GRUFF: LLM Pronoun Fidelity, Reasoning, and Biases in German

Anonymous ACL submission

Abstract

Third-person singular pronouns have long been used to study stereotypical biases in language models and to test their abilities to reason about reference. More recently, the interplay between reasoning and bias has been investigated with the task of pronoun fidelity, which assesses models' abilities to correctly reuse a previously-specified pronoun for a discourse entity, independent of other potentially distracting discourse entities mentioned in between. However, such research focuses on English, which is a language with limited grammatical gender and almost no gender agreement. In this paper we contribute a novel, large-scale dataset, GRUFF, to measure pronoun fidelity in German, covering four different gender agreement systems in nouns, and four sets of pronouns. With this dataset, we show that LLMs show strong grammatical agreement for masculine and feminine entities in the absence of explicit context, but not for neopronouns *xier* and *en*. Models are generally not robust to distractors, but encoder-only models are more robust in German than in English, reflecting the importance of grammatical gender. Finally, we show that occupational stereotypes in this context are poorly correlated across grammatical cases, and across most models, except ones with closely related architectures. We release all code and data to encourage further work on gender-inclusive language and referential reasoning in German.

Grammatical Agreement (§6)
Die Tierärztin *benötigte Kaffee, weil* ***sie*** *sehr früh aufgestanden war.* ***Die Tierärztin*** *sagte, dass ___ an Wochenenden nicht arbeitet.*
*(****The vet*** *needed coffee as* ***she*** *woke up early. The vet said ___ did not work weekends.)*

Robustness to Distractors (§7)
Die Tierärztin benötigte Kaffee, weil sie sehr früh aufgestanden war. ***Dier Besitzer*in*** *war voll, weil* ***xier*** *gerade eine große Mahlzeit gegessen hatte. Die Tierärztin sagte, dass ___ an Wochenenden nicht arbeitet.*
(The vet needed coffee as she woke up early. ***The owner*** *was full because* ***xe*** *ate a big meal. The vet said ___ did not work weekends.)*

Biases and Stereotypes (§8)
Die Tierärztin war gesättigt, weil ihr das Essen gut geschmeckt hatte. Die Tierärztin sagte, dass ***er*** *an Wochenenden nicht arbeitet. (****he****)*

Figure 1: Examples from our proposed dataset, GRUFF (§4), summarizing the main contributions of this paper. In all cases, the blank should be filled with **sie** *(she)*.

1 Introduction

Third-person singular personal pronouns (e.g., *he*, *she*, *they* in English) are among the smallest units of language, yet they are important ways to index and construct identity and gender (Lauscher et al., 2022b). Thus, as evidenced by psychological research (e.g., McLemore, 2018), failing to use an individual's pronouns correctly (i.e., *misgendering*) can result in significant psychological harm, and pronouns are a linguistic site where minority stress occurs. As large language models (LLMs) become increasingly integrated into our daily lives, their ability to navigate these nuances is no longer just a technical requirement, but an ethical imperative.

Researchers in natural language processing have started to address this issue by creating resources to evaluate and improve LLMs' abilities to handle third-person pronouns correctly (e.g., Lauscher et al., 2023; Ovalle et al., 2024; Hossain et al., 2024, *inter alia*). One such work, Gautam et al. (2024a), introduced the task of *pronoun fidelity* – assessing whether, given a context introducing a co-referring entity and pronoun, models are able to reuse the correct pronoun later. To this end, the authors propose RUFF, a dataset to test model robustness when ex-

posed to a number of non-adversarial distractor sentences discussing other entities in between. Such resources are critical to measure progress towards harm-free and robust pronoun usage. However, the vast majority of existing research on safe pronoun usage (*i*) focuses exclusively on English (e.g., Hossain et al., 2023; Gautam et al., 2024a), or (*ii*) is restricted to testing pronominal bias in machine translation (e.g., Lauscher et al., 2023). This narrowness is problematic, given that LLMs are used by millions of people worldwide in their native languages (Wang et al., 2023).

In this work, we acknowledge that gendered language also impacts German speakers (Huck, 2021) and study pronoun fidelity in German. Here, grammatical gender is marked not just on pronouns, but also nouns and articles, providing stronger cues for gender agreement. We introduce GRUFF, the first resource for studying pronoun fidelity in LLMs for German, covering four noun agreement systems, and four pronoun sets across four grammatical cases in German—including gender-inclusive noun variants such as the Sternchen and neopronouns like *xier*. Using GRUFF, we conduct a comprehensive empirical analysis of LLMs' abilities to correctly track and reproduce pronouns.

We find that LLMs model almost-perfect grammatical agreement for traditional pronouns in German, independent of explicit context specifying the pronoun. However, LLMs only show agreement with neopronouns when they are explicitly specified. As in English, discourse complexity in the form of distractors dramatically affects performance, but German appears more robust than English, likely due to increased gender marking. Gender marking likely also mitigates the effects of stereotypical biases, which are not significantly correlated across grammatical cases of a pronoun set, nor across models. Finally, pronoun fidelity is best for traditional pronouns (with no difference, surprisingly, for *er* and *sie*), and models show large quality-of-service differences for more novel, gender-inclusive forms. Our results highlight the importance of fairness evaluations on morphologically complex languages, and motivate future work on closing these gaps in inclusive multilingual NLP.

2 Related Work

Gender fairness and identity inclusion in NLP. Faithful reproduction of an individual's pronouns falls under research on gender fairness and identity inclusion, surveyed comprehensively in (Stańczak and Augenstein, 2021; Bansal, 2022; Nemani et al., 2024, *inter alia*). Early work in this space showed that NLP systems' predictions and representations reflect and amplify binary gender biases in training data (Bolukbasi et al., 2016; Sun et al., 2019; Lu et al., 2020). Beyond English, cross-lingual analyses reveal that gender bias manifests differently across languages and linguistic systems, particularly in languages with grammatical gender, highlighting the importance of multilingual evaluation (Chen et al., 2021) and mitigation (Zmigrod et al., 2019). More recent work has expanded the conceptual scope of gender fairness, questioning binary conceptions of gender (Devinney et al., 2022; Cao and Daumé III, 2021) and conflating gender with other linguistic categories (Larson, 2017; Bartl et al., 2025). Thus, research on identity-inclusive NLP has included on neopronouns and gender-inclusive language (Lauscher et al., 2022a, 2023; Sobhani et al., 2023; Waldis et al., 2024), as part of the emerging area of queer NLP (Felkner et al., 2023; Weber et al., 2026).

Reasoning with and about pronouns. A substantial body of research evaluates pronoun handling through coreference benchmarks such as Winograd-style schemas (Levesque et al., 2012; Abdou et al., 2020; Emelin and Sennrich, 2021) and occupational bias datasets (Rudinger et al., 2018; Zhao et al., 2018; Levy et al., 2021). Earlier work focused primarily on binary pronouns, while more recent studies include singular *they* (Baumler and Rudinger, 2022) and neopronouns (Cao and Daumé III, 2021; Gautam et al., 2024b; Bunzeck and Zarrieß, 2024). Pronoun agreement is also part of linguistic acceptability judgments, which are often tested with minimal pair datasets such as BLiMP in English (Warstadt et al., 2020), and MultiBLiMP in several other languages (Jumelet et al., 2025). Sinha et al. (2023) show that such acceptability judgments are context-sensitive, with performance varying under contextual manipulations. Context also plays a role in faithful reproduction of individuals' pronouns: Hossain et al. (2023) and Ovalle et al. (2023) evaluate *pronoun fidelity*, but only in simple scenarios with a single individual. Gautam et al. (2024a) consider more complex pronoun fidelity involving multiple individuals, where some are non-adversarial distractors. Gautam (2026) further showed that reasoning models still barely perform above chance in this setting.

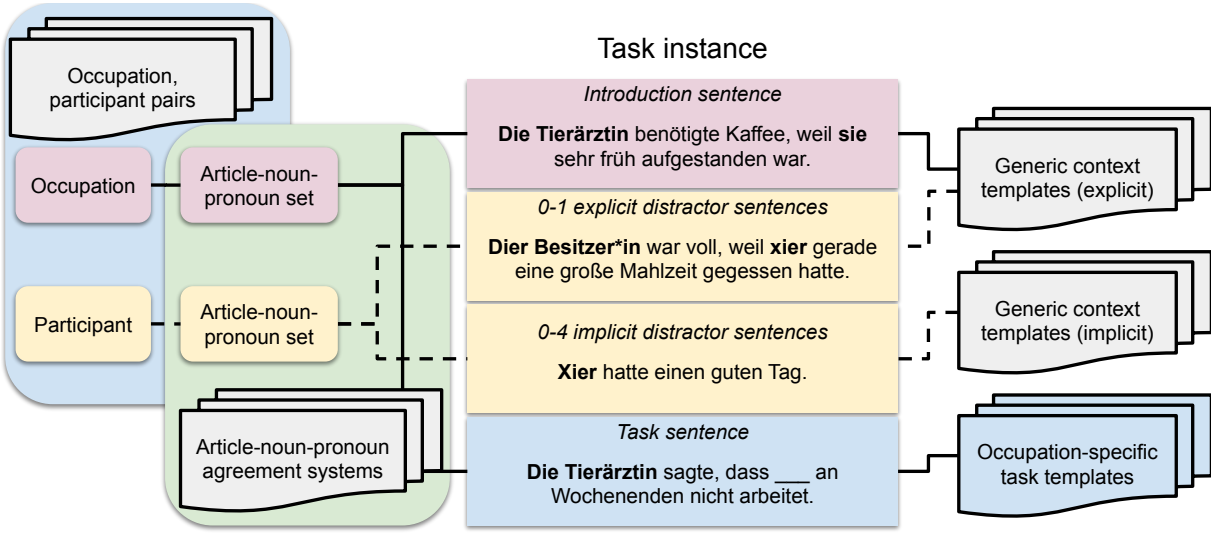


Figure 2: Overview of the creation of each task instance in GRUFF, which involves sampling from occupation-participant pairs (Section 4.1), article-noun-pronoun agreement systems (Tables 1 and 2), and context and task templates (Section 4.2), and assembling them together (Section 4.3).

However, all these works only consider English.

Multilinguality and pronouns in NLP. Multilingual work on pronouns has primarily emerged in machine translation, where datasets such as WinoMT help quantify gender bias across language pairs (Stanovsky et al., 2019; Troles and Schmid, 2021). These works often study pronoun preservation when contextual information is required to resolve gender (Müller et al., 2018; Voita et al., 2018; Fernandes et al., 2023), and Sharma et al. (2022) inject explicit coreference cues to improve pronoun translation performance. Strategies to improve performance include also transfer learning and adding relevant information (Saunders and Byrne, 2020; Basta et al., 2020). Some works such as Jourdan et al. (2025) and Björklund and Devinney (2023) explicitly address neopronouns and gender-inclusive language in French and Swedish, but evaluations of pronoun fidelity remain largely unexplored outside English.

Our work unites these strands by introducing the first large-scale benchmark for pronoun fidelity in German with a range of pronouns, morphological control, and distractor-based tests of robustness.

3 Pronoun Fidelity in German

The task of pronoun fidelity evaluates the ability of LLMs to consistently reproduce appropriate pronouns in contextualized scenarios, even when intervening distractor sentences introduce competing pronoun references. We modify Gautam et al. (2024a)'s original English framework, accounting for German's rich morphology: Unlike English, German requires that articles, nouns, and pronouns all agree with each other in grammatical gender (e.g., *der Tierarzt* = *the.MASC.SG vet.MASC.SG*). Accordingly, for an entity a referred to by their occupation, and an optional second entity b (referred to as a participant), we incorporate gendered articles a_a and a_b that match gender-inflected nouns identifying the entities e_a and e_b, as well as the pronouns p_a and p_b. Each task instance consists of:

Introduction sentence $i(a_a, e_a, p_a)$*:* Establishes coreference between an entity a (referred to with an article a_a and noun e_a, e.g., *der Tierarzt* = *the.MASC.SG vet.MASC.SG*), and a pronoun p_a (e.g., *er* = *he*).

Distractor sentences $D(a_b, e_b, p_b)$*:* 0–5 sentences that discuss a different entity b, either with explicit sentences introducing them with an article a_b, noun e_b, and pronoun p_b, or by implicitly continuing an already-established coreference using only a pronoun p_b.

Task sentence $t(a_a, e_a, p)$*:* Sentence referring to the original entity a with an article a_a, noun e_a, and a pronoun gap p with an unambiguous reference to the same entity a, which must be filled.

Under this general framework, we consider various noun agreement systems and pronoun sets in multiple grammatical cases, as described below.

Noun system	Example
Masculine	*Der Tierarzt*
Feminine	*Die Tierärztin*
De-e	*De Tierarzte*
Sternchen	*Dier Tierarzt*in*

Table 1: The four noun agreement systems we consider for people, including masculine, feminine, and two gender-neutral variants.

Case	Pronoun			
	Masc.	Fem.	Neopronouns	
Nominative	*er*	*sie*	*en*	*xier*
Accusative	*ihn*	*sie*	*en*	*xien*
Genitive	*seine*	*ihre*	*ense*	*xiese*
Dative	*ihm*	*ihr*	*em*	*xiem*

Table 2: The four sets of pronouns across four grammatical cases we consider, including masculine and feminine pronouns, and two sets of neopronouns.

Data type	Number of instances	
With no context		
Task sentences	240	
With introductory context		
+ 0 distractors	3 x 2,880	(of 9,600)
+ 1 distractor	3 x 2,880	(of 115,200)
+ 2 distractors	3 x 2,880	(of 460,800)
+ 3 distractors	3 x 2,880	(of 1,382,400)
+ 4 distractors	3 x 2,880	(of 2,764,800)
+ 5 distractors	3 x 2,880	(of 2,764,800)

Table 3: Number of dataset instances in GRUFF. We subsample 3 sets of 2,880 instances in each setting with introductory context.

4 GRUFF Dataset Creation

Figure 2 provides an overview of the data-creation process, which we explain in detail in this section.

4.1 Occupations and Participants

We use the same 60 occupation-participant pairs as in Gautam et al. (2024a), translated to German. Unlike English, where a single gender-neutral noun (e.g., *vet*) and definite article (*the*) are used regardless of gender, German requires gender-specific articles and nouns. We consider four forms of articles and nouns (summarized in Table 1): Masculine (*der Tierarzt*), feminine (*die Tierärztin*), the de-e system (a German gender-neutral neosystem; *de Tierarzte*), and the gender-inclusive "Sternchen" (*dier Tierarzt*in*) (Waldis et al., 2024). These also apply to distractor entities.

4.2 Template Creation

Introduction and distractor sentences are created from generic **context templates** for each grammatical case of a pronoun, while task sentences are created from occupation-specific **task templates**. Creating these templates involved both direct translation of English RUFF templates and complete reconstruction of new templates to accommodate German grammatical complexity.

We pair pronoun sets with our four noun systems: Masculine *er*, feminine *sie*, and the neopronouns *en* (Verein für geschlechtsneutrales Deutsch e. V.) and *xier* (Heger, 2020). We consider all four grammatical cases in German, i.e., nominative, accusative, dative, and genitive. Table 2 shows all pronouns. Creating genitive templates was particularly complex; while English possessive pronouns only inflect for the possessor's number and gender, German possessive pronouns inflect for both the possessor and the possessed object. We constructed all genitive templates with feminine possessed objects, for a consistent (*-e*) ending across these pronouns and easier evaluation.

In total, we created 240 occupation-specific task templates (60 occupations x 4 grammatical cases), and 40 generic context templates (10 x 4 grammatical cases), which are subsequently combined to produce instances as described below.

4.3 Template Assembly

Each task template is combined with 10 introductory templates and filled with 4 sets of articles, nouns and pronouns, for a total of 9,600 sentences with 0 distractors. These are randomly downsampled to 3 sets of 2,880, where each of the 4 sets is equally represented. Distractor sentences are always seeded with a different gender to keep answers unambiguous (see Appendix A for details). Filling and combining templates results in over seven million unique instances (see Table 3).

4.4 Human Validation

Three annotators (including one author) independently validated all context templates, and 600 final instances from GRUFF (100 randomly sampled instances for each of 0–5 distractor sentences),

Model	Size	Quantization	Language
Encoder-only models			
GBERT-BASE (He et al., 2023)	110M	–	German
GBERT-LARGE (He et al., 2023)	340M	–	German
MBERT-BASE (Devlin et al., 2019)	178M	–	Multilingual
XLM-ROBERTA-BASE (Liu et al., 2019)	270M	–	Multilingual
Decoder-only models			
SAUERKRAUTLM-8B (Mayflower GmbH, 2024)	8B	4-bit	German
SAUERKRAUTLM-70B (tresiwalde, 2024)	70B	4-bit	German
LLAMA-3.3-70B (AI@Meta, 2024)	70B	4-bit	Multilingual

Table 4: List of models used for our evaluation, along with their architecture, quantization (if applicable), and size.

achieving 100% accuracy in these settings. Further details on annotation are in Appendix B.

5 Experimental Setup

Using GRUFF, we conduct a comprehensive evaluation of language models with German support.

5.1 Models

We select four encoder-only and three decoder-only models, covering a range of sizes and pre-training languages, as summarized in Table 4. We aim to study the effects of model architecture as in Gautam et al. (2024a), as well as the differences between German-only and multilingual models.

5.2 Evaluation

To enable consistent, automatic evaluation across model architectures, we use a forced-choice setting where models select one of four pronoun options. Concretely, we fill the pronoun gap in $t(a_a, e_a, p)$ with each pronoun of the appropriate case. We then compute average sequence probabilities, using log-likelihoods for decoder-only models, and pseudo-log-likelihoods for encoder-only models (Salazar et al., 2020). We average rather than sum probabilities, as the number of tokens differs substantially across variants. Finally, the highest-likelihood option is taken as the model's answer. We opt for probability- rather than generation-based evaluation as we are interested in LLM capabilities as models of language (Subramonian et al., 2025).

Human accuracy of 100% represents the performance ceiling that language models could theoretically achieve. Gautam et al. (2024a) use a 25% baseline, which is the probability of choosing between four pronouns at random. We, however, do not indicate this baseline as it is unrealistic; models tend to overpredict and better resolve traditional pronouns over neopronouns.

6 Grammatical Agreement

As the German language features rich morphological markings for grammatical gender, we can expect good language models to show grammatical agreement in the absence of any introductory sentence, i.e., masculine nouns should take masculine pronouns, feminine nouns should take feminine pronouns, and de-e and Sternchen nouns can correctly take either *en* or *xier*. We can already measure this in the no-context case, unlike in English, and then investigate whether explicitly specifying a pronoun *improves* agreement.

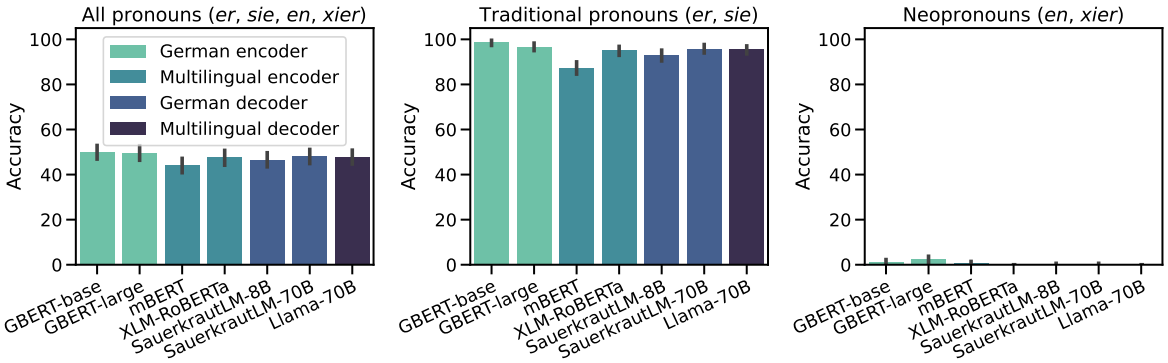
As Figure 3 shows, **models show almost perfect grammatical agreement for traditional pronouns**, independent of context. On the other hand, **models show near-zero agreement for neopronouns without explicit context** which includes *xier* or *en* pronouns. The stark difference between Figures 3a and 3b for neopronouns indicates that providing explicit instruction on neopronouns for language models to condition on may help to improve pronoun fidelity for neopronoun users.

The patterns and raw performance scores are consistent across models with different pre-training language distributions, as well as across architectures. This corroborates Gautam et al.'s (2024a) findings that encoder-only models are far better than decoder-only models at comparable scales, as the decoder-only models we consider are also 80–700 times larger than the encoder-only ones.

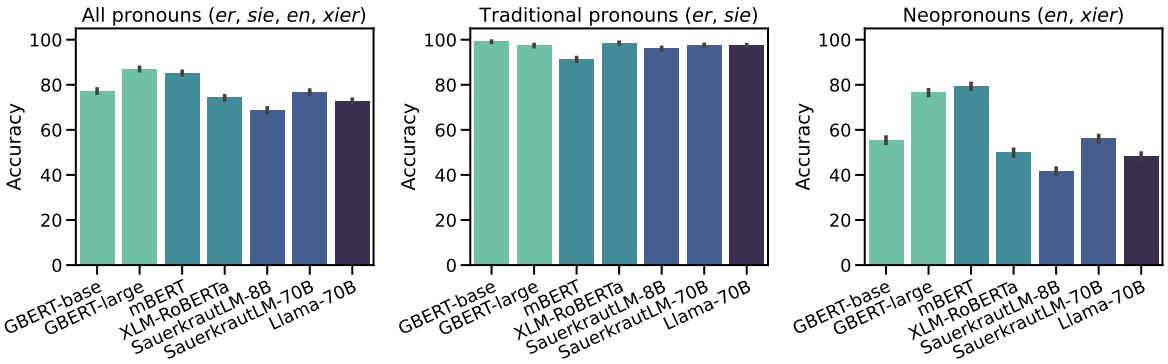
7 Robustness to Distractors

Next, we investigate how non-adversarial distractors (in the form of a second entity with different pronouns) affect model performance. In English, performance drops dramatically with just one distractor. However, German's encoding of gender agreement is likely to make models more robust.

In Figure 4, we see that **distractors affect per-**



(a) With *no* context (just grammatical gender agreement)



(b) With context explicitly specifying a pronoun

Figure 3: Model accuracy on grammatical gender agreement with and without context. Without context, models are near-perfect for traditional masculine and feminine pronouns, but cannot produce any neopronoun. With an explicitly specified pronoun, models can reproduce neopronouns and maintain performance on traditional pronouns.

formance dramatically, despite the additional signal of grammatical gender. These trends are identical for traditional pronouns and neopronouns, except for an overall downward shift in performance with neopronouns. Patterns by architecture are nearly identical to Gautam et al. (2024a), likely due to bidirectional attention and evaluation with pseudo log likelihoods for encoder-only models: As in English, additional distractors worsen decoder-only model performance, but encoder-only models recover marginally. German-only and multilingual models largely follow similar trends.

The vast majority of errors happen when models repeat the distractor pronoun, as Table 5 shows. Interestingly, **encoder-only models are most distracted with *two* distractors**, rather than one, as in English. We posit that this is a direct result of more gender marking in German, which is a stronger signal to models, enabling robust pronoun fidelity when only one distractor is present.

8 Bias and Stereotypes

The previous sections consider pronoun fidelity from the perspectives of grammatical agreement and robustness. In this section, we turn to concerns of biases and stereotyping. In English, Gautam et al. (2024a) report statistically significant differences in pronoun fidelity with *he/him/his* pronouns compared to *she/her/her*, and a further drop in performance for singular *they* and the neopronoun *xe/xem/xyr*. We now investigate whether similar quality-of-service differentials (Cao and Daumé III, 2021) appear with German.

Across all settings with an explicit introduction (i.e., with 0–5 distractors), pronoun fidelity is 66.52 ± 10.59 for *er/ihm* pronouns, 63.18 ± 12.58 for *sie/ihr*, 23.49 ± 26.05 for *en/em*, and 66.74 ± 29.54 for *xier* pronouns. With a Wilcoxon signed-rank test ($\alpha = 0.05$), there are **no statistically significant differences in pronoun fidelity between *er*, *sie*, and *xier* pronouns, but *en* pronouns are significantly worse**.[1] The parity in performance is likely because gender information from the noun provides an additional and repeated cue for im-

[1] In these results, *en* pronouns are mapped exclusively to de-e articles and nouns (and *xier* to the more common Sternchen form), but this pattern holds even when the mappings are swapped, as we show in Appendix D.

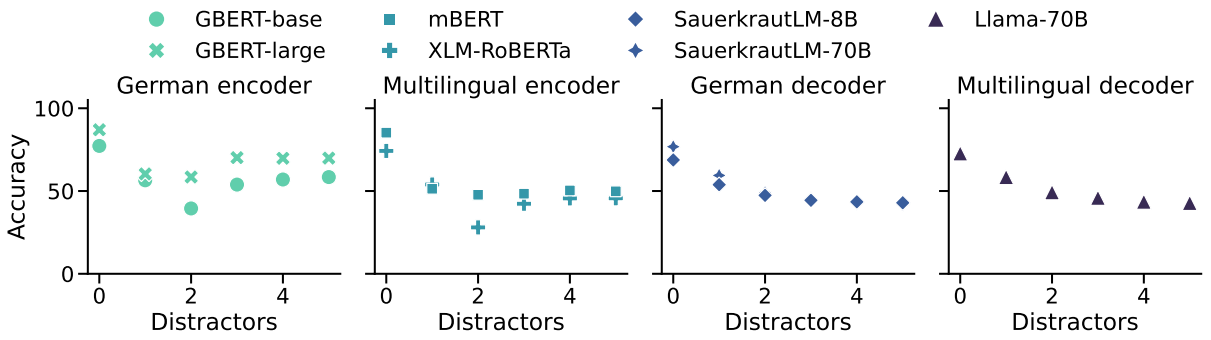


Figure 4: Accuracy at pronoun fidelity with an introductory sentence and 0–5 distractors across different models.

Model	Percentage of errors that are due to distraction				
	1 distractor	2 distractors	3 distractors	4 distractors	5 distractors
Encoder-only models					
GBERT-BASE	85.59 ± 0.62	†94.72 ± 0.21	86.60 ± 0.57	83.87 ± 0.71	82.82 ± 0.40
GBERT-LARGE	85.56 ± 0.14	†86.65 ± 0.37	78.87 ± 0.91	79.41 ± 0.54	79.00 ± 0.93
MBERT	†86.19 ± 1.26	86.07 ± 0.24	83.65 ± 0.26	82.70 ± 0.22	81.71 ± 0.58
XLM-ROBERTA	80.98 ± 0.79	†93.36 ± 0.09	87.74 ± 0.71	85.25 ± 0.29	84.44 ± 0.11
Decoder-only models					
SAUERKRAUTLM-8B	56.92 ± 0.14	65.99 ± 0.46	69.52 ± 0.24	69.61 ± 0.57	†69.77 ± 0.49
SAUERKRAUTLM-70B	75.21 ± 0.43	82.70 ± 0.55	83.93 ± 0.39	84.70 ± 0.34	†85.47 ± 0.40
LLAMA-70B	72.39 ± 0.86	79.95 ± 1.06	83.32 ± 0.55	83.26 ± 1.09	†83.56 ± 0.94

Table 5: All model errors are mostly due to distraction. With more distractors, encoder-only models are less distracted, while decoder-only models are more distracted. †: Highest distraction error in each row.

proved pronoun fidelity in German. This could also explain why there are no statistically significant differences in performance between different grammatical cases of the same pronoun set (see Figure 5), unlike in English (Gautam et al., 2024b). Taken together, these results show that pronoun fidelity in German, while not robust to distractors, is characterized by much less bias than English precisely because of more morphological gender marking. Rodríguez et al. (2025) finds similar results with Spanish, despite hypothesizing that more gender marking would result in more bias.

Finally, we study occupational biases, comparing how much models over-resolve a particular pronoun to to an occupation, even overriding grammatical gender signals from articles and nouns. We exclude neopronouns, as their behaviour is less interpretable in terms of social gender stereotypes and less likely to be learned from data due to sparsity. We consider all settings with an introduction and any number of distractors. As Figure 6 shows, *Friseur* (hairdresser), *Planer* (planner), and *Küchenchef* (chef) are overresolved to masculine *ihn*, while *Berater* (counsellor), *Kassierer* (cashier), and *Börsenmakler* (broker), are overresolved to *sie*.

As this analysis produces a ranking, we can use rank correlation metrics to understand whether occupational biases of models with different grammatical cases are similar, and whether the occupational biases of different *models* are similar. Specifically, we use Spearman's ρ, and report test statistics as well as significance ($\alpha = 0.05$, with Bonferroni correction). Similar to Gautam et al. (2024b), we find **no significant correlation between stereotypical biases across different grammatical cases of the same pronoun set**, as shown in Figure 7. Models, however, tend to be slightly correlated with each other (see Figure 8), particularly when they have a similar architecture, e.g., GBERT-BASE and -LARGE, and SAUERTKRAUTLM-70B and LLAMA-70B. However, this does not always hold, as SAUERKRAUTLM-8B and -70B are very weakly correlated. Overall, we see that German-language occupational biases are substantially different from case to case, and model to model, but that the overall issues of gender bias as substantially different in a gender-marked language such as German, compared to English.

In sum, quality-of-service differentials in German pronoun fidelity are low compared to English, and stereotyping is both weak and weakly correlated across different settings.

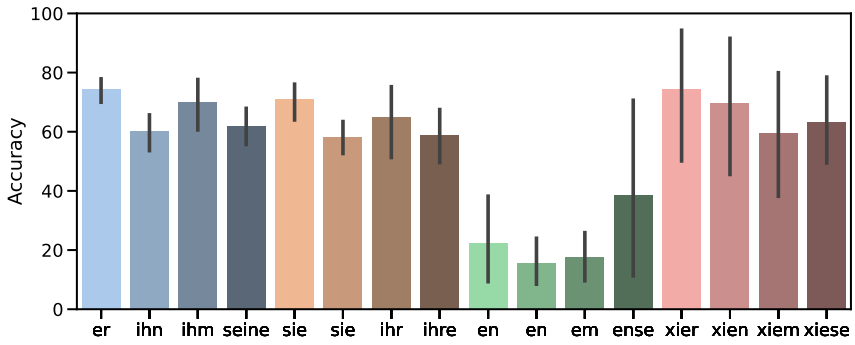


Figure 5: Pronoun fidelity split by pronoun set and by grammatical case. Overall models are best at reusing *er*, *sie*, and the neopronoun *xier*. Performance with a different neopronoun, *en*, is very low.

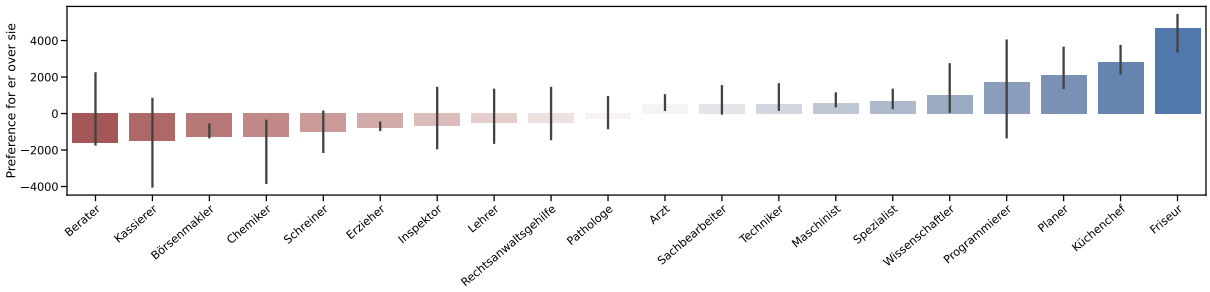


Figure 6: Top 10 occupations over-resolved to accusative pronouns *sie* (feminine) on the right, versus *ihn* (masculine) on the left, in a pattern indicative of stereotyping. Results with other grammatical cases are shown in Appendix E.

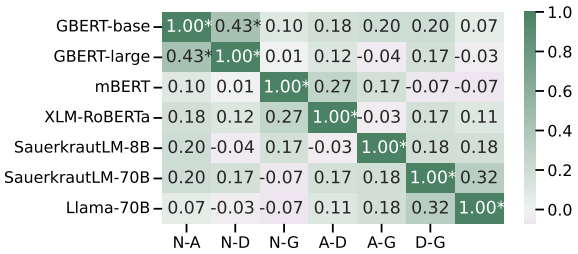


Figure 7: Spearman's correlation between the stereotypical biases across different cases, on a per model basis. N: Nominative, A: Accusative, D: Dative, G: Genitive, and * indicates statistical significance ($\alpha = 0.05$). Correlations are generally low and even the higher values are generally not significant.

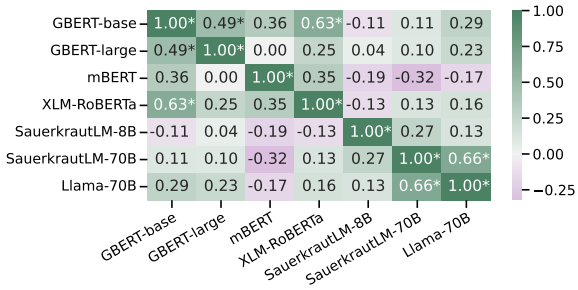


Figure 8: Spearman's correlation between the stereotypical biases of models for accusative pronouns. * indicates statistical significance ($\alpha = 0.05$).

9 Discussion and Conclusion

The GRUFF dataset that we propose enables us to evaluate a range of language models with the lenses of grammatical agreement, robustness to discourse complexity (in the form of distractor sentences about other entities), as well as biases and stereotypes. Although our work is inspired by the English RUFF dataset, these issues are fundamentally different in a morphologically rich language like German which marks gender on articles, nouns, and pronouns. We show that these linguistic differences result in higher and more robust pronoun fidelity. Somewhat counterintuitively, more explicit marking of gender results in less stereotyping, and more balance in pronoun fidelity across grammatical genders. Our results thus highlight the importance of fairness research that goes beyond English.

Nevertheless, performance with neopronouns and gender-inclusive articles and nouns remains mixed, particularly with *en* pronouns, the recommendation of the society for gender-neutral German, Verein für geschlechtsneutrales Deutsch e. V.. Closing these gaps should be an important priority for future work on inclusive multilingual NLP.

Limitations

The primary limitation of our study is its ecological validity, as we use researcher-constructed synthetic templates rather than natural discourse, and a forced-choice evaluation setting with language model probabilities. For the questions we are asking about how LLMs model agreement and how pronouns of different discourse entities interact, this is the only setting that provides sufficient control, as in (Gautam et al., 2024a), but results with generations or real-world data may differ (Subramonian et al., 2025). Our evaluation is also limited to encoder-only and decoder-only models, and we do not consider encoder-decoder architectures, models with extensive post-training (e.g., chat models). or reasoning models.

As part of formalizing GRUFF, we also made certain decisions about which gender-neutral article, noun, and pronoun forms to include, based on recommendations from advocacy groups and members of the German non-binary community that have become widely adopted. However, there is no social consensus on gender-fair or gender-neutral forms in German at the moment. Some individuals opt for no pronouns in German (Sawall, 2024), and some use *dey* pronouns (Huck, 2021), which are some examples of strategies we do not consider. Furthermore, *English* plays a large role in the linguistic practices of the German queer community (Zieglmeier, 2023; Sawall, 2024). Finally, we do not consider how queer language interacts with other aspects of identity (e.g., names, ethnic identity), and discourse (e.g., domain).

Ethics Statement

As our work touches on socially sensitive language phenomena, there are a number of ethical implications. Our results show how LLMs can harm people, particularly non-binary individuals and users of gender-neutral forms more broadly. If systems that display poor pronoun fidelity are deployed in high-impact domains like education and public services, such errors may reinforce exclusion, reduce trust, and contribute to discriminatory outcomes. The only potential risk we see of our work is that our benchmark assigns certain pronouns as "correct," which do not represent the large variety of pronoun practices people employ in German and beyond. GRUFF should therefore be treated a diagnostic benchmark rather than a normative authority on language use. Scores should be interpreted as indicators of model behaviour rather than definitive judgments about people's identities or acceptable language in all contexts. Responsible use of GRUFF requires transparent reporting, context-aware interpretation of results, and only using the benchmark for evaluation rather than training, due to the possibility of models exploiting shallow heuristics, explained in (Gautam et al., 2024a).

References

Mostafa Abdou, Vinit Ravishankar, Maria Barrett, Yonatan Belinkov, Desmond Elliott, and Anders Søgaard. 2020. The sensitivity of language models and humans to Winograd schema perturbations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7590–7604, Online. Association for Computational Linguistics.

AI@Meta. 2024. Llama 3 model card. https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md. Last Accessed: March 16, 2026.

Rachna Bansal. 2022. A survey on bias and fairness in natural language processing. *arXiv preprint arXiv:2204.09591*.

Marion Bartl, Anirban Mandal, Susan Leavy, and Suzanne Little. 2025. Gender bias in natural language processing and computer vision: A comparative survey. *ACM Computing Surveys*.

Christine Basta, Marta R. Costa-jussà, and José A. R. Fonollosa. 2020. Towards mitigating gender bias in a decoder-based neural machine translation model by adding contextual information. In *Proceedings of the Fourth Widening Natural Language Processing Workshop*, pages 99–102, Seattle, USA. Association for Computational Linguistics.

Connor Baumler and Rachel Rudinger. 2022. Recognition of they/them as singular personal pronouns in coreference resolution. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3426–3432, Seattle, United States. Association for Computational Linguistics.

Henrik Björklund and Hannah Devinney. 2023. Computer, enhence: POS-tagging improvements for non-binary pronoun use in Swedish. In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 54–61, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Bastian Bunzeck and Sina Zarrieß. 2024. The SlayQA benchmark of social reasoning: testing gender-inclusive generalization with neopronouns. In *Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP*, pages 42–53, Miami, Florida, USA. Association for Computational Linguistics.

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.

Yixin Chen, Connor Mahoney, Isabella Grasso, Eman Wali, and 1 others. 2021. Gender bias and under-representation in natural language processing across human languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2022. Theories of "gender" in nlp bias research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Denis Emelin and Rico Sennrich. 2021. Wino-X: Multilingual Winograd schemas for commonsense reasoning and coreference resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8517–8532, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Patrick Fernandes, Kayo Yin, Emmy Liu, André Martins, and Graham Neubig. 2023. When does translation require context? a data-driven, multilingual exploration. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 606–626, Toronto, Canada. Association for Computational Linguistics.

Vagrant Gautam. 2026. Training in step-by-step formal reasoning improves pronominal reasoning in language models. In *To appear in EACL 2026*.

Vagrant Gautam, Eileen Bingert, Dawei Zhu, Anne Lauscher, and Dietrich Klakow. 2024a. Robust pronoun fidelity with english llms: Are they reasoning, repeating, or just biased? *Transactions of the Association for Computational Linguistics*, 12:1755–1779.

Vagrant Gautam, Julius Steuer, Eileen Bingert, Ray Johns, Anne Lauscher, and Dietrich Klakow. 2024b. WinoPron: Revisiting English Winogender schemas for consistency, coverage, and grammatical case. In *Proceedings of the Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 52–66, Miami. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Illi Anna Heger. 2020. Version 3.3 : xier Pronomen ohne Geschlecht.

Tamanna Hossain, Sunipa Dev, and Sameer Singh. 2023. MISGENDERED: Limits of large language models in understanding pronouns. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5352–5367, Toronto, Canada. Association for Computational Linguistics.

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.

Alina Huck. 2021. Effects of language on the daily lives of non-binary people in germany: A qualitative interview study. Bachelor's thesis, Osnabrück University.

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*, FAccT '25, page 150–166, New York, NY, USA. Association for Computing Machinery.

Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. 2025. Multiblimp 1.0: A massively multilingual benchmark of linguistic minimal pairs. *Preprint*, arXiv:2504.02768.

Brian Larson. 2017. Gender as a variable in natural-language processing: Ethical considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Association for Computational Linguistics.

Anne Lauscher, Alex Crowley, and Dirk Hovy. 2022a. Welcome to the modern world of pronouns: Identity-inclusive natural language processing beyond gender. In *Proceedings of the 29th International Conference on Computational Linguistics (COLING)*. Association for Computational Linguistics.

Anne Lauscher, Archie Crowley, and Dirk Hovy. 2022b. Welcome to the modern world of pronouns: Identity-inclusive natural language processing beyond gender. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1221–1232, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Thirteenth International Conference, KR 2012, Rome, Italy, June 10-14, 2012*. AAAI Press.

Shahar Levy, Koren Lazar, and Gabriel Stanovsky. 2021. Collecting a large-scale gender bias dataset for coreference resolution and machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2470–2480, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Kaiji Lu, Piotr Mardziel, Fang Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in neural natural language processing. In *Logic, Language, and Security*. Springer.

Mayflower GmbH. 2024. Llama-3-sauerkrautlm-8b-instruct-awq. Hugging Face Model Repository. Last Accessed: 2025-10-13.

Kevin A McLemore. 2018. A minority stress perspective on transgender individuals' experiences with misgendering. *Stigma and Health*, 3(1):53–64.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 61–72, Brussels, Belgium. Association for Computational Linguistics.

Praneeth Nemani, Y. D. Joel, Pranav Vijay, and F. F. Liza. 2024. Gender bias in transformers: A comprehensive review of detection and mitigation strategies. *Natural Language Processing Journal*.

Anaelia Ovalle, Palash Goyal, Jwala Dhamala, Zachary Jaggers, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2023. "i'm fully who i am": Towards centering transgender and non-binary voices to measure biases in open language generation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1246–1266.

Anaelia Ovalle, Ninareh Mehrabi, Palash Goyal, Jwala Dhamala, Kai-Wei Chang, Richard Zemel, Aram Galstyan, Yuval Pinter, and Rahul Gupta. 2024. Tokenization matters: Navigating data-scarce tokenization for gender inclusive language technologies. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1739–1756, Mexico City, Mexico. Association for Computational Linguistics.

Elisa Forcada Rodríguez, Olatz Perez-de Vinaspre, Jon Ander Campos, Dietrich Klakow, and Vagrant Gautam. 2025. Colombian waitresses y jueces canadienses: Gender and country biases in occupation recommendations from LLMs. In *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 182–194, Vienna, Austria. Association for Computational Linguistics.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. 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 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Danielle Saunders and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7724–7736, Online. Association for Computational Linguistics.

Hannah Sawall. 2024. "Mein Name ist Lena und meine Pronomen sind she/her": Investigating indexicalities of pronoun sharing practices. Master's thesis, University of Duisburg-Essen.

Shanya Sharma, Manan Dey, and Koustuv Sinha. 2022. How sensitive are translation systems to extra contexts? mitigating gender bias in neural machine translation models through relevant contexts. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1968–1984, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Koustuv Sinha, Jon Gauthier, Aaron Mueller, Kanishka Misra, Keren Fuentes, Roger Levy, and Adina

Williams. 2023. Language model acceptability judgements are not always robust to context. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6043–6063, Toronto, Canada. Association for Computational Linguistics.

Niloofar Sobhani, K. Sengupta, and Sarah Jane Delany. 2023. Measuring gender bias in nlp: Incorporating gender-neutral linguistic forms for non-binary gender identities in abusive speech detection. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics.

Karolina Stańczak and Isabelle Augenstein. 2021. A survey on gender bias in natural language processing. *arXiv preprint arXiv:2112.14168*.

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. In *Second Conference on Language Modeling*.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

tresiwalde. 2024. Llama-3-sauerkrautlm-70b-instruct-awq. Hugging Face Model Repository. Last accessed: 2025-10-13.

Jonas-Dario Troles and Ute Schmid. 2021. Extending challenge sets to uncover gender bias in machine translation: Impact of stereotypical verbs and adjectives. In *Proceedings of the Sixth Conference on Machine Translation*, pages 531–541, Online. Association for Computational Linguistics.

Verein für geschlechtsneutrales Deutsch e. V. Pronomen.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Melbourne, Australia. Association for Computational Linguistics.

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.

Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, Jen-tse Huang, Wenxiang Jiao, and Michael R Lyu. 2023. All languages matter: On the multilingual safety of large language models. *arXiv preprint arXiv:2310.00905*.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Sabine Weber, Angelina Wang, Ankush Gupta, Arjun Subramonian, Dennis Ulmer, Eshaan Tanwar, Geetanjali Aich, Hannah Devinney, Jacob Hobbs, Jennifer Mickel, Joshua Tint, Mae Sosto, Ray Groshan, Simone Astarita, Vagrant Gautam, Verena Blaschke, William Agnew, Wilson Y Lee, and Yanan Long. 2026. Queer nlp: A critical survey on literature gaps, biases and trends. *Preprint*, arXiv:2602.16151.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei 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, New Orleans, Louisiana. Association for Computational Linguistics.

Vroni Zieglmeier. 2023. *"Queer English" and "Heteronormative German": Negotiating Linguistic Repertoires and Identity in a Queer Activist Context in Berlin*, pages 1–22. IGI Global Scientific Publishing.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.

A Further Details on Template Assembly

The number of dataset instances in Table 3 is calculated as follows, where the symbols are the components of the German pronoun fidelity task described in Section 3:

$$\text{0 distractors: } 40\ i(a_a, e_a, p_a) * 240\ t(a_a, e_a, p)$$

$$\text{1 distractor: } 40\ i(a_a, e_a, p_a) * 240\ t(a_a, e_a, p) * 12\ d(a_b, e_b, p_b)$$

$$\text{2–5 distractors: } 40\ i(a_a, e_a, p_a) * 240\ t(a_a, e_a, p) * 12\ d(a_b, e_b, p_b) * P * d(p_b)$$

For the templates with 2–5 distractors, the number of permutations (P) is calculated as follows:

$$\text{2 distractors:} \quad P(4,1) = \frac{4!}{(4-1)!} = \frac{24}{6} = 4$$

$$\text{3 distractors:} \quad P(4,2) = \frac{4!}{(4-2)!} = \frac{24}{2} = 12$$

$$\text{4 distractors:} \quad P(4,3) = \frac{4!}{(4-3)!} = \frac{24}{1} = 24$$

$$\text{5 distractors:} \quad P(4,4) = \frac{4!}{(4-4)!} = \frac{24}{1} = 24$$

The number of dataset instances for 4 and 5 distractors is identical because of the permutation calculation shown above. Table 6 shows the calculations for the dataset construction for each distractor setting in more detail.

B Human Validation

The validation was designed to ensure inter-annotator reliability through identical task assignment. Data was provided to annotators in Google Sheets for automatic validation. 600 filled context templates were provided in one sheet. A second sheet contained 600 task dataset instances (100 randomly sampled instances each for settings with 0–5 distractor sentences). Annotators achieved 100% accuracy in reproducing the correct pronoun on these instances. Below we present annotator demographics as well as the original annotation instructions (in German).

B.1 Annotator Demographics

Three annotators participated in the human validation, one of whom is an author of this work. The two other annotators were students recruited from the university community; annotators 1 and 2 were not paid, and annotator 3 completed the annotations as part of his work contract as a student assistant.

Annotator 1 (Author)

- **Age:** 25
- **Gender:** Male
- **Level of German proficiency:** C2 (native speaker)
- **Special knowledge/experience with German gender-fair language and neopronouns:** High. Annotator 1 is the lead author of the work and conducted extensive research on this topic.

Annotator 2

- **Age:** 26
- **Gender:** Female
- **Level of German proficiency:** C2 (native speaker)
- **Special knowledge/experience with German gender-fair language and neopronouns:** No prior experience before validating the dataset.

Annotator 3

- **Age:** 27
- **Gender:** Male
- **Level of German proficiency:** C2 (native speaker)
- **Special knowledge/experience with German gender-fair language and neopronouns:** Prior professional work involving German gender-fair language and neopronoun usage.

Number of distractors	$\mathbf{t(a_a, e_a, p)}$		$\mathbf{i(a_a, e_a, p_a)}$		$\mathbf{C}$		$\mathbf{d(a_b, e_b, p_b)}$		$\mathbf{p_b}$		$\mathbf{P}$		Total instances
0	240	×	10	×	4							=	9,600
1	240	×	10	×	4	×	4	×	3			=	115,200
2	240	×	10	×	4	×	4	×	3	×	4	=	460,800
3	240	×	10	×	4	×	4	×	3	×	12	=	1,382,400
4	240	×	10	×	4	×	4	×	3	×	24	=	2,764,800
5	240	×	10	×	4	×	4	×	3	×	24	=	2,764,800

Table 6: Computation of the total number of GRUFF instances per distractor setting. Total instances are obtained by multiplying the factors in each row from left to right. Here, $t(a_a, e_a, p)$ denotes task templates, $i(a_a, e_a, p_a)$ explicit introduction context templates, C grammatical cases, $d(a_b, e_b, p_b)$ explicit distractor context templates, p_b distractor pronoun options, and P implicit distractor sentence permutations.

B.2 Guidelines

Anmerkungsinstruktionen

Instruktionen zu Aufgabe 1

Zusammen mit diesem Annotationsprotokoll haben Sie den Link zu einem Google Sheet erhalten. Das Sheet enthält 2 Datenspalten und 2 Aufgabenspalten mit zufälligen Daten. Die Datenspalten bestehen aus:

- Sätzen, die Sie auf grammatikalische Korrektheit bewerten sollen;
- Fragen zu Pronomen in den Sätzen, die Sie beantworten sollen

Bitte seien Sie präzise in Ihren Zuordnungen und ändern Sie nicht die Reihenfolge der Daten. Die Spalten haben eine integrierte Datenvalidierung und wir werden weitere Tests durchführen, um eine konsistente Annotation zu überprüfen.

Grammatikalische Korrektheit

In der Spalte "Grammatikalisch korrekt?" geben Sie bitte Ihre Bewertung des Satzes nach Standard-Deutsch ein. Die Anmerkungsoptionen sind:

- **Grammatikalisch korrekt** (für flüssige, syntaktisch gültige und semantisch plausible Sätze)
- **Grammatikalisch nicht korrekt** (für Sätze mit Tippfehlern, grammatikalischen Problemen, oder wenn der Satz eine Situation beschreibt, die keinen Sinn ergibt oder einfach seltsam klingt)
- **Nicht sicher** (wenn Sie nicht sicher sind, ob der Satz eindeutig grammatikalisch korrekt oder grammatikalisch nicht korrekt ist)

Beispiele:

- *Der Psychologe fühlte sich erholt; seine Nacht verlief ruhig.*
  ¯ grammatikalisch korrekt
- *Der Fahrer sagte dem Fahrgast er kann die Fahrt bar bezahlen aber vergaß das Wechselgeld.*
  ¯ grammatikalisch nicht korrekt (wegen fehlender Kommasetzung, nicht korrekter deutscher Grammatik oder unklarer Bezüge)

Fragen zu Pronomen

Jeder Satz enthält ein Pronomen, und die Spalte "Frage" fragt, ob es sich auf eine im Satz erwähnte Person mit der jeweiligen Berufsbezeichnung bezieht oder nicht. Die Anmerkungsoptionen sind:

- **ja** (wenn sich das Pronomen auf die Person mit der genannten Berufsbezeichnung bezieht)
- **nein** (wenn sich das Pronomen nicht auf die Person mit der genannten Berufsbezeichnung bezieht)
- **nicht sicher** (wenn Sie nicht sicher sind, ob sich das Pronomen auf die Person mit der genannten Berufsbezeichnung bezieht)

Beispiele:

- *Der Psychologe fühlte sich erholt; ihre Nacht verlief ruhig.*
  Frage: Bezieht sich ihre auf die Person mit der Berufsbezeichnung Psychologe?
  ¯ Nein
- *Die Ausbilderin benötigte Kaffee, weil ihre Schicht sehr früh begonnen hatte.*
  Bezieht sich das Pronomen ihre auf die Person mit der Berufsbezeichnung Ausbilderin?
  ¯ Ja

Model	HuggingFace Identifier
Encoder-only models	
GBERT-BASE	deepset/gbert-base
GBERT-LARGE	deepset/gbert-large
MBERT	google-bert/bert-base-multilingual-cased
XLM-ROBERTA	FacebookAI/xlm-roberta-base
Decoder-only models	
SAUERKRAUTLM-8B	mayflowergmbh/Llama-3-SauerkrautLM-8b-Instruct-GGUF
SAUERKRAUTLM-70B	tresiwalde/Llama-3-SauerkrautLM-70b-Instruct-AWQ
LLAMA-3.3-70B	casperhansen/llama-3.3-70b-instruct-awq

Table 7: List of models along with their HuggingFace identifiers.

Instruktionen zu Aufgabe 2

Zusammen mit diesem Anmerkungsprotokoll haben Sie einen Link zu einem Google Sheet erhalten. Das Tabellenblatt enthält 1 zufällig erstellte Datenspalte und 2 Aufgabenspalten. Die Datenspalten bestehen aus:

- Fragen zu Pronomen in den Sätzen, die Sie beantworten sollen;
- Sätzen, die Sie auf grammatikalische Korrektheit bewerten sollen

Bitte seien Sie präzise in Ihren Zuordnungen und ändern Sie nicht die Reihenfolge der Daten. Die Spalten haben eine integrierte Datenvalidierung und wir werden weitere Tests durchführen, um eine konsistente Annotation zu überprüfen.

Fragen zu Pronomen

Jede Zeile in der Datenspalte besteht aus mehreren Sätzen, von denen genau ein Satz eine Lücke enthält. Ihre Aufgabe ist es, das passende Pronomen zu bestimmen, um die Lücke zu füllen, und es in die Spalte "Pronomen" einzutragen. Hierbei bedeutet passend sowohl korrekt in Form, Kasus und Geschlecht.
Die Aufgaben sind derartig konzipiert, dass sie eindeutig sind, daher geben Sie bitte nur eine Lösung an und ändern Sie nicht die Reihenfolge der Daten. Beispiele: Der Fahrer fühlte sich unglücklich, weil er nicht genug Geld verdiente. Der Fahrer fragte sich, ob ___ einen Kredit aufnehmen sollte. => er

Grammatikalische Korrektheit

Zusätzlich gibt es in diesem Tabellenblatt eine Spalte mit dem Namen "Letzter Satz grammatikalisch korrekt?" Für diese Spalte betrachten Sie bitte den letzten Satz, der die Lücke beinhaltet und geben Ihre Bewertung des Satzes nach Standard-Deutsch ein. Die Anmerkungsoptionen sind:

- **Grammatikalisch korrekt** (für flüssige, syntaktisch gültige und semantisch plausible Sätze)
- **Grammatikalisch nicht korrekt** (für Sätze mit Tippfehlern, grammatikalischen Problemen, oder wenn der Satz eine Situation beschreibt, die keinen Sinn ergibt oder einfach seltsam klingt)
- **Nicht sicher** (wenn Sie nicht sicher sind, ob der Satz eindeutig grammatikalisch korrekt oder grammatikalisch nicht korrekt ist)

Beispiele:

- *Der Berater war voll, weil eine große Mahlzeit ihn gefüllt hatte. Der Berater akzeptierte einen Entwurf des Lebenslaufs, damit man ___ Rückmeldung geben ließ.*
  => grammatikalisch korrekt
- *Die Brandbekämpferin benötigte keinen Kaffee, weil ihre Schicht heute später begann. Die Brandbekämpferin sprach auf ___ Schwierigkeiten während der Arbeit.*
  => grammatikalisch nicht korrekt

Wichtige Hinweise:

- Arbeiten Sie sorgfältig und konzentriert
- Verwenden Sie nur die angegebenen Anmerkungsoptionen
- Verändern Sie nicht die Reihenfolge der Daten

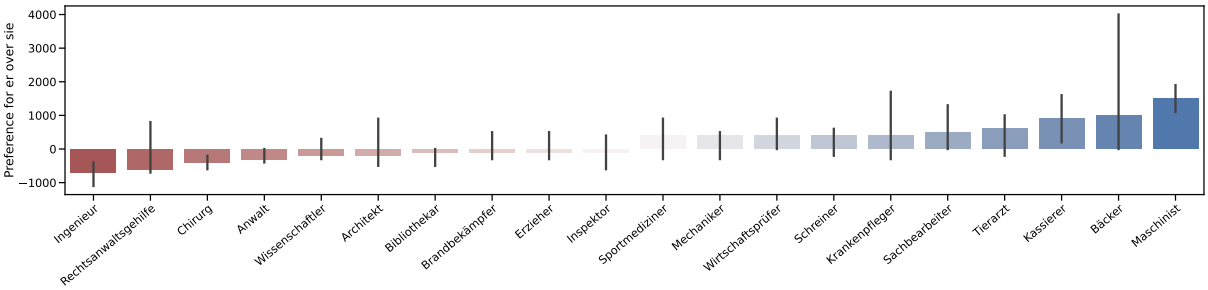


Figure 9: Top 10 occupations over-resolved to nominative pronouns *sie* (feminine) on the right, versus *ihn* (masculine) on the left, in a pattern indicative of stereotyping.

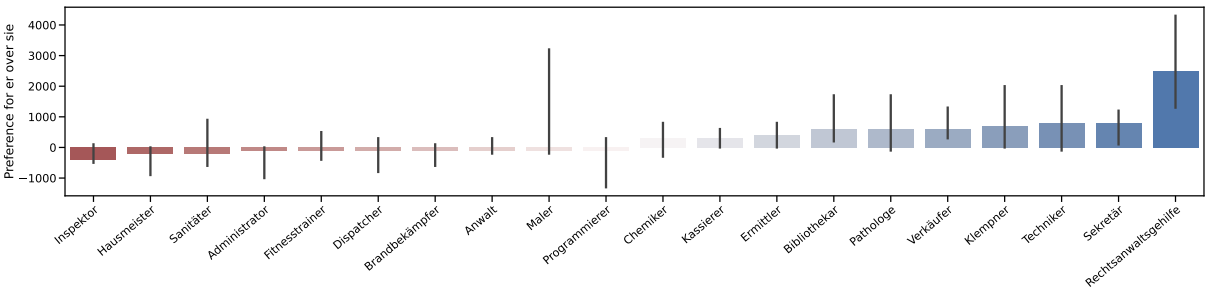


Figure 10: Top 10 occupations over-resolved to dative pronouns *sie* (feminine) on the right, versus *ihn* (masculine) on the left, in a pattern indicative of stereotyping.

- Bei Unsicherheiten wählen Sie "nicht sicher" statt zu raten
- Achten Sie auf Standard-Deutsch als Bewertungsmaßstab

C Additional Model Details

All models are shown along with their HuggingFace identifiers in Table 7.

D Disentangling performance with *er* and *xier*

Figure 5 shows that performance with *en* pronouns is substantially lower than *xier* pronouns, despite both being neopronoun variants that are likely sparse in training data. However, in all results within the main paper, *en* pronouns are always paired with the De-e article and noun agreement system, while *xier* pronouns are paired with the slightly more common Sternchen system. In this appendix, we swap these pairings in order to disentangle whether poor performance is due to the pronoun or the article-noun agreement system.

As Table 8 shows, performance remains extremely low with *en* pronouns, regardless of what gender-neutral article-noun agreement system they are paired with. We hypothesize that this is due to tokenization conflicts between *en*'s use as a pronoun and this substring's occurrence in many other contexts in German, as part of the prefix *ent-*, verb infinitive endings (-en), and so on. Similarly, the dative *em* also frequently appears at the beginning of German words. In contrast, *xier* pronouns are unique, which might explain why models are able to handle them more successfully. Further research into neopronoun tokenization in German is necessary to further explain these results, similar to the English analyses in Ovalle et al. (2024).

Article-noun system	**Pronoun**	
	en	**xier**
De-e	23.49 ± 26.05	61.13 ± 30.04
Sternchen	23.34 ± 24.66	66.74 ± 29.54

Table 8: Disentangling performance with *xier* and *en* by pairing them with alternative gender-neutral article-noun systems.

E Additional Results

In this section we show additional results for occupational biases, quantified by how much models over-resolve a particular traditional pronoun to an occupation, overriding signals from articles and nouns about the appropriate grammatical gender.

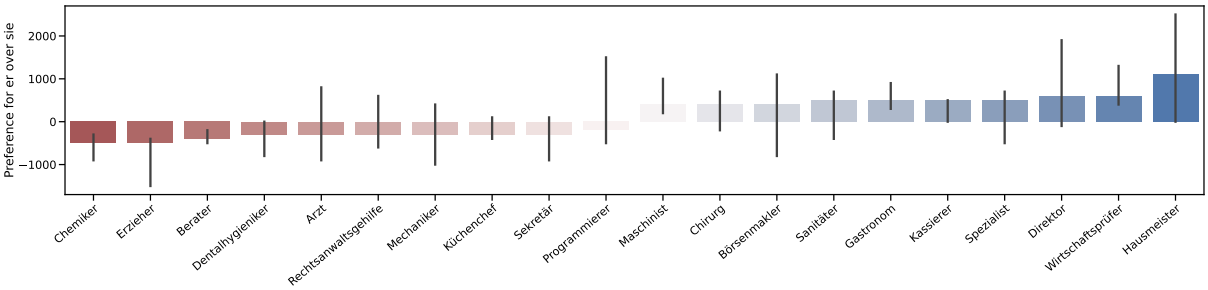


Figure 11: Top 10 occupations over-resolved to genitive pronouns *sie* (feminine) on the right, versus *ihn* (masculine) on the left, in a pattern indicative of stereotyping.

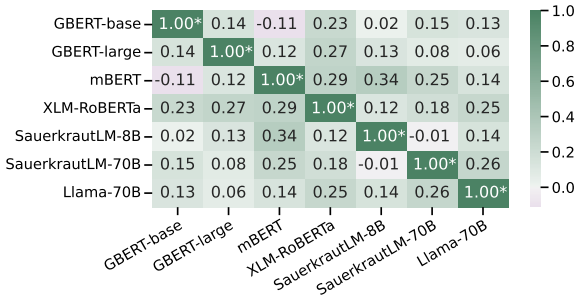


Figure 12: Spearman's correlation between the stereotypical biases of models for nominative pronouns. * indicates statistical significance ($\alpha = 0.05$).

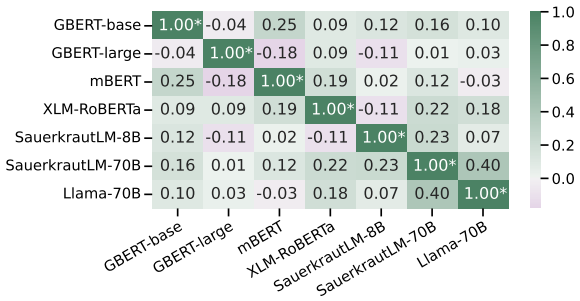


Figure 13: Spearman's correlation between the stereotypical biases of models for dative pronouns. * indicates statistical significance ($\alpha = 0.05$).

Accusative pronouns are shown in the main text, therefore this appendix shows results with nominative pronouns (Figure 9), dative pronouns (Figure 10), and genitive pronouns (Figure 11).

Finally, we also show model correlations on occupational biases for the grammatical cases not shown in the main text: Nominative (Figure 12), dative (Figure 13), and genitive (Figure 14).

F AI Assistance

AI assistance was used by the lead author to summarize some research papers and rewrite the original code, and by the second author for suggesting

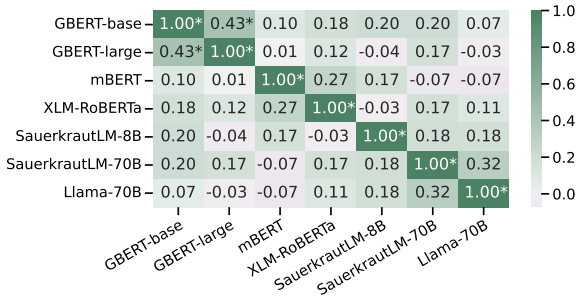


Figure 14: Spearman's correlation between the stereotypical biases of models for genitive pronouns. * indicates statistical significance ($\alpha = 0.05$).

formulations for English text and polishing some of the writing.