Scaling WebFAQ 2.0 Dataset Size and Its Impact on MTEB Retrievers for Low-Resource Languages

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 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: How does the scaling of WebFAQ 2.0's dataset size (198M vs. smaller subsets) influence the trade-off between MTEB retrieval scores and inference efficiency of dense retrievers in low-resource languages?.

2 Methodology

Systematic literature search across multiple databases yielded 8 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.7/10.

3 Results

8 papers retrieved. 20 claims extracted; 16 independently verified. Quality review score: 7.7/10.

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
WebFAQ is used to construct a set of QA-aligned bilingual corpora spanning over 1000 language pairs.	✓	0.24
The construction of WebFAQ bilingual corpora utilizes state-of-the-art bitext mining and automated LLM-assessed translat	✓	0.22
The resulting WebFAQ bilingual corpora demonstrate higher translation quality compared to similar datasets.	✓	0.24
WebFAQ and all associated resources are publicly available on GitHub and HuggingFace.	✓	0.21
Dataset-specific fine-tuning applied to an in-domain pretrained XLM-RoBERTa model using WebFAQ data achieves substantial	✓	0.28
Performance gains from fine-tuning on WebFAQ data generalize to other multilingual retrieval datasets.	✓	0.17
The authors constructed 1,001 bilingual datasets containing a total of 1.5 million aligned QAs.	×	0.14
Each of the 1001 language pairs in the WebFAQ bilingual datasets comprises at least 100 QA pairs.	×	0.13
The aligned text sequences of the final WebFAQ bitext corpora exhibit high translation quality compared to human-curated	✓	0.23
The Web Data Commons (WDC) project focuses on large-scale extraction of structured data from the Common Crawl corpus.	✓	0.21
WDC extracts data by parsing and organizing schema.org annotations embedded as JSON-LD, Microdata, RDF, or Microformats.	✓	0.18
CCQA is an open-domain question answering dataset from Meta AI utilizing QA pairs extracted from Common Crawl.	✓	0.24
CCQA comprises approximately 55 million unique QAs, including 24 million English samples.	×	0.12
CCQA data was gathered from 13 distinct web snapshots.	✓	0.16
Huber et al. demonstrated the effectiveness of CCQA for in-domain pre-training on Closed-Book Question Answering (CBQA)	✓	0.27
Kocmi et al. introduced GEMBA, a GPT-based metric for translation evaluation.	✓	0.16
GEMBA demonstrates that LLMs can assess translation quality on par with human evaluators.	✓	0.16
WMT 2019 is a dataset of 124 million bitext pairs spanning nine language combinations.	✓	0.16
Tatoeba is a community-driven collection of sen	✓	0.16

References

- http://arxiv.org/abs/2602.17327v1
- http://arxiv.org/abs/2502.20936v1
- http://arxiv.org/abs/2605.07210v2