Non-Archival Abstract

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

A Pranav
University of Hamburg, Germany

Overview

English-to-German translation systems show bias towards masculine generic forms even when source passages specify women or non-binary referents (Choubey et al., 2021; Savoldi et al., 2022; Lardelli et al., 2024). This bias intensifies in multi-sentence input, where gender cues sit outside the sentence carrying the translated noun and current systems fail to track them. We address this with GLITTER, a benchmark of 909 multi-sentence English passages paired with 1,785 professionally post-edited German references, each source provided in three gender-fair target forms (neutral rewording, gender star, -ens form).

Resource Creation. We construct GLITTER (Figure 1) from 115 gender-ambiguous English plural nouns (e.g., *deputies*), drawing source passages from two streams: filtered Wikipedia text and human-validated synthetic generations. Each passage spans four sentences, with disambiguating gender cues placed at different positions relative to the seed term so that contextual reasoning is measurable. Professional translators post-edit each passage into three German references: a neutral rewording (e.g., *Beratungspersonen*), a gender-star form (*Berater*innen*), and an -ens form (*Beratens*); annotators add span-level labels for referential gender and ambiguity.

Results

We benchmark EUROLLM, GEMMA 3 27B, QWEN 2.5 72B, and GPT-4.1 under five prompting strategies. We use GPT-4.1 as an LLM-as-judge after validating it against human labels (F1 0.88), and report the percentage of gender-specific translations per system.

LMs fail at choosing when to use gender-fair forms. EUROLLM often consistently uses gendered forms (72–81%); GEMMA 3 27B and QWEN 2.5 72B overly rely on gender-fair forms (75–92%).

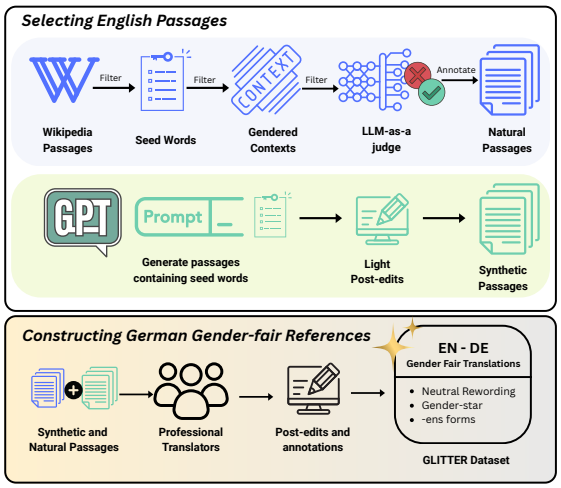

Figure 1: Construction of GLITTER: natural and synthetic English passages are filtered, annotated for gender phenomena, and post-edited into three parallel German gender-fair references.

LMs struggle more with disambiguating gender in synthetic passages. With multi-turn chain-of-thought prompting, EUROLLM's gendered outputs drop from 82.4% to 36.1% on synthetic passages, but only from 81.7% to 68.7% on natural ones.

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. GEMMA 3 27B uses gendered forms for only 12–32% of queer passages compared to 23–41% for non-queer content.

Conclusion. GLITTER brings several contributions 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.

References

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.

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.

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.