Invisible Speakers? Gender Disparity in German AI Discourse and Its Reflection in Language Models

Milena Belosevic
Bielefeld University
Faculty of Linguistics and Literary Studies
milena.belosevic@uni-bielefeld.de

Abstract

This paper investigates how language models (LMs) reproduce the existing gender disparity found in German media discourse about artificial intelligence (AI). Building on a human-annotated corpus of quotations from German media discourse on AI, we first quantify the frequency with which male and female speakers are directly cited across domains and speaker roles. We then train LLäMmlein (Pfister et al., 2025), a state-of-the-art German-only language model, GBERT, and a logistic regression model using only the quoted text as input and without providing any gender cues to classify the quotation as originating from a male or female speaker. By comparing model predictions with corpus-based gold labels, we find that male voices dominate both the corpus and the model predictions. Balancing the data mitigates but does not fully eliminate this disparity, indicating that the strong male-default tendency of transformer models cannot be explained by corpus skew alone, but also by their priors from pretraining. The study contributes to the interpretability of language models' output for DH-related tasks, adaptation of NLP tools to domain-specific humanities corpora, and knowledge modelling in the humanities.

1 Introduction

Current language models are trained on large corpora in which male speakers are typically more visible than female speakers (Brennen et al., 2019), particularly in domains stereotypically associated with men, where male speakers are more frequently cited than female speakers. We hypothesise that such models replicate patterns from their training data and tend to assign the male label more regularly than the female label. This model behaviour can be tested in cases where gender is not explicitly marked in the text. However, to our knowledge, this kind of model performance has not been investigated on German-only language models.

In this paper, we draw on a gender classification task as a diagnostic probe. Therefore, the goal of the paper is not to build a "gender predictor" in which the model learns gender from linguistic cues in the quotes, nor to create a system that accurately infers the "true" gender of quoted speakers, but to test how the training distribution (including bias) affects the model's behaviour and how pre-existing biases in language models surface when they are applied to quotations. By removing overt gender markers and evaluating on a balanced test set, any systematic male–female asymmetry reflects learned associations in the model and/or in the training distribution rather than explicit cues. This makes the setup meaningful for DH use cases where such models may be used to (semi-)automatically tag speaker metadata in corpora.

Specifically, we focus on two research questions: (1) How are gender and social roles distributed in German AI discourse?, and (2) Do model predictions reproduce the empirical gender imbalance in our dataset? We compile a set of direct quotations from the German media discourse about AI, for a stratified subset in which the speaker's gender can be identified from external sources, and we compare the corpus-based gender distribution of cited speakers with the gender predictions produced by the fine-tuned German-only model LLäMmlein (Pfister et al., 2025). We compare its performance with GBERT and logistic regression.

After providing a theoretical background (Section 2) and methods (Section 3), we present results in Section 4 and discuss their implications in Section 5.

2 Related work

We use *disparity* as the main term for the corpus and model outputs because our primary measurements are descriptive and observable (Barocas

Proceedings of SIGHUM 2026, pages 66–79

et al., 2023): (1) the unequal representation of male vs. female quoted speakers in the dataset, and (2) asymmetric error rates of models. We do not claim to identify the underlying causal mechanisms (e.g., intentional discrimination, editorial practices, or broader societal inequities) and therefore avoid the more normatively loaded label *bias* for the empirical counts themselves. At the same time, we connect our findings to the broader bias-in-NLP literature (Bartl et al., 2025; Gallegos et al., 2024) by treating these disparities as downstream risks. When used as annotation assistants, models can systematically under-tag female speakers.

The topic has received growing attention with the advancement of language models (Jenny et al., 2024; Ho et al., 2025) and use mainly English data (Yang et al., 2025). Recent studies focus on detecting gender bias in literary texts or song lyrics, using English data and word embeddings or topic modelling methods (Chen et al., 2025). Another line of research is centred around identifying gender disparities in LLM-generated responses (Rhue et al., 2024; Wan et al., 2023; Fulgu and Capraro, 2024), creating alignment datasets (Zhang et al., 2025) or proposing (LLM-based) methods for identifying gender imbalances in training data (Derner et al., 2025). In German, other types of bias, such as epistemological bias in parliamentary debates, were investigated in Rehbein et al. 2024. Brennen et al. 2019 examine which AI experts are granted authority to shape narratives in the media over 30 years in the US and the UK, noting that the ten most-cited AI scholars account for 70% of the media space in both countries. They also find a strong bias against female experts in their dataset. Nguyen and Hekman 2022 found that reporting about AI is dominated by a small group of 'AI alpha journalists' who account for a disproportionate number of articles. Ryazanov et al. 2024 observe a significant increase in mentions of government agencies and people in leadership positions (via the Leadership frame), suggesting that industrial leaders and governments play a greater role in AI discourse post-ChatGPT.

Quotation detection is only part of our methodology in Section 3, and we do not aim to develop a new approach to this technique. However, it is noteworthy that many studies address methods for automatic quotation identification and attribution, but usually do not focus on gender asymmetries (e.g., Petersen-Frey and Biemann, 2024b; Brunner, 2013; Tekir et al., 2023; Petersen-Frey and Biemann, 2024a) and use literary dialogues as a dataset (Van Cranenburgh and Van Den Berg 2023, Underwood et al. 2018).

3 Methodology

Data collection and preprocessing. We queried the subcorpus Webmonitor from the Digitales Wörterbuch der deutschen Sprache (DWDS)[1] for the token *KI* (künstliche Intelligenz 'artificial intelligence') for the period 01.01.2025-01.12.2025. The Webmonitor subcorpus (Nolda et al., 2023) is a daily-updated collection of German-language web sources. It is therefore particularly suitable for observing current linguistic trends. It comprises online texts, blogs, and news articles on various topics (politics, sports, lifestyle). Given the high number of hits (131,094) for this period, we restricted the search results to one month (31.10.2025-30.11.2025) to make human annotation manageable. This restriction yields 15,003 hits. Based on the search results, we constructed a corpus of articles that explicitly use this term. Henceforth, we refer to this corpus as the AI Corpus. Note that we define AI discourse as quotations from news and web articles whose main topic is AI. Many quotations do not explicitly mention AI or use the term, but they address its consequences, risks, governance, or related economic and social issues. We exported hit lists with source URLs from DWDS. Since DWDS does not support full-text retrieval, we crawled and cleaned full-text articles using Trafilatura (Barbaresi, 2021). We retrieved 6,059 full articles. Importantly, DWDS was used only to export hit lists including article URLs, not to retrieve full texts. Full texts were downloaded from the linked publisher pages and extracted with Trafilatura based on the exported URL lists.

Quotes subset and preprocessing. We automatically extracted candidate quotations and speaker candidates from the full articles using a combined regex–spaCy heuristic. First, we identified all text segments enclosed in German quotation marks („…“, “…“, »…«) in each article and retained a context window of ±200 characters around each segment. On these context snippets, we applied the German spaCy model (de_core_news_md) to obtain tokenisation, POS tags, lemmata and dependency parses. We then searched each snippet for a reporting verb whose lemma belongs to a manually

[1] https://www.dwds.de/

compiled list of communication verbs and whose grammatical subject (dependency labels sb/nsubj) refers to a speaker candidate. Our reporting-verb list is intentionally conservative: we prioritise precision in speaker attribution over exhaustive coverage, and accept lower recall because candidates are subsequently sampled and manually filtered for relevance.

As potential speaker candidates, we treated (i) proper names, (ii) personal pronouns and (iii) noun subjects whose dependency subtree contains person-like elements, capturing patterns like *die Sprecherin von Bitkom, Anja Müller* 'the speaker of Bitkom, Anna Müller'. For each quotation, we stored a boolean flag indicating whether such a reporting-verb–subject configuration was present, as well as the surface form of the subject as a *speaker_candidate*. Only quotations with at least one speaker candidate near a reporting verb were retained as candidates for subsequent manual annotation. The quotations coming from the same speaker were not merged. In this way, we extracted 7,142 quotations.

Quotations that were clearly not relevant to our research question (e.g., quoted dialogues from book and song titles, or ironical quotes like KI "understands") were manually excluded by the author (3174 cases). This cleaned subset, comprising 3968 quotations, was provided to human annotators. We refer to this subset as the Quotation Corpus.

For all modelling experiments, we used only the quotation text as input and systematically removed explicit gender cues referring to the speaker (e.g. names, pronouns, or gendered titles). Thus, the models were exposed to quotations that, as far as could be automatically and manually detected, contained no direct gender information about the quoted person. Indirect correlations between wording and gender (e.g. topic or style differences) may remain. We did not alter non-speaker-related gendered language inside the quote.

Human annotation. Two annotators first examined the quotations from the corpus for their relevance to AI discourse, providing a binary decision for each quotation (relevant or irrelevant) based on annotation guidelines[2]. Annotators saw the quotation together with the surrounding reporting clause and ±1 sentence of context. Because of cognitive load and potential inconsistencies (annotators may read different amounts of text), we retain full articles only for complex cases during the adjudication step. The agreement between annotators was substantial (Cohen's κ = 0.72). Cases of disagreement were resolved through majority vote performed by the author. Of 3968 quotations, 3195 were annotated as relevant, and 773 were labelled as irrelevant after adjudication.

The author annotated these 3195 quotations for the perceived gender of the quoted speaker (male, female, or neutral/unknown). Because gender was not identifiable in all quotations and the annotators were not provided with the full texts, the speaker's gender was annotated by the author based on context or, in ambiguous cases, on the full article. To estimate inter-annotator reliability, 200 of the 3,195 quotations with clearly identifiable speaker gender were randomly sampled and independently labelled by a second annotator, yielding Cohen's $\kappa = 0.98$ (almost perfect agreement with the author's labels). After adjudication, these quotations remained in the corpus and are included in the final dataset used for modelling. We treat gender as a property of how speakers are represented in the media, not as a claim about their identity, and explicitly acknowledge the limitations of binary gender labels. Quotations annotated as *neutral/unknown* (277 of 3,195, e.g., multiple speakers or only a surname mentioned in the full text) were excluded, leaving 2,918 quotations for modelling.

The two annotators who assessed the relevance of quotations for AI discourse also annotated the 2,918 quotations for *social_role* (company_executive, expert, researcher, politician, journalist, spokesperson, worker, user, artist, cleric, activist, celebrity) and *domain* (work, everyday use, IT domain, finance, culture and arts, data_privacy, education, health, cybersecurity, military) according to the annotation guidelines. Domain labels were derived from outlet section metadata (via DWDS hit lists) and harmonised into a unified scheme. Source-role labels follow the taxonomy proposed by Asr et al. 2021b. The author resolved disagreements to obtain a gold standard. Inter-annotator agreement was very high, with Cohen's $\kappa = 0.982$ for social roles and $\kappa = 0.922$ for domains, providing a robust basis for the subsequent modelling experiments. The *speaker_candidate* field was automatically identified (see above). The author reviewed these candidates and, where necessary, drew on the full article to provide information about the speaker's job or role.

[2]Annotation guidelines and code are available at: https://osf.io/m5xqw/overview

Baselines. As a trivial reference point, we include a majority baseline that always predicts the majority class ("male") observed in the annotated dataset. This baseline reflects the global skew in the data. Evaluating all models against this baseline on the balanced test set allows us to see whether they learn anything beyond reproducing the overall male dominance. As a standard supervised NLP baseline, we train a logistic regression classifier on TF–IDF representations of the quotation text. Unlike pretrained language models, this classifier has no external-world knowledge and no pretraining corpus. Any systematic preference for male or female labels would therefore directly reflect the labelled AI data. Next, we fine-tune the GBERT base (deepset/gbert-base, Chan et al. 2020) transformer model as a binary sequence classifier with labels *male* and *female*. We select GBERT as a strong German encoder baseline because it is a widely used BERT-style model pretrained on large German corpora (news, web text, Wikipedia), thereby providing a well-established reference point for German text classification.

Main model. We use the LLäMmlein 1B (LSX-UniWue/LLaMmlein_1B, Pfister et al. 2025), an open German-only decoder-only language model trained from scratch on German data. In contrast to GBERT's encoder architecture, LLäMmlein is a generative transformer that is typically used for text generation. We adapt it to our classification task by fine-tuning it with QLoRA, a memory-efficient adapter method well-suited to small datasets (Dettmers et al., 2023). By comparing GBERT and LLäMmlein, we can test whether the observed male-default tendency persists across different transformer types. The TF–IDF/logistic baseline serves as a non-pretrained control. The key hyperparameters for GBERT-N/B, LLäMmlein-N/B, and the TF–IDF/logistic regression baseline are provided in Appendix Table 14. We also explored several prompting-based approaches (zero- and three-shot prompting with the unfine-tuned LLäMmlein-1B model), but all proved methodologically unreliable. For all models, the inputs are degendered quotations, without broader article context. We excluded quotations with fewer than four tokens (191 of the 2,918 manually annotated quotations) from modelling experiments, as they provide too little linguistic material for gender classification. From the 2,727 remaining quotations, we first drew a fixed, held-out test set of 171 quotations (86 female, 85 male) by stratified random sampling on gender only. All models are evaluated on this same balanced test set. The remaining 2,556 quotations form the pool for training and development. In the natural condition (-N), we randomly split this pool into 2,044 training instances and 512 development instances, preserving the observed gender skew. In the balanced condition (-B), we downsample the non-test items only by gender to create a 50/50 balanced training set of 796 quotations and a separate balanced development set of 171 quotations. Thus, -N and -B are two alternative ways of partitioning the same 2,556 non-test items. In both cases, the balanced test set is never used for downsampling or model selection. Alternative strategies such as class-weighted loss functions were considered but not pursued, as our primary aim was to contrast a "natural" gender distribution with an explicitly balanced training regime in a transparent and easily interpretable way. Exploring more fine-grained debiasing techniques (e.g. class weights, focal loss) is left for future work. The -N/-B design was applied only to the two transformer models to assess how they reflect the gender disparity in our dataset and how changes in the training distribution affect their performance. Logistic regression is trained and tuned on the same balanced train/validation splits as the -B transformer models and evaluated on the same balanced test set.

Evaluation metrics include overall accuracy, macro-averaged F1 score, per-gender precision and recall, and the proportion of quotations predicted as male versus female. The qualitative error analysis focuses on misclassifications across social roles and domains.

4 Results

Descriptive statistics. Of 2,918 manually annotated quotations, 79% are attributed to male speakers (2,304 instances) and only 21% to female speakers (614 cases), indicating a substantial gender disparity in the dataset and supporting results obtained on English data (Asr et al., 2021a).

The distribution of social roles is highly uneven (see Appendix Table 8). Half of all quotations are attributed to company executives (1,459 instances), followed by experts (15%), researchers (11%) and politicians (7.8%); all other roles (artist, user, worker, spokesperson, journalist, cleric, activist, celebrity) each account for less than 4% of the corpus. In almost all roles, male speakers pre-

dominate: for example, 85.3% of company executive quotations and 77.9% of expert quotations are attributed to men, and men almost exclusively hold roles such as activist (95% male). Female speakers dominate only in a few comparatively small categories, such as journalists (64% female), and are closer to parity among spokespersons (44.8% female), users (36.4% female), and workers (38.6% female, all row-wise percentages). Because several of these roles have low absolute counts, we treat such patterns as descriptive tendencies rather than robust statistical differences. Human annotations serve as a gold standard to assess how different models reproduce this gender disparity.

The two largest domains are work (31.2% of all quotations) and everyday use (29.8%), followed by IT (17.6%). Other domains, such as finance, education, culture, or health, account for much smaller shares (Appendix Table 11). Within almost all domains, male speakers clearly predominate. For example, 84.3% of quotations in the work domain and 86.9% in the IT domain are attributed to men, with similarly high male shares in finance (76.4%), data_privacy (81.3%) and especially cybersecurity (94.1%). Female speakers constitute a higher relative proportion only in a few areas, most notably education (49.5% female, essentially parity), culture (35.3% female), and health (36.1% female). Column-wise percentages show that male quotations are concentrated in work, everyday use and IT. In contrast, female quotations are more dispersed and relatively more frequent in everyday use, education, culture and health. As with social roles, some of these domain-specific patterns involve small absolute numbers and should be interpreted cautiously, but overall they reinforce the picture of an AI discourse in which male voices dominate across most topics. Tables 1 and 2 summarise the gender distribution across most frequent social roles and domains (for a more detailed breakdown, see Appendix Tables 9,10,12, and 13).

Finally, we examined gender based differences in quotation length. Overall, quotations in the corpus are relatively short (median length 16 tokens, IQR 9–26), with a long tail of longer quotations (up to 188 tokens). Female quotations have a mean length of 19.8 tokens (median 15), whereas male quotations have a mean length of 20.4 tokens (median 16). These minor differences suggest that the strong gender imbalance observed in our data is not an artefact of men being quoted at much greater length.

Social role	N	**Female** n **(%)**	**Male** n **(%)**
company ex.	1459	215 (14.7%)	1244 (85.3%)
expert	439	97 (22.1%)	342 (77.9%)
researcher	320	89 (27.8%)	231 (72.2%)
politician	228	56 (24.6%)	172 (75.4%)
other[3]	472	157 (33.3%)	315 (66.7%)
Total	**2918**	**614 (21.0%)**	**2304 (79.0%)**

Table 1: Gender distribution across major social roles in the Quotation Corpus (row-wise percentages: for each social role, they indicate the proportion of female vs. male quotations within that role).

Domain	N	**Female** n **(%)**	**Male** n **(%)**
work	909	143 (15.7%)	766 (84.3%)
everyday use	869	216 (24.9%)	653 (75.1%)
IT	514	67 (13.0%)	447 (87.0%)
other[4]	626	188 (30.0%)	438 (70.0%)
Total	**2918**	**614 (21.0%)**	**2304 (79.0%)**

Table 2: Gender distribution across major domains in the Quotation Corpus (row-wise percentages).

Majority baseline. On the balanced test set, this baseline reaches an accuracy of 0.50 but completely fails to recognise female quotations: male instances are classified correctly in 100% of cases, while all 86 female quotations are misclassified as male (F1 = 0.00 for female; confusion matrix: 85 male→male, 86 female→male). This behaviour mirrors a pure "male-by-default" strategy and provides a lower bound against which we can compare more sophisticated models.

Logistic regression. On the balanced test set, this model provides a comparatively neutral benchmark: accuracy and macro-F1 are 0.57, with almost identical performance for male and female quotations (F1 $\approx$ 0.57 for both classes). Compared with the majority baseline, the TF–IDF model moves beyond reproducing the global skew and instead offers a relatively balanced treatment of male and female speakers (37 female→male vs. 36 male→female misclassifications on the balanced test set), providing a useful classical benchmark for transformer- and LLM-based models.

[3]Aggregated category comprising artist, user, worker, spokesperson, journalist, cleric, activist, and celebrity.

[4]Aggregated category comprising finance, culture, data_privacy, education, health, cybersecurity and military.

As shown in Appendix Tables 16 and 15, within individual roles and domains, row-wise proportions fluctuate (e.g., a slight dominance of male→female errors for company executive and expert, and more female→male errors in finance and culture), but these patterns are based on small absolute counts and remain relatively balanced overall. For example, the model incorrectly predicts that a quotation by a female researcher in finance is authored by a male. However, it also predicts a female speaker for a quotation by a male expert in the IT domain, which contrasts with stereotypical associations with typical male and female professions and expertise (see Table 3[5]). This error profile supports our interpretation of the TF–IDF model as a comparatively "neutral" baseline on the balanced data, against which the more strongly gender-skewed behaviour of GBERT and LLäMmlein can be contrasted.

Quote	Gold	Pred	Role / Domain
It appears that investors are taking profits on these AI-related stocks.	female	male	finance / expert
AI is trained, not programmed.	male	female	expert / IT

Table 3: Logistic regression misclassification examples.

GBERT. Under natural training (-N), GBERT exhibits a strong male-default behaviour on the balanced test set: a good performance on the validation set (val. accuracy ≈0.81, macro-F1 ≈0.53 after three epochs) changes completely when we evaluate on the balanced test set, where overall accuracy drops to 0.49 (chance level), with an F1-score of 0.65 for male but only 0.06 for female. The confusion matrix shows that GBERT correctly identifies 81 of 85 male quotations (recall 0.95) but misclassifies 83 of 86 female quotations as male (recall 0.03). In total, it assigns the male label to 164 of 171 test instances (96%) despite the 50/50 class balance. Taken together, GBERT-N largely reproduces the training skew and heavily underrepresents female speakers in quote-only classification.

The qualitative error analysis shows highly asymmetric error patterns. All misclassified male quotes cluster around the use of AI in domains stereotypically associated with women, such as children, parenting, and education (see Table 4). Similarly, GBERT-N frequently misclassifies female roles stereotypically associated with men, such as female company executives or experts (see Appendix Table 16 for a quantitative summary). A similar pattern appears at the domain level. Female speakers in the core AI domains, such as IT, are almost exclusively misclassified as male. Only in everyday use do we observe a small number of male→female errors at all. In contrast to the comparatively symmetric errors of the logistic regression baseline, these results show that GBERT-N does not merely make random errors on the balanced test set. Instead, it exhibits a male default, which is particularly evident among female speakers in high-status roles and in central AI domains.

Quote	Gold	Pred	Role / Domain
The models reproduce these mental patterns and reinforce them.	female	male	researcher / IT
AI cannot replace the relationship with friends or parents.	male	female	expert / use

Table 4: GBERT-N misclassification examples.

Balanced fine-tuning of GBERT (GBERT-B) improves over the naturally trained model, but a male-default tendency remains. After three epochs, validation accuracy stabilises at ≈0.59 (macro-F1 = 0.59). On the balanced test set, GBERT-B achieves an accuracy of 0.56 and a macro-F1 of 0.54. However, the confusion matrix shows a clear asymmetry: 67/85 male quotations are classified correctly (recall 0.79) but only 29/86 female quotations (recall 0.34), with female→male errors still more than three times as frequent as male→female (57 vs. 18).

Qualitative examples (Table 5) illustrate that misclassified male examples are related to domains stereotypically associated with women (e.g., feelings), and vice versa: topics such as robotics are wrongly associated with male speakers (see examples in Table 5). Across social roles, female company executives, researchers, and experts are much more often misclassified as male than vice versa (see Appendix Table 15 for a quantitative summary). Overall, balancing mitigates but does not fully remove the model's inherited gender dis-

[5]In all tables, examples are translated from German, glosses are omitted for space reasons.

parity.

Quote	Gold	Pred	Role / Domain
When robots are equipped with AI, AI gets a body.	female	male	journalist / IT
AI does not feel sad when it experiences 'pain'.	male	female	comp. exec. / IT

Table 5: GBERT-B misclassification examples.

LLäMmlein (natural skew). Under natural training (-N), LLäMmlein also exhibits an evident male-default tendency on the balanced test set. After three epochs, it reaches an accuracy of 0.55 and a macro-F1 of 0.47. Class-wise, it performs well on male quotations (F1 = 0.68, recall = 0.94), but much worse on female quotations (F1 = 0.27, recall = 0.16). The confusion matrix shows that it correctly identifies 80 of 85 male quotes, but misclassifies 72 of 86 female quotes as male, with only 14 female quotes correctly recognised.

The error analysis again reveals a strongly asymmetric pattern. Especially female company executives are systematically "pulled" towards the male label (see examples in Table 6). Across domains, we observe the same disparity: female speakers in IT and finance are overwhelmingly misclassified as male, with male-to-female errors occurring only sporadically in everyday use and work (see Appendix Table 16 for a quantitative summary). Compared with the more symmetric error profile of the logistic regression baseline, LLäMmlein-N thus behaves very similarly to GBERT-N.

Quote	Gold	Pred	Role / Domain
The algorithm itself plays an important role (...).	female	male	expert / use
AI chatbots pose a serious threat to our children (...).	male	female	politician / use

Table 6: LLäMmlein-N misclassification examples.

LLäMmlein (balanced fine-tuning). Balanced fine-tuning of LLäMmlein (LLäMmlein-B) leads to a much more symmetric behaviour than the naturally trained variant. On the balanced test set, the model achieves an accuracy of 0.57 and a macro-F1 of 0.57, with performance similar for male (F1 = 0.58, recall = 0.61) and female quotations (F1 = 0.55, recall = 0.52). Overall, despite a slight tendency to overpredict male labels, balancing the training data substantially reduces the strong male default observed in LLäMmlein-N.

However, domain-linked gender associations remain. For example, the model assumes a female speaker for educational contexts and a male speaker in the health domain (see Table 7). The errors are almost evenly distributed (41 female→male and 33 male→female). The symmetry in social roles is most evident among company executives, where female and male speakers are misclassified as the opposite gender at equal rates. In other roles, however, female researchers and experts are still more often "pulled" towards the male label, whereas some smaller categories (e.g., journalists) show only female→male errors. Across domains, LLäMmlein-B behaves more evenly: in IT and culture, errors are roughly balanced, and in everyday use male→female errors even outnumber female→male ones, whereas in finance, health, and military, all misclassifications still go from female to male (see Appendix Tables 16 and 15 for a quantitative summary).

Quote	Gold	Pred	Role / Domain
With AI, we bring precision to the surgery room.	female	male	politician/health
This is how we combine digital innovation with real pedagogical impact.	male	female	expert/education

Table 7: LLäMmlein-B misclassification examples.

Comparing GBERT and LLäMmlein shows that per-gender gaps (Δ = male − female for F1 and recall) are large under -N: GBERT-N yields ΔF1= 0.59 and ΔRecall= 0.92, and LLäMmlein-N ΔF1= 0.41 and ΔRecall= 0.78. Under -B, LLäMmlein becomes close to symmetric (ΔF1= 0.04, ΔRecall= 0.09), whereas GBERT still shows a notable gap (ΔF1= 0.21, ΔRecall= 0.45). A simple non-parametric bootstrap over the 171 test instances (B=10,000; percentile 95% CI) confirms that the -N gaps are far larger than sampling noise (e.g., GBERT-N ΔF1 CI [0.49, 0.67]; LLäMmlein-N [0.30, 0.53]), while the balanced LLäMmlein-B gap is compatible with zero (CI includes 0).

Exploratory tests find no statistically significant domain dependence of error direction: only LLäMmlein-B shows a modest omnibus associ-

ation (permutation p=0.024), with finance being suggestive (Fisher p=0.006; FDR q=0.057). For social roles, an omnibus permutation test indicates an association only for GBERT-N (p=0.026); male→female errors are rare and confined to few roles, and no role remains significant after FDR.

5 Conclusions

This paper investigated the extent to which German language models reproduce the existing gender disparities found in German media discussions about AI. The approach accounts for the need to contextualise fairness criteria used to evaluate biases in LLM systems (Anthis et al., 2025) by focusing on a specific context (public AI discourse).

On a balanced test set (50% male / 50% female), neutral behaviour would correspond to roughly symmetric performance across genders. Instead, we observe that both GBERT-N and LLäMmlein-N systematically overpredict male speakers when female speakers are equally represented in the gold data. Therefore, gender disparity in their predictions cannot be reduced to the annotated data alone but is also shaped by their pretraining and architecture. This result is also supported by the observation that a logistic regression baseline performs on par with, or slightly better than, GBERT and LLäMmlein on the balanced test set and exhibits a highly symmetric treatment of male and female quotations. The value of more complex architectures in our study lies, therefore, less in raw performance and more in their ability to reveal the interaction between training distributions (natural vs. balanced) and gender disparities in predictions. Hence, if DH researchers use GBERT or LLäMmlein as an annotation assistant for speaker gender in AI corpora, they risk systematic under-tagging of female speakers, especially in high-status roles. To mitigate skew, they should monitor per-group errors and rely on human validation.

In addition to the mitigation strategies (see e.g., Ferrara 2023), we recommend raising researchers' awareness of such biases by cultivating healthy distrust (Paaßen et al., 2025) toward model outputs.

Retraining on a balanced subset of the same material improves the performance of both models, yet in different ways: GBERT-B becomes a stronger classifier overall but still systematically prefers the male label, while LLäMmlein-B not only gains in macro-F1 but also moves much closer to a symmetric treatment of male and female speakers, with comparable recall and more balanced error patterns. Therefore, model choice and training regime matter for downstream DH analyses, such as authorship attribution (Rybicki, 2025). From a DH perspective, this also suggests that simple data balancing cannot completely eliminate gender disparities in model outputs, and that different German architectures respond differently to such corrective interventions, which is highly relevant for the use of these models in the study of public discourse. Additionally, the comparison between a natural and a balanced training illustrates that fine-tuning is not a neutral operation: depending on how we curate training data, we can obtain models with different bias profiles, which is highly relevant for DH scenarios where such models might be used as annotation assistants.

Overall, this paper makes three contributions: (1) a quote corpus on German AI media discourse annotated with gender, social roles, and domains (2) a diagnostic probe setup for gender prediction, enabling interpretable analysis of model priors, and (3) empirical evidence that transformer-based models can exhibit a male-default tendency that persists under balancing, with implications for using LMs as annotation tools in DH pipelines.

Limitations and future work

The study has several limitations. First, the analysis is based on a one-month snapshot of texts from a single dataset. As a result, the observed gender distributions and model behaviours reflect this specific period and text type. They should not be interpreted as long-term trends in German media as a whole. Future work will extend the time span and include additional domains to examine temporal variation and generalizability beyond AI discourse.

While we treat gender as the target variable, a complementary line of work could investigate implicit gender disparity, i.e., whether models treat male and female speakers differently on other tasks such as credibility assessment, stance detection or topic classification. This would involve using gender only as a group attribute and analysing differences in model behaviour across gender groups. Furthermore, reliance on binary gender labels may limit the generalizability of our findings to non-binary or gender-diverse speakers. Next, our extraction heuristic is intentionally conservative (high precision for speaker attribution) and does not aim to capture all quotations. As estimating overall re-

call would require full manual annotation of quotes, we leave it for future work.

Finally, since we trained and evaluated the models on quotations without their local context and speaker-related gender markers, future work should examine model behaviour by including (de-gendered) local context.

References

Jacy Reese Anthis, Kristian Lum, Michael Ekstrand, Avi Feller, and Chenhao Tan. 2025. The impossibility of fair LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 105–120, Vienna, Austria. Association for Computational Linguistics.

Fatemeh Torabi Asr, Mohammad Mazraeh, Alexandre Lopes, Vagrant Gautam, Junette Gonzales, Prashanth Rao, and Maite Taboada. 2021a. The gender gap tracker: Using natural language processing to measure gender bias in media. *PLOS ONE*, 16(1):e0245533.

Fatemeh Torabi Asr, Mohammad Mazraeh, Alexandre Lopes, Vasundhara Gautam, Junette Gonzales, Prashanth Rao, and Maite Taboada. 2021b. The gender gap tracker: Using natural language processing to measure gender bias in media. *PLoS One*, 16(1):e0245533.

Adrien Barbaresi. 2021. Trafilatura: A web scraping library and command-line tool for text discovery and extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 122–131, Online. Association for Computational Linguistics.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.

Marion Bartl, Abhishek Mandal, Susan Leavy, and Suzanne Little. 2025. Gender bias in natural language processing and computer vision: A comparative survey. *ACM Computing Surveys*, 57(6).

J Brennen, Anne Schulz, Philip Howard, and R Nielsen. 2019. Industry, experts, or industry experts? Academic sourcing in news coverage of AI. *RIJS Factsheets*.

Annelen Brunner. 2013. Automatic recognition of speech, thought, and writing representation in german narrative texts. *Literary and Linguistic Computing*, 28(4):563–575.

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.

Danqing Chen, Adithi Satish, Rasul Khanbayov, Carolin Schuster, and Georg Groh. 2025. Tuning into bias: A computational study of gender bias in song lyrics. In *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*, pages 117–129, Albuquerque, New Mexico. Association for Computational Linguistics.

Erik Derner, Sara Sansalvador De La Fuente, Yoan Gutierrez, Paloma Moreda Pozo, and Nuria M Oliver. 2025. Leveraging large language models to measure gender representation bias in gendered language corpora. In *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 468–483, Vienna, Austria. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized LLMs. *Preprint*, arXiv:2305.14314.

Emilio Ferrara. 2023. Should ChatGPT be biased? Challenges and risks of bias in large language models. *First Monday*.

Raluca Alexandra Fulgu and Valerio Capraro. 2024. Surprising gender biases in GPT. *Computers in Human Behavior Reports*, 16:100533.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Jerlyn Q.H. Ho, Andree Hartanto, Andrew Koh, and Nadyanna M. Majeed. 2025. Gender biases within artificial intelligence and chatgpt: Evidence, sources of biases and solutions. *Computers in Human Behavior: Artificial Humans*, 4:100145.

David F. Jenny, Yann Billeter, Bernhard Schölkopf, and Zhijing Jin. 2024. Exploring the jungle of bias: Political bias attribution in language models via dependency analysis. In *Proceedings of the Third Workshop on NLP for Positive Impact*, pages 152–178, Miami, Florida, USA. Association for Computational Linguistics.

Dennis Nguyen and Erik Hekman. 2022. The news framing of artificial intelligence: a critical exploration of how media discourses make sense of automation. *AI & Society*.

Andreas Nolda, Adrien Barbaresi, and Alexander Geyken. 2023. *Korpora für die lexikographische Beschreibung diatopischer Variation in der deutschen Standardsprache*, pages 29–52. De Gruyter, Berlin, Boston.

Benjamin Paaßen, Suzana Alpsancar, Tobias Matzner, and Ingrid Scharlau. 2025. Healthy distrust in AI systems.

Fynn Petersen-Frey and Chris Biemann. 2024a. Dataset of quotation attribution in German news articles. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4412–4422, Torino, Italia. ELRA and ICCL.

Fynn Petersen-Frey and Chris Biemann. 2024b. Fine-grained quotation detection and attribution in German news articles. In *Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024)*, pages 196–208, Vienna, Austria. Association for Computational Linguistics.

Jan Pfister, Julia Wunderle, and Andreas Hotho. 2025. LLäMmlein: Transparent, compact and competitive German-only language models from scratch. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2227–2246, Vienna, Austria. Association for Computational Linguistics.

Ines Rehbein, Josef Ruppenhofer, Annelen Brunner, and Simone Paolo Ponzetto. 2024. Out of the mouths of MPs: Speaker attribution in parliamentary debates. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12553–12563, Torino, Italia. ELRA and ICCL.

Lauren Rhue, Sofie Goethals, and Arun Sundararajan. 2024. Evaluating LLMs for gender disparities in notable persons.

Igor Ryazanov, Carl Öhman, and Johanna Björklund. 2024. How ChatGPT changed the media's narratives on AI: A semi-automated narrative analysis through frame semantics. *Minds & Machines*, 35(1).

Jan Rybicki. 2025. Can machine translation of literary texts fool stylometry? *Digital Scholarship in the Humanities*, 40(1):268–276.

Selma Tekir, Aybüke Güzel, Samet Tenekeci, and Bekir Haman. 2023. Quote detection: A new task and dataset for NLP. In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 21–27, Dubrovnik, Croatia. Association for Computational Linguistics.

Ted Underwood, David Bamman, and Sabrina Lee. 2018. The transformation of gender in English-language fiction. *Journal of Cultural Analytics*.

Andreas Van Cranenburgh and Frank Van Den Berg. 2023. Direct speech quote attribution for Dutch literature. In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 45–62, Dubrovnik, Croatia. Association for Computational Linguistics.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. "Kelly is a warm person, Joseph is a role model": Gender biases in LLM-generated reference letters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3730–3748, Singapore. Association for Computational Linguistics.

Jinrui Yang, Xudong Han, and Timothy Baldwin. 2025. Demographics and democracy: Benchmarking LLMs' gender bias and political leaning in European parliament. In *Proceedings of the 8th International Conference on Natural Language and Speech Processing (ICNLSP-2025)*, pages 416–439, Southern Denmark University, Odense, Denmark. Association for Computational Linguistics.

Tao Zhang, Ziqian Zeng, YuxiangXiao YuxiangXiao, Huiping Zhuang, Cen Chen, James R. Foulds, and Shimei Pan. 2025. GenderAlign: An alignment dataset for mitigating gender bias in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11293–11311, Vienna, Austria. Association for Computational Linguistics.

A Appendix

A.1 Overall distribution of social roles

Social role	Count	Percent of all quotes
company_executive	1459	50.00%
expert	439	15.04%
researcher	320	10.97%
politician	228	7.81%
artist	115	3.94%
user	88	3.02%
worker	70	2.40%
spokesperson	67	2.30%
journalist	50	1.71%
cleric	47	1.61%
activist	20	0.69%
celebrity	15	0.51%
Total	**2918**	**100.00%**

Table 8: Distribution of social roles in the annotated AI Quotation Corpus.

A.2 Gender distribution by social role (row-wise percentages)

Social role	N	Female n	Male n	Female %	Male %
company_executive	1459	215	1244	14.74%	85.26%
expert	439	97	342	22.10%	77.90%
researcher	320	89	231	27.81%	72.19%
politician	228	56	172	24.56%	75.44%
artist	115	31	84	26.96%	73.04%
user	88	32	56	36.36%	63.64%
worker	70	27	43	38.57%	61.43%
spokesperson	67	30	37	44.78%	55.22%
journalist	50	32	18	64.00%	36.00%
cleric	47	0	47	0.00%	100.00%
activist	20	1	19	5.00%	95.00%
celebrity	15	4	11	26.67%	73.33%

Table 9: Gender distribution by social role in the annotated AI Quotation Corpus (row-wise percentages).

A.3 Gender distribution by social role (column-wise percentages)

Social role	Fem. % of all fem.	Male % of all males
company_executive	35.02%	53.99%
expert	15.80%	14.84%
researcher	14.50%	10.03%
politician	9.12%	7.47%
artist	5.05%	3.65%
journalist	5.21%	0.78%
user	5.21%	2.43%
worker	4.40%	1.87%
spokesperson	4.89%	1.61%
celebrity	0.65%	0.48%
activist	0.16%	0.82%
cleric	0.00%	2.04%

Table 10: Gender distribution by social role in the annotated AI Quotation Corpus (column-wise percentages).

A.4 Overall domain distribution

Domain	Count	Percent of all quotes
work	909	31.15%
everyday use	869	29.78%
IT	514	17.61%
finance	144	4.93%
culture	116	3.98%
data_privacy	112	3.84%
education	101	3.46%
health	86	2.95%
cybersecurity	34	1.17%
military	33	1.13%
Total	**2918**	**100.00%**

Table 11: Distribution of domains in the annotated AI Quotation Corpus.

A.5 Gender distribution by domain (row-wise percentages)

Domain	N	Female n	Male n	Female %	Male %
work	909	143	766	15.73%	84.27%
everyday use	869	216	653	24.86%	75.14%
IT	514	67	447	13.04%	86.96%
finance	144	34	110	23.61%	76.39%
culture	116	41	75	35.34%	64.66%
data_privacy	112	21	91	18.75%	81.25%
education	101	50	51	49.50%	50.50%
health	86	31	55	36.05%	63.95%
cybersecurity	34	2	32	5.88%	94.12%
military	33	9	24	27.27%	72.73%

Table 12: Gender distribution by domain in the annotated AI Quotation Corpus (row-wise percentages).

A.6 Gender distribution by domain (column-wise percentages)

Domain	Female % of all females	Male % of all males
work	23.29%	33.25%
everyday use	35.18%	28.34%
IT	10.91%	19.40%
finance	5.54%	4.77%
culture	6.68%	3.26%
data_privacy	3.42%	3.95%
education	8.14%	2.21%
health	5.05%	2.39%
cybersecurity	0.33%	1.39%
military	1.47%	1.04%

Table 13: Distribution of domains within each gender (column-wise percentages). Percentages indicate the distribution of each gender across domains.

A.7 Key hyperparameters and configurations for all models

Model	Variants	LR	Epochs	$\text{Batch}_{\text{train}}$	Max seq
GBERT	-N, -B	2×10^{-5}	3	16	128
LLäMmlein	-N, -B	2×10^{-4}	3	4	128
TF–IDF+LogReg	-B only	–	–	–	128[a]
Majority	–	–	–	–	–

Table 14: Key hyperparameters for all models. GBERT-N/B and LLäMmlein-N/B share the same optimisation settings within each architecture and differ only in the training distribution (natural vs. downsampled balanced).

A.8 Misclassifications by social role and model

	LogReg		GBERT-N		GBERT-B		LLäMmlein-N		LLäMmlein-B	
Social role	F→M	M→F	F→M	M→F	F→M	M→F	F→M	M→F	F→M	M→F
company_executive	11	15	30	0	22	7	25	1	16	16
expert	5	8	9	2	9	2	8	2	5	3
researcher	5	3	15	0	11	2	12	0	8	1
politician	6	4	10	1	6	2	11	1	7	5
journalist	3	0	5	0	3	0	4	0	2	0
spokesperson	2	0	3	0	2	0	3	0	2	0
user	2	1	5	0	2	0	3	0	1	1
artist	2	2	5	0	2	2	5	0	2	4
worker	1	0	1	0	0	0	1	0	0	0
activist	0	0	0	0	0	0	0	0	0	1

Table 15: Misclassifications by social role and model. F→M = female quotation predicted as male; M→F = male quotation predicted as female. Empty cells indicate zero errors for that combination.

A.9 Misclassifications by domain and model

	LogReg		GBERT-N		GBERT-B		LLäMmlein-N		LLäMmlein-B	
Domain	F→M	M→F	F→M	M→F	F→M	M→F	F→M	M→F	F→M	M→F
IT	3	6	8	0	7	2	5	0	4	4
finance	5	2	11	0	9	0	11	0	8	0
work	7	8	22	0	16	5	20	1	13	9
everyday use	15	17	27	4	12	9	22	4	7	14
education	0	1	2	0	2	1	2	0	1	2
culture	2	1	4	0	3	1	4	0	2	2
health	3	1	5	0	4	0	5	0	3	0
military	2	0	3	0	3	0	3	0	2	0
data_privacy	0	0	1	0	1	0	1	0	1	2

Table 16: Misclassifications by domain and model. F→M = female quotation predicted as male; M→F = male quotation predicted as female. Empty cells indicate zero errors for that combination.