Qeios

Peer Review

Review of: "ToxiLab: How Well Do Open-Source LLMs Generate Synthetic Toxicity Data?"

Samsul Arifin[1]

1. Independent researcher

Title

The title, "ToxiLab: How Well Do Open-Source LLMs Generate Synthetic Toxicity Data?", is clear and informative. It effectively captures the essence of the study, focusing on the evaluation of open-source LLMs for synthetic toxic data generation. However, it could be more concise by removing "How Well Do" to streamline readability.

Abstract

The abstract provides a good overview of the study's objectives, methodology, and findings. It highlights key contributions such as evaluating open-source LLMs and discussing ethical considerations. However, it could better emphasize specific quantitative results to strengthen its impact.

Introduction

The introduction is well-structured and provides a compelling motivation for the study. It clearly identifies challenges in toxic content detection and the need for synthetic data. However, it could benefit from a more detailed explanation of why open-source models are preferable over proprietary ones in terms of scalability and accessibility.

Related Works

The related works section is comprehensive but lacks depth in comparing the proposed methods to previous approaches. The authors should elaborate on how their methodology improves upon existing frameworks like ToxiGen or Toxicraft, particularly in terms of cost-effectiveness and dataset diversity.

Methodology

The methodology is detailed and logically structured, with clear descriptions of prompt engineering and

doi.org/10.32388/6MQ90O

supervised fine-tuning stages. However, the paper could improve by providing more specifics on hyperparameter tuning during fine-tuning and discussing potential biases introduced by proprietary datasets.

Results

The results are presented with adequate metrics such as F1-score and accuracy across multiple datasets. While Mistral's superiority is highlighted, the discussion could delve deeper into why certain models underperformed and how these insights can guide future improvements.

References

The references are relevant and well-cited but could include more recent works on the ethical implications of synthetic data generation to enrich the discussion on responsible AI deployment.

Readability

The manuscript is generally readable but occasionally uses overly technical jargon without sufficient explanation (e.g., "LoRA methods"). Simplifying the language or adding brief explanations would make it more accessible to a broader audience.

Open Problem

The paper identifies open problems such as ethical risks and real-world deployment challenges but does not propose actionable solutions. Including recommendations for mitigating these issues would significantly enhance its practical relevance.

Declarations

Potential competing interests: No potential competing interests to declare.