✨ GLITTER: A Multi-Sentence, Multi-Reference Benchmark for Gender-Fair German Machine Translation

A Pranav[*], 🚢, **Janiça Hackenbuchner**[*], 🔶,
Giuseppe Attanasio🐟, **Manuel Lardelli**🍕, **Anne Lauscher**🚢
🚢University of Hamburg, Germany 🔶Ghent University, Belgium
🐟Instituto de Telecomunicações, Lisbon 🍕University of Padua, Italy
{pranav.agrawal, anne.lauscher}@uni-hamburg.de, janica.hackenbuchner@ugent.be
manuel.lardelli@unipd.it, giuseppe.attanasio@lx.it.pt

Abstract

Machine translation (MT) research addressing gender inclusivity has gained attention for promoting non-exclusionary language representing all genders. However, existing resources are limited in size, most often consisting of single sentences, or single gender-fair formulation types, leaving questions about MT models' ability to use context and diverse inclusive forms. We introduce ✨ GLITTER, an English-German benchmark featuring extended passages with professional translations implementing three gender-fair alternatives: neutral rewording, typographical solutions (gender star), and neologistic forms (-ens forms). Our experiments reveal significant limitations in state-of-the-art language models, which default to masculine generics, struggle to interpret explicit gender cues in context, and rarely produce gender-fair translations. Through a systematic prompting analysis designed to elicit fair language, we demonstrate that these limitations stem from models' fundamental misunderstanding of gender phenomena, as they fail to implement inclusive forms even when explicitly instructed. GLITTER establishes a challenging benchmark, advancing research in gender-fair English-German MT. It highlights substantial room for improvement among leading models and can guide the development of future MT models capable of accurately representing gender diversity.

https://github.com/pranav-ust/glitter

https://huggingface.co/datasets/mbvslowride/glitter

1 Introduction

Language and society are deeply interrelated (Montgomery, 2008), with language both reflecting and reinforcing societal norms and power imbalances. This relationship is particularly evident in gender representation (Hord, 2016), where, for example, the (supposedly) generic masculine—a form often used to address all genders—fails to be interpreted inclusively (Gygax et al., 2008). Instead, it significantly influences readers' mental representations of gender across various contexts (Sato et al., 2025; Fatfouta and Sczesny, 2023).

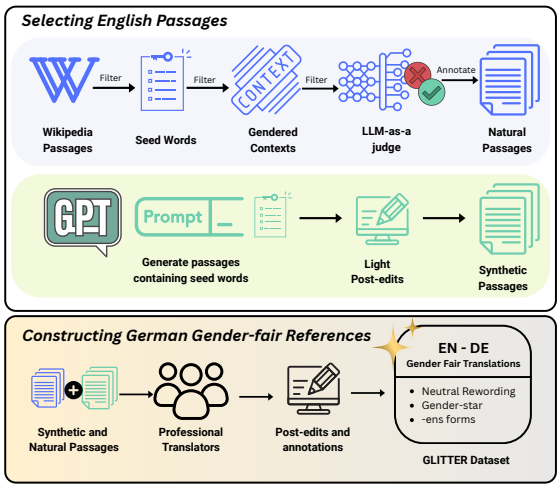

Figure 1: Construction of ✨ GLITTER involves selecting natural and synthetic English passages and post-editing into multi-form German GFL references.

Motivated by growing societal awareness and institutional recognition (EIGE, 2019), recent work in natural language processing (NLP) has focused on developing resources and methods that promote gender-inclusive language technologies (Piergentili et al., 2023; Lardelli et al., 2024; Waldis et al., 2024; Bartl and Leavy, 2024). Machine translation (MT) presents a critical domain for these efforts, as gender-fair forms[1] provide alternatives to masculine generics that otherwise diminish the visibility of women and non-binary individuals in translations (Choubey et al., 2021; Lardelli, 2023). This challenge is particularly pro-

[*]The first two authors contributed equally.

[1]We use "gender-fair" as an umbrella term to encompass both gender-neutral rewording and gender-inclusive forms such as typographical characters and neomorphemes.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 18450–18477

nounced when translating from notional gender languages (e.g., English) into grammatical gender languages requiring extensive gender markers and morphological agreement (e.g., romance languages and German) (Savoldi et al., 2022).

Recent research has introduced benchmark datasets aimed at measuring how effectively current MT systems support translation beyond binary gender representation (e.g., Piergentili et al., 2023; Lardelli et al., 2024; Friðriksdóttir, 2024). However, these resources remain limited in two critical aspects: they typically focus on short sources—often single sentences—and they address only one type of gender-fair formulation (e.g., neutralization through rewording). Despite being a solid starting point, this approach overlooks other inclusive strategies gaining traction among German speakers, such as typographical solutions that explicitly indicate all genders (Waldendorf, 2023). Thus, **fundamental questions remain about the ability of MT models to leverage contextual cues in extended passages and implement diverse, gender-fair forms appropriate to the context**.

To address these limitations, we introduce GLITTER (acronym for Gender Language Inclusivity and Translation Testing Evaluation Resource) a comprehensive English-German translation benchmark featuring multi-sentence passages with professionally post-edited gender-fair translations. This resource introduces two key innovations: (**1**) its focus on extensive context beyond sentence-level evaluation, and (**2**) its provision of multiple professionally curated gender-fair translations resulting in three parallel outputs—gender-neutral rewording, gender star (*), and ens-form translations. We benchmark five open-weight and commercial language models on GLITTER and discover a widespread bias towards masculine forms regardless of contextual cues, and almost absent gender-fair translations. Moreover, we experiment with in-context learning to steer models towards gender-fair forms. However, even with explicit instructions, models struggle to correctly identify contexts that warrant inclusive language and implement corresponding gender-fair alternatives, resulting in inappropriate application of gender-fair translation techniques. These findings underscore the need for more focused research to enhance language models' ability to accurately represent all genders when used for MT.

2 Gender-Fair Language in German

Grammatical gender languages such as German require gender marking in several word classes, e.g., nouns, pronouns, articles, and adjectives. In the case of person-referring terms, there is an overlap between the term's grammatical gender and the extra-linguistic referent's gender (Corbett, 1991). This may hinder the linguistic visibility and inclusion of gender identities beyond the binary because, for instance, most European languages have two or three grammatical genders only, i.e., masculine, feminine, and neuter.[2] Moreover, said languages rely on masculine forms as generics, reducing the visibility of women and other genders in linguistic and cognitive representation (Sato et al., 2025). This masculine-default bias perpetuates inequality, while gender-fair alternatives promote more inclusive representations (Fatfouta and Sczesny, 2023).

German has developed several approaches to gender fairness: neutralization replaces gendered terms with neutral alternatives (*"die Studierenden"* instead of *"die Studenten"*); coordination explicitly includes masculine and feminine forms (*"Studentinnen und Studenten"*); typographical solutions incorporate inclusive characters like the Gendersternchen (*"Student*innen"*), Gender-Doppelpunkt (*"Student:innen"*), or Gendergap (*"Student_innen"*); gender-fair neosystems, e.g., the ens-forms (*"Studens"*), introduce a new set of morphemes inclusive of all genders (Lardelli and Gromann, 2023b; Dick et al., 2024). These strategies vary in acceptance, reflecting ongoing evolution toward greater inclusivity.

3 Related Work

Gender Bias in MT. A long record of research on gender and MT has established that modern technologies are biased (Savoldi et al., 2025b), has designed resources to evaluate (binary) gender bias (Zhao et al., 2018; Stanovsky et al., 2019; Vanmassenhove and Monti, 2021; Currey et al., 2022; Rarrick et al., 2023; Piazzolla et al., 2024; Rarrick et al., 2024; Robinson et al., 2024, among others), and proposed methodological approaches to mitigate it (Saunders and Byrne, 2020; Saunders et al., 2020; Stafanovičs et al., 2020; Attanasio et al., 2023; Garg et al., 2024, among others). This extensive body of work has primarily focused

[2]The neuter for people is rare as it is typically associated with a negative social-evaluative function (Lind, 2022).

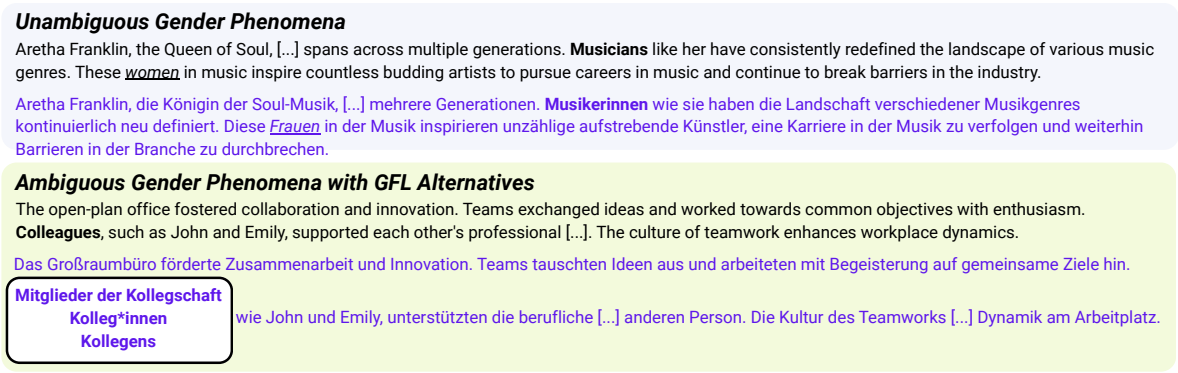

Figure 2: **Unmbiguous and ambiguous examples from GLITTER.** EN source sentence in black, DE post-edited reference in blue. For the unambiguous example, the referential gender of the seed noun is disambiguated by contextual cues in the trailing context. We provide three GFL alternatives (gender-star, -ens ending, and reworded neutral) for ambiguous terms.

on binary gender bias, examining how MT systems incorrectly translate between masculine and feminine forms or default to masculine generics. However, it does not address the broader challenge of gender-fair translation that encompasses non-binary and gender-neutral language forms.

Gender-Fair Language Resources. More recent research, questioning the old binary framework for gender, has sought to investigate NLP methods and resources for enhancing the representativeness of all genders. Classic NLP problems have been re-framed to include non-binary perspectives, e.g., text classification (Waldis et al., 2024) or (biased) language modeling (Felkner et al., 2023). Hossain et al. (2024) addressed misgendering of transgender and non-binary individuals and released a dataset with misgendering corrections. Inquiries into translation technologies revealed that MT models struggle to generate fair translations, particularly when handling neopronouns (Lauscher et al., 2023), even with leading commercial models (Savoldi et al., 2024b). GLITTER falls in this category, providing a comprehensive resource to evaluate gender-fair MT.

New MT benchmarks test whether modern translation systems can implement gender-fair translation from English into Italian (Piergentili et al., 2023, GeNTE), (Savoldi et al., 2025a, mGeNTE), (Piergentili et al., 2024, NEO-GATE), Icelandic (Friðriksdóttir, 2024, GenderQueer), French (Jourdan et al., 2025, FairTranslate), Spanish (Savoldi et al., 2025a), and German (Lardelli et al., 2024; Savoldi et al., 2025a). Most of these resources propose parallel corpora based on single sentence pairs (e.g., GeNTE/mGeNTE) or short passages (GenderQueer). GLITTER overcomes this limitation, providing four-sentence long passages with gender cues located in different parts of the context. Moreover, previous benchmarks propose fair translations, typically using a single gender-fair strategy. GeNTE and mGeNTE rely on gender-neutral rephrasing, while NEO-GATE uses neomorphemes to replace traditional gender markings. GenderQueer leverages Icelandic's neuter gender, and FairTranslate uses typographical marks and neomorphemes/neologisms. GLITTER improves upon existing benchmarks, provisioning three gender-fair alternatives, all curated by professional translators.

4 Constructing GLITTER

We create GLITTER through a three-step process (Figure 1): *i*) extracting and augmenting English sources to ensure comprehensive coverage of gender phenomena, *ii*) annotating gender contexts for nuanced evaluation, and *iii*) developing professionally post-edited references in three distinct gender-fair forms for each source passage. The complete data statement (Bender and Friedman, 2018) is given in Appendix A.

4.1 Collecting Source Sequences

Wikipedia Collection. We extract English source passages from Wikipedia,[3] using 115 gender-ambiguous plural nouns as seeds (e.g., *deputies*) from Lardelli et al. (2024) and our own additions. For each seed occurrence, we extract the matching sentence along with two preceding

[3] https://huggingface.co/datasets/wikimedia/wikipedia, version: 20231101.en, "train" split.

and one trailing sentence (when available) to capture both intra- and extra-sentential gender phenomena. Therefore, each passage is composed of four sentences. We apply quality filters (requiring proper POS tagging and minimum word counts) and retain passages containing gendered referential expressions (e.g., *she*, *mother*) that might disambiguate the gender of seed nouns, following the approach in Currey et al. (2022). Refer to Appendices A, B.1 and B.2 for more details. We ensured dataset diversity by extracting Wikipedia passages with varied styles, including quotations and direct speech. The multi-contributor nature of Wikipedia content provides natural variation in lexical choice, writing style, and topical coverage.

Synthetic Augmentation. To address underrepresented scenarios in our natural data, particularly unambiguous instances with gender cues in the trailing context, we augment the collection with synthetic data. This dataset is specifically designed to complement the encyclopedic style, extending the data source and style of our dataset; thus adding more diversity and variety to our benchmark. We prompt GPT-4o (`gpt-4o-2024-08-06`) to generate multi-sentence passages, then perform light post-editing to improve quality and diversity (see full details in Appendix B.4). Post-editing was necessary as LLM-generated passages predominantly defaulted to male gender references and intra-sentential disambiguation. Through targeted editing, we ensure balanced representation of disambiguation locations (preceding, matching, and trailing sentences) and diverse gender references, see Figure 2.

Adding Queer Content. MT systems often misgender individuals (e.g., via incorrect pronouns) or degender them (applying gender-neutral terms where gender-specific forms would be appropriate)[4] (Subramonian et al., 2025; Tomasev et al., 2021; Robinson et al., 2024). To address this representation issue, we deliberately incorporated passages from Wikipedia pages tagged with LGBTQ+ topics, thus collecting 112 additional instances (56 ambiguous and unambiguous gender contexts). This sampling ensures queer contexts are adequately represented and allows us to assess whether MT handles gender differently when dealing with queer-related content—where correct gendering matters particularly.

[4]E.g., using *"Patient*innen"* for a group of transgender women rather than a feminine form. This erasure of gender identity is particularly harmful for transgender individuals whose gender recognition is critical to their dignity.

4.2 Annotating Gender Phenomena

Ambiguity in Gender. We define *ambiguous* passages as those lacking gender cues for the seed anywhere in the context, where models should ideally avoid defaulting to specific gender forms. *Unambiguous* passages contain instead explicit gender cues disambiguating the seed word. Refer to Figure 2 for an example. We classify sources into mutually-exclusive scenarios: `Ambiguous` passages and `Unambiguous` passages with subcategories `Female`, `Male`, and `All genders` (indicating diverse gender representation beyond binary, e.g., *"Chairpeople, encompassing a diverse gender representation, led the discussions..."* or a combination of genders *"Voters, which includes women and non-binary folks..."*). The breakdown of the categories in GLITTER is given in Table 1.

Pre-annotation via LMs. Given the substantial scale of our dataset, we leverage LMs as pre-annotators to categorize passages by ambiguity type efficiently. Before pre-annotation, we first validate which LM strongly correlates with human judgment. We evaluate two strong multilingual models: OLMO 2 INSTRUCT (OLMo et al., 2024) and QWEN 2.5 72B (Yang et al., 2024). For validation, we randomly select 115 stratified instances covering diverse seed phrases and manually annotate them for gender ambiguity. Comparing these gold annotations against model predictions, QWEN 2.5 72B demonstrates a substantially higher agreement (weighted $F_1 = 0.781$, $\kappa = 0.62$) than OLMO 2 INSTRUCT (weighted $F_1 = 0.653$, $\kappa = 0.39$), particularly for correctly identifying unambiguous gender contexts (refer to Appendix C.2 for more details). Based on these results, we adopt QWEN 2.5 72B to pre-annotate our entire candidate pool. Using these pre-annotations, we sample 2,620 candidate passages, stratifying by seed phrase and balancing across ambiguity scenarios. For unambiguous passages, we additionally annotate which part of the context contains the gender cues (preceding, matching, or trailing sentences). Through iterative manual verification, we finalize 909 passages balanced across all scenarios.

Ambiguity	Synth (%)	#
Yes	43	458
No (F)	50	173
No (M)	39	208
No (All)	45	67

Table 1: **EN sources statistics.** Counts by seed gender ambiguity type and synthetic instance ratio per split.

Model	CometKiwi (↑)	MetricX (↓)
NLLB 3.3B	0.415	5.983
GEMMA 3 27B	0.858	1.458
EUROLLM 9B	0.860	1.452
VESUVIUS	**0.875**	**1.346**

Table 2: Overall translation quality results on GLITTER.

4.3 Collecting German References

We automatically translate the final pool of passages using Tower Vesuvius,[5] a leading commercial translation system. We hire four professional translators with expertise in gender-fair German to annotate the translation hypotheses on several aspects, including whether the seed was correctly translated and which gender form was used by the model. The translators collect span-level annotations for precise analysis of disambiguation patterns in the translations. We task these professionals with post-editing the passages to provide three gender-fair reference variants through minimal changes: gender-neutral rewording, gender star (*), and ens-forms. The resulting corpus comprises **1,785 EN-DE parallel pairs**, where **909 unique English sources** are translated and professionally post-edited into **three different gender-fair forms** whenever possible. Refer to the Appendix A for the annotation guidelines.

5 Benchmarking Translation Systems

We conduct an empirical analysis to assess several MT models' performance on GLITTER. In particular, we are interested in assessing whether contemporary translation systems are robust to the various setups we propose—ambiguous seeds that require gender-fair forms, unambiguous scenarios with different genders of the referent, location of disambiguation gender cues, and synthetic vs. naturalistic contexts.

5.1 Overall Translation Quality

We begin by verifying whether MT systems translate the passages in GLITTER reasonably well. Besides gender-related aspects, length makes our passages more challenging than existing gender benchmarks—the median length in GLITTER's sources is 71 words, compared to GeNTE's 26 (Piergentili et al., 2023), and GenderQueer's 15 (Friðriksdóttir, 2024).

[5] `https://www.widn.ai` accessed on February 20, 2025.

Setup. We test the overall translation quality of four state-of-the-art multilingual models: NLLB 3.3B (Team et al., 2022), an encoder-decoder model specialized for MT, two general-purpose open-weight autoregressive LMs in their instruct version, EUROLLM 9B (Martins et al., 2025) and GEMMA 3 27B (Team et al., 2025), and we also include the original translation from VESUVIUS (see §4.3). To assess quality, we use two reference-free metrics that ranked among the top performers in recent editions of the WMT QE Shared Task (Blain et al., 2023): CometKiwi 23 XXL (Rei et al., 2023), producing sentence-level quality scores ranging from 0 to 1, and MetricX QE 23 XXL (Juraska et al., 2023) with scores ranging from 0 to 25 and representing errors (the lower the better). We used reference-free metrics as we want to assess fluency and adequacy independently from post-edited references. The details on the experimental setup are given in the Appendix B.5.

Results. Table 2 shows the evaluation results for all models on English-to-German zero-shot translation. VESUVIUS performs best, followed by GEMMA 3 27B and EUROLLM 9B that share similar performance. Notably, EUROLLM 9B is slightly better than GEMMA 3 27B despite being one third of its size. NLLB 3.3B performs significantly worse and is hence not suitable for the task.

5.2 Manual Analysis of Gender Forms

We continue our analysis on VESUVIUS and EUROLLM 9B, the two models with the highest overall quality. In particular, we are interested in assessing how the seed noun is translated across GLITTER's diverse source scenarios.

Manual Analysis. We manually annotate all translations to characterize the seed gender in the German target passage, assigning one of the eight labels: (1-3) `gendered`, either `Female`, `Male`, or `Both` (e.g., *"Teilnehmerinnen und Teilnehmer"*), (4) `non-binary` (e.g., use of typographical solutions like the gender-star), (5) `neutral (all)` (i.e., use of gender-neutral alternatives (e.g., *"die*

Teilnehmemenden"), (6) `untranslated` (i.e., the seed appears in English as part of larger expressions such as *"Assembly of Participants"*), (7) `reworded` (e.g., use of synonyms), or (8) `error` (i.e., semantically wrong translation or omissions). Note that, for these annotations, we focus on the translated seed only. For instance, *"weibliche Verwandte"* (female relatives) is `neutral (all)`, considering the neutral seed word *"Verwandte"*. We provide all results in Appendix B.3. In what follows, we describe the general trends and then dive into the detailed results for the diverse context-related scenarios. The detailed results are in Table 3. Readers should note that this analysis was done on zero-shot translations without reference to our post-edited data.

5.2.1 Findings

Results show that **the majority of passages have been translated as `gendered male`** by both VESUVIUS and EUROLLM 9B. This finding is in line with the previous results of Lardelli et al. (2024), obtained on the initial set of seed nouns.

When source gender is ambiguous, LMs default to masculine forms rather than using gender-neutral alternatives. Table 3 (column **Ambiguity**) shows that both models predominantly translated ambiguous source texts into `gendered male` (84%), except for a few translated as `gendered female` (∼3%) or `neutral (all)` (∼7%).

Even with explicit female gender markers in the source, LMs struggle to produce appropriate feminine forms consistently. For unambiguous female contexts, VESUVIUS produced feminine translations in 51.9% of cases versus EUROLLM 9B's 39.7%. This result reveals a concerning aspect, with models' biases towards the masculine weighing more than explicit gender markers.

LMs persistently default to masculine forms even when source texts explicitly indicate inclusive contexts. For unambiguous "all genders" contexts (which ideally trigger inclusive forms), neutral translations reached only 10.3% while masculine forms still dominated (80.5–83.9%). This demonstrates that current MT systems default to masculine translations even when inclusivity is explicitly indicated.

LMs exhibit a clear positional bias in gender cue interpretation. Table 3 (column **Context Cue**) shows both models more accurately interpret gender cues in preceding/matching sentences (∼62–72% masculine) than in trailing contexts, where seed nouns are predominantly translated as masculine (79–85%). VESUVIUS shows stronger capability than EUROLLM 9B in interpreting contextual gender cues.

5.3 Automating Evaluations on GLITTER

To facilitate future research on the benchmark, we explore whether LLM-as-a-judge approaches can automate the evaluation on GLITTER. Following recent work on automatic gender-neutral translation evaluation (Piergentili et al., 2025), we are interested in detecting which gender form a hypothetical MT system uses when translating the source passages. In practice, we test whether general-purpose LMs can approximate human judgment to a reasonable extent. This idea echoes those of several prior works that explored the use of large-scale models to automate evaluation of NLP tasks (Bavaresco et al., 2025), including MT (Kocmi and Federmann, 2023; Vu et al., 2024), and, more specifically, gender-fair translation (Piergentili et al., 2025).

Setup. We prompt three open-weight instruct models—GEMMA 3 27B (Team et al., 2025), QWEN 2.5 72B (Yang et al., 2024), and QWEN 3 32B (Yang et al., 2025)—and GPT-4.1 (`gpt-4.1-2025-04-1`) to produce one of the labels to characterize the gender of the German translation (§5.2). We craft several prompts that allow an instruction-following model to predict such labels. Following Piergentili et al. (2025), we vary the information available to the LLM critic (e.g., showing the seed phrase or not, providing the source English passage) to generate an assessment. Moreover, we explore the effectiveness of in-context learning by providing eight randomly shuffled examples (see details in Appendix B.5).

Results. Table 4 reports the F1 and Recall scores of the four judges across all the types of sources with the best configuration found. **GPT-4.1 is the best model by a large margin**, achieving solid overall scores (F1: 0.88) and a peak recall for unambiguous sources (all genders) of 0.91. The second-best judge is QWEN 2.5 72B (F1: 0.72, Recall: 0.66), while smaller GEMMA 3 27B and QWEN 3 32B fall short despite the claimed multilingual capabilities. Across all models, we report higher results on average when the prompt also includes the source passage, in line with Pier-

		Overall	Type		Ambiguity				Context Cue		
		Total	Nat.	Syn.	Amb.	Male	Female	All	Prec.	Match.	Trail.
Male	VESUVIUS	76.65	75.00	78.80	84.30	92.70	36.00	80.50	62.60	69.30	79.10
	EUROLLM 9B	78.46	75.80	82.10	84.30	90.40	46.60	83.90	71.90	68.90	84.90
Female	VESUVIUS	12.12	10.40	14.40	3.60	0.50	51.90	4.60	26.50	19.70	14.00
	EUROLLM 9B	9.72	10.40	8.80	3.40	0.90	39.70	3.40	16.80	18.90	7.00
All	VESUVIUS	0.10	0.00	0.20	0.00	0.00	0.00	1.10	0.00	0.40	0.00
	EUROLLM 9B	0.20	0.20	0.20	0.40	0.00	0.00	0.00	0.00	0.00	0.00
Neutral	VESUVIUS	7.31	7.90	6.50	7.80	4.60	7.90	10.30	8.40	6.30	5.80
	EUROLLM 9B	6.61	5.60	7.70	6.60	5.00	6.90	10.30	7.10	6.30	7.00

Table 3: **Manual Analysis: Distribution of gender forms in translations (values in percentages) across different conditions.** Headers: **Overall** (percentage across all instances), **Type** (Natural vs. Synthetic text), **Ambiguity** (Ambiguous or Unambiguous gender with Male/Female/All subcategories), and **Context Cue** (location of gender disambiguation: Preceding, Matching, or Trailing sentence). Refer §5.2 for details.

Judge	F1-score	Recall	Ambiguous	Unambiguous (F)	Unambiguous (M)	Unambiguous (All)
GEMMA 3 27B	0.5703	0.4865	0.3857	0.6720	0.5936	0.3953
QWEN 2.5 72B	0.7232	0.6640	0.6083	0.7407	0.8174	0.4302
QWEN 3 32B	0.6630	0.5587	0.5109	0.5926	0.6530	0.5233
GPT-4.1	**0.8769**	**0.8586**	**0.8410**	**0.8571**	**0.8813**	**0.9070**

Table 4: **Judges alignment with human labels.** Weighted F1-score and Recall on GLITTER (left); Recall separately by scenario type for the English source (right): ambiguous and unambiguous (*female* (F), *male* (M), *all*). Results against human annotations on translations generated by VESUVIUS.

gentili et al. (2025). The results are consistent with EUROLLM 9B's outputs (see all results in Appendix C.2) and highlight how the chosen open-weight models are weaker judges compared to commercial systems. Overall, we find the LLM-as-a-judge approach, if used with care, to be a feasible alternative for human judgements on GLITTER.

6 Eliciting Gender-Fair Translation

We showed that MT systems consistently default to masculine forms when translating gender-ambiguous English multi-sentence text into German. Here, we explore whether LMs can be prompted to produce gender-fair translations that appropriately respond to contextual cues.

6.1 Setup

Models and prompts. We evaluate three instruct language models: EUROLLM 9B, GEMMA 3 27B, QWEN 2.5 72B and five prompting strategies: (1) **Zero-Shot (ZS)**, providing basic instruction to use gender-fair language when appropriate; (2) **Few-Shot (FS)**, adding examples of gender-star, ens-forms, and neutral rewording alternatives; (3) **Contrastive (Contrast FS)**, explicitly comparing masculine-generic translations with gender-fair alternatives; (4) **Chain-of-Thought (CoT)**, demonstrating reasoning about gender-ambiguous terms; and (5) **Multi-Turn CoT**, implementing a three-step translation, analysis, and refinement.

Evaluation Methods. We evaluate translations both for quality and gender treatment. For quality, we use reference-free metrics CometKiwi (Rei et al., 2023) and MetricX (Juraska et al., 2023) along with reference-based metrics BLEU, COMET, and chrF using professional post-edits as references (results in Appendix D). Since prior work has shown these metrics fail to accurately capture gender phenomena (Zaranis et al., 2025), we employed GPT-4.1 as a judge (in §5.3) to classify how each seed term was translated. We then simplified the analysis by grouping the detailed classification labels into two coarse categories: gender-specific (including `gendered male/female` labels), gender-fair (including `gendered both`, `non-binary`, `neutral (all)`, `untranslated`, and `reworded`). This categorization allowed us to more effectively assess whether translations maintained gender neutrality when appropriate or defaulted to gender-specific forms. Detailed prompting approaches and experimental results are available in Appendix D.

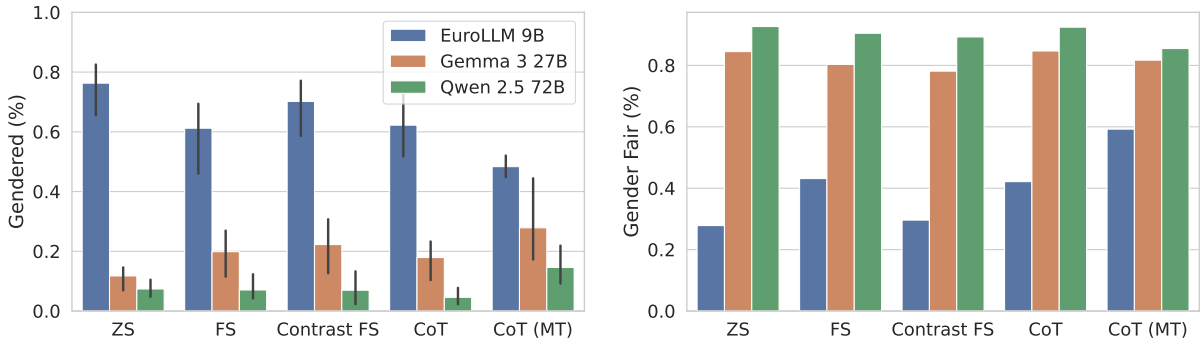

Figure 3: **Ratio of gender forms used across prompt strategies and models.** We test five strategies: zero-shot (ZS), few-shot (FS), contrastive few-shot (Contrast FS), chain-of-thought (CoT), multi-turn chain-of-thought (CoT MT). Left: unambiguous (female / male/ all genders) instances. Right: ambiguous instances. Higher is better.

6.2 Results

Our findings reveal several key patterns about LMs' gender-fair translation capabilities:

LMs fail at choosing when to use gender-fair forms. Figure 3 highlights two opposite behaviors. On the one hand, EuroLLM 9B often consistently uses gendered forms (72-81%), serving poorly ambiguous scenarios (right-end chart). On the other hand, Gemma 3 27B and Qwen 2.5 72B overly rely on gender-fair forms (75-92%) as a result of our prompts, rendering them unusable for gender-specific passages (left-end chart). Moreover, larger models (Gemma 3 27B and Qwen 2.5 72B) can generate gender-fair forms already in a zero-shot setup, leading to no improvements for more complex prompts. However, prompt complexity helps EuroLLM 9B, with FS, CoT, and multi-turn CoT yielding more gender-fair forms.

LMs struggle more with disambiguating gender in synthetic passages. Table 5 contrasts how models handle synthetic versus natural passages in unambiguous contexts. On synthetic text, EuroLLM 9B shows a large reduction in gender-specific translations with CoT-MT (82.4% → 36.1%), whereas gains on natural text are modest (81.7% → 68.7%). Gemma 3 27B and Qwen 2.5 72B exhibit the same pattern, yielding fewer gender-specific translations on synthetic passages across all prompting strategies. We attribute this gap to input complexity: synthetic passages contain fewer competing cues, so the instruction can dominate. Natural passages, by contrast, activate stronger world-knowledge and frequency-based priors (e.g., names, occupations, long-distance coreference), and these priors strengthen with model size, making default gendering harder to override. As a result, a prompt that works well on synthetic text works less well on natural text; this difference is larger for bigger models.

Model	Prompt	Wiki	Synt	Queer	Non-Q
EuroLLM	ZS	81.7	82.4	80.0	82.0
	FS	73.4	56.0	62.0	75.9
	CFS	79.3	72.2	79.6	79.3
	CoT	78.4	53.7	72.0	79.8
	CoT-MT	68.7	36.1	68.0	68.9
Gemma 3	ZS	21.6	3.7	12.0	23.7
	FS	36.7	6.5	32.0	37.7
	CFS	37.3	14.8	26.5	39.6
	CoT	31.7	6.5	12.0	36.0
	CoT-MT	39.2	30.6	32.0	40.8
Qwen 2.5	ZS	15.1	2.3	8.0	16.7
	FS	15.5	2.3	6.0	17.5
	CFS	15.9	2.8	12.2	16.7
	CoT	9.0	1.9	6.0	9.6
	CoT-MT	25.2	5.6	20.0	26.3

Table 5: **Percentage of gender-specific translations in unambiguous contexts.** (Higher is preferred.) Results across different prompting strategies—zero-shot (ZS), few-shot (FS), contrastive few-shot (CFS), chain-of-thought (CoT), and multi-turn CoT (CoT-MT)—and source categories: naturalistic text from Wikipedia, synthetically generated text, queer and non-queer related content.

Queer content elicits more gender-fair language despite gender cues. All models produce more gender-fair translations when handling queer-related content, even when explicit gender cues are present in the source text. EuroLLM 9B shows lower gender-specific translation rates for queer content (62-80%) versus non-queer content (68-82%) across all prompting strategies. Similarly, Gemma 3 27B uses gendered forms for only 12-32% of queer passages compared to 23-41% for non-queer content. In contrast, Qwen 2.5 72B

maintains this pattern with even stronger preference for gender-fair forms (6-20% gendered for queer content). This systematic difference reveals that models have learned to associate queer topics with gender-inclusive language conventions, suggesting that content domain influences translation choices independent of grammatical gender cues.

Taken all together, these findings suggest that leading multilingual open-weight models struggle to produce consistent and reliable gender-fair translations, further underscoring the importance of GLITTER as a challenging benchmark.

7 Conclusion

We introduced GLITTER, a dataset comprising 909 sources and three parallel, professionally curated German translations, totaling 1,785 translation pairs—the largest benchmark for testing gender-fair English-German MT to date. GLITTER brings several innovations including *i*) longer passages and annotations to study the impact of contextual information, *ii*) human-written references with three gender-fair reformulations (neutral, gender-star, -ens form) , and a *iii*) fine-grained span-level annotation layer for analyzing gender-related phenomena. To complement the resource, we presented a language model-based automatic metric and baseline results testing several translation systems and prompting strategies. All our results underscore the same compelling finding: contemporary translation systems still struggle with gender-fair language, fail in understanding when and how to use it (even if prompted with extensive guidelines on how to do so or if provided with explicit gender cues in the extensive source context), and generally default to (generic) masculine forms.

Limitations

Two primary limitations of this work warrant consideration in future research.

First, GLITTER draws its initial corpus exclusively from Wikipedia, a single naturalistic source. The dataset primarily covers gender-related attributes in encyclopedic contexts where named entities and personal names appear frequently. Nevertheless, many comparable studies have used Wikipedia as a standard starting point for high-quality text (e.g., Currey et al., 2022; Savoldi et al., 2024a; Piergentili et al., 2025). During the annotation process for ambiguity and referential gender, we deliberately avoided inferences based on personal names and provided professional translators with consistent guidelines. Additionally, we enhanced GLITTER with human-validated synthetic passages to improve source diversity.

Second, large-scale automated evaluation on GLITTER requires language model-based judges, which presents two notable constraints. This approach depends on commercial APIs (with our optimal metric relying on OpenAI's GPT 4.1), introducing additional costs and potential inconsistencies when models receive updates. Furthermore, our experimental results show imperfect correlation with human judgment, potentially leading to inaccurate MT system quality assessments. To address this limitation, we tested multiple metrics and prompting strategies, documenting both the capabilities and limitations of open-weight and commercial language model judges for this specific evaluation task.

Ethical Considerations

This research addresses ethical dimensions in LMs' translation capabilities. The dataset and evaluation metric we present emphasize the importance of non-exclusionary language that acknowledges gender diversity beyond binary constructs. We recognize the broad variation in gender-fair linguistic strategies and do not prescribe any specific approach as universally optimal. Accordingly, our methodology incorporates multiple gender-inclusive solutions for each source text: neutral rephrasing that minimizes unnecessary gender markers, typographical alternatives, and neologistic systems designed to represent all genders. This comprehensive approach acknowledges the linguistic and cultural complexity of gender representation in translation.

Acknowledgments

We gratefully acknowledge the financial support from the European Association for Machine Translation (EAMT) through its sponsorship of activities in 2024, which made it possible to compensate the professional translators involved in this project. We also wish to sincerely thank said translators for their expertise and dedication in carrying out the dataset annotation and gender-fair post-edits. The work of Anne Lauscher and Pranav A is funded under the Excellence Strategy of the German Federal Government and States. Janiça Hackenbuchner was funded by The Research Foundation – Flan-

ders (FWO), research project 1SH5V24N (from 01.11.2023 until 31.10.2027) and used computational resources (Stevin Supercomputer Infrastructure) and services provided by the VSC (Flemish Supercomputer Center), funded by Ghent University, FWO and the Flemish Government – department EWI. Giuseppe Attanasio is supported by the Portuguese Recovery and Resilience Plan through projects C645008882-00000055 (Center for Responsible AI) and UID/50008: Instituto de Telecomunicações. Manuel Lardelli was supported by the European Union – NextGenerationEU, within the framework of the Italian National Recovery and Resilience Plan, Mission 4 "Education and Research", Component 2 "From Research to Enterprise", Investment 1.2 "Funding of projects submitted by young post-doctoral researchers".

References

Giuseppe Attanasio, Flor Miriam Plaza del Arco, Debora Nozza, and Anne Lauscher. 2023. A tale of pronouns: Interpretability informs gender bias mitigation for fairer instruction-tuned machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3996–4014, Singapore. Association for Computational Linguistics.

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

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria. Association for Computational Linguistics.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Frederic Blain, Chrysoula Zerva, Ricardo Rei, Nuno M. Guerreiro, Diptesh Kanojia, José G. C. de Souza, Beatriz Silva, Tânia Vaz, Yan Jingxuan, Fatemeh Azadi, Constantin Orasan, and André Martins. 2023. Findings of the WMT 2023 shared task on quality estimation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 629–653, Singapore. Association for Computational Linguistics.

Prafulla Kumar Choubey, Anna Currey, Prashant Mathur, and Georgiana Dinu. 2021. GFST: Gender-filtered self-training for more accurate gender in translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1640–1654, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Greville G Corbett. 1991. *Gender*. Cambridge University Press.

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

Anna-Katharina Dick, Matthias Drews, Valentin Pickard, and Victoria Pierz. 2024. GIL-GALaD: Gender inclusive language - German auto-assembled large database. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7740–7745, Torino, Italia. ELRA and ICCL.

EIGE. 2019. *Toolkit on Gender-sensitive Communication: A Resource for Policymakers, Legislators, Media and Anyone Else with an Interest in Making Their Communication More Inclusive*. European Institute for Gender Equality (EIGE).

Ramzi Fatfouta and Sabine Sczesny. 2023. Unconscious bias in job titles: Implicit associations between four different linguistic forms with women and men. *Sex Roles*, 89(11):774–785.

Virginia Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. 2023. WinoQueer: A community-in-the-loop benchmark for anti-LGBTQ+ bias in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9126–9140, Toronto, Canada. Association for Computational Linguistics.

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

Sarthak Garg, Mozhdeh Gheini, Clara Emmanuel, Tatiana Likhomanenko, Qin Gao, and Matthias Paulik. 2024. Generating gender alternatives in machine

translation. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 237–254, Bangkok, Thailand. Association for Computational Linguistics.

P. Gygax, U. Gabriel, O. Sarrasin, J. Oakhill, and A. Garnham. 2008. Generically intended, but specifically interpreted: When beauticians, musicians, and mechanics are all men. In *LANGUAGE AND COGNITIVE PROCESSES*, volume 23:3, pages 464–485.

Levi C. R. Hord. 2016. Gender neutral language in english, swedish, french, and german. In *Bucking the Linguistic Binary*.

Tamanna Hossain, Sunipa Dev, and Sameer Singh. 2024. MisgenderMender: A community-informed approach to interventions for misgendering. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7538–7558, Mexico City, Mexico. Association for Computational Linguistics.

Fanny Jourdan, Yannick Chevalier, and Cécile Favre. 2025. Fairtranslate: An english-french dataset for gender bias evaluation in machine translation by overcoming gender binarity. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 150–166.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

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

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

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

Manuel Lardelli and Dagmar Gromann. 2023a. Gender-fair post-editing: A case study beyond the binary. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 251–260.

Manuel Lardelli 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.

Anne Lauscher, Debora Nozza, Ehm Miltersen, Archie Crowley, and Dirk Hovy. 2023. What about "em"? how commercial machine translation fails to handle (neo-)pronouns. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 377–392, Toronto, Canada. Association for Computational Linguistics.

Miriam Lind. 2022. Liminalität, Transdifferenz und Geschlecht: Sprachliche Praktiken jenseits von Zweigeschlechtlichkeit. *Zeitschrift für Literaturwissenschaft und Linguistik*, 52(4):631–649.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M Guerreiro, Ricardo Rei, Duarte M Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, and 1 others. 2025. Eurollm: Multilingual language models for europe. *Procedia Computer Science*, 255:53–62.

Martin Montgomery. 2008. *An introduction to language and society*. Routledge.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, and 1 others. 2024. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*.

Silvia Alma Piazzolla, Beatrice Savoldi, and Luisa Bentivogli. 2024. Good, but not always fair: An evaluation of gender bias for three commercial machine translation systems. *HERMES-Journal of Language and Communication in Business*, (63):209–225.

Andrea Piergentili, Beatrice Savoldi, Dennis Fucci, Matteo Negri, and Luisa Bentivogli. 2023. Hi guys or hi folks? benchmarking gender-neutral machine translation with the GeNTE corpus. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14124–14140, Singapore. Association for Computational Linguistics.

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

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

Spencer Rarrick, Ranjita Naik, Varun Mathur, Sundar Poudel, and Vishal Chowdhary. 2023. Gate: A challenge set for gender-ambiguous translation examples. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 845–854.

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

Ricardo Rei, Nuno M. Guerreiro, JosÃ© Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. Association for Computational Linguistics.

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

Sayaka Sato, Pascal Mark Gygax, Ute Gabriel, Jane Oakhill, and Lucie Escasain. 2025. Does inclusive language increase the visibility of women, or does it simply decrease the visibility of men? a missing piece of the inclusive language jigsaw. *Collabra: Psychology*, 11(1).

D. Saunders and B. 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*, page 7724–7736. Association for Computational Linguistics.

Danielle Saunders, Rosie Sallis, and Bill Byrne. 2020. Neural machine translation doesn't translate gender coreference right unless you make it. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 35–43, Barcelona, Spain (Online). Association for Computational Linguistics.

Beatrice Savoldi, 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 *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.

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

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

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

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

Artūrs Stafanovičs, Toms Bergmanis, and Mārcis Pinnis. 2020. Mitigating gender bias in machine translation with target gender annotations. In *Proceedings of the Fifth Conference on Machine Translation*, pages 629–638, Online. Association for Computational Linguistics.

Gabriel Stanovsky, 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, Florence, Italy. Association for Computational Linguistics.

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

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Nenad Tomasev, Kevin R. McKee, Jackie Kay, and Shakir Mohamed. 2021. Fairness for unobserved characteristics: Insights from technological impacts on queer communities. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 254–265, New York, NY, USA. Association for Computing Machinery.

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

Tu Vu, Kalpesh Krishna, Salaheddin Alzubi, Chris Tar, Manaal Faruqui, and Yun-Hsuan Sung. 2024. Foundational autoraters: Taming large language models for better automatic evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17086–17105, Miami, Florida, USA. Association for Computational Linguistics.

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

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Emmanouil Zaranis, Giuseppe Attanasio, Sweta Agrawal, and Andre Martins. 2025. Watching the watchers: Exposing gender disparities in machine translation quality estimation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25261–25284, Vienna, Austria. Association for Computational Linguistics.

J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K. W. Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 2 (Short Papers), pages 15–20. Association for Computational Linguistics.

A GLITTER Data Statement

A.1 Executive Summary

GLITTER is a benchmark dataset for evaluating gender-fair MT from English to German. It contains approximately 1000 English source passages with one to three parallel professionally post-edited German translations for each source (2010 translations total), implementing different gender-fair strategies: gender-neutral rewording, gender star (*), and ens-forms. The dataset uniquely features extended multi-sentence passages with gender disambiguation across different context positions and combines both natural (Wikipedia) and synthetic passages. Each passage has been carefully annotated for gender phenomena and ambiguity.

A.2 Curation Rationale

The GLITTER dataset aims to evaluate how MT systems handle gender representation when translating from English (with relatively few gender markers) to German (requiring extensive gender marking). The dataset was designed to address the lack of comprehensive resources for gender-fair English-to-German translation that go beyond single sentences and binary gender representation.

Each data instance consists of a four-sentence passage: two preceding sentences, one matching sentence containing a seed noun, and one trailing sentence. The dataset is structured to represent various gender ambiguity scenarios, allowing for the evaluation of how context influences translation decisions. The dataset was deliberately designed to balance ambiguous cases (where gender is not specified and gender-fair language would be appropriate) and unambiguous cases with gender cues in different positions (preceding, matching, or trailing sentences).

The dataset balances natural and synthetic data, with stratification across seed nouns and different gender phenomena (ambiguous, unambiguous female, unambiguous male, unambiguous mixed/all genders).

A.3 Documentation for Source Datasets

The natural passages in GLITTER were extracted from the English Wikipedia (version: 20231101.en, "train" split, available at https://huggingface.co/datasets/wikimedia/wikipedia). The synthetic data was generated using GPT-4o (`gpt-4o-2024-11-20`) and post-edited by human annotators to improve quality and ensure an appropriate distribution of gender cues.

A.4 Language Varieties

- **Source language:** en-US (English as used in Wikipedia, primarily standard English with encyclopedic style)
- **Target language:** de-DE (Standard German, including various gender-fair language forms: neutral rewording, gender star (*), and ens-forms)

A.5 Language User Demographic

Natural data: The source texts stem from Wikipedia, written by a diverse community of editors whose exact demographic characteristics are unknown but are generally representative of Wikipedia contributors (historically skewing toward educated adults from North America and Europe, with efforts toward greater diversity in recent years).

Synthetic data: The synthetic texts were generated by GPT-4o, then lightly post-edited by research team members to better represent diverse gender configurations (e.g., gender substitution to increase the number of unambiguous feminine passages) and disambiguation patterns (e.g., to increase the number of passages including disambiguating cues in the trailing context).

A.6 Annotator Demographic

Four professional translators, all native German speakers with high proficiency in English, were hired to perform the annotation and post-editing tasks. All translators had previous experience with gender-fair language in German and were specifically selected based on their expertise in this area. Translators were compensated at a rate of EUR 50 per hour. Each translator was assigned approximately 255 passages for annotation and post-editing.

A.7 Linguistic Situation and Text Characteristics

- **Time and place:** The natural data stems from Wikipedia articles as of November 2023. Synthetic data was generated in December 2024. The dataset was created between December 2024 and May 2025.
- **Modality:** Written text.

- **Genre:** Encyclopedic text (natural) and similar encyclopedic-style text (synthetic).
- **Structure:** Each passage consists of four sentences: two preceding sentences providing context, one matching sentence containing the seed noun, and one trailing sentence.
- **Topics:** Diverse, based on Wikipedia content and synthetic generation. Some passages include queer-related content, which showed distinct patterns in translation.
- **Context:** Passages include explicit gender cues in different positions (preceding, matching, or trailing sentences) for unambiguous cases.

A.8 Preprocessing and Data Formatting

The construction of GLITTER followed these steps:

1. **Seed word selection:** 115 gender-ambiguous plural noun phrases in English (e.g., "counsellors," "economists") were selected, based on and expanded from previous work.
2. **Wikipedia passage extraction:** For each seed noun, sentences containing the seed were extracted from Wikipedia. For each matching sentence, two preceding sentences and one trailing sentence were also extracted to form a complete passage.
3. **Quality filtering:** Quality filters were applied to extract coherent passages, including POS tagging requirements, minimum word count, and removal of formatting patterns.
4. **Gender disambiguation filtering:** Passages were filtered to include only those with at least one gender-specific term from a predefined list of gendered words.
5. **Ambiguity annotation:** Language models (QWEN 2.5 72B) were used to pre-annotate whether the gender of the seed noun was ambiguous or unambiguous, followed by human validation.
6. **Synthetic data generation:** To balance representation of different gender scenarios, additional synthetic data was generated using GPT-4o and manually post-edited.
7. **Machine translation:** The resulting passages were translated using Tower Vesuvius.
8. **Professional post-editing:** Professional translators annotated the translations and created three gender-fair variants for each passage where appropriate.

A.9 Annotation and Post-Editing Details

Annotation process. We provide here the condensed annotation guidelines used in our project. We report the full annotation guidelines and screenshots in our official repository at https://github.com/pranav-ust/glitter. We conducted the annotation using Label Studio[6]. Each annotation item included a seed noun and a source passage, segmented into preceding context, the matching sentence, and trailing context. Annotators were instructed to label all human entities in the matching sentence, including the seed noun. Subsequently, they were asked to determine whether the gender of the seed noun was ambiguous. If the gender was ambiguous, they could proceed to the next step. Otherwise, they were required to annotate the disambiguating span(s) and specify the gender. Annotators also had the option to leave comments on the source passage, for instance, to flag potential errors.

Annotators then evaluated the MT of the source passage. Both the original source and the corresponding VESUVIUS MT output were shown. As in the previous step, annotators assessed whether the gender of the seed noun was ambiguous. If not, they were asked to indicate the gender and assess whether the gender expressed in the MT matched the gender of the seed in the source.

Post-editing process. Where the gender of the seed noun in the MT did not match that of the source, annotators were instructed to post-edit the translation. Specifically, they were asked to provide:

- a German translation in the masculine form if the seed noun in the source was masculine.
- a German translation in the feminine form if the seed noun in the source was feminine.
- three gender-fair alternatives if the seed noun referred to non-binary individuals, a mixed-gender group, or all genders.
- three gender-fair alternatives if the gender of the seed noun was ambiguous.

[6] https://labelstud.io

The three gender-fair alternatives, i.e. gender-neutral rewording, gender star (*), and the -ens forms, were selected based on previous literature on gender-fair post-editing (Lardelli and Gromann, 2023a).

Translators' recruitment. Four professional translators were hired to perform the annotation and post-editing tasks for GLITTER. They were selected from the authors' professional network, having previously collaborated on research projects concerning gender-fair German. All translators were native German speakers with high proficiency in English. They were compensated at a rate of EUR 50 per hour.

To ensure the quality and consistency of the dataset, we provided detailed annotation and post-editing guidelines, along with a reference handout on gender-fair German. One of the authors held individual onboarding sessions with each translator to demonstrate the use of Label Studio and address any initial questions or concerns.We report the full annotation guidelines and screenshots in our official repository at https://github.com/pranav-ust/glitter.

Each translator was assigned 255 passages for annotation and post-editing. The entire process was closely supervised by one of the authors, who provided ongoing support and clarification as needed throughout the project.

A.10 Distribution

The dataset will be publicly available through a GitHub repository. The dataset is provided under an open license that allows for academic and research use with appropriate attribution.

The dataset will be maintained by the research team, with updates and errata published through the GitHub repository. Questions and issues can be directed through the repository's issue tracker.

A.11 Capture Quality and Translators' Feedback

Three of the four professional translators provided qualitative feedback on the machine translations. Overall, they noted that the output was of generally high grammatical quality, with only minor issues in syntax or fluency. However, some common shortcomings were identified:

- **Idiomatic and Figurative Language:** all translators observed that idioms and metaphors were often translated literally, impairing the clarity and appropriateness of the German output.
- **Gender Representation:** the translations frequently defaulted to masculine forms when gender was ambiguous in the source. In some cases, where gender was explicit, the MT system redundantly applied multiple gender-marking strategies (e.g., *weibliche Schriftstellerinnen*, EN: female writers), requiring post-editing.
- **Seed Noun Selection:** One translator noted that some seed terms (e.g., book titles or organization names) were improperly translated or treated inconsistently as proper names in the target language.

These insights underline both the strengths and limitations of current MT systems in handling gender representation and semantic nuance.

A.12 Glossary

- **Gender-fair language (GFL):** Language that represents, includes, and addresses all genders equally, avoiding gender-exclusive expressions (particularly masculine generics).
- **Gender star (*):** A typographical solution in German to represent all genders in written language (e.g., "Student*innen").
- **Ens-forms:** A neologistic gender-fair system in German that introduces a new set of morphemes (e.g., "Studens" instead of "Student/Studentin").
- **Gender-neutral rewording:** The replacement of gendered terms with gender-neutral alternatives (e.g., "die Studierenden" instead of "die Studenten").
- **Seed noun:** The target gender-ambiguous plural noun in each passage that serves as the focus for gender disambiguation and translation evaluation.
- **Ambiguous:** Passages where the gender of the seed noun is not specified in the context.
- **Unambiguous:** Passages where the gender of the seed noun is explicitly specified through contextual cues.

A.13 Examples from GLITTER

This section provides annotated examples from our dataset, showcasing different gender disambiguation scenarios and their corresponding gender-fair translation alternatives. The seed words are highlighted in **bold**, and gender disambiguating context is underlined.

Ambiguous gender example from Wikipedia: Seed word is interpreters.

Source (English): North Korean fiction provides insights into how foreigners, in particular Russians are viewed. During the 1940s and 1950s Soviet Russians were portrayed as ideological guides of Koreans. In literature from the 2000s, the tables have turned and now Russians look up to Koreans as the **interpreters** of socialist values and initiative. For instance, Rim Hwawon's representative short story, "The Fifth Photo" follows the ordeal of a Russian girl in a post-Soviet world.

Gender-neutral translation: Nordkoreanische Belletristik gibt Aufschluss darüber, wie das Ausland, insbesondere Russland, wahrgenommen wird. In den 1940er- und 1950er-Jahren wurden das sowjetische Russland als ideologische Führung von Nordkorea dargestellt. In der Literatur der 2000er-Jahre haben sich die Rollen geändert, und nun schaut Russland zu Nordkorea auf, das als **Interpret** der sozialistischen Werte und Initiativen gilt. Zum Beispiel erzählt Rim Hwawons repräsentative Kurzgeschichte „Das fünfte Foto" von den Leiden eines russischen Mädchens in einer post-sowjetischen Welt.

Gender-star translation: Nordkoreanische Belletristik gibt Aufschluss darüber, wie Ausländer*innen, insbesondere Russ*innen, wahrgenommen werden. In den 1940er- und 1950er-Jahren wurden sowjetische Russ*innen als ideologische Führer*innen der Koreaner*innen dargestellt. In der Literatur der 2000er-Jahre haben sich die Rollen geändert, und nun schauen die Russ*innen zu den Koreaner*innen auf, die als **Interpret*innen** der sozialistischen Werte und Initiativen gelten. Zum Beispiel erzählt Rim Hwawons repräsentative Kurzgeschichte „Das fünfte Foto" von den Leiden eines russischen Mädchens in einer post-sowjetischen Welt.

Ens-form translation: Nordkoreanische Belletristik gibt Aufschluss darüber, wie Ausländens, insbesondere Russens, wahrgenommen werden. In den 1940er- und 1950er-Jahren wurden sowjetischens Russens als ideologischens Führens dens Koreanens dargestellt. In der Literatur der 2000er-Jahre haben sich die Rollen geändert, und nun schauen dens Russens zu dens Koreanens auf, dens als **Interpretens** der sozialistischen Werte und Initiativen gelten. Zum Beispiel erzählt Rim Hwawons repräsentative Kurzgeschichte „Das fünfte Foto" von den Leiden eines russischen Mädchens in einer post-sowjetischen Welt.

Unambiguous gender example from Wikipedia: Seed word is patients.

Source (English): There is also an increased risk of heart disease, hypothyroidism such as Hashimoto's thyroiditis, Addison's disease, and other autoimmune disorders. Emotional health The most common words women use to describe how they felt in the two hours after being given the diagnosis of POI are "devastated", "shocked," and "confused." The diagnosis is more than infertility and affects a woman's physical and emotional well-being. **Patients** face the acute shock of the diagnosis, associated stigma of infertility, [...] symptoms of estrogen deficiency, worry over the associated potential medical sequelae such as reduced bone density and cardiovascular risk, and the uncertain future that all of these factors create. Women diagnosed with POI in their 20s have disproportionately reported experiencing dismissiveness, bias, and "not being taken seriously" by healthcare professionals.

Gender-specific translation: Es besteht auch ein erhöhtes Risiko für Herzkrankheiten, Hypothyreose wie die Hashimoto-Thyreoiditis, die Addison-Krankheit und andere Autoimmunerkrankungen. Emotionale Gesundheit Die häufigsten Wörter, die Frauen verwenden, um zu beschreiben, wie sie sich in den zwei Stunden nach der Diagnose POI fühlten, sind „am Boden zerstört", „geschockt" und „verwirrt". Die Diagnose bedeutet mehr als nur Unfruchtbarkeit und beeinträchtigt das körperliche und emotionale Wohlbefinden einer Frau. **Patientinnen** stehen vor dem akuten Schock der Diagnose, dem damit verbundenen Stigma der Unfruchtbarkeit, [...], Symptome eines Östrogenmangels, Sorgen über die damit verbundenen möglichen medizinischen Folgen wie reduzierte Knochendichte und kardiovaskulärisches Risiko und die unsichere Zukunft, die all diese Faktoren mit sich bringen. Frauen, bei denen POI in ihren Zwanzigern diagnostiziert wurde, berichteten unverhältnismäßig oft von Missachtung, Vorurteilen und „nicht ernst genommen werden" durch medizinisches Fachpersonal.

Unambiguous gender synthetic example: The seed word is travellers.

Source (English): The tourism industry saw a resurgence as global travel restrictions eased. Destinations lured visitors with unique experiences. Resourceful **travellers** embraced the spirit of adventure, exploring unfamiliar places and cultures with curiosity. The journeys of these women enriched personal perspectives, paving the way for cross-cultural understanding.

Gender-neutral translation: Die Tourismusbranche erlebte ein Comeback, als die weltweiten Reisebeschränkungen gelockert wurden. Reiseziele lockten Besucherinnen mit einzigartigen Erlebnissen. Findige **Reisende** folgten dem Abenteuergeist und erkundeten neugierig unbekannte Orte und Kulturen. Die Reisen dieser Frauen bereicherten persönliche Perspektiven und ebneten den Weg für interkulturelles Verständnis.

B Experimental Details

B.1 List of Seed Phrases

Our initial Wikipedia sample is based on extracting passages that mention specific seed phrases. We start the list of nouns introduced by Lardelli et al. (2024) and manually extend it to increase coverage and diversity. The full list is: addressers, administrators, advocates, background actors, beginners, bloggers, builders, businesspeople, call center operators, camera operators, ceramists, chairpeople, children's day carers, civil servants, clients, colleagues, commentators, consumer advisors, contemporaries, contractors, coordinators, counsellors, course participants, custodians, deputies, diabetics, directors, dispatchers, donors, economists, employee representatives, employees, employers, enforcement debtors, enthusiasts, experts, extremists, fellow citizens, fellow runners, forest keepers, freelancers, hairdressers, hosts, illustrators, influencers, intermediaries, interpreters, investigators, investors, invoice recipients, jews, jurists, kindergarten teachers, lawyers, laypeople, losers, mechanics, ministers, mood animators, moun-

taineers, neighbors, newcomers, notaries, operators, opinion leaders, organ donors, pacifists, painters, participants, partners, party leaders, pastors, patients, pensioners, performers, personnel development managers, practitioners, presidents, primary school pupils, private patients, producers, project leaders, prosecutors, readers, recipients, reformers, relatives, respondents, sailors, secretaries, senders, settlers, signees, slave holders, social workers, speakers, specialists, speculators, sportspeople, stonecutters, student representatives, supervisors, supremacists, tax consultants, team leaders, trainers, travellers, tutors, users, veterans, workers, xylophone players.

B.2 List of Gendered Words

The full list of gendered words is[7]: actress, archduke, ashi, ashwapati, auntie, bachelor, bahadur, baron, b*st*rd, baugrygr, begum, bitch, boy, boyfriend, brother, burgher, chhatrapati, count, dad, daddy, damapati, dame, darbar, dewan, duke, earl, emperor, empress, father, female, feminine, firewoman, firewomen, fräulein, gentleman, gentlemen, gentlewoman, girl, girlfriend, goodman, goodwife, grandfather, grandmother, granny, heir, heiress, hero, heroine, hostess, husband, inamdar, jarls, kabiraj, kaviraj, khan, khanum, lad, ladies, lady, lalla, landlady, lass, lord, madam, mademoiselle, male, mamsell, man, margrave, marquess, master, maternal, mayoress, men, mesne, miss, mistress, mom, mommy, monseigneur, monsieur, mother, mr, mrs, ms, mum, mummy, nephew, niece, noyan, pandit, papa, paternal, policewoman, policewomen, prince, qanungoh, rai, rao, ritter, saheb, sahib, sahibratna, senapati, senhor, shaikh, shrimati, sidi, sir, sister, sl*t, spinster, stewardess, stud, uncle, vaidya, viscount, waitress, wife, woman, women.

B.3 GLITTER Statistics based on Gender Form Use

Complementing Section 5.2, Table 3 shows the detailed analysis of annotation translations for models VESUVIUS and EUROLLM 9B. Figure 7 shows how queer-related topics in the source text influence gender in EUROLLM 9B and VESUVIUS translations.

[7]Derived from https://github.com/gregology/biased-words

B.4 Synthetic Data Creation

We prompted GPT-4o (`gpt-4o-2024-11-20`) to generate synthetic passages, complementary to the naturally sampled data. Since GPT-4o struggled in generating requested samples, where the gender was adequately disambiguated at all or anywhere but the matching sentence, we compiled our synthetic data based on a combination of a few different few-shot prompts. These followed the same overall pattern and ensured that we cover a range of linguistic features (ambiguous vs. unambiguous, for which the gender was either disambiguated in preceding, matching or trailing context) and gender phenomena (female, male or inclusive) as needed for this dataset. An example prompt used can be found in our https://github.com/pranav-ust/glitter.

B.5 Automatic Evaluation

We prompted Gemma 3 27B (https://huggingface.co/google/gemma-3-27b-it, Qwen 2.5 72B (https://huggingface.co/Qwen/Qwen2.5-72B-Instruct), and Qwen 3 32B (https://huggingface.co/Qwen/Qwen3-32B) using a temperature of 0, while we used standard parameters for GPT-4.1 (`gpt-4.1-2025-04-1`). We formatted the input using each model's chat template. We provided eight in-context exemplar shots, one for each label, randomly shuffled for each passage to translate, and formatted as the first eight conversation turns. We used guided decoding to force the output into valid JSON strings. Figure 8 reports the exact prompt.

We used model and code implementation from transformers (Wolf et al., 2020) and vLLM as the inference engine (Kwon et al., 2023).

C Translation Model Evaluation Results

C.1 Translation Quality Analysis

We evaluate four machine translation models on GLITTER using reference-free metrics to assess their performance across different content types. This section provides additional details on experimental setup and presents a comprehensive analysis of translation quality variations.

Experimental Setup. We used the following model checkpoints: NLLB 3.3B (Meta AI's No Language Left Behind 3.3B model), GEMMA 3 27B (Google DeepMind's Gemma 3 27B Instruct

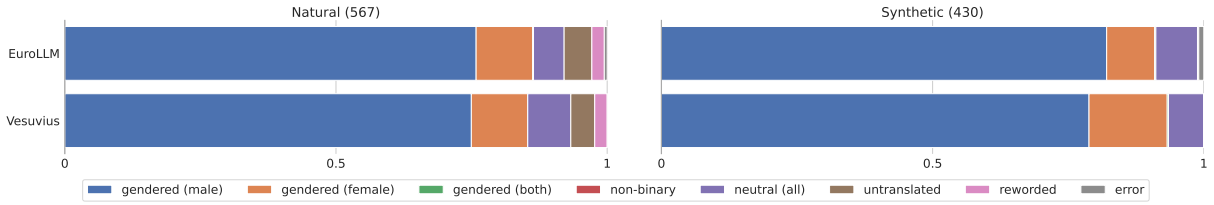

Figure 4: **Gender form used across source type: natural vs. synthetic.** Bars indicate the ratio of human-validated labels (§5.2) found per category. Category numerosity in each subtitle. Complementary to Table 3.

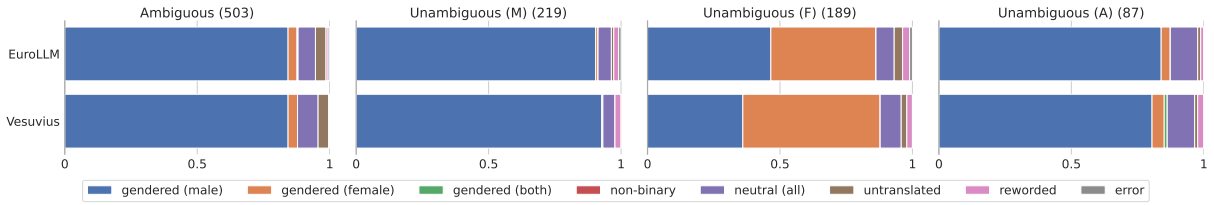

Figure 5: **Gender form used across source ambiguity type.** Bars indicate the ratio of human-validated labels (§5.2) found per category. Category numerosity in each subtitle. Complementary to Table 3.

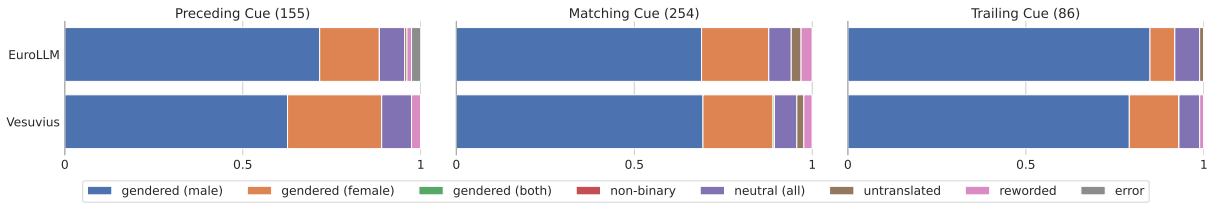

Figure 6: **Gender form used across source unambiguous types (preceding, matching, trailing).** Bars indicate the ratio of human-validated labels (§5.2) found per category. Category numerosity in each subtitle. Complementary to Table 3.

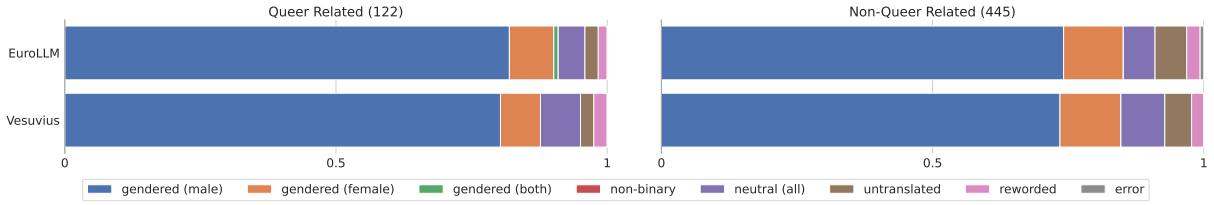

Figure 7: **Gender form used across source content type: queer-related or not.** Bars indicate the ratio of human-validated labels (§5.2) found per category. Category numerosity in each subtitle. Complementary to Table 3.

```
You are an expert language annotator and evaluator of gender-fair translation from English into German.

I will provide you with an English seed noun phrase, a source passage in English mentioning the seed, and an automatic translation of that passage in German.
Your task is to identify the translation of the English phrase and assess its referential gender.

We define referential gender as the actual gender identity of the referents in the real world. For instance, the word "Mensch" in German has masculine grammatical gender but is semantically gender-neutral.

Identify ONE of the following referential gender labels:
- GENDERED FEMALE: the translated seed noun has feminine grammatical gender and it refers specifically to individuals who identify as women, e.g., "Teilnehmerinnen".
- GENDERED MALE: the translated seed noun has masculine grammatical gender and it refers specifically to individuals who identify as men, e.g., "Teilnehmer".
- GENDERED BOTH: the seed noun was translated by mentioning both feminine and masculine grammatical gender, e.g., "Teilnehmerinnen und Teilnehmer" or "Teilnehmer und Teilnehmerinnen".
- NON-BINARY: the translated seed noun specifically refers to non-binary individuals, e.g., using typographical solutions such as the Gendersternchen ("Teilnhemer*innen"), Genderdoppelpunkt ("Teilnehmer:innen), Underscore (Teilnehmer_innen) or gender-fair neosystems such as the ens-forms ("Studens").
- NEUTRAL (ALL): the translated seed noun has gender-neutral semantic gender when it refers to individuals of any gender, without specifying or implying gender, e.g., "die Teilnehmemenden" is used as a neutral alternative to generic masculine "die Teilnehmer", or where terms such as "Vorgesetzte" are a literal neutral translation of "supervisors".
- UNTRANSLATED: the seed noun has been left in English because it was part of a larger expression, e.g. "Assembly of Participants", that did not necessarily need to be translated.
- REWORDED: the seed noun has not been directly translated but the meaning has been preserved by using an adjective or a verb instead, e.g. "Sie haben am Turnier teilgenommen" instead of "Sie waren Teilnehmer des Turniers".
- ERROR: the seed noun was translated with a semantically wrong term or omitted.

Rules:
- Reply in JSON format with two fields: "translated_seed", which indicates the seed translation, and "gender," which indicates one of the allowed values listed above. If "gender" is ERROR, set "translated_seed" to ERROR.
- If "gender" is anything but "ERROR", the field "translation_seed" MUST be a phrase found in the German translation.
```

```
Seed: {seed}
Source: {source}
Translation: {translation}
```

Figure 8: **LLM-as-a-judge Prompt for gender evaluation prompt.** Bottom: Template of the user conversation turn used for in-context examples.

version), EuroLLM 9B (9B Instruct variant), and Vesuvius (Tower Vesuvius commercial system). For reference-free evaluation, we employed CometKiwi (`Unbabel/wmt23-cometkiwi-da-xxl`) and MetricX (`google/metricx-24-hybrid-xl-v2p6`). All inference was performed on NVIDIA A100 40GB GPUs using the vLLM inference engine with a temperature of 0 and greedy decoding.

Content Type Analysis. Table 6 shows that all models perform slightly worse on queer-related content compared to non-queer content, with a more substantial gap for NLLB 3.3B. The commercial Vesuvius system demonstrates the highest robustness across both content types, achieving the best scores in all metrics. Interestingly, when examining MetricX scores, we observe a larger performance disparity for NLLB 3.3B (2.221 point difference) compared to more advanced models like Gemma 3 27B (0.550), suggesting that larger,

Model	CometKiwi		MetricX	
	Queer	Non-queer	Queer	Non-queer
NLLB 3.3B	0.397	0.417	7.851	5.630
GEMMA 3 27B	0.815	0.865	1.935	1.385
EUROLLM 9B	0.815	0.868	1.920	1.378
VESUVIUS	**0.824**	**0.882**	**1.924**	**1.271**

Table 6: Translation quality on queer versus non-queer content, showing performance differences across content types. Higher CometKiwi and lower MetricX scores indicate better performance.

Model	CometKiwi		MetricX	
	Wikipedia	LM-generated	Wikipedia	LM-generated
NLLB 3.3B	0.361	0.478	8.026	3.410
GEMMA 3 27B	0.815	0.911	2.100	0.694
EUROLLM 9B	0.820	0.910	2.003	0.789
VESUVIUS	**0.835**	**0.922**	**1.916**	**0.700**

Table 7: Translation quality comparison between Wikipedia and LLM-generated content, highlighting variation in model robustness. Higher CometKiwi and lower MetricX scores indicate better performance.

more recent models handle diverse content more consistently.

Source Type Impact. As shown in Table 7, all models perform significantly better on LM-generated content compared to Wikipedia passages. This pattern aligns with our dataset construction process, where synthetic examples were deliberately created to be more straightforward. The most dramatic difference appears in NLLB 3.3B's MetricX scores (8.026 for Wikipedia vs. 3.410 for LM-generated), indicating that older specialized MT models struggle particularly with naturalistic examples. The performance gap narrows for larger instruction-tuned models, suggesting they better handle the complexities of natural language.

Ambiguity Handling. Table 8 reveals that, contrary to what might be expected, ambiguous and non-ambiguous contexts yield remarkably similar translation quality scores across all models. This surprising finding suggests that MT systems prioritize general fluency and adequacy over gender-specific disambiguation challenges. VESUVIUS maintains the highest consistency between both context types, while NLLB 3.3B shows a slight preference for ambiguous contexts—possibly because ambiguous contexts allow the model to default to masculine forms without contradicting explicit gender cues.

These detailed analyses complement the main findings presented in Section 5.1, providing deeper insights into how different content characteristics affect translation quality across model architectures.

C.2 LLM-as-Judge Evaluations

To facilitate evaluation and reduce the need for manual annotation on future research, we explored whether LLM-as-judge approaches can reliably automate the evaluation of gender phenomena in our dataset. We conducted two distinct experiments focusing on different aspects of gender evaluation.

C.2.1 Ambiguity Detection

First, we investigated whether language models could accurately determine if the gender of a seed noun is ambiguous in a given context. We evaluated OLMO 2 INSTRUCT and QWEN 2.5 72B on their ability to detect gender ambiguity in our English source passages.

Setup. We randomly selected 100 instances stratifying by the seed, manually annotated them for gender ambiguity, and used this gold standard to evaluate the models. Each model was prompted to classify passages as either ambiguous or unambiguous with respect to the gender of the specified seed noun. For unambiguous cases, models were

Model	CometKiwi		MetricX	
	Ambiguous	Non-ambiguous	Ambiguous	Non-ambiguous
NLLB 3.3B	0.435	0.398	5.712	6.056
GEMMA 3 27B	0.861	0.858	1.409	1.488
EUROLLM 9B	0.867	0.857	1.412	1.470
VESUVIUS	**0.876**	**0.876**	**1.340**	**1.349**

Table 8: Translation performance on ambiguous versus non-ambiguous content, showing model behavior with different contextual clarity. Higher CometKiwi and lower MetricX scores indicate better performance.

Model	Ambiguous	Unamb (F)	Unamb (M)	Unamb (Both)	Overall
OLMo 2	0.729	0.625	0.444	0.381	0.653
Qwen 2.5	**0.818**	**0.870**	**0.581**	**0.545**	**0.781**

Table 9: Performance comparison of LLMs for gender ambiguity detection. The table shows weighted F1 scores for ambiguous passages and for each unambiguous category (Female, Male, and Both). Bold values indicate best performance.

further asked to identify the specific gender (Female, Male, or Both/All genders).

Results. As shown in Table 9, QWEN 2.5 72B significantly outperformed OLMO2 across all categories, achieving an overall weighted F1 score of 0.781 compared to OLMO 2 INSTRUCT's 0.653. QWEN 2.5 72B demonstrated particularly strong performance in identifying female gender contexts (F1: 0.870), while both models showed comparative weakness in recognizing mixed-gender ("Both") contexts, suggesting this remains a challenging task for current LLMs.

C.2.2 Translation Gender Form Classification

In our second experiment, we evaluated the ability of various LLMs to classify the gender form used in German translations, comparing them against human annotations.

Setup. We prompted four models—GEMMA 3 27B, QWEN 2.5 72B, QWEN 3 32B, and GPT-4O—to classify the gender form of seed nouns in German translations according to eight categories: masculine, feminine, both (coordination), non-binary (typographical solutions), neutral (all), untranslated, reworded, or error. We tested each model with and without providing the source English passage alongside the German translation.

We used in-context learning by providing eight randomly shuffled examples (one for each possible label) and utilized guided decoding to ensure valid JSON output format. For open-weight models, we used a temperature of 0, while for GPT-4O we used default parameters.

Results. As shown in Table 10, GPT-4O significantly outperformed all open-weight models, achieving a weighted F1 score of 0.877 and recall of 0.859 when provided with both source and target text. This performance was consistent across different source ambiguity types, with particularly strong performance on unambiguous "all gender" contexts (recall: 0.907).

Among open-weight models, QWEN 2.5 72B showed the strongest performance (F1: 0.732, recall: 0.664), followed by QWEN 3 32B and GEMMA 3 27B. Interestingly, including the source text alongside the translation improved performance for GPT-4O and GEMMA 3 27B, but had mixed effects on the Qwen models.

All models demonstrated better performance on unambiguous male contexts than on ambiguous or unambiguous female/all-gender contexts, suggesting persistent challenges in recognizing and evaluating gender-fair language alternatives.

The results were consistent across translations from both VESUVIUS and EUROLLM 9B, confirming the robustness of our evaluation approach. These findings highlight that while commercial LLMs can serve as reliable judges for gender phenomena in translation, open-weight models still lag behind significantly in this capability.

Judge	F1	Recall	Amb	Unamb (F)	Unamb (M)	Unamb (All)
GEMMA 3 27B	0.5520	0.4614	0.3857	0.5132	0.6621	0.2791
+ source	0.6283	0.5386	0.4970	0.6085	0.6484	0.3488
QWEN 2.5 72B	0.7337	0.6680	0.6223	0.7037	0.8447	0.4070
+ source	0.7332	0.6640	0.6302	0.6984	0.8174	0.3953
QWEN 3 32B	0.7590	0.7001	0.6640	0.7460	0.7900	0.5814
+ source	0.7046	0.6108	0.5626	0.6825	0.6849	0.5465
GPT-4O	0.8314	0.8044	**0.8012**	0.7778	**0.8813**	0.6860
+ source	**0.8533**	**0.8295**	0.7833	**0.8836**	0.8676	**0.8837**
GEMMA 3 27B	0.6044	0.5256	0.4433	0.6667	0.6712	0.3256
+ source	0.5703	0.4865	0.3857	0.6720	0.5936	0.3953
QWEN 2.5 72B	0.7423	0.6810	0.6441	0.7566	0.7945	0.4419
+ source	0.7232	0.6640	0.6083	0.7407	0.8174	0.4302
QWEN 3 32B	0.7786	0.7202	0.6938	0.7937	0.7534	0.6279
+ source	0.6630	0.5587	0.5109	0.5926	0.6530	0.5233
GPT-4O	0.8450	0.8185	0.8151	0.7989	0.8676	0.7558
+ source	**0.8769**	**0.8586**	**0.8410**	**0.8571**	**0.8813**	**0.9070**

Table 10: **Full results of LLM Judges.** Weighted F1 and Recall on GLITTER (left) and separately by ambiguous (Amb) and unambiguous (Unamb) scenarios for the English source. Results against human annotations on EUROLLM (top) and VESUVIUS (bottom).

D Gender-Fair Prompting Experiments

This appendix section provides a detailed expansion of the gender-fair translation experiments discussed in Section 6. We present a comprehensive analysis of prompting strategies to elicit gender-fair translations from English to German.

D.1 Experimental Setup Details

We evaluated three multilingual LLMs of increasing size on their ability to produce gender-fair translations: EUROLLM 9B (`utter-project/eurollm-7b-instruct`), GEMMA 3 27B (`google/gemma-3-27b-it`), QWEN 2.5 72B (`Qwen/Qwen2.5-72B-Instruct`).

All experiments were conducted on 4x NVIDIA A6000 GPUs using the vLLM inference engine (Kwon et al., 2023) with the Hugging Face Transformers library (Wolf et al., 2020). Generation was performed with greedy decoding (temperature=0) to ensure reproducibility, with a maximum sequence length of 2048 tokens and a maximum new token count of 1024.

For automatic gender classification of translations, we employed GPT-4.1 as a judge (as described in Section 5.3). Quality evaluation utilized both CometKiwi (Rei et al., 2023) and MetricX (Juraska et al., 2023) metrics.

D.2 Detailed Results Analysis

Here's the fixed table with corrected model names, consistent bolding, and more compact headers:

Table 11 provides a comprehensive view of gender form distribution and translation quality across models and prompting strategies. Several key observations emerge:

Size-dependent gender form preferences: EUROLLM 9B (9B) shows a strong preference for gendered masculine forms (62% in zero-shot) with minimal use of non-binary forms (2%). In stark contrast, Qwen (72B) defaults to non-binary forms (74% in zero-shot) with very limited masculine gender usage (6%). Gemma (27B) sits between these extremes but still heavily favors non-binary forms (65%).

Impact of prompting on model behavior: For EuroLLM, the iterative approach most effectively reduces masculine forms (from 62% to 37%) while increasing non-binary forms (from 2% to 32%), suggesting smaller models are more malleable to prompting. For Gemma and Qwen, while prompting does shift distribution somewhat, the effect is less dramatic and doesn't fundamentally change their preference for non-binary forms.

Female gender representation: Notably, all models use female gender forms at very low rates (1-13%) across all prompting strategies. This suggests that models either default to masculine (EuroLLM) or bypass binary gender entirely for gender-neutral/non-binary alternatives (Gemma, Qwen).

Translation quality: CometKiwi scores remain consistently high (0.81-0.88) across all models and strategies, with EUROLLM 9B often achieving the highest quality in zero-shot settings (0.88). MetricX scores show more variation, with Gemma typ-

Strategy	Model	Male	Female	Neutral	NB	MeX ↓	CoKw ↑
ZS	EuroLLM 9B	0.62	0.13	0.15	0.02	**1.46**	**0.88**
	Gemma 3 27B	0.12	0.02	0.17	0.65	1.68	0.86
	Qwen 2.5 72B	0.06	0.01	0.13	0.74	3.14	0.81
FS	EuroLLM 9B	0.49	0.12	0.14	0.20	**1.59**	**0.87**
	Gemma 3 27B	0.17	0.03	0.20	0.55	1.67	**0.87**
	Qwen 2.5 72B	0.07	0.01	0.26	0.61	2.44	0.83
CoT	EuroLLM 9B	0.51	0.10	0.13	0.20	**1.79**	0.86
	Gemma 3 27B	0.13	0.04	0.15	0.63	1.88	0.86
	Qwen 2.5 72B	0.05	0.01	0.13	0.78	2.25	**0.85**
CFS	EuroLLM 9B	0.58	0.13	0.15	0.06	3.31	0.84
	Gemma 3 27B	0.18	0.05	0.20	0.51	**1.53**	**0.87**
	Qwen 2.5 72B	0.08	0.01	0.27	0.58	2.26	0.84
CoT-MT	EuroLLM 9B	0.37	0.07	0.15	0.32	2.26	0.84
	Gemma 3 27B	0.15	0.09	0.23	0.43	**1.44**	**0.87**
	Qwen 2.5 72B	0.14	0.02	0.09	0.73	1.64	**0.87**

Table 11: Performance comparison of gender-fair translation strategies (ZS: Zero-shot, FS: Few-shot, CFS: Contrastive, CoT: Chain-of-Thought, CoT-MT: Multi-turn CoT) across models. **Male** and **Female** columns show the proportion of translations using gendered expressions. **Neutral** shows the proportion of gender-neutral translations, while **NB** shows the proportion of non-binary translations. **MeX** (MetricX score, lower is better) and **CoKw** (CometKiwi score, higher is better) measure translation quality. Bold values indicate the best performance for each quality metric within each strategy.

ically yielding the fewest errors. Importantly, there is no clear quality penalty for gender-fair translations.

Table 12 breaks down performance by source type and ambiguity status. This reveals several additional insights:

Natural vs. synthetic content: All models demonstrate significantly higher gender-fair performance on synthetic (LM-generated) content compared to natural content. This effect is most dramatic with EuroLLM, which produces gender-fair translations for only 18.3% of unambiguous natural content (zero-shot) but 63.9% of unambiguous synthetic content (iterative). This suggests syntactic complexity and linguistic patterns in natural text may reinforce gendered translation patterns.

Ambiguity sensitivity: Ideally, models should use gender-fair forms for ambiguous content but respect gender cues in unambiguous content. However, the data shows that larger models fail to make this distinction. For example, Qwen (zero-shot) produces gender-fair translations for 87.9% of ambiguous natural content but also for 84.9% of unambiguous natural content, demonstrating it overrides explicit gender cues in favor of gender-fair defaults.

Prompting effectiveness across contexts: The CoT-MT approach shows particularly strong improvements for EuroLLM 9B with synthetic content, increasing gender-fair translations from 21.5% to 74.3% for ambiguous synthetic passages. However, even this most effective strategy only increases gender-fair translations to 31.3% for unambiguous natural content, suggesting entrenched patterns are difficult to overcome in complex text.

Table 13 offers a novel analysis of how queer-related content influences gender-fair translation patterns:

Queer content bias: Across all models and most prompting strategies, queer-related content receives more gender-fair translations than non-queer content in unambiguous scenarios. For example, with Gemma (CoT strategy), unambiguous queer content receives gender-fair translations 88% of the time compared to only 64% for non-queer content. This suggests models have learned associations between LGBTQ+ topics and gender-inclusive language.

Resilient patterns across prompting: The

		Overall		Amb. (nat.)		Unamb. (nat.)		Amb. (Synt.)		Unamb. (Synt.)	
Strategy	Model	**GF**	**GD**	**GF**	**GD**	**GF**	**GD**	**GF**	**GD**	**GF**	**GD**
ZS	EuroLLM 9B	22.1	**77.9**	29.4	70.6	18.3	**81.7**	21.5	78.5	17.6	**82.4**
	Gemma 3 27B	85.3	14.7	76.8	23.2	78.4	21.6	94.4	5.6	96.3	3.7
	Qwen 2.5 72B	91.5	8.5	87.9	12.1	84.9	15.1	**98.6**	1.4	97.7	2.3
FS	EuroLLM 9B	38.4	61.6	33.6	66.4	26.6	73.4	54.7	45.3	44.0	56.0
	Gemma 3 27B	78.2	21.8	68.5	31.5	63.3	36.7	95.3	4.7	93.5	6.5
	Qwen 2.5 72B	90.3	9.7	86.2	13.8	84.5	15.5	95.8	4.2	97.7	2.3
CFS	EuroLLM 9B	25.8	74.2	28.4	71.6	20.7	79.3	27.1	72.9	27.8	72.2
	Gemma 3 27B	75.1	24.9	64.7	35.3	62.7	37.3	94.9	5.1	85.2	14.8
	Qwen 2.5 72B	89.5	10.5	83.4	16.6	84.1	15.9	97.2	2.8	97.2	2.8
CoT	EuroLLM 9B	37.2	62.8	29.1	70.9	21.6	78.4	59.3	40.7	46.3	53.7
	Gemma 3 27B	81.8	18.2	74.0	26.0	68.3	31.7	98.1	1.9	93.5	6.5
	Qwen 2.5 72B	93.3	6.7	**88.2**	11.8	91.0	9.0	98.1	1.9	98.1	1.9
CoT-MT	EuroLLM 9B	51.2	48.8	43.6	56.4	31.3	68.7	74.3	25.7	63.9	36.1
	Gemma 3 27B	71.1	28.9	72.0	28.0	60.8	39.2	85.0	15.0	69.4	30.6
	Qwen 2.5 72B	84.5	15.5	78.5	21.5	74.8	25.2	94.9	5.1	94.4	5.6

Table 12: Gender-fair (GF) versus gender specific (GD) translation percentages across prompting strategies (ZS: Zero-shot, FS: Few-shot, CFS: Contrastive, CoT: Chain-of-Thought, CoT-MT: Multi-turn CoT) and data types. For ambiguous contexts, higher GF values are better (highest bolded). For unambiguous contexts, higher GD values can be appropriate (highest bolded). Results are separated by context ambiguity (Amb./Unamb.) and data source (natural/synthetic).

		Overall		Amb. (Q)		Unamb. (Q)		Amb. (Not Q)		Unamb. (Not Q)	
Strategy	Model	**GF**	**GD**	**GF**	**GD**	**GF**	**GD**	**GF**	**GD**	**GF**	**GD**
ZS	EuroLLM 9B	24.0	76.0	25.0	75.0	20.0	**80.0**	30.9	69.1	18.0	**82.0**
	Gemma 3 27B	77.6	22.4	80.6	19.4	88.0	12.0	75.6	24.4	76.3	23.7
	Qwen 2.5 72B	86.4	13.6	90.3	9.7	92.0	8.0	87.1	12.9	83.3	16.7
FS	EuroLLM 9B	30.2	69.8	29.2	70.8	38.0	62.0	35.0	65.0	24.1	75.9
	Gemma 3 27B	66.0	34.0	70.8	29.2	68.0	32.0	67.7	32.3	62.3	37.7
	Qwen 2.5 72B	85.4	14.6	91.7	8.3	**94.0**	6.0	84.3	15.7	82.5	17.5
CFS	EuroLLM 9B	24.6	75.4	26.4	73.6	20.4	79.6	29.0	71.0	20.7	79.3
	Gemma 3 27B	63.7	36.3	68.1	31.9	73.5	26.5	63.6	36.4	60.4	39.6
	Qwen 2.5 72B	83.7	16.3	86.1	13.9	87.8	12.2	82.5	17.5	83.3	16.7
CoT	EuroLLM 9B	25.4	74.6	23.6	76.4	28.0	72.0	30.9	69.1	20.2	79.8
	Gemma 3 27B	71.3	28.7	75.0	25.0	88.0	12.0	73.7	26.3	64.0	36.0
	Qwen 2.5 72B	89.6	10.4	**90.3**	9.7	94.0	6.0	87.6	12.4	90.4	9.6
CoT-mt	EuroLLM 9B	37.6	62.4	37.5	62.5	32.0	68.0	45.6	54.4	31.1	68.9
	Gemma 3 27B	66.5	33.5	72.2	27.8	68.0	32.0	71.9	28.1	59.2	**40.8**
	Qwen 2.5 72B	76.7	23.3	72.2	27.8	80.0	20.0	**80.6**	19.4	73.7	26.3

Table 13: Gender-fair (GF) and gendered (GD) percentages for queer-related (Q) and non-queer content across prompting strategies (ZS: Zero-shot, FS: Few-shot, CFS: Contrastive, CoT: Chain-of-Thought, CoT-mt: Multi-turn CoT). For ambiguous contexts, higher GF values are better (highest bolded). For unambiguous contexts, higher GD values can be appropriate (highest bolded).

Strategy	Model	BLEU		chrF		COMET	
		Avg	Max	Avg	Max	Avg	Max
ZS	EuroLLM 9B	**0.586**	**0.597**	**80.26**	**80.88**	**0.875**	**0.879**
	Gemma 3 27B	0.476	0.495	74.46	75.27	0.866	0.871
	Qwen 2.5 72B	0.414	0.431	72.08	72.87	0.837	0.843
FS	EuroLLM 9B	**0.587**	**0.602**	**80.55**	**81.29**	**0.875**	**0.880**
	Gemma 3 27B	0.500	0.518	75.78	76.58	0.867	0.872
	Qwen 2.5 72B	0.452	0.469	73.56	74.36	0.854	0.858
CFS	EuroLLM 9B	**0.554**	**0.565**	**78.23**	**78.87**	**0.857**	**0.861**
	Gemma 3 27B	0.510	0.527	76.15	76.93	0.867	0.872
	Qwen 2.5 72B	0.460	0.477	73.79	74.57	0.856	0.861
CoT	EuroLLM 9B	**0.579**	**0.593**	**79.91**	**80.62**	**0.872**	**0.876**
	Gemma 3 27B	0.510	0.528	76.50	77.33	0.866	0.871
	Qwen 2.5 72B	0.469	0.490	74.59	75.50	0.859	0.864
CoT-MT	EuroLLM 9B	**0.576**	**0.591**	**79.52**	**80.25**	**0.867**	**0.872**
	Gemma 3 27B	0.520	0.536	76.78	77.53	0.870	0.875
	Qwen 2.5 72B	0.491	0.511	75.63	76.49	0.868	0.873

Table 14: Reference-based translation quality metrics across prompting strategies and models, using professional post-edits as references. For each strategy, the highest values for each metric are in bold. Strategies: ZS: Zero-shot, FS: Few-shot, CFS: Contrastive, CoT: Chain-of-Thought, CoT-MT: Multi-turn CoT.

queer content effect persists across prompting strategies, demonstrating a robust learned association rather than a superficial response to instructions. Even in EuroLLM, which heavily favors gendered forms overall, unambiguous queer content receives gender-fair translations at higher rates (20-38%) than non-queer content (18-31%).

Model size and topical sensitivity: Larger models show stronger differentiation between queer and non-queer content. Qwen (few-shot) shows a 11.5 percentage point difference in gender-fair translation rates between unambiguous queer content (94%) and non-queer content (82.5%), while EuroLLM 9B shows a 13.9 point difference.

D.3 Detailed Prompts

Below are the exact prompts used for each strategy in our experiments:

Zero-Shot (ZS) Prompt

You are an experienced translator specializing in gender-fair language. In the following English passage, different word classes such as nouns and adjectives need to be inflected for gender.

Translate the passage into German, using a gender-fair strategy when the gender identity of the referents in the passage is unknown, non-binary or it encompasses more than one gender identity. When translating, you might want to modify gendered referents by either using the gender star (*), ending with -ens, rephrasing with participles, using inherently neutral words, or keeping unchanged if the gender context is already specified in the sentence. Provide the translation only.

English passage:

Few-Shot (FS) Prompt

You are an experienced translator specializing in gender-fair language. In the following English passage, different word classes such as nouns and adjectives need to be inflected for gender.

Translate the passage into German, using a gender-fair strategy when the gender identity of the referents in the passage is unknown, non-binary or it encompasses more than one gender identity. When translating, you might want to modify gendered referents by either using the gender star (*), ending with -ens, rephrasing with participles, using inherently neutral words, or keeping unchanged if the gender context is already specified in the sentence.

Here are a few examples on how you can translate an English passage into German with said strategies.

English: ”The ascent challenges both skill and endurance. Weather conditions added another layer of complexity. Experienced mountaineers undertook the task despite the risks involved. Their journey will serve as inspiration for future climbers.”

Gender star: ”Der Aufstieg stellt sowohl Geschick als auch Ausdauer auf die Probe. Die Wetterbedingungen fügten eine weitere Komplex-

itätsebene hinzu. Erfahrene Bergsteiger*innen unternahmen die Aufgabe trotz der damit verbundenen Risiken. Ihre Reise wird als Inspiration für zukünftige Bergsteiger*innen dienen."

Ens-forms: "Der Aufstieg stellt sowohl Geschick als auch Ausdauer auf die Probe. Die Wetterbedingungen fügten eine weitere Komplexitätsebene hinzu. Erfahren Bergsteigens unternahmen die Aufgabe trotz der damit verbundenen Risiken. Ihrens Reise wird als Inspiration für zukünftig Bergsteigens dienen."

Gender-neutral rewording: "Der Aufstieg stellt sowohl Geschick als auch Ausdauer auf die Probe. Die Wetterbedingungen fügten eine weitere Komplexitätsebene hinzu. Erfahrene Bergsteigende unternahmen die Aufgabe trotz der damit verbundenen Risiken. Ihre Reise wird als Inspiration für zukünftige Bergsteigende dienen."

Now translate the following English passage and provide only the gender-fair German translation using any of the strategies above.

Contrastive Few Shot (CFS) Prompt

You are an experienced translator specializing in gender-fair language. In the following English passage, different word classes such as nouns and adjectives need to be inflected for gender.

Translate the passage into German, using a gender-fair strategy when the gender identity of the referents in the passage is unknown, non-binary or it encompasses more than one gender identity. When translating, you might want to modify gendered referents by either using the gender star (*), ending with -ens, rephrasing with participles, using inherently neutral words, or keeping unchanged if the gender context is already specified in the sentence.

Consider this example:

English: "The ascent challenges both skill and endurance. Weather conditions added another layer of complexity. Experienced mountaineers undertook the task despite the risks involved. Their journey will serve as inspiration for future climbers."

The standard translation uses the masculine generic and is: "Der Aufstieg stellt sowohl Geschick als auch Ausdauer auf die Probe. Die Wetterbedingungen fügten eine weitere Komplexitätsebene hinzu. Erfahrene Bergsteiger unternahmen die Aufgabe trotz der damit verbundenen Risiken. Ihre Reise wird als Inspiration für zukünftige Bergsteiger dienen."

However, there are no clues as to what the gender identity of the referents is. To avoid linguistic sexism, gender-fair language should be preferred over generic masculine.

A gender-fair version might look like any of this:

Gender star: "Der Aufstieg stellt sowohl Geschick als auch Ausdauer auf die Probe. Die Wetterbedingungen fügten eine weitere Komplexitätsebene hinzu. Erfahrene Bergsteiger*innen unternahmen die Aufgabe trotz der damit verbundenen Risiken. Ihre Reise wird als Inspiration für zukünftige Bergsteiger*innen dienen."

Ens-forms: "Der Aufstieg stellt sowohl Geschick als auch Ausdauer auf die Probe. Die Wetterbedingungen fügten eine weitere Komplexitätsebene hinzu. Erfahren Bergsteigens unternahmen die Aufgabe trotz der damit verbundenen Risiken. Ihrens Reise wird als Inspiration für zukünftig Bergsteigens dienen."

Gender-neutral rewording: "Der Aufstieg stellt sowohl Geschick als auch Ausdauer auf die Probe. Die Wetterbedingungen fügten eine weitere Komplexitätsebene hinzu. Erfahrene Bergsteigende unternahmen die Aufgabe trotz der damit verbundenen Risiken. Ihre Reise wird als Inspiration für zukünftige Bergsteigende dienen."

Now translate the following English passage and provide only the gender-fair German translation using any of the strategies above.

Chain-of-Thought (CoT) Prompt

Translate the following English passage into German. The passage contains different word classes such as nouns and adjectives that need to be inflected for gender. When translating the passage into German, use a gender-fair strategy when the gender identity of the referents in the passage is unknown, non-binary or it encompasses more than one gender identity. When translating, you might want to modify gendered referents by either using the gender star (*), ending with -ens, rephrasing with participles, using inherently neutral words, or keeping unchanged if the gender context is already specified in the sentence. Let's think it through, considering the following example:

English text: The ascent challenges both skill and endurance. Weather conditions added another layer of complexity. Experienced mountaineers undertook the task despite the risks involved. Their journey will serve as inspiration for future climbers.

SEED WORD: The noun in this passage is

”mountaineers”.

REASONING: In the passage, there are no clues as to what the gender of ”mountaineers”. In such cases, the generic masculine ”Bergsteiger” is a common translation in German. This, however, is an example of gender-exclusive language. To avoid exclusion, one could opt for gender-fair alternatives such as gender star as in ”Bergsteiger*innen”, ens-forms as in ”Bergsteigens”, or a neutral alternative as in ”Bergsteigende”.

FINAL TRANSLATION: Der Aufstieg stellt sowohl Geschick als auch Ausdauer auf die Probe. Die Wetterbedingungen fügten eine weitere Komplexitätsebene hinzu. Erfahrene Bergsteiger*innen unternahmen die Aufgabe trotz der damit verbundenen Risiken. Ihre Reise wird als Inspiration für zukünftige Bergsteiger*innen dienen.

Now, translate the following English passage into German. First, identify any words that might need gender-fair translations in German. Then provide gender-fair alternatives for each identified term (if needed), and create a complete German translation. English text:

Chain-of-Thought Iterative (CoT-MT) Prompt

Turn 1: Translate the following passage into German. Consider this example:

English: ”The ascent challenges both skill and endurance. Weather conditions added another layer of complexity. Experienced mountaineers undertook the task despite the risks involved. Their journey will serve as inspiration for future climbers.”

The standard translation is: ”Der Aufstieg stellt sowohl Geschick als auch Ausdauer auf die Probe. Die Wetterbedingungen fügten eine weitere Komplexitätsebene hinzu. Erfahrene Bergsteiger unternahmen die Aufgabe trotz der damit verbundenen Risiken. Ihre Reise wird als Inspiration für zukünftige Bergsteiger dienen.”

Now translate the following passage into German.

Turn 2: Consider the English passage: ’text_to_translate’ and your German translation: ’initial_translation’.

In the English passage, identify any words or phrases (seed words) that might require gender-fair language considerations when translated into German. For example, ”mountaineers” could be translated to ”Bergsteiger” (masculine), ”Bergsteiger*innen” (gender star), ”Bergsteigens” (ens-form), or ”Bergsteigende” (neutral participle).

For your identified seed words from the original English text, explain your reasoning for the gender-fair choices or why a particular form was chosen in your initial translation, and list potential gender-fair alternatives if applicable. Structure your response clearly. Respond in under 150 words.

Turn 3: Based on the original English passage ’text_to_translate’, your initial translation, and your reasoning about gender-fair alternatives: ’reasoning_and_alternatives’.

Now, provide the final, revised German translation incorporating the most appropriate one. For example, if ”Bergsteiger” was identified and ”Bergsteiger*innen” was chosen as a better alternative because the gender cue is ambiguous, use that in the final translation.

Respond only with the complete final German translation, no other text.

E AI Assistant Statement

We used GitHub’s Copilot and Claude Code to support coding experiments and Claude 3.7 Sonnet for lightweight editing and rephrasing in the manuscript.