Tackling the Generic Masculine: Evaluating Gender Neutrality of German AI-Generated Texts

Jasmin Schmidt and Claudia Hess
IU International University of Applied Sciences, Erfurt, Germany

jasmin.schmidt56@iu-study.org
claudia.hess@iu.org

Abstract: When using large language models (LLMs)—artificial intelligence (AI) systems trained to generate and interpret human language—gender-specific biases in AI-generated text represent a key challenge. Particularly in grammatically gendered languages such as German, this often results in texts outputs using the so-called generic masculine as an allegedly gender neutral default form. Consequently, generated text is neither neutral nor inclusive, and gender stereotypes are perpetuated. Organizations seeking to offer LLM-based systems that generate inclusive language by default typically rely on system prompts or specific model configurations to steer the model's responses. However, for German, methods for automatically, systematically, and objectively assessing whether such approaches enhance gender neutrality and inclusivity remain limited and underexplored. To address this gap, a framework was developed that applies the concept of LLM-as-a-judge. This approach involves using an LLM to systematically evaluate the outputs of another, thereby enabling automated and replicable assessments of features such as gender neutrality and inclusivity. The paper presents the development and evaluation of a prototype of this framework, designed specifically for German, following a Design Science Research approach. Using the framework, the effectiveness of configurations or system prompts can be evaluated. To enable this in a systematic and replicable manner, a catalogue of 150 prompts in German was developed, adapting and extending approaches from other languages. The outputs generated by an LLM in response to these prompts are then assessed by an evaluation module: Linguistic analysis identifies gendered forms and grammatical structures, while scoring metrics quantify the degree of gender neutrality and inclusivity. To demonstrate the framework, it was applied in several test runs using an iteratively developed system prompt designed to elicit gender neutral responses. The resulting metrics allowed assessment of whether a given prompt effectively enhances the neutrality of generated outputs and reduces gender-specific bias. Potential applications of the framework in organisational settings, as well as its relevance for the development of responsible AI systems, are outlined.

Keywords: Artificial intelligence, Gender neutrality, Large language models, Fairness, Ethics, Responsible AI

1. Introduction

Large language models (LLMs) are increasingly being used to generate texts for communication, education and the public sector. They learn statistical patterns from their training corpora, which include prevailing linguistic conventions and social biases (Zhang, 2024). In grammatically gendered languages like German, this includes the supposedly neutral generic masculine. As a result, LLMs often generate the generic masculine, thereby reproducing and reinforcing the underlying gender bias found in the data. Consequently, women and non-binary people are referenced less and mainly through masculine-marked expressions, reducing linguistic visibility and reinforcing stereotypes (Misersky et al., 2019; Sczesny et al., 2016). In light of these effects, gender neutrality is understood linguistically: personal expressions are formulated to avoid morphological male or female markers wherever possible. In German, this mainly concerns personal names, pronouns and gender markers, and differs from broader social or identity-related concepts of gender.

Responsible use of LLMs requires promoting gender-neutral or inclusive language directly at the system level. In practice, two approaches are common: system prompts, i.e., persistent, top-level instructions setting the LLM's role, tone, and constraints (e.g., "avoid the generic masculine and consistently choose inclusive expressions"), and model configurations, i.e., the settings and policies that steer generation (e.g. preferred or blocked terms). Evaluating their effectiveness requires automatic, objective, reproducible measurement of gender neutrality (Doyen & Todirascu, 2025). Although benchmarks and classifiers assess gender bias in German and other languages, none measure grammatical gender neutrality in open-ended German LLM outputs, a gap repeatedly noted in recent work (Doyen & Todirascu, 2025). Existing German benchmarks target gender bias in model predictions but neither evaluate the grammatical gender neutrality of generated text nor provide fully automated, reproducible procedures (Kraft *et al.*, 2022; Satheesh *et al.*, 2025). English approaches don't transfer because grammatical gender in German is morphologically encoded and affects core linguistic structures (Diewald and Nübling, 2022). Consequently, no systematic method exists to evaluate or optimise gender neutrality in German LLM outputs. The research presented addresses this gap by introducing a framework for the automated measurement of gender neutrality in German-language LLM outputs. With this aim in mind, the paper is organised as follows. Section 2 outlines the methodology. Section 3 reviews the linguistic and empirical

background. Building on these foundations, the framework is presented and evaluated in Section 4. Then, section 5 discusses results and limitations. Section 6 outlines potential applications, and Section 7 concludes.

2. Methodology

As outlined in the introduction, there is a need for approaches that can evaluate gender neutrality in German-language LLM outputs automatically and in a way that reflects the specific linguistic properties of German. Given that LLMs themselves can serve as evaluative tools, this leads to the central research question:

How can LLMs be used to automatically and objectively evaluate the gender neutrality of German text outputs using suitable metrics?

Addressing this research question requires the development of a suitable artifact. Accordingly, this research follows the principles of design science research (DSR) according to Österle et al. (2010), in which the design, justification, and evaluation of an artifact are integral to the research process. This approach is appropriate because no established solution exists for this task. Based on this rationale, the development followed the phases of the DSR cycle.

In the analysis phase, the problem space and the need for a German-specific evaluation approach were clarified. This involved reviewing existing frameworks and benchmarks that analyse gendered language in LLM outputs and clarifying how gender neutrality should be conceptualised for German. The key findings are presented in Section 3. Based on potential application scenarios in the researchers' organisation, relevant stakeholders were identified and prioritised, and their needs were translated into functional and non-functional software requirements for the artifact. These requirements were made testable through general acceptance criteria and requirements-engineering techniques. While this software-engineering process is not the focus of the paper, it ensured that the artifact was developed on a sound and transparent methodological basis. In the design phase, the conceptual structure of the artifact—later named GenScore-DE—was developed and the artifact was implemented in two iterations as a working prototype. Its architecture and the different components will be presented in detail in Section 4. In this phase, the identified requirements were translated into a coherent framework suited to German linguistic patterns. In the evaluation phase, the artifact was then reviewed against the defined requirements and its usefulness validated through end-to-end runs, including test-retest checks and documentation of all results. In this way, the framework is developed as a proof of concept that is both theoretically sound and practically applicable. The subsequent diffusion of the artifact included making the framework available on GitHub and presenting it in various formats.

3. Gender Bias in German AI-generated Text

The theoretical foundation of the framework combines linguistic insights into how German encodes gender, empirical findings from research on gender bias in language use and in LLMs, and methods for automatically analysing and evaluating text.

3.1 Linguistic Structures That Promote Bias

The German language categorises gender on three linguistic levels:

- grammatical gender (Genus) as a formal, non-biological characteristic of nouns,
- biological gender (Sexus) as a biologically determined characteristic independent of grammar, and
- social gender (Gender), which expresses roles, identities, and expectations (Diewald, 2018; Diewald & Nübling, 2022).

While gender is grammatically marked by articles (der, die, das) and suffixes, sex and gender refer to the semantic level (Diewald and Nübling, 2022). Since a noun can be grammatically masculine without referring to a male person, this distinction makes gender neutrality in German particularly complex (Diewald, 2018; Diewald & Nübling, 2022). Due to this interaction, formally masculine forms may be intended as generic, yet they are not empirically interpreted as neutral. Although these categories differ, grammatical gender cues (articles, endings) influence how personal expressions and pronouns are interpreted in context (Stahlberg and Sczesny, 2001; Braun, Sczesny and Stahlberg, 2005). This interplay is crucial for assessing gender neutrality in LLM outputs.

In German, gender is expressed through several linguistic markers. This study distinguishes between the following forms: neutral forms without gender markers (e.g., Lehrkraft [teacher], Mitarbeitende [employee]), inclusive forms that explicitly refer to multiple genders (e.g., Lehrer:innen, Lehrer*innen), binary forms that are

morphologically masculine or feminine (e.g., Lehrer, Lehrerin), and paired forms (e.g., Lehrer & Lehrerinnen) (Ivanov, Lange and Tiemeyer, 2018; Diewald and Nübling, 2022).

These markers reveal how gender is encoded and enable systematic, sound evaluation of gender neutrality in model outputs. A further mechanism particularly relevant in German is the generic masculine: masculine nouns function as unmarked generic forms, regardless of their actual gender. However, they are read as male in practice, reducing linguistic visibility of women and non-binary people and reinforcing gendered perception patterns (Stahlberg and Sczesny, 2001; Braun, Sczesny and Stahlberg, 2005; Misersky, Majid and Snijders, 2019).

3.2 Reproduction of Gender Biases by LLMs

LLMs are trained on extensive text collections that reflect social norms, historical perspectives and traditional role models. As a result, LLMs often reproduce existing stereotypes and imbalance (Gonen & Goldberg, 2019; Zhang, 2024). For example, technical or analytical roles are more often associated with men, whereas caregiving or communicative roles tend to be assigned to females. These biases are also evident in competency assessments which often favour male personas (Bartl & Leavy, 2024; Chen et al., 2025; Kotek et al., 2023; Ranjan et al., 2024). In German, this effect arises from the grammatical system, and in particular from the generic masculine (Stahlberg and Sczesny, 2001; Braun, Sczesny and Stahlberg, 2005; Misersky, Majid and Snijders, 2019). As a result, LLMs generate male designations by default, such as the developer, the doctor, the teacher or the employee, even though neutral alternatives would be possible (Doyen & Todirascu, 2025). This can set up a self-reinforcing pattern: because the generic masculine is prevalent in German source corpora and common in everyday usage, users often prompt with masculine forms, and LLMs mirror that style, making inclusive variants rarer and therefore less likely to be produced. Such AI-generated text may enter future training data, for example via AI-generated text published on websites or logged interactions used for fine-tuning, perpetuating the dominance of the masculine form in German (Drake, 2025).

3.3 Evaluation of Gender Neutrality

Gender-neutral language aims to avoid privileging any specific gender through grammatical marking and to minimise gender-specific linguistic forms. Several international benchmarks can be used to analyse gender-specific language patterns, including OCCUGENDER (Chen *et al.*, 2025), GenderPair (Tang *et al.*, 2024) and GG-BBQ (Satheesh *et al.*, 2025). They provide structured test scenarios that reveal how gender is expressed in model outputs, and which linguistic phenomena must be considered in systematic evaluations. Each benchmark has a distinct focus: OCCUGENDER examines job and role descriptions, GenderPair targets stereotypical role attributions and contrasting pairs, and GG-BBQ addresses everyday scenarios and pronoun resolution. Although these benchmarks are designed for English or multilingual contexts, they offer useful methodological guidance. A German evaluation framework should capture the same core dimensions. Moreover, automatic evaluation of gender neutrality requires clear, operational criteria defining what counts as a gendered expression. In German, gender is often marked on the surface level of words as discussed in Section 3.1., and these cues (masculine/feminine morphology, typographic inclusives, and neutral forms) must be detected consistently and independently of any specific LLM. These aspects form core requirements for the framework developed.

4. The GenScore-DE-Framework

The artifact produced through the DSR methodology is a framework, which enables automated and objective evaluation of gender neutrality in German-language LLM outputs. The name GenScore-DE reflects its core intention: ‘Gen’ stands for gender neutrality, ‘Score’ for the quantitative evaluation of linguistic characteristics, and ‘DE’ for the fact that the framework was developed specifically for the German language.

4.1 Overview of the Framework

To illustrate how the framework operates, a typical application scenario is considered. Suppose the aim is to introduce a system prompt that increases gender neutrality in an LLM’s outputs. First, a baseline is established without any system prompt, forming the reference for later comparison. In a second run, the system prompt is added, and the results of both runs are compared. Each run follows the same procedure: all prompts from the standardised prompt catalogue are sent to the LLM, which generates an output for each prompt. These outputs are then classified by the evaluation module, and scores are computed for defined metrics. As both runs are processed in exactly the same way, it can be directly assessed whether and to what extent the optimised system prompt increases the gender neutrality of the LLM outputs. This results in a standardised and reproducible evaluation process, which is depicted in Figure 1.

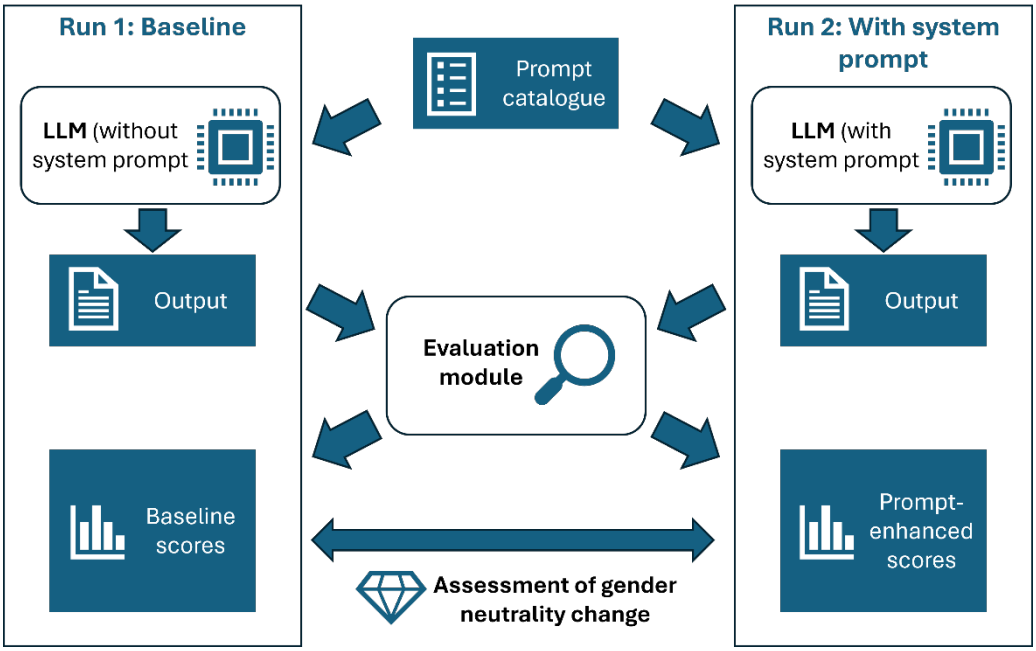

Figure 1: Overview of the framework

The GenScore-DE framework is designed in a modular way. The current version of the framework comprises the following implemented components: the curated, standardised prompt catalogue, the evaluation module that classifies the LLM outputs, an orchestration layer that runs the evaluation steps in a consistent order, and structured data storage to keep all results transparent and traceable. Details on the prompt catalogue and evaluation module are provided in the next sections. The assessment of the evaluation results still has to be carried out manually, but it could be automated in a future iteration of the DSR methodology. The modular structure ensures flexibility for future adjustments, allows the use of different LLMs, and keeps every step well documented. In this way, the framework meets scientific standards of validity and reliability, as well as practical requirements such as extensibility and reusability.

GenScore-DE is available as open-source resources on GitHub under the MIT licence: https://github.com/Lacotto/GenScore-DE. The GenScore-DE framework is implemented as a functional prototype. The prompt catalogue, evaluation module, orchestration logic and data storage are implemented in Python with fixed decision rules in prompts and reproducible execution. All components are available in the associated GitHub repository (including scripts, evaluation logic, and sample runs), demonstrating that the framework is fully functional and not just conceptually defined.

4.2 Standardised Prompt Catalogue

The prompt catalogue provides a uniform test basis for all evaluations. It builds on insights from existing studies of gender expression in LLM outputs, as discussed in Section 3.3 and covers 50 situations across 5 categories that represent typical contexts in which gender neutrality plays a role. Including prompts from different domains is essential because varying scenarios and framings can activate different linguistic structures. Based on the international benchmarks, the following categories were defined:

- Work context (cf. OCCUGENDER)
- Professional roles (cf. OCCUGENDER)
- People and characteristics (cf. GenderPair)
- Everyday situations (cf. GG-BBQ)
- Personas and roles (This category captures prompts that explicitly assign the model an identity or interaction role (e.g., “You are a student …”, “As a teacher …”), which can influence whether gender-marked role nouns appear in the output.)

Note that the category ‘Personas and Roles’ does not originate from an existing benchmark, but from general prompt design. This category was introduced specifically in GenScore-DE because it is close to everyday life and thus reflects more realistic usage behaviour.

The prompt catalogue builds on prior analyses of gender expression in LLM outputs, which informed the selection of linguistic structures. For each of the 50 situations (10 per category), masculine, feminine, and neutral variants are formulated, resulting in 150 prompts that are submitted to the LLM that is going to be tested for gender neutrality. This approach was conceptually derived from German language theory, not from benchmarks. It ensures that gender marking can be experimentally isolated. Having both gendered and neutral prompts ensures that the model can be tested across different linguistic conditions specific for the German language as Section 4.3.2 will show. Although GenScore-DE is designed for German, a short comparison with English shows how it works. In English, there is no gender marker for nouns, yet gender bias is evident in job titles such as actor vs. actress (Kotek, Dockum and Sun, 2023). A comparable prompt from the prompt catalogue in the 'Professional Roles' category would be:

- 'Describe the daily tasks of an actor.' (male)
- 'Describe the daily tasks of an actress.' (female)
- 'Describe the daily tasks of a performer.' (neutral)

This comparison shows why German needs a more comprehensive catalogue. In English, gender is mainly marked by individual words, whereas in German it is encoded through grammatical patterns like word endings. Because such gender marking is the norm rather than the exception (Diewald and Nübling, 2022), broad coverage of relevant linguistic structures is essential, and systematic test coverage is particularly important.

4.3 The Evaluation Module

The GenScore-DE evaluation module works in two steps. First, linguistic classification is performed using an LLM-as-a-Judge procedure. In the second step, metrics for gender neutrality are calculated.

4.3.1 LLM-as-a-judge

LLMs are increasingly used not only to generate text but also to perform evaluation tasks. This approach is referred to as LLM-as-a-Judge. It uses an LLM as an evaluator that classifies outputs based on predefined criteria, providing a scalable and automated alternative to manual assessments while retaining semantic flexibility (Gu *et al.*, 2025). As research has shown that LLMs can achieve high consistency with human judgements in standardised linguistic evaluation tasks (Li et al., 2025), this approach is adopted in GenScore-DE to classify gender markers in German LLM outputs.

The Judge-LLM processes the 150 text snippets generated as response to the standardised prompt catalogue. Based on explicit linguistic rules encoded in the evaluation prompt, it assigns each snippet to one of four categories, masculine, feminine, neutral and mixed. The rules reflect the German linguistic structures described in Section 3.1: Morphological masculine or feminine forms are classified as such, gender-unmarked forms such as 'Studierende' [student] or 'Lehrkraft' [teacher] are categorised as neutral, and typographical variants and paired forms are classified as mixed but not neutral. To ensure consistent behaviour, the evaluation prompt sets fixed rules for decision-making that the Judge-LLM must follow. This rule-based structure supports reproducible classifications and reduces variation between runs.

4.3.2 Metrics for gender neutrality

Two metrics were developed specifically for assessing gender neutrality. Because each prompt in the catalogue exists in masculine, feminine, and neutral variants, the following two dimensions can be evaluated independently.

- Gender Prompt Match (GPM) measures how well the output follows the gender specification given in the prompt—namely, whether the model uses the gendered form requested. Therefore, the metric compares the gender indicated in the prompt (masculine, feminine, or neutral) with the gender category assigned to the output by the Judge-LLM.
- Gender Neutrality Score (GNS) measures whether the output adopts grammatically gender-neutral forms (neutral) or explicitly gender-inclusive strategies (inclusive) regardless of the prompt's instruction. It reflects the grammatical gender neutrality category assigned by the LLM-as-a-Judge component. The GNS acts as an aggregate metric, it records both neutrality (using invariant forms like 'Lehrkraft') and inclusivity (using typographic markers like 'Lehrer*innen'). While theoretically distinct, both strategies serve the common goal of reducing masculine defaults in the framework's scoring

Using these metrics, changes in output behaviour—whether between runs with and without a system prompt or between different system prompts—can be recorded objectively and translated into quantitative indicators.

4.4 Evaluation

To evaluate the functionality of the GenScore-DE framework, test runs were conducted. Two open-source LLMs were used: a Mistral model (mistral:7b) for text generation and a Meta model (llama-3.3-70b-instruct) as the LLM-as-a-Judge, combined with varied system prompts. All GenScore-DE outputs, i.e., generated texts, results of the LLM-as-a-judge and calculated metrics, were then systematically analysed to assess the framework's ability to capture differences in gender neutrality. Across all test runs, the GenScore-DE framework revealed clear differences between the baseline outputs and those generated with an optimised system prompt.

The following end-to-end run serves as an illustrative example. Two conditions were tested: a baseline without a system prompt and an optimised condition using a gender-neutral system prompt instructing the model to avoid the generic masculine, favour gender-neutral German structures, specifically substantivized participles (e.g., 'Mitarbeitende' or 'Studierende'). Unlike the generic masculine, these plural forms are grammatically neutral in German as they are derived from the present participle and do not carry a gendered suffix. Additionally, the prompt required balanced naming (i.e., when first names are required, providing an equal number of female and male names). All 150 results were processed without technical errors, automatically classified and manually checked for plausibility. Results show that the model without a system prompt predominantly used gender-specific formulations. The gender neutrality score (GNS) was only 12%, while gender prompt consistency (GPM) was comparatively high at 47%. With the system prompt, the output behaviour changed markedly. The GNS increased to 42%, while the GPM fell to 24%, as expected. The frequency analysis confirmed a substantial rise of neutral and mixed forms and a corresponding decrease in generic masculine formulations. A test-retest check produced 99% identical classifications, confirming the reliability of the evaluation module.

5. Discussion

Applying GenScore-DE across several test runs has revealed consistent patterns. In their baseline state, the tested LLMs frequently produced gender-specific forms, especially the generic masculine, a pattern already well documented in previous research (Kotek et al., 2023; Ranjan et al., 2024). At the same time, neutral forms remain underrepresented. Consequently, LLMs often reproduce or amplify gender-typical patterns present in their training data (Gonen & Goldberg, 2019). Through systematic analysis of generated texts and their classifications, it can be analysed under which conditions LLMs choose gender-specific forms. This contributes to an objective discussion on gender-neutral language, whose effect on perception and inclusion has been empirically demonstrated (Stahlberg and Sczesny, 2001; Braun, Sczesny and Stahlberg, 2005; Sczesny, Formanowicz and Moser, 2016).

While the baseline behaviour reveals systematic patterns, the analysis also shows that these patterns can be modified. A system prompt requiring gender neutrality significantly increased the number of neutral or mixed outputs. As neutrality increased, compliance with prompts decreased, illustrating the expected trade-off: greater neutrality typically reduces adherence to explicitly gendered instructions.

The test–retest comparison conducted showed a high level of consistency in the classifications. This demonstrates the robustness of separating the LLM-as-a-Judge classification from the rule-based scoring. It also reduces dependencies on the specific evaluation model. Overall, the framework produced a stable, data-based picture of the distribution of masculine, feminine, neutral, and mixed forms in the LLM outputs.

The framework has certain limitations. It deliberately remains analytical and focuses on superficial linguistic gender markers rather than semantic or implicit stereotypes. It evaluates the degree of grammatical neutrality in the outputs, not whether a particular variant is socially 'correct'. In its current form, the framework is a working prototype that requires further development and broader testing before large-scale application. Moreover, the framework is designed specifically for German; applying it to another language would require adapting the evaluation logic and revising the prompt catalogue accordingly.

6. Implications for Practice

GenScore-DE can be used in various application contexts in which AI-supported text generation influences linguistic representation and is therefore relevant for many organisations, both in internal communication and in customer-facing applications such as chatbots, assistance systems, automated text generators, and knowledge bases. Many organisations for which linguistic consistency, fairness and compliance are key criteria already follow internal guidelines for gender neutral or inclusive language, yet they lack appropriate tools to verify whether LLM-based systems adhere to these standards. GenScore-DE closes this gap by enabling

systematic comparisons of different model configurations or system prompts. It helps identify configurations that promote gender neutral output without requiring changes to, or retraining of the underlying model. In this way, the framework supports the monitoring and refinement of language standards as well as fairness and compliance requirements. Beyond supporting organisational language practices, these functions highlight the broader value of the framework. The framework contributes to responsible AI development because it shows developers and AI specialists where and how an LLM uses gendered language, turning what was previously based on personal impressions or anecdotal evidence into clear, analysable data. This makes it possible to adjust system behaviour where necessary.

7. Conclusion

This paper demonstrates that gender-neutral language generation in German-language LLM outputs can be measured objectively, systematically and reproducibly. GenScore-DE provides an automated framework for evaluating linguistic structures such as gender markers, personal pronouns and neutral formulations using a curated prompt catalogue and a dedicated evaluation component. The application demonstrates the strong impact of system prompts on the gender neutrality. In test runs, the use of gender-neutrality-oriented system prompts markedly increased neutral and mixed forms while reducing the dominance of the generic masculine. This demonstrates that gender neutrality in LLMs is both measurable and steerable. These findings have important implications for research and practice. GenScore-DE enables data-based decisions about language settings and supports the evaluation of LLMs by making gender-related effects visible. The framework is analytical rather than normative and offers a robust basis for discussing gender-equitable AI development and the effects of linguistic interventions. Future work should examine its application to other LLMs and across different use cases. Testing the framework in public- and private-sector settings could further broaden its relevance and practical value.

Ethics Declaration: This study does not involve human participants or any personal or sensitive data. All analyses were performed exclusively on results generated by publicly available large language models using synthetic test prompts. No identifiable data was collected or processed, and the study design does not pose any ethical risks. As the work only examines the linguistic properties of AI-generated texts, no ethical approval was required.

AI Declaration: AI tools (ChatGPT-5, DeepL) were used for language editing and improving clarity. The authors confirm that all intellectual content, interpretation, and conclusions were developed by the authors and that AI outputs were reviewed and edited by the authors for accuracy and appropriateness. AI tools (ChatGPT-5, PerplexityAI) were also used to expand literature search terms, suggest related keywords, and identify candidate publications. All searches were validated through academic databases, all inclusion/exclusion decisions were made by the authors. AI was neither used for statistical analysis nor figure/table generation.

References

Bartl, M. and Leavy, S. (2024) "From Showgirls to Performers: Fine-tuning with Gender-inclusive Language for Bias Reduction in LLMs", *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, Bangkok, Thailand: Association for Computational Linguistics, pp. 280–294. doi:10.18653/v1/2024.gebnlp-1.18.

Braun, F., Sczesny, S. and Stahlberg, D. (2005) "Cognitive effects of masculine generics in German: An overview of empirical findings", *Communications*, 30(1), pp. 1–21. doi:10.1515/comm.2005.30.1.1.

Chen, Y., Mattern, J., Mihalcea, R. and Jin, Z. (2024) "Causally testing gender bias in LLMs: A case study on occupational bias", *arXiv preprint*, doi:10.48550/arXiv.2212.10678.

Diewald, G. (2018) "Zur Diskussion: Geschlechtergerechte Sprache als Thema der germanistischen Linguistik – exemplarisch exerziert am Streit um das sogenannte generische Maskulinum", *Zeitschrift für germanistische Linguistik*, 46(2), pp. 283–299. doi:10.1515/zgl-2018-0016.

Diewald, G. and Nübling, D. (eds.) (2022) Genus – Sexus – Gender, De Gruyter, Berlin/Boston. doi:10.1515/9783110746396.

Doyen, E. and Todirascu, A. (2025) "Man made language models? Evaluating LLMs' perpetuation of masculine generics bias", *arXiv preprint*, doi:10.48550/arXiv.2502.10577.

Drake, S. (2025) "Neural howlround in large language models: A self-reinforcing bias phenomenon, and a dynamic attenuation solution", *arXiv preprint*, doi:10.48550/arXiv.2504.07992.

Gonen, H. and Goldberg, Y. (2019) "Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them", *Proceedings of NAACL-HLT 2019*, Minneapolis, Minnesota: Association for Computational Linguistics, pp. 609–614.

Gu, J. et al. (2025) "A survey on LLM-as-a-Judge", *arXiv preprint*, doi:10.48550/arXiv.2411.15594.

Ivanov, C., Lange, M.B. and Tiemeyer, T. (2018) "Geschlechtergerechte Personenbezeichnungen in deutscher Wissenschaftssprache", *Suvremena lingvistika*, 44(86), pp. 261–290. doi:10.22210/suvlin.2018.086.05.

Kotek, H., Dockum, R. and Sun, D. (2023) "Gender bias and stereotypes in Large Language Models", Proceedings of The ACM Collective Intelligence Conference (CI '23), Delft, Netherlands: ACM, pp. 12–24. doi:10.1145/3582269.3615599.

Kraft, A. et al. (2022) "Measuring gender bias in German language generation", in Demmler, D. et al. (eds.) *INFORMATIK 2022*, Bonn: Gesellschaft für Informatik, pp. 1257–1274. doi:10.18420/inf2022_108.

Li, Q. et al. (2025) "Evaluating scoring bias in LLM-as-a-Judge", *arXiv preprint*, doi:10.48550/arXiv.2506.22316.

Misersky, J., Majid, A. and Snijders, T.M. (2019) "Grammatical Gender in German Influences How Role-Nouns Are Interpreted: Evidence from ERPs", *Discourse Processes*, 56(8), pp. 643–654. doi:10.1080/0163853X.2018.1541382.

Österle, H. et al. (2010) "Memorandum zur gestaltungsorientierten Wirtschaftsinformatik", *Schmalenbachs Zeitschrift für betriebswirtschaftliche Forschung*, 62(6), pp. 664–672. doi:10.1007/bf03372838.

Ranjan, R., Gupta, S. and Singh, S.N. (2024) "Gender Biases in LLMs: Higher intelligence in LLM does not necessarily solve gender bias and stereotyping", *arXiv preprint*, doi:10.48550/ARXIV.2409.19959.

Satheesh, S. et al. (2025) "GG-BBQ: German gender bias benchmark for question answering", *arXiv preprint*, doi:10.48550/arXiv.2507.16410.

Sczesny, S., Formanowicz, M. and Moser, F. (2016) "Can Gender-Fair Language Reduce Gender Stereotyping and Discrimination?", *Frontiers in Psychology*, 7, doi:10.3389/fpsyg.2016.00025.

Stahlberg, D. and Sczesny, S. (2001) "Effekte des generischen Maskulinums und alternativer Sprachformen auf den gedanklichen Einbezug von Frauen", *Psychologische Rundschau*, 52(3), pp. 131–140. doi:10.1026//0033-3042.52.3.131.

Tang, K. et al. (2024) "GenderCARE: A Comprehensive Framework for Assessing and Reducing Gender Bias in Large Language Models", *Proceedings of the 2024 ACM SIGSAC Conference*, Salt Lake City, UT: ACM, pp. 1196–1210. doi:10.1145/3658644.3670284.

Zhang, C. (2024) "Exploration, detection, and mitigation: Unveiling gender bias in NLP", *Applied and Computational Engineering*, 52(1), pp. 62–68. doi:10.54254/2755-2721/52/20241234.