The Lou Dataset
Exploring the Impact of Gender-Fair Language in German Text Classification

Andreas Waldis* [1,2], **Joel Birrer**[2], **Anne Lauscher**[3], **Iryna Gurevych**[1]
[1]Ubiquitous Knowledge Processing Lab (UKP Lab)
Department of Computer Science and Hessian Center for AI (hessian.AI)
Technical University of Darmstadt
[2]Information Systems Research Lab, Lucerne University of Applied Sciences and Arts
[3]Data Science Group, University of Hamburg
www.ukp.tu-darmstadt.de www.hslu.ch

Abstract

Gender-fair language, an evolving German linguistic variation, fosters inclusion by addressing all genders or using neutral forms. Nevertheless, there is a significant lack of resources to assess the impact of this linguistic shift on classification using language models (LMs), which are probably not trained on such variations. To address this gap, we present Lou, the first dataset featuring high-quality reformulations for German text classification covering seven tasks, like stance detection and toxicity classification. Evaluating 16 mono- and multi-lingual LMs on Lou shows that gender-fair language substantially impacts predictions by flipping labels, reducing certainty, and altering attention patterns. However, existing evaluations remain valid, as LM rankings of original and reformulated instances do not significantly differ. While we offer initial insights on the effect on German text classification, the findings likely apply to other languages, as consistent patterns were observed in multi-lingual and English LMs.[1]

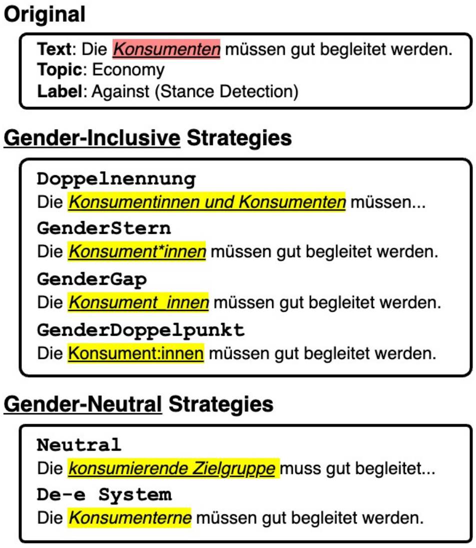
Warning: This paper contains offensive text.

huggingface.co/datasets/tresiwalde/lou

UKPLab/lou-gender-fair-reformulations

1 Introduction

Language is constantly evolving. This change involves dialect variations of specific localities or the slang of certain generations (Keidar et al., 2022; Sun and Xu, 2022). Such linguistic variations reflect societal changes, where negotiations emerge and influence how individuals speak. Notable shifts are gender-fair formulations in languages with feminine and masculine genders, like German or French. The societal importance of such linguistic variations is reflected in concrete guidelines, like those from the EU Parliament.[2] Concretely, inclusive and neutral strategies (Lardelli and Gromann, 2023), illustrated in Figure 1 for German, serve as tools to reduce gender stereotypes and discrimination (Sczesny et al., 2016) and to meet the UN goal of gender equality.[3] Inclusive strategies (Doppelnennung) address both genders explicitly (*Konsumentinnen und Konsumenten*, *consumer.FEM.PL*[4] *and consumer.MASC.PL*), or any gender using special



Original

Text: Die Konsumenten müssen gut begleitet werden.
Topic: Economy
Label: Against (Stance Detection)

Gender-Inclusive Strategies

Doppelnennung
Die Konsumentinnen und Konsumenten müssen...
GenderStern
Die Konsument*innen müssen gut begleitet werden.
GenderGap
Die Konsument_innen müssen gut begleitet werden.
GenderDoppelpunkt
Die Konsument:innen müssen gut begleitet werden.

Gender-Neutral Strategies

Neutral
Die konsumierende Zielgruppe muss gut begleitet...
De-e System
Die Konsumenterne müssen gut begleitet werden.

Figure 1: A German stance detection instance from the Lou dataset. We reformulate the masculine formulation *Konsumenten* (*consumers*) regarding six inclusive or neutral strategies, highlighted in yellow. Translation: *Consumers must be well supported.*

*The corresponding author is andreas.waldis@live.com

[1] Data is also available in a online archive.

[2] Available online.

[3] https://sdgs.un.org/goals/goal5.

[4] In English, we indicate the German gender following the Lepzig Glossing Rules. We use *FEM.MASC.NEUT* for formulations neutralizing or addressing any gender.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 10604–10624
November 12-16, 2024 ©2024 Association for Computational Linguistics

characters (GenderStern, *Konsument*innen*, consumer.FEM.MASC.NEUT.PL). Neutral strategies avoid specific genders using neutral terms (*konsumierende Zielgruppe, consuming target group*).

As language models (LMs) are inherently biased from training on data from the past (Kurita et al., 2019; Srivastava et al., 2022; Attanasio et al., 2023), current research increasingly addresses the question '*How does gender-fair language impact LMs?*'. This includes examining the gender bias of machine translation regarding gender-fair language (Paolucci et al., 2023; Piergentili et al., 2023) and pronouns (Lauscher et al., 2023; Amrhein et al., 2023), or fundamentally concentrate on LMs' limitations when interpreting pronouns (Brandl et al., 2022; Hossain et al., 2023; Gautam et al., 2024). Despite the widespread application of text classification (Wang et al., 2018; Zhang et al., 2015), there is a notable lack of resources and scientific effort to examine the impact of gender-fair language on classification systems. Consequently, it remains unclear whether LMs perform consistently without unwanted side effects when processing gender-fair language under the inclusive aim.

To address this research gap, we introduce Lou (§ 3), a German dataset featuring high-quality text reformulations following six reformulation strategies (Figure 1). While creating these reformulations, we also examine the reliability of amateurs with moderate experience using gender-fair language compared to professionals with linguistic backgrounds. Using the resulting 3.6k reformulated instances from seven classification tasks, we systematically evaluate the impact of gender-fair language and specific reformulation strategies for fine-tuning and zero-shot classification setups covering 16 LMs (§ 4). We then compare these results with the original instances (§ 5 and § 6) to address four key research questions.

RQ1: Do amateurs produce gender-fair reformulations with sufficient quality? No. Amateurs struggle to consistently apply reformulation strategies, with an error rate of up to 31% for GenderStern, hinting at a lack of societal establishment and standardization.

RQ2: Does gender-fair language impact German text classification during inference? Yes. Gender-fair language leads to task performance variations in macro F_1 score, ranging from -1.0 to +4.0, and flips up to 10.9% of individual predictions. However, the effects of distinct reformulation strategies vary. Those making minimal sentence adjustments, such as GenderStern, tend to enhance performance. At the same time, neutralization-focused strategies, like De-e or Neutral, which substantially alter the text, generally lead to lower performance.

RQ3: Do LMs process gender-fair language differently? Yes. Gender-fair language notably impacts how the lower LM layers process reformulated instances compared to the original ones. Further, altered attention patterns and decreased prediction certainty lead to the observed label flips.

RQ4: What are the practical implications of encountering gender-fair language? Existing datasets and evaluations remain valid (consistent LM rankings), but significant label flips occur, especially in tasks with lower absolute performance. Observing mostly syntactic and consistent effects across German, English, and multi-lingual LMs, our findings are relevant to other languages with similar reformulation strategies, such as using the interpoint (·) as an inclusion character in French.

Contributions We lay the foundation for studying the impact of gender-fair language on classification and make three key contributions: **1)** We present the first high-quality dataset of reformulated text instances for German classification tasks and provide insights into the practical annotation challenges. **2)** A systematic evaluation underscores the practical value of German-specialized LMs and reveals the substantial impact of gender-fair language and individual reformulation strategies on individual predictions. **3)** We offer concrete guidance on how LMs process gender-fair language differently, highlighting the necessity to consider such fine-grained linguistic variations.

2 Preliminaries

2.1 Gender-Fair Language

We define *gender-fair* language as a specific linguistic phenomenon that replaces the generic formulations, either the feminine or the predominant masculine one. With alternative formulations, this linguistic shift reduces gender stereotypes and discrimination by comprehensively addressing people (Sczesny et al., 2016). As in Lardelli and Gromann (2023), gender-fair embodies both *inclusive* and *neutral* language (Figure 1). Inclusive language addresses either the masculine or feminine gender explicitly or uses characters like the gender star (*) (German) to address everyone on the gender spectrum, including those identifying with no gender. Differently, neutral language prevents gender-specific formulations with alternative terms.

2.2 Gender-Fair Reformulation Strategies

Different strategies guide the formulation of gender-fair language. Specifically, we consider the following inclusive or neutral ones.[5]

i) Binary Gender Inclusion (Doppelnennung) explicitly mentions the feminine and masculine but ignores others like agender. For example, *Ärzte* (doctor.MASC.PL) is transformed into *Ärztinnen und Ärzte* (*doctor.FEM.PL and doctor.MASC.PL*).

ii) All Gender Inclusion explicitly addresses every gender, including agender, non-binary, or demi-gender, using a gender gap character pronounced with a small pause. In this work, we consider three commonly used gender characters: GenderStern (*), GenderDoppelpunkt (:), and GenderGap (_). For example, *Ärzte* (*doctor.MASC.PL*) is turned into *Ärzt*innen*, *Ärzt:innen*, or *Ärzt_innen* (doctor.FEM.MASC.NEUT.PL).

iii) Gender Neutralization (Neutral) avoids naming a particular gender using neutral terms, like *ärztliche Fachperson* (*medical professional*).

iv) De-e (Neosystem) is a well-specified system that emerged from a significant community-driven effort.[6] It introduces a fourth gender, including new pronouns, articles, and suffixes. For example, *der Arzt* (*the doctor.MASC.SG*) is changed to *de Arzte* (*the doctor.FEM.MASC.NEUT.SG*).

3 The Lou Dataset

Lou marks the largest collection of reformulated instances for German text classification. With 3.6k reformulations following six reformulation strategies, Lou enables thoroughly assessing the impact of gender-fair language and the individual strategies across seven classification tasks. In the following, we discuss the used data (§ 3.1) before focusing on the reformulation study (§ 3.2).

3.1 Data

We start from three German classification datasets: Detox (Demus et al., 2022), GermEval-2021 (Risch et al., 2021), and X-Stance (Vamvas and Sennrich, 2020). We select them since they cover established tasks and minimize the reformulation effort because single instances are annotated with multiple labels. For example, Detox provides labels for sentiment analysis, hate-speech, and toxicity detection. Further details and statistics of the datasets are provided in the Appendix § A.4.

X-Stance (Vamvas and Sennrich, 2020) annotates multi-lingual texts (*de*, *fr*, *it*) with their stance (*favor* or *against*) regarding 12 topics.

GermEval-2021 (Risch et al., 2021) annotates social media texts with three binary properties: toxicity, fact-claiming, and engaging.

Detox (Demus et al., 2022) annotates social media texts regarding sentiment, hate-speech, and toxicity. Following original instructions, we derive classification labels from the provided raw annotations. Because the additional training data used in the original paper is unavailable, we sub-sample a more label-balanced train set.

3.2 Reformulation Study

For every dataset, we sampled 200 test instances containing at least one gender-specific term, identified via Diversifix[7]. We employ an iterative approach involving both eight amateurs and two professionals to ensure the quality of gender-fair reformulations. While amateurs have an average self-determined moderate experience of using gender-fair language (more details in Appendix § A.2), professionals have a linguistic background and use it daily. Within this study, we ensure *high-quality*, meaning that specific reformulation strategies are correctly applied without grammatical errors, and *consistency* as semantics and annotated task labels of the original instances are preserved. Therefore, we avoid using large LMs for annotation as they do not produce gender-fair language with sufficient quality (Savoldi et al., 2024).

i) Amateur Annotators First, we ask each of the eight amateurs to reformulate 50 distinct instances from X-Stance and GermEval-2021 regarding the Doppelnennung, GenderStern, and Neutral strategies, leading to 1.2k distinct reformulations. The annotators need to fulfill the reformulation according to a given strategy. Other grammatical errors should be ignored to ensure

[5] As these strategies are proper names, we do not translate from German to English.

[6] Find more details online at https://geschlechtsneutral.net

[7] A tool for gender-fair language (https://diversifix.org/).

Task	Instance	Label
	X-Stance (Vamvas and Sennrich, 2020)	
Stance	**Topic**: Integration, **Text**: Integration ist das A und O im Umgang mit Ausländischen Mitbürger*innen $_{\text{GenderStern}}$.	favor
Translation	**Topic**: Integration, **Text**: Integration is the be-all and end-all when dealing with foreign citizens.	
	GermEval-2021 (Risch et al., 2021)	
Engaging	**Text**: Die Möglichkeit, dass Trump gewinnt ist groß, weil seine Konkurrenz $_{\text{Neutral}}$ so schwach ist.	engaging
Fact-Claiming	**Text**: Die Möglichkeit, dass Trump gewinnt ist groß, weil ens Gegnere $_{\text{De-e}}$ so schwach ist.	no fact claimed
Toxicity	**Text**: Die Möglichkeit, dass Trump gewinnt ist groß, weil seine Gegnerin oder Gegner $_{\text{Doppelnennung}}$ so schwach ist.	not toxic
Translation	**Text**: The possibility that Trump will win is high because of his opponents.	
	Detox (Demus et al., 2022)	
Hate-Speech	**Text**: NRW Lusche ihr seid scheiße nein du bist es! Ein Freund $_{\text{Masculine}}$ aller Schwulen Spahnferkels.	hate-speech
Sentiment	**Text**: NRW Lusche ihr seid scheiße nein du bist es! Ein:e Freund:in $_{\text{GenderDoppelpunkt}}$ aller Schwulen Spahnferkels.	negative
Toxicity	**Text**: NRW Lusche ihr seid scheiße nein du bist es! Ein_e Freund_in $_{\text{GenderGap}}$ aller Schwulen Spahnferkels.	toxic
Translation	**Text**: NRW losers you suck, no you are! A friend of all gay Spahn pigs.	

Table 1: Example of the seven German classification tasks in Lou, along with their translations. Gender-fair reformulation strategies (subscript) are highlighted in yellow, and masculine formulations are in orange.

	Doppelnennung	GenderStern	Neutral	Avg.
X-Stance	7.5%	10.5%	10.0%	9.3%
Germeval-2021	21.0%	31.0%	21.5%	24.5%
Avg.	14.3%	20.8%	15.6%	16.9%

Table 2: Percentage of proofreading corrections compared to amateur reformulations.

the dataset's validity. We speed up the reformulation with automatic suggestions from Diversifix and highlight the relevant parts (as in Figure 1) and provide examples in the interface.[8]

ii) Professional Proofreading We validate the amateur reformulations with professional proofreading (**P1**) to ensure *high-quality* of the reformulations. Table 2 shows substantial corrections were necessary by **P1**, **hinting at the substantially degraded quality of amateur reformulations** (**RQ1**). Corrections were necessary in 16.9% of the cases and up to 31.0% for GenderStern on the GermEval-2021 data. At the same time, the nature of the original text matters as GermEval-2021 seems more challenging for amateurs than X-Stance, as its texts are generally longer and less grammatically consistent (social media). Further categorization of corrections (find details in Appendix § A.3) shows that amateurs particularly struggle, among others, when adapting pronouns (*ein* into *ein*e*) or handling the grammatical number (*Studenten*innen* instead of *Student*innen*).

iii) Proofreading Verification Due to the substantial corrections during proofreading, we verify the reliability of **P1** with another verification round using a subset of 20% of the instances. Those were verified by another professional proofreader (**P2**). We find a high agreement of 95% between **P1** and **P2**, confirming the reliability of **P1**.

iv) Detox Dataset and De-e Strategy Based on their high reliability, we conduct a fourth iteration with **P1** including 200 instances from the Detox and the De-e strategy for all three datasets.

3.3 Dataset Composition

Using the reformulations, we compose the final Lou dataset with instances for the seven tasks of X-Stance, GermEval-2021, and Detox (200 ones each). For all 600 instances, Lou provides reformulations for Doppelnennung, GenderStern, GenderGap, GenderDoppelpunkt, Neutral, and De-e leading to 3.6k distinct reformulations. Note, we use a regular expression to generate instances for GenderGap automatically and GenderDoppelpunkt based on GenderStern by replacing the star character (*****) with a colon (**:**) or gap (**_**). Ultimately, we ensure *consistency* of the reformulations and manually verify a subset of them to ensure that task labels remain valid. We find that semantics and the specific task label of instances are unchanged, confirming their validity for analysis (Appendix § A.9).

4 Experimental Setup

The following section outlines the experimental setups used to assess the impact of gender-fair language on classification during inference, including learning paradigms (§ 4.1), used encoder and decoder LMs (§ 4.2), and evaluation (§ 4.3).

[8] Annotation guidelines are available online.

4.1 Learning Paradigm

Fine-Tuning We tune encoder LMs on the original train and dev (without reformulations) set of the seven Lou tasks for five epochs with early stopping. We select the best batch size $\{8, 16, 32\}$ and learning rate $\{5 \cdot 10^{-5}, 2 \cdot 10^{-5}, 1 \cdot 10^{-5}\}$ based on the dev performance for every LM and task across three random seeds. Then, we tune LMs across ten random seeds to ensure numeric stability.

In-Context Learning (ICL) We evaluate open and closed decoder LMs using their textual response to zero-shot prompts. We reuse prompting templates of previous work whenever available and evaluate additional three paraphrased templates to account for variabilities (Mizrahi et al., 2024).[9]

4.2 Models

We evaluate the following five LM types, including ten encoders and six decoders.[10] Apart from German-specialized and multi-lingual LMs, we further consider English-specialized ones. As they were mainly trained in English text, they represent the lower bound without a fine-grained understanding of the German language. Thus, we assume LMs mainly capture lexical features if English LMs perform competitively.

German We tune four German encoder LMs (Chan et al., 2020): GBERT-base, GBERT-large, GELECTRA-base, and GELECTRA-large.

Multi-Lingual We consider three mulit-lingual encoder LMs: mBERT-base (Devlin et al., 2019), XLM-R-base (Conneau et al., 2020), and mDeBERTa-base v3 (He et al., 2023).

English We evaluate three English LMs: BERT-base (Devlin et al., 2019), RoBERTa-base (Liu et al., 2019), and DeBERTa-base v3 (He et al., 2023).

Instruction-Tuned (IT) For ICL, we consider four decoder LMs Llama-3-8B and Llama-3-70B (AI@Meta, 2024), gpt-3.5-turbo, and gpt-4o (Ouyang et al., 2022).

German IT In addition, we consider two German specialized large LMs based on Llama-3: Sauerkraut-8B and Sauerkraut-70B.[11]

4.3 Evaluation

We assess the impact of gender-fair language by comparing predictions on the original test instances with the reformulated ones per LM. Specifically, we analyze the impact on task level using the F_1 macro score and on the instance level by counting prediction flips under gender-fair language. We report average and standard deviation across ten random seeds. We report results on the Lou subset of 200 test instances per task. Results on these subsets significantly aligned with the full test set, with a Pearson correlation of $\rho = 0.97$.

5 Results

We discuss results obtained across the seven Lou tasks. First, we establish our baseline with results on the original samples (i). Next, we focus on **RQ2** and the substantial impact of gender-fair language on aggregated evaluation (ii, iii) and individual predictions (iv). Addressing **RQ4**, we confirm that existing datasets and evaluations retain their validity under gender-fair language (v).

i) The value of German specialized LMs. Figure 2 shows the aggregated Lou performance, emphasizing the necessity of specialized German LMs to achieve competitive results. On average, German decoders (53.7) outperform general ones by 2.1 points. Similarly, German and multilingual encoders (60.9, 56.9) surpass their English counterparts by 10.5 and 6.1 points, respectively. Notably, **mDeBERTa demonstrates its practical value for German tasks, marginally outperforming the German-specific encoders**, particularly in challenging scenarios with highly label imbalances like in the Detox Hate-Speech task (see Table 5 in the Appendix). The surprisingly strong performance of its English counterpart (DeBERTa) suggests that these LMs may rely more on lexical features than on a nuanced linguistic understanding of the German language. This assumption is supported by the substantially larger vocabulary sizes of mDeBERTa (250k) and DeBERTa (128k) compared to the 31k tokens of GBERT and GELECTRA. Interestingly, model size appears less critical, as GBERT-base and GELECTRA-base do not significantly underperform compared to their larger versions. However, in ICL, model size plays a crucial role, as Llama-3-70B gains 9 points over Llama-3-8B, and GPT-4o 9.7 points over GPT-3.5-turbo. Interestingly, this trend does not hold for German decoders, where Sauerkraut-8B remains notably competitive

[9] More detail in the Appendix § A.6

[10] More details are in Appendix § A.7

[11] Available on Huggingface.

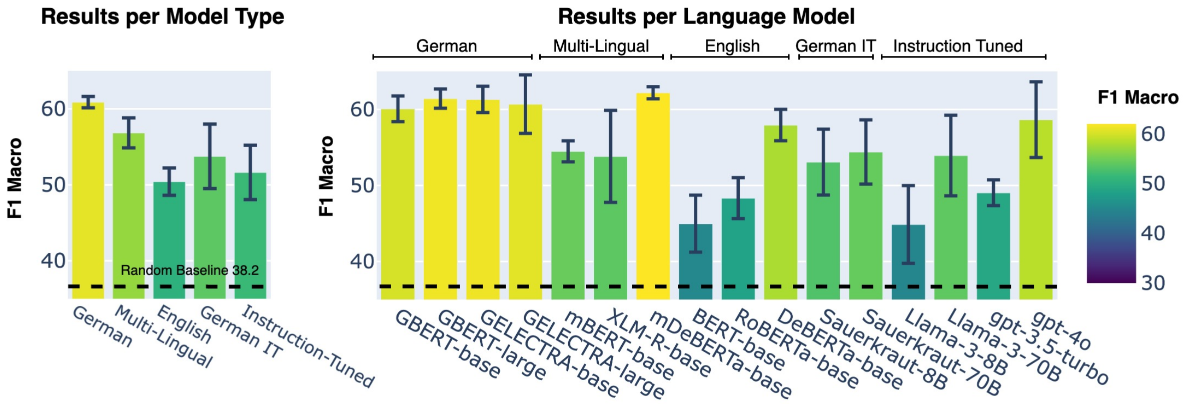


Figure 2: Mean performance and standard deviation, averaged over the seven Lou tasks and seeds (fine-tuning) or prompting templates (ICL) by model type (*left*) or specific LM (*right*).

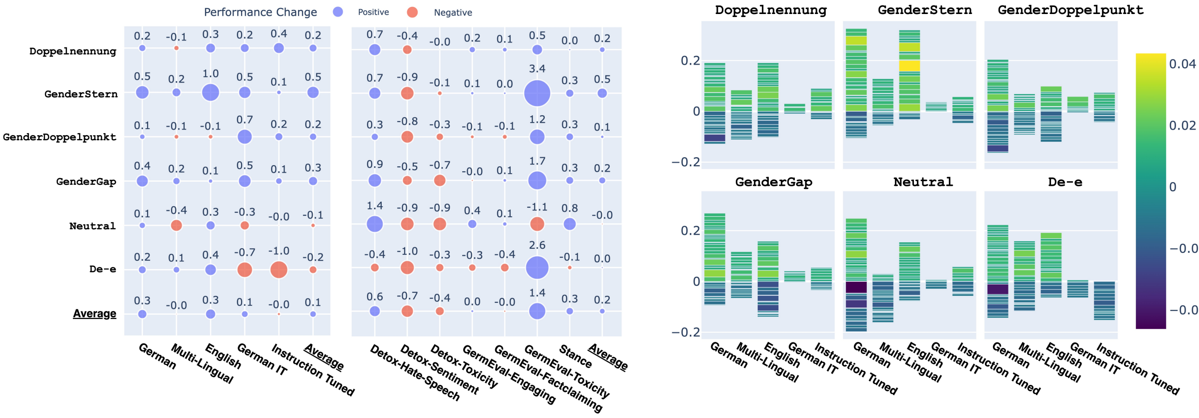


Figure 3: Difference between original and reformulated instances for strategies, model types, and tasks in average F_1 macro (*left*). The size and the color indicate the difference, whether positive (blue) or negative (red). On the *right*, we stack the average difference per LM and seed or prompt template for the model types and strategies.

with Sauerkraut-70B, showing only a 1.3-point difference. Finally, we compare decoders using ICL and fine-tuned encoders. They excel in different types of tasks (see Table 7 in the Appendix). For example, ICL outperforms fine-tuning when datasets embody apparent difficulties, like imbalanced labels in Detox Hate-Speech. Overall, our results generalize previous findings from English to German: **specialized encoders outperform decoders (Mosbach et al., 2023), and ICL and fine-tuning are supplementary learning paradigms, as demonstrated in Waldis et al. (2023)**.

ii) Gender-fair language substantially impacts the performance. Figure 3 (*left*) focuses on the task-level influence of gender-fair language and shows the average difference between the original performance and the six reformulation strategies. Surprisingly, reformulations tend to improve measurable performance, especially with inclusive strategies, showing 17 improvements out of 20 cases. In contrast, neutralization (Neutral and De-e) tends to harm performance on average while only improving the performance in 5 out of 10 cases. Further, GenderStern provides the most improvement, while De-e exhibits the largest performance degradation. Interestingly, while GenderStern, GenderDoppelpunkt, and GenderGap minimally differ from each other (more details in Appendix § A.5), their performance considerably varies. This observation suggests that specific special characters (*, :, and _) semantically differ and shows, again, that LMs rely on lexical features rather than on linguistic specialties of the German language. Comparing the Lou tasks, Detox ones are more impacted than others, and offensive tasks show more impact compared to Germeval-Engaging, Germeval-Factclaiming, and Stance. Specifically, reformulations of GermEval-Toxicity show a significant impact. Notably, LMs perform at a lower level on these sensitive tasks, hinting that task difficulty and the impact of gender-

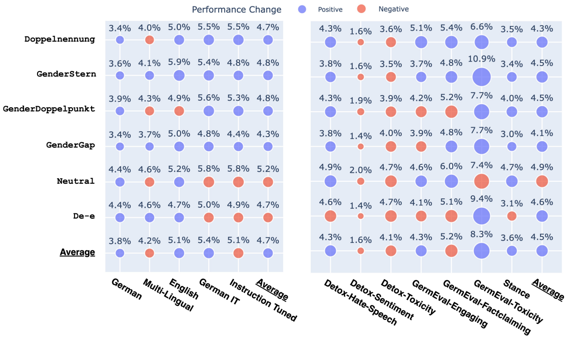


Figure 4: Label flip fractions for strategies, model types, and tasks. Size indicates the label flip fraction under gender-fair language and the color positive (blue) or negative (red) effect on aggregated performance.

fair language are connected. These insights show that **even minor changes have big effects, in particular for challenging tasks.**

iii) Aggregation across tasks may hide the impact of gender-fair language. We stack in Figure 3 (*right)* the differences of every LM and seed or prompting template (German IT and Instruction Tuned) separately. This detailed analysis shows that the impact of gender-fair language vanishes when aggregating across tasks. While the Neutral strategy showed a small impact (-0.1 points F_1 macro), the stacked analysis reveals substantial positive and negative effects. These insights highlight that **only a detailed analysis provides the full picture of the impact of gender-fair language.**

iv) Gender-fair language triggers significant label flips. We analyze the impact on individual predictions as the fraction of label flips under gender-fair language. Figure 4 shows reformulations flipping labels on average in 4.6%. Analyzing the model types (*left*) shows less variability but fewer flips for encoders than decoders, in particular for German specialized ones (German < Multi-Lingual < English). In contrast, the flip fraction is more spread across tasks (*right*). While detox-sentiment shows the smallest flip fraction, germinal-toxic exhibits the largest one up to 10.9% in combination with GenderDoppelpunkt. Relating to previously discussed results, GenderGap shows, again, a different pattern (less flips) than GenderStern and GenderDoppelpunkt for German, multi-lingual, and English LMs. This consistent finding demonstrates that even **minimal syntactic variations of gender-fair language significantly impact single predictions.** Comparing with the performance differences in Figure 3 (*left*) reveals that the label flip fractions provide a different perspective on the impact of gender-fair language. These two measures are moderately correlated ($\rho = 0.47, p < 0.05$) and show substantially different relations to the absolute performance in Figure 5. However, both measures tend to be less pronounced when LMs perform on a lower level, hinting again at a connection between task difficulty and the impact of gender-fair language.

v) The consistency of evaluations under gender-fair language. We compare the model rankings when evaluating the original or reformulated instances. We find significant ($p < 0.05$) high correlations ($\rho \geq 0.95$), meaning that LM rankings are consistent among original and reformulated instances. As a result, **existing datasets retain their validity for evaluations focusing on the supremacy of specific LMs.**

6 Analysis

Focusing on **RQ3**, we discuss the pronounced effect of reformulations on lower model layers (i) and find reformulations significantly alter attention patterns and decrease prediction certainty (ii), and these properties are crucial for label flips (iii).

i) Gender-fair language affects lower LM layers. We analyze how LMs process gender-fair language internally by computing layer-wise average embeddings (Reimers and Gurevych, 2019) of the original and reformulated text (s and s'). Afterward, we isolate the reformulation within these embeddings as $r = s - s'$. Then, we test how well we can distinguish the different strategies with r using KMeans (Lloyd, 1982) clustering for every layer separately. Across all LMs, we find statistically significant (p<0.05) negative correlations between the layer numbers and the cluster performance, rand index (ρ=-0.40), mutual information (ρ=-0.56), completeness (ρ=-0.55), and homogeneity (ρ=-0.54). As lower layers account for syntactic information and their degree of contextualization is lower (Tenney et al., 2019), **gender-fair language has a syntactic impact.**

Next, we qualitatively analyze and show in Figure 6 that the six strategies are better distinguishable on lower layers by projecting r for all layers to 1D using T-SNE. While these plots focus on GBERT-base only, we observe similar patterns for other LMs (Appendix § A.10).Doppelnennung, Neutral, and De-e are more different, while

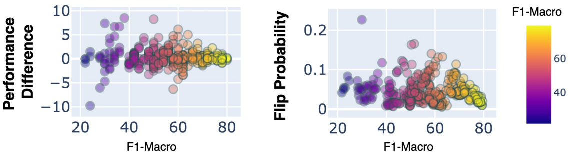


Figure 5: Performance difference and flip fraction against LMs' F_1 macros score of each task and strategy.

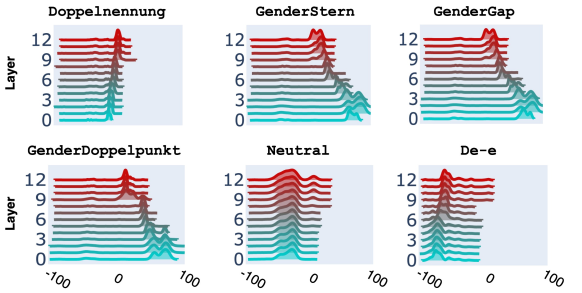


Figure 6: Kernel-density plot of the 1D projected reformulation embeddings r using t-SNE for all six strategies and 13 layers (x-axis) of GBERT-base, including the embedding layer (0).

	Overall	Flip		Correct	
		No	Yes	No	Yes
Norm. Flesch	-2.01	-2.00	-2.3	-2.08	-1.44
Norm. Length	+1.73	+1.74	+1.69	+1.76	+1.42
Prediction Certainty	-0.14	+0.35	-10.0	+0.49	-2.03
Attention Max	+0.41	+0.37	+1.14	+0.33	+0.56
Attention Variation	+0.10	+0.08	+0.23	+0.07	+0.13

Table 3: Change (stat. sig. at $p < 0.05$) between original and reformulated properties, overall, when instances flip or not, or are correct or not.

strategies using gender character (GenderStern, GenderGap, and GenderDoppelpunkt) overlap. Noteworthy, the specific special characters are again crucial, GenderDoppelpunkt (:) differs from GenderStern (*) and GenderGap (_). These insights confirm again that the impact of gender-fair language is primarily syntactic.

ii) Reformulations change instances and how LMs process them. Next, we examine the impact of reformulation on input instances and how language models (LMs) process them. We focus on surface properties such as normalized instance length and normalized Flesch score (readability, Flesch (1948)), prediction certainty (for encoders only), and LMs' attention patterns. We normalize both instance length and Flesch score between zero and one for each task independently. Attention patterns are characterized by the maximum attention and its variation (standard deviation) across input tokens. Specifically, we analyze how the prediction token attends input tokens: either the classification token (*[CLS]* or) for encoders or the first *next-token* for decoders. To ensure comparability, we exclude tokens affected by the reformulation to ensure consistent attention vector lengths of original and reformulated instances.

Table 3 shows that reformulated instances are longer and less readable (lower Flesch). These differences are less pronounced for correct predictions, shorter, and more readable than others. These surface-level changes are known to impact LMs (Ovalle et al., 2023). Next, LMs show less certainty for reformulated instances, mainly when they cause a label flip (-10.0) or are correct (-2.03). Consequently, LMs are even less sure when reformulations flip to the correct label (-11.1) and tend to increase attention variation and maximal attention. This effect is, again, most pronounced for reformulations causing a flip and/or are correct. These insights show that **reformulations alter attention patterns and potentially reduce the impact of spurious correlations**, a known drawback for tasks like in Hate-Speeech (Attanasio et al., 2022) or stance detection (Thorn Jakobsen et al., 2021; Beck et al., 2023).

iii) The surface properties of instances cause flips. Table 4 shows that the predicted labels of reformulated instances flip when the original ones are shorter, less readable (lower Flesch), and when LMs show lower prediction certainty. From higher attention maximum (5.1 vs. 1.4) and variation (1.4 and 0.8) of flipped instances, LMs give higher attention to single tokens, potentially causing a drop in certainty. These observations align with our previous results and analyses, which show that the influence of gender-fair language is stronger when the task is difficult, and LMs tend to be less sure. Specifically, we found an average certainty of 92.4 (no flip) and 89.2 (flip) for GermEval-Toxicity, with a particularly strong impact of gender-fair language. Figure 7 shows the relation between the different properties and the label flip fraction in more detail. While the Flesch score shows less pronounced effects, the flip fraction tends to be higher (up to 6%) for shorter instances. Further, the flip fraction is crucially higher when an LM predicts with less certainty (up to 15%) and exhibits a high attention maximum or variation, up to 15% and 20%).

	Flip==No	Flip==Yes
Norm. Flesch	71.1±11.4	70.7±11.7
Norm. Length	27.6±16.9	26.8±16.8
Prediction Certainty	95.7±9.5	90.6±14.3
Attention Max	10.3±7.9	12.0±9.6
Attention Variation	1.8±1.6	2.1±1.9

Table 4: Properties of instances when their reformulation causes a label flip or not, including surface properties (Flesch and length), prediction certainty, and attention patterns. All differences are stat. sig. ($p < 0.05$).

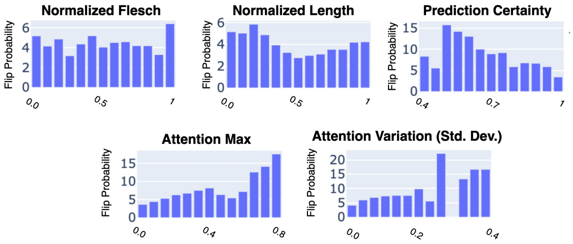


Figure 7: Distribution of instance properties and label flip fractions are statistically significant ($p < 0.05$).

7 Related Work

Previous work shows LMs embodying substantially stereotypical bias (Kurita et al., 2019; Nadeem et al., 2021; Srivastava et al., 2022) regarding gender, profession, race, and religion. In particular, gender bias gained more attention recently (Sun et al., 2019; Hardmeier et al., 2022). While some works focus on analyzing the existence of gender bias (Zhao et al., 2019, 2020; Kaneko et al., 2022), others aim to reduce this bias (Qian et al., 2019; Ravfogel et al., 2020; Ranaldi et al., 2024). Another line of work examines the effects of gender-fair language, such as how LMs process (neo)pronouns (Brandl et al., 2022; Hossain et al., 2023; Gautam et al., 2024). Gender-fair language is also broadly studied in translation (Vanmassenhove et al., 2023), focusing on the effects of gender bias (Stanovsky et al., 2019). This includes analyzing the acceptance of gender-fair formulations (Attanasio et al., 2023), gender neutralization (Piergentili et al., 2023), the use of interpretability methods (Attanasio et al., 2023), and the impact of pronouns on translation (Lauscher et al., 2023; Amrhein et al., 2023). Unlike previous work, we focus on the impact of gender-fair language on classification inference. Specifically, we present with Lou the first dataset of its kind, to assess the impact of gender-fair language on text classification regarding seven German tasks and analyze this impact in detail.

8 Conclusion

We comprehensively assess the impact of gender-fair language on German text classification tasks. Specifically, we introduce Lou, a high-quality dataset of parallel annotated reformulations that employ various gender-fair strategies. Our systematic evaluation and analysis reveal that aggregated evaluations of original data maintain their validity under gender-fair language. However, absolute performance tends to increase, while predicted labels can flip with a probability of up to 10.9%, particularly due to significantly reduced prediction certainty and altered attention patterns. These findings highlight the importance of considering this linguistic variation, especially since even minor syntactic changes can critically alter how LMs process individual instances. Moving forward, we plan to extend Lou to other languages that employ similar gender-fair reformulation strategies, such as Italian and French, and work on adopting LMs for this linguistic variation.

Limitations

The Focus on German This work solely focuses on gender-fair language in German. However, we assume our evaluation and analytical pipeline is adaptable to other languages. Furthermore, we see empirical insights that the impact is mostly due to syntactic variations of gender-fair language in LMs that can be transferred to another language. This is especially plausible since these patterns are consistent across German, multi-lingual, and English LMs.

Selected Reformulation Strategies We select a set of six reformulation strategies to reflect the diversity of options. However, we acknowledge the incompleteness as other strategies exist. For example, the use of neo-pronouns or the addressing the feminine and masculine gender using the slash character, for example, *Schüler/in*.

Dataset Selection The selected German datasets reflect a subset of the available ones. With them, we aim to cover diverse tasks while optimizing reformulation efforts. For example, Detox and GermEval-2021 provide multiple annotations. However, we do not claim completeness.

Licensing For Lou, we adopt the licensing of the underlying datasets and make reformulated instances for X-Stance and GermEval-2021 freely available. For Detox, please contact the corresponding author or request the data via the online archive along with a confirmation of the original dataset access.[12]

Ethical Considerations

With Lou, we cover a broad selection of German text classification tasks. This collection includes some datasets with offensive content, like text instances from the GermEval-2021 or the Detox datasets. Addressing this issue during reformulation, we collected the explicit consent and willingness to annotate this type of text. This includes informing them that potential triggers could arise and that they can stop or skip reformulation without giving reasons when they feel uncomfortable.

Acknowledgements

We thank Vagrant Gautam, Qian Ruan, Tianyu Yang, Nils Dycke, and the anonymous reviews for the helpful feedback, and Mara Haas and Chantal Amrhein for the valuable input on gender-fair writing strategies. Andreas Waldis has been funded by the Hasler Foundation Grant No. 21024. The work of Anne Lauscher is funded under the Excellence Strategy of the German Federal Government and States.

References

AI@Meta. 2024. Llama 3 model card.

Chantal Amrhein, Florian Schottmann, Rico Sennrich, and Samuel Läubli. 2023. Exploiting biased models to de-bias text: A gender-fair rewriting model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4486–4506, Toronto, Canada. Association for Computational Linguistics.

Giuseppe Attanasio, Debora Nozza, Dirk Hovy, and Elena Baralis. 2022. Entropy-based attention regularization frees unintended bias mitigation from lists. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1105–1119, Dublin, Ireland. Association for Computational Linguistics.

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

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2024. Sensitivity, performance, robustness: Deconstructing the effect of sociodemographic prompting. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2589–2615, St. Julian's, Malta. Association for Computational Linguistics.

Tilman Beck, Andreas Waldis, and Iryna Gurevych. 2023. Robust integration of contextual information for cross-target stance detection. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 494–511, Toronto, Canada. 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.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Christoph Demus, Jonas Pitz, Mina Schütz, Nadine Probol, Melanie Siegel, and Dirk Labudde. 2022. Detox: A comprehensive dataset for German offensive language and conversation analysis. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 143–153, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Rudolf Franz Flesch. 1948. A new readability yardstick. *The Journal of applied psychology*, 32(3):221–233.

[12] Information are available online under https://github.com/hdaSprachtechnologie/detox

Vagrant Gautam, Eileen Bingert, Dawei Zhu, Anne Lauscher, and Dietrich Klakow. 2024. Robust pronoun fidelity with english llms: Are they reasoning, repeating, or just biased? *ArXiv preprint*, abs/2404.03134.

Christian Hardmeier, Christine Basta, Marta R. Costa-jussà, Gabriel Stanovsky, and Hila Gonen, editors. 2022. *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*. Association for Computational Linguistics, Seattle, Washington.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

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.

Masahiro Kaneko, Aizhan Imankulova, Danushka Bollegala, and Naoaki Okazaki. 2022. Gender bias in masked language models for multiple languages. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2740–2750, Seattle, United States. Association for Computational Linguistics.

Daphna Keidar, Andreas Opedal, Zhijing Jin, and Mrinmaya Sachan. 2022. Slangvolution: A causal analysis of semantic change and frequency dynamics in slang. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1422–1442, Dublin, Ireland. Association for Computational Linguistics.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Manuel Lardelli and Dagmar Gromann. 2023. Translating non-binary coming-out reports: Gender-fair language strategies and use in news articles. *The Journal of Specialised Translation*, (40):213–240.

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

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. *ArXiv preprint*, abs/1907.11692.

Stuart P. Lloyd. 1982. Least squares quantization in PCM. *IEEE Trans. Inf. Theory*, 28(2):129–136.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of what art? A call for multi-prompt LLM evaluation. *ArXiv preprint*, abs/2401.00595.

Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, Dietrich Klakow, and Yanai Elazar. 2023. Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12284–12314, Toronto, Canada. Association for Computational Linguistics.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Anaelia Ovalle, Ninareh Mehrabi, Palash Goyal, Jwala Dhamala, Kai-Wei Chang, Richard S. Zemel, Aram Galstyan, and Rahul Gupta. 2023. Are you talking to ['xem'] or ['x', 'em']? on tokenization and addressing misgendering in llms with pronoun tokenization parity. *ArXiv preprint*, abs/2312.11779.

Angela Balducci Paolucci, Manuel Lardelli, and Dagmar Gromann. 2023. Gender-fair language in translation: A case study. In *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 13–23, Tampere, Finland. European Association for Machine Translation.

Andrea Piergentili, Dennis Fucci, Beatrice Savoldi, Luisa Bentivogli, and Matteo Negri. 2023. Gender neutralization for an inclusive machine translation: from theoretical foundations to open challenges. In *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 71–83, Tampere, Finland. European Association for Machine Translation.

Yusu Qian, Urwa Muaz, Ben Zhang, and Jae Won Hyun. 2019. Reducing gender bias in word-level language models with a gender-equalizing loss function. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 223–228, Florence, Italy. Association for Computational Linguistics.

Leonardo Ranaldi, Elena Ruzzetti, Davide Venditti, Dario Onorati, and Fabio Zanzotto. 2024. A trip towards fairness: Bias and de-biasing in large language models. In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 372–384, Mexico City, Mexico. Association for Computational Linguistics.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Julian Risch, Anke Stoll, Lena Wilms, and Michael Wiegand. 2021. Overview of the GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments. In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, pages 1–12, Duesseldorf, Germany. Association for Computational Linguistics.

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

Sabine Sczesny, Magda Formanowicz, and Franziska Moser. 2016. Can gender-fair language reduce gender stereotyping and discrimination? *Frontiers in Psychology*, 7.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew K. Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakas, and et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *ArXiv preprint*, abs/2206.04615.

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.

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*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Zhewei Sun and Yang Xu. 2022. Tracing semantic variation in slang. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1299–1313, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Terne Sasha Thorn Jakobsen, Maria Barrett, and Anders Søgaard. 2021. Spurious correlations in cross-topic argument mining. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 263–277, Online. Association for Computational Linguistics.

Jannis Vamvas and Rico Sennrich. 2020. X-Stance: A multilingual multi-target dataset for stance detection. In *Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS)*, Zurich, Switzerland.

Eva Vanmassenhove, Beatrice Savoldi, Luisa Bentivogli, Joke Daems, and Janiça Hackenbuchner, editors. 2023. *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*. European Association for Machine Translation, Tampere, Finland.

Andreas Waldis, Yufang Hou, and Iryna Gurevych. 2023. How to handle different types of out-of-distribution scenarios in computational argumentation? a comprehensive and fine-grained field study. *ArXiv preprint*, abs/2309.08316.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 649–657.

Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. 2020. Gender bias in multilingual embeddings and cross-lingual transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2896–2907, Online. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

A Appendix

A.1 The Use of AI Assistants

We use ChatGPT to rework this paper regarding grammatical correctness and clarity.

A.2 Additional Information about the Reformulation Study

Annotators During our iterative annotation study, we distinguish between eight amateurs and two professional (**P1** and **P2**) annotators. The amateur annotators do not have a linguistic background but are native German-speaking, They determine their experience in applying gender-fair language from 1 (no experience) to 5 (professional experience) with low (2) to advanced (4) and an average moderate (3). The professional proofreaders have both a linguistic background as they studied the German language (*Germanistik*, *German studies*) and work in proofreading (**P1**) or in text agency (**P2**).

Payment The amateur annotators did not receive a payment as they conducted the annotations voluntarily. In contrast, we pay the principal professional annotator (**P1**) an hourly rate of 56$ and the second one (**P2**) 167$.

A.3 Error Analysis of Amateur Annotators

As we found substantial difficulties for amateur annotators in applying gender-fair language with sufficient quality, we analyzed these errors in more detail. Specifically, the principal professional annotator (**P1**) categorised the errors regarding seven categories:

1. Personfication When it is clear that a gender-specific phrase corresponds to a person with a specific gender, gender-fair language is not applicable. For example, Präsident (English *president.MASC.SG*) when it is clear that the text refers to Donald Trump.

2. Neutral Substantive When gender-fair reformulation is unnecessary because the substantive is neutral, like *Gäste* (English *customers.NEUT.PL*).

3. Numerus Inconsistency in singular and plural in the reformulation. For example, the phrase *die Künstlerin oder den Künstler* should be in plural *die Künstlerinnen oder Künstler* (*the artist.FEM.SG or the artist.MASC.PL*).

4. And/Or The use of *oder* (English *or*) instead of *und* (English *and*) in `Doppelnennung`, as *und* is more inclusive an appropriate at this point.

5. Pronoun If pronouns were not changed accordingly. For example, *keiner* (English *nobody*) needs to be reformulated into *keine**r* for the strategy `GenderStern`.

6. Compounds Errors in compounded words like *Zuschauerreaktionen* (English *audience reactions*). This should be reformulated as *Zuschauer**innenreaktionen* considering the `GenderStern` strategy.

7. Word root Errors in the word's root form. For example, when considering `GenderStern` *Experte**in* (*expert.NEUT.SG*) is not correct, it has to be *Expert**in*.

8. Other A collection of other errors. For example, overlooked reformulations, less common neutral formulations like *Deutsche Personen aus Regierungskreisen* instead of *Deutsche Regierende*

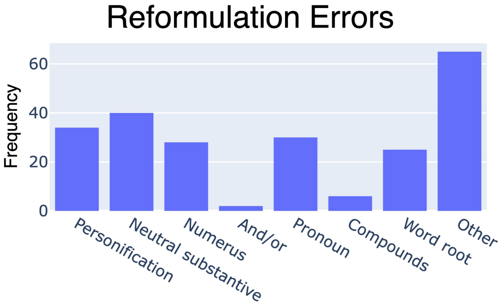


Figure 8: Overview of the categorization frequency when analyzing the errors of the amateur annotators.

(*german person from the government*), or other grammatical errors.

We show in Figure 8 and Figure 9 the frequency of these categories aggregated and per dataset and strategy. Despite *And/or* and *Compounds*, all categories have a similar frequency. *Other* is the most frequent category, summarizing many errors. However, it is particularly frequent for `Neutral`, where amateur often used over-complicated and non-usual neutral formulations. Regarding the dataset, amateur struggled more with the longer text from GermEval-2021, which often included grammatical errors. Concerning the different strategies, `GenderStern` seems to cause the most errors. Particularly prominent are *Personification*, *Neutral Substantive*, *Numerus*, *Pronoun*, and *Word Root*. These errors show that people struggle to consistently apply gender-fair language, highlighting the need for standardization for broad establishment. As their frequency heavily depends on the type and complexity of the text, our insights suggest enrolling in more sophisticated tutoring when solely relying on amateur annotations. While being more costly, professional annotators show a clear advantage in providing high-quality reformulations.

A.4 Additional Information about the Data

Table 5 show additional information about the considered datasets in `Lou`. This includes the average number of tokens (length) and readability (Flesch score), the number of samples per dataset, label distribution, and how train and test label distribution agree using KL divergence.

A.5 The Effect of Tokenization for Gender-Fair Language

Table 6 shows how many additional tokens the different reformulation strategies add to the input sentence regarding the various strategies and LMs. From these results, `Doppelnennung` adds the most tokens (6.6 on average), `De-e` the least with on average 1.6 more tokens, and the other between 3.4 (`GenderStern`) to 3.1 tokens (`Neutral`). Noteworthy, we see apparent differences between `GenderStern` & `GenderDoppelpunkt` and `GenderGap` for decoder LMs, hinting at the different semantic meanings of these special characters. Regarding the LM difference, we note that within and across the model type, LMs with a more extensive vocabulary size tend to add fewer tokens than those with a smaller number of distinct tokens. Comparing the results, we do not find a clear correlation between the additional number of tokens and the impact of gender-fair language on an aggregated level or for individual predictions.

A.6 Additional Details In-Context Learning

Similar to using random seeds when fine-tuning LMs, we use four different prompts to measure the LMs' task capabilities thoroughly. Following, we provide examples of these templates for the fact-claiming task. For the first template, we follow the previous when task prompts are available and translate them into German, such as hate-speech, toxicity, or stance detection in Beck et al. (2024). Composing templates two and three, we rephrase the task instructions. For the fourth template, we restructure the prompt and embed the example within the task instructions.

Prompt Template 1 Geben ist der folgenden Satz, wird in diesem Tatsachen behauptet oder nicht? Mögliche Antworten sind 'ja', falls im Satz Tatsachen behauptet werden oder 'nein' falls nicht. Antworte nur mit einem dieser Möglichkeiten und ohne Erklärung!
Text: Die Möglichkeit, dass Trump gewinnt ist groß, weil seine Gegner*innen so schwach ist. Tatsachen erwähnt: **nein**

Prompt Template 2 Die Aufgabe ist es zu erkennen ob im folgenden Satz Tatsachen behauptet werden oder nicht. Mögliche Antworten sind 'ja', falls im Satz Tatsachen behauptet werden oder 'nein' falls nicht. Antworte nur mit einem dieser Möglichkeiten und ohne Erklärung!
Text: Die Möglichkeit, dass Trump gewinnt ist groß, weil ens Gegnere so schwach ist. Tatsachen erwähnt: **nein**

Prompt Template 3 Betrachten wir den folgenden Satz. Wird in diesem Tatsachen behauptet?

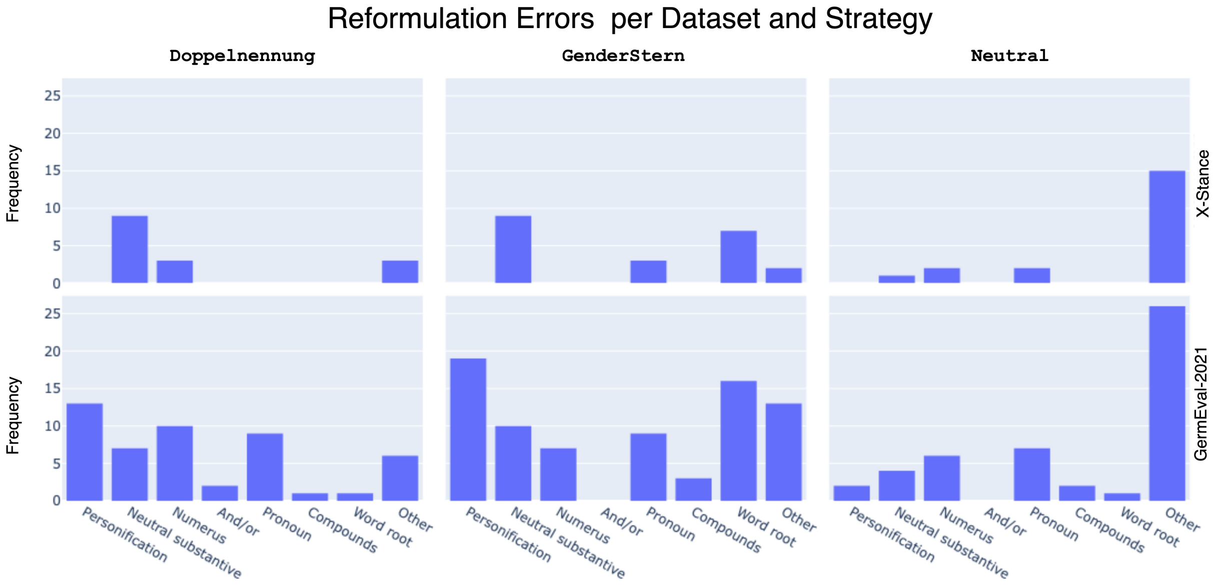


Figure 9: Detailed overview of the categorization frequency when analyzing the errors of the amateur annotators per dataset (row) and strategy (rows).

		GermEval-2021			Detox		
	Stance	Engaging	Fact-Claiming	Toxicity	Hate-Speech	Sentiment	Toxicity
Length	24.5±15.6		30.6±41.6		28.4±12.7	27.8±12.8	28.6±12.7
Flesch	46.3±9.6		59.9±103.9		50.2±21.9	49.9±22.7	50.3±21.9
# Samples	36,921		2,276		4,100	8,348	4,035
Labels	favor: 18,227 against: 18,694	yes: 635 no: 1,641	yes: 809 no: 1,467	yes: 801 no: 1,475	yes: 928 no: 3,172	negative: 6,253 neutral: 1,737 positive: 358	yes: 808 no: 3,227
# KL Div.	0.01	0.25	0.03	0.08	0.06	0.04	0.04

Table 5: Further insights about the seven Lou tasks, including surface properties, size, and label distribution.

Antworte mit einem Wort: 'ja' falls der Satz Tatsachen behauptet oder 'nein' falls nicht.
Text: Die Möglichkeit, dass Trump gewinnt ist groß, weil seine Konkurreenz so schwach ist. Tatsachen erwähnt: **nein**

Prompt Template 4 Werden im Satz "Die Möglichkeit, dass Trump gewinnt ist groß, weil seine Gegner so schwach ist." tatsachen behauptet oder nicht? Antworte mit "ja", falls der Satz toxisch ist oder "nein" falls nicht.
nein

A.7 Used Language Models

We run all of our experiments using Nvidia RTX A6000 GPUs. Every GPU provides 48GB of memory and 10752 CUDA Cores. We use the following models from the huggingface model hub:

- deepset/gbert-base
- deepset/gbert-large
- deepset/gelectra-base
- deepset/gelectra-large
- bert-base-multilingual-cased
- FacebookAI/xlm-roberta-base
- microsoft/mdeberta-v3-base
- bert-base-uncased
- roberta-base
- microsoft/deberta-v3-base
- TechxGenus/Meta-Llama-3-70B-Instruct-AWQ
- TechxGenus/Meta-Llama-3-8B-Instruct-AWQ

	Vocab. Size	Doppelnennung	GenderStern	GenderDoppelpunkt	GenderGap	Neutral	De-e	Average
GBERT	31k	5.3	3.5	3.5	3.5	2.7	1.8	3.4
GELECTRA	31k	5.3	3.5	3.5	3.5	2.7	1.8	3.4
mBERT	120k	6.2	3.5	3.5	3.5	2.9	1.7	3.6
XLM-R	250k	6.0	3.5	3.5	3.5	2.6	1.5	3.4
mDeBERTa-v3-base	250k	6.0	3.4	3.4	3.4	2.4	1.1	3.3
BERT	31k	8.3	4.4	4.4	4.4	4.2	1.5	4.5
RoBERTa	50k	9.9	4.4	4.4	4.4	4.6	1.7	4.9
DeBERTa-v3-base	128k	6.9	3.4	3.4	3.4	3.2	1.2	3.6
Sauerkraut	128k	7.1	3.2	3.2	2.7	3.2	1.6	3.5
Llama-3	128k	7.1	3.2	3.2	2.7	3.2	1.6	3.5
gpt-3.5-turbo	100k	7.2	3.2	3.2	2.7	3.2	1.6	3.5
gpt-4o	250k	6.1	2.3	2.2	2.8	2.7	1.5	2.9
Average		6.6	3.4	3.4	3.3	3.1	1.6	3.6

Table 6: Number of additional tokens when comparing the reformulated examples with the original ones. Average across all tasks regarding models and strategies.

- mayflowergmbh/Llama-3-SauerkrautLM-8b-Instruct-AWQ
- tresiwalde/Llama-3-SauerkrautLM-70b-Instruct-AWQ

A.8 Detailed Results

Table 7, Figure 11, Figure 12, Figure 13, Figure 14, Figure 15, Figure 16, and Figure 17 shows the detailed baseline results covering all the seven Lou tasks for the 16 considered results. Note that we evaluated the original examples without reformulations.

A.9 Label Verification

We list in Table 8 manually check examples. We found that gender-fair language does not invalidate any annotated task label.

A.10 Detailed Reformulation Distribution

Figure 10 shows the distribution of the reformulation representation r for every model, reformulation strategy and model layer. Similar patterns, as previously discussed, can be observed: strategies are more distinguishable for lower layers, and noteworthy differences between GenderDoppelpunkt and GenderStern and GenderGap. Further, LMs with a higher performance level (like German LMs) tend to show more variation among the layers, hinting at their better semantic understanding of the German language.

	Detox			**GermEval-2021**				
	Hate-Speech	Sentiment	Toxicity	Engaging	Fact-Claiming	Toxicity	Stance	Average
GBERT-base	51.7±4.1	59.6±3.7	51.1±2.5	57.8±1.8	70.3±1.1	54.0±9.2	76.0±2.1	60.1
GBERT-large	47.6±4.6	**63.0**±4.8	**54.3**±3.1	**62.1**±2.3	69.9±2.0	54.6±3.4	**78.5**±1.7	**61.4**
GELECTRA-base	51.4±3.4	**64.7**±4.5	**52.0**±3.1	59.4±1.8	**70.9**±1.9	53.7±7.1	77.1±1.9	**61.3**
GELECTRA-large	54.1±3.6	**62.0**±5.1	47.3±17.3	**61.6**±2.1	68.6±1.8	53.0±12.1	**78.2**±1.4	60.7
mBERT-base	40.4±5.6	41.7±1.7	50.7±4.3	59.2±2.1	70.4±2.2	47.0±7.4	72.0±2.4	54.5
XLM-R-base	46.0±3.0	45.6±9.0	46.8±17.0	59.5±1.8	70.4±1.7	34.2±29.6	74.1±1.4	53.8
mDeBERTa-base	53.1±4.5	63.1±3.1	50.5±2.6	**60.3**±2.1	**72.6**±1.8	58.0±4.9	77.8±1.6	**62.2**
BERT-base	28.0±6.9	41.2±1.7	34.5±14.1	57.7±3.1	66.8±2.2	22.5±19.8	64.2± 3.2	45.0
RoBERTa-base	41.6±5.9	43.0±1.2	31.9±19.2	57.7±1.5	67.8±2.8	29.7±5.3	66.6±1.7	48.3
DeBERTa-base	52.3±3.5	50.1±7.7	50.8±2.9	59.3±2.2	**71.7**±2.5	42.0±18.2	**79.3**±2.6	57.9
Sauerkraut-8B	**54.7**±2.8	44.1±11.4	49.9±10.0	56.2±3.3	52.7±7.0	58.7±4.5	55.3±1.8	53.1
Sauerkraut-70B	**56.1**±1.4	43.2±10.6	46.6±7.4	50.7±3.4	59.3±13.2	**67.9**±3.2	56.9±7.0	54.4
Llama-3-8B	42.0±5.8	35.1±23.5	27.9±5.9	46.0±7.0	59.7±4.4	51.0±8.3	52.5±11.7	44.9
Llama-3-70B	57.2±3.9	38.8±13.3	41.9±6.9	50.7±2.6	62.9±12.9	**68.7**±2.8	57.4±5.6	53.9
gpt-3.5-turbo	55.3±3.9	56.6±3.2	26.7±6.7	50.3±1.2	44.4±3.9	57.0±3.3	52.8±3.0	49.0
gpt-4o	**64.7**±6.2	52.5±4.4	**53.0**±8.7	43.8±2.7	66.5±14.4	**67.9**±0.8	62.1±6.5	58.7
Average	48.2	51.8	45.8	57.6	67.5	48.2	70.9	55.7

Table 7: Detailed performance on the seven Lou tasks for all the analyzed LMs on the original examples, without reformulations.

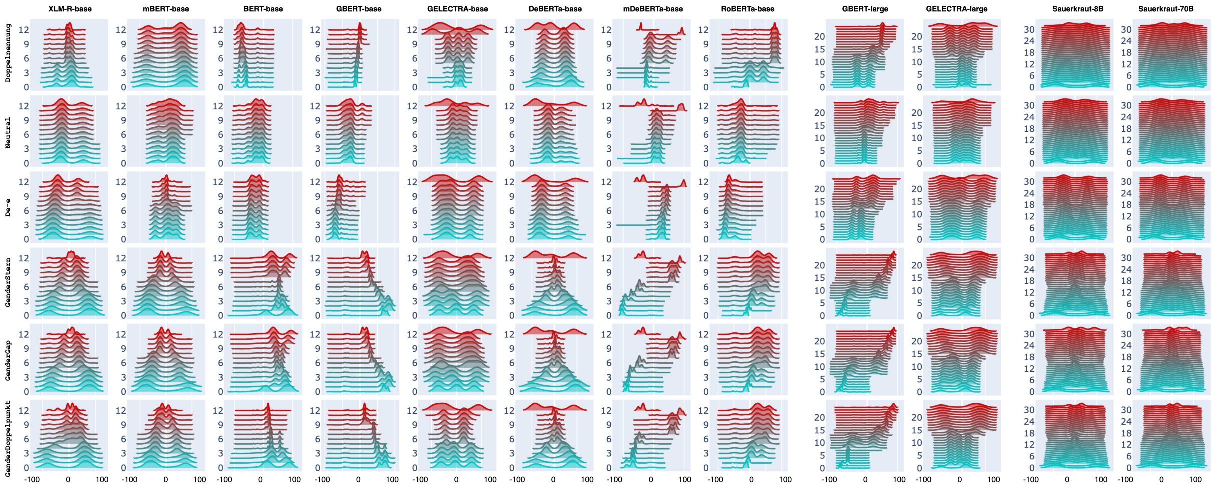


Figure 10: Projection (1D using T-SNE) of the vector difference between the average embeddings of the reformulated examples and the original ones for all six strategies and 13 layers (x-axis) of GBERT-base, including the embedding layer (0).

Task	Text	Topic	Label
Stance	Staatlicher Zwang ist falsch. Das ist Sache zwischen Arbeitgeber*in und Arbeitnehmer*in	Welfare	`against`
Stance	Freie Wirtschaft für freie Bürger*innen. Weltweiter Freihandel ohne Schranken ist erstrebenswert.	Foreign Policy	`favor`
Stance	Nicht unter einem*r Präsident*in, welcher die Rechte anderer mit Füssen tritt und Respektlos gegenüber ändern ist.	Foreign Policy	`against`
Stance	Jede anbietende Person soll an seinem Standort selber entscheiden können wie lange geöffnet sein soll	Economy	`favor`
Stance	Das wäre kontraproduktiv. Das Problem, dass ältere Arbeitnehmer*innen keine Stelle mehr finden, würde dadurch verschärft.	Economy	`against`
Stance	Es konnte kein Rückgang bei kriminellen Straftaten festgestellt werden. `favor` geraten bislang unbescholtene Bürgerne zunehmend unter Generalerneverdacht. Dieser Entwicklung ist Einhalt zu gebieten.	Security	`against`
Stance	Es sollen Anreize geschaffen werden (z.B. via BVG-Beiträge) damit es für Arbeitgeberne attraktiv bleibt, ältere Angestellte im Betrieb zu behalten. Ein Kündigungsschutz setzt falsche Anreize.	Economy	`against`
Stance	Es sollen die gleichen Spielregeln für alle gelten – die Online Anbieter bewegen sich oft noch im Grau-Bereich. Die Angebote sollen aber nicht durch Regulierungen verunmöglicht werden.	Digitisation	`favor`
Stance	Die Schweiz bietet den internationalen Unternehmen anderweitig genug gewichtige Vorteile (politische Stabilität, Einstellung der Mitarbeiterinnen und Mitarbeiter, Infrastruktur...)	Finances	`against`
Stance	Das Rentenalter soll flexibel sein, so kann jeder eigenverantwortlich bestimmen, Arbeitskräfte wie Unternehmen	Welfare	`favor`
Stance	Wichtig ist , dass alle am Markt teilnehmenden Personen gleich lange Spiesse haben.	Digitisation	`favor`
Fact-Claiming	@USER stimmt. Die Russ*innen hatten wenigstens diese Ossis unter Kontrolle		`no fact claimed`
Fact-Claiming	Tja, nur weil das bei uns so gehandhabt wird wenn die Wahl zum Staatsoberhaupt nicht passt, heißt das noch lange nicht das rs überall auf der Welt so läuft		`fact claimed`
Fact-Claiming	Das hoffen die allermeisten meiner amerikanischen Lieblingsmenschen allerdings nicht. Vote him out.		`no fact claimed`
Fact-Claiming	Republicans Overseas haben sich echt nicht positiv hervorgetan die letzten Wochen, das ist an Peinlichkeit kaum zu überbieten. Ankündigungen und Lügen. Super Staatsoberhaupt habt ihr da.		`no fact claimed`
Fact-Claiming	@USER, bin ich Politikere und verdiene jede Menge? Nein.		`no fact claimed`
Fact-Claiming	@USER und eigentlich, ja eigentlich hätte sie gewonnen, wenn nicht das amerikanische Wahlsystem keine eigentliche Gewinnerin oder eigentlichen Gewinner kennen würde!		`fact claimed`
Fact-Claiming	Warum ist danach Schluss? F. D. Roosevelt war auch 3 Amtszeiten Staatsoberhaupt. Das Gesetz wird der Orange wohl auch noch einmal ändern		`fact claimed`
Fact-Claiming	Er ist der aller schlimmste Präsident den Amerika je hatte		`no fact claimed`
Fact-Claiming	Hackt nicht nimmer auf den Fussball rum. Bei allen Sportarten ist wieder Publikum erlaubt. Hygienekonzept vorausgesetzt.		`fact claimed`
Fact-Claiming	Impfkritische Personen haha,wir haben einfach nur angst um unsere Gesundheit. Ich will nicht durch ne Impfung krank werden in 2 oder 3 Jahren.		`no fact claimed`
Engaging	Genau wegen solchen Eingeladenen in dieser Sendung schaue ich mir @MEDIUM nicht mehr an, eine Zumutung ist das.		`engaging`
Engaging	George Weinberg war so ignorant und abstoßend, dass er nen gutes Sinnbild für Trumps Lager abgegeben hat. Die Republikaner*innen sind seit Bush in einer stetigen menschlichen Abwärtsspirale.		`not engaging`
Engaging	Wie ist das mit der Zwangsabgabe- Rundfunkbeitrag? Zahlt die doch direkt an die Schauspielerne und Künstlerne aus! Dann sind sie sinnvoller eingesetzt!		`engaging`
Engaging	Regierende mit Verstand ?...und das in Zusammenhang mit dem Namen Trump ?		`not engaging`
Engaging	Was heisst dort lassen sollen? Der ist nur nach Deutschland gekommen weil es die Russ*innen so erlaubt haben. Deutschland hatte da garnicht's zu entscheiden.		`engaging`
Engaging	Fussballclubs bestehen nicht nur aus ein paar Kickerinnen oder Kickern und Trainerinnen oder Trainern. Was Sie fordern würde Tausende Angestellte treffen.		`not engaging`
Engaging	@USER Auffallend viele Republikanerne können mit enserm Präsidente nichts (mehr) anfangen.		`not engaging`
Hate-Speech	@ZDFheute Mit dem Bericht aus Israel, wo gegen Israel gehetzt wird, zeigt sich, wer hier antisemitisch ist. Ist jetzt Aktivistin ein neuer Beruf für schreiende, ungezogene, rechthaberische Mädchen?		`no hate-speech`
Hate-Speech	@TsiJulian @janboehm Promovierte Journalistinnen und Journalisten. Na und?		`no hate-speech`
Hate-Speech	@MichaelDhr1 @kristen_peter @Xrxtn @Alice_Weidel @UweJunge Steht ebend nicht drin, meine Göttin oder mein Gott. Die Binde muss nur zu erkennen sein.		`no hate-speech`
Hate-Speech	@ubskm_de @AfD_RBK Sehr gut. Und bitte noch die diversen Schlupflöcher weitgehend zukitten, damit kein linksgrün verstrahlte "Sachverständigere" de Tätere am Ende doch noch zum Opfer macht.		`hate-speech`
Hate-Speech	@AliCologne Die Demo-Teilnehmerne verhöhnen die Corona-Opfer ebenso wie #Querdenkerne und #Coronaschwurblerne. Es ist ein Schlag ins Gesicht der Pflegekräfte und Arzterne in den Kliniken.		`no hate-speech`
Hate-Speech	@Hoellenaufsicht @eduardt_s @Alice_Weidel @UweJunge Benennt diese AFD Leute als das was sie sind Faschisterne, nicht rechts radikal oder sonstiges rumgeiere. Diese Leute sind brandgefährlich Polizei		`hate-speech`
Hate-Speech	Ab 1. Juni hebt #Israel viele Vorteile für #greenpass Besitzerinnen und Besitzer auf und gestattet weitestgehende #Lockerungen auch für Nicht- #Geimpfte . #Corona https://t.co/wlZ5tmt3HJ		`no hate-speech`
Hate-Speech	@Svenfpunkt @LillyBlaudszun @janboehm 76 Wählerinnen und Wähler sind jetzt nicht die Welt, aber das Dorf sollte man bei Gelegenheit mal entseuchen.		`hate-speech`
Hate-Speech	@_FriedrichMerz Die vielen Rechtschreib- und Orthographie Fehler in den Anti-Merz-Tweets beweisen das geistige Kleinrentnerinnen- und Kleinrentnertum der linksgrünen Absenderinnen und Absender.		`hate-speech`
Toxicity	Ich hoffe ihre Gäste werden Recht behalten und Jo Biden wird Präsident. Noch mal 4 Jahre Trump und der bringt die ganze Welt durcheinander		`not toxic`
Toxicity	kein Wunder bei diesen Haufen lügenden und betrügenden Personen die sich Titeln kaufen und ein auf schlau machen !!!		`toxic`
Toxicity	@Anna48338034 @mista98berk Frag mal Arye Sharuz Shalicar, er ist hier auf Twitter und Sprecher der israelischen Armee, in Deutschland geboren, er kann das bestätigen		`not toxic`
Toxicity	Tja, nur weil das bei uns so gehandhabt wird wenn die Wahl zur Ministerpräsidentin oder Ministerpräsidenten nicht passt, heißt das noch lange nicht das rs überall auf der Welt so läuft		`toxic`
Toxicity	@USER Die nächste Eiszeit kommt. Derzeit leben wir in einer Zwischenwarmzeit. Durchaus seriöse forschende Personen halten ein schnelles Abreisen des Golfstroms für sehr wahrscheinlich.		`not toxic`
Toxicity	https://t.co/cs5KoavBp8 NRW Lusche ihr seid scheiße nein du bist es! Eine Bekanntschaft aller Schwulen Spahnferkels. thx		`toxic`
Toxicity	Sein Wort in Gött*in Gehörgang		`not toxic`
Toxicity	@willensbrecher Willensbrecher du bist ein kranker und ekelhafter Spinner oder eine kranke und ekelhafte Spinnerin verzieh dich aus Twitter oder ich melde dein Tweet den Behörden du Nazi		`toxic`
Toxicity	Da bin ich ja mal gespannt. Denn das, was ich von den Lernenden und jungen Leuten vor allem mitbekomme ist, dass sie sich wie die Versuchskaninchen für das Virus vorkommen.		`not toxic`
Toxicity	@MEDIUM Warum denn mit fiebern? "Die Amerikanerinnen und Amerikaner werden schon das Richtige machen!.... nachdem sie alles andere Ausprobiert haben."		`toxic`
Toxicity	Die USA würden sich viel Stress ersparen, wenn sie einfach das nächste Staatsoberhaupt vom @MEDIUM und @MEDIUM wählen lassen würden.		`toxic`
Toxicity	@Svenfpunkt @LillyBlaudszun @janboehm 76 Wählende sind jetzt nicht die Welt, aber das Dorf sollte man bei Gelegenheit mal entseuchen.		`toxic`
Toxicity	@TiloJung Als regierungssprechende Person muss man dumm sein. BÖSE und DUMM. #niewiederCDU #fckcdu		`toxic`
Sentiment	@BastardHegels @iknrr @ainyrockstar Alter der Syrer saß da ganz friedlich im Bus wie jeder andere auch und der Nazi Spast kommt an und attackiert ihn wtf ist falsch bei dir		`negative`
Sentiment	@sschyvonne @AndySpirig @Karl_Lauterbach Herr, lass Hirn über Frau*Herr Richter*in regnen!		`negative`
Sentiment	@MarkusWerner18 @c_lindner Der Georg Thile hat bekommen , was er bekommen soll, wer GEZ nicht bezahlt bricht das Gesetz, also eine Perso , die Verbrechen begeht.		`negative`
Sentiment	@Svenfpunkt @LillyBlaudszun @janboehm 76 Wählerne sind jetzt nicht die Welt, aber das Dorf sollte mensch bei Gelegenheit mal entseuchen.		`negative`

Table 8: Overview of the label verification. We randomly chose these examples from the original and the reformulated examples. We manually checked them and found that the annotated task labels do not change.

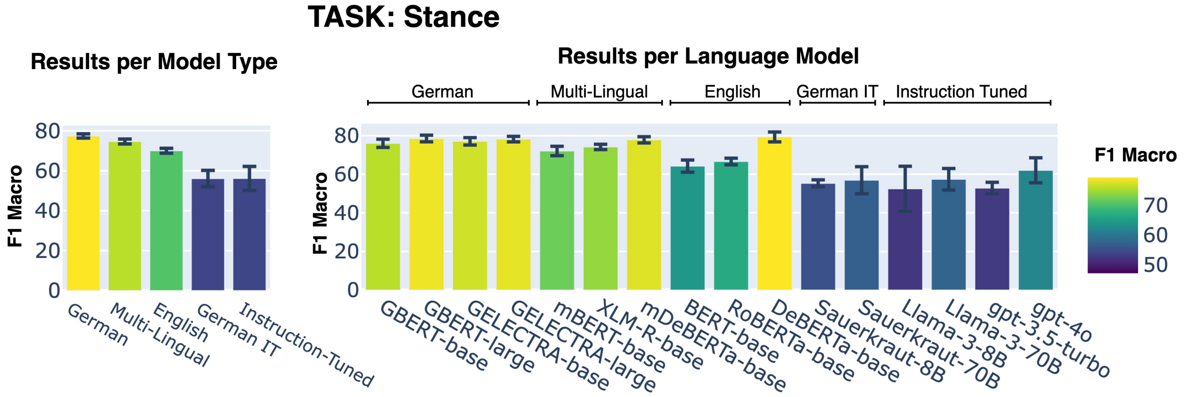


Figure 11: Mean performance and standard deviation for the `stance` task averaged over and seeds (fine-tuning) or prompting templates (ICL) by model type (*left*) or specific LM (*right*).

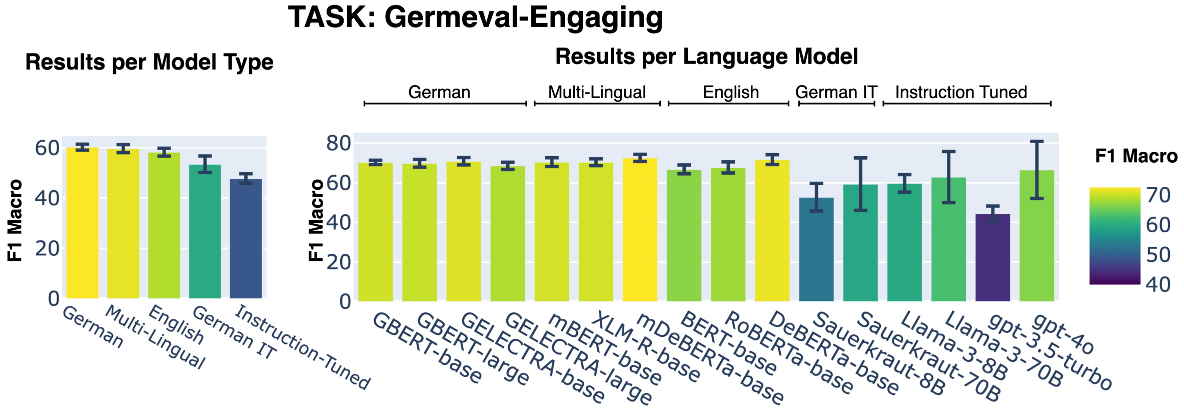


Figure 12: Mean performance and standard deviation for the `GermEval-Engaging` task averaged over and seeds (fine-tuning) or prompting templates (ICL) by model type (*left*) or specific LM (*right*).

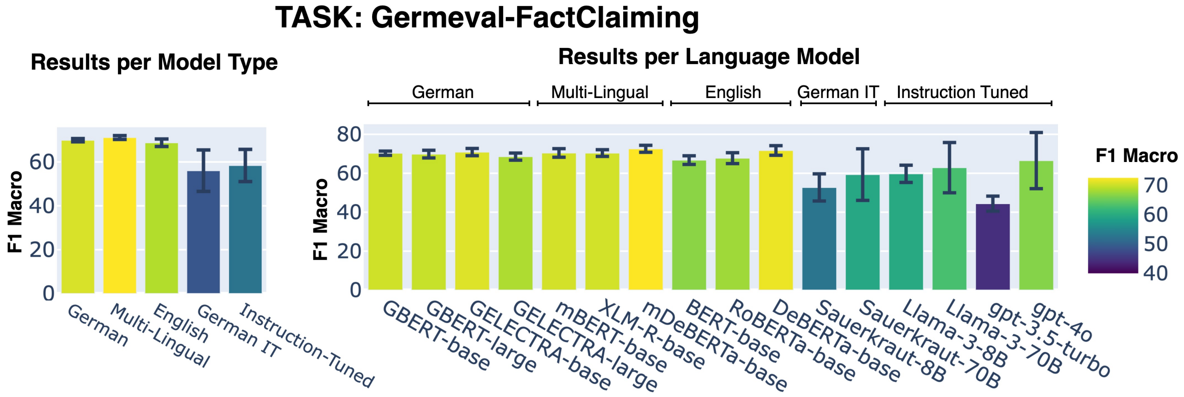


Figure 13: Mean performance and standard deviation for the `GermEval-Fact-Claiming` task averaged over and seeds (fine-tuning) or prompting templates (ICL) by model type (*left*) or specific LM (*right*).

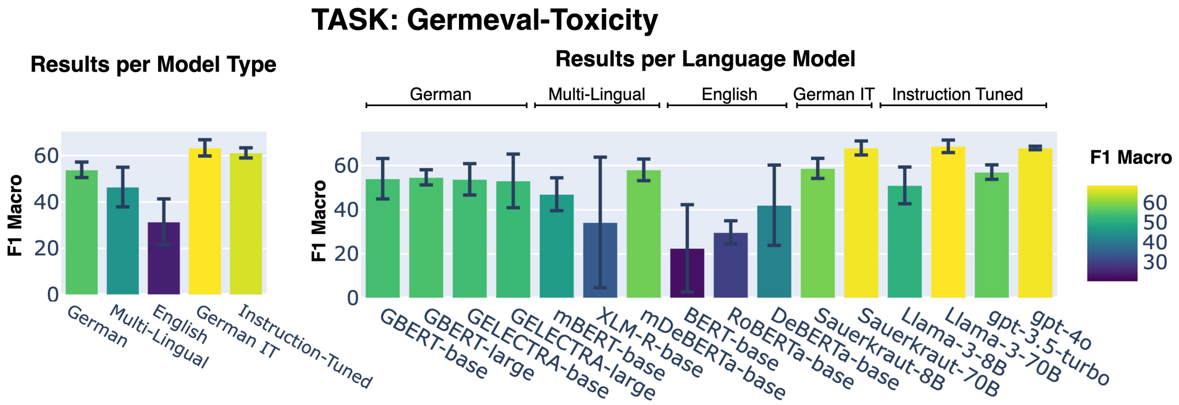


Figure 14: Mean performance and standard deviation for the `GermEval-Toxicity` task averaged over and seeds (fine-tuning) or prompting templates (ICL) by model type (*left*) or specific LM (*right*).

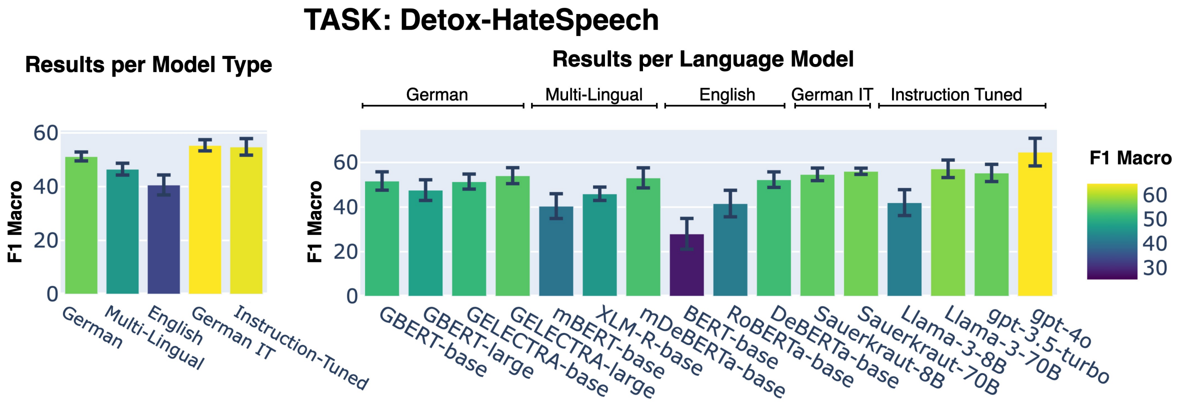


Figure 15: Mean performance and standard deviation for the `Detox-Hate-Speech` task averaged over and seeds (fine-tuning) or prompting templates (ICL) by model type (*left*) or specific LM (*right*).

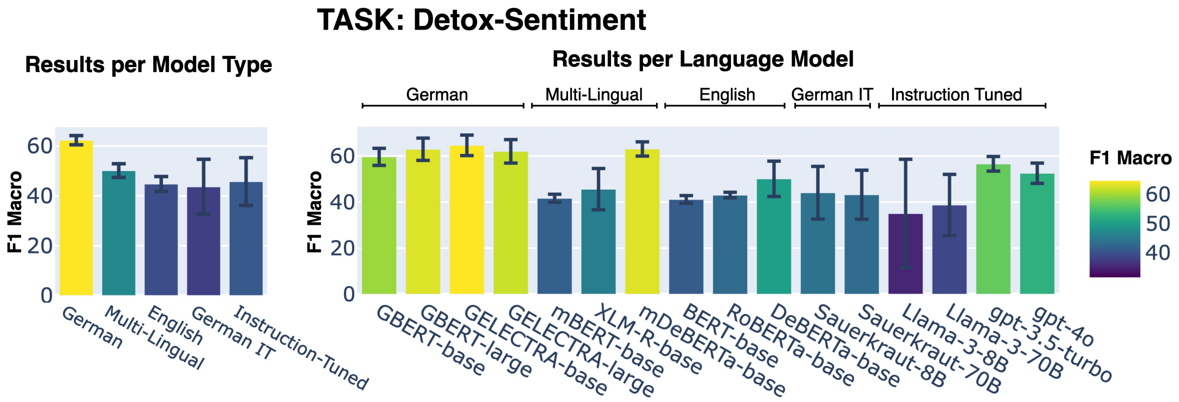


Figure 16: Mean performance and standard deviation for the `Detox-Sentiment` task averaged over and seeds (fine-tuning) or prompting templates (ICL) by model type (*left*) or specific LM (*right*).

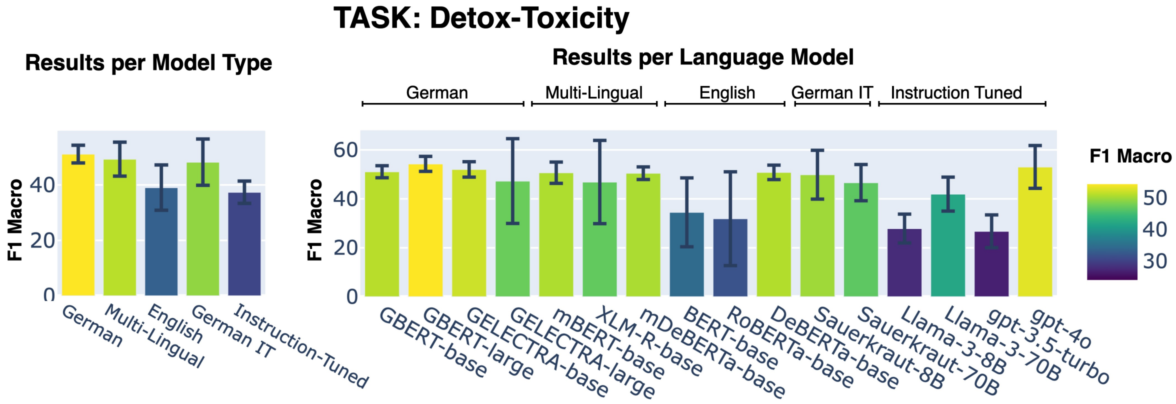


Figure 17: Mean performance and standard deviation for the `Detox-Toxicity` task averaged over and seeds (fine-tuning) or prompting templates (ICL) by model type (*left*) or specific LM (*right*).