Multimodal Pre-Training Impact on Llama-2 Robustness in MBPP Pro Benchmark

Assignee Research

June 3, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: What is the comparative robustness of Llama-2 models with and without multimodal pre-training when evaluated on non-adversarial versus adversarial inputs in the MBPP Pro benchmark, measured by. 14 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Enhancing clinical decision support with physiological waveforms – a multimodal benchmark in emergency care. Research question: What is the comparative robustness of Llama-2 models with and without multimodal pre-training when evaluated on non-adversarial versus adversarial inputs in the MBPP Pro benchmark, measured by accuracy degradation?.

2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.7/10.

3 Results

11 papers retrieved. 14 claims extracted; 0 independently verified. Quality review score: 3.7/10.

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 multimodal model integrating ECG waveforms and clinical routine data achieved a macro-AUROC of 0.8256 (0.8222, 0.828	×	0.08
The multimodal model integrating ECG waveforms and clinical routine data achieved an AUROC of 0.9115 (0.8991, 0.9222) fo	×	0.09
The model's predictive performance for diagnoses ranges from an AUROC of 0.7405 for the musculoskeletal system and conne	×	0.04
The model predicted 609 out of 1,428 individual ICD diagnoses with high accuracy, defined as conditions where the lower	×	0.05
The inclusion of ECG waveforms alongside clinical routine data consistently improves performance over ECG features and c	×	0.04
The highest gains from including ECG waveforms are observed for XII (Skin, 13.06%), VI (Nervous, 10.75%), and XIX (Injur	×	0.02
Smaller improvements are seen for III (Blood, 2.36%) and IV (Endocrine, 5.25%), indicating that static ECG features migh	×	0.01
For clinical deterioration, the model achieves an AUROC of 0.9070.	×	0.11
For ICU admissions, the model reports an overall AUROC of 0.9063.	×	0.03
For mortality predictions, the model exhibits an overall AUROC of 0.9168.	×	0.03
The MDS-ED pipeline involves feature collection encompassing patient demographics, biometrics, vital parameters and tren	×	0.11
The MDS-ED pipeline predicts patient discharge diagnoses out of 1428 cardiac and non-cardiac ICD10-CM codes and predicts	×	0.12
The MDS-ED pipeline collects features from a window of 90 minutes from the patient's arrival at the ED.	×	0.01
The MDS-ED dataset was created by linking ECG waveforms from the MIMIC-IV-ECG dataset to clinical features and outcomes	×	0.04

References

- http://arxiv.org/abs/2412.21199v2
- http://arxiv.org/abs/1901.09960v5
- http://arxiv.org/abs/2407.17856v4