Mistral-7B and Llama-3-8B-128K Throughput-Accuracy Trade-offs on HumanEval in Multi-Threaded Settings

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 8 peer-reviewed papers addressing the following research question: What is the trade-off between throughput and code generation accuracy when comparing Mistral-7B and Llama-3-8B-128K in multi-threaded environments using the HumanEval benchmark. As machine learning models are increasingly embedded into society through high-stakes decision-making, selecting the right algorithm for a given task, audience, and sector presents a critical challenge, particularly in the context of fairness. Traditional assessments of model. 9 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Selecting for Less Discriminatory Algorithms: A Relational Search Framework for Navigating Fairness-Accuracy Trade-offs in Practice. Research question: What is the trade-off between throughput and code generation accuracy when comparing Mistral-7B and Llama-3-8B-128K in multi-threaded environments using the HumanEval benchmark?.

2 Methodology

Systematic literature search across multiple databases yielded 8 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.8/10.

3 Results

8 papers retrieved. 9 claims extracted; 0 independently verified. Quality review score: 3.8/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
The findings demonstrate the robustness of Lee and Floridi's fairness assessment framework in visualizing algorithmic tr	×	0.12
Racial disparity measures such as those in Lee and Floridi [29] do not account for historical bias prevalent in mortgage	×	0.04
Machine learning models evaluated include LR, KNN, CART, NB, and RF.	×	0.06
Fairness metrics evaluated include Equal Opportunity (EOP), False Positive Error Rate Balance (FPERB), Equal Odds (EO),	×	0.02
The metrics indicate the difference in outcomes between Non-Hispanic White applicants and Black or People of Color appli	×	0.02
Negative values in the fairness metric comparisons indicate that White applicants have lower fairness metrics than Black	×	0.05
The LR model shows an Equal Opportunity (EOP) disparity of 6% between White and Black applicants.	×	0.02
The RF model shows the highest Equal Opportunity (EOP) disparity at 24% between White and Black applicants.	×	0.01
The NB model shows perfect parity (0%) for Equal Opportunity (EOP), False Positive Error Rate Balance (FPERB), and Equal	×	0.01

References

- http://arxiv.org/abs/2406.14712v1
- http://arxiv.org/abs/2310.06825v1
- http://arxiv.org/abs/2506.01594v2