Fine-Tuning Gemma2-2B on Low-Resource Italian Datasets and Zero-Shot English IR Performance

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: What is the impact of fine-tuning Gemma2-2B on low-resource Italian datasets like Italian SQuAD on its zero-shot performance in English IR tasks measured by MRR or nDCG. 10 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Learning Cross-Lingual IR from an English Retriever. Research question: What is the impact of fine-tuning Gemma2-2B on low-resource Italian datasets like Italian SQuAD on its zero-shot performance in English IR tasks measured by MRR or nDCG?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.2/10.

3 Results

15 papers retrieved. 10 claims extracted; 0 independently verified. Quality review score: 3.2/10.

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
The XOR-TyDi dataset contains examples in seven typologically diverse languages: Arabic (Ar), Bengali (Bn), Finnish (Fi)	×	0.02
The XOR-TyDi test set contains 2,113 questions.	×	0.04
The MKQA dataset is used for training and validation in zero-shot experiments.	×	0.07
The Natural Questions (NQ) dataset is used for English pre-training of the baseline model.	×	0.05
The CLIR baseline used in the experiments is ColBERT with an underlying XLM-R PLM.	×	0.05
The DR.DECR model is initialized with the parameter weights of the ColBERT baseline.	×	0.11
The KD teacher is a ColBERT model fine-tuned with only English triples.	×	0.09
Evaluation metrics used are Recall at t tokens for t $\in$ {2000, 5000}, i.e., R@2kt and R@5kt.	×	0.04
Pre-training the baseline model with English IR triples from the NQ train set substantially boosts its performance in bo	×	0.07
DR.DECR yields an improvement of 25.4 points over the baseline in the in-domain setting.	×	0.11

References

- http://arxiv.org/abs/2509.12382v1
- http://arxiv.org/abs/2112.08185v3
- http://arxiv.org/abs/2404.14700v4