Performance of Dense Retrievers on Synthetic Typo-Injected Dutch Data vs. English Baselines in BEIR-NL Benchmark

Assignee Research

June 12, 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 performance of dense retrievers trained on synthetic typo-injected Dutch data compare to English baselines on the BEIR-NL benchmark across low-resource domains?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.5/10.

3 Results

12 papers retrieved. 22 claims extracted; 14 independently verified. Quality review score: 7.5/10.

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-NL is a zero-shot information retrieval benchmark for the Dutch language.	✓	0.26
BEIR-NL facilitates zero-shot IR evaluation and supports the development of retrieval models tailored to Dutch.	✓	0.22
BEIR-NL is available on the Hugging Face hub.	✓	0.18
BEIR-NL inherits the same licenses as the datasets from BEIR.	✓	0.16
The BEIR-NL benchmark was created by translating datasets from BEIR into Dutch.	✓	0.16
Extensive evaluations of small and mid-range multilingual IR models, which support Dutch, were conducted, including dens	✓	0.24
The e5-multilingual-small model has 118M parameters, a dimension of 384, a max input of 512, and is IR finetuned.	×	0.13
The e5-multilingual-base model has 278M parameters, a dimension of 768, a max input of 512, and is IR finetuned.	✓	0.17
The e5-multilingual-large model has 560M parameters, a dimension of 1024, a max input of 512, and is IR finetuned.	✓	0.16
The e5-multilingual-large-instruct model has 560M parameters, a dimension of 1024, a max input of 512, and is IR finetun	✓	0.17
The gte-multilingual-base model has 305M parameters, a dimension of 768, a max input of 8192, and is IR finetuned.	✓	0.16
The jina-embeddings-v3 model has 572M parameters, a dimension of 1024, a max input of 8192, and is IR finetuned.	×	0.13
The bge-m3 model has 568M parameters, a dimension of 1024, a max input of 8192, and is IR finetuned.	×	0.13
The dpr-xm model has 852M parameters (277M during inference), a dimension of 768, a max input of 512, and is IR finetune	×	0.12
The LEALLA-small model has 69M parameters, a dimension of 128, a max input of 512, and is not IR finetuned.	×	0.10
The LEALLA-base model has 107M parameters, a dimension of 192, a max input of 512, and is not IR finetuned.	×	0.11
The LaBSE model has 471M parameters, a dimension of 768, a max input of 512, and is not IR finetuned.	×	0.11
The mContriever model has 179M parameters, a dimension of 768, a max input of 512, and is not IR finetuned.	×	0.10
The bge-reranker-v2-m3 model has 568M pa-	✓	0.16

References

- http://arxiv.org/abs/2412.08329v1
- http://arxiv.org/abs/2301.12566v1
- http://arxiv.org/abs/2410.13153v1