BEIR-NL: Zero-shot Information Retrieval Benchmark for the Dutch Language

Nikolay Banar* **Ehsan Lotfi*** **Walter Daelemans**

CLiPS, University of Antwerp, Belgium
{nicolae.banari, ehsan.lotfi, walter.daelemans}@uantwerpen.be

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 wide range of multilingual dense ranking and reranking models, as well as the lexical BM25 method. Our experiments show that BM25 remains a competitive baseline, and is only outperformed by the larger dense models trained for retrieval. When combined with reranking models, BM25 achieves performance on par with the best dense ranking models. In addition, we explored the impact of translation on the data by back-translating a selection of datasets to English, and observed a performance drop for both dense and lexical methods, indicating the limitations of translation for creating benchmarks. BEIR-NL is publicly available on the Hugging Face hub[1].

1 Introduction

An increasing number of natural language processing (NLP) tasks require an information retrieval (IR) step to identify relevant pieces of text in a large corpus of documents. Therefore, IR models are crucial in various use cases, including question-answering (Chen et al., 2017), claim-verification (Thorne et al., 2018), and retrieval-augmented generation (Lewis et al., 2020).

Recently, IR has witnessed significant progress, driven mainly by advancements in large language models (LLMs; Zhao et al., 2024). Pre-trained on large corpora, these models can generate high-quality contextualized textual embeddings that capture semantic relationships beyond surface-level features like keywords. The produced vector representations demonstrate strong performance in IR tasks, as well as in other problems (Muennighoff et al., 2023) such as classification and clustering.

Benchmarking and evaluating such models is essential in sustaining advances in NLP research. Comprehensive benchmarks provide a standardized framework to assess the performance of models, identify their limitations, and guide the direction of future work. BEIR (Benchmarking IR; Thakur et al., 2021) was introduced to address this need in IR and became a standard benchmark in zero-shot evaluation, enabling the comparison of retrieval models in a unified framework. BEIR offers a diverse and heterogeneous collection of datasets covering various domains from biomedical and financial texts to general web content, and recently has been integrated into the broader MTEB benchmark (Massive Text Embedding Benchmark; Muennighoff et al., 2023), which measures the performance of textual embeddings on a broad range of tasks. While BEIR has substantially advanced the evaluation of IR models, its main limitation lies in the monolingual structure, which restricts its application for other languages.

In this work, we focus on extending the BEIR benchmark to Dutch, a resource-scarce language in IR research. By translating datasets from BEIR into Dutch, we aim to provide a foundation for evaluating IR models in this language. Our benchmark BEIR-NL facilitates zero-shot IR evaluation and supports the development of retrieval models tailored to Dutch. In addition, we conduct extensive evaluations of small and mid-range multilingual IR models, which support Dutch, including dense ranking and reranking models. We make the BEIR-NL benchmark available on the Hugging Face hub,

*indicates equal contribution

[1]https://huggingface.co/collections/clips/beir-nl-6756c81a8ebab4432d922a08

Proceedings of the 18th Building and Using Comparable Corpora Workshop, pages 36–45
January 20, 2025. ©2025 Association for Computational Linguistics

ensuring that it inherits the same licenses as the datasets from BEIR (Appendix A).

2 Related Work

Recently, increasing efforts have been directed towards extending English or multilingual benchmarks to cover more languages. These efforts are primarily divided into two categories: (i) the existing (or to-be) human-annotated datasets are compiled into benchmarks, or (ii) existing benchmarks are automatically translated into new languages. The first approach provides high-quality datasets but requires substantial time and financial investment. The second approach is faster and more cost-effective, but the quality of translations can affect the overall quality of the benchmark and potentially lead to inaccurate model evaluations (Engländer et al., 2024). However, the recent availability of relatively cheap and high-quality machine translation solutions (thanks mainly to the LLM developments and advances) has made this an attractive and commercially feasible option, especially for large datasets and benchmarks. Below we outline relevant work focused on extending existing benchmarks to additional languages.

In generative benchmarking, Lai et al. (2023) utilized ChatGPT to translate three widely-used benchmark datasets for LLMs into 26 languages, to evaluate the performance of models for the Okapi framework. These datasets include ARC (Clark et al., 2018), HellaSwag (Zellers et al., 2019), and MMLU (Hendrycks et al.). Vanroy (2023) extended these datasets, along with TruthfulQA (Lin et al., 2022), to Dutch using ChatGPT. Subsequently, Thellmann et al. (2024) added GSM8K (Cobbe et al., 2021) to the mentioned benchmarking datasets and translated the entire collection into 21 European languages using DeepL.

Another branch of work focuses on extending MTEB (Muennighoff et al., 2023), which evaluates the quality of textual embeddings across multiple tasks. Xiao et al. (2023) extended this benchmark to Chinese (C-MTEB) by collecting 35 publicly-available Chinese datasets. MTEB-French (Ciancone et al., 2024) added 18 datasets in French to MTEB, including both original and DeepL-translated data. Building on MTEB, Wehrli et al. (2024) introduced six benchmarking datasets for clustering text embeddings in German. A Polish version, MTEB-PL (Poświata et al., 2024), consists of 28 datasets, with its retrieval part sourced from BEIR-PL (Wojtasik et al., 2024). ruMTEB (Snegirev et al., 2024) comprises 23 tasks in the MTEB format, with primarily original datasets in Russian, and with one translated using DeepL. SEB (Scandinavian Embedding Benchmark; Enevoldsen et al., 2024) represents 24 evaluation tasks for Scandinavian languages, incorporating a portion of existing translated datasets from MTEB.

Finally in IR, mMARCO (Bonifacio et al., 2021) extended the popular MSMARCO dataset (Bajaj et al., 2016) to multiple languages by translating queries and passages using Google Translate and Helsinki-NLP models (Tiedemann and Thottingal, 2020). Most related to our work, BEIR-PL (Wojtasik et al., 2024) translated a subset of the BEIR benchmark to Polish using Google Translate.

These efforts highlight the necessity of extending existing benchmarks to a multilingual context, enabling the evaluation of models across a wide range of languages. Building on the previous work, our study extends the BEIR benchmark to Dutch using machine translation, providing a valuable resource for evaluating IR models in this language.

3 Dataset

The original BEIR benchmark (Thakur et al., 2021) comprises 18 datasets, covering 9 different information retrieval tasks. Of these, 4 datasets are not publicly available, and therefore are removed from our selection for BEIR-NL. The remaining 14 datasets are listed in Table 1 along with their selected features and statistics. Since most retrieval models are trained on MSMARCO (Bajaj et al., 2016), we also report on its Dutch-translated version from mMARCO (Bonifacio et al., 2021), but do not include it for translation. We refer the reader to the BEIR paper (Thakur et al., 2021) for further descriptions and more details on each dataset.

3.1 Translation

The next step is translating the selected 14 datasets from English to Dutch. After considering commonly used options, we opted for Gemini-1.5-flash[2] which offers a good balance of speed, cost, and translation quality. We prompted the model to translate the inputs, providing it with the input type (query or document), and domain (4th column in Table 1) as context. We used the API in batch mode, which lowers the total cost to less than 450

[2]A small portion of translations were done using GPT-4o-mini and Google Translate, as Gemini declined to translate certain content and had occasional issues with tags in prompts.

Task	Dataset	Source	Domain	#Queries	#Docs	Avg. D/Q
Biomedical IR	TREC-COVID	Voorhees et al. (2021)	Biomedical	50	171K	493.5
	NFCorpus	Boteva et al. (2016)	Biomedical	323	3.63K	38.2
Question Answering	NQ	Kwiatkowski et al. (2019)	Wikipedia	3,452	2.68M	1.2
	HotpotQA	Yang et al. (2018)	Wikipedia	7,405	5.23M	2.0
	FiQA-2018	Maia et al. (2018)	Financial	648	57.6K	2.6
Argument Retrieval	ArguAna	Wachsmuth et al. (2018)	Miscellaneous	1,406	8.67K	1.0
	Touche-2020	Bondarenko et al. (2020)	Miscellaneous	49	383K	19.0
Duplicate-Question Retrieval	CQADupstack	Hoogeveen et al. (2015)	StackExchange	13,145	457K	1.4
	Quora	Thakur et al. (2021)	Quora	10,000	522K	1.6
Entity Retrieval	DBPedia	Hasibi et al. (2017)	Wikipedia	400	4.64M	38.2
Citation Prediction	SciDocs	Cohan et al. (2020)	Scientific	1,000	25.7K	4.9
Fact Checking	SciFact	Wadden et al. (2020)	Scientific	300	5.18K	1.1
	FEVER	Thorne et al. (2018)	Wikipedia	6,666	5.42M	1.2
	Climate-FEVER	Diggelmann et al. (2020)	Wikipedia	1,535	5.42M	3.0
Passage Retrieval	mMARCO	Bonifacio et al. (2021)	Miscellaneous	6,980	8.84M	1.1

Table 1: Statistics of datasets included in the BEIR-NL benchmark (plus mMARCO). The table highlights the number of queries and documents, as well as the average number of relevant documents per query (Avg. D/Q) (from Thakur et al. (2021)).

Euro. The exact prompts can be found in Appendix B.

To assess the translation quality, we randomly sampled 10 items from each dataset (140 in total) and asked a native Dutch speaker to check the translations against the original English text, and annotate instances for major (i.e. translation includes semantic addition or omission) or minor (i.e. translation is correct but too literal) issues. The results show major and minor issues in 2.2% and 14.8% of samples respectively, which means that almost 98% of the translated samples can be trusted for semantic accuracy. We will revisit this issue in the discussion section.

4 Experimental Setup

This section provides an overview of the experimental setup used to assess the performance of different models on BEIR-NL. We mostly follow the BEIR official repository[3] for zero-shot evaluation, using the provided code as much as possible but occasionally adapt it to specific requirements of the evaluated models. In the following, we describe the models, data processing steps, and evaluation metrics used in our experiments.

4.1 Models

We include models from three categories: lexical models, dense ranking models, and dense reranking models.

[3] https://github.com/beir-cellar/beir

4.1.1 Lexical models

As the most popular lexical retrieval solution, BM25 (Robertson et al., 1994) relies on keyword matching and utilizes empirical word (or token) weighting schemes to determine the relevance of documents to a given query. Despite lexical gap issues, where the vocabulary used in queries can differ from that of relevant documents, BM25 remains a robust baseline for many retrieval tasks and was outperformed only recently by E5 (Wang et al., 2022) on the BEIR retrieval benchmark (Thakur et al., 2021) in zero-shot setting. Similarly to Wojtasik et al. (2024), we utilize the BM25 implementation from Elasticsearch for Dutch.

4.1.2 Dense ranking models

Dense ranking (or embedding) models encode an input sequence into a dense vector, which can be used to calculate similarity or relevance between sequences (query and document in our case). Inspired by recent related studies and the MTEB leaderboard[4], we select the following multilingual retrieval models for our zero-shot experiments[5]: mContriever (Izacard et al., 2022), LaBSE (Feng et al., 2022), LEALLA (Mao and Nakagawa, 2023), mE5 (Wang et al., 2024), BGE-M3 (Chen et al., 2024), DPR-XM (Louis et al., 2024), jina-embeddings-v3 (Sturua et al., 2024), and mGTE (Zhang et al., 2024). Table 2 lists these models along with a number of relevant features. Follow-

[4] https://huggingface.co/spaces/mteb/leaderboard

[5] Due to computational limitations, we exclude larger models like e5-mistral-7b-instruct and bge-multilingual-gemma2.

Model	Based on	#Parameters	Dim	Max input	IR Finetuned
e5-multilingual-small	Multilingual-MiniLM	118M	384	512	Yes
e5-multilingual-base	XLMRoberta-base	278M	768	512	Yes
e5-multilingual-large	XLMRoberta-large	560M	1024	512	Yes
e5-multilingual-large-instruct	XLMRoberta-large	560M	1024	512	Yes
gte-multilingual-base	-	305M	768	8192	Yes
jina-embeddings-v3	XLMRoberta-large	572M	1024	8192	Yes
bge-m3	XLMRoberta-large	568M	1024	8192	Yes
dpr-xm	XMOD	852M (277M†)	768	512	Yes
LEALLA-small	LaBSE (distilled)	69M	128	512	No
LEALLA-base	LaBSE (distilled)	107M	192	512	No
LaBSE	-	471M	768	512	No
mContriever	Bert-multilingual-base	179M	768	512	No
bge-reranker-v2-m3	bge-m3	568M	1024	8192	Yes
jina-reranker-v2-base-multilingual	XLMRoberta-base	278M	768	1024	Yes
gte-multilingual-reranker-base	gte-multilingual-base	305M	768	8192	Yes

Table 2: Dense ranking (top) and reranking (bottom) models used in our experiments. 'Dim' is the dimension of the output embedding vector. LaBSE and gte-multilingual-base are trained from scratch. LEALLA is distilled from LaBSE, and the rest are fine-tuned from the model mentioned in the second column. †: dpr-xm is modular and uses 277M parameters during inference.

ing the convention, we do not impose any limits on the input length for these models, allowing them to handle truncation if necessary[6]. In all cases, cosine similarity is employed to score similarity between the normalized embeddings.

4.1.3 Zero-shot reranking models

Unlike ranking models that are employed in a bi-encoder setting, reranking models rely on cross-encoding the query and document, which can provide more accurate results at a higher computational cost. Consequently, reranking models are usually applied on the top outputs of a fast ranking model such as BM25.

We examine three popular multilingual reranking models, namely bge-reranker-v2-m3 (Chen et al., 2024), jina-reranker-v2-base-multilingual (Sturua et al., 2024), and gte-multilingual-reranker-base (Zhang et al., 2024) (see Table 2-bottom). Following the convention (Thakur et al., 2021), we apply these models on the top-100 documents retrieved by BM25, and evaluate the reranked output. We do not restrict the input length for the reranking models, leaving them to manage truncation.

4.2 Metrics

To assess the performance of our models, we employ two standard retrieval metrics: nDCG@10 and Recall@100. NDCG (normalized discounted cumulative gain) is a ranking-aware metric often used to report retrieval performance, especially on graded (non-binary) labels (Thakur et al., 2021). We also report recall, which, although ranking-agnostic, is a useful and relevant metric for practical settings like retrieval-augmented generation.

[6]Considering the average document length in BEIR datasets, truncation is rarely needed for any of these models.

5 Results and Discussion

5.1 Retrieval Performance on BEIR-NL

Table 3 shows the retrieval performance of the selected models on the 14 subsets of BEIR-NL, in addition to MSMARCO. As mentioned before, MSMARCO is not part of our dataset, but considering its popularity in retrieval training, we include it in the evaluations (based on the Dutch-translated version from mMARCO (Bonifacio et al., 2021)).

The results show that BM25 still provides a competitive baseline, and in many cases is only outperformed by the larger dense models. The four recently released multilingual-e5-large-instruct, gte-multilingual-base, jina-embeddings-v3 and bge-m3 achieve the best overall performances, with multilingual-e5-large-instruct getting the highest Recall@100 on half of the datasets. We also observe a sizeable gap between the older 'sentence embedding' models, and the new generation of trained-for-retrieval models (see the last column in Table 2), with the latter achieving substantially higher results. However, based on their published metadata, the majority of these models have been at least partially exposed to BEIR datasets in their training process, which makes the comparison unfair (The corresponding *potentially inflated* results

Model		MSMARCO	TREC-COVID	NFCorpus	NQ	HotpotQA	FiQA-2018	ArguAna	Touche-2020	CQADupstack	Quora	DBPedia	SciDocs	SciFact	FEVER	Climate-FEVER
BM25	NDCG@10	16.87	63.37	30.54	25.09	53.62	18.73	41.76	28.15	27.77	65.92	25.46	11.44	61.13	60.65	12.09
multilingual-e5-small	NDCG@10	30.85†	41.74	24.10	27.03†	53.30†	20.39	44.76	16.04	28.51	79.85†	25.89	6.58	58.82	56.69†	14.08
multilingual-e5-base	NDCG@10	32.79†	40.68	24.17	36.06†	60.87†	23.76	47.06	10.29	30.36	81.02†	28.74	10.53	67.23	58.52†	16.31
multilingual-e5-large	NDCG@10	**37.51**†	69.72	28.06	49.15†	67.95†	31.84	48.90	22.18	31.92	82.01†	**38.67**	11.95	68.38	72.73†	13.76
multilingual-e5-large-instruct	NDCG@10	34.35†	71.22	31.08	**55.79**†	65.97†	**37.93**	50.32	26.67	36.95	83.54†	38.24	**18.07**	69.10	79.39†	21.05
gte-multilingual-base	NDCG@10	27.19†	53.36	27.97†	47.42†	58.53†	29.45	**52.85**†	22.60 †	31.59†	81.25†	36.46†	15.86	64.41	82.68†	17.53
jina-embeddings-v3	NDCG@10	26.05†	54.46	29.84	37.26†	51.82	35.71	52.23	15.05	36.16	82.92	30.71	18.42	64.90	68.88	19.54
bge-m3	NDCG@10	31.96†	48.22	27.90	51.92†	65.20†	32.60	52.16	22.68	34.75	**83.72**	35.46	14.41	62.83	76.08	**26.39**
dpr-xm	NDCG@10	28.46†	40.86	18.58	28.56	26.34	13.98	26.91	15.99	18.73	74.70	21.07	8.64	34.29	49.46	11.16
LEALLA-small	NDCG@10	3.95	13.32	5.56	5.11	12.18	3.41	19.25	5.65	13.14	68.50	9.60	3.70	12.98	7.08	0.34
LEALLA-base	NDCG@10	5.60	14.44	6.09	7.77	17.46	3.75	24.97	5.00	14.34	70.87	13.40	3.09	7.13	7.46	1.15
LaBSE	NDCG@10	6.87	18.50	13.54	11.24	18.64	7.38	39.15	4.67	19.66	75.55	15.27	6.32	39.07	12.51	3.85
mContriever	NDCG@10	7.46†	17.51	13.36	10.50	27.84	5.41	39.60	6.15	12.81	72.90	15.58	4.93	37.89	21.51	3.08
BM25 + bge-reranker	NDCG@10	31.80†	76.47	**33.78**	51.28†	**71.78**†	30.41	47.27	**33.78**	31.70	76.81	37.84	13.88	69.94	84.17	25.60
BM25 + jina-reranker	NDCG@10	31.93†	**76.83**	33.19	49.07†	70.57	30.86	48.53	30.96	34.06	79.44	36.26	14.49	**70.68**	**85.17**	22.56
BM25 + gte-reranker	NDCG@10	28.90†	76.24	28.26†	47.85†	70.43†	24.13	46.74†	28.26†	25.69†	74.95†	36.67†	13.22	68.37	85.13†	22.96
BM25	Recall@100	51.20	10.52	22.16	65.57	70.54	42.83	92.32	44.16	54.77	88.66	36.92	26.49	83.42	89.20	30.42
multilingual-e5-small	Recall@100	74.63†	7.89	23.56	60.70†	69.45†	47.10	94.59	38.18	56.99	97.51†	35.83	22.93	87.67	85.83†	40.47
multilingual-e5-base	Recall@100	77.39†	6.58	22.09	73.61†	76.24†	55.02	95.59	32.96	60.65	97.93†	39.40	29.78	91.00	89.98†	42.69
multilingual-e5-large	Recall@100	**82.71**†	13.31	27.34	83.49†	**82.21**†	61.81	96.37	43.65	63.30	98.66†	47.26	30.42	92.27	93.08†	32.68
multilingual-e5-large-instruct	Recall@100	80.89†	**14.48**	**28.88**	**92.39**†	80.55†	68.70	98.86	46.97	70.56	98.83†	**49.66**	40.80	**93.67**	**94.53**†	**46.05**
gte-multilingual-base	Recall@100	70.29†	10.74	27.89†	85.39†	70.08†	61.53	97.87†	41.12†	66.14†	98.12†	44.11†	37.43	91.00	94.32†	40.40
jina-embeddings-v3	Recall@100	73.43†	11.74	26.50	84.43†	68.04	**69.98**	**98.93**	37.69	**72.62**	98.58	42.22	**42.64**	91.17	93.04	44.98
bge-m3	Recall@100	77.71†	9.43	25.20	89.62†	80.20†	63.41	97.44	**48.70**	66.89	**98.85**	46.30	35.02	91.93	94.11	56.54
dpr-xm	Recall@100	67.77†	5.78	17.95	62.42	38.31	33.81	78.73	36.46	41.94	93.41	22.25	19.36	67.26	76.17	28.54
LEALLA-small	Recall@100	15.99	1.44	9.12	19.99	23.48	12.19	56.47	9.89	32.58	91.41	13.38	12.62	42.81	14.79	1.81
LEALLA-base	Recall@100	22.12	1.61	9.92	27.45	30.32	13.04	61.30	8.39	33.89	93.13	18.80	10.73	34.18	14.97	2.61
LaBSE	Recall@100	26.71	1.97	16.05	41.68	33.56	25.57	87.98	10.09	47.06	95.87	22.91	21.50	74.67	36.48	15.24
mContriever	Recall@100	32.06†	1.71	16.81	40.42	45.97	20.36	91.61	12.06	35.91	94.48	25.25	18.56	74.24	48.31	10.29

Table 3: Performance of selected models on the BEIR-NL benchmark (plus MSMARCO), measured by NDCG@10 (top) and Recall@100 (bottom).† indicates results that are (or are highly likely to be) inflated because of potential contamination of the model with in-domain data for a given dataset, based on available descriptions from the corresponding work (i.e. they are highly unlikely to be zero-shot). bge-reranker, jina-reranker, and gte-reranker refer to bge-reranker-v2-m3, jina-reranker-v2-base-multilingual, and gte-multilingual-reranker models, respectively.

are marked with a † in the table.). In other words, in these cases the evaluation could not be considered proper zero-shot.

Finally, the last three rows of the top section in Table 3 (NDCG@10 results) show the performance of the reranking models when used in combination with BM25 as the first-step ranker. As demonstrated, this approach can often offer a competitive edge over the best ranking models.

5.2 Comparison with BEIR and BEIR-PL

Since BEIR-NL is a translated benchmark, we can compare the performance of the retrieval methods on parallel subsets in different languages, including the (translated) Polish version, BEIR-PL (Wojtasik et al., 2024).

Tables 4 and 5 show this comparison for BM25 and gte-multilingual-base, across the subsets for which performance data is publicly available[7]. As Table 4 reveals, BM25 performs comparably on BEIR-NL and BEIR-PL subsets, with a marginal overall advantage for BEIR-NL. However, these numbers lag behind the BM25 performance on the original BEIR dataset by 6-7 points in NDCG@10 and Recall@100. One potential reason for this drop is the lexical mismatch between the translated query and relevant passages since queries and passages are translated independently[8] (Bonifacio et al., 2021). Table 5 shows that the performance difference persists with dense models (e.g. gte-multilingual-base). Here, the discrepancy can be attributed to both the data (translation quality) and model (higher competence in English compared to other languages).

5.3 Impact of Translation

To isolate the semantic effect of translation (from that of the model/language) we back-translate a subset of 5 BEIR-NL datasets to English using the same translation pipeline, and compare the performance of lexical and dense models on this version against the original one. Table 6 shows the results (NDCG@10), which indicate an average drop of 1.9 and 2.6 points for the lexical (BM25) and dense model (gte-multilingual-base) respectively. Since the model-language competence factor is absent here, this drop can be considered a proxy for the impact of translation on the benchmark quality and/or reliability.

[7]BEIR-PL only covers 10 of the 14 public BEIR datasets.

[8]Assuming a uniform BM25 performance for different languages, which is not trivial.

6 Conclusions and Future Work

In this work, we introduced BEIR-NL, an automatically translated version of the BEIR benchmark into Dutch, which aims to address the need for the evaluation of IR models in this language. Using BEIR-NL, we conducted extensive zero-shot evaluations for various models, including one lexical model as well as small and mid-range dense retrieval and reranking models. These experiments showed that larger dense IR models generally outperform BM25, while BM25 remains a competitive baseline for smaller models. Furthermore, combining BM25 with reranking models results in performance comparable to the best dense retrieval models.

We also observed several challenges, including the impact of translation on retrieval performance and the risk of in-domain data contamination in IR models. These issues might affect the reliability of zero-shot evaluations on this benchmark and highlight the need for creating native Dutch resources, which we leave for future work.

BEIR-NL fills a critical gap in the evaluation of Dutch IR models and sets a foundation for further development of IR benchmarks in Dutch. By making BEIR-NL publicly available, we aim to support future research and encourage the development of retrieval models for this language.

Limitations

Besides the issues originated from translation (which we briefly addressed before), here we discuss other important limitations pertinent to this work.

Native Dutch Resources. While BEIR-NL provides a benchmark for evaluating IR models in Dutch, it relies on translations from the original BEIR, which is exclusively in English. This lack of native Dutch datasets limits the ability of BEIR-NL to fully represent and reflect the linguistic nuances and cultural context of the language, and therefore the complexities of Dutch IR, especially in domain-specific contexts with local terminology and knowledge.

Data Contamination. Many modern IR models are trained on massive corpora that might include content from BEIR. Table 3 indicates multiple models that have (or might have) been exposed to in-domain contamination for a given dataset. This can result in inflated performance –as models might have already seen the relevant data during

Metric	Benchmark	TREC-COVID	NFCorpus	NQ	HotpotQA	FiQA-2018	ArguAna	CQADupstack	DBPedia	SciDocs	SciFact	Average
NDCG@10	BEIR-NL	63.4	30.5	25.1	53.6	18.7	41.8	27.8	25.6	11.4	61.1	35.9
	BEIR-PL	61.0	31.9	20.1	49.2	19.0	41.4	28.4	22.9	14.1	62.5	35.1
	BEIR (EN)	68.9	34.3	32.6	60.2	25.4	47.2	32.5	32.1	16.5	69.1	41.9
Recall@100	BEIR-NL	10.5	22.2	65.6	70.5	42.8	92.3	54.8	36.9	26.5	83.4	50.6
	BEIR-PL	10.1	24.6	57.9	67.1	44.1	93.5	53.9	30.1	33.0	88.4	50.3
	BEIR (EN)	11.7	26.0	78.3	76.3	54.9	95.2	62.1	43.5	36.8	92.0	57.7

Table 4: BM25 performance on the overlapping subset of BEIR-NL, BEIR-PL, and original BEIR, for which performance data is publicly available. Results for BEIR-PL and BEIR are from Wojtasik et al. (2024).

Metric	Benchmark	TREC-COVID	NFCorpus	NQ	HotpotQA	FiQA-2018	ArguAna	DBPedia	SciDocs	SciFact	Average
NDCG@10	BEIR-NL	53.4	28.0	47.4	58.5	29.4	52.9	36.5	15.9	64.4	42.9
	BEIR-PL	59.4	26.8	43.1	56.9	29.0	53.2	32.5	14.2	58.9	41.6
	BEIR (EN)	57.6	36.6	58.1	63.0	45.0	58.2	40.1	18.2	73.4	50.0

Table 5: Performance of gte-multilingual-base on the overlapping subset of BEIR-NL, BEIR-PL, and original BEIR, for which performance data is publicly available. Results for BEIR-PL and BEIR are sourced from the MTEB leaderboard.

Model	BEIR	NFCorpus	FiQA-2018	ArguAna	SciDocs	SciFact	Average	Δ_{tr}
BM25	original	34.3	25.4	47.2	16.5	69.1	38.5	-
	back-translated	32.4	22.0	45.2	15.1	68.2	36.6	-1.9
gte-multilingual-base	original	36.7	45.0	58.2	18.2	73.4	46.3	-
	back-translated	32.6	40.7	55.0	18.3	71.7	43.7	-2.6

Table 6: NDCG@10 results for BM25 and gte-multilingual-base on selected datasets from the original BEIR, and their back-translated version (from Dutch to English). Δ_{tr} is the change in average performance due to back translation.

different phases of training– raising concerns about the validity of zero-shot evaluations. Ensuring a truly zero-shot evaluation is a difficult challenge, as many IR models lack transparency regarding the exact composition of training corpora.

Benchmark Validity Over Time. BEIR has become a standard benchmark to evaluate the performance of IR models, attracting a large number of evaluations over time. This extensive usage introduces the risk of overfitting, as researchers might unintentionally train models tailored to perform well on BEIR rather than on broader IR tasks. In addition, advances in IR models and evaluation needs might outpace the benchmark, making it less representative and less relevant. As a result, the relevance and validity of BEIR as well as BEIR-NL may diminish over time.

Acknowledgments

This research received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme. We would like to thank Jens Van Nooten for assessing the quality of the translations. In addition, we acknowledge the use of the GPT-4o model for assisting with error checking and proofreading of this paper.

References

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Alexander Bondarenko, Maik Fröbe, Meriem Beloucif, Lukas Gienapp, Yamen Ajjour, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, et al. 2020. Overview of touché 2020: argument retrieval. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22–25, 2020, Proceedings 11*, pages 384–395. Springer.

Luiz Henrique Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, , Roberto Lotufo, and Rodrigo Nogueira. 2021. mmarco: A multilingual version of ms marco passage ranking dataset. *Preprint*, arXiv:2108.13897.

Vera Boteva, Demian Gholipour, Artem Sokolov, and Stefan Riezler. 2016. A full-text learning to rank dataset for medical information retrieval. In *Advances in Information Retrieval: 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20–23, 2016. Proceedings 38*, pages 716–722. Springer.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Mathieu Ciancone, Imene Kerboua, Marion Schaeffer, and Wissam Siblini. 2024. Mteb-french: Resources for french sentence embedding evaluation and analysis. *arXiv preprint arXiv:2405.20468*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-fever: A dataset for verification of real-world climate claims. *arXiv preprint arXiv:2012.00614*.

Kenneth Enevoldsen, Márton Kardos, Niklas Muennighoff, and Kristoffer Laigaard Nielbo. 2024. The scandinavian embedding benchmarks: Comprehensive assessment of multilingual and monolingual text embedding. *arXiv preprint arXiv:2406.02396*.

Leon Engländer, Hannah Sterz, Clifton Poth, Jonas Pfeiffer, Ilia Kuznetsov, and Iryna Gurevych. 2024. M2qa: Multi-domain multilingual question answering. *arXiv preprint arXiv:2407.01091*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891.

Faegheh Hasibi, Fedor Nikolaev, Chenyan Xiong, Krisztian Balog, Svein Erik Bratsberg, Alexander Kotov, and Jamie Callan. 2017. Dbpedia-entity v2: a test collection for entity search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1265–1268.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Doris Hoogeveen, Karin M Verspoor, and Timothy Baldwin. 2015. Cqadupstack: A benchmark data set for community question-answering research. In *Proceedings of the 20th Australasian document computing symposium*, pages 1–8.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Viet Lai, Chien Nguyen, Nghia Ngo, Thuật Nguyễn, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327.

Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *ArXiv*, abs/2005.11401.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252.

Antoine Louis, Vageesh Saxena, Gijs van Dijck, and Gerasimos Spanakis. 2024. Colbert-xm: A modular multi-vector representation model for zero-shot multilingual information retrieval. *arXiv preprint arXiv:2402.15059*.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: financial opinion mining and question answering. In *Companion proceedings of the the web conference 2018*, pages 1941–1942.

Zhuoyuan Mao and Tetsuji Nakagawa. 2023. LEALLA: Learning lightweight language-agnostic sentence embeddings with knowledge distillation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1886–1894, Dubrovnik, Croatia. Association for Computational Linguistics.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. Mteb: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037.

Rafał Poświata, Sławomir Dadas, and Michał Perełkiewicz. 2024. Pl-mteb: Polish massive text embedding benchmark. *arXiv preprint arXiv:2405.10138*.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994*, volume 500-225 of *NIST Special Publication*, pages 109–126. National Institute of Standards and Technology (NIST).

Artem Snegirev, Maria Tikhonova, Anna Maksimova, Alena Fenogenova, and Alexander Abramov. 2024. The russian-focused embedders' exploration: rumteb benchmark and russian embedding model design. *arXiv preprint arXiv:2408.12503*.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, et al. 2024. jina-embeddings-v3: Multilingual embeddings with task lora. *arXiv preprint arXiv:2409.10173*.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Klaudia Thellmann, Bernhard Stadler, Michael Fromm, Jasper Schulze Buschhoff, Alex Jude, Fabio Barth, Johannes Leveling, Nicolas Flores-Herr, Joachim Köhler, René Jäkel, et al. 2024. Towards multilingual llm evaluation for european languages. *arXiv preprint arXiv:2410.08928*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819.

Jörg Tiedemann and Santhosh Thottingal. 2020. Opus-mt–building open translation services for the world. In *Proceedings of the 22nd annual conference of the European Association for Machine Translation*, pages 479–480.

Bram Vanroy. 2023. Language resources for dutch large language modelling.

Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2021. Trec-covid: constructing a pandemic information retrieval test collection. In *ACM SIGIR Forum*, volume 54, pages 1–12. ACM New York, NY, USA.

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018. Retrieval of the best counterargument without prior topic knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 241–251.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Silvan Wehrli, Bert Arnrich, and Christopher Irrgang. 2024. German text embedding clustering benchmark. *arXiv preprint arXiv:2401.02709*.

Konrad Wojtasik, Kacper Wołowiec, Vadim Shishkin, Arkadiusz Janz, and Maciej Piasecki. 2024. Beir-pl: Zero shot information retrieval benchmark for the polish language. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2149–2160.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighof. 2023. C-pack: Packaged resources to advance general chinese embedding. *arXiv preprint arXiv:2309.07597*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval. *Preprint*, arXiv:2407.19669.

Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. 2024. Dense text retrieval based on pretrained language models: A survey. *ACM Transactions on Information Systems*, 42(4):1–60.

A Appendix: Licenses

The BEIR repository on Hugging Face[9] reports that the following datasets are distributed under the CC BY-SA 4.0 license: NFCorpus, FiQA-2018, Quora, Climate-Fever, FEVER, NQ, DBPedia, ArguAna, Touché-2020, SciFact, SCIDOCS, HotpotQA, TREC-COVID. The only one exception is CQADupStack[10] with the Apache License 2.0 license.

B Appendix: Translation Prompts

We prompt Gemini-1.5-flash with the following instructions (temperature = 0).

Query Prompt:"Translate to English the QUERY from the {domain} domain. Provide only the translation. QUERY:\n ['{query}']".

Document Prompt:"Translate to English the DOCUMENT from the {domain} domain. Provide only the translation. DOCUMENT:\n ['<title> {title} <title\> <body> {document} <body\>']".

[9] https://huggingface.co/datasets/BeIR/
[10] https://github.com/D1Doris/CQADupStack