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.