# Cross-Lingual F1-Score Gaps in Gemma2 Models on Adversarial QA Datasets

Assignee Research

June 6, 2026

**Abstract**

This report synthesises findings from 10 peer-reviewed papers addressing the following research question: How does the F1-score gap between English and Italian QA tasks scale when comparing Gemma2-2B and Gemma2-7B on adversarial cross-lingual datasets generated via beam search. 13 claims were extracted from source literature; 4 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.7/10. 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 F1-score gap between English and Italian QA tasks scale when comparing Gemma2-2B and Gemma2-7B on adversarial cross-lingual datasets generated via beam search?.

## 2 Methodology

Systematic literature search across multiple databases yielded 10 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.7/10.

## 3 Results

10 papers retrieved. 13 claims extracted; 4 independently verified. Quality review score: 5.7/10.

## 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 |
|---|---|---|
| Linguistically related languages share similar performance patterns and benefit from targeted post-training. | ✓ | 0.28 |
| English is the primary training language for LLMs, and they generally perform best in English. | × | 0.05 |
| Cross-lingual weakness is defined as a model answering correctly in English but incorrectly in at least one other langua | × | 0.10 |
| The proposed beam search-based methodology efficiently uncovers cross-lingual weaknesses in LLMs. | ✓ | 0.22 |
| The code for the study is available at https://github.com/xzx34/Cross-Lingual-Pitfalls. | ✓ | 0.26 |
| The proficiency demonstrated in English often fails to generalize to other languages, resulting in errors in other lingu | × | 0.03 |
| The pipeline initiates with sampling English questions and creating bilingual pairs. | × | 0.07 |
| Iterative perturbation, driven by a beam search strategy and guided by LLM-based simulation scores, refines these pairs | × | 0.13 |
| The resulting candidate list of question pairs is designed to highlight inherent cross-lingual weaknesses in LLMs. | ✓ | 0.18 |
| The accuracy scores for various languages and datasets are provided in the benchmark tables. | × | 0.03 |
| The percentage scores for various languages and datasets are provided in the benchmark tables. | × | 0.03 |
| The correlation scores for various languages and datasets are provided in the benchmark tables. | × | 0.03 |
| The similarity scores for Chinese, Japanese, and Korean are provided in the benchmark tables. | × | 0.02 |

## References


- http://arxiv.org/abs/2508.09516v1
- http://arxiv.org/abs/2505.18673v1

- http://arxiv.org/abs/2106.16020v1