

# Fine-Tuning on WebFAQ Curated Data for Robust Multilingual Dense Retrieval Under Domain Shift

Assignee Research

June 12, 2026

## Abstract

We present WebFAQ, a large-scale collection of open-domain question answering datasets derived from FAQ-style schema.org annotations. In total, the data collection consists of 96 million natural question-answer (QA) pairs across 75 languages, including 47 million (49%) non-English samples. WebFAQ further serves as the foundation for 20 monolingual retrieval benchmarks with a total size of 11.2 million QA pairs (5.9 million non-English). These datasets are carefully curated through refined filtering and near-duplicate detection, yielding high-quality resources for training and evaluating multil

## 1 Introduction

This paper examines: WebFAQ: A Multilingual Collection of Natural Q&A Datasets for Dense Retrieval. Research question: Does fine-tuning on WebFAQ’s curated FAQ-style data improve robustness against domain shift in multilingual dense retrieval tasks compared to training on synthetic QA pairs?.

## 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 7.2/10.

## 3 Results

15 papers retrieved. 10 claims extracted; 8 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
WebFAQ is used to construct a set of QA-aligned bilingual corpora spanning over 1000 language pairs.	✓	0.24
The constructed bilingual corpora contain a total of 1.5 million aligned QAs.	×	0.10
Each of the 1001 language pairs in the WebFAQ bilingual corpora comprises at least 100 QA pairs.	×	0.11
WebFAQ and all associated resources are publicly available on GitHub and HuggingFace.	✓	0.21
Fine-tuning an in-domain pretrained XLM-RoBERTa model using WebFAQ data leads to substantial performance gains that gene	✓	0.24
The CCQA dataset comprises approximately 55M unique QAs, including 24M English samples.	✓	0.19
The CCQA dataset was gathered from 13 distinct web snapshots.	✓	0.16
The WMT 2019 dataset contains 124M bitext pairs spanning nine language combinations.	✓	0.19
The BUCC 2018 dataset contains 35k bitext pairs in four language combinations.	✓	0.15
GEMBA is a GPT-based metric for translation evaluation that assesses translation quality on par with human evaluators.	✓	0.18

## References

- <http://arxiv.org/abs/2311.05800v2>
- <http://arxiv.org/abs/2602.17327v1>
- <http://arxiv.org/abs/2502.20936v1>