Robustness Analysis of Multilingual Dense Retrievers Across BEIR Domains

Assignee Research

June 11, 2026

Abstract

Zero-shot evaluation of information retrieval (IR) models is often performed using BEIR; a large and heterogeneous benchmark composed of multiple datasets, covering different retrieval tasks across various domains. Although BEIR has become a standard benchmark for the zero-shot setup, its exclusively English content reduces its utility for underrepresented languages in IR, including Dutch. To address this limitation and encourage the development of Dutch IR models, we introduce BEIR-NL by automatically translating the publicly accessible BEIR datasets into Dutch. Using BEIR-NL, we evaluated a

1 Introduction

This paper examines: BEIR-NL: Zero-shot Information Retrieval Benchmark for the Dutch Language. Research question: How does the robustness of multilingual dense retrievers, trained on SWIM-IR with different language subsets, vary across different domains (e.g., biomedical, legal, news) in the BEIR benchmark when measured by retrieval accuracy and latency?.

2 Methodology

Systematic literature search across multiple databases yielded 9 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.4/10.

3 Results

9 papers retrieved. 23 claims extracted; 14 independently verified. Quality review score: 7.4/10.

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
BEIR measures the performance of textual embeddings on a broad range of tasks.	×	0.15
The main limitation of BEIR lies in its monolingual structure.	×	0.13
BEIR-NL was created by translating datasets from BEIR into Dutch.	×	0.13
BEIR-NL facilitates zero-shot Information Retrieval (IR) evaluation.	✓	0.20
The BEIR-NL benchmark is available on the Hugging Face hub.	✓	0.19
BEIR-NL inherits the same licenses as the datasets from BEIR.	✓	0.16
Compiling existing human-annotated datasets into benchmarks requires substantial time and financial investment.	×	0.14
Automatically translating existing benchmarks is faster and more cost-effective than compiling human-annotated datasets.	×	0.13
Lai et al. (2023) utilized ChatGPT to translate ARC, HellaSwag, and MMLU datasets.	✓	0.21
Vanroy (2023) extended datasets including ARC, HellaSwag, MMLU, and TruthfulQA to Dutch using ChatGPT.	✓	0.16
Thellmann et al. (2024) translated a collection of benchmarking datasets into 21 European languages using DeepL.	✓	0.26
Xiao et al. (2023) extended MTEB with 35 publicly-available Chinese datasets.	✓	0.24
MTEB-French added 18 datasets in French to MTEB.	✓	0.18
The e5-multilingual-small model has 118M parameters and an output embedding dimension of 384.	×	0.13
The e5-multilingual-base model is based on XLMRoberta-base and has 278M parameters.	✓	0.20
The e5-multilingual-large model has a maximum input length of 512 tokens.	×	0.12
The gte-multilingual-base model supports a maximum input length of 8192 tokens.	×	0.13
The jina-embeddings-v3 model has 572M parameters and is fine-tuned from XLMRoberta-large.	✓	0.16
The dpr-xm model uses 277M parameters during inference.	×	0.13
LEALLA-small and LEALLA-base are distilled from LaBSE.	✓	0.15
LaBSE and gte-multilingual-base are trained from scratch.	✓	0.20
The bge-reranker-v2-m3 model is based on the bge-m3 architecture.	✓	0.16
Cosine similarity is employed to score similarity	✓	0.18

References

- http://arxiv.org/abs/2412.08329v1
- http://arxiv.org/abs/2311.05800v2
- http://arxiv.org/abs/2509.22472v1