How does the cross-lingual pretraining of OpenPangu-MLA impact its performance on paralinguistic feature extra

Assignee Research

June 10, 2026

Abstract

Multilingual language models (MLLMs) are crucial for handling text across various languages, yet they often show performance disparities due to differences in resource availability and linguistic characteristics. While the impact of pre-train data percentage and model size on performance is well-known, our study reveals additional critical factors that significantly influence MLLM effectiveness. Analyzing a wide range of features, including geographical, linguistic, and resource-related aspects, we focus on the SIB-200 dataset for classification and the Flores-200 dataset for machine translati

1 Introduction

This paper examines: Beyond Data Quantity: Key Factors Driving Performance in Multilingual Language Models. Research question: How does the cross-lingual pretraining of OpenPangu-MLA impact its performance on paralinguistic feature extraction in low-resource languages compared to high-resource languages when evaluated on the MMSU benchmark?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 2.2/10.

3 Results

14 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 2.2/10.

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.

References

- http://arxiv.org/abs/2602.05599v1
- http://arxiv.org/abs/2412.12500v1
- http://arxiv.org/abs/2505.18673v1