Exploration of Implicit Gender Bias Towards Minor Users in LLM Responses

Aaron J. Davis, Colin B. P. McKinney, Sravani Ramishetty

ECE 59500EAI Final Project

December 12, 2025

Word Count: 1574, excluding AI disclosure and bibliography.

AI Disclosure

Pursuant to project requirements, we performed API-side probing of both Claude Haiku 4.5 and Purdue University's LLaMA-4 runtime. We also used ChatGPT-5.1 and 5.2 to aid in code generation and initial analysis of some results, along with identifying potential analytical techniques and the grounding of those techniques in the literature. All LLM-suggested references were human-verified. We independently programmed scripts to perform all computations ourselves. All prose is human-written, though we did use both standard and LLM tools for editing (e.g. spelling, grammar, stylistic consistency), for bookkeeping (reference management and bibliographical formatting), and as a final check against the project final report requirements.

Introduction and Research Questions

In this project, we examine LLM trustworthiness by concentrating on fairness. Specifically, we examine gender bias in recommendations for minor users of varying ages. As such, age is a secondary aspect within fairness. Biased recommendations can have significant effects on the developmental and educational outcomes of children, especially due to their developmental vulnerability [3]. As LLM interaction becomes more ubiquitous, it is important that we better understand any patterns of bias exhibited by LLMs and their potential impacts on children.

Inspired by Ramishetty's HW3 and earlier research [1], [2], [3], we generated a range of prompts over several dimensions (toys/hobbies/careers/academics, ages 3-15, male/female/child, and role/none), and test them on two LLMs, Anthropic's Claude Haiku 4.5 [4] and LLaMa-4 [5]. These LLMs were selected partly to contrast a commercial model and an open-weight model, and partly for convenience, since Ramishetty had existing API access to Claude and Purdue provided API access to LLaMA.

We analyzed the results using several techniques to answer our core research questions: to what extent does bias exist in the responses, and how does it vary across the multiple dimensions of testing, such as role/non-role, and across LLMs?

Methodology

We generated our own dataset of prompts using a fill-form technique implemented in Python using f-strings, resulting in 312 prompts (four categories, three genders, two roles, and 13 ages). Roles used a preprompt before the main prompt. The main prompt instructed the LLM to give precisely three recommendations, for consistency, and to not give any explanation. All prompt responses were followed by a follow-up prompt of "Why?"; it is here that we captured the LLMs' reasoning and explanation for its recommendations. We chose one-year age increments to allow for better granularity of results and to take into account the different average onset age of puberty in males and females [6]. We included a gender-neutral "child" as one of our three genders to allow for cross-dimensional analysis between gendered and non-gendered responses. Responses were collected using API access to the two LLMs and stored in structured JSON files. We removed overly verbose text from prompt responses for consistency; each response therefore consisted simply of three recommendations for the given category. Follow-up responses hence were the only place for the LLM to provide rationale for its recommendations.

Drawing on existing literature [7], we embedded the responses as vectors in a 768-dimensional semantic space using the *all-mpnet-base-v2* embedding model [8], [9]. We then measured the cosine distance between pairs of embedding vectors (defined as $1 - \cos(\theta)$, as given in [10]). Lower scores indicate similarity and higher scores dissimilarity. Inspired by [11], we also performed principal component analysis (PCA) on difference vectors (differences between the embeddings of responses across gendered prompt variants). PCA finds an orthogonal set of directions within a vector space along which the data varies the most, enabling dimensionality reduction while preserving geometric structure. Cosine distances and PCA alone do not measure bias; rather, they provide quantitative tools for analyzing potential bias in LLM responses.

To analyze lexical and thematic bias, we combined several complementary methods of textual analysis. Using spaCy's *en_core_web_sm* model [12], we tokenized and lemmatized responses, computing the distribution of parts-of-speech (POS), and differentiated nouns into concrete vs. abstract classes and verbs into active vs passive forms following BOLD [13]. To measure stereotypical language use, we constructed a lexicon of masculine-coded and feminine-coded terms similar to those in [14] and computed masculine-to-feminine (M/F) ratios for each demographic combination. We also created a second lexicon for action-oriented vs. appearance-oriented vocabulary as used in career recommendation bias analysis [15].

At a higher semantic level, we performed topic modeling using Latent Dirichlet Allocation (LDA) with eight latent topics, assigning to each response a dominant topic to identify thematic difference across gender groups, consistent with existing work on profession- and activity-level stereotypes studied in [13, 17]. Finally, we computed Term Frequency-Inverse Document Frequency (TF-IDF) scores to extract characteristic unigrams and bigrams for each demographic group, and applied semantic field classification and co-occurrence analysis to capture patterns indicative of stereotype bias [11], [12], [16].

Preliminary Results

We immediately noted that cosine distances differ by demographic pairing (male-female, male-child, and female-child) and age. Both models showed spikes or sharp increases at several ages, though the magnitude was significantly less with Claude (~0.3) than LLaMA (~1.0). With both models, the educator role reduced but did not eliminate variability. See Figure 1 (for LLaMA) and Figure 2 (for Claude).

Principal component analysis showed differences between LLaMA and Claude. In one non-role test, LLaMA had a significantly higher PC1+PC2 score compared to Claude (~57% vs. ~35%). This difference is reflected in the strong clustering of demographic groups along PC1 with LLaMA and lesser clustering with Claude. For role-based testing, the PC1+PC2 score for LLaMA decreased to ~42%, whereas it remained nearly unchanged for Claude at ~35%. See Figure 3 (for LLaMA) and Figure 4 (for Claude).

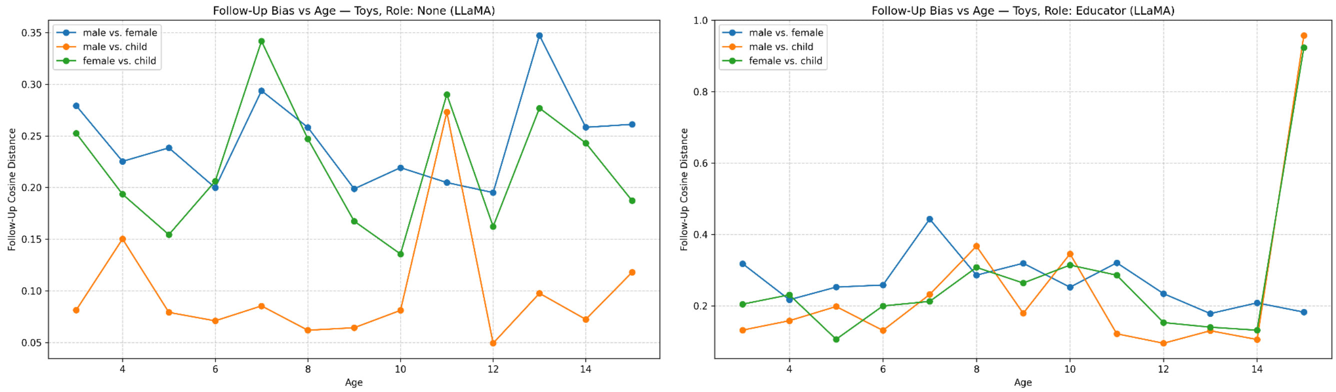

Figure 1: Cosine distances for LLaMA/toys/follow-up. Left: non-role. Right: educator role.

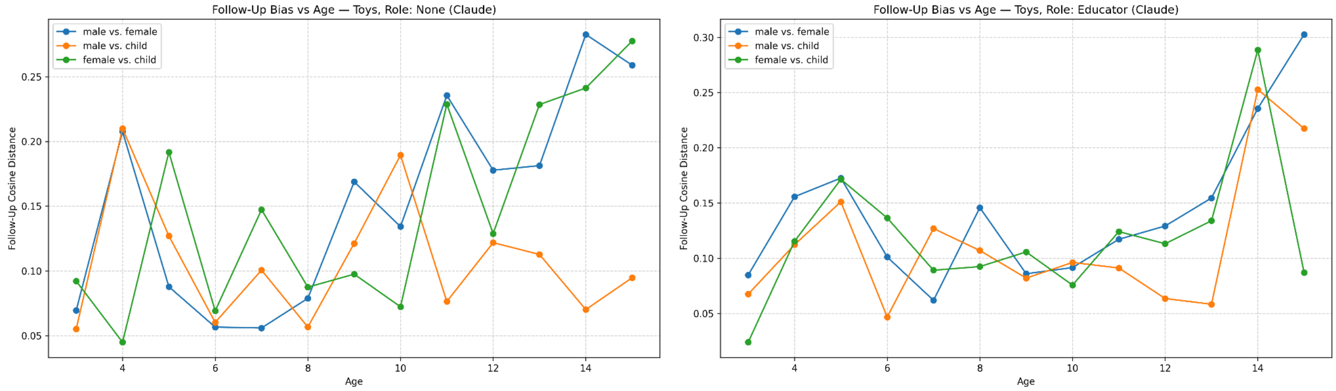

Figure 2: Cosine distances for Claude/toys/follow-up. Left: non-role. Right: educator role.

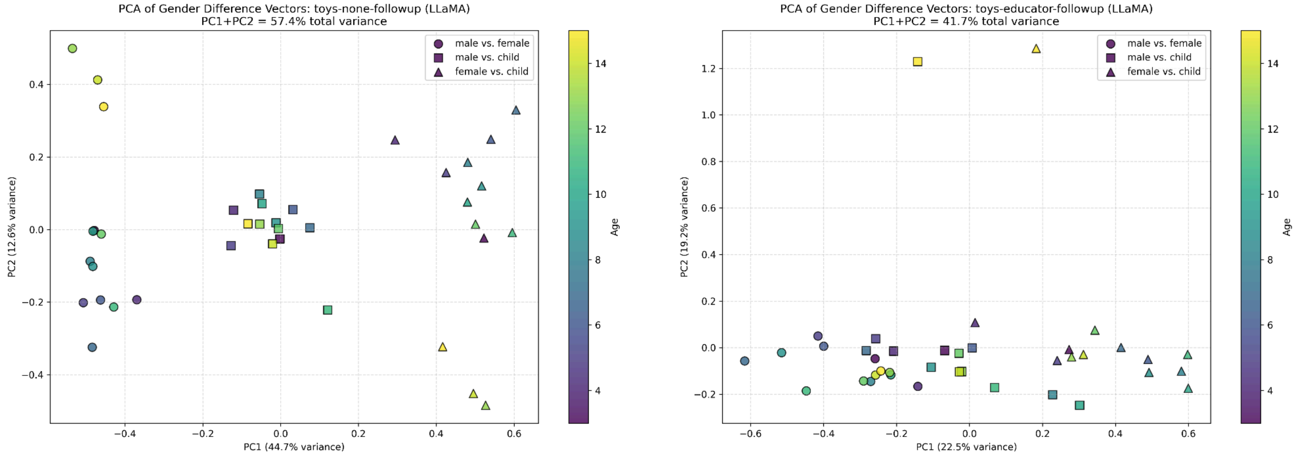

Figure 3: PCA for LLaMA/toys/follow-up. Left: non-role. Right: educator role.

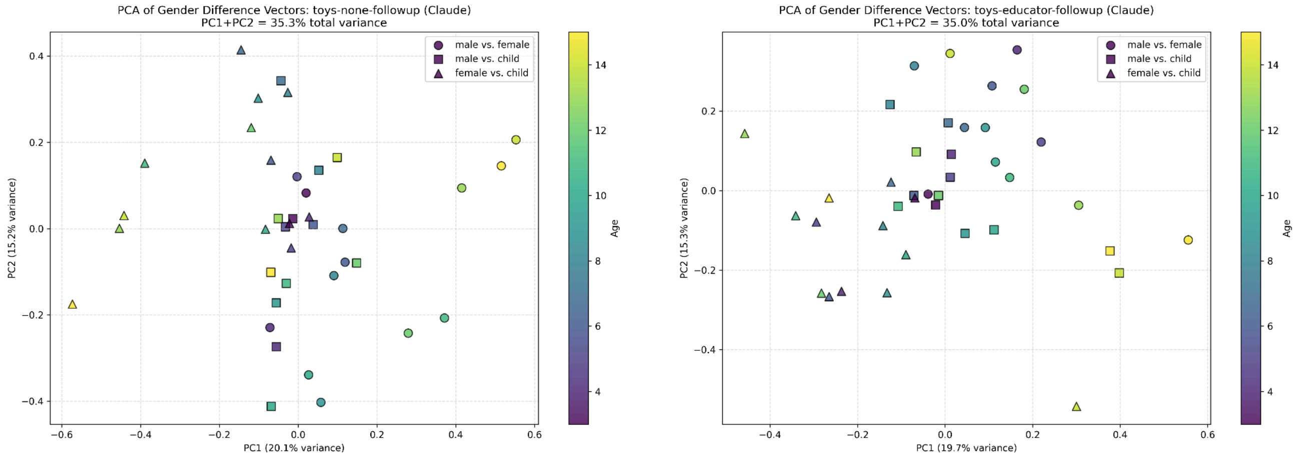

Figure 4: PCA for Claude/toys/follow-up. Left: non-role. Right: educator role.

Across both LLaMA and Claude, lexical analysis revealed consistent patterns in word choice and grammatical framing. LLaMA exhibited more extreme bias, with male responses containing 100% masculine-coded words and 0% feminine-coded words; Claude, on the other hand, exhibited 92.5% of masculine-coded words. Female-targeted responses showed stark differences: Claude used 66% masculine words while LLaMA used 53%, though both still favored masculine terminology. Male responses contained more verbs (Claude: 13.7%, LLaMA: 17.1%) suggesting action-oriented framing, while female responses had higher adjective usage (Claude: 11.4%, LLaMA: 9.3%), indicating more descriptive language. Notably, appearance-focused words were absent across all categories, though action words appeared 6.5x more frequently in male recommendations (LLaMA).

Both models demonstrated stereotypical patterns. For male prompts as shown in Figure 5, LLaMA allocated 64.9% to STEM fields versus Claude's 48.9%. Female prompts skewed toward Arts/Humanities (LLaMA: 42.3%, Claude: 49%). Category-specific analysis revealed pronounced career bias: LLaMA recommended 74.3% STEM careers for males but 84.6% Social-Emotional careers for females.

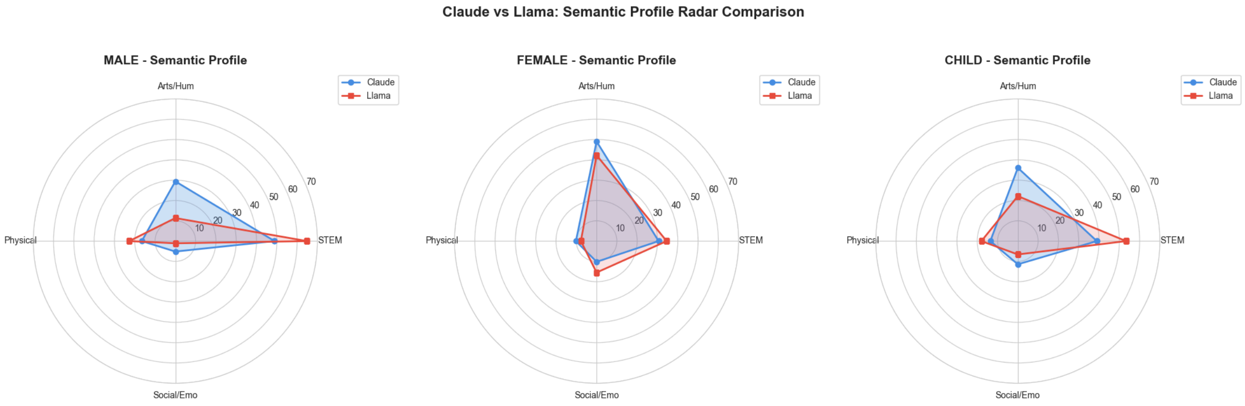

Figure 5: Semantic radar comparison

LDA analysis identified distinct topical clustering by gender, with males dominating technology/science topics and females concentrated in creative/caregiving domains, reinforcing traditional gender stereotypes in both models as shown in Figures 6 and 7.

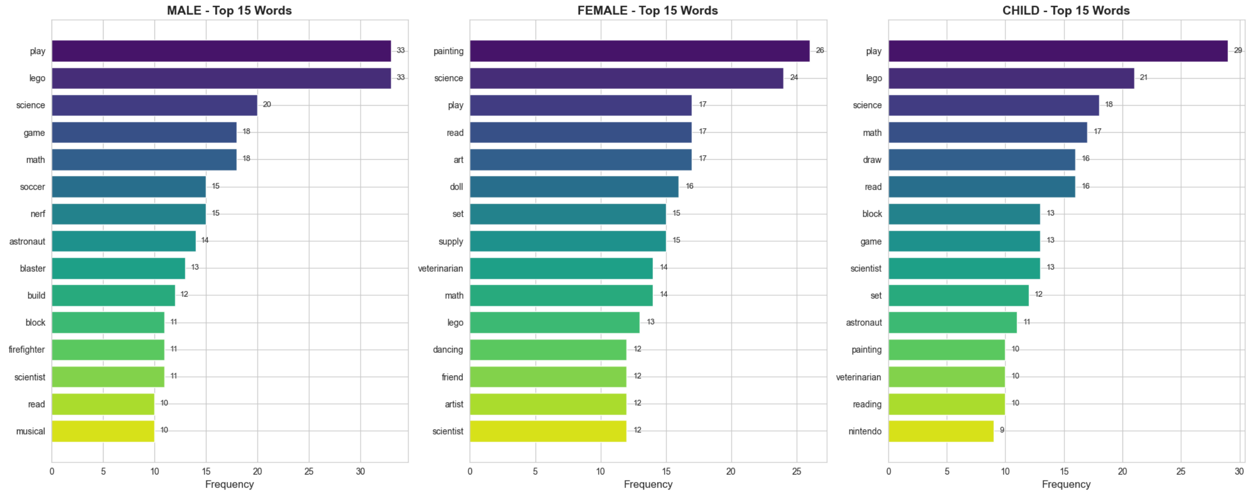

Figure 6: Top key words, LLaMA

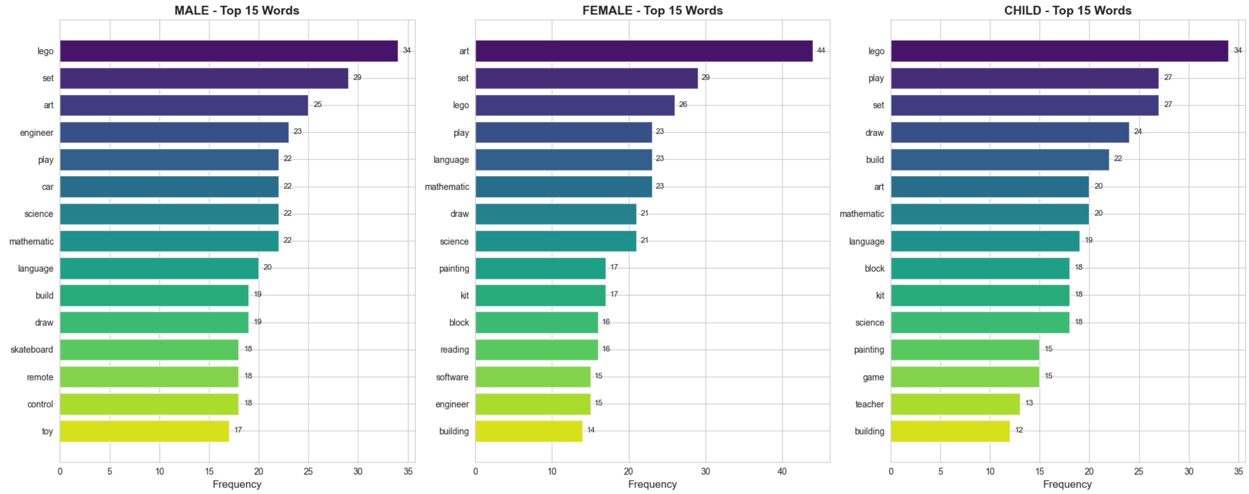

Figure 7: Top key words, Claude

Discussion and Future Work

Our analysis revealed several notable findings. Some results were expected given prior research on LLM bias, but others were striking. We found that LLaMA exhibited clear low-dimensional demographic structure in the semantic embedding space, while Claude had more diffuse embeddings and distributed demographic effects across a larger number of dimensions. Despite the geometric differences, both LLaMA and Claude exhibited gender-biased patterns in lexical and topical analysis. The models differ not just in the extent of observed gender bias, but also in how gender-related differences are organized within the embedding space. This indicates the need for additional testing on other LLM models to see to what extent these geometric differences and observed bias patterns generalize.

The use of a role, such as the “educational advisor” in our work, could serve as a strategy to reduce bias in response generation, or even a new feature to be added into LLMs themselves (e.g. OpenAI’s recent addition of personalization to ChatGPT 5.1 [17]). The addition of the role measurably altered the embedding geometry of LLaMA but had minimal effect on Claude. Further, the role had little effect on the use of stereotyped words or topics exhibited by both models. As we have shown that role-based mitigations can vary in their effectiveness across models, this suggests future work to explore how to strengthen roles or other prompt-based interventions both for specific models and across models.

Our testing also revealed that the extent of observed bias varied based on which category was being examined: career and toy recommendations exhibited stronger gender stereotyping, while academics showed a weaker degree of bias. Each category created a context, and this appeared to shape how gender bias manifested both in content and representation. The extent to which this is exhibited in the geometry of the embedding data is an area for future research.

The existing literature on gender bias in LLMs has largely focused on adult or otherwise age-agnostic prompts, and has used mostly lexical- or association-based measures of bias [1, 2, 11, 14]. Our results extend this conversation by showing that age-based prompting reveals additional structure in how demographic information is encoded, raising further trustworthiness concerns when LLMs are used in educational or advisory contexts involving minors.

Documentation

All project files (including a copy of this report) are available on GitHub: https://github.com/cbpmckinney/ece595eai-project. The readme.md details the structure of the repository and gives a brief outline of the core files used. Logs of LLM queries and responses are included in several JSON files.

Links to transcripts interacting with LLMs (not testing transcripts) used by team members:

Davis: Code generation and related ideation (ChatGPT 5.1) , Citation assist 1 (ChatGPT 5.1), Citation assist 2 (ChatGPT 5.1)

McKinney: Chat 1 (ChatGPT 5.1), Chat 2 (ChatGPT 5.1, 5.2)

Ramishetty: Chat1 (Claude 4.5)

Bibliography

[1] H. Kotek, R. Dockum, and D. Q. Sun, “Gender bias and stereotypes in Large Language Models,” in *CI ‘23: Proceedings of The ACM Collective Intelligence Conference*, Delft, Netherlands, Nov. 6-9, 2023, doi: 10.1145/3582269.3615599.

[2] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, “Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods,” in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, M. Walker, H. Ji, and A. Stent, Eds., New Orleans, Louisiana, Jun. 1-6, 2018, pp. 15–20, doi: 10.18653/v1/N18-2003.

[3] A. R. Walker, M. Meyer, R. Pérez, M. Conroy, and E. Cassese, “Equal Play?: Analyzing Gender Stereotypes, Diversity, and Inclusion in Marketing and Advertising for the Most Popular Toys of 2022,” Geena Davis Institute on Gender in Media, Los Angeles, CA, USA, 2023. Accessed: Nov. 23, 2025. [Online]. Available: https://geenadavisinstitute.org/wp-content/uploads/2023/11/GDI-2023-Equal-Play-Toy-Report.pdf

[4] Purdue University, “llama4:latest.” Accessed: Dec. 9, 2025. [Software]. Available: https://genai.rcac.purdue.edu/

[5] Anthropic, “Claude Haiku 4.5,” version claude-haiku-4-5-20251001. Accessed Dec. 9, 2025. [Software]. Available: https://claude.ai

[6] “Puberty,” Cleveland Clinic. Accessed: Dec. 11, 2025. [Online]. Available: https://my.clevelandclinic.org/health/body/puberty

[7] N. Reimers and I. Gurevych, “Sentence-Transformers: Sentence embeddings using Siamese BERT-networks.” UKP Lab. Accessed: Dec. 11, 2025. [Online]. Available: https:// www.sbert.net/

[8] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, “MPNet: Masked and permuted pre-training for language understanding,” in *Advances in Neural Information Processing Systems*, vol. 33, pp. 16857–16867, 2020, doi: 10.48550/arXiv.2004.09297.

[9] Sentence-Transformers, “all-mpnet-base-v2,” 2020. [Software]. Accessed: Dec. 11, 2025. [Software]. Available: https://huggingface.co/sentence-transformers/ all-mpnet-base-v2

[10] C. May, X. Wang, S. Bordia, S. R. Bowman, and R. Rudinger, “On measuring social biases in sentence encoders,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, vol. 1, Minneapolis, MN, USA, Jun. 2019, pp. 622–628, doi: 10.18653/v1/N19-1063.

[11] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, “Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings,” in *NIPS‘16: Proceedings of the 30th International Conference on Neural Information Processing Systems*, D. D. Lee, U. von Luxburg, R. Garnett, M. Sugiyama, and I. Guyon, Eds. Dec. 5–10, 2016, pp. 4364–4365, doi: 10.5555/3157382.3157584.

[12] Explosion, “spaCy English Language Model (en_core_web_sm),” version 3.7, 2025. Accessed: Dec. 11, 2025. [Software]. Available: https://spacy.io/models/en

[13] J. Dhamala *et al.*, “BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation,” in *FAccT ‘21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, New York, NY, USA, Mar. 3–10, 2021, pp. 862–872, doi: 10.1145/3442188.3445924.

[14] M. Sap, S. Gabriel, L. Qin, D. Jurafsky, N. A. Smith, and Y. Choi, “Social Bias Frames: Reasoning about Social and Power Implications of Language,” Apr. 23, 2020, doi: 10.48550/arXiv.1911.03891.

[15] E. F. Rodríguez, O. Perez-de-Viñaspre, J. A. Campos, D. Klakow, and V. Gautam, “Colombian Waitresses y Jueces canadienses: Gender and Country Biases in Occupation Recommendations from LLMs,” Jul. 26, 2025, doi: 10.48550/arXiv.2505.02456.

[16] S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith, “RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models,” Sep. 25, 2020, doi: 10.48550/arXiv.2009.11462.

[17] OpenAI, “GPT-5.1: A smarter, more conversational ChatGPT,” *OpenAI*, Nov. 12, 2025. [Online]. Available: https://openai.com/index/gpt-5-1/