Evaluating Bias Detection in Lightweight LLMs

Veronika Bryskina, Milita Songailaitė, Justina Mandravickaitė

Vytautas Magnus University,
Universiteto str. 10, 53361 Akademija
veronika.bryskina@vdu.lt, milita.songailaite@vdu.lt,
justina.mandravickaite@vdu.lt

Abstract. This study evaluates the ability of lightweight open-source large language models (LLMs) to detect bias in text. Eleven models of six popular LLM families were tested in a zero-shot setting on a unified dataset of 8,745 sentences derived from three selected sources, covering gender, race, religion, and appearance bias. Results showed that none of the models exceeded 70% accuracy, which highlighted limitations of lightweight LLMs and existing challenges related to current bias detection datasets.

Keywords: bias detection, LLM, benchmarking, open-source LLMs, evaluation.

1 Introduction

For the longest time, human-based methods of data evaluation have remained one of the most effective methods of bias identification [1]. But the amount of time and labor necessary to process vast datasets in resource-restricted research environments, calls for more automated approach to data annotation. Thus, the goal of this research is to evaluate the performance of popular open-source models on selected bias detection benchmarks in a zero-shot setting.

Bias refers to information representation that supports a specific viewpoint, which can shape how people understand or interpret events or issues [3]. Bias detection includes variety of methods, some of them for reduction and migration of previously identified bias [6], other times research leans to bias detection, e.g., testing bias detection on MBIB [3] dataset using GPT-3.5 and multiple fine-tuned models [2], or identifying bias in clinical notes [4], in which GPT-4 evaluated and categorized biased language in health records. To improve bias detection, a diverse collection of datasets was assembled over time, for various methods and bias dimensions. For example, RedditBias [5] represents actual human conversations from

DOI: https://doi.org/10.15388/LMITT.2026.4

Reddit that contain racial, religious, gender and queerness bias, while MBIB [3] consolidates 22 bias datasets, all of which are valuable resources in media bias identification.

2 Data and Methodology

The proposed methodology for study of bias detection evaluation includes three datasets (*Crows-Pairs* [7], *Grep-BiasIR* [8] and *StereoSet* [9]), that represent four bias categories (gender, race, religion, appearance) and a selection of the most popular open-source, lightweight LLMs (Table 1)), where lightweight refers to models with lower computational requirements and relatively small parameter sizes. Models used in given research include: *llama3.1* 8b, *gemma3* 4, 12 and 27b, *qwen3* 4, 8 and 24b, *mistral-small3.2* 24b, *ministral-3* 8 and 14b, *gpt-oss* 20b. Datasets were rearranged to include full sentences, unify labeling and drop anti-stereotypes statements as separate category, resulting in 8745 unique sentences, 54.5% of which classified as biased. Models were running on the *Ollama* platform and instructed, with zero-shot prompt, to classify the presented sentence as biased or unbiased and to specify type of bias if present. Evaluation of models' performance was conducted with standard accuracy, precision, recall and F1 score metrics.

3 Results

As we can see in the results (Table 1), *gemma3:4b* shows the best accuracy and F1 score, being balanced towards both classes, and outperforming both 12b and 27b variants. *Llama3.1:8b* follows as the second-best model with very similar metrics. Meanwhile, *qwen* models' performance was the worst. Even though they were getting better with more parameters, accuracy stayed less than 50%. On the other hand, *qwen* models show the best recall which means they tend to classify data as unbiased. *Gpt-oss* model follows closely, passing 50% threshold, but being not much better than simple guessing.

Despite vast improvements, in zero-shot scenario lightweight LLMs, as our study shows, struggle to achieve bias classification with over 70% accuracy rate, at some instances even failing to achieve accuracy similar to a coin flip. Closer look at the misclassified data revealed that examples were often confusing by being too vague, for example, "Jamal made a real athletic play that gave the team their first lead of the game" was labelled

as racial stereotype, and “Bob had a hard time trying to learn how to cook” was labelled as gender stereotype. By themselves those sentences can be biased or just stating the fact, missing necessary contextual information for clearer labelling decisions.

Table 1. Results sorted by accuracy.

Model	Accuracy	Precision	Recall	F1
gemma3:4b	0.679	0.803	0.815	0.809
llama3.1:8b	0.678	0.763	0.860	0.808
mistral-small3.2:24b	0.669	0.751	0.860	0.802
ministral-3:14b	0.661	0.776	0.818	0.796
gemma3:27b	0.649	0.774	0.801	0.787
gemma3:12b	0.617	0.705	0.831	0.763
ministral-3:8b	0.589	0.659	0.847	0.741
gpt-oss:20b	0.523	0.562	0.882	0.687
qwen3:30b	0.498	0.521	0.917	0.665
qwen3:8b	0.425	0.442	0.920	0.597
qwen3:4b	0.303	0.308	0.943	0.465

4 Conclusion

Lightweight open-source LLMs show limited reliability for bias detection, with none of the evaluated models exceeding 70% accuracy and some performing close to random classification. Error analysis also suggests that ambiguous or context-dependent examples in existing datasets may contribute to misclassification. This highlights existing challenges in both model capability and dataset design. While some models show balanced performance, others lean towards classification of all information as unbiased, undermining their practical usability.

References

[1] Chen, Guiming Hardy, et al. Humans or LLMs as the Judge? A Study on Judgement Biases. 2024, https://arxiv.org/abs/2402.10669.

[2] Wen, Zehao, and Rabih Younes. “ChatGPT v.s. Media Bias: A Comparative Study of GPT-3.5 and Fine-Tuned Language Models.” Applied and Computational Engineering, vol. 21, no. 1, EWA Publishing, Oct. 2023, pp. 249–57, doi:10.54254/2755-2721/21/20231153.

[3] Wessel, Martin, et al. “Introducing MBIB - The First Media Bias Identification Benchmark Task and Dataset Collection.” Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2023, pp. 2765–74, https://arxiv.org/abs/2304.13148.
[4] Apakama, Donald U., et al. “Identifying Bias at Scale in Clinical Notes Using Large Language Models.” Mayo Clinic Proceedings: Digital Health, vol. 3, no. 4, 2025, p. 100296, doi:https://doi.org/10.1016/j.mcpdig.2025.100296.
[5] Barikeri, Soumya, et al. RedditBias: A Real-World Resource for Bias Evaluation and Debiasing of Conversational Language Models. 2021, https://arxiv.org/abs/2106.03521.
[6] Schick, Timo, et al. Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP. 2021, https://arxiv.org/abs/2103.00453.
[7] Nangia, Nikita, et al. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. 2020, https://arxiv.org/abs/2010.00133.
[8] Krieg, Klara, et al. Grep-BiasIR: A Dataset for Investigating Gender Representation-Bias in Information Retrieval Results. 2023, https://arxiv.org/abs/2201.07754.
[9] Nadeem, Moin, et al. StereoSet: Measuring Stereotypical Bias in Pretrained Language Models. 2020, https://arxiv.org/abs/2004.09456.