Fine-Tuning LLMs on Artificial Code-Switched Data for Enhanced Zero-Shot Cross-Lingual Reasoning

Assignee Research

July 8, 2026

Abstract

Transferring information retrieval (IR) models from a high-resource language (typically English) to other languages in a zero-shot fashion has become a widely adopted approach. In this work, we show that the effectiveness of zero-shot rankers diminishes when queries and documents are present in different languages. Motivated by this, we propose to train ranking models on artificially code-switched data instead, which we generate by utilizing bilingual lexicons. To this end, we experiment with lexicons induced from (1) cross-lingual word embeddings and (2) parallel Wikipedia page titles. We use

1 Introduction

This paper examines: Boosting Zero-shot Cross-lingual Retrieval by Training on Artificially Code-Switched Data. Research question: Can fine-tuning LLMs on artificially code-switched data improve their zero-shot cross-lingual reasoning capabilities, as measured by MMLU or TyDiQA benchmarks, compared to standard multilingual training?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.2/10.

3 Results

13 papers retrieved. 22 claims extracted; 16 independently verified. Quality review score: 7.2/10.

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.

5 Extracted Claims

Claim	Verified	Confidence
Code-switching improves cross-lingual and multilingual re-ranking performance.	×	0.14
Code-switching training does not impede monolingual (MoIR) setup performance.	×	0.06
The average MoIR zero-shot performance is substantially higher than CLIR with 15.7 MRR@10.	✓	0.18
The average MoIR zero-shot performance is substantially higher than MLIR with 16.6 MRR@10.	✓	0.16
In CLIR, the performance drop when transferring models is larger for setups involving typologically distant languages li	✓	0.26
The performance gap between zero-shot and fine-tuning on translated data is +4 MRR@10 in MoIR.	✓	0.23
The performance gap between zero-shot and fine-tuning on translated data is +11.1 MRR@10 in CLIR.	✓	0.25
The performance gap between zero-shot and fine-tuning on translated data is +8.3 MRR@10 in MLIR.	✓	0.24
Training on code-switched data consistently outperforms zero-shot models in CLIR and MLIR.	✓	0.25
In the AR–IT pair, code-switching training improved MRR@10 from 7.7 to 15.6.	×	0.13
In the AR–RU pair, code-switching training improved MRR@10 from 7.1 to 14.1.	×	0.14
The difference in MoIR performance between code-switching approaches (BL-CS and ML-CS) and Zero-shot is not statisticall	✓	0.20
Specializing one zero-shot model for multiple CLIR language pairs (ML-CS, Wiki-CS) performs almost on par with specializ	✓	0.31
Wiki-CS results are slightly worse in MoIR compared to other approaches.	✓	0.17
Wiki-CS results are on par with ML-CS on MLIR and CLIR tasks.	✓	0.20
In MoIR, Zero-shot Translate Test and ML-CS Translate Test underperform compared to other approaches.	✓	0.23
Zero-shot rankers perform better on clean monolingual data in the target language than on noisy monolingual data in Engl	✓	0.22
In CLIR, Translate Test yields improvements of +0.2 and +2.2 MRR@10.	×	0.14
In both MoIR and CLIR, Translate Test consistently falls behind code-switching at training time.	✓	0.23
Code-switching gains remain virtually unchanged when moving from six seen languages	✓	0.23

References

- http://arxiv.org/abs/2208.08629v1
- http://arxiv.org/abs/2305.05295v2
- http://arxiv.org/abs/2506.15415v1