimating the magnitude of causal effects (Blodgett et al., 2020; Feder et al., 2022).

Turkish presents a uniquely informative testbed for this question. As a morphologically rich, agglutinative language with a gender-neutral third-person pronoun (*O*), Turkish encodes no grammatical gender in the source text. Any stereotype-consistent output must therefore reflect the MT system's own internalized priors — not grammatical inheritance from the source. Prior work confirms that bias in Turkish NLP stems from cultural norms and data imbalances rather than grammatical structure (Ciora et al., 2021; Altinok, 2024; Köksal et al., 2023). Despite this typological advantage, no prior work has applied a causal framework to MT bias evaluation from a gender-neutral source language.

This paper addresses these gaps through three research questions aligned with Pearl's causal hierarchy (Pearl and Mackenzie, 2018): **(RQ1)** do MT systems default to stereotypical gender assignments in the absence of explicit cues? **(RQ2)** does cue direction (pro- vs. anti-stereotypical) causally affect translation output? **(RQ3)** does cue presence versus absence produce a measurable causal effect? To answer these questions, we apply Double Machine Learning (Chernozhukov et al., 2018; Imbens and Rubin, 2015), which isolates causal ef-

fects of gender signals while controlling for high-dimensional covariates. We introduce **WinoTR**, a Turkish adaptation of the WinoMT benchmark (Stanovsky et al., 2019), consisting of 4,752 sentences across pro-stereotypical, anti-stereotypical, and neutral conditions, translated into German and Spanish using three MT systems: DeepL, Google Translate, and OpenAI.

2 Related Work

Gender bias in machine translation has been systematically documented since Stanovsky et al. (2019) introduced WinoMT, a benchmark combining WinoGender (Rudinger et al., 2018) and WinoBias (Zhao et al., 2018) to evaluate gender bias across eight target languages. Subsequent work extended this evaluation to additional language pairs and morphologically rich target languages (Kocmi et al., 2020; Vanmassenhove and Monti, 2021; Ramesh et al., 2021), and to instruction-tuned LLMs (Attanasio et al., 2023). Mitigation efforts have focused on debiasing word embeddings (Bolukbasi et al., 2016), data augmentation (Sun et al., 2019), and constrained decoding. However, these approaches remain largely correlational — they document disparities without identifying the underlying mechanisms (Feder et al., 2022; Blodgett et al., 2020).

Turkish has received growing attention as a typologically distinctive case for bias research. As a grammatically gender-neutral, morphologically agglutinative language, Turkish encodes no gender in third-person pronouns, making it an ideal testbed for isolating model-internal priors from source-language cues. Ciora et al. (Ciora et al., 2021) examined overt and covert gender bias in Turkish–English MT, highlighting the importance of language-specific and culturally informed methodology. Altinok (Altinok, 2024) demonstrated that occupational titles in Turkish word embeddings are predominantly linked to masculine forms, reflecting cultural norms rather than grammatical gender. Köksal et al. (Köksal et al., 2023) applied a language-agnostic bias probing method across six languages including Turkish, finding that bias correlates with historical and cultural contexts rather than grammatical structure alone. These findings motivate a causal rather than correlational analysis of MT gender bias from Turkish.

Causal inference has recently emerged as a principled framework for NLP bias research. Pearl's causal hierarchy (Pearl and Mackenzie, 2018) distinguishes association, intervention, and counterfactual reasoning, enabling researchers to ask not only whether bias exists but why it arises. The potential outcomes framework (Imbens and Rubin, 2015) provides the statistical foundation for treatment effect estimation. Double Machine Learning (Chernozhukov et al., 2018) extends this framework to high-dimensional settings by combining flexible ML nuisance estimation with econometric orthogonalization and cross-fitting, producing asymptotically unbiased treatment effect estimates. Recent work has applied DML and related causal estimators to NLP tasks (Feder et al., 2022; Veljanovski and Wood-Doughty, 2024), but causal estimation of gender bias in MT — particularly from a gender-neutral source language — remains unexplored.

Both Saunders and Olsen (2023) and Manna et al. (2026) adopt an approach that leverages gender-neutral source languages to better expose translation bias toward gendered target languages. Saunders and Olsen demonstrate that cases where the source provides no gender marker yet the target requires a gender choice are far more prevalent than typically assumed — a scenario that mirrors the structural properties of Turkish. Manna et al. introduce Minimal Pair Accuracy (MPA) to measure whether NMT models systematically draw on contextual gender cues in English-Italian translation, finding that models largely revert to stereotypical defaults and integrate male cues more consistently than female ones. Both findings converge with our results: the near-zero ATE in RQ3 supports the view that Turkish source sentences constitute a natural testbed for ambiguous gender resolution, while the asymmetry between pro and anti cue effects in RQ2 echoes the male-female cue asymmetry reported by Manna et al.

3 Dataset: WinoTR

WinoTR is a Turkish adaptation of the WinoMT benchmark (Stanovsky et al., 2019), manually constructed by a native Turkish speaker. The dataset consists of 4,752 sentences across three conditions: **trpro** (pro-stereotypical), **tranti** (anti-stereotypical), and **trneutral** (no gender cue), covering 40 occupations drawn from the original WinoMT profession list (see Table 1).

Each sentence involves two occupations with an implicit coreference structure. The sentences were

manually translated and adapted from the original English WinoMT sentences by a native Turkish speaker, preserving the syntactic profession–pronoun relationship while ensuring Turkish semantic integrity. Since Turkish uses the gender-neutral third-person pronoun *o* for all references, no grammatical gender is encoded in the source text. To create the pro and anti conditions, explicit lexical gender markers — *kadın* (woman) and *adam* (man) — were introduced in the subject position of subordinate clauses. In the **pro** condition, the marker reinforces the societal stereotype (e.g., *adam doktor* / male doctor); in the **anti** condition, it contradicts it (e.g., *kadın doktor* / female doctor). The **neutral** condition contains no marker, relying entirely on the gender-neutral Turkish pronoun. Representative examples from each condition are shown in Table 2.

Subset	Male	Female	Neutral	Total
trpro	792	792	—	1,584
tranti	792	792	—	1,584
trneutral	—	—	1,584	1,584
Total	**1,584**	**1,584**	**1,584**	**4,752**

Table 1: WinoTR corpus structure.

Each source sentence was translated once using three MT systems: Google Cloud Translation API (May 2025), DeepL API (June 2025), and OpenAI GPT-4o (July 2025), into two target languages: German and Spanish. No post-editing was applied to any translation output. Profession tokens were identified using a bilingual Turkish–German–Spanish lexicon. Word alignment was performed using SimAlign (Sabet et al., 2020) and AwesomeAlign (Dou and Neubig, 2021). Gender labels were extracted via morphological analysis using spaCy and deterministic lexicon lookup. After filtering for successfully aligned and gender-labeled instances, the final dataset contains 16,331 rows (alignment coverage: ≈57%). This coverage gap reflects the structural challenges of Turkish morphology: agglutinative suffixation produces surface forms that diverge substantially from canonical lemmas, complicating token-level alignment with German and Spanish (Bisazza and Federico, 2009). Additionally, Turkish flexible word order differs fundamentally from the more rigid structures of German and Spanish, further compounding alignment ambiguity.

4 Methodology

4.1 Causal Framework

We structure our analysis using Pearl's causal hierarchy (Pearl and Mackenzie, 2018), mapping each research question to a distinct level of causal reasoning, as summarized in Table 3.

The outcome variable Y is defined consistently across all three RQs as stereotype-consistency: $Y = 1$ if the translated profession gender matches the predefined stereotype map; $Y = 0$ otherwise. In RQ2, the treatment $D \in \{0{=}\, anti, 1{=}\, pro\}$. In RQ3, $D \in \{0{=}\, neutral, 1{=}\, signal\}$. Pre-treatment covariates X include target profession, target language, MT tool, syntactic position (subject/object), and source sentence length. Features directly affected by cue insertion were excluded to avoid post-treatment bias (Chernozhukov et al., 2018).

While WinoTR's matched-pair design approximates counterfactual reasoning — each profession pair appears under both cued and uncued conditions within the same MT system — RQ3 does not achieve individual-level counterfactual identification, as neutral sentences differ syntactically from their cued counterparts. RQ3 is therefore interpreted as a population-level Average Treatment Effect (ATE) of cue presence, consistent with Pearl's intervention level.

4.2 Double Machine Learning

We apply the Interactive Regression Model (IRM) variant of Double Machine Learning (Chernozhukov et al., 2018), implemented via the DoubleML Python package (Bach et al., 2022). IRM is appropriate here because both the outcome Y and treatment D are binary.

Nuisance models. We use `RandomForestClassifier` (500 trees, `min_samples_leaf`=10) for the outcome model $g(D, X) = E[Y|D, X]$, and `LogisticRegression` for the propensity model $m(X) = E[D|X]$. This combination balances flexibility in outcome estimation with stability in propensity estimation.

Cross-fitting. We apply 5-fold cross-fitting throughout, following chernozhukov2018double. This mitigates overfitting in nuisance estimation given $N > 10{,}000$.

Subset	Gender	Source Sentence (Turkish)	Profession
trpro	male	Geliştirici, tasarımcıyla tartıştı çünkü adam tasarımı beğenmedi.	geliştirici
	female	Geliştirici, tasarımcı ile tartıştı çünkü kadının fikri uygulanamazdı.	tasarımcı
tranti	female	Geliştirici, tasarımcıyla tartıştı çünkü kadın tasarımı beğenmedi.	geliştirici
	male	Geliştirici, tasarımcı ile tartıştı çünkü adamın fikri uygulanamazdı.	tasarımcı
trneutral	—	Geliştirici, tasarımı beğenmedi için tasarımcıyla tartıştı.	geliştirici
	—	Geliştirici, tasarımcı ile tartıştı çünkü tasarımcının fikri uygulanamazdı.	tasarımcı

Table 2: Representative examples from WinoTR. Gender markers (*adam*/*kadın*) are underlined in the pro and anti conditions. The neutral condition contains no explicit gender marker.

Level	Research Question
Association (RQ1)	Do MT systems default to stereotypes without cues?
Intervention (RQ2)	Does cue *direction* (pro vs. anti) affect stereotype-consistent output?
Intervention (RQ3)	Does cue *presence* vs. absence affect output?

Table 3: Research questions mapped to Pearl's causal hierarchy.

Overlap. To verify the overlap assumption, we estimated propensity scores on observed pretreatment covariates. Scores ranged from 0.483 to 0.504 (mean = 0.497), confirming near-perfect balance between conditions — consistent with WinoTR's balanced-by-design structure.

ATE estimation. The Average Treatment Effect is recovered via the Neyman-orthogonal score:

$$ATE = E[Y(1) - Y(0)] \tag{1}$$

Profession-, language-, and tool-level heterogeneity is reported descriptively via accuracy differences across subgroups. Formal Conditional Average Treatment Effect (CATE) estimation is left for future work.

5 Results

5.1 RQ1: Neutral Baseline (Association Level)

Stereotype-consistency is defined as a binary outcome: a translation is stereotype-consistent ($Y = 1$) if the gender of the translated profession matches the predefined stereotype map (e.g., male for *engineer*, female for *nurse*), and stereotype-inconsistent ($Y = 0$) otherwise. Without any explicit gender cue, MT systems produce stereotype-consistent translations in 62.9% of cases overall.

Without any explicit gender cue, MT systems produce stereotype-consistent translations in 62.9% of cases overall. Female-stereotyped professions receive a feminine translation in only 17.3% of neutral sentences, far below the 50% expected under a gender-neutral system.

Tool	N	Female share	Stereotype support
DeepL	1,838	0.180	0.644
Google	1,828	0.173	0.635
OpenAI	1,754	0.165	0.608
Overall	**5,420**	**0.173**	**0.629**

Table 4: RQ1: Stereotype support in neutral condition by MT tool.

Male-stereotyped professions show near-perfect stereotype support (*avukat* 1.00, *tamirci* 0.995, *doktor* 0.995), while several female-stereotyped professions receive no feminine translation in the neutral condition (*editör*, *terzi*, *yazar*: 0.00), suggesting a mismatch between the inherited stereotype map and MT system priors for these occupations.

5.2 RQ2: Cue Direction (Intervention Level)

Tool	Pro acc.	Anti acc.	Difference
DeepL	0.790	0.550	0.240
Google	0.770	0.497	0.273
OpenAI	0.697	0.472	0.225
Overall	**0.754**	**0.507**	**0.247**

Table 5: RQ2: Stereotype-consistency by cue direction and MT tool.

DML estimation yields ATE $= 0.245$ (SE $=$ 0.009, $t = 27.67$, $p < 0.001$, $N =$ 10,911). The pro condition produces stereotype-consistent translations in 75.4% of cases, compared to 50.7% in the anti condition. The effect is consistent across all three MT systems and both target languages (German: $\Delta = 0.254$; Spanish: $\Delta = 0.240$), suggesting this reflects a general property of current MT architectures rather than a system-specific artifact.

5.3 RQ3: Cue Presence (Second Intervention Level)

Tool	Neutral acc.	Signal acc.	Difference
DeepL	0.644	0.671	+0.027
Google	0.635	0.634	−0.001
OpenAI	0.608	0.582	−0.026
Overall	**0.629**	**0.630**	**+0.001**

Table 6: RQ3: Stereotype-consistency by cue presence and MT tool.

DML estimation yields ATE $= +0.005$ (SE $= 0.007$, $t = 0.639$, $p = 0.523$, $N = 16{,}331$), statistically indistinguishable from zero. The near-zero overall effect masks divergent tool-level behaviour: DeepL shows a positive signal–neutral difference ($+0.027$) while OpenAI shows a negative one (-0.026), suggesting that different MT architectures respond differently to cue presence independent of cue direction.

6 Discussion

Our results reveal a striking asymmetry between the two causal questions examined. Cue direction has a large and statistically significant effect on stereotype-consistent output (ATE $= 0.245$, $p < 0.001$), while cue presence alone produces virtually no effect (ATE $= +0.005$, $p = 0.523$). Together with the high stereotype-consistency rate in the neutral condition (62.9%), these findings yield a coherent three-level narrative aligned with Pearl's causal hierarchy (Pearl and Mackenzie, 2018):

At the **association level** (RQ1), MT systems are not gender-neutral by default: male forms dominate as the unmarked baseline, while female forms remain exceptions. This reflects a learned mapping from Turkish professions — which lack grammatical gender — to gendered equivalents in German and Spanish, consistent with prior evidence from WinoMT and related corpora (Stanovsky et al., 2019; Kocmi et al., 2020).

At the **intervention level** (RQ2), cue direction strongly determines whether translations align with stereotypes. Pro cues amplify stereotype-consistency (75.4%), while anti cues suppress it (50.7%). Systems are more accurate when gender cues reinforce expectations than when they challenge them, suggesting that lexical interventions alone are insufficient to counteract entrenched priors (Bolukbasi et al., 2016). The effect is consistent across all three MT systems and both target languages, suggesting it reflects a general property of current MT architectures rather than a system-specific artifact.

At the **second intervention level** (RQ3), the presence of a signal alone is insufficient: neutral and signalled cases perform nearly identically. The negligible overall ATE masks divergent tool-level behaviour: DeepL shows a positive signal–neutral difference ($+0.027$) while OpenAI shows a negative one (-0.026), suggesting different MT architectures encode occupational gender priors differently. This aligns with previous work showing that debiasing attempts at the embedding level often fail to eliminate systematic bias (Zhao et al., 2018; Attanasio et al., 2023).

Representation-level persistence. The near-zero ATE in RQ3 suggests that the learned representations in MT systems exert a strong inductive bias toward stereotype-consistent translations. Explicit lexical cues do not re-anchor the model sufficiently: profession tokens remain strongly associated with culturally dominant gender patterns regardless of the input signal. The application of DML strengthens this interpretation by showing that even after controlling for high-dimensional covariates, the effect of cue presence remains statistically indistinguishable from zero (Chernozhukov et al., 2018; Bach et al., 2022).

Crucially, Turkish provides a unique test case: its lack of grammatical gender means that gendered translations must be inferred entirely from learned social regularities rather than grammatical constraints (Ciora et al., 2021; Altinok, 2024). By leveraging this typological property, our causal analysis reveals that gender bias in contemporary MT and LLM-based translation systems persists as an embedded prior, independent of any explicit gender signal in the input. This suggests that meaningful bias reduction may require addressing model-internal priors at the level of training data or model architecture, rather than surface-level input signals.

7 Conclusion

This paper introduced WinoTR, a Turkish adaptation of the WinoMT benchmark, and applied Double Machine Learning to estimate the causal effect of gender cues on stereotype-consistent MT output. Our findings reveal a clear asymmetry: cue direction produces a large causal effect (ATE $= 0.245$, $p < 0.001$), while cue presence alone

produces none (ATE $= +0.005$, $p = 0.523$). Even without any gender signal, MT systems default to stereotype-consistent translations in 62.9% of cases. By leveraging Turkish as a gender-neutral source language, our causal analysis provides evidence that gender bias in contemporary MT and LLM-based translation systems reflects occupational gender priors embedded in the models themselves, persisting independently of explicit input signals. We compare three MT systems across two target languages, finding consistent patterns for cue direction but divergent tool-level responses to cue presence. These results suggest that meaningful bias reduction may require addressing model-internal priors rather than surface-level input signals.

WinoTR and all analysis code will be made publicly available upon acceptance.

8 Limitations

Deterministic treatment assignment. WinoTR's pro/anti conditions are constructed by design, not randomized. Causal identification relies on conditional independence given observed covariates. Propensity score analysis confirms near-perfect overlap (0.483–0.504), providing empirical support for this assumption.

Alignment coverage. Approximately 43% of translation instances could not be reliably gender-labeled, introducing potential selection bias. This gap reflects structural challenges inherent to Turkish morphology and syntax: agglutinative suffixation produces surface forms that diverge substantially from canonical lemmas, while Turkish flexible word order — which differs fundamentally from the more rigid Subject-Verb-Object structure of German and Spanish — further compounds alignment ambiguity (Bisazza and Federico, 2009). Standard alignment models implicitly assume more fixed syntactic structures, making Turkish a particularly challenging source language for token-level alignment. Six professions showed particularly high alignment failure rates in the neutral condition, resulting in approximately 28% data loss for RQ1.

Alignment fallback. In cases where all alignment methods failed, a masculine default was assigned as a tie-breaker. This conservative choice may slightly inflate male-default rates in the neutral condition.

Stereotype map validity. The profession stereotype map is inherited from WinoMT and reflects English-language occupational associations, which may not fully correspond to Turkish cultural norms. Several professions mapped as female-stereotyped received 0% feminine translations in the neutral condition, suggesting a mismatch between the inherited map and MT system priors.

RQ3 causal level. While WinoTR's matched-pair design approximates counterfactual reasoning, RQ3 does not achieve individual-level counterfactual identification, as neutral sentences differ syntactically from their cued counterparts. RQ3 is therefore interpreted as a second population-level intervention estimate.

Future work. Several directions remain for future research. First, formal Conditional Average Treatment Effect (CATE) estimation using CausalForest or the R-learner would allow profession-level heterogeneity to be quantified with statistical guarantees. Second, the stereotype map could be validated against Turkish cultural norms through native speaker annotation, replacing the inherited English-language associations. Third, following Troles and Schmid (Troles and Schmid, 2021), WinoTR could be extended to include stereotypical adjectives and verbs in addition to occupational nouns, providing a richer probe of gender bias in Turkish MT. Fourth, Turkish-specific gender-marked adjectives and morphological agreement patterns could be incorporated into the dataset to capture more subtle bias signals. Finally, alignment quality could be improved through Turkish-specific tokenization and morphological pre-processing pipelines, potentially recovering a substantial portion of the currently excluded instances.

Ethical Considerations

This work represents an adaptation and extension of WinoMT, a foundational benchmark in gender bias research, to Turkish, deepening the scope of existing work. Gender bias in MT systems becomes particularly pronounced when translating from grammatically gender-neutral source languages such as Turkish into grammatically gendered target languages such as German and Spanish. For this reason, evaluating such bias not only through algorithmic correlations but also through causal methods represents an important step to-

ward both its accurate detection and reduction. We believe that continuing this line of research will make meaningful contributions to addressing structural gender inequalities in MT systems. The WinoTR dataset is released publicly to support future work in this direction.

WinoTR and all analysis code are publicly available at `https://github.com/denizalbyrk/winotr`.

Acknowledgments

The author would like to thank Prof. Dr. Martin Spindler for valuable guidance and feedback during the thesis work that formed the basis of this paper.

References

Altinok, Duygu. 2024. Gender bias in turkish word embeddings: A comprehensive study of syntax, semantics and morphology across domains. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 203–218.

Attanasio, Giuseppe, Debora Nozza, Dirk Hovy, and Eleonora Baranzini. 2023. A tale of pronouns: Interpretability informs gender bias mitigation for fairer instruction-tuned machine translation. In *Proceedings of EMNLP*.

Bach, Philipp, Victor Chernozhukov, Malte S Kurz, and Martin Spindler. 2022. Doubleml: An object-oriented implementation of double machine learning in python. In *Journal of Machine Learning Research*, volume 23, pages 1–6.

Bisazza, Arianna and Marcello Federico. 2009. Morphological pre-processing for turkish to english statistical machine translation. In *Proceedings of IWSLT*.

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.

Bolukbasi, Tolga, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. Double/debiased machine learning for treatment and structural parameters.

Ciora, Chloe, Nur Iren, and Malihe Alikhani. 2021. Examining covert gender bias: A case study in turkish and english machine translation models.

Dou, Zi-Yi and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128.

Feder, Amir, Katherine A Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E Roberts, et al. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.

Imbens, Guido W and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press.

Kocmi, Tom, Tomasz Limisiewicz, and Gabriel Stanovsky. 2020. Gender coreference and bias evaluation at wmt 2020. In *Proceedings of the Fifth Conference on Machine Translation*.

Köksal, Abdullatif, Omer Faruk Yalcin, Ahmet Akbiyik, M Tahir Kilavuz, Anna Korhonen, and Hinrich Schuetze. 2023. Language-agnostic bias detection in language models with bias probing. *arXiv preprint arXiv:2305.13302*.

Manna, Chiara, Hosein Mohebbi, Afra Alishahi, Frédéric Blain, and Eva Vanmassenhove. 2026. Gender disambiguation in machine translation: Diagnostic evaluation in decoder-only architectures. *arXiv preprint arXiv:2603.17952*.

Pearl, Judea and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect*. Basic books.

Prates, Marcelo OR, Pedro H Avelar, and Luís C Lamb. 2020. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, 32(10):6363–6381.

Ramesh, Shanya, Manish Kumar, and Mitesh M Khapra. 2021. Evaluating gender bias in hindi-english machine translation. In *Proceedings of the 1st Workshop on NLP for Positive Impact*.

Rudinger, Rachel, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of NAACL-HLT*.

Sabet, Masoud Jalili, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. Simalign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643.

Saunders, Danielle and Katrina Olsen. 2023. Gender, names and other mysteries: Towards the ambiguous for gender-inclusive translation. In *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 85–93.

Stanovsky, Gabriel, Noah A Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1679–1684.

Sun, Tony, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of ACL*.

Troles, Jonas-Dario and Ute Schmid. 2021. Extending challenge sets to uncover gender bias in machine translation: Impact of stereotypical verbs and adjectives. In Barrault, Loic, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz, editors, *Proceedings of the Sixth Conference on Machine Translation*, pages 531–541, Online, November. Association for Computational Linguistics.

Vanmassenhove, Eva and Johanna Monti. 2021. Neural machine translation: the gender pro and the gender con. In *Proceedings of Recent Advances in Natural Language Processing*.

Veljanovski, Filip and Zach Wood-Doughty. 2024. Doublelingo: Causal estimation with large language models. *arXiv preprint arXiv:2404.04291*.

Zhao, Jieyu, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of NAACL-HLT*.