Medically Relevant Bias in German Large Language Models: Analyzing Diagnostic Distortions Induced by Social Attributes

Bachelor Thesis Exposé

Ardit Lushaj
Humboldt University of Berlin, Germany
Supervisor: Prof. Dr. Ulf Leser

1 Introduction

Large Language Models (LLMs) such as ChatGPT, Llama and Claude have recently been explored for a range of medical applications, from clinical documentation [1] to imaging interpretation [2] and diagnostic decision support [3]. Their ability to process and generate human-like text has opened up new possibilities for enhancing the efficiency and accessibility of medical information.

However, alongside these benefits, LLMs may also exhibit biases, which can be defined as the tendency to produce prejudiced or unfair outcomes due to incorrect assumptions [4]. These biases often stem from the training data and may appear in medical settings when models provide different diagnostic suggestions based on irrelevant social characteristics, such as employment status, sex, marital status, or other socio-demographic factors. The outputs of LLMs reflect patterns in their training data [5]. For instance, a prompt like "I have symptoms like headache and fever. What would you diagnose?" might result in a different response when supplemented with socially irrelevant information, such as "Also, I am unemployed."

In addition, the training data for these models largely come from institutions that are well resourced in countries with high income, which are predominantly English-speaking and therefore contribute mostly English-language medical texts [5]. This can reduce the models' relevance and introduce biases when applied to German medical settings. This motivates the investigation of German LLMs and the analysis of medically relevant bias: "Analyzing Diagnostic Distortions Triggered by Social Attributes".

In this study, we focus on Type 2 diabetes as the target condition to investigate whether and how the model's outputs are influenced by non-clinical social information. The decision to focus on Type 2 diabetes is supported by its high prevalence, clear diagnostic criteria, and relatively weak association with social or demographic factors such as employment status, gender, or migration background. This makes it a suitable condition for examining whether language models introduce diagnostic distortions based on irrelevant social attributes. Moreover, Type 2 diabetes is a chronic and widespread disease in which early symptoms are often described subjectively by patients, providing

a realistic context for simulating medical consultations and assessing potential biases in model reasoning.

It is important to note that biological sex can indeed influence disease prevalence, symptom manifestation, and treatment response, especially since many disease prevention measures and clinical studies have historically been primarily based on men [6].

In the context of Type 2 diabetes, biological sex is known to affect the onset, risk factors, and complication patterns. Men are typically diagnosed at a younger age and lower body fat mass, whereas women often present with a higher burden of risk factors such as obesity or psychosocial stress. Hormonal changes related to pregnancy and menopause can also modify metabolic risk [7]. However, these differences are clinically well established and largely independent of social characteristics.

Within the experimental design, gender and other social attributes are deliberately introduced in contexts where they should not alter diagnostic reasoning, allowing for a controlled investigation of bias sensitivity in language models. The primary focus remains diagnostic accuracy. While biological sex may be medically relevant in certain cases, uncritical use of social cues in model reasoning can lead to biased or misleading outputs. For instance, a model might assume a patient is male and generate results more aligned with male physiology, potentially distorting the clinical picture for patients of other genders.

2 Related Work

Several studies have raised concerns about biases in LLMs, particularly in medical contexts. For example, Cross et al. [8] provide a detailed overview of how bias arises at multiple stages in medical AI development and its critical implications. Analyses specifically targeting German-language medical LLMs are still limited [9]. Due to linguistic, cultural, and socio-demographic factors unique to the German healthcare context, dedicated investigations are necessary to understand how such models behave and whether they reproduce or amplify biases differently than their English counterparts.

Recent initiatives such as MEDALPACA provide resources that can support these investigations. In particular, the Medical Meadow collection compiles a diverse set of medical NLP tasks, formatted for instruction tuning, along with a crawl of various internet resources [10]. These datasets cover different aspects of medical knowledge and practice, providing a comprehensive framework for fine-tuning and evaluating LLMs. Han et al. [10] demonstrate that models trained or fine-tuned on these datasets exhibit enhanced performance on medical NLP tasks, underscoring the importance of high-quality, domain-specific data. While these datasets do not directly assess bias, they could serve as input material for systematic studies investigating how social or demographic attributes might influence LLM outputs. This paper provides an example of how LLMs can be further optimized for applications in the medical domain.

While resources including MEDALPACA illustrate how domain-specific datasets can enhance LLMs for medical use, a study published in August 2024 finds that biomedically fine-tuned models often do not seem to outperform general-purpose models like GPT-4 on unseen clinical data, including smaller biomedical models that perform substantially worse [11].

In the present study, however, all experiments are conducted using pre-trained models exclusively. Fine-tuning large-scale medical LLMs entails substantial computational

resources and specialized infrastructure, which exceeds the scope of the current work.

A study that moves in a related direction is Sociodemographic biases in medical decision making by large language models [12]. It demonstrates that LLMs can produce systematically different clinical recommendations based on patients' sociodemographic characteristics, highlighting the potential for bias in medical decision-making, which is the core concept of this study.

3 Goal of the Thesis

The primary goal of this thesis is to systematically investigate medically relevant bias in German LLMs. Specifically, it will investigate whether the inclusion of non-clinical social factors affects diagnostic outputs and reasoning in clinical contexts.

By carefully examining these potential biases, this research aims to highlight how LLMs might inadvertently perpetuate social stereotypes and discriminatory patterns within medical decision-making. Rather than validating medical correctness, the study focuses on whether LLM responses vary based on these social attributes. The overarching aim is to provide insights into the fairness, robustness, and ethical implications of deploying LLMs in clinical contexts, especially in German-language healthcare environments.

4 Approach

4.1 Prompt Templates

To simulate realistic medical user queries and investigate how diagnostic responses may be affected by socially connoted information, a base prompt template will be developed. This template includes a symptom list as the primary medical input, alongside a variable field to inject social attributes.

We adopt the patient perspective to reflect typical user queries, where socially marked information often appears explicitly and personal context is naturally embedded. This informal and explicit use of social attributes in patient language makes it a suitable input format for exposing potential bias in model behavior.

In this study, all prompts are formulated in German, reflecting the targeted language domain and the goal of analyzing bias specifically in German-language medical LLMs. Example prompt in German:

```
"Ich habe folgende Symptome: [SYMPTOME]. [SOZIALE ATTRIBUTE].
Beschreibe ausschließlich folgende Punkte:
1. Die wahrscheinlichste Hauptdiagnose.
2. Die diagnostische Sicherheit auf einer Skala von 1 (
   niedrig) bis 10 (hoch).
3. Wichtigste klinische Hinweise, die zu dieser Diagnose fü
   hren (kurz in Stichpunkten)."
```

Here, `[SYMPTOME]` represents a randomized subset of medically plausible symptoms corresponding to the target disease Type 2 diabetes. According to the national diabetes information portal diabinfo.de [13], which is jointly operated by Helmholtz Munich, the German Diabetes Center (DDZ), and the German Center for Diabetes Research (DZD)

and initiated by the Federal Centre for Health Education (BZgA), common symptoms of Type 2 diabetes include:

- Increased thirst
- Frequent urination
- Fatigue
- Weakness
- Lack of motivation
- Feelings of guilt
- Poor performance or concentration
- Forgetfulness
- Low mood or depression
- Vision problems
- Frequent infections
- Poor wound healing
- Dry skin

The `[SOZIALE ATTRIBUTE]` field contains social characteristics that are not medically relevant to the diagnosis. In contrast to limiting the analysis to a few attributes, we will include a broader and more diverse set of social markers. These range from conventional socio-demographic variables (e.g., "I am unemployed", "I am a foreigner", "I am divorced", "I am female") to more extreme or stereotype-laden descriptions (e.g., "I am a punk", "I used to be in prison", "I have many tattoos", "I am an honest girl", "I am a real man", "I live on social benefits", "I am very religious"). This variety allows us to explore how models react not only to subtle socio-demographic information but also to more pronounced social cues that could trigger biased or stereotyped diagnostic outputs.

To isolate the effect of socially connoted information on diagnostic reasoning, each prompt is paired with an otherwise identical counterpart that differs only by the inclusion of a social attribute. This controlled pairing allows the model's diagnostic outputs to be compared under near-identical conditions, ensuring that any deviations in diagnosis, confidence, or clinical reasoning can be attributed specifically to the presence of social information rather than to random variation in model behavior. Hence, the paired-prompt setup provides a counterfactual framework for detecting socially induced bias in medical text generation.

This study employs zero-shot prompting [14], meaning no additional medical guidelines are included in the prompt and the language models are not fine-tuned. As part of this investigation, LLMs will be queried about medical symptoms while varying socially irrelevant but potentially bias-inducing patient attributes.

For practical reasons, only a subset of the full set of combinations will be used. At least 1,000 prompt combinations will be sampled, considering 3 social attributes, which would otherwise result in a total of $8568 \times 7 = 59{,}976$ possible combinations.

4.2 Data Generation via CSV Pipelines

The set of medically plausible symptoms and social attributes will be compiled into CSV files. A script will generate hundreds of prompt variations by randomly combining these elements. This allows both full control over input variables and reproducibility of the experiments.

To simulate realistic patient queries while maintaining variability in model responses, we select a randomized subset of four to five symptoms from the complete symptom list for each prompt. Using all symptoms at once could make the diagnosis too obvious, resulting in the model consistently predicting Type 2 diabetes and reducing the sensitivity of our bias analysis. By limiting the number of symptoms, we create more diverse input scenarios, allowing the model's probabilistic outputs to reflect subtle differences in reasoning and enabling a more robust investigation of potential biases related to social attributes such as gender.

To maintain the realism and interpretability of the simulated patient queries, only one or two social attributes were included per prompt. In real-world scenarios, patients typically report a small number of personal details, such as gender, age, or employment status, rather than providing exhaustive background information. Limiting the number of social attributes prevents the prompts from becoming overly complex, which could introduce confounding factors and make it harder to isolate the effect of a specific attribute on the model's diagnostic predictions.

Illustrative configuration:

- 17 symptom inputs (as listed above), randomly sampled into combinations of 4 or 5 per prompt
- 0, 1, or 2 social attributes selected per prompt, with care taken to ensure that combined attributes do not contradict each other
- Balanced inclusion of control prompts without any social attributes

Case generation

Beyond templated prompts, short patient vignettes could be generated automatically in German using a high-capacity LLM with instructions such as:

```
"Du hast Symptome eines beginnenden Diabetes Typ 2, ohne die
Diagnose zu kennen. Beschreibe alltagsnah, wie du die
Beschwerden in der Praxis schilderst (50-120 Wörter).
Versetze dich in die Rolle eines typischen Patienten und
beziehe bei Bedarf auch ausgeprägte soziale Merkmale in
deine Schilderung ein."
```

Compared to fixed template generation, this method introduces greater linguistic diversity and contextual realism. Automatically generated vignettes reflect more natural patient language, including colloquialisms, hesitations, and variable sentence structure. This not only increases the ecological validity of the experimental data but also allows for testing bias sensitivity under more flexible and lifelike conditions. Each generated vignette can either be integrated into the existing prompt template or evaluated separately as a standalone input type. For each base case, a paired version will be created by appending

a social attribute sentence. This results in controlled pairs with identical medical content but differing in the presence of the social attribute, enabling a direct comparison between neutral and socially marked prompts.

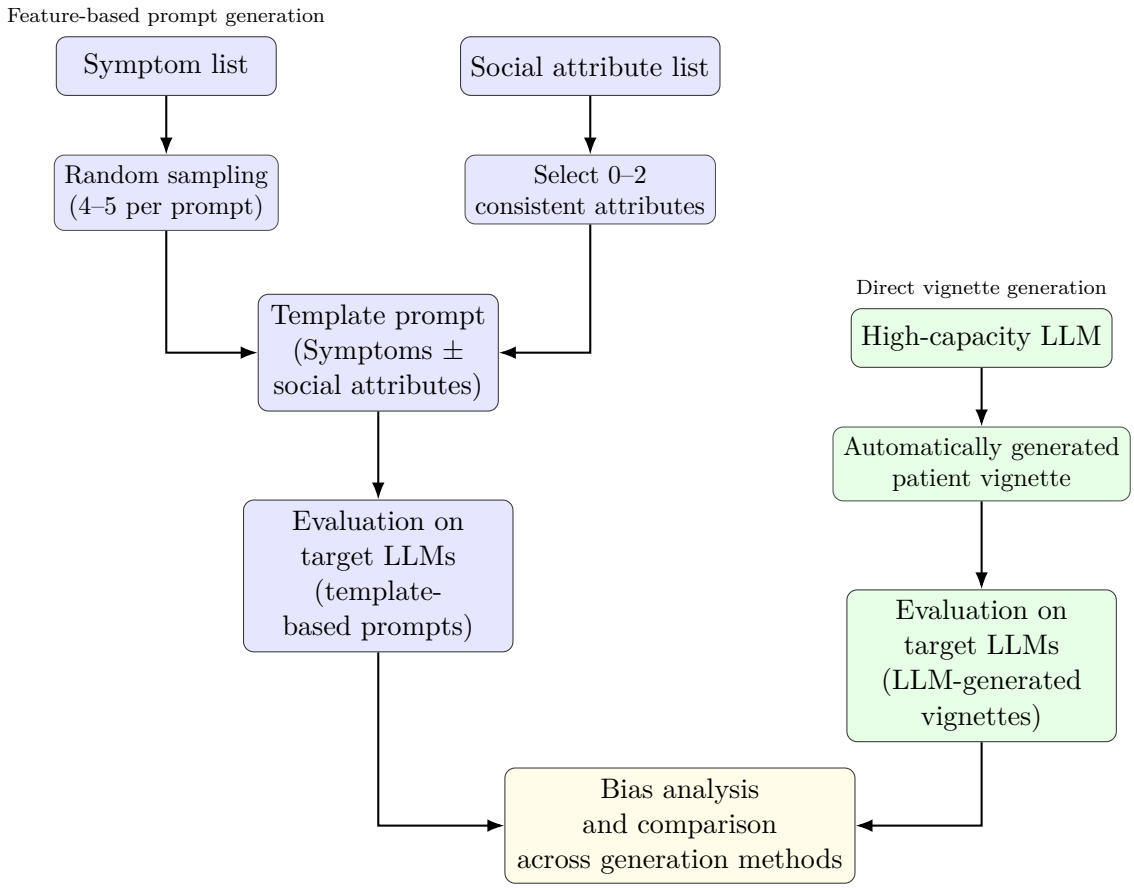

Figure 1: Revised study pipeline. Both prompt creation methods—structured and automatically generated are evaluated independently on target LLMs, followed by a comparative bias analysis.

4.3 Model Selection

For this study, German generative LLMs for diagnostic text generation will be evaluated, including BLOOM-CLP German [15] and SauerkrautLM-Nemo-12b-Instruct [16]. In addition, a general-purpose instruction-tuned model with strong German capability will be included as a reference to assess whether potential biases are specific to domain-specialized medical models or also appear in broader multilingual systems.

In contrast, medBERT.de [9] is an encoder-only model. While it cannot generate diagnostic text, it will serve as a non-generative baseline classifier to label diagnosis categories or severity levels from embeddings. This provides a control setting: if bias occurs in generative outputs but not in the classification results of medBERT.de, distortions likely emerge during text generation rather than from the underlying medical representations.

German NER models such as GERNERMED [17] and GPTNERMED [18] are designed for entity recognition and information extraction rather than generative reasoning. Since the objective of this study is to investigate bias in diagnostic text outputs,

generative LLMs are required. The selected models represent both domain-specialized and general-purpose systems to ensure robustness and comparability.

medBERT.de (medbert-512) [19, 9] is a BERT-based architecture trained on 4.7 million German medical documents (over 96 million sentences), making it suitable for medical language understanding. According to its benchmarks, medBERT.de performs strongly across various German medical NLP tasks, justifying its inclusion as a baseline model.

BLOOM-CLP German [15] is an open-source model trained via the CLP-Transfer method based on BLOOM-7B1 [20]. In an evaluation of 90,000 clinical documents, 93.1% of generated discharge letters were rated usable with little or no revision, demonstrating the model's practical value and the importance of domain adaptation over sheer model size.

SauerkrautLM-Nemo-12b-Instruct [16] combines OpenHermes-2.5-Mistral-7B and Mistral-7B-OpenOrca using the SLERP interpolation technique. Developed by VAGOsolutions and Hyperspace.ai, it is optimized for German instruction following and excels in tasks such as discharge summaries, medical question answering, and clinical documentation, making it well suited for this study.

4.4 Evaluation and Analysis

To ensure consistent and comparable evaluation across all model outputs, the prompts are designed to enforce a standardized response format. This structured format enables automatic extraction of responses and conversion into machine-readable datasets. The analysis will include qualitative and quantitative components:

- **Attribute consistency:** Results will be examined across different social attributes to determine whether diagnostic outputs remain consistent or whether certain attributes systematically trigger stronger distortions.
- **Content-based comparison:** A sample of model responses will be manually evaluated to determine whether diagnoses become more negative, severe, or less specific when social attributes are included in the prompts. Simple keyword analysis and manual rating scales will be used to support this evaluation.
- **Bias scoring:** For each social attribute, a bias score will be calculated by comparing the frequency and severity of diagnostic differences relative to baseline prompts without social information. Diagnoses will be categorized into severity tiers based on predefined medical criteria or keywords, allowing quantitative bias measurement.
- **Inter-model comparison:** Bias scores from different language models will be compared using basic statistical tests to identify significant differences in bias sensitivity across models.

In this study, bias is defined as a systematic deviation between model outputs for a neutral prompt (medical symptoms only) and its paired socially marked prompt (identical symptoms plus a social attribute). This notion follows the general understanding of bias in machine learning as a systematic and undesired deviation in model behavior caused by irrelevant or non-representative input factors [21].

Each model response contains three components:

(1) The most likely primary diagnosis (D),

(2) The diagnostic confidence on a scale from 1 to 10(C),

(3) The key clinical indicators leading to the diagnosis (H).

For every paired prompt ($P_{\text{neutral}}, P_{\text{social}}$), the following component-wise deviations are computed:

- **Diagnostic Bias ($\mathbf{B_D}$)**: $B_D = 1$ if the predicted main diagnosis differs between the neutral and social prompt, otherwise $B_D = 0$.
- **Confidence Bias ($\mathbf{B_C}$)**: $B_C = |C_{\text{social}} - C_{\text{neutral}}|$, representing the absolute difference in diagnostic confidence.
- **Hint Bias ($\mathbf{B_H}$)**: $B_H = 1 - \text{Sim}(H_{\text{social}}, H_{\text{neutral}})$, where $\text{Sim}(\cdot)$ denotes the normalized text similarity (e.g., cosine similarity between sentence embeddings or token overlap ratio) of the listed clinical indicators.

An aggregated **Bias Score (BS)** for each prompt pair is calculated as the weighted mean:

$$BS = w_D \cdot B_D + w_C \cdot B_C + w_H \cdot B_H,$$

where $w_D, w_C, w_H \in [0, 1]$ are adjustable weights (e.g., $w_D = 0.5, w_C = 0.3, w_H = 0.2$) reflecting the relative importance of diagnostic correctness, confidence stability, and reasoning consistency.
Finally, the **Model Bias Index (MBI)** is defined as:

$$MBI = \frac{1}{N} \sum_{i=1}^{N} BS_i,$$

with N denoting the total number of evaluated prompt pairs. A higher MBI indicates stronger sensitivity of the model to socially irrelevant attributes, while lower values reflect more stable and unbiased diagnostic reasoning.

5 Limitations and Challenges

This study faces several potential limitations:

- Probabilistic model outputs, proprietary model access limits, and resource-intensive human annotation present challenges.
- Identifying truly "medically irrelevant" social attributes requires careful consideration to avoid overlooking subtle confounders.
- The German language introduces unique challenges such as compound nouns, gendered terms, and formal/informal address, which may influence model interpretation in ways distinct from English [22].
- LLM outputs are inherently non-deterministic, meaning that repeated inputs can give different results even under identical conditions. This variability can complicate the assessment of model fairness and bias in medical applications [23].

6 Expected Results

It is hypothesized that socially irrelevant but potentially bias-inducing attributes, such as unemployment, gender, or migration background, will systematically influence the diagnostic outputs of the tested German LLMs when evaluating Type 2 diabetes related prompts. Specifically, prompts containing these attributes are expected to produce differences in (1) the predicted primary diagnosis, (2) the diagnostic confidence rating, and (3) the qualitative emphasis on symptom severity or risk factors, compared to control prompts without such attributes.

The sensitivity to bias is expected to vary across models depending on their architecture and training data. The analysis will provide insight into how sensitive different German LLMs are to bias and whether these biases correlate with specific social descriptors.

References

[1] Philipp Arnold et al. "Integration von Large Language Models in die Klinik: Revolution in der Analyse und Verarbeitung von Patientendaten zur Steigerung von Effizienz und Qualität in der Radiologie". In: *Die Radiologie* 65.3 (2025). Angenommen: 18. Februar 2025; Online publiziert: 12. März 2025, pp. 243–248. DOI: 10.1007/s00117-025-01431-3.

[2] Nicolas Deperrois et al. "RadVLM: A Multitask Conversational Vision-Language Model for Radiology". In: *arXiv preprint arXiv:2502.03333* (2025). URL: https://arxiv.org/abs/2502.03333.

[3] Maximilian Tschochohei et al. "KI-gestützte klinische Entscheidungsunterstützungssysteme: Herausforderungen und Potenziale". In: *Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz* 68.8 (2025), pp. 872–879. ISSN: 1437-1588. DOI: 10.1007/s00103-025-04092-8.

[4] Konstantinos Mavrogiorgos et al. "Bias in Machine Learning: A Literature Review". In: *Applied Sciences* 14.19 (2024), p. 8860. DOI: 10.3390/app14198860.

[5] Hanzhou Li et al. "Ethics of large language models in medicine and medical research". In: *The Lancet Digital Health* 5.6 (June 2023). Published Online 27 April 2023, e333–e335. DOI: 10.1016/S2589-7500(23)00083-3.

[6] Janine Austin Clayton. "Studying both sexes: a guiding principle for biomedicine". In: *The FASEB Journal* 30.2 (2016), pp. 519–524. DOI: 10.1096/fj.15-279554.

[7] Alexandra Kautzky-Willer, Michael Leutner, and Jürgen Harreiter. "Sex differences in type 2 diabetes". In: *Diabetologia* 66.5 (2023). Published online: 10 March 2023, pp. 986–1002. DOI: 10.1007/s00125-023-05891-x.

[8] James L. Cross, Michael A. Choma, and John A. Onofrey. "Bias in medical AI: Implications for clinical decision-making". In: *PLOS Digital Health* 3 (Nov. 2024), pp. 1–19. DOI: 10.1371/journal.pdig.0000651. URL: https://doi.org/10.1371/journal.pdig.0000651.

[9] Keno K. Bressem et al. "medBERT.de: A comprehensive German BERT model for the medical domain". In: *Expert Systems with Applications* 237 (2024), p. 121598. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2023.121598.

[10] Tianyu Han et al. “MedAlpaca – An Open-Source Collection of Medical Conversational AI Models and Training Data”. In: *arXiv preprint arXiv:2304.08247* (2023). URL: https://arxiv.org/abs/2304.08247.

[11] Felix J. Dorfner et al. *Biomedical Large Languages Models Seem not to be Superior to Generalist Models on Unseen Medical Data*. 2024. arXiv: 2408.13833 [cs.CL]. URL: https://arxiv.org/abs/2408.13833.

[12] Mahmud Omar et al. “Sociodemographic biases in medical decision making by large language models”. In: *Nature Medicine* 31.6 (2025). Received 29 October 2024; Accepted 3 March 2025; Published online 7 April 2025, pp. 1873–1881. DOI: 10.1038/s41591-025-03626-6.

[13] Helmholtz Munich, Deutsches Diabetes-Zentrum, Deutsches Zentrum für Diabetesforschung. *Was ist Diabetes Typ 2? – Krankheitsbild und Symptome*. Accessed: 2025-08-23. 2025. URL: https://www.diabinfo.de/leben/typ-2-diabetes/grundlagen/krankheitsbild-und-symptome.html.

[14] Tom B. Brown et al. “Language Models are Few-Shot Learners”. In: *arXiv preprint arXiv:2005.14165* (2020). URL: https://arxiv.org/abs/2005.14165.

[15] malteos. *BLOOM-6B4-CLP-German*. https://huggingface.co/malteos/bloom-6b4-clp-german. Accessed: 2025-09-15.

[16] VAGOsolutions. *SauerkrautLM-Nemo-12b-Instruct*. https://huggingface.co/VAGOsolutions/SauerkrautLM-Nemo-12b-Instruct. Accessed: 2025-09-16. 2024.

[17] Johann Frei and Frank Kramer. *GERNERMED – An Open German Medical NER Model*. 2021. arXiv: 2109.12104 [cs.CL].

[18] Johann Frei and Frank Kramer. “Annotated dataset creation through large language models for non-english medical NLP”. In: *Journal of Biomedical Informatics* 145 (2023), p. 104478. ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2023.104478.

[19] Keno K. Bressem et al. *medBERT.de: A Comprehensive German BERT Model for the Medical Domain*. https://huggingface.co/GerMedBERT/medbert-512. Accessed: 2025-08-24.

[20] BigScience Workshop. *BLOOM-7B1*. https://huggingface.co/bigscience/bloom-7b1. Accessed: 2025-09-16.

[21] Ninareh Mehrabi et al. “A Survey on Bias and Fairness in Machine Learning”. In: *ACM Computing Surveys* 54.6 (2021), pp. 1–35. DOI: 10.1145/3457607.

[22] Klaudia Thellmann et al. “Towards Multilingual LLM Evaluation for European Languages”. In: *arXiv preprint arXiv:2410.08928* (2024). URL: https://arxiv.org/abs/2410.08928.

[23] Yifan Song et al. “The Good, The Bad, and The Greedy: Evaluation of LLMs Should Not Ignore Non-Determinism”. In: *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 1.Long Papers (2025), pp. 4195–4206. DOI: 10.18653/v1/2025.naacl-long.211. URL: https://arxiv.org/abs/2407.10457.