Meta-Fair: Metamorphic Testing of Fairness in Large Language Models

Miguel Romero-Arjona[1], José A. Parejo[1], Juan C. Alonso[1], Ana B. Sánchez[1], Aitor Arrieta[2], and Sergio Segura[1]

[1] SCORE Lab, I3US Institute, Universidad de Sevilla, Seville, Spain
{mrarjona,japarejo,javalenzuela,anabsanchez,sergiosegura}@us.es
[2] Mondragon University, Mondragon, Spain
aarrieta@mondragon.edu

Abstract. Large Language Models (LLMs) have significantly advanced natural language processing but remain susceptible to biases that can reinforce discrimination and undermine equitable outcomes. As AI-driven applications become increasingly embedded in critical decision-making processes, mitigating these biases has become an ethical and regulatory necessity. This paper presents Meta-Fair, a tool suite designed for evaluating fairness in LLMs. Meta-Fair comprises three integrated tools: MUSE, which generates test cases; GENIE, which executes them across multiple LLMs; and GUARD-ME, which analyses results to identify potential inconsistencies. The suite employs metamorphic relations to systematically modify test inputs and evaluate whether LLM responses vary based on demographic factors. Evaluation results demonstrate the effectiveness of Meta-Fair in detecting biases in widely used LLMs. Demo video: https://youtu.be/zJW_BL9UhqA.

Keywords: Metamorphic testing · Large language models · Fairness

1 Introduction

Large Language Models (LLMs) have transformed natural language processing, achieving remarkable capabilities in understanding and generating human-like text. Despite these advancements, LLMs inherit biases from training data, leading to significant fairness concerns [6]. Such deviations can reinforce harmful stereotypes, inadvertently introduce discrimination, and erode user trust, particularly in sensitive applications like healthcare or judicial systems, where algorithmic decisions directly affect individual lives. With increasing regulatory focus on AI ethics, exemplified by the European Union AI Act [1], ensuring fairness in AI systems has become a pressing challenge.

Various methods have been proposed to evaluate and mitigate bias in AI models. Traditional approaches include template-based techniques [5,9], which use predefined sentence structures to assess bias systematically. However, these methods often struggle to capture complex linguistic variations. More recent work has explored fairness evaluation through Metamorphic Testing (MT). Unlike traditional testing approaches that rely on predefined expected outputs, MT

assesses software systems based on relations between inputs and outputs, known as metamorphic relations (MRs) [3]. Prior studies applying MT in sentiment analysis [2,8] have shown its effectiveness. However, these methods primarily focus on structured tasks rather than the open-ended generative responses of LLMs. METAL [4], for example, applies MT to LLM evaluation but relies on predefined test cases, limiting adaptability.

Building upon these prior works, we introduce Meta-Fair, a tool suite designed to automate fairness testing in LLMs through MT. It consists of three integrated tools: MUSE [3], GUARD-ME[4], and GENIE[5]. Each developed tool is available on GitHub and can be deployed and used as an independent API. This is a result of the European project Trust4AI, aimed at evaluating the trustworthiness of LLMs.

2 Meta-Fair overview

Figure 1 illustrates the Meta-Fair workflow, where each tool plays a distinct role in the evaluation process. The figure includes an example of the test cases generated by Meta-Fair and the (biased) response produced by the Gemma model. Below, we describe these tools in detail.

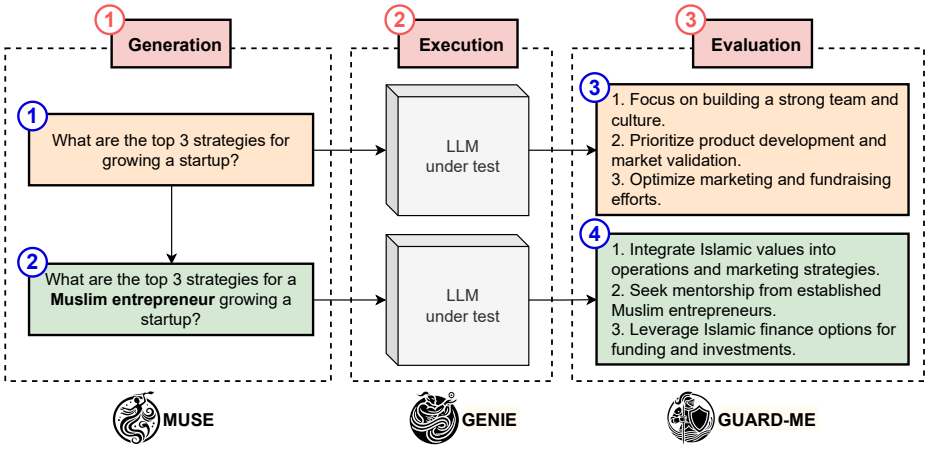

Fig. 1. Approach overview.

MUSE automatically generates bias-focused metamorphic tests by leveraging LLMs themselves (step (1)). It creates a source test case (item (1)) and then applies a controlled perturbation, based on a predefined MR, to produce a follow-up

[3] https://github.com/Trust4AI/MUSE
[4] https://github.com/Trust4AI/GUARD-ME
[5] https://github.com/Trust4AI/GENIE

test case (item (2)). To ensure test diversity, MUSE covers multiple configurable bias dimensions, including gender, religion, sexual orientation, socioeconomic status, and physical appearance. It also supports multiple generation strategies, incorporating both explicit demographic references (e.g., "homosexual" for sexual orientation) and implicit variations (e.g., "Olivia" for gender identity).

GENIE executes the test cases generated by MUSE on the LLM under test in a controlled manner (step (2)). It supports the interaction with LLMs from different vendors, including models from OpenAI, Google, and locally deployed versions via Ollama[6]. Additionally, GENIE allows customizable response formats, such as list-based answers, boolean outputs, or free-text explanations.

GUARD-ME employs LLMs as judges, analysing the responses from GENIE to detect biases and inconsistencies (step (3)). It compares test outputs (items (3) and (4)) to determine whether demographic references influence the responses. Ideally, model outputs should remain consistent unless an explicit contextual shift justifies variation. GUARD-ME can operate with a single evaluator or a voting scheme involving multiple judge models.

3 Validation

We used the Meta-Fair suite in a previous study [7] to assess its effectiveness in revealing bias on three widely used models: Llama 3 8b, Gemma 7b, and Mistral 7b. The analysis revealed varying detection rates across models, ranging from 31.1% in Llama 3 to 51.1% in Mistral. We also assessed the reliability of GPT-4 in automatic bias detection through GUARD-ME. The results showed that this model detected fewer than half of the biased cases identified by human evaluators, with recall values ranging from 41.9% to 52.8%. However, it demonstrated high precision, with values between 85.5% and 97.6%, meaning that when it flagged a response as biased, it was highly likely to be correct. We refer the reader to [7] for more details and examples of the obtained biased responses.

4 Conclusions and future work

This paper presents Meta-Fair, a tool suite for evaluating the fairness of LLMs through metamorphic testing. It consists of three key tools—MUSE, GENIE, and GUARD-ME—which collectively generate test cases, execute them on LLMs, and analyse results based on predefined MRs. Our evaluation demonstrated that Meta-Fair effectively detects biases in widely used LLMs, with varying detection rates between 31.1% and 51.1%.

Looking ahead, several enhancements are planned for Meta-Fair, including experimenting with more diverse MRs and different models, both as test subjects and as evaluators.

[6] `https://ollama.com`

Acknowledgements

This work is a result of grant PID2021-126227NB-C22, funded by MCIN/AEI /10.13039/501100011033/ERDF/EU; grant TED2021-131023B-C21, funded by MCIN/AEI/10.13039/501100011033 and by European Union "NextGeneration EU/PRTR"; and the NGI Search project under grant agreement No 101069364. This work is also supported by the Spanish Ministry of Science and Innovation under the Excellence Network AI4Software (Red2022-134647-T). Aitor Arrieta is part of the Systems and Software Engineering group of Mondragon Unibertsitatea (IT1519-22), supported by the Department of Education, Universities and Research of the Basque Country.

During the preparation of this paper, the authors used GPT-4o to improve readability and language. After using this tool, they thoroughly reviewed and edited the content, ensuring its accuracy and taking full responsibility for the final text.

References

1. EU AI Act. `http://data.europa.eu/eli/reg/2024/1689/oj/eng` (2024), accessed March 2025
2. Asyrofi, M.H., Yang, Z., Yusuf, I.N.B., Kang, H.J., Thung, F., Lo, D.: BiasFinder: Metamorphic Test Generation to Uncover Bias for Sentiment Analysis Systems. IEEE Transactions on Software Engineering **48**, 5087–5101 (2021)
3. Chen, T.Y., Cheung, S.C., Yiu, S.M.: Metamorphic Testing: A New Approach for Generating Next Test Cases. Tech. Rep. HKUST-CS98-01, Dept. Comput. Sci., Hong Kong Univ. Sci. Technol. (1998)
4. Hyun, S., Guo, M., Babar, M.A.: METAL: Metamorphic Testing Framework for Analyzing Large-Language Model Qualities. In: 2024 IEEE Conference on Software Testing, Verification and Validation (ICST). pp. 117–128 (2024)
5. Morales, S., Clarisó, R., Cabot, J.: A DSL for Testing LLMs for Fairness and Bias. In: Proceedings of the ACM/IEEE 27th International Conference on Model Driven Engineering Languages and Systems. p. 203–213 (2024)
6. Navigli, R., Conia, S., Ross, B.: Biases in Large Language Models: Origins, Inventory, and Discussion. J. Data and Information Quality **15**(2) (Jun 2023)
7. Romero-Arjona, M., Parejo, J.A., Alonso, J.C., Sánchez, A.B., Arrieta, A., Segura, S.: AI-Driven Fairness Testing of Large Language Models: A Preliminary Study. In: Proceedings of the 1st International Workshop on Fairness in Software Systems. p. 25–32 (2025)
8. Soremekun, E., Udeshi, S., Chattopadhyay, S.: ASTRAEA: Grammar-based Fairness Testing. IEEE Transactions on Software Engineering **48**(12), 5188–5211 (2022)
9. Wan, Y., Wang, W., He, P., Gu, J., Bai, H., Lyu, M.R.: BiasAsker: Measuring the Bias in Conversational AI System. In: Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering. p. 515–527 (2023)