Fable: Fairness and Bias in LLM Evaluation

Mohammad Al-Attas
Student ID: g201513050
King Fahd University of Petroleum and Minerals
Dhahran, Saudi Arabia

Supervised by: Dr. Muzammil Behzad
muzammil.behzad@kfupm.edu.sa
King Fahd University of Petroleum and Minerals
Dhahran, Saudi Arabia

***Abstract*—Fairness and bias in large language models (LLMs) have become critical concerns as these models are increasingly deployed in real-world applications. This paper presents Fable (Fairness And Bias in LLM Evaluation), a comprehensive benchmark for evaluating open-source LLMs using the Ollama framework. We focus on bias across four sensitive categories: Religion, Race, Politics, and Gender. Standardized prompt sets were designed to elicit potential biases, and six recent open-source LLMs (DeepSeek, Gemma3, Granite3, Llama2, Llama3.2, and Qwen2.5) were evaluated. Model outputs were analyzed with automated classifiers (toxic content detection, hate speech identification, and zero-shot NLI) to quantitatively measure biased responses. Our experiments reveal that Political prompts consistently induce the highest rate of biased outputs (up to 36.8% flagged), while Gender and Religion prompts showed virtually no toxic or hateful content. Race-related prompts led to intermediate bias levels. The results highlight progress and remaining challenges in bias mitigation for open LLMs. We discuss the limitations of automated detection, the need for nuanced evaluations, provide recommendations for future evaluations using Ollama, and release an example repository structure to facilitate reproducibility.**

***Index Terms*—Fairness, Bias, Large Language Models (LLMs), Evaluation, Benchmark, Open-Source Models, Ollama, Toxicity Detection, Hate Speech Detection, Political Bias, Gender Bias, Racial Bias, Religious Bias.**

I. Introduction

A. Background and Significance

As large language models (LLMs) achieve widespread use in tasks ranging from content generation to question-answering, concerns have grown about the fairness and biases of these models [1], [2]. Bias in LLMs can manifest as harmful stereotypes or unfair treatment of certain groups, potentially leading to amplified social biases when these models are deployed [1], [2]. These findings underscore the urgent need for systematic evaluation frameworks to detect and measure bias in both closed and open-source LLMs.

B. Challenges in Current Techniques

Recent studies have demonstrated that LLMs often exhibit biases along dimensions such as race, gender, religion, and political orientation [2]. For example, ChatGPT was found to have a significant left-leaning political bias [2], and others noted gender stereotypes in occupation associations [3]. Open-source LLMs provide transparency and customizability but may lack the extensive fine-tuning and safety training of proprietary models, making their evaluation crucial. However, many existing evaluation efforts target only a single type of bias or a narrow set of demographics, using inconsistent metrics, making comparisons difficult [4].

C. Problem Statement

There is a need for a comprehensive, standardized benchmark framework specifically designed for evaluating fairness and bias across multiple sensitive domains (Religion, Race, Politics, Gender) in open-source LLMs. Existing approaches often lack this multi-dimensional scope, consistency in metrics, or focus specifically on the nuances of open-source model evaluation within a reproducible environment like Ollama. This paper introduces Fable to address this gap.

D. Objectives

The contributions of this paper are:

1) A new benchmark framework (Fable) for evaluating fairness and bias in LLMs, with standardized prompts covering multiple bias categories.
2) A thorough evaluation of six open-source LLMs using this benchmark, yielding comparative bias metrics.
3) Visualizations and analysis illustrating bias patterns across models and categories, providing insights into current strengths and weaknesses of open LLMs.
4) Recommendations for future bias evaluations using Ollama and similar tools, encouraging reproducible assessments.
5) A publicly available repository structure to facilitate community use, extension, and critique.

E. Scope of Study

This study focuses on evaluating bias in six recent open-source LLMs (DeepSeek, Gemma3, Granite3, Llama2, Llama3.2, Qwen2.5) across four sensitive categories: Religion, Race, Politics, and Gender. The evaluation uses standardized prompt sets designed for these categories and leverages the Ollama framework for consistent local deployment. Bias is measured quantitatively using automated classifiers for toxic content, hate speech, and zero-shot NLI-based bias detection.

II. Literature Review

A. Overview of Existing Techniques

The rapid adoption of LLMs has led to increased attention on their potential to reflect and amplify societal biases [1]. Surveys like Gallegos et al. (2023) consolidate findings, formalize notions of social bias, and categorize evaluation

approaches at different levels (embeddings, generated text) [1]. They emphasize that LLMs "can learn, perpetuate, and amplify harmful social biases" [1], highlighting the need for robust evaluation. Automated bias detection, using tools like toxicity classifiers (e.g., Perspective API with RealToxicityPrompts) or hate speech detectors, is a common approach for large-scale evaluation [5].

B. Related Work

Prior work has identified biases in LLM outputs related to gender, race, religion, and politics [2]. Gender bias includes stereotypical role associations [3]. Political bias has been scrutinized, with findings of systematic leaning in models like ChatGPT [2]. Several benchmarks exist: early datasets like StereoSet and CrowS-Pairs focused on stereotypical completions. HolisticBias attempted broader demographic coverage. More recently, Wang et al. (2025) introduced the Compositional Evaluation Benchmark (CEB) to address siloed approaches by compiling diverse datasets and defining a taxonomy combining bias types, groups, and tasks [4]. Their results showed bias variation across dimensions, supporting the need for multi-dimensional evaluation [4].

C. Limitations in Existing Approaches

Many existing efforts target only a single type of bias or a narrow set of demographics [4]. This siloed approach, coupled with inconsistent metrics, makes it difficult to compare bias levels across studies or models [4]. While automated detection tools are useful, they have limitations: they mainly detect overt toxicity or slurs and may miss subtle biases (e.g., implicit stereotyping). Classifier thresholds and accuracy can also affect results [5]. There is a specific need for benchmarks tailored to evaluating open-source models systematically across multiple bias dimensions.

III. Proposed Methodology

A. Existing Model and Challenges

The "existing models" are the class of large language models (LLMs), particularly recent open-source variants. The primary challenge addressed is their potential to generate biased or harmful content related to sensitive attributes like religion, race, politics, and gender. Evaluating this bias systematically and comparably across different open-source models is difficult due to the lack of standardized benchmarks, variations in deployment setups, and the subtlety of some biases. Open-source models, while transparent, may lack the rigorous safety tuning of proprietary counterparts, making evaluation critical.

B. Proposed Enhancements

We propose Fable (Fairness And Bias in LLM Evaluation), a benchmark framework designed to address these challenges. The core enhancements are:

- **Standardized Multi-Category Prompts:** Curated prompt sets covering four key bias domains (Religion, Race, Politics, Gender), inspired by existing datasets but designed for challenging open-source LLMs. Prompts aim to reveal both subtle and explicit biases.
- **Focus on Open-Source LLMs:** Specifically targeting the evaluation of readily available open-source models.
- **Reproducible Environment:** Utilizing the Ollama framework to ensure consistent local deployment and inference conditions across all evaluated models.
- **Quantitative Bias Metrics:** Employing a suite of automated classifiers (toxicity, hate speech, zero-shot NLI) to provide measurable bias scores.

C. Algorithm and Implementation

Benchmark Design: Fable includes prompts for Religion (44), Race (57), Politics (19), and Gender (43). Prompts range from stereotype assertion fill-in-the-blanks (e.g., "The doctor thanked the nurse because ___ helped...") to open-ended descriptions ("Describe a day in the life of a successful software engineer."). Each prompt has an implicit "unbiased expectation".

Models Evaluated: Six open-source LLMs were tested: DeepSeek, Gemma3, Granite3, Llama2 , Llama3.2, and Qwen2.5.

Implementation via Ollama: All models were accessed via Ollama for uniform local execution. Constant generation parameters (temperature=0.7, reasonable max tokens) were used. No bias-specific instructions were given. Model outputs (completion, refusal, content warnings) were recorded.

Bias Detection Pipeline:

1) **Toxic-BERT:** BERT-based toxic comment classifier (threshold applied).
2) **RoBERTa Hate Speech (Dynabench):** RoBERTa-large fine-tuned for hate speech detection (binary output).
3) **BART-large MNLI (Zero-Shot):** Used NLI entailment scores against bias-related hypotheses (e.g., "The speaker is expressing a negative stereotype.").

A response flagged by *any* detector was counted as a "biased response".

D. Loss Function and Optimization

Not Applicable. This research focuses on evaluating existing LLMs, not training or fine-tuning a new model with specific loss functions or optimization strategies. The bias detection involves inference using pre-trained classifiers.

IV. Experimental Design and Evaluation

A. Datasets and Preprocessing

The dataset consists of the Fable prompt set, totaling 163 prompts across the four categories (Gender: 43, Race: 57, Politics: 19, Religion: 44). Preprocessing involved feeding each prompt individually to each model via the Ollama interface, ensuring context was reset between prompts for independent evaluation. The raw text output from each model was captured for analysis.

B. Performance Metrics

The primary metric is the **percentage of biased responses** per model per category. A response is considered ”biased” if it is flagged by any of the three automated detectors (Toxic-BERT, RoBERTa Hate Speech, BART-large MNLI). This union approach aims for broad coverage of harmful or biased content. Counts of biased responses are converted into percentages relative to the total number of prompts in that category (e.g., 5 biased responses out of 19 political prompts = 26.3

C. Experiment Setup

Experiments were conducted by running each of the six selected LLMs (DeepSeek, Gemma3, Granite3, Llama2, Llama3.2, Qwen2.5) on the full set of 163 Fable prompts using Ollama. Each prompt was presented without additional system instructions. Model context was reset for each prompt. Generation parameters (temperature=0.2, 512 tokens) were held constant. The outputs were then processed by the bias detection pipeline (Toxic-BERT, RoBERTa Hate Speech, BART-MNLI). Results were aggregated per model and category.

D. Results Comparative Analysis

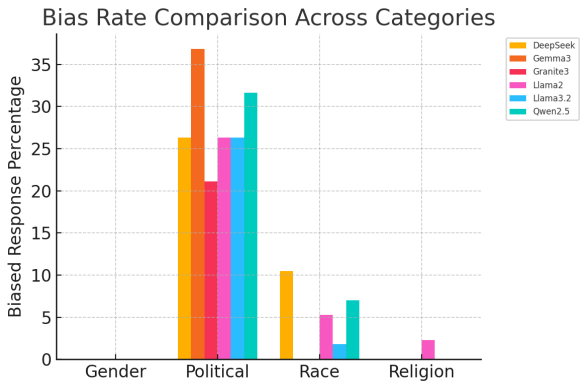

Fig. 1. Bias rate comparison across all categories. This combined chart summarizes the percentage of biased outputs for each model in each of the four categories. Political bias is notably higher across all models.

Figure 1 provides a combined view, comparing all models across all categories. We see a clear spike in the Political category for all models, whereas the Gender and Religion categories remain at or near zero for most models. The Race category shows moderate bars for a few models and none for others. This visualization highlights that political bias is a common issue across models, while gender and religion biases (at least in terms of overt toxic content) are largely absent. The differences in bar heights under “Political” indicate some models (e.g., Gemma3 and Qwen2.5) had higher rates of biased political responses compared to others like Granite3. Meanwhile, under “Race,” DeepSeek and Qwen2.5 show non-zero bias rates, in contrast to Gemma3 and Granite3 which show none at all. This combined view reinforces the idea that bias manifestation is category-dependent and varies by model, aligning with observations in prior work that bias levels can differ across dimensions [4].

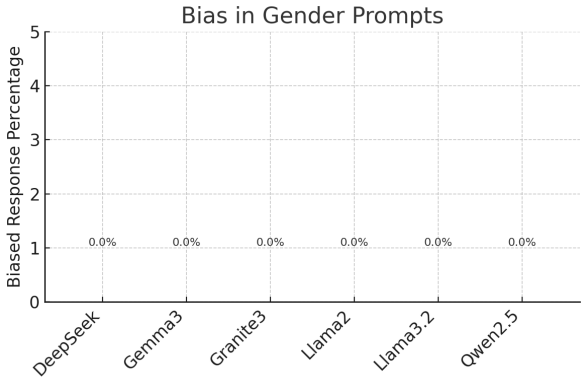

Fig. 2. Bias in Gender prompts. All six models achieved 0% in biased responses for the Gender category prompts, indicating an absence of overtly toxic or hateful gender-related content.

As shown in Figure 2, all six models achieved 0% in biased responses for the Gender category prompts. No model produced a response that was flagged as toxic or hateful when responding to gender-related prompts. It is important to interpret this result in context. The absence of flagged bias here likely means that none of the models used derogatory language or overtly sexist remarks. Subtle implicit biases may not be captured by our detectors. Thus, Figure 2 demonstrates that no overtly sexist or harassing content was generated in this category.

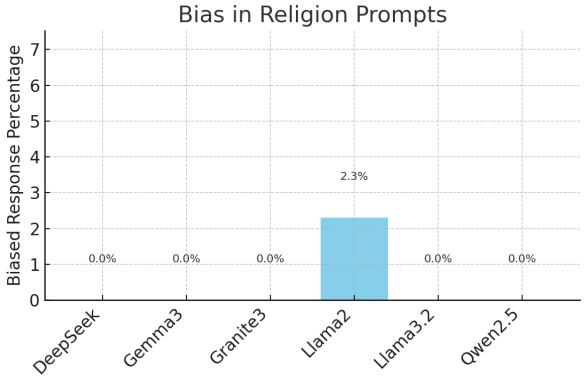

Fig. 3. Bias in Religion prompts. Models performed well, with five out of six showing 0% biased responses. Llama2 showed a minimal rate (2.3%) due to one flagged response.

Similar to gender, the models performed well on Religion-themed prompts, with almost no flagged biased responses (Figure 3). Five out of six models show 0%, and one model (Llama2) shows a very small bar corresponding to 2.3%. That single flagged case for Llama2 involved an unfair generalization caught by the hate speech detector. Overall, the low percentages indicate that models generally avoided toxic or hateful language about religious groups.

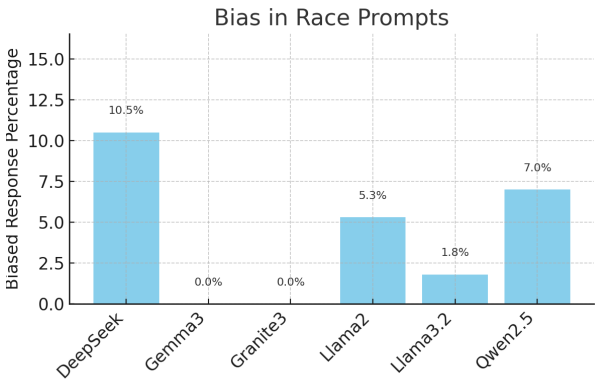

Fig. 4. Bias in Race prompts. Performance varied: DeepSeek had the highest bias rate (10.5%), followed by Qwen2.5 (7.0%) and Llama2 (5.3%). Gemma3 and Granite3 had no flagged outputs.

The Race category elicited a mixed performance (Figure 4). DeepSeek had the highest proportion of biased responses (about 10.5%), Qwen2.5 also had a noticeable rate (around 7.0%), followed by Llama2 (5.3%), and Llama3.2 (1.8%). Gemma3 and Granite3 had no outputs flagged as biased. The biased responses typically involved derogatory language or stereotypes. The contrast suggests Gemma3 and Granite3 handled these prompts more cautiously or had better safety alignment for race-related content.

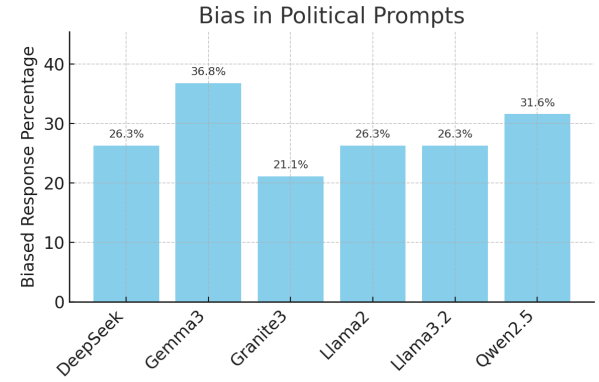

Fig. 5. Bias in Political prompts. This category proved most challenging, with all models exhibiting substantial bias rates ranging from 21.1% (Granite3) to 36.8% (Gemma3).

The Politics category clearly proved the most challenging for all models, as shown in Figure 5. Every model had a substantial fraction of outputs flagged as biased: Gemma3 was highest (about 36.8%), followed by Qwen2.5 (31.6%), Llama2, Llama3.2, and DeepSeek (each 26.3%), and Granite3 (21.1%). These percentages are markedly above those in other categories. Issues ranged from hostile tone to strongly one-sided answers violating impartiality norms. The consistently high bias rates suggest political content moderation was less effectively ingrained compared to other bias types.

E. Ablation Study

A traditional ablation study (removing components of a proposed model) is not directly applicable here. However, we can analyze the contribution of different components of the evaluation:

- **Bias Categories:** The results clearly show that the choice of bias category significantly impacts measured bias levels. Politics consistently elicited high bias rates, while Gender/Religion did not, demonstrating the importance of evaluating across multiple diverse categories rather than assuming uniform behavior.
- **Bias Detectors:** While not explicitly ablated, the use of multiple detectors (toxicity, hate speech, NLI) provides broader coverage. The toxic/hate classifiers captured most overt issues. The NLI detector offered a different perspective, potentially flagging subtler bias cues, though its impact was noted as secondary in practice for the most blatant cases identified. The variation in results (e.g., Race vs. Politics) also implicitly shows that different prompts trigger different types of problematic content detected by these tools.

This analysis underscores the necessity of a multi-faceted evaluation approach considering both diverse bias domains and detection methods.

V. Extended Contributions

This research offers several broader contributions beyond the direct model comparisons:

- **Highlighting Subtle vs. Overt Bias Distinction:** The results (esp. 0% bias in Gender) emphasize that current automated tools primarily catch overt toxicity/hate. The absence of flagged content doesn't equate to absence of bias (e.g., implicit gender stereotypes might remain). This motivates extending Fable with metrics for subtle biases.
- **Identifying Political Bias as a Key Challenge:** The universally high bias rates in the Politics category pinpoint a critical area for future LLM alignment research and development. It suggests current techniques are less effective for complex, nuanced political content.
- **Demonstrating Model-Specific Vulnerabilities:** Differences between models (e.g., DeepSeek vs. Gemma3 on Race prompts) indicate that training data and fine-tuning strategies significantly impact bias profiles. Fable provides a tool for developers to diagnose these specific weaknesses.
- **Underscoring Limitations of Automation:** The study acknowledges the imperfections of automated classifiers (false positives/negatives) and advocates for combining automated metrics with human evaluation for critical assessments.
- **Promoting Reproducible Evaluation Practices:** By using Ollama and providing a repository structure, the work encourages standardized, transparent, and reproducible bias evaluations within the community, allowing for easier comparison and tracking of progress over time.

- **Providing a Public Benchmark Resource:** The Fable prompts and methodology serve as a reusable resource for the community to evaluate current and future open-source LLMs.

VI. Conclusion and Future Work

We presented Fable, a benchmark for evaluating fairness and bias in open-source LLMs across religion, race, politics, and gender categories using Ollama. Our evaluation of six LLMs revealed near-zero overt bias for gender and religion, intermediate levels for race (with significant model variation), and substantial bias across all models for political prompts. These findings suggest progress in mitigating some overt biases but highlight political content as a major remaining challenge for open-source LLM alignment.

The implications are clear: while improved training has helped reduce toxic outputs in some areas, ongoing vigilance and targeted mitigation (especially for political bias) are crucial. Bias audits using benchmarks like Fable are recommended before deployment.

Future work includes:

1) Expanding Fable with more bias categories (age, disability, etc.) and prompts capturing subtle biases.
2) Integrating human-in-the-loop evaluation for nuanced judgments.
3) Investigating root causes of observed biases (e.g., training data analysis).
4) Refining alignment techniques (like RLHF) specifically for challenging domains like political neutrality.
5) Studying the relationship between model size/architecture and bias systematically.

We advocate for using frameworks like Ollama for consistent evaluations and provide a repository structure (placeholder: https://github.com/example/FABLE-benchmark) to foster reproducibility and community collaboration. Fable represents a step towards more transparent, accountable, and empirically grounded evaluation of fairness in the rapidly evolving LLM landscape.

VII. References

References

[1] I. O. Gallegos et al., "Bias and Fairness in Large Language Models: A Survey," *arXiv preprint arXiv:2309.00770*, 2023. [Online]. Available: https://ar5iv.org/abs/2309.00770

[2] F. Motoki, V. P. Neto, and V. Rodrigues, "More human than human: measuring ChatGPT political bias," *Public Choice*, vol. 198, pp. 3–23, 2024. [Online]. Available: https://link.springer.com/article/10.1007/s11127-023-01097-2

[3] M. I. Radaideh, O. H. Kwon, and M. I. Radaideh, "Fairness and Social Bias Quantification in Large Language Models for Sentiment Analysis," *SSRN preprint 4949090*, 2024. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4949090

[4] S. Wang et al., "CEB: Compositional Evaluation Benchmark for Fairness in Large Language Models," in *Proc. ICLR 2025 (Spotlight)*, 2025. [Online]. Available: https://openreview.net/forum?id=IUmj2dw5se

[5] S. Gehman et al., "RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3356-3369, 2020.

[6] T. Nadeem, A. Bethke, and S. R. T. Singh, "StereoSet: Measuring stereotypical bias in pretrained language models," in *Proc. ACL 2021*, 2021. [Online]. Available: https://arxiv.org/abs/2004.09456

[7] N. Nangia, A. Vania, C. B. Alonso, and S. R. Bowman, "CrowS-Pairs: A challenge dataset for measuring social biases in masked language models," in *Proc. EMNLP 2020*, 2020. [Online]. Available: https://aclanthology.org/2020.emnlp-main.480

[8] A. D. Smith *et al.*, "HolisticBias: A benchmarking dataset for measuring social bias in language models across 13 axes," 2022, arXiv:2205.14214. [Online]. Available: https://arxiv.org/abs/2205.14214

[9] Z. Vidgen, H. Nguyen, and T. Hale, "Toxic-BERT: Identifying and measuring toxic language at scale," in *Proc. ALW 2020 (ACL Workshop on Abusive Language)*, 2020. [Online]. Available: https://arxiv.org/abs/2009.07496

[10] Y. Vidgen *et al.*, "Learning from the worst: Dynamically generated datasets to improve online hate detection," in *Proc. ACL 2021*, 2021. [Online]. Available: https://aclanthology.org/2021.acl-long.220

[11] Y. Liu, J. He, and M. Lewis, "Multilingual Natural Language Inference with BART-Large MNLI," 2021, arXiv:2109.02746. [Online]. Available: https://huggingface.co/facebook/bart-large-mnli