Investigating how LLMs propagate gender stereotypes: Comparing what models "say" via prompts with what they "internally represent" in their embeddings

Submitted on: **24-04-2025**

Andrea Valderrey
andreavalderreyn@gmail.com
University of Amsterdam
Amsterdam, The Netherlands

Jelke Bloem
j.bloem@uva.nl
University of Amsterdam
Amsterdam, The Netherlands

1 INTRODUCTION

Large Language Models (LLMs) are being increasingly deployed in various high-stakes applications, leading to growing concerns regarding their potential to propagate social biases, particularly gender stereotypes [14]. These biases, stemming from the large datasets they are trained on, can significantly influence both the models' generated responses and their internal representations, potentially reinforcing harmful assumptions about gender roles and societal expectations [2, 5]. Research has shown that LLMs can exhibit and even amplify existing gender biases, sometimes more so than traditional Neural Machine Translation (NMT) models, by associating genders with specific occupations or attributes in stereotypical ways [22]. Even when not explicitly prompted with gendered terms, LLMs can reveal implicit gender biases in their generated text, and their performance on gender bias evaluation datasets like WinoMT, Gold BUG, and MuST-SHE highlights these tendencies [22]. This raises critical ethical considerations about the safety and transparency of LLMs in sensitive domains.

Recent studies have shown that LLMs frequently produce stereotypical completions when prompted with gendered cues, particularly in tasks like occupation prediction or machine translation[14, 22]. Other research has explored how gender associations are encoded within internal representations, such as embeddings or neuron activations[23, 24]. However, these two lines of inquiry—prompt-level and embedding-level analysis—are typically studied in isolation. Instruction tuning, which is designed to align model outputs with human expectations, may suppress biased responses without necessarily addressing the underlying representational biases. As a result, models may appear fair and unbiased in their outputs while still retaining problematic associations beneath the surface. This disconnect between a model's outputs and its internal representations raises serious concerns, especially in contexts where such biases could influence sensitive or high-stakes decisions. Developing an integrated understanding of how instruction tuning shapes both generated responses and internal representations is essential for evaluating the ethical reliability and transparency of LLMs.

This thesis aims to fill that gap by comparing what models say via prompts with what they internally represent in their embeddings.

The study will use two complementary datasets: GEST [16], which includes statements labelled by stereotype category (e.g. "Men are strong"), and StereoSet [17], which provides contextual triples (stereotype, anti-stereotype, and unrelated sentences). These datasets will be merged to increase coverage and diversity across 16 stereotype categories. Using this merged corpus, the same set of phrases will be used both as prompts to elicit completions and as inputs to extract embeddings. The methodology involves evaluating bias at two levels. At the output level, LLMs will be prompted with stereotype-relevant phrases, and their continuations will be scored using rule-based criteria and sentiment analysis. At the embedding level, vector representations for gendered and occupational terms will be extracted and analysed using cosine similarity and WEAT.

The preliminary research question that will be answered during this thesis is as follows:

- To what extent do LLMs express and encode gender stereotypes, and how does instruction tuning affect the relationship between output-level and embedding-level bias?

To address this question, the study is further divided into the following sub-questions:

- How do different open-source LLMs compare in their expression and representation of gender bias?
- Does instruction tuning reduce output-level gender bias while leaving internal representations largely unchanged, or even reinforcing them?
- Are certain categories of gender stereotypes more likely to be expressed or encoded than others, and does this vary across models and between output and embedding evaluations?

This research contributes to the growing body of work focused on mitigating the harmful effects of gender bias and stereotypes in language models. By examining both output-level and embedding-level behaviour across open-source LLMs, it aims to provide a more comprehensive understanding of how bias manifests and evolves, particularly under instruction tuning. The findings are intended to inform future research on bias evaluation, model alignment, and the development of fairer and more transparent AI systems.

2 RELATED WORK

This section reviews prior work on output-level biases, internal representations, instruction tuning, and evaluation methods, highlighting the gaps this thesis seeks to address.

2.1 Gender Bias in LLM Outputs

Studies have explored gender bias in large language model (LLM) output through prompt-based analyzes that reveal how models reproduce and amplify stereotypes in generated text. Kotek et al. [14] investigated stereotypes by prompting LLMs with sentences pairing gendered pronouns with occupations. For example, when given the sentence "The doctor phoned the nurse because she was late,"

LLMs were significantly more likely to assume "she" referred to the nurse, aligning with the stereotype. They discovered that LLMs were significantly more likely to associate pronouns with stereotypically gendered occupations, and this pattern exceeded both human perceptions and labour statistics. The study identified a "siloing effect" for women, where LLMs selected stereotypically female occupations more frequently than expected. This highlights output-level bias in how LLMs associate gender with professions, as their pronoun resolutions and explanations consistently favoured stereotypical pairings, even when sentence structure did not explicitly require such interpretations. Kapoor and Narayanan [12] examined GPT-3.5 and GPT-4 using the WinoBias benchmark, which tests pronoun resolution in stereotype-sensitive contexts. Models performed substantially worse on anti-stereotypical examples, with GPT-3.5 and GPT-4 being 2.8 and 3.2 times more likely, respectively, to answer incorrectly, sometimes producing completions like claiming that lawyers cannot be pregnant. Sant et al. [22] investigated the effectiveness of prompts in mitigating gender bias in LLMs used for machine translation. They applied techniques such as few-shot prompting, context inclusion, and chain-of-thought reasoning, finding that well-crafted prompts reduced gender bias by up to 12% on WinoMT. Their results demonstrate that prompting strategies play a direct role in modulating biased outputs. Zhao et al. [25] examined the distinction between implicit and explicit gender biases in LLMs using a self-evaluation methodology. Working with LLaMA and GPT-4, they first prompted models to complete masked sentence templates, uncovering implicit gender stereotypes. They then asked the same models to assess the appropriateness of their own completions, capturing explicit bias. The key finding was a human-like inconsistency: while the models exhibited strong gender stereotypes in implicit outputs, they often showed weaker or even rejecting responses during explicit evaluation. This finding highlights the complex nature of gender bias in LLMs and indicates that focusing solely on explicit bias risks missing deeper, implicit patterns encoded in the model.

These studies reveal that gender bias in LLM outputs remains persistent across tasks, models, and prompting contexts, and that mitigating such bias requires not only technical interventions but also a good understanding of how different forms of bias manifest.

2.2 Internal Representations and Embedding Bias

While much of the research on gender bias in LLMs has focused on outputs, a growing body of work investigates how these biases are encoded in the models' internal representations, particularly in their embeddings. Basta et al. [1] studied the underlying gender bias in contextualised word embeddings, adapting measures originally developed for static embeddings. Their results showed that ELMo[1]'s contextual representations are significantly less biased than traditional embeddings, even after the latter have been debiased. This suggests that contextualization helps reduce, but does not eliminate, bias, offering a promising direction for mitigating stereotypical associations at the representation level. Similarly, Schuster et al. [23] examined gender bias in the contextual embeddings of twelve different large language models by transforming these embeddings based on the stereotype content model (SCM) from social psychology. Their analysis demonstrated statistically significant bias for gender-associated names, with a trend of female names being relatively associated with higher warmth and male names with higher competence, aligning with human stereotypes. More recently, Yu and Ananiadou [24] focused on understanding and mitigating gender bias by analysing the neuron-level information flow within LLMs. They introduced the CommonWords dataset to evaluate gender bias and identified specific neuron circuits, including "gender neurons" and "general neurons," responsible for this behaviour, noting that even editing a small number of general neurons can significantly impact the model's overall capabilities.

These studies demonstrate that gender bias in LLMs is not only reflected in output but also deeply embedded in their internal representations, calling for more interpretable and targeted interventions.

2.3 Instruction Tuning and Bias Mitigation

Instruction tuning has become a standard method for aligning LLMs with human-like behaviour, but its effects on bias propagation remain complex and not fully understood. Haller et al. [8] introduced OpinionGPT, a web demo built to make these biases explicit. By fine-tuning LLaMA 2 with Reddit-sourced responses reflecting 11 demographic perspectives, they created an MoE model that lets users select which bias to reflect in generated responses. Although this model showcases how instruction tuning can surface and control for bias, the use of Reddit introduces new sources of bias, raising questions about the underlying training data and the broader ethical implications. In another line of work, Itzhak et al. [10] investigated whether instruction tuning and reinforcement learning from human feedback (RLHF) introduce cognitive biases into models. Focusing on decision-making fallacies like the decoy effect and belief bias, they found that instruction-tuned models (e.g., GPT-3.5, Flan-T5) consistently exhibited stronger cognitive bias than their pre-trained versions. This suggests that while instruction tuning improves usability and safety, it may simultaneously amplify latent biases or introduce new ones. Shifting focus from individual models, Guo et al. [7] conducted a comprehensive review of bias mitigation techniques applied at various stages of model training and deployment. They evaluated strategies ranging from counterfactual data augmentation to causal prompting, and categorised interventions into pre-, intra-, and post-model stages. Their work underscores the growing toolkit available for mitigating bias but also highlights the lack of consistent standards for evaluating these techniques. Together, these studies show that while instruction tuning offers powerful alignment capabilities, it can also deepen or reshape existing biases. Notably, few works investigate whether instruction tuning suppresses stereotypes only at the output level or whether it truly modifies the model's internal representations, precisely the gap this thesis aims to address by comparing embedding- and prompt-level bias before and after instruction tuning.

2.4 Benchmarking and Evaluation Methods

Recent work has focused on developing and comparing evaluation frameworks for measuring gender bias in LLMs, particularly as

[1] ELMo (Embeddings from Language Models) is a deep contextualized word representation model that generates word embeddings based on the entire context in which a word appears, capturing both syntactic and semantic information [20].

it affects downstream task performance. Kaneko et al. [11] introduced the Multi-Step Gender Bias Reasoning (MGBR) benchmark, designed for unscalable reasoning tasks. MGBR evaluates gender bias by counting gendered words in model outputs and aims to capture unconscious social bias in multi-step reasoning. Their findings showed that Chain-of-Thought (CoT) prompting helped reduce bias by encouraging models to reflect on their reasoning. They also found that MGBR correlated strongly with extrinsic bias metrics, such as the Bias Benchmark for Question Answering (BBQ)[2] and the Bias Benchmark for Natural Language Inference (BNLI)[3], but not with intrinsic ones like CrowS-Pairs [18] and StereoSet [17]. This suggests MGBR may better capture biases that affect real-world tasks. Kumar et al. [15] addressed bias evaluation by introducing automated methods, including using an attacker LLM (Llama3) to generate adversarial prompts to elicit biased responses and employing LLMs (GPT-4o) as judges to identify and measure bias (LLM-as-a-Judge). They also analysed existing automatic evaluation metrics like Perspective API and Regard, comparing them to human evaluations. Kumar et al. concluded that their LLM-as-a-Judge approach, which employs GPT-4o to evaluate model outputs, showed the strongest alignment with human judgments for detecting gender bias in LLMs. In contrast, their analysis of existing automated metrics, such as Perspective API and Regard, highlighted notable limitations in how well these tools capture bias. Perspective API occasionally assigned higher toxicity scores to female-oriented responses, while Regard's sentiment-based scoring often failed to reflect the context-dependent nature of gender bias. LLM-as-a-Judge offered a more comprehensive and context-aware evaluation.

These studies reflect ongoing efforts to build more robust and context-sensitive bias evaluation pipelines.

3 METHODOLOGY

3.1 Dataset Construction

The study will utilize a merged corpus of two established datasets to capture a broad spectrum of gender stereotypes:

- **GEST**: The Gender Stereotype (GEST) [13] dataset consists of 3,565 user-generated sentences, each labeled with one of 16 distinct gender stereotype categories (e.g., "Men are strong," "Women are submissive").
- **StereoSet (gender subset)**: StereoSet [17] measures stereotypical biases across gender, profession, race, and religion. This research focuses on the gender subset, comprising 726 contextual triples: stereotype, anti-stereotype, and unrelated continuations.

For robustness, 242 gender-stereotypical sentences from StereoSet will be manually mapped to one of GEST's 16 categories. The mappings will be conducted by the researcher and reviewed by the project supervisor, a domain expert in linguistics and LLMs. The two datasets will be merged to increase phrasal and contextual diversity within each stereotype category. GEST provides more formal, declarative stereotype statements, while StereoSet contributes some contextual and possibly conversational instances; together they cover a wider spectrum of ways a stereotype might appear. For consistency, sentences from the combined dataset will be used both as prompts for LLM completions and as inputs for embedding extraction.

3.2 Model Selection

GPT-2 [4] will serve as the baseline model, as it is not instruction-tuned and offers a well-documented, open-source architecture. For comparison, a LLaMA variant [6] (LLaMA 2 or LLaMA 3) will be used due to its instruction-tuned capabilities and the availability of its model weights. One or more additional open-source models will be incorporated to broaden the scope of the analysis, depending on time and computational resources.

3.3 Embedding and Prompt-Based Analysis

Embedding-Based: To examine how gender stereotypes are internally encoded by different LLMs, contextual embeddings will be extracted from a subset of sentences in the merged dataset (e.g., "She is pretty and kind"). Each sentence will be analyzed on two levels using cosine similarity [21]: pronoun–adjective pairs and pronoun–phrase pairs.

- **At the token level:** Cosine similarity will be computed between embeddings of gendered pronouns (e.g., "he," "she") and stereotype-relevant adjectives (e.g., "kind," "pretty"). These associations will help quantify the degree to which models encode gendered connotations for individual traits.
- **At the phrase level:** Cosine similarity will also be calculated between full sentence embeddings (e.g., "She is pretty and kind" vs "He is pretty and kind"). Sentence representations will be pooled using either average pooling or [CLS] token embeddings, depending on the model's architecture.

These measurements will allow for a detailed analysis of how closely gendered terms are associated with specific stereotypes across models. Embedding results will be aggregated by stereotype category to explore whether certain biases are more deeply encoded and how they vary with instruction tuning.

Prompt-Based: To examine how gender bias manifests in model outputs, the same subset of sentences used in the embedding analysis will be reformulated into prompts by masking the gendered subject (e.g., "She is pretty and kind" becomes "___ is pretty and kind"). These prompts will be submitted to each model (GPT-2, LLaMA, etc.), which will then generate a completion. The completions will be evaluated using rule-based scoring and human annotation.

Rule-Based Scoring: Completions will be assigned a bias score of 1 (stereotypical), 0 (neutral), or -1 (anti-stereotypical), based on their alignment with one of the 16 predefined stereotype categories. Scores will be aggregated by model and stereotype category to identify patterns in bias expression. Cross-validation will be applied to ensure robustness.

Human Annotation: To address subjectivity, the completions will be rated by 20 annotators—10 with data science background and 10 from other backgrounds. Annotators will follow a set of predefined

[2]The Bias Benchmark for Question Answering (BBQ) is a hand-crafted dataset designed to evaluate social biases in QA systems by testing whether models prefer stereotypical over anti-stereotypical answers across multiple dimensions such as gender, race, and nationality [19].

[3]The Bias Benchmark for Natural Language Inference (BNLI) is a dataset for evaluating social biases in NLI models across three identity categories: gender, race, and religion. It measures whether models favor biased inferences in entailment classification tasks [9].

guidelines, and final labels will be determined by majority consensus. These annotations will serve both as an evaluation subset and as a means to validate the consistency of the rule-based scoring scheme.

3.4 Comparative Analysis and Tools

This study will compare how gender bias manifests across embedding-level and output-level representations in different LLMs, with a particular focus on the effects of instruction tuning. Bias differences between LLaMA and other open-source models, with GPT-2 serving as the baseline, will be analysed at both levels. The analysis will explore which of the 16 stereotype categories—such as emotional traits or professional roles—are more likely to be internally encoded or externally expressed by each model. Anti-stereotypical and unrelated examples from StereoSet will be included as control inputs to test contrastive behaviour. Cross-validation will be used to ensure robustness and reduce overfitting to specific examples.

To assess the relationship between internal and external bias, the study will measure categorical alignment between prompt-based and embedding-based results. Cosine similarity scores from the embedding analysis will be thresholded using percentile-based cutoffs and mapped to a three-point scale: 1 (stereotypical), 0 (neutral), or -1 (anti-stereotypical). This enables direct comparability with prompt-based scores. Pearson correlation [3] and other statistical metrics may be used to quantify the degree of alignment between the two evaluation levels.

Implementation will rely on PyTorch, Hugging Face Transformers, and scikit-learn, with visualizations such as heatmaps, UMAP/t-SNE plots, and bar charts to illustrate patterns of stereotype encoding and expression.

4 RISK ASSESSMENT

The project presents several potential risks that could impact its execution. These risks are categorised into dataset biases and limitations, computational constraints and model compatibility. A mitigation plan is outlined for each risk.

Dataset Biases and Limitations: The merged dataset (GEST and StereoSet) may not represent all gender stereotype categories equally, potentially resulting in imbalanced coverage or the underrepresentation of certain themes. To address this,category-balancing strategies will be applied during data preparation. This approach ensures that each of the 16 stereotype categories is represented by a comparable number of examples across models and evaluation tasks.

Computational Constraints: Extracting contextual embeddings and generating completions across multiple LLMs can be computationally intensive. To address this, I will use my Google Colab Pro subscription to access GPU-enabled runtimes, which will accelerate embedding extraction and inference processes.

Model Compatibility and Representation Layer Access: Different LLM architectures vary in how accessible and interpretable their internal representations are, particularly when extracting contextual embeddings. For instance, not all models provide clear access to final hidden states or use the same pooling strategies (e.g., [CLS] token vs. mean pooling). This could complicate the consistency and comparability of embedding-based analyses across models. To mitigate this risk, I will carefully review model documentation and select models with accessible, well-documented APIs via Hugging Face. In addition, I will standardise representation extraction methods across models as much as possible to ensure methodological consistency.

5 PROJECT PLAN

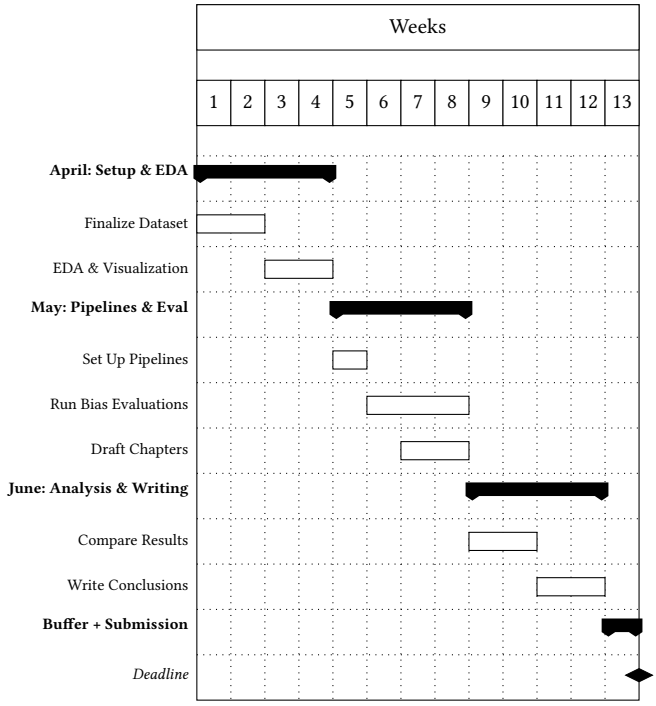
Although I had to restart my thesis project in April due to unforeseen circumstances, I am now working on it full time and believe I might still be able to meet the original deadline.

I have already finalized the merged dataset, including the manual mapping of 242 gender-stereotypical sentences from StereoSet into GEST's 16 stereotype categories. The remainder of April will be dedicated to exploratory data analysis (EDA) to examine the distributional properties, detect imbalances across categories, and generate visualizations that will guide subsequent model analysis.

In May, I will set up and test pipelines for both embedding extraction and prompt-based generation using GPT-2, LLaMA variant and additional LLMs. The month will focus on conducting both embedding-based and prompt-based analysis and evaluations across the selected models. Throughout this period, I will also document methodological decisions and begin drafting the core analytical chapters of the thesis.

June will be dedicated to comparative analysis, synthesizing my findings from both embedding-level and output-level evaluations. Particular attention will be given to divergences between models and the influence of instruction tuning. I will also refine visualizations, formalize conclusions, and continue writing, incorporating feedback from my supervisor and iterating on earlier drafts.

The official deadline remains June 27. However, I have been granted an extension, allowing the first week of July as a buffer to finalize the writing, polish formatting, and ensure a complete and well-rounded submission.



REFERENCES

[1] Christine Basta, Marta R. Costa-jussà, and Noe Casas. 2019. Evaluating the Underlying Gender Bias in Contextualized Word Embeddings. arXiv:1904.08783 [cs.CL] https://arxiv.org/abs/1904.08783

[2] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. 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)*. ACM, 610–623. https://doi.org/10.1145/3442188.3445922

[3] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson Correlation Coefficient. In *Noise Reduction in Speech Processing*. Springer, 1–4. https://doi.org/10.1007/978-3-642-00296-0_5

[4] OpenAI Community. 2023. GPT-2 Model on Hugging Face. https://huggingface.co/openai-community/gpt2. Accessed: 2025-04-22.

[5] Fangyu Dong, Pavan Ammanamanchi, Ying Zhang, Hwaran Kim, Sebastian Riedel, and Pontus Stenetorp. 2024. Disclosure and Mitigation of Gender Bias in LLMs. arXiv:2402.11190 [cs.CL] https://arxiv.org/abs/2402.11190

[6] Hugging Face. 2023. LLaMA 2 on Hugging Face: State-of-the-art open-access language models. https://huggingface.co/blog/llama2. Accessed: 2025-04-22.

[7] Yufei Guo, Muzhe Guo, Juntao Su, Zhou Yang, Mengqiu Zhu, Hongfei Li, Mengyang Qiu, and Shuo Shuo Liu. 2024. Bias in Large Language Models: Origin, Evaluation, and Mitigation. arXiv:2411.10915 [cs.CL] https://arxiv.org/abs/2411.10915

[8] Patrick Haller, Ansar Aynetdinov, and Alan Akbik. 2023. OpinionGPT: Modelling Explicit Biases in Instruction-Tuned LLMs. arXiv:2309.03876 [cs.CL] https://arxiv.org/abs/2309.03876

[9] Jason Huang, Ximing Lu, Zijian Liu, Yiwei Sun, Hai Zhao, Rajiv Shah, and Steven C.H. Hoi. 2023. BBNLI: A Bias Benchmark for Natural Language Inference. arXiv:2305.12620 [cs.CL] https://arxiv.org/abs/2305.12620

[10] Itay Itzhak, Gabriel Stanovsky, Nir Rosenfeld, and Yonatan Belinkov. 2024. Instructed to Bias: Instruction-Tuned Language Models Exhibit Emergent Cognitive Bias. arXiv:2308.00225 [cs.AI] https://arxiv.org/abs/2308.00225

[11] Masahiro Kaneko, Danushka Bollegala, Naoaki Okazaki, and Timothy Baldwin. 2024. Evaluating Gender Bias in Large Language Models via Chain-of-Thought Prompting. arXiv:2401.15585 [cs.CL] https://arxiv.org/abs/2401.15585

[12] Shirin R. Kapoor and Arvind Narayanan. 2023. Quantifying ChatGPT's Gender Bias. https://www.aisnakeoil.com/p/quantifying-chatgpts-gender-bias AI Snake Oil Blog, accessed April 17, 2025.

[13] Kinit. 2023. GEST - Gender Stereotype Dataset. https://huggingface.co/datasets/kinit/gest. Accessed: 2025-04-22.

[14] Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender Bias and Stereotypes in Large Language Models. In *Proceedings of The ACM Collective Intelligence Conference (CI '23)*. Association for Computing Machinery, Delft, Netherlands, 12–24. https://doi.org/10.1145/3582269.3615599

[15] Shachi H Kumar, Saurav Sahay, Sahisnu Mazumder, Eda Okur, Ramesh Manuvinakurike, Nicole Beckage, Hsuan Su, Hung yi Lee, and Lama Nachman. 2024. Decoding Biases: Automated Methods and LLM Judges for Gender Bias Detection in Language Models. arXiv:2408.03907 [cs.CL] https://arxiv.org/abs/2408.03907

[16] Swapnil Mishra, Preksha Nagaraj, Anandhavel Subramanian, Anupam Jamatia, Anudeep Tummalapenta, Bharathi Raja Chakravarthi, and Ritesh Kumar. 2023. GEST: Gender Stereotype Dataset. https://huggingface.co/datasets/kinit/gest. Accessed: 2025-04-17.

[17] Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*. 5356–5371. https://arxiv.org/abs/2004.09456

[18] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 1953–1967. https://doi.org/10.18653/v1/2020.emnlp-main.154

[19] Alicia Parrish, Paul Röttger, Margaret Smith, Nithum Thain, Lucas Dixon, Jack Sorensen, Ben Hutchinson, Kellie Webster, Prasanna Jagadeesan, Hailey Devaney, et al. 2022. BBQ: A Hand-Built Bias Benchmark for Question Answering. In *Findings of the Association for Computational Linguistics: ACL 2022*. 2161–2179. https://aclanthology.org/2022.findings-acl.165

[20] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*. 2227–2237.

[21] Gerard Salton and Michael J. McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill.

[22] Aleix Sant, Carlos Escolano, Audrey Mash, Francesca De Luca Fornaciari, and Maite Melero. 2024. The power of Prompts: Evaluating and Mitigating Gender Bias in MT with LLMs. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, Agnieszka Faleńska, Christine Basta, Marta Costa-jussà, Seraphina Goldfarb-Tarrant, and Debora Nozza (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 94–139. https://doi.org/10.18653/v1/2024.gebnlp-1.7

[23] Carolin M. Schuster, Maria-Alexandra Dinisor, Shashwat Ghatiwala, and Georg Groh. 2025. Profiling Bias in LLMs: Stereotype Dimensions in Contextual Word Embeddings. arXiv:2411.16527 [cs.CL] https://arxiv.org/abs/2411.16527

[24] Zeping Yu and Sophia Ananiadou. 2025. Understanding and Mitigating Gender Bias in LLMs via Interpretable Neuron Editing. arXiv:2501.14457 [cs.CL] https://arxiv.org/abs/2501.14457

[25] Yachao Zhao, Bo Wang, Yan Wang, Dongming Zhao, Xiaojia Jin, Jijun Zhang, Ruifang He, and Yuexian Hou. 2024. A Comparative Study of Explicit and Implicit Gender Biases in Large Language Models via Self-evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (Eds.). ELRA and ICCL, Torino, Italia, 186–198. https://aclanthology.org/2024.lrec-main.17/