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

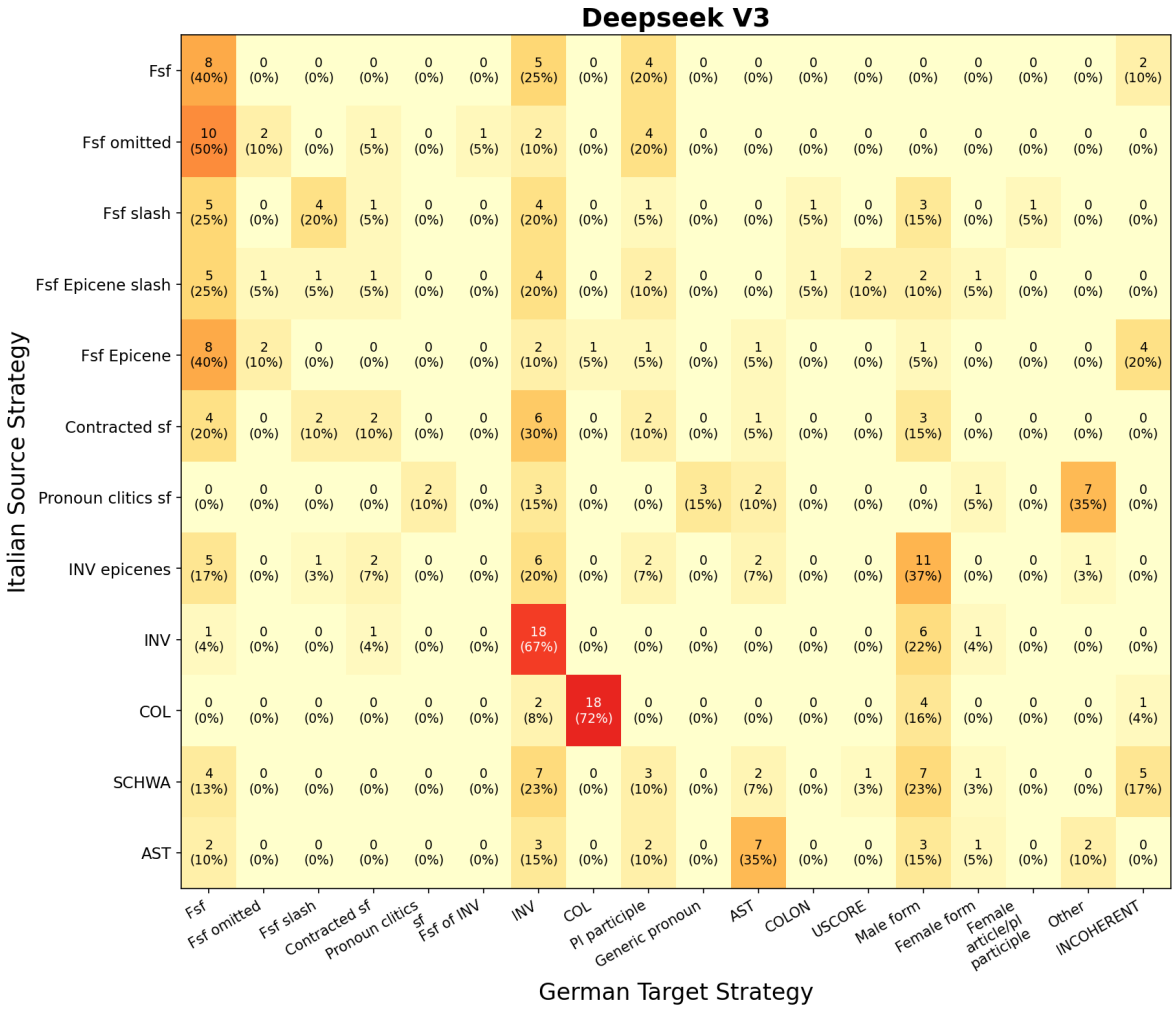

Figure 1: Follows in the next page

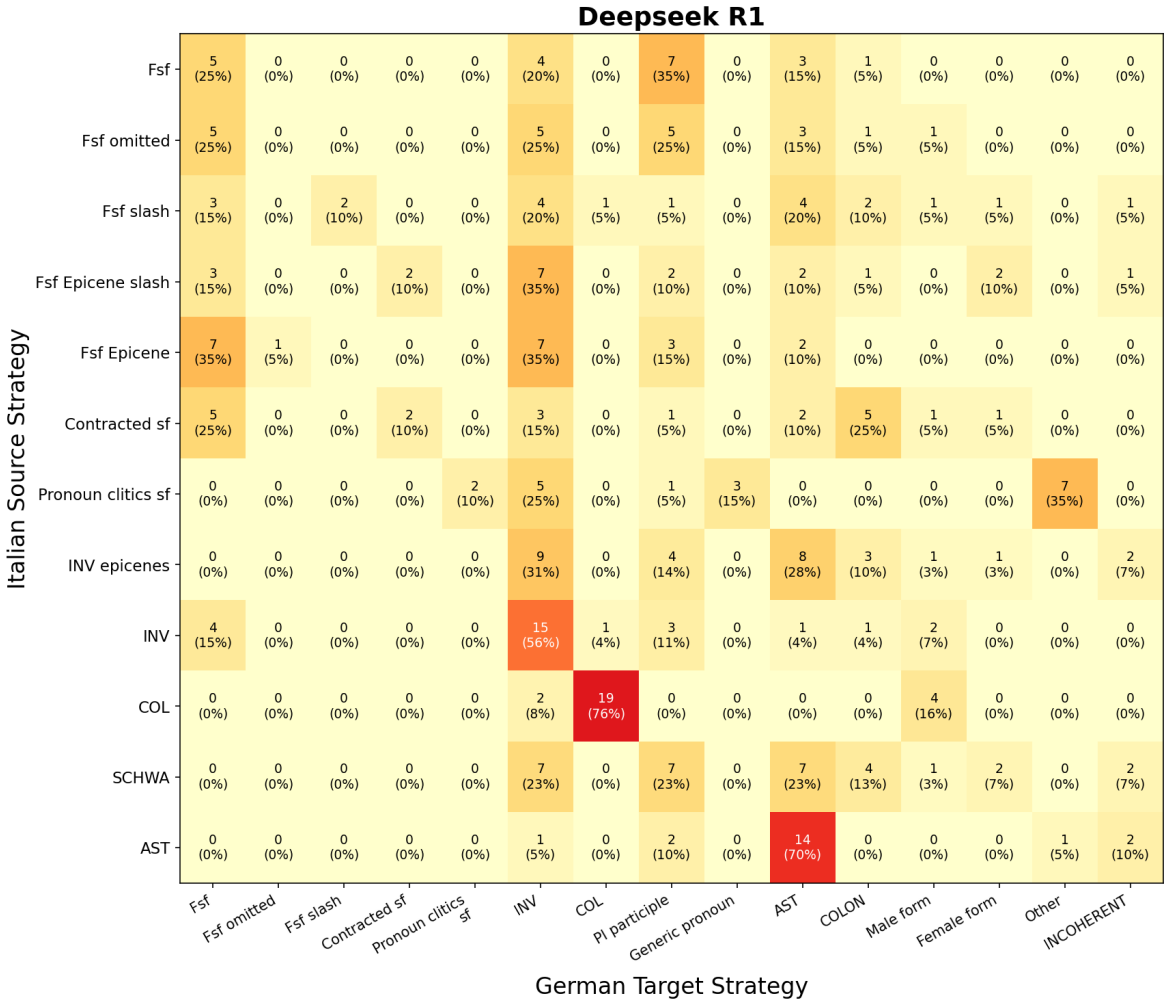

Figure 2: Annotations of the frequency of German strategies adopted to translate each Italian source strategy.

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
}
```