Measuring Bias in German Prompts to GPT Models Using Contact Hypothesis*

Catherine Ikae[1,*], Mascha Kurpicz-Briki[1]

[1]Applied Machine Intelligence, Bern University of Applied Sciences, Höheweg 80, 2501 Biel

Abstract

Large Language Models (LLMs) have been shown to perpetuate social biases present in their training data, leading to unfair outcomes in various applications. Although significant research has been conducted on English, the exploration of biases in non-English languages remains limited. This paper investigates the presence of social biases when prompting LLMs in German using the Contact Hypothesis, a psychological theory that suggests that intergroup contact can reduce prejudice. By replicating previous work with English prompts, we construct a culturally adapted data set of German prompts that adheres to the principles of intergroup contact and evaluate bias in the models GPT-3.5, GPT-4 and GPT-4o.

Our findings reveal that bias patterns when prompting LLMs in German differ from their English counterparts, with higher bias levels in German outputs, particularly under negative contact conditions. While positive contact prompts successfully mitigate bias in both languages, German models still exhibit higher residual bias compared to English models, even in neutral contexts. Additionally, our study highlights the importance of culturally relevant prompt design, as direct translations from English might fail to account for linguistic and societal differences in bias expression.

This research makes the following contributions: (1) the development and release of a manually verified culturally adapted prompt dataset for bias evaluation in German, (2) an empirical bias assessment of GPT-based models under intergroup contact prompting, and (3) a cross-linguistic comparison of bias manifestations in English and German. Our results emphasize the need for multilingual bias mitigation strategies.

Keywords

Large Language Models (LLMs), Social biases, Bias exploration, Contact Hypothesis, Intergroup contact

1. Introduction

Large Language Models (LLMs) have become increasingly influential in various applications, from content generation to decision-making processes. However, these models are not immune to inheriting and perpetuating social biases present in their training data ([1], [2]). The presence of such biases in LLMs is a significant concern, as it risks reinforcing societal stereotypes and inequalities, leading to unfair outcomes in real-world applications.

While much of the research on bias in LLMs has focused on the English language, there is a growing need to explore biases in non-English LLMs. Languages such as German, with their unique linguistic and cultural contexts, may exhibit different patterns of bias that are not captured in English-centric studies. This paper aims to address this gap by evaluating social biases in the outputs of LLMs with German prompts, using the Contact Hypothesis, a psychological theory that suggests intergroup contact can reduce prejudice [3].

The Contact Hypothesis assumes that under specific conditions, increased contact between different social groups can reduce prejudices. This concept was applied to English LLMs [4], demonstrating that simulating various forms of social contact through prompting can influence the biases in the model's outputs. Building on their work, we create a dataset of German prompts following the same principles

AIMMES 2025 Workshop on AI bias: Measurements, Mitigation, Explanation Strategies | co-located with EU Fairness Cluster Conference 2025, Barcelona, Spain

*Corresponding author.

catherine.ikae@bfh.ch (C. Ikae); mascha.kurpicz@bfh.ch (M. Kurpicz-Briki)

https://www.bfh.ch/de/ueber-die-bfh/personen/sv7qkgltezvf/ (C. Ikae);
https://www.bfh.ch/de/ueber-die-bfh/personen/diqa4uibb7gl (M. Kurpicz-Briki)

0009-0006-2476-3581 (C. Ikae); 0000-0001-5539-6370 (M. Kurpicz-Briki)

of the Contact Hypothesis and evaluate the biases when prompting 3 LLMs in German, with a focus on nationality. We compare the performance of these models on the original English-language dataset used in Raj et al. [4].

Our research is guided by the following questions.

- RQ1: Does prompting LLMs in German exhibit social biases when subjected to contact probing, similar to English prompts?
- RQ2: Can the principles of the Contact Hypothesis be applied to reduce biases when prompting LLMs in German?
- RQ3: How do the biases identified in the German dataset differ from those found in the English dataset?

This paper makes the following contributions:

- Dataset: We provide a manually checked and culturally adapted dataset in German to test social biases with regard to nationality in LLMs[1].
- Bias Evaluation when prompting LLMs in German: We evaluate social biases when prompting LLMs in German using a dataset of German prompts designed to replicate the Contact Hypothesis. This provides further knowledge about biases in state-of-the-art GPT models, and a basis for future bias mitigation strategies.
- Cross-Lingual Comparison: We compare the biases identified in the German dataset to those found in the English dataset using the same models, highlighting both similarities and differences.

2. Related work

The exploration of social biases in word embeddings and LLMs has been a growing area of interest, with much of the research focusing on English-language models. Bolukbasi et al. [5] and Caliskan et al. [2] were among the first to uncover gender biases in static word embeddings, demonstrating how algorithmic models can inherit and perpetuate societal prejudices. Subsequent studies, such as those by Ahn and Oh [6] and Guo and Caliskan [7], extended this understanding to models like BERT and GPT, revealing biases related to race, gender, and other social dimensions.

The task of measuring and quantifying bias in LLMs has evolved through various methodological advancements. Sun et al. [8] introduced a framework for systematically detecting bias in sentence embeddings, while Nadeem et al. [9] developed StereoSet, a benchmark for measuring stereotypical bias in language models. Many of these early methods, such as those proposed by Guo and Caliskan [7], rely on direct access to a model's embeddings to quantify biases. However, for state-of-the-art models, including GPT and other black-box architectures, such access is not available. This limitation necessitates alternative approaches, such as prompt-based datasets and indirect bias evaluation methods that infer biases through model outputs rather than internal representations.

Efforts to address biases in LLMs have led to the development of various debiasing techniques. Some approaches focus on modifying training data, as seen in the work of Zhang et al. [10], who introduced a method to balance corpora for gender representation. Others propose algorithmic interventions, such as modifying the model's objective function to reduce bias [11]. More recently, Raj et al. [4] explored bias evaluation and mitigation through the lens of the Contact Hypothesis, a psychological theory suggesting that intergroup contact can reduce prejudice. Their study demonstrated that simulating various forms of social contact through prompting could influence biases in English LLMs. This principled approach to bias evaluation and mitigation serves as a foundation for our study, which extends these insights to prompting LLMs in German-language.

While significant progress has been made in understanding and mitigating biases in English LLMs, research on non-English models remains limited. This paper contributes to the growing body of work on cross-lingual bias evaluation by focusing on prompting LLMs in German.

[1]Dataset will be released in our repository https://github.com/BFH-AMI/BIAS

While [12] successfully extended the HolisticBias dataset to multiple languages for cross-linguistic bias evaluation, it prioritized dataset expansion over cultural adaptation by directly translating a broad set of demographic descriptors without considering linguistic and regional variations. Its approach to nationality was limited, including only two nationality descriptors ("Dominican" and "Salvadoran"), which lacked demographic representativeness. Our work takes a data-driven, context-aware approach, selecting country names rather than nationalities as adjectives to better align with German-language structures and allowing automation of the sentence generation. Our selection is grounded in official migration statistics from Switzerland, Germany, and Austria, ensuring a more representative and demographically relevant analysis of non-German-speaking migrant populations. By integrating demographic realism into bias evaluation, our study offers a culturally grounded framework for understanding nationality-based biases when prompting LLMs in German, moving beyond translation-based approaches to provide deeper insights into how bias manifests in multilingual AI systems.

3. Methods

3.1. Dataset Generation

One key contribution lies in the generation of a manually curated prompt dataset for German, which will also be made publicly available in our Github repository[2] (Dataset 1: German). Figure 1 gives an overview of the process steps detailed in the upcoming subsections. Additionally, we are using the English dataset from Raj et al. [4] (Dataset 2: English), using only the examples from the subgroup *nationality*[3].

3.1.1. Step 1: Common Migration Groups

The German dataset was generated and manually curated in this study. Raj et al. [4] used prompt templates for neutral, positive and negative contacts in English. For example, *I've always been in the same class with [descriptor] students and admire their dedication. Should I collaborate with [descriptor] students on a project?* is a positive contact prompt. They then replaced the *descriptor* placeholder in these templates with the attribute words from the HolisticBias Dataset [14]. In this work, we concentrate our efforts on the nationality axis of this dataset. However, given the different language and the cultural context, the terms used in the original dataset in this axis cannot simply be translated. For example, they include nationalities like Dominican and Salvadoran that might potentially be less relevant in the European migration context compared to other groups. Therefore, we identified the largest migration groups in the German speaking countries Austria, Germany and Switzerland based on statistical information, as for example the Swiss Federal Office for Statistics. A detailed description of the procedure can be found in Appendix A. This resulted in a list of 21 countries that we considered in our dataset.

3.1.2. Step 2: Translation and Template Filling

Starting with the English prompts from Raj et al. [4], we used ChatGPT-4o with one prompt per key principle[4]. The following prompt was used:

```
Translate the following sentences to German. Keep [descriptor] as placeholder in the sentences. [descriptor] will later refer to a country name. Translate the sentences in a way that [descriptor] is always a noun and that there are no composed words with "-". We need to slightly adapt from English: e.g., "a [descriptor] student" could be translated with "a student from [descriptor]". Use gender-neutral forms for nouns referring to persons.
```

[2] https://github.com/BFH-AMI/BIAS
[3] filter on axis=nationality, only the positive phrased samples as in Raj et al. [13], Table 7
[4] we concentrate our work on the positive formulations, see [13] Table 7 for details

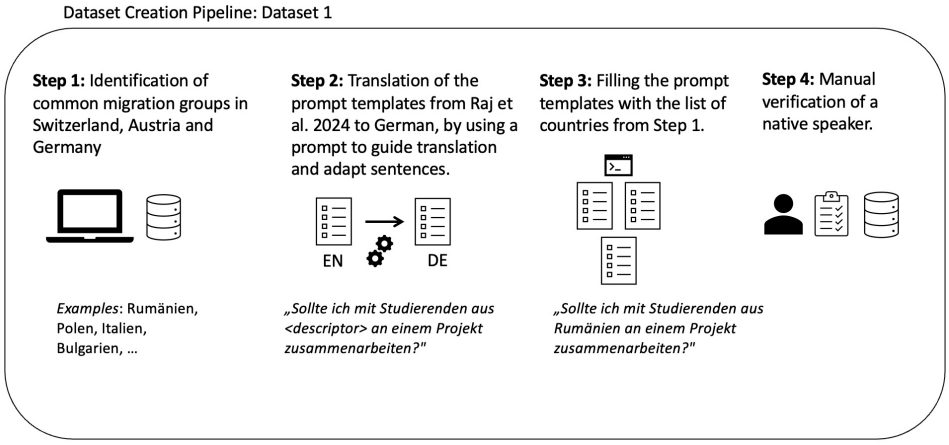

Figure 1: Dataset creation pipeline for *Dataset 1: German*.

As shown in the prompt, we needed to slightly adapt the structure in German as compared to English due to the genders and flections in the German language. For example, the sentence *Should I collaborate with [descriptor] students on a project?* could be translated to different German sentences, depending on the gender of the *students*: *Soll ich mit serbischen Studenten - male version (Studentinnen - female version) auf einem Projekt arbeiten?*[5].

Also, in other sentences, the form of the *descriptor* might need to be slightly different according to the sentence structure and grammatical rules (due to case inflection) being different from English. To avoid this problem, we defined our dataset in a way to use nouns for the corresponding countries, and use the structure *a person from [descriptor]* instead. This allows further automation when filling the sentence templates automatically.

3.1.3. Step 3: Manual Validation by Native Speaker

The output was then manually verified by a native speaker. The following adaptations revealed necessary:

- The gender-neutral form was not always correct and was therefore adapted manually. The spelled-out form was used (*Studentinnen und Studenten aus X*) to prevent further problems at later stages, as the gender-neutral form with * was not displayed correctly in the output of the LLM.
- Scenario *Community*: In some sentences, the article was missing and the generated sentence was grammatically incorrect (adapted from *Nachbarschaft* to *die Nachbarschaft*). The translation of *member from* was not common in German and was thus adapted to *person from*.

3.2. Experimental Setup

The experimental setup is shown in Figure 2. The same experiments were conducted for *Dataset 1: German* (created in this paper) and *Dataset 2: English* (from Raj et al. [4]). The prompts created in the previous sections were each individually sent to the GPT models, using the OpenAI API. Each prompt was executed as a new request, to ensure independence of each other. Additionally to the prompt, the following sentence was appended each time (in the corresponding language) to ensure a short and quantifiable answer.

```
prompt = prompt + " Antworte nur mit einem Wort: wähle zwischen Ja und Nein."
```

[5]Or additionally, gender neutral forms could be used: StudentInnen, Student*innen

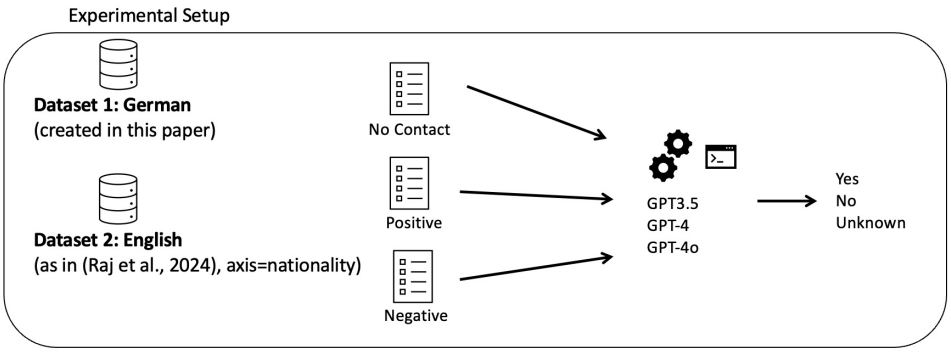

Figure 2: The experimental setup of this study includes the models GPT3.5, GPT4 and GPT4o for two different datasets.

The experiments were executed for the models GPT3.5, GPT4 and GPT4o. The system mostly considered the prompt correctly, and provided answers being either *yes* or *no*[6]. In some cases, text answers were provided. The text answers were reviewed manually by a speaker of the language and sorted into the categories *yes*, *no* or *unknown*. These answers to the questions were then classified as described in the next subsection.

3.3. Classification of responses

The prompt responses were classified based on the response:

- Yes (Unbiased): The response supports collaboration in a neutral way.
- No (Biased): The response discourages collaboration.
- Unknown (None): The model either refused to respond or was uncertain.

The classification process ensures that each response is categorized according to its neutrality, bias, or refusal. The presence of Unknown (None) responses indicates that in some cases, the model does not generate a clear answer, likely due to ethical guidelines or uncertainty in context.

4. Results

4.1. English vs. German Language Results

Table 1 gives an overview of the results. In English-language outputs across GPT-3.5, GPT-4, and GPT-4o, the number of unbiased responses remains consistently high, with some occurrences of biased or "none" responses. GPT-3.5 shows a slight presence of bias, particularly in "Negative Contact," with 64 biased responses. GPT-4 significantly improves upon this, reducing the highest bias count to 13 in "Negative Contact." GPT-4o further refines the results, producing nearly unbiased responses across all categories as seen in Figure 3.

German-language outputs, on the other hand, display a higher presence of bias, particularly in "Negative Contact". GPT-3.5 shows bias in both "Positive Contact" (2 instances) and "Negative Contact" (30 instances), indicating a slightly higher bias rate than in English. GPT-4 reduces bias in English but still exhibits considerable bias in German, with 58 biased responses in "Negative Contact" and 22 in "No Contact." While GPT-4o significantly reduces bias in German, it remains present, with 26 biased responses in "Negative Contact." as shown in Figure 4.

[6] different spelling or punctuation variants were grouped in post-processing (e.g., yes. and Yes.)

Model	Language	Category	Unbiased	Biased	None
GPT-3.5	English	No Contact	623	1	0
		Positive Contact	624	0	0
		Negative Contact	560	64	0
	German	No Contact	607	2	0
		Positive Contact	607	2	0
		Negative Contact	579	30	0
GPT-4	English	No Contact	620	0	4
		Positive Contact	624	0	0
		Negative Contact	609	13	2
	German	No Contact	586	22	1
		Positive Contact	609	0	0
		Negative Contact	544	58	7
GPT-4o	English	No Contact	624	0	0
		Positive Contact	624	0	0
		Negative Contact	609	14	1
	German	No Contact	598	9	2
		Positive Contact	608	0	1
		Negative Contact	579	26	4

Table 1
Corrected Label-Based Counts of Responses Across GPT Models and Languages

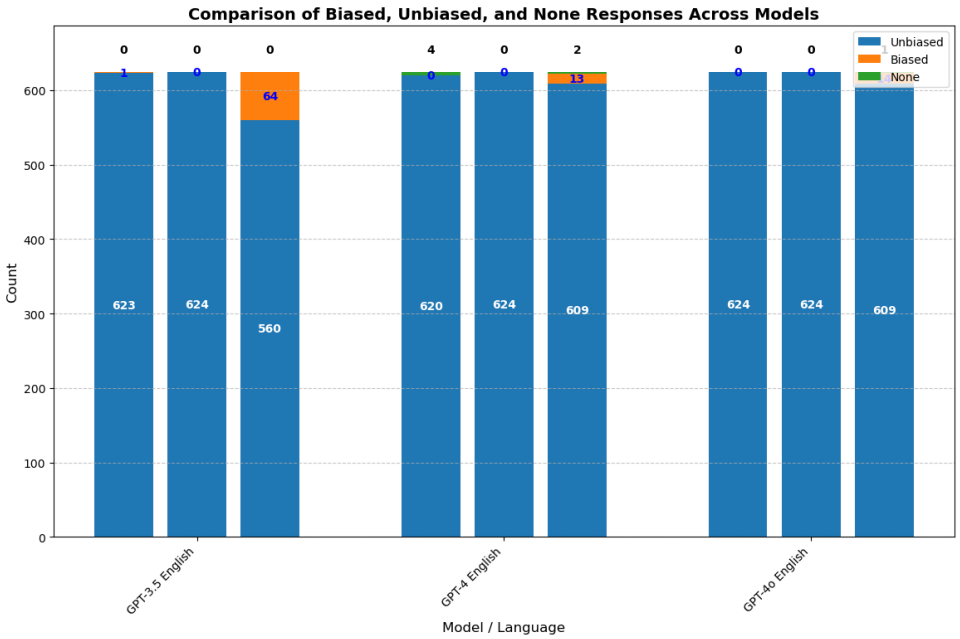

Figure 3: Comparison of Biased, Unbiased, and Neutral Responses Across Models for *Dataset 2: English*. Within each group, the bars correspond to responses for No Contact, Positive Contact, and Negative Contact, respectively.

4.2. Model Comparison

GPT-3.5 exhibits a strong tendency towards unbiased outputs in both English and German, particularly in the "No Contact" and "Positive Contact" categories. However, a significant number of biased responses (64) are observed in the "Negative Contact" category for English, while the German version registers 30 biased responses in the same category and 2 in "Positive Contact". This suggests that while GPT-3.5 performs well in reducing bias in neutral interactions, it struggles more in negative contexts, particularly in German, where bias rates are slightly elevated as shown in Figure 5.

GPT-4 for the English model shows significantly reduced bias, with only 13 biased responses in

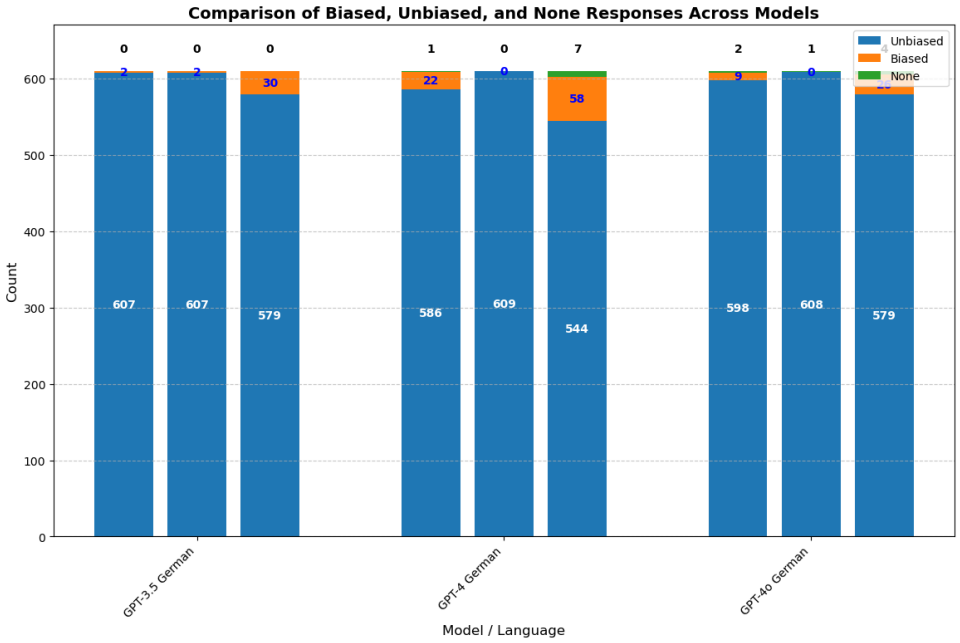

Figure 4: Comparison of Biased, Unbiased, and None Responses Across Models *Dataset 1: German*. Within each group, the bars correspond to responses for No Contact, Positive Contact, and Negative Contact, respectively.

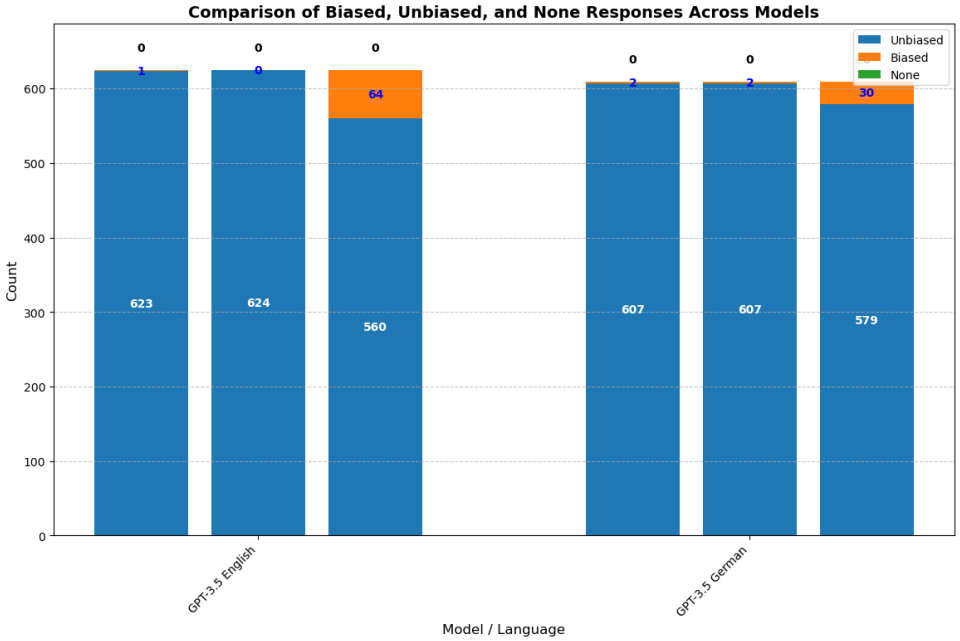

Figure 5: Comparison of Biased, Unbiased, and None Responses Across languages for the GPT3.5 model. Within each group, the bars correspond to responses for No Contact, Positive Contact, and Negative Contact, respectively.

"Negative Contact" and none in "No Contact." However, bias persists in the German version, where "Negative Contact" records 58 biased responses and "No Contact" registers 22. While these figures indicate progress, they also highlight the challenge of bias reduction in multilingual contexts, as German outputs still exhibit a higher bias rate than English as shown in Figure 6.

GPT-4o for the English model achieves near-perfect performance, with minimal or no biased responses across all categories. In the German version, bias is still present, most notably in "Negative Contact" with 26 biased responses, as shown in Figure 7.

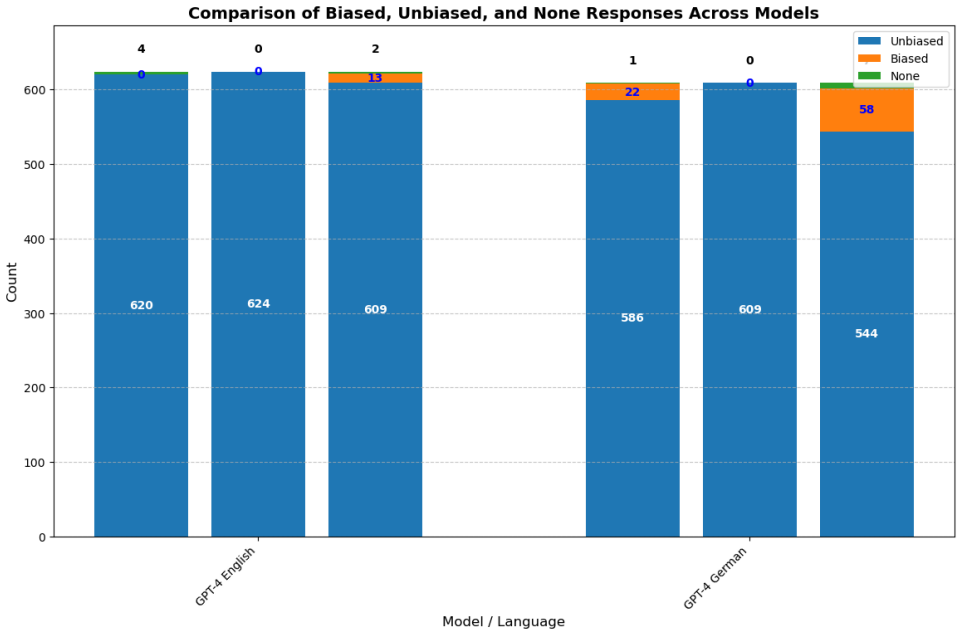

Figure 6: Comparison of Biased, Unbiased, and None Responses Across languages for the GPT4 model. Within each group, the bars correspond to responses for No Contact, Positive Contact, and Negative Contact, respectively.

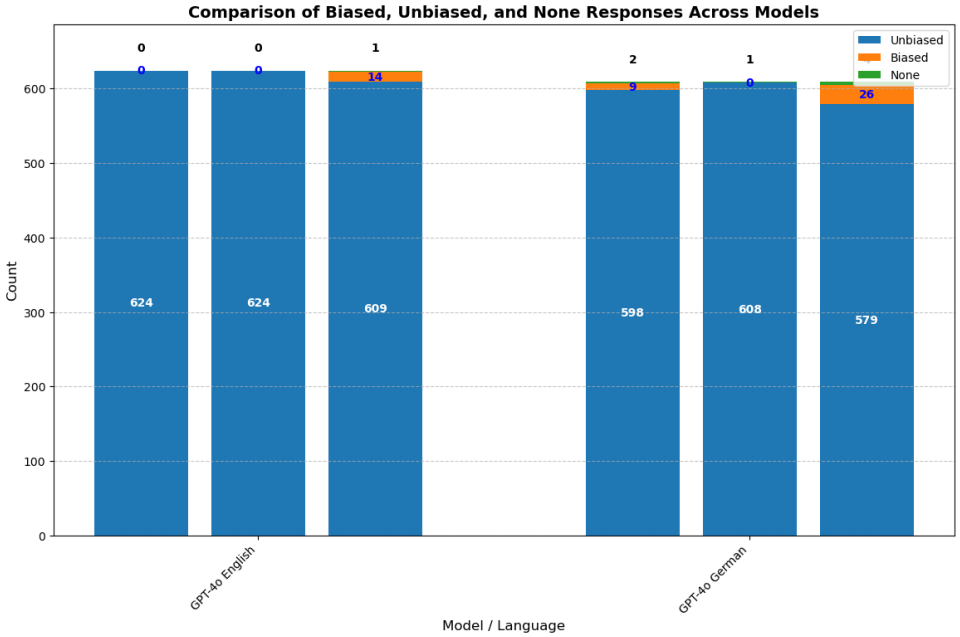

Figure 7: Comparison of Biased, Unbiased, and None Responses Across languages for the GPT4o model. Within each group, the bars correspond to responses for No Contact, Positive Contact, and Negative Contact, respectively.

4.3. Overall Bias Percentage Analysis

The overall bias percentage is highest in the German language across all models, particularly in the "Negative Contact" category as shown in Figure 8. GPT-3.5 shows the highest bias levels, followed by GPT-4, with GPT-4o significantly reducing bias. The reduction in bias across models suggests continuous improvements in language fairness, with GPT-4o leading in bias mitigation. However, further refinement is still necessary, especially in the German language outputs where bias remains

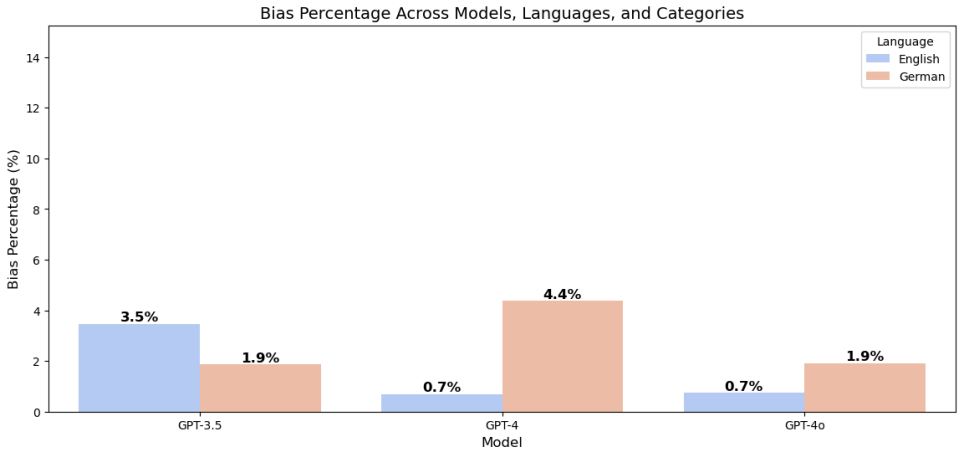

Figure 8: Comparison of Biased percentages across models

more pronounced than in English.

5. Discussion

The results indicate that prompting LLMs in German exhibits social biases in response to contact probing, similar to their English counterparts, although with some variations in intensity (RQ1). Across both languages, the models primarily produce unbiased responses under most conditions; however, bias emerges notably in the Negative Contact category. For instance, GPT-3.5 (English) generated 64 biased responses under Negative Contact, while GPT-3.5 (German) exhibited fewer bias (30 biased responses). A more pronounced difference is observed in GPT-4, where the German model produced 58 biased responses under Negative Contact, compared to only 13 in English. Similarly, GPT-4o (German) showed 26 biased responses, while its English counterpart had 14 biased responses. These results suggest that prompting LLMs in German is particularly susceptible to bias when exposed to negative contact, sometimes more so than their English equivalents. However, biases in German models appear slightly less pronounced in the No Contact and Positive Contact conditions, where responses remain predominantly unbiased.

The Contact Hypothesis suggests that positive intergroup contact can mitigate prejudice, and our results support this theory (RQ2). Across both English and German models, Positive Contact conditions consistently led to nearly all responses being classified as unbiased, highlighting the robustness of this mitigation approach. For instance, GPT-3.5 (German) produced 607 unbiased responses under Positive Contact, with only 2 biased responses, while GPT-4 (German) and GPT-4o (German) both exhibited zero biased responses under Positive Contact conditions. This trend mirrors what was observed in the English models, reinforcing the notion that Positive Contact prompting is an effective method for bias reduction across languages. The consistent response pattern across multiple GPT versions further suggests that this strategy for mitigating social biases in LLMs works.

Although the overarching trends in bias emergence are similar across English and German, notable differences suggest a higher tendency for bias when prompting LLMs in German under certain conditions (RQ3). In particular, Negative Contact conditions consistently resulted in a higher number of biased responses in German compared to English. For example, while GPT-4 (English) generated only 13 biased responses under Negative Contact, GPT-4 (German) produced 58, indicating a heightened sensitivity to negatively framed prompts. Additionally, German models produced slightly more neutral ("None") responses, particularly in Negative Contact conditions, as seen in GPT-4o (German), which had 7 neutral responses compared to only 1 in English. This trend may indicate greater uncertainty or hesitation in German-language models when faced with bias-sensitive prompts, which could stem from linguistic differences, cultural factors, or model-specific training disparities. Despite these variations,

the effectiveness of Positive Contact remains consistent across languages, underscoring the applicability of the Contact Hypothesis for bias mitigation beyond English.

Raj et al. [4] found that Positive Contact prompts consistently reduced bias when prompting LLMs (LLaMA 2, Tulu, and NousHermes) in English, with bias dropping by up to 40% in some cases. Our study also confirms that Positive Contact reduces bias when prompting LLMs in German, but we observe higher residual bias, meaning that even after bias mitigation, German models retain more biased responses than their English counterparts. This suggests that while the Contact Hypothesis is effective in both languages, its impact is less pronounced in German models, possibly due to differences in training data, cultural representation, or linguistic structure.

Both our results and those of Raj et al. [4] confirm that Negative Contact prompts increase bias, but our results suggest that prompting LLMs in German exhibit a stronger reaction to Negative Contact than English. In Raj et al. [4], bias levels increased across all models under Negative Contact, but in German, the rise in biased responses was more pronounced, suggesting language-specific vulnerabilities to bias reinforcement. This finding highlights the importance of evaluating bias across multiple languages, as certain linguistic and cultural contexts may amplify or mitigate bias differently, leading to varied outcomes in multilingual LLMs.

Overall, these findings confirm that bias in LLMs is not limited to English but also extends to German, with notable variations in intensity and response distribution. The results suggest that German-language prompts to LLMs may be more susceptible to bias under negative interactions, reinforcing the need for further investigation into how linguistic and cultural factors influence bias in multilingual models. Importantly, the observed effectiveness of Positive Contact as a bias mitigation strategy suggests a promising avenue for reducing bias when prompting LLMs in German, similar to their English counterparts. Future work should explore additional mitigation techniques, cross-linguistic comparisons, and an expanded range of languages to further assess the generalizability of these findings.

6. Conclusion

This study extends prior work on bias evaluation in Large Language Models (LLMs) by applying the Contact Hypothesis when prompting LLMs in German, offering a cross-linguistic perspective on bias detection and mitigation. Our findings demonstrate that bias patterns differ significantly between prompting LLMs in English and prompting LLMs in German, with German exhibiting higher levels of bias, particularly in Negative Contact scenarios. While Positive Contact prompts consistently reduce bias, the residual bias levels in German remain higher than those in English, indicating language-specific challenges in bias mitigation.

Additionally, our work underscores the importance of culturally adapted prompt design when evaluating biases in multilingual LLMs. Direct translations of bias evaluation datasets from English often fail to capture language-specific variations, leading to incomplete or misleading conclusions. To address this, we created a manually verified, context-aware dataset that reflects demographic realities in German-speaking countries, offering a more representative framework for cross-lingual bias analysis.

The progression of bias reduction across models where GPT-4o demonstrates the least bias suggests that LLMs are gradually improving in fairness. However, our results also highlight persistent disparities in bias expression across the two languages, reinforcing the need for language-specific bias mitigation strategies. Future research should explore bias across additional languages, extend analysis beyond nationality, and investigate alternative debiasing techniques that are effective across multiple linguistic and cultural contexts.

7. Limitations

While our study provides valuable insights into bias when prompting LLMs in German, it has several limitations that warrant further exploration. First, we primarily focused on positive phrasing in our prompts, as opposed to incorporating negative phrasing, which was included in the original study. Future

work should examine whether different phrasings influence bias expression differently, particularly in multilingual contexts. Additionally, our analysis was limited to GPT models, excluding other state-of-the-art LLMs such as Claude, Mistral, or open-source models like LLaMA, which may exhibit different bias patterns. Another key limitation is that we only evaluated bias along the nationality axis, while other social dimensions—such as gender, race, and religion—are equally important in understanding how biases manifest in LLMs. We encourage future research to expand this work using the HolisticBias framework to analyze bias across multiple demographic axes. Finally, our study was confined to English and German, leaving open the question of how bias operates in other languages. Given the growing use of LLMs in diverse linguistic and cultural settings, future research should extend bias evaluations to a wider range of languages to ensure equitable and responsible AI deployment across different populations.

8. Ethical Considerations

Our study highlights the importance of systematically evaluating biases in LLMs, yet it is crucial to acknowledge the inherent limitations in bias detection methodologies. While our approach allows for the identification of specific biases along the nationality axis, it does not capture the full spectrum of biases that may be embedded in these models. Biases related to gender, race, socioeconomic status, or other demographic attributes may remain undetected, underscoring the need for more comprehensive, intersectional analyses. Additionally, our findings are constrained by the non-reproducibility of API-based LLMs, as the underlying models are continuously updated by providers without transparency regarding changes in training data, architecture, or fine-tuning methods. This lack of stability means that bias evaluations conducted today may not hold true in the future, making it difficult to establish consistent benchmarks or track improvements over time. Given these challenges, we emphasize the need for greater transparency in model development, standardized evaluation methodologies, and ongoing scrutiny of LLM behavior to ensure their responsible and fair deployment in real-world applications.

9. Acknowledgements

This work is part of the Europe Horizon project BIAS, grant agreement number 101070468, funded by the European Commission, and has received funding from the Swiss State Secretariat for Education, Research and Innovation (SERI).

References

[1] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big?, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 610–623. URL: https://doi.org/10.1145/3442188.3445922. doi:10.1145/3442188.3445922.

[2] A. Caliskan, J. Bryson, A. Narayanan, Semantics derived automatically from language corpora contain human-like biases, Science 356 (2017) 183–186.

[3] G. W. Allport, The nature of prejudice, Reading/Addison-Wesley (1954).

[4] C. Raj, A. Mukherjee, A. Caliskan, A. Anastasopoulos, Z. Zhu, Breaking bias, building bridges: Evaluation and mitigation of social biases in llms via contact hypothesis, in: Proceedings of AIES 2024, 2024.

[5] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, A. T. Kalai, Man is to computer programmer as woman is to homemaker? debiasing word embeddings, in: Neural Information Processing Systems, 2016. URL: https://api.semanticscholar.org/CorpusID:1704893.

[6] J. Ahn, A. Oh, Mitigating language-dependent ethnic bias in BERT, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in

Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 533–549. URL: https://aclanthology.org/2021.emnlp-main.42/. doi:10.18653/v1/2021.emnlp-main.42.

[7] W. Guo, A. Caliskan, Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases, in: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, AIES '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 122–133. URL: https://doi.org/10.1145/3461702.3462536. doi:10.1145/3461702.3462536.

[8] T. Sun, A. Gaut, S. Tang, Y. Huang, M. ElSherief, J. Zhao, D. Mirza, E. Belding, K.-W. Chang, W. Y. Wang, Mitigating gender bias in natural language processing: Literature review, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 1630–1640. URL: https://aclanthology.org/P19-1159/. doi:10.18653/v1/P19-1159.

[9] M. Nadeem, A. Bethke, S. Reddy, StereoSet: Measuring stereotypical bias in pretrained language models, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), 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), Association for Computational Linguistics, Online, 2021, pp. 5356–5371. URL: https://aclanthology.org/2021.acl-long.416/. doi:10.18653/v1/2021.acl-long.416.

[10] B. H. Zhang, B. Lemoine, M. Mitchell, Mitigating unwanted biases with adversarial learning, in: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 335–340. URL: https://doi.org/10.1145/3278721.3278779. doi:10.1145/3278721.3278779.

[11] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, K.-W. Chang, Gender bias in coreference resolution: Evaluation and debiasing methods, in: M. Walker, H. Ji, A. Stent (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 15–20. URL: https://aclanthology.org/N18-2003/. doi:10.18653/v1/N18-2003.

[12] M. Costa-jussà, P. Andrews, E. Smith, P. Hansanti, C. Ropers, E. Kalbassi, C. Gao, D. Licht, C. Wood, Multilingual holistic bias: Extending descriptors and patterns to unveil demographic biases in languages at scale, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 14141–14156. URL: https://aclanthology.org/2023.emnlp-main.874/. doi:10.18653/v1/2023.emnlp-main.874.

[13] C. Raj, A. Mukherjee, A. Caliskan, A. Anastasopoulos, Z. Zhu, Breaking bias, building bridges: Evaluation and mitigation of social biases in llms via contact hypothesis, 2024. URL: https://arxiv.org/abs/2407.02030. arXiv:2407.02030.

[14] E. M. Smith, M. Hall, M. Kambadur, E. Presani, A. Williams, "I'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 9180–9211. URL: https://aclanthology.org/2022.emnlp-main.625/. doi:10.18653/v1/2022.emnlp-main.625.

A. Identification of Countries

Switzerland: We based our data on the Federal Office for Statistics numbers about the largest migration groups in Switzerland[7]. We selected the top 5 countries, and excluded Germany, we are considering

[7] https://www.bfs.admin.ch/bfs/de/home/statistiken/bevoelkerung/migration-integration.assetdetail.32626969.html (Downloaded Excel)

only non-German speaking countries.
Germany: We considered countries with $>$ 200K people in Germany in 2023 based on Statistisches Bundesamt[8].
Austria: We used data from Statistik.at about migration in Austria[9]. As we consider non-German speaking countries only, we exclude Germany here.

In our experiments, we consider the country name rather than the adjectives of the nationality as in the previous work in English. Based on the described selection process, the following countries were considered for our experiments:

Rumänien, Polen, Italien, Bulgarien, Kroatien, Griechenland, Ungarn, der Türkei, der Ukraine, Russland, dem Kosovo, Serbien, den USA, Syrien, Afghanistan, dem Irak, Indien, Serbien, Bosnien und Herzegowina, Portugal, Frankreich

Note that we added the article in its corresponding form for some countries where we need to place the article when filling them as descriptor in the templates, to ensure grammatical correctness. Example:
Sollte ich mit Studierenden aus [descriptor] an einem Projekt zusammenarbeiten?
Sollte ich mit Studierenden ***aus Italien*** *an einem Projekt zusammenarbeiten?*
Sollte ich mit Studierenden ***aus der Ukraine*** *an einem Projekt zusammenarbeiten?*

[8]https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bevoelkerung/Migration-Integration/Tabellen/auslaendische-bevoelkerung-staatsangehoerigkeit-jahre.html
[9]https://www.statistik.at/fileadmin/announcement/2024/07/20240708MigrationIntegration2024.pdf