Impact of Training Language Scaling in WebFAQ on Zero-Shot Retrieval for Unseen Non-English Subsets

Assignee Research

June 11, 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 49 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: To what extent does scaling the number of training languages in WebFAQ improve zero-shot retrieval performance on unseen non-English subsets?.

2 Methodology

Systematic literature search across multiple databases yielded 3 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 9.2/10.

3 Results

3 papers retrieved. 10 claims extracted; 10 independently verified. Quality review score: 9.2/10.

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 a large-scale collection of open-domain question answering datasets derived from FAQ-style schema.org annotati	✓	0.32
WebFAQ consists of 96 million natural question-answer (QA) pairs across 75 languages.	✓	0.28
WebFAQ includes 47 million (49%) non-English samples.	✓	0.19
WebFAQ serves as the foundation for 49 monolingual retrieval benchmarks with a total size of 11.2 million QA pairs (5.9	✓	0.36
WebFAQ datasets are carefully curated through refined filtering and near-duplicate detection, yielding high-quality reso	✓	0.35
Fine-tuning an in-domain pretrained XLM-RoBERTa model using WebFAQ achieves significant retrieval performance gains.	✓	0.24
The retrieval performance gains from fine-tuning with WebFAQ generalize to other multilingual retrieval benchmarks evalu	✓	0.23
WebFAQ is used to construct a set of QA-aligned bilingual corpora spanning over 1000 language pairs using state-of-the-a	✓	0.36
The resulting bilingual corpora from WebFAQ demonstrate higher translation quality compared to similar datasets.	✓	0.26
WebFAQ and all associated resources are publicly available on GitHub and HuggingFace.	✓	0.20

References

- https://openalex.org/W7162606467

- https://openalex.org/W7139144950
- https://doi.org/10.1145/3726302.3731934