Retrieval Accuracy Gap Between Dense and Sparse Models Across Low- and High-Resource Languages in WebFAQ

Assignee Research

June 11, 2026

Abstract

Abstract Progress in cross-lingual modeling depends on challenging, realistic, and diverse evaluation sets. We introduce Multilingual Knowledge Questions and Answers (MKQA), an open- domain question answering evaluation set comprising 10k question-answer pairs aligned across 26 typologically diverse languages (260k question-answer pairs in total). Answers are based on heavily curated, language- independent data representation, making results comparable across languages and independent of language-specific passages. With 26 languages, this dataset supplies the widest range of languages to-date

1 Introduction

This paper examines: MKQA: A Linguistically Diverse Benchmark for Multilingual Open Domain Question Answering. Research question: How does the retrieval accuracy gap between dense and sparse models vary across low-resource versus high-resource languages in the WebFAQ benchmark?.

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 8.7/10.

3 Results

13 papers retrieved. 5 claims extracted; 5 independently verified. Quality review score: 8.7/10.

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
MKQA is an open-domain question answering evaluation set comprising 10k question-answer pairs aligned across 26 typologi	✓	0.52
Answers in MKQA are based on heavily curated, language-independent data representation, making results comparable across	✓	0.39
MKQA supplies the widest range of languages to-date for evaluating question answering, with 26 languages.	✓	0.32
The paper benchmarks a variety of state-of-the-art methods and baselines for generative and extractive question answerin	✓	0.37
Results from the benchmarking indicate that the MKQA dataset is challenging even in English, but especially in low-resou	✓	0.25

References

- https://doi.org/10.18653/v1/2024.acl-long.642
- https://doi.org/10.1007/s11704-026-60308-3
- https://doi.org/10.1162/tacl_a_00433