DEval-MT: An Intersectional Gender Bias Evaluation Framework for German Machine Translation

Michelle Kappl[1], Leon Schreiber[1], Stefan Hillmann[1]

[1]Technische Universität Berlin

Correspondence: michelle.kappl@tu-berlin.de

Abstract

Machine translation systems can reproduce gender bias, but most evaluation resources are limited to English as the source language. We introduce *DEval-MT*, a template-based framework for evaluating gender bias in German-source machine translation, comprising 6,618 sentences across six sentence styles with an intersectional template design that can be extended with additional demographic attributes and support for nonbinary neopronouns. Alongside it, we provide a Python package for translation, alignment, gender prediction, and analysis. Evaluating six different MT systems on seven target languages using this dataset and pipeline, we find evidence of representational harms across multiple demographic groups.

1 Introduction

Machine Translation (MT) systems have become an integral part of our daily lives, facilitating communication between languages (Vieira et al., 2021). However, research has shown that these systems are prone to reproducing and amplifying societal biases. One of the most extensively studied forms of bias in MT is gender bias. It refers to the systematic and unfair treatment of individuals based on their gender and is of rising interest within the research of natural language processing systems (Costa-jussà, 2019; Blodgett et al., 2020; Savoldi et al., 2025). As an example, the German sentence *Die Professorin liebt ihre Arbeit.*[1] unambiguously marks its subject as feminine, yet Google Translate (September 02, 2026) renders it into Spanish as: *El profesor ama su trabajo.*[2], defaulting to the masculine *el profesor* despite the clear indicator.

Most prior work on gender bias in MT has focused on English as the source language (Savoldi et al., 2025), leaving languages with grammatical gender on the source side, such as German, comparatively understudied. This is a notable gap as German morphology explicitly marks the gender of many role nouns (e.g. *Professor*$_m$ vs. *Professorin*$_f$), providing unambiguous gender information that MT systems could, in principle, use. To address this gap, we introduce DEval-MT, a template-based evaluation framework for gender bias in German.

Bias Statement (Blodgett et al., 2020). We address representational harms caused by gender bias in MT. When systems fail to preserve unambiguous gender and default to masculine forms, they reproduce and reinforce stereotypical beliefs about gender roles and abilities (Vervecken and Hannover, 2015; Horvath et al., 2016; Sterling et al., 2020). Children are especially vulnerable to this kind of bias: exposure to it can influence how difficult, prestigious, or attainable they judge a given career to be (Vervecken and Hannover, 2015). This directly influences people beyond childhood, as self-efficacy in STEM fields has been linked to the gender pay gap that persists in those occupations (Sterling et al., 2020). Beyond occupational stereotyping, biased translations also cause direct representational harm to trans, non-binary, and queer individuals. Misgendering by language technologies, including MT, is a well-documented harm that can intensify the difficulties trans and non-binary people already face in having their identities recognised and result in underrepresentation (Cao and Daumé III, 2021; Dev et al., 2021). Similarly, when systems assume that romantic relationships are heterosexual by default, they reproduce a heteronormative standard that renders same-gender relationships invisible.

Our Contributions:

- **DEval-MT**, a template-based, extensible German gender-bias testset of 6,618 sentences, built around occupational sentences. Beyond

[1]The professor$_f$ loves her work.

[2]The professor$_m$ loves their work.

Proceedings of the 7th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 181–192
November 9, 2026 ©2026 Association for Computational Linguistics

a binary-gendered baseline, it includes a non-binary neopronoun class, an intersectional template design ready for further demographic attributes, a version of each sentence with an explicit pronoun added, a heteronormativity probe, a stereotypicality predictor derived from German labour statistics, and gendered and gender-neutral first names as an additional gender-bias probe.

- An automatic evaluation pipeline released as an installable Python package designed to be extended with new languages, analysers, and datasets.
- An evaluation using our framework of six MT systems (Google Translate, Google Translate LLM, Microsoft Translator, DeepL, SYSTRAN, and GPT-4o) on seven typologically diverse target languages (Spanish, French, Italian, Arabic, Russian, Ukrainian, Hebrew), validated against human judgments.

2 Related Work

Gender bias in machine translation. Stanovsky et al. (2019) introduced WinoMT, the alignment-based evaluation pipeline this work builds on and finds substantial bias within English MT. Costa-jussà et al. (2023) confirm this bias across 50 languages, finding masculine translations score higher in nearly all gendered languages. Finally, Savoldi et al. (2025) survey a decade of MT bias research and find most benchmarks remain Anglocentric, with little coverage of morphologically rich languages. Kappl (2025) addresses part of this gap by adapting the WinoMT approach to German, establishing a masculine-default and stereotype-congruence bias for German MT.

Beyond a binary gender frame. Gender bias evaluation should not be limited to a binary frame. Cao and Daumé III (2021) make this case for coreference resolution, introducing a gender-inclusive benchmark covering singular they and neopronouns. Brandl et al. (2022) provide converging evidence at the model level: encoder LMs handle gender-neutral pronouns worse than gendered ones in Danish, English, and Swedish, with higher perplexity and lower accuracy.

Person names offer a way of probing social bias, since a name alone can carry inferred gender, race, and ethnicity without any explicit demographic marker in the source sentence. Wang et al. (2022) exploit this with a name-substitution mechanism and by appending demographic adjectives to certain gendered nouns (e.g. "Asian man") to investigate the intersectional effect of race and gender. Furthermore, Sandoval et al. (2023) use names as a proxy for race, showing that MT systems mistranslate and misgender referents based on social attributes implicit in names.

Heteronormativity. Bias against non-heterosexual relationships remains an under-explored aspect of gender bias research. Stewart and Mihalcea (2024) address this gap by constructing minimal-pair sentences that vary only a stated partner's gender (e.g. "wife" vs. "husband"), finding that same-gender pairings are systematically mistranslated more often and attributing this to an implicit heterosexual default rather than source-side ambiguity.

3 Methodology

3.1 DEval-MT Framework

We introduce DEval-MT, a German gender-bias evaluation framework, which is publicly available at michellekappl/DEval-MT (CC-BY 4.0) as an installable, easy-to-use Python package. It includes a German dataset comprised of 6,618 sentences for the evaluation of gender bias in different context scenarios and styles. Furthermore, it provides a pipeline for automatic translation, alignment, and gender annotation. Lastly, it includes a set of evaluation functions to quantify and visualise bias given the bias evaluation results. The framework is built to easily reproduce the results presented in this work and to extend it: dedicated tools allow users to add new templates, attributes, names, and occupation nouns, and to translate the dataset using the APIs of the MT systems under evaluation.

3.1.1 Dataset

The dataset consists of six sentence styles (Table 1), each built from templates that were written and annotated by a native German speaker and then reviewed by a second native speaker for correctness and naturalness. Each template centres on a single occupational title, which serves as the subject of interest across all sentence styles. For each sentence, we annotated the index of the subject, denoting the position within the sentence, and its gender label (M/F/D). The occupational titles were obtained from the Federal Employment Agency (Statistik der Bundesagentur für Arbeit, 2020). This list was

restricted to occupations with at least 1,000 employees whose job title is not a composite of three or more words. Lastly, two native German speakers independently checked the naturalness of the resulting list and, if possible, added gender-neutral phrasings following "geschickt gendern – das Genderwörterbuch"[3]. Furthermore, each sentence was annotated with the percentage of male vs. female employees in that occupation. This information was derived from German labour statistics by the Federal Employment Agency (Statistik der Bundesagentur für Arbeit, 2020).

Context	Description	M / F / D	n
1	baseline	530 / 530 / 58	1118
2	+ pronoun	530 / 530 / 58	1118
3	+ adjective	530 / 530 / 0	1060
4	+ adjective + pronoun	530 / 530 / 0	1060
5	romantic context	530 / 530 / 58	1118
6	name context	530 / 530 / 84	1144

Table 1: Overview of the six sentence Styles, with subject-gender gold label counts (M/F/D = masculine/feminine/diverse) and total sentences per Style.

Styles 1 & 2 (baseline / pronoun). Style 1 forms the baseline: a simple, everyday sentence stating an occupation-holder's action or preference, with the subject referred to only by the occupational noun and no pronoun. Styles 1 and 2 form a minimal pair as seen in (1): Style 2 adds an explicit pronoun referring back to the subject. We use male and female pronouns as well as the German neopronoun *dey/deren* for diverse subjects, as a study found it to be the most frequently self-chosen neopronoun among non-binary speakers (Verein für geschlechtsneutrales Deutsch e.V., 2022).

(1) Der Bauer liebt die Arbeit.
'The farmer$_{\text{m}}$ loves the work.'
Der Bauer liebt *seine* Arbeit.
'The farmer$_{\text{m}}$ loves *his* work.'

Styles 3 & 4 (intersectional adjectives). Style 3 mirrors Style 1 but adds an adjective to the subject (e.g. *Der **trans** Bauer liebt die Arbeit*[4]), so the sentence explicitly mentions an additional attribute following Wang et al. (2022). Style 4 mirrors Style 2 the same way, adding both the adjective and the pronoun. Styles 3 and 4 thus form a second minimal pair built from the same templates as Styles 1 and 2, but with an intersecting attribute slot that can later be filled with additional descriptors (nationality, race, ...).

Style 5 (romantic). Style 5 as seen in (2) places the occupation-holder in a romantic context: each sentence mentions both the subject and a romantic partner, whose gender is always explicitly stated, and their gold gender is annotated within the dataset. This sentence Style is based on the approach of Stewart and Mihalcea (2024).

(2) Der Bauer küsst seine Liebhaberin.
'The farmer$_{\text{m}}$ kisses his lover$_{\text{f}}$.'
Der Bauer küsst seinen Liebhaber.
'The farmer$_{\text{m}}$ kisses his lover$_{\text{m}}$.'

Style 6 (names). Style 6 as shown in (3) states the name of the occupation-holder: we use the most common male and female names as of 2024 (tagesschau.de, 2024) as well as six gender-neutral names (Bielefeld, 2025) following prior work on name-based gender-bias (Wang et al., 2022; Sandoval et al., 2023). A masculine name is paired with the masculine and, where available, the gender-neutral occupational noun. The feminine name is paired analogously. Each gender-neutral name is paired with all possible gendered forms of the occupational noun.

(3) Leon ist Bauer.
'Leon is a farmer$_{\text{m}}$.' (masc. name)
Bente ist Bauer.
'Bente is a farmer$_{\text{m}}$.' (neutral name)

As it is based on a templating approach, the dataset is designed to be further extended with additional templates, attributes, names, and occupational nouns.

3.1.2 Pipeline

The evaluation pipeline mainly builds on previous work by Stanovsky et al. (2019) and Kappl (2025) and includes three main steps: translation, alignment, and gender prediction.

Translation. Each DEval-MT sentence is translated into the target language using the MT systems under evaluation. This can be done automatically using custom translation functions, or existing translations can be attached directly.

Alignment. Each source word is mapped to its translation using `SimAlign` (Sabet et al., 2020),

[3] https://geschicktgendern.de
[4] The trans farmer loves the work.

rather than the `fast-align` (Dyer et al., 2013) used by Stanovsky et al. (2019). `SimAlign` aligns source and target words by the cosine similarity of their multilingual embeddings (e.g. mBERT), which does not need a parallel corpus, and outperforms statistical aligners for several languages, including German (Sabet et al., 2020). Using the annotated subject index together with the resulting alignment, the translated subject of interest is extracted from the target sentence, with its article and, if present, adjective.

Gender Prediction. Given the extracted subject phrase in the target sentence, the final step is to infer its grammatical gender. For languages with an available `spaCy` pipeline (German, French, Spanish, Italian, Russian, Ukrainian), its built-in morphologizer is used to analyse the tokens. For Arabic, the Qalsadi morphological analyser (Zerrouki, 2012) is utilised. Following the implementation from Stanovsky et al. (2019) we use a simple word-ending heuristic for Hebrew. All approaches return token-level gender features, which are then aggregated into a single gender label for the entire subject phrase by majority vote. In case of a tie, the label *unknown* is assigned. As with alignment, analysers can easily be swapped out or added to support further languages.

3.1.3 Evaluation

DEval-MT provides three main evaluation functions:

- `Error analysis` reports gender accuracy, Δ_G (a gender-gap metric introduced by Stanovsky et al. (2019)), and a fine-grained breakdown of error types
- `confusion matrix` reports per-gender precision, recall, and F1
- `logistic regression` relates a chosen covariate column, e.g. occupational stereotypicality, to prediction accuracy

Beyond these core functions, the package includes appropriate significance tests, comparison and summary scripts, and automated plotting of results, allowing users to easily reproduce the results presented in this work and to extend them with new analyses.

4 Experimental Setup

MT Systems. We evaluate six MT systems: Google Translate, Microsoft Translator and DeepL as traditional neural MT models, GPT-4o and Google Translate LLM as large language models and SYSTRAN, a hybrid model that combines rules with a statistical approach. Note that Google Translate and Google Translate LLM produce near-identical output for this dataset – only 33 of 4355 Spanish translations differ between the two.

Languages. We evaluate seven target languages: Spanish, French, Italian, Arabic, Russian, Ukrainian, and Hebrew, chosen to cover distinct language families (Romance, Slavic, Semitic) with different gendered grammatical structures.

Scope. Of the demographic attributes that the template system supports (§3), only *trans* is activated as an exemplary adjective in this study. The sentences with the gold label DIVERSE are excluded entirely from the results section, which is further discussed in Limitations.

5 Results

5.1 General Analysis

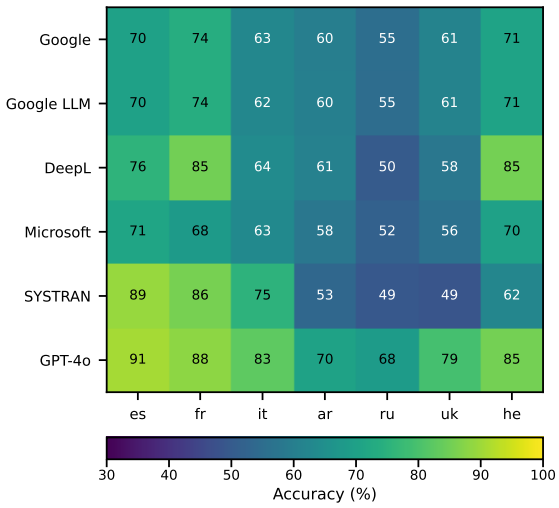


Figure 1: Translation accuracy by MT system and target language, pooled across all sentence styles.

Accuracy. We first assess how well a model preserves the original gender. The accuracy results, displayed in Figure 1, range from 49% to 91% and show considerable differences across models and languages. GPT-4o achieves the highest and most consistent accuracy, scoring between 68.3% and 90.7%. The traditional neural models range more widely, from 50.1% (DeepL, ru) to 85.3% (DeepL, he). SYSTRAN, the only hybrid model evaluated, shows the widest range (48.5%

– 89.1%) but performs particularly well in the Romance family, outperforming all neural models there. Accuracy is generally weakest within the Slavic languages, where traditional models average only ∼55%, barely above chance. Microsoft is the weakest system overall, averaging 62.6% across all seven languages.

Error Counts. Where translation errors occur, they overwhelmingly point in one direction: feminine-to-masculine misgendering dominates in 35 of 42 evaluated language-system pairs, illustrated in Figure 2 for Google, GPT-4o, and SYSTRAN (the remaining models show comparable patterns). The error rate is smallest for the Romance languages, where GPT-4o and SYSTRAN clearly outperform Google, and largest for the Slavic languages, with SYSTRAN misgendering 74% of feminine-source occupations as masculine – the largest error rate observed in this study. Arabic shows a similarly pronounced skew, accompanied by a higher rate of unknown predictions. The main exception is Hebrew, where the error direction reverses for 5 of the 6 systems.

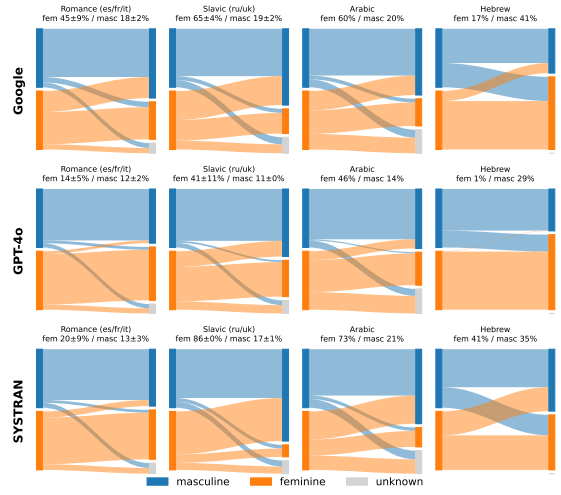


Figure 2: Gender flow from gold label to predicted translation, for Google, GPT-4o and SYSTRAN, pooled by language family. The orange band tracks feminine-gold instances and the blue band masculine-gold instances. Grey marks instances the pipeline could not resolve (UNKNOWN). Per-panel labels report the mean ± std feminine and masculine misclassification rate across languages within each grouping.

Stereotypicality. Logistic regression was used to measure the correlation between the stereotypicality of an occupation, which denotes the percentage of male vs. female employees, and the accuracy of its translation. In Table 2 the results of GPT-4o and Google are displayed. For GPT-4o, odds ratios range from 1.22 to 1.40 ($p < 0.001$) in 6 of 7 languages, indicating that more stereotype-congruent instances are translated more accurately. Hebrew reverses this relationship (OR=0.72, $p < 0.001$). Google's odds ratios are larger than GPT-4o's in every language except Hebrew, where the reversal effect is weaker (OR=0.94, $p = 0.023$) and range from 1.33 to 1.91.

Lang.	OR (GPT-4o)	OR (Google)
es	1.22***	1.90***
fr	1.26***	1.59***
it	1.34***	1.91***
ar	1.26***	1.33***
ru	1.40***	1.61***
uk	1.30***	1.51***
he	0.72***	0.94*

Table 2: Odds ratio (OR) of a correct translation given stereotype-congruence, per language, GPT-4o vs. Google. OR >1: more accurate when stereotype-congruent. Full report in 6. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Figure 3 further illustrates this pattern, plotting real-world vs. translated gender distribution per occupation category, pooled across all languages and sentence styles. The correlation is moderate (Pearson's $r = 0.50$, $p < 0.001$), with translations shifted towards male across nearly all categories. GPT-4o falls within the correct-translation range most often among the evaluated systems.

5.2 Effects of Contextual Style

Beyond the plain baseline (Style 1, see §3.1.1), two additional gender cues were tested: adding an explicit pronoun clause (Style 2) and explicitly marking the referent as trans (Style 3) and combining both (Style 4). The results are summarised in Figure 4.

Pronoun-Elaboration Effect. Adding an explicit pronoun clause (Style 2) improves accuracy over the plain baseline (Style 1) for every model evaluated. Google shows the largest gain, improving from 55% to 72% (+17.1 percentage points (pp)), while GPT-4o shows the smallest gain of the six systems, at just 0.7pp. Pooled across all models and languages, accuracy rises from 62.1% to 73.1%, an overall gain of 11.0pp.

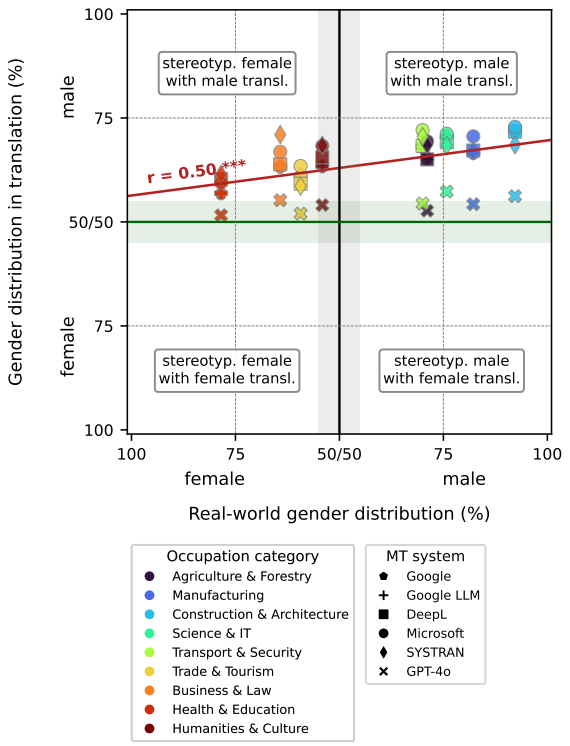


Figure 3: Real-world gender distribution of an occupation (x-axis) plotted against the gender distribution produced in translation for that occupation (y-axis), pooled across all six systems, seven languages, and all sentence styles. The green band marks the correct-translation margin around the 50/50 line as our dataset is balanced and the black line marks jobs with a 50/50 real-world gender distribution. The red line is the fitted trend (Pearson's r=0.50, p<0.001).

Trans-Identity Marking. Trans-marked sentences (Style 3) also beat the plain baseline, and, as with the pronoun clause, all models gain accuracy, though by a smaller margin (+2.3pp). The largest individual gains occur for GPT-4o (ru +14.2pp, uk +11.2pp, ar +12.1pp), concentrated in its weakest languages, while it slightly regresses in its strongest (es, fr, it). Combining trans-marking with the pronoun clause (Style 4) more than stacks the two effects: pooled accuracy reaches 78.2%, a +16.1pp gain over the baseline. Therefore, it is above the 13.3pp expected under a purely additive effect.

Gender-Neutral Names. Figure 5 holds the sentence and job form constant and varies only the name – a masculine or feminine name with a matching job title vs. a gender-neutral name with the same job title. Accuracy drops consistently once a neutral name is introduced. A masculine name paired with a masculine-form job noun reaches 72.8% accuracy, compared to 56.8% when the

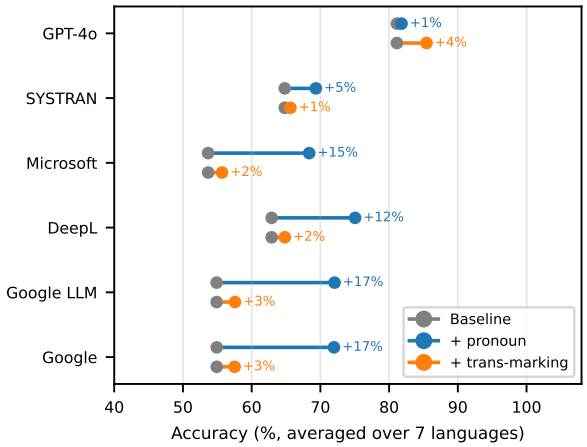


Figure 4: Accuracy, pooled over seven languages, for each MT system under the plain baseline (grey) versus with an added pronoun clause (blue) or trans-identity marking (orange). The labels indicate the accuracy gain in percentage points (pp).

name is neutral (-16.0pp). A feminine name paired with a feminine job noun reaches 54.2%, compared to 40.6% when the name is neutral (-13.6pp).

To estimate how strong a disambiguation cue a name is, we compare these figures against Styles 1 and 2. A gendered name closes part of the accuracy gap between masculine and feminine forms similarly to the pronoun clause, but less effectively. A masculine name falls below the masculine baseline (72.8% vs. 81.9%), while a feminine name improves on the feminine baseline (54.2% vs. 42.2%), but still trails the accuracy of 61.1% reached with the pronoun clause. A gender-neutral name, on the contrary, underperforms for all Styles: pooled over both job forms, neutral names reach only 48.7% accuracy, below the 62.1% pooled accuracy of the plain, pronoun-free baseline.

Heteronormativity Probe. Finally, we test whether accuracy on a grammatically unambiguous subject noun depends on whether the referenced partner is of a different or the same gender. Each subject sentence exists in both partner-gender versions, allowing a paired McNemar's test on matched pairs. The results are displayed in Table 3. The effect is positive and significant for most systems, which means that the accuracy is worse when the partner is of the same gender. The gap is most pronounced for DeepL and Google (+4.2pp) and smallest for SYSTRAN (+0.8pp). GPT-4o is the exception with no significant gap (-0.8pp, $p = 0.12$).

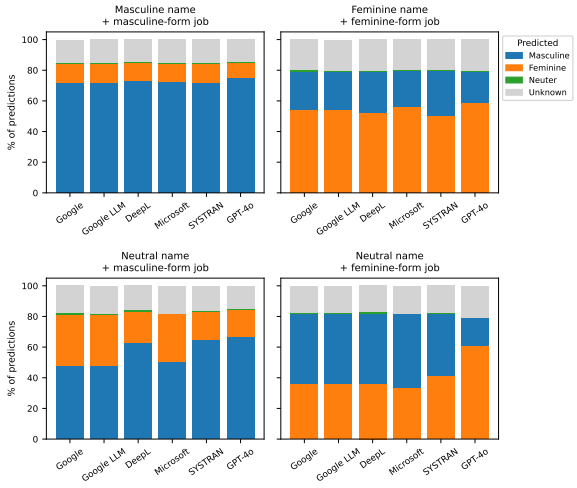


Figure 5: Distribution of predicted translation gender, pooled across all seven languages, for four name-job-form pairings. Each bar represents one system and the stacking shows the share of predictions resolved as masculine, feminine, neuter, or unknown. In an ideal system, the masculine-job and feminine-job panels would each be entirely one colour matching the job's gender.

5.3 Human Validation

To validate our approach described in §3 underlying the results above, linguistics students from *Freie Universität Berlin*, who were reimbursed according to German minimum wage, checked its alignment and gender-prediction steps for six of the seven target languages. We exclude Arabic as only one annotator completed the task. Annotation was carried out on a sample of 100 sentences per language, randomly drawn from Styles 1–4 only. The pipeline-extracted subject expression was confirmed correct by annotators 87.1% of the time, ranging from 99.0% for Spanish to 74.2% for Hebrew. The latter's low fidelity aligns with our results, where the Hebrew translation accuracy was often a striking outlier compared to other languages. We discuss this further in Limitations. To assess whether the pipeline is as reliable as a human annotator, we computed the same downstream system accuracy twice, once using the human majority-vote gender judgment and once using the pipeline's own prediction, each compared independently against the German gold label as seen in Table 4. Pooled across languages, the human-derived system accuracy reaches 82.1%, compared to 79.3% for the pipeline-derived accuracy, a gap of 2.8pp. This gap is not uniform across systems: GPT-4o's pipeline-derived accuracy matches its human-derived accuracy exactly, while SYSTRAN shows the largest divergence (+7.4pp). Following WinoMT's (Stanovsky et al., 2019) own validation statistic, human annotation compared directly to the pipeline's output yields 89.6% agreement – exceeding inter-annotator agreement (84.9% pairwise), so the pipeline diverges from human judgment no more than two annotators diverge from each other.

System	Diff.	Same	Gap
Google	74.4%	70.2%	+4.2***
DeepL	76.6%	72.4%	+4.2***
Microsoft	70.4%	68.0%	+2.4***
SYSTRAN	69.3%	68.5%	+0.8*
GPT-4o	82.8%	83.7%	-0.8

Table 3: Accuracy measured on the occupation noun when the romantic partner's gender matches (Same) vs. differs from (Diff.) the subject's gender, pooled over all 7 languages per system. Google Translate LLM performs near identical to Google Translate and is therefore excluded. Significance via McNemar's test on the matched pairs. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Lang.	n	Human	Pipeline	Gap (pp)
es	92	91.3%	91.%	+0.0
fr	93	91.4%	92.5%	-1.1
it	85	81.2%	78.8%	+2.4
ru	85	63.5%	62.4%	+1.2
uk	86	76.7%	73.3%	+3.5
he	96	86.5%	76.0%	+10.4
all	537	82.1%	79.3%	+2.8

Table 4: Per-language accuracy on the human-eval sample: human majority-vote gender judgment (Question 2) vs. the pipeline's own gold-vs-predicted accuracy, on the identical sentences. Both are scored against the same German gold, not against each other.

6 Discussion

Our results suggest a general tendency in MT systems to default to masculine translations when encountering words that exist in different gendered forms, confirming behaviour reported throughout related literature (Stanovsky et al., 2019; Prates et al., 2019; Savoldi et al., 2021). Since all our German sentences are explicitly gendered, the bias can be assumed to stem from the target-language generation step itself, rather than from source-side ambiguity. The stereotypicality analysis results

similarly fall in line with prior findings (Stanovsky et al., 2019; Prates et al., 2019; Kappl, 2025). A correlation between translation error frequency and occupational gender distributions (with translations skewed towards male across nearly all occupational categories) strengthens the notion that MT systems not only default to masculine forms but also underestimate the frequency of female occupations beyond what occupation statistics alone would suggest.

Despite being one of the most widely used commercial MT systems[5], Google Translate ranks only marginally ahead of Microsoft Translator, the weakest system overall. Moreover, Google's own Translation LLM performs nearly identically to its traditional NMT system, whereas GPT-4o, a general-purpose LLM, consistently yields the most accurate translations. However, GPT-4o is not bias-free, performing weakest in Russian. SYSTRAN, the only hybrid system evaluated, performs remarkably well across the Romance languages, suggesting that explicit grammatical rules can mitigate gender bias specifically in German-to-Romance translation. It also performs worst within the Slavic languages. This shortfall likely reflects the fact that these are the two highest-complexity languages in our gender-complexity ranking (Appendix A). The pattern holds consistently in our results, suggesting that more complex gender-marking systems pose a genuine obstacle for MT systems in general.

The improvements when adding a pronoun clause (Style 2) and trans-identity marking (Style 3) suggest that translation accuracy can increase with explicit gender cues and systems move away from the masculine default. Notably, this does not hold for names, as both masculine and gender-neutral ones perform worse than no name at all, and feminine names improve only marginally against the baseline (§5.2). This runs against the intuition that a first name often is a direct gender cue. A possible explanation is that German first names, gender-neutral names in particular, are underrepresented in the training data these systems rely on. The heteronormativity probe reveals an additional bias: the partner's gender affects how the subject is translated. For individuals portrayed in a same-gender relationship accuracy drops consistently across five of six systems. A plausible explanation is that same-gender relationships are both underrepresented and stereotyped in the training data, so that systems associate romantic contexts with a heterosexual default and translate same-gender pairings less reliably.

[5] https://blog.google/products-and-platforms/products/translate/ten-years-of-google-translate/

Future Work. Our results point to several concrete directions for future work. First, the Hebrew gender prediction in this study is based on a simple word-ending heuristic (§3) and a dedicated Hebrew morphological analyser would likely improve both the alignment and the accuracy of the gender prediction for this language specifically. Although we provide an initial framework for the evaluation of non-binary persons, our approach needs to be extended with more gender-neutral occupational nouns and further neopronoun paradigms. Furthermore, accepted gender-neutral strategies per target language (e.g. Spanish *-e* forms) could be added as a rule or analysers which are able to detect inclusive morphology could be implemented. Furthermore, DEval-MT's template system already supports an additional demographic axis within the same intersectional attribute slot used for trans-identity marking in this study, but this axis is not empirically tested in the present paper. Five other attributes are implemented: *polnisch*, *türkisch*, *syrisch*, and *französisch* for nationality, and *Schwarz* for race. These are specifically chosen for a German context: Polish and Turkish as Germany's two largest immigrant communities (Bundeszentrale für politische Bildung, 2026), Syrian for its prominence in refugee-politics discourse, French as a Western-European baseline not typically exposed to racist stereotyping, and Black following the race categories used by Wang et al. (2022). Activating these would let future work test whether the masculine-default and stereotypicality effects reported here interact with a second protected attribute.

7 Conclusion

To address bias in MT and the harms it can cause, a systematic evaluation of MT systems is crucial. DEval-MT provides such a foundation for German MT, offering a dataset and evaluation methodology tailored to the German language. The evaluation of five MT systems and one general-purpose LLM reveals bias as a consistent pattern of underrepresentation and stereotyping of individuals. These findings underscore the need for MT development that prioritizes fairness along with accuracy, so that translations are correct and equitable for all users.

Limitations

Despite providing a systematic resource for examining intersectional gender bias in German machine translation, this evaluation has several limitations.

We want to highlight that, although we provide an initial framework for the evaluation of MT beyond a binary gender, we cover only a very small part of the gender spectrum. A single neopronoun paradigm, one adjective for trans-identity marking, and a small set of gender-neutral nouns cannot represent the range of identities, linguistic practices, and lived experiences of trans, non-binary, and queer people. Dedicated research on these topics, ideally together with the communities concerned, is necessary, and our work can serve as a basis for it.

Especially non-binary representation is covered only to a limited extent. The DIVERSE label applies to 258 of 6,618 sentences (∼3.9%), built from 28 gender-neutral occupation nouns and a single neopronoun paradigm (*dey/deren*), which is too little data for reliable per-language or per-system conclusions. Systems almost never produce a DIVERSE label (0.0% in Spanish, French, Italian, Arabic, and Hebrew; 3.2% in Russian and 4.1% in Ukrainian), which suggests model failure, but can also be attributed to the fact that spaCy does not provide a tag for gender-neutral or inclusive forms and only detects a true grammatical neuter, which is insufficient as many target languages offer no standardized neutral form at all.

Hebrew appears as a striking outlier in multiple gender translation accuracy metrics and the human evaluation step. We believe this to be a result of the morphological analysis step that we implemented. Since Hebrew frequently encodes gender within the verb as well as the sentence subject and our pipeline was not specifically adapted for this case. We recommend not taking these results at face-value and refer to §6 for possible adaptations of our method to solve this issue.

The heteronormativity probe (§5.2) treats the romantic partner only as a binary label of the same / different gender. Therefore, it does not analyse the partner noun's own grammatical gender as a variable, and the size or direction of the observed gap could change once that is accounted for.

The human validation (§5.3) is incomplete in coverage: it excludes Arabic entirely because only a single annotator completed that language.

Furthermore, certain biases, such as semantic derogation, have not been investigated (Savoldi et al., 2021). For example, when translating "teacher" into Spanish, Google Translate translates the masculine noun as *el profesor* ("high school/college teacher") but the feminine subject as *la maestra* ("elementary/middle school teacher"). Since both forms are grammatically correctly gendered, this lexical asymmetry does not register as an error in our evaluation.

Finally, the dataset is fully synthetic and template-based. This enables controlled, minimal-pair comparisons, but it does not measure how these biases manifest in naturally occurring, non-templated text.

Ethical Considerations

Our support for the German neopronoun *dey/deren* reflects one commonly self-chosen option among non-binary German speakers (Verein für geschlechtsneutrales Deutsch e.V., 2022), but non-binary people use a range of self-referential forms, and our neopronoun design should not be read as representative of non-binary language use in general. Our romantic-context sentences (Style 5) describe only monogamous, two-person relationships. This reflects a normative assumption about the structure of relationships that does not capture how all romantic partnerships are actually organised. Lastly, none of the patterns we report reflects malicious or intentional design on the part of the MT systems evaluated, but our results highlight that these systems are not gender-neutral in their output and can reinforce occupational stereotypes and underrepresent non-masculine referents, which in turn can produce wrong and harmful translations.

References

Knud Bielefeld. 2025. Die top 50-hitliste der echten unisexnamen.

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

Stephanie Brandl, Ruixiang Cui, and Anders Søgaard. 2022. How conservative are language models? adapting to the introduction of gender-neutral pronouns. In *Proceedings of the 2022 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3624–3630, Seattle, United States. Association for Computational Linguistics.

Bundeszentrale für politische Bildung. 2026. Bevölkerung mit Migrationshintergrund | Die soziale Situation in Deutschland.

Yang Trista Cao and Hal Daumé III. 2021. Toward gender-inclusive coreference resolution: An analysis of gender and bias throughout the machine learning lifecycle. *Computational Linguistics*, 47(3):615–661.

Marta R. Costa-jussà, Pierre Andrews, Eric Smith, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Daniel Licht, and Carleigh Wood. 2023. Multilingual holistic bias: Extending descriptors and patterns to unveil demographic biases in languages at scale. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14141–14156, Singapore. Association for Computational Linguistics.

Marta R. Costa-jussà. 2019. An analysis of gender bias studies in natural language processing. *Nature Machine Intelligence*, 1(11):495–496.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Lisa K. Horvath, Elisa F. Merkel, Anne Maass, and Sabine Sczesny. 2016. Does Gender-Fair Language Pay Off? The Social Perception of Professions from a Cross-Linguistic Perspective. *Frontiers in Psychology*, 6.

Michelle Kappl. 2025. Are All Spanish Doctors Male? Evaluating Gender Bias in German Machine Translation. *arXiv preprint*. ArXiv:2502.19104 [cs.CL].

Mikael Parkvall. 2008. The simplicity of creoles in a cross-linguistic perspective. In *Language Complexity: Typology, Contact, Change*, pages 265–285. John Benjamins Publishing Company.

Marcelo O. R. Prates, Pedro H. C. Avelar, and Luis Lamb. 2019. Assessing gender bias in machine translation – a case study with google translate. *Preprint*, arXiv:1809.02208.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High Quality Word Alignments without Parallel Training Data using Static and Contextualized Embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643. Association for Computational Linguistics.

Sandra Sandoval, Jieyu Zhao, Marine Carpuat, and Hal Daumé III. 2023. A rose by any other name would not smell as sweet: Social bias in names mistranslation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3933–3945, Singapore. Association for Computational Linguistics.

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

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2025. A decade of gender bias in machine translation. *Patterns*, 6(6):101257.

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.

Statistik der Bundesagentur für Arbeit. 2020. Klassifikation der Berufe 2010 – überarbeitete Fassung 2020.

Adina D. Sterling, Marissa E. Thompson, Shiya Wang, Abisola Kusimo, Shannon Gilmartin, and Sheri Sheppard. 2020. The confidence gap predicts the gender pay gap among STEM graduates. *Proceedings of the National Academy of Sciences of the United States of America*, 117(48):30303–30308.

Ian Stewart and Rada Mihalcea. 2024. Whose wife is it anyway? assessing bias against same-gender relationships in machine translation. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 365–375, Bangkok, Thailand. Association for Computational Linguistics.

tagesschau.de. 2024. Die beliebtesten vornamen 2024.

Verein für geschlechtsneutrales Deutsch e.V. 2022. Ergebnisse der zweiten öffentlichen umfrage zu geschlechtsneutralem deutsch.

Dries Vervecken and Bettina Hannover. 2015. Yes I Can! Effects of Gender Fair Job Descriptions on Children's Perceptions of Job Status, Job Difficulty, and Vocational Self-Efficacy. *Social Psychology*, 46:76–92.

Lucas Nunes Vieira, Minako O'Hagan, and Carol O'Sullivan. 2021. Understanding the societal impacts of machine translation: a critical review of the

literature on medical and legal use cases. *Information, Communication & Society*, 24(11):1515–1532.

Jun Wang, Benjamin Rubinstein, and Trevor Cohn. 2022. Measuring and mitigating name biases in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2576–2590, Dublin, Ireland. Association for Computational Linguistics.

Taha Zerrouki. 2012. qalsadi, arabic mophological analyzer library for python.

A Language Gender-Complexity Scores

Table 5 reports a gender-complexity score for each of the seven target languages of this paper plus German (the source language). The scores follow the method of Parkvall (2008), aggregating seven WALS[6] typological features, each normalised to $[0, 1]$ and combined into a single weighted score: five directly describe gender marking (number of genders, sex- vs. non-sex-based gender assignment, systems of gender assignment, gender distinctions in personal pronouns, and the relation between third-person pronouns and demonstratives), while two are general morphological features with a comparatively lower weight in the combined score (prefixing vs. suffixing in inflectional morphology, inflectional synthesis of the verb) rather than gender-specific measures on their own. We report this ranking as a supplementary typological context for the language-level pattern discussed in §5.

Language	Complexity score
Ukrainian	1.000
Russian	0.869
Spanish	0.853
Arabic	0.849
German (source)	0.849
Hebrew	0.776
French	0.776
Italian	0.708

Table 5: WALS-based gender-complexity scores (0-1, higher = more complex) for this paper's seven target languages and German.

B Full Logistic Regression Report

The complete logistic regression report is shown in Table 6.

[6] https://wals.info

	GPT-4o			Google		
Lang.	**OR**	**95% CI**	p	**OR**	**95% CI**	p
es	1.22***	[1.12, 1.33]	<0.001	1.90***	[1.80, 2.02]	<0.001
fr	1.26***	[1.17, 1.37]	<0.001	1.59***	[1.50, 1.69]	<0.001
it	1.34***	[1.25, 1.43]	<0.001	1.91***	[1.81, 2.02]	<0.001
ar	1.26***	[1.19, 1.33]	<0.001	1.33***	[1.26, 1.40]	<0.001
ru	1.40***	[1.32, 1.47]	<0.001	1.61***	[1.53, 1.70]	<0.001
uk	1.30***	[1.22, 1.38]	<0.001	1.51***	[1.44, 1.60]	<0.001
he	0.72***	[0.67, 0.77]	<0.001	0.94*	[0.89, 0.99]	0.023

Table 6: Full logistic regression results behind Table 2: odds ratio (OR) of a correct translation given stereotype-congruence, per language, for GPT-4o and Google, with 95% confidence interval and exact p-value ($n = 6,348$). OR >1 means more accurate when stereotype-congruent. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.