Gemma2 Scaling Behavior in Monolingual and Multilingual Question Answering

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: How does the scaling behavior of Gemma2-2B's F1-score differ between monolingual (English) and multilingual (Italian) QA tasks as model size increases from 2B to 7B parameters. 9 claims were extracted from source literature; 4 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 6.4/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Cross-Lingual Pitfalls: Automatic Probing Cross-Lingual Weakness of Multilingual Large Language Models. Research question: How does the scaling behavior of Gemma2-2B's F1-score differ between monolingual (English) and multilingual (Italian) QA tasks as model size increases from 2B to 7B parameters?.

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 6.4/10.

3 Results

14 papers retrieved. 9 claims extracted; 4 independently verified. Quality review score: 6.4/10.

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 paper defines cross-lingual weakness as a scenario where a model answers a question correctly in English but incorre	×	0.10
English is the primary training language for Large Language Models (LLMs).	×	0.14
LLMs generally perform best in English compared to other languages.	×	0.05
Linguistically related languages share similar cross-lingual performance patterns.	✓	0.27
Linguistically related languages benefit from targeted post-training.	✓	0.19
The proposed methodology uses a beam search strategy guided by LLM-based simulation scores to refine bilingual question	✓	0.16
The methodology utilizes source datasets including ARC-Challenge, CommonsenseQA, MMLU, TruthfulQA, and SciQ.	×	0.03
The code for the study is available at https://github.com/xzx34/Cross-Lingual-Pitfalls.	✓	0.26
Failure across all languages for a given question indicates a knowledge-related limitation rather than a cross-lingual w	×	0.11

References

- http://arxiv.org/abs/2404.01331v2
- http://arxiv.org/abs/2505.18673v1
- http://arxiv.org/abs/2403.08295v4