Parameter Count Impact on Zero-Shot Cross-Lingual Semantic Parsing Accuracy for Low-Resource Languages

Assignee Research

June 13, 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 increasing parameter count affect zero-shot cross-lingual semantic parsing accuracy on Wikidata compared to Freebase for low-resource languages?.

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 7.3/10.

3 Results

14 papers retrieved. 20 claims extracted; 16 independently verified. Quality review score: 7.3/10.

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 test set includes 204 languages, each with 204 instances.	✓	0.23
SIB-200, based on Flores-200, is an open-source benchmark for topic classification across 200+ languages and dialects.	✓	0.24
Its test set also covers 204 languages, with 204 instances per language.	✓	0.24
Bloom model achieved R2 = 0.645, MSE = 0.005 in Classification Zero-Shot task.	✓	0.16
BloomZ model achieved R2 = 0.903, MSE = 0.001 in Classification Zero-Shot task.	✓	0.16
XGLM model achieved R2 = 0.855, MSE = 0.003 in Classification Zero-Shot task.	✓	0.18
Bloom model achieved R2 = 0.847, MSE = 0.007 in Classification Two-Shot task.	×	0.14
BloomZ model achieved R2 = 0.754, MSE = 0.009 in Classification Two-Shot task.	×	0.14
XGLM model achieved R2 = 0.902, MSE = 0.003 in Classification Two-Shot task.	✓	0.18
Bloom model achieved R2 = 0.553, MSE = 8.037 in Generation Zero-Shot task.	✓	0.16
BloomZ model achieved R2 = 0.918, MSE = 37.443 in Generation Zero-Shot task.	✓	0.18
XGLM model achieved R2 = 0.902, MSE = 3.365 in Generation Zero-Shot task.	✓	0.19
Bloom model achieved R2 = 0.866, MSE = 6.322 in Generation Two-Shot task.	×	0.15
BloomZ model achieved R2 = 0.950, MSE = 18.687 in Generation Two-Shot task.	✓	0.17
XGLM model achieved R2 = 0.801, MSE = 2.950 in Generation Two-Shot task.	✓	0.15
Simpler models like SVR, K-Nearest Neighbors, and Lasso Regression generally performed poorly, often yielding negative R	✓	0.33
Ensemble models such as Random Forest, Gradient Boosting, and XGBoost consistently excelled, demonstrating strong predic	✓	0.22
The study spans both classification and generation tasks, assessed in zero-shot and two-shot learning settings.	✓	0.24
The code for this study is publicly available at https://github.com/PortNLP/SHAP-MLLM-Analysis.	✓	0.23
SHAP (SHapley Additive exPlanations) values are used to quantify the importance of each feature.	×	0.10

References

- http://arxiv.org/abs/2310.09917v3
- http://arxiv.org/abs/1908.10461v1
- http://arxiv.org/abs/2412.12500v1