

BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models

Nandan Thakur, Nils Reimers, Andreas Rücklé*, Abhishek Srivastava, Iryna Gurevych
 Ubiquitous Knowledge Processing Lab (UKP-TUDA)
 Department of Computer Science, Technische Universität Darmstadt
www.ukp.tu-darmstadt.de

Abstract

Existing neural information retrieval (IR) models have often been studied in homogeneous and narrow settings, which has considerably limited insights into their out-of-distribution (OOD) generalization capabilities. To address this, and to facilitate researchers to broadly evaluate the effectiveness of their models, we introduce **Benchmarking-IR (BEIR)**, a robust and heterogeneous evaluation benchmark for information retrieval. We leverage a careful selection of 18 publicly available datasets from diverse text retrieval tasks and domains and evaluate 10 state-of-the-art retrieval systems including lexical, sparse, dense, late-interaction and re-ranking architectures on the BEIR benchmark. **Our results show BM25 is a robust baseline and re-ranking and late-interaction based models on average achieve the best zero-shot performances, however, at high computational costs. In contrast, dense and sparse-retrieval models are computationally more efficient but often underperform other approaches, highlighting the considerable room for improvement in their generalization capabilities.** We hope this framework allows us to better evaluate and understand existing retrieval systems, and contributes to accelerating progress towards better robust and generalizable systems in the future. BEIR is publicly available at <https://github.com/UKPLab/beir>.

1 Introduction

Major natural language processing (NLP) problems rely on a practical and efficient retrieval component as a first step to find relevant information. Challenging problems include open-domain question-answering [8], claim-verification [60], duplicate question detection [78], and many more. Traditionally, retrieval has been dominated by lexical approaches like TF-IDF or BM25 [55]. However, these approaches suffer from lexical gap [5] and are able to only retrieve documents containing keywords present within the query. Further, lexical approaches treat queries and documents as bag-of-words by not taking word ordering into consideration.

Recently, deep learning and in particular pre-trained Transformer models like BERT [12] have become popular in information retrieval [37]. These neural retrieval systems can be used in many fundamentally different ways to improve retrieval performance. We provide an brief overview of the systems in Section 2.1. Many prior work train neural retrieval systems on large datasets like Natural Questions (NQ) [34] (133k training examples) or MS MARCO [45] (533k training examples), which both focus on passage retrieval given a question or short keyword-based query. In most prior work, approaches are afterward evaluated on the same dataset, where significant performance gains over lexical approaches like BM25 are demonstrated [15, 31, 46].

However, creating a large training corpus is often time-consuming and expensive and hence many retrieval systems are applied in a **zero-shot setup**, with no available training data to train the system.

*Contributions made prior to joining Amazon.

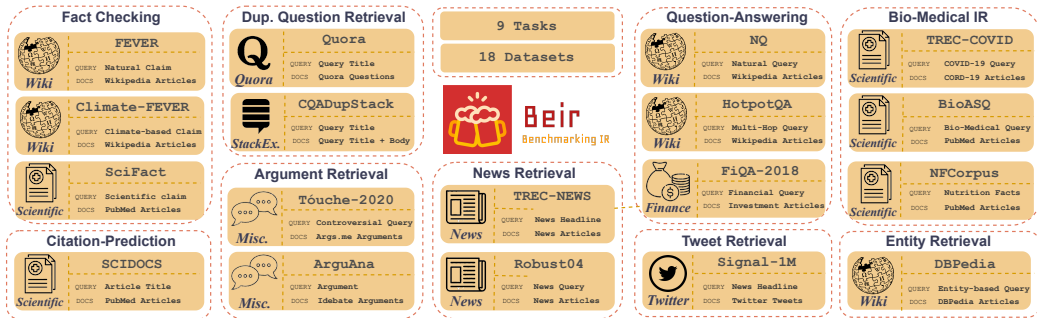


Figure 1: An overview of the diverse tasks and datasets in BEIR benchmark.

So far, it is unclear how well existing trained neural models will perform for other text domains or textual retrieval tasks. Even more important, it is unclear how well different approaches, like sparse embeddings vs. dense embeddings, generalize to out-of-distribution data.

In this work, we present a novel robust and heterogeneous benchmark called **BEIR (Benchmarking IR)**, comprising of 18 retrieval datasets for comparison and evaluation of model generalization. Prior retrieval benchmarks [19, 50] have issues of a comparatively narrow evaluation focusing either only on a single task, like question-answering, or on a certain domain. In BEIR, we focus on **Diversity**, we include nine different retrieval tasks: Fact checking, citation prediction, duplicate question retrieval, argument retrieval, news retrieval, question answering, tweet retrieval, bio-medical IR, and entity retrieval. Further, we include datasets from diverse text domains, datasets that cover broad topics (like Wikipedia) and specialized topics (like COVID-19 publications), different text types (news articles vs. Tweets), datasets of various sizes (3.6k - 15M documents), and datasets with different query lengths (average query length between 3 and 192 words) and document lengths (average document length between 11 and 635 words).

We use BEIR to evaluate **ten diverse retrieval methods** from five broad architectures: lexical, sparse, dense, late interaction, and re-ranking. From our analysis, we find that no single approach consistently outperforms other approaches on all datasets. Further, we notice that the in-domain performance of a model does not correlate well with its generalization capabilities: models fine-tuned with identical training data might generalize differently. In terms of efficiency, we find a trade-off between the performances and the computational cost: computationally expensive models, like re-ranking models and late interaction model perform the best. More efficient approaches e.g. based on dense or sparse embeddings can substantially underperform traditional lexical models like BM25. Overall, BM25 remains a strong baseline for zero-shot text retrieval.

Finally, we notice that there can be a strong lexical bias present in datasets included within the benchmark, likely as lexical models are pre-dominantly used during the annotation or creation of datasets. This can give an unfair disadvantage to non-lexical approaches. We analyze this for the TREC-COVID [65] dataset: We manually annotate the missing relevance judgements for the tested systems and see a significant performance improvement for non-lexical approaches. Hence, future work requires better unbiased datasets that allow a fair comparison for all types of retrieval systems.

With BEIR, we take an important step towards a single and unified benchmark to evaluate the zero-shot capabilities of retrieval systems. It allows to study when and why certain approaches perform well, and hopefully steers innovation to more robust retrieval systems. We release BEIR and an integration of diverse retrieval systems and datasets in a well-documented, easy to use and extensible open-source package. BEIR is model-agnostic, welcomes methods of all kinds, and also allows easy integration of new tasks and datasets. More details are available at <https://github.com/UKPLab/beir>.

2 Related Work and Background

To our knowledge, BEIR is the first broad, zero-shot information retrieval benchmark. Existing works [19, 50] do not evaluate retrieval in a zero-shot setting in depth, they either focus over a single task, small corpora or on a certain domain. This setting hinders for investigation of model generalization across diverse set of domains and task types. MultiReQA [19] consists of eight Question-Answering (QA) datasets and evaluates sentence-level answer retrieval given a question. It only tests a single task and five out of eight datasets are from Wikipedia. Further, MultiReQA evaluates retrieval over rather small corpora: six out of eight tasks have less than 100k candidate sentences, which benefits dense retrieval over lexical as previously shown [54]. KILT [50] consists of five knowledge-intensive

tasks including a total of eleven datasets. The tasks involve retrieval, but it is not the primary task. Further, KILT retrieves documents only from Wikipedia.

2.1 Neural Retrieval

Information retrieval is the process of searching and returning relevant documents for a query from a collection. In our paper, we focus on text retrieval and use *document* as a cover term for text of any length in the given collection and *query* for the user input, which can be of any length as well. Traditionally, lexical approaches like TF-IDF and BM25 [55] have dominated textual information retrieval. Recently, there is a strong interest in using neural networks to improve or replace these lexical approaches. In this section, we highlight a few neural-based approaches and we refer the reader to Lin et al. [37] for a recent survey in neural retrieval.

Retriever-based Lexical approaches suffer from the lexical gap [5]. To overcome this, earlier techniques proposed to improve lexical retrieval systems with neural networks. Sparse methods such as docT5query [48] identified document expansion terms using a sequence-to-sequence model that generated possible queries for which the given document would be relevant. DeepCT [11] on the other hand used a BERT [13] model to learn relevant term weights in a document and generated a pseudo-document representation. Both methods still rely on BM25 for the remaining parts. Similarly, SPARTA [79] learned token-level contextualized representations with BERT and converted the document into an efficient inverse index. More recently, dense retrieval approaches were proposed. They are capable of capturing semantic matches and try to overcome the (potential) lexical gap. Dense retrievers map queries and documents in a shared, dense vector space [18]. This allowed the document representation to be pre-computed and indexed. A bi-encoder neural architecture based on pre-trained Transformers has shown strong performance for various open-domain question-answering tasks [19, 31, 35, 43]. This dense approach was recently extended by hybrid lexical-dense approaches which aims to combine the strengths of both approaches [17, 57, 42]. Another parallel line of work proposed an unsupervised domain-adaption approach [35, 43] for training dense retrievers by generating synthetic queries on a target domain. Lastly, ColBERT [32] (Contextualized late interaction over BERT) computes multiple contextualized embeddings on a token level for queries and documents and uses an maximum-similarity function for retrieving relevant documents.

Re-ranking-based Neural re-ranking approaches use the output of a first-stage retrieval system, often BM25, and re-ranks the documents to create a better comparison of the retrieved documents. Significant improvement in performance was achieved with the cross-attention mechanism of BERT [46]. However, at a disadvantage of a high computational overhead [53].

3 The BEIR Benchmark

BEIR aims to provide a one-stop zero-shot evaluation benchmark for all diverse retrieval tasks. To construct a comprehensive evaluation benchmark, the selection methodology is crucial to collect tasks and datasets with desired properties. For BEIR, the methodology is motivated by the following three factors: (i) **Diverse tasks**: Information retrieval is a versatile task and the lengths of queries and indexed documents can differ between tasks. Sometimes, queries are short, like a keyword, while in other cases, they can be long like a news article. Similarly, indexed documents can sometimes be long, and for other tasks, short like a tweet. (ii) **Diverse domains**: Retrieval systems should be evaluated in various types of domains. From broad ones like News or Wikipedia, to highly specialized ones such as scientific publications in one particular field. Hence, we include domains which provide a representation of real-world problems and are diverse ranging from generic to specialized. (iii) **Task difficulties**: Our benchmark is challenging and the *difficulty* of a task included has to be sufficient. If a task is easily solved by any algorithm, it will not be useful to compare various models used for evaluation. We evaluated several tasks based on existing literature and selected popular tasks which we believe are recently developed, challenging and are not yet fully solved with existing approaches. (iv) **Diverse annotation strategies**: Creating retrieval datasets are inherently complex and are subject to *annotation biases* (see Section 6 for details), which hinders a fair comparison of approaches. To reduce the impact of such biases, we selected datasets which have been created in many different ways: Some where annotated by crowd-workers, others by experts, and others are based on the feedback from large online communities.

In total, we include 18 English zero-shot evaluation datasets from 9 heterogeneous retrieval tasks. As the majority of the evaluated approaches are trained on the MS MARCO [45] dataset, we also report performances on this dataset, but don't include the outcome in our zero-shot comparison. We would like to refer the reader to Appendix D where we motivate each one of the 9 retrieval tasks and 18

Split (→)					Train		Dev	Test			Avg. Word Lengths	
Task (↓)	Domain (↓)	Dataset (↓)	Title	Relevancy	#Pairs	#Query	#Query	#Corpus	Avg. D / Q		Query	Document
Passage-Retrieval	Misc.	MS MARCO [45]	✗	Binary	532,761	—	6,980	8,841,823	1.1		5.96	55.98
Bio-Medical Information Retrieval (IR)	Bio-Medical	TREC-COVID [65]	✓	3-level	—	—	50	171,332	493.5		10.60	160.77
	Bio-Medical	NFCorpus [7]	✓	3-level	110,575	324	323	3,633	38.2		3.30	232.26
	Bio-Medical	BioASQ [61]	✓	Binary	32,916	—	500	14,914,602	4.7		8.05	202.61
Question Answering (QA)	Wikipedia	NQ [34]	✓	Binary	132,803	—	3,452	2,681,468	1.2		9.16	78.88
	Wikipedia	HotpotQA [76]	✓	Binary	170,000	5,447	7,405	5,233,329	2.0		17.61	46.30
	Finance	FiQA-2018 [44]	✗	Binary	14,166	500	648	57,638	2.6		10.77	132.32
Tweet-Retrieval	Twitter	Signal-1M (RT) [59]	✗	3-level	—	—	97	2,866,316	19.6		9.30	13.93
News Retrieval	News	TREC-NEWS [58]	✓	5-level	—	—	57	594,977	19.6		11.14	634.79
	News	Robust04 [64]	✗	3-level	—	—	249	528,155	69.9		15.27	466.40
Argument Retrieval	Misc.	ArguAna [67]	✓	Binary	—	—	1,406	8,674	1.0		192.98	166.80
	Misc.	Touché-2020 [6]	✓	3-level	—	—	49	382,545	19.0		6.55	292.37
Duplicate-Question Retrieval	StackEx.	CQADupStack [25]	✓	Binary	—	—	13,145	457,199	1.4		8.59	129.09
	Quora	Quora	✗	Binary	—	5,000	10,000	522,931	1.6		9.53	11.44
Entity-Retrieval	Wikipedia	DBPedia [21]	✓	3-level	—	67	400	4,635,922	38.2		5.39	49.68
Citation-Prediction	Scientific	SCIDOCs [9]	✓	Binary	—	—	1,000	25,657	4.9		9.38	176.19
Fact Checking	Wikipedia	FEVER [60]	✓	Binary	140,085	6,666	6,666	5,416,568	1.2		8.13	84.76
	Wikipedia	Climate-FEVER [14]	✓	Binary	—	—	1,535	5,416,593	3.0		20.13	84.76
	Scientific	SciFact [68]	✓	Binary	920	—	300	5,183	1.1		12.37	213.63

Table 1: Statistics of datasets in BEIR benchmark. Few datasets contain documents without titles. Relevancy indicates the query-document relation: binary (relevant, non-relevant) or graded into sub-levels. Avg. D/Q indicates the average relevant documents per query.

datasets in depth. Examples for each dataset are listed in Table 8. We additionally provide dataset licenses in Appendix E, and links to the datasets in Table 5.

Table 1 summarizes the statistics of the datasets provided in BEIR. A majority of datasets contain binary relevancy judgements, i.e. relevant or non-relevant, and a few contain fine-grained relevancy judgements. Some datasets contain few relevant documents for a query (< 2), while other datasets like TREC-COVID [65] can contain up to even 500 relevant documents for a query. Only 8 out of 19 datasets (including MS MARCO) have training data denoting the practical importance for zero-shot retrieval benchmarking. All datasets except ArguAna [67] have short queries (either a single sentence or 2-3 keywords). Figure 1 shows an overview of the tasks and datasets in the BEIR benchmark.

Information Retrieval (IR) is ubiquitous, there are lots of datasets available within each task and further even more tasks with retrieval. However, it is not feasible to include all datasets within the benchmark for evaluation. We tried to cover a balanced mixture of a wide range of tasks and datasets and paid importance not to overweight a specific task like question-answering. Future datasets can easily be integrated in BEIR, and existing models can be evaluated on any new dataset quickly. The BEIR website will host an actively maintained leaderboard² with all datasets and models.

3.1 Dataset and Diversity Analysis

The datasets present in BEIR are selected from diverse domains ranging from Wikipedia, scientific publications, Twitter, news, to online user communities, and many more. To measure the diversity in domains, we compute the domain overlap between the pairwise datasets using a pairwise weighted Jaccard similarity [26] score on unigram word overlap between all dataset pairs. For more details on the theoretical formulation of the similarity score, please refer to Appendix F. Figure 2 shows a heatmap denoting the pairwise weighted jaccard scores and the clustered force-directed placement diagram. Nodes (or datasets) close in this graph have a high word overlap, while nodes far away in the graph have a low overlap. From Figure 2, we observe a rather low weighted Jaccard word overlap across different domains, indicating that BEIR is a challenging benchmark where approaches must generalize well to diverse out-of-distribution domains.

3.2 BEIR Software and Framework

The BEIR software³ provides an is an easy to use Python framework (`pip install beir`) for model evaluation. It contains extensive wrappers to replicate experiments and evaluate models from well-known repositories including Sentence-Transformers [53], Transformers [72], Anserini [74], DPR [31], Elasticsearch, ColBERT [32], and Universal Sentence Encoder [75]. This makes the software useful for both academia and industry. The software also provides you with all IR-based metrics from Precision, Recall, MAP (Mean Average Precision), MRR (Mean Reciprocal Rate) to nDCG

²BEIR Leaderboard: <https://tinyurl.com/beir-leaderboard>

³BEIR Code & documentation: <https://github.com/UKPLab/beir>

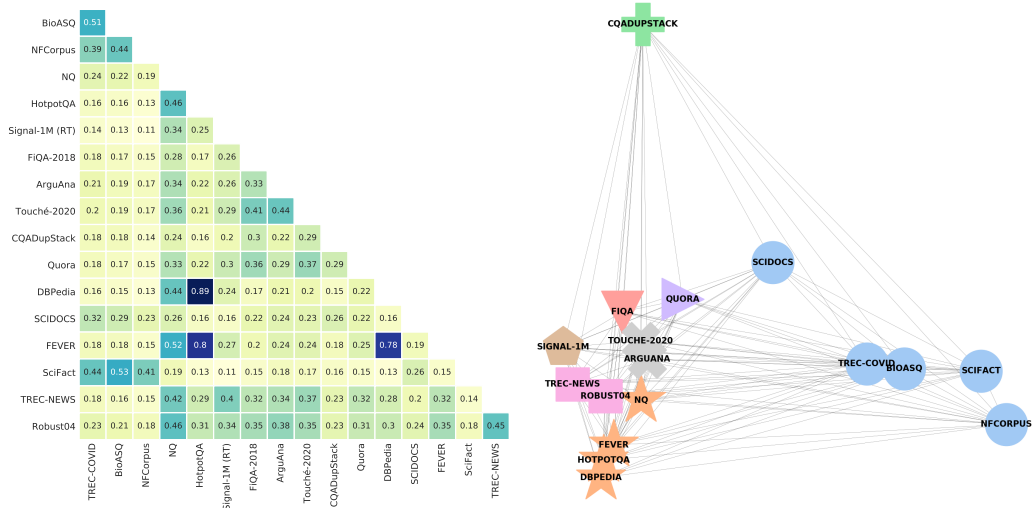


Figure 2: Domain overlap across each pairwise dataset in the BEIR benchmark. Heatmap (left) shows the pairwise weighted jaccard similarity scores between BEIR datasets. 2D representation (right) using a force-directed placement algorithm with NetworkX [20]. We color and mark datasets differently for different domains.

(Normalised Cumulative Discount Gain) for any top-k hits. One can use the BEIR benchmark for evaluating existing models on new retrieval datasets and for evaluating new models on the included datasets.

Datasets are often scattered online and are provided in various file-formats, making the evaluation of models on various datasets difficult. BEIR introduces a standard format (corpus, queries and qrels) and converts existing datasets in this easy universal data format, allowing to evaluate faster on an increasing number of datasets.

3.3 Evaluation Metric

Depending upon the nature and requirements of real-world applications, retrieval tasks can be either be precision or recall focused. To obtain comparable results across models and datasets in BEIR, we argue that it is important to leverage a single evaluation metric that can be computed comparably across all tasks. Decision support metrics such as Precision and Recall which are both rank unaware are not suitable. Binary rank-aware metrics such as MRR (Mean Reciprocal Rate) and MAP (Mean Average Precision) fail to evaluate tasks with graded relevance judgements. We find that **Normalised Cumulative Discount Gain** (nDCG@k) provides a good balance suitable for both tasks involving binary and graded relevance judgements. We refer the reader to Wang et al. [71] for understanding the theoretical advantages of the metric. For our experiments, we utilize the Python interface of the official TREC evaluation tool [63] and compute nDCG@10 for all datasets.

4 Experimental Setup

We use BEIR to compare diverse, recent, state-of-the-art retrieval architectures with a focus on transformer-based neural approaches. We evaluate on publicly available pre-trained checkpoints, which we provide in Table 6. Due to the length limitations of transformer-based networks, we use only the first 512 word pieces within all documents in our experiments across all neural architectures.

We group the models based on their architecture: (i) lexical, (ii) sparse, (iii) dense, (iv) late-interaction, and (v) re-ranking. Besides the included models, the BEIR benchmark is model agnostic and in future different model configurations can be easily incorporated within the benchmark.

(i) Lexical Retrieval: (a) **BM25** [55] is a commonly-used bag-of-words retrieval function based on token-matching between two high-dimensional sparse vectors with TF-IDF token weights. We use Anserini [36] with the default Lucene parameters ($k=0.9$ and $b=0.4$). We index the title (if available) and passage as separate fields for documents. In our leaderboard, we also tested Elasticsearch BM25 and Anserini + RM3 expansion, but found Anserini BM25 to perform the best.

(ii) **Sparse Retrieval:** (a) **DeepCT** [11] uses a bert-base-uncased model trained on MS MARCO to learn the term weight frequencies (tf). It generates a pseudo-document with keywords multiplied with the learnt term-frequencies. We use the original setup of Dai and Callan [11] in combination with BM25 with default Anserini parameters which we empirically found to perform better over the tuned MS MARCO parameters. (b) **SPARTA** [79] computes similarity scores between the non-contextualized query embeddings from BERT with the contextualized document embeddings. These scores can be pre-computed for a given document, which results in a 30k dimensional sparse vector. As the original implementation is not publicly available, we re-implemented the approach. We fine-tune a DistilBERT [56] model on the MS MARCO dataset and use sparse-vectors with 2,000 non-zero entries. (c) **DocT5query** [47] is a popular document expansion technique using a T5 (base) [52] model trained on MS MARCO to generate synthetic queries and append them to the original document for lexical search. We replicate the setup of Nogueira and Lin [47] and generate 40 queries for each document and use BM25 with default Anserini parameters.

(iii) **Dense Retrieval:** (a) **DPR** [31] is a two-tower bi-encoder trained with a single BM25 hard negative and in-batch negatives. We found the open-sourced Multi model to perform better over the single NQ model in our setting. The Multi-DPR model is a bert-base-uncased model trained on four QA datasets (including titles): NQ [34], TriviaQA [30], WebQuestions [4] and CuratedTREC [3]. (b) **ANCE** [73] is a bi-encoder constructing hard negatives from an Approximate Nearest Neighbor (ANN) index of the corpus, which in parallel updates to select hard negative training instances during fine-tuning of the model. We use the publicly available RoBERTa [41] model trained on MS MARCO [45] for 600K steps for our experiments. (c) **TAS-B** [23] is a bi-encoder trained with Balanced Topic Aware Sampling using dual supervision from a cross-encoder and a ColBERT model. The model was trained with a combination of both a pairwise Margin-MSE [24] loss and an in-batch negative loss function. (d) **GenQ**: is an unsupervised domain-adaption approach for dense retrieval models by training on synthetically generated data. First, we fine-tune a T5 (base) [52] model on MS MARCO for 2 epochs. Then, for a target dataset we generate 5 queries for each document using a combination of top-k and nucleus-sampling (top-k: 25; top-p: 0.95). Due to resource constraints, we cap the maximum number of target documents in each dataset to 100K. For retrieval, we continue to fine-tune the TAS-B model using in-batch negatives on the synthetic queries and document pair data. Note, GenQ creates an independent model for each task.

(iv) **Late-Interaction:** (a) **ColBERT** [32] encodes and represents the query and passage into a bag of multiple contextualized token embeddings. The late-interactions are aggregated with sum of the max-pooling query term and a dot-product across all passage terms. We use the ColBERT model as a dense-retriever (end-to-end retrieval as defined [32]): first top-k candidates are retrieved using ANN with faiss [29] (faiss depth = 100) and ColBERT re-ranks by computing the late aggregated interactions. We train a bert-base-uncased model, with maximum sequence length of 300 on the MS MARCO dataset for 300K steps.

(v) **Re-ranking model:** (a) **BM25 + CE** [70] reranks the top-100 retrieved hits from a first-stage BM25 (Anserini) model. We evaluated 14 different cross-attentional re-ranking models that are publicly available on the HuggingFace model hub and found that a 6-layer, 384-h MiniLM [70] cross-encoder model offers the best performance on MS MARCO. The model was trained on MS MARCO using a knowledge distillation setup with an ensemble of three teacher models: BERT-base, BERT-large, and ALBERT-large models following the setup in Hofstätter et al. [24].

5 Results and Analysis

In this section, we evaluate and analyze how retrieval models perform on the BEIR benchmark. Table 2 reports the results of all evaluated systems on the selected benchmark datasets. **As a baseline, we compare our retrieval systems against BM25.** Figure 3 shows, on how many datasets a respective model is able to perform better or worse than BM25.

1. In-domain performance is not a good indicator for out-of-domain generalization. We observe BM25 heavily underperforms neural approaches by 7-18 points on in-domain MS MARCO. However, BEIR reveals it to be a strong baseline for generalization and generally outperforming many other, more complex approaches. This stresses the point, that retrieval methods must be evaluated on a broad range of datasets.

2. Term-weighting fails, document expansion captures out-of-domain keyword vocabulary. DeepCT and SPARTA both use a transformer network to learn term weighting. While both methods

Model (→)	Lexical	Sparse			Dense				Late-Interaction	Re-ranking
Dataset (↓)	BM25	DeepCT	SPARTA	docT5query	DPR	ANCE	TAS-B	GenQ	ColBERT	BM25+CE
MS MARCO	0.228	0.296 [‡]	0.351 [‡]	0.338 [‡]	0.177	0.388 [‡]	0.408 [‡]	0.408 [‡]	<u>0.401[‡]</u>	0.413[‡]
TREC-COVID	0.656	0.406	0.538	<u>0.713</u>	0.332	0.654	0.481	0.619	0.677	0.757
BioASQ	0.465	0.407	0.351	0.431	0.127	0.306	0.383	0.398	<u>0.474</u>	0.523
NFCorpus	0.325	0.283	0.301	<u>0.328</u>	0.189	0.237	0.319	0.319	0.305	0.350
NQ	0.329	0.188	0.398	0.399	0.474 [‡]	0.446	0.463	0.358	<u>0.524</u>	0.533
HotpotQA	<u>0.603</u>	0.503	0.492	0.580	0.391	0.456	0.584	0.534	0.593	0.707
FiQA-2018	0.236	0.191	0.198	0.291	0.112	0.295	0.300	0.308	<u>0.317</u>	0.347
Signal-1M (RT)	<u>0.330</u>	0.269	0.252	0.307	0.155	0.249	0.289	0.281	0.274	0.338
TREC-NEWS	0.398	0.220	0.258	<u>0.420</u>	0.161	0.382	0.377	0.396	0.393	0.431
Robust04	0.408	0.287	0.276	<u>0.437</u>	0.252	0.392	0.427	0.362	0.391	0.475
ArguAna	0.315	0.309	0.279	0.349	0.175	0.415	<u>0.429</u>	0.493	0.233	0.311
Touché-2020	0.367	0.156	0.175	<u>0.347</u>	0.131	0.240	0.162	0.182	0.202	0.271
CQADupStack	0.299	0.268	0.257	0.325	0.153	0.296	0.314	0.347	<u>0.350</u>	0.370
Quora	0.789	0.691	0.630	0.802	0.248	<u>0.852</u>	0.835	0.830	0.854	0.825
DBPedia	0.313	0.177	0.314	0.331	0.263	0.281	0.384	0.328	<u>0.392</u>	0.409
SCIDOCS	0.158	0.124	0.126	<u>0.162</u>	0.077	0.122	0.149	0.143	0.145	0.166
FEVER	0.753	0.353	0.596	0.714	0.562	0.669	0.700	0.669	<u>0.771</u>	0.819
Climate-FEVER	0.213	0.066	0.082	0.201	0.148	0.198	<u>0.228</u>	0.175	0.184	0.253
SciFact	0.665	0.630	0.582	<u>0.675</u>	0.318	0.507	0.643	0.644	0.671	0.688
Avg. Performance vs. BM25		- 27.9%	- 20.3%	+ 1.6%	- 47.7%	- 7.4%	- 2.8%	- 3.6%	+ 2.5%	+ 11%

Table 2: In-domain and zero-shot performances on BEIR benchmark. All scores denote **nDCG@10**. The best score on a given dataset is marked in **bold**, and the second best is underlined. Corresponding Recall@100 performances can be found in Table 9. ‡ indicates the in-domain performances.

perform well in-domain on MS MARCO, they completely fail to generalize well by under performing BM25 on nearly all datasets. In contrast, document expansion based docT5query is able to add new relevant keywords to a document and performs strong on the BEIR datasets. It outperforms BM25 on 11/18 datasets while providing a competitive performance on the remaining datasets.

3. Dense retrieval models with issues for out-of-distribution data. Dense retrieval models (esp. ANCE and TAS-B), that map queries and documents independently to vector spaces, perform strongly on certain datasets, while on many other datasets perform significantly worse than BM25. For example, dense retrievers are observed to underperform on datasets with a large domain shift compared from what they have been trained on, like in BioASQ, or task-shifts like in Touché-2020. DPR, the only non-MSMARCO trained dataset overall performs the worst in generalization on the benchmark.

4. Re-ranking and Late-Interaction models generalize well to out-of-distribution data. The cross-attentional re-ranking model (BM25+CE) performs the best and is able to outperform BM25 on almost all (16/18) datasets. It only fails on ArguAna and Touché-2020, two retrieval tasks that are extremely different to the MS MARCO training dataset. The late-interaction model ColBERT computes token embeddings independently for the query and document, and scores (query, document)-pairs by a cross-attentional like MaxSim operation. It performs a bit weaker than the cross-attentional re-ranking model, but is still able to outperform BM25 on 9/18 datasets. It appears that cross-attention and cross-attentional like operations are important for a good out-of-distribution generalization.

5. Strong training losses for dense retrieval leads to better out-of-distribution performances. TAS-B provides the best zero-shot generalization performance among its dense counterparts. It outperforms ANCE on 14/18 and DPR on 17/18 datasets respectively. We speculate that the reason lies in a strong training setup in combination of both in-domain batch negatives and Margin-MSE losses for the TAS-B model. This training loss function (with strong ensemble teachers in a Knowledge Distillation setup) shows strong generalization performances.

6. TAS-B model prefers to retrieve documents with shorter lengths. TAS-B underperforms ANCE on two datasets: TREC-COVID by 17.3 points and Touché-2020 by 7.8 points. We observed that these models retrieve documents with vastly different lengths as shown in Figure 4. On TREC-COVID, TAS-B retrieves documents with a median length of mere 10 words versus ANCE with 160 words. Similarly on Touché-2020, 14 words vs. 89 words with TAS-B and ANCE respectively. As discussed in Appendix H, this preference for shorter or longer documents is due to the used loss function.

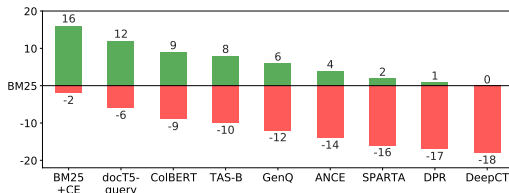


Figure 3: Comparison of zero-shot neural retrieval performances with BM25. Re-ranking based models, i.e., BM25+CE and sparse model: docT5query outperform BM25 on more than half the BEIR evaluation datasets.

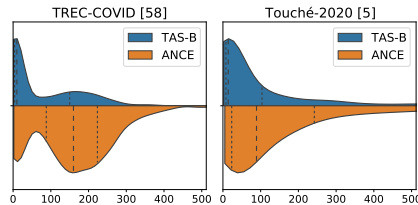


Figure 4: Distribution plots [22] for top-10 retrieved document lengths (in words) using TAS-B (blue, top) or ANCE (orange, bottom). TAS-B has a preference towards shorter documents in BEIR.

7. Does domain adaptation help improve generalization of dense-retrievers? We evaluated GenQ, which further fine-tunes the TAS-B model on synthetic query data. It outperforms the TAS-B model on specialized domains like scientific publications, finance or StackExchange. On broader and more generic domains, like Wikipedia, it performs weaker than the original TAS-B model.

5.1 Efficiency: Retrieval Latency and Index Sizes

Models need to potentially compare a single query against millions of documents at inference, hence, a high computational speed for retrieving results in real-time is desired. Besides speed, index sizes are vital and are often stored entirely in memory. We randomly sample 1 million documents from DBPedia [21] and evaluate latency. For dense models, we use exact search, while for ColBERT we follow the original setup [32] and use approximate nearest neighbor search. Performances on CPU were measured with an 8 core Intel Xeon Platinum 8168 CPU @ 2.70GHz and on GPU using a single Nvidia Tesla V100, CUDA 11.0.

Tradeoff between performance and retrieval latency

The best out-of-distribution generalization performances by re-ranking top-100 BM25 documents and with late-interaction models come at the cost of high latency (> 350 ms), being slowest at inference. In contrast, dense retrievers are 20-30x faster (< 20ms) compared to the re-ranking models and follow a low-latency pattern. On CPU, the sparse models dominate in terms of speed (20-25ms).

Tradeoff between performance and index sizes

Lexical, re-ranking and dense methods have the smallest index sizes (< 3GB) to store 1M documents from DBPedia. SPARTA requires the second largest index to store a 30k dim sparse vector while ColBERT requires the largest index as it stores multiple 128 dim dense vectors for a single document. Index sizes are especially relevant when document sizes scale higher: ColBERT requires ~900GB to store the BioASQ (~15M documents) index, whereas BM25 only requires 18GB.

DBPedia [21] (1 Million)			Retrieval Latency		Index
Rank	Model	Dim.	GPU	CPU	Size
(1)	BM25+CE	–	450ms	6100ms	0.4GB
(2)	ColBERT	128	350ms	–	20GB
(3)	docT5query	–	–	30ms	0.4GB
(4)	BM25	–	–	20ms	0.4GB
(5)	TAS-B	768	14ms	125ms	3GB
(6)	GenQ	768	14ms	125ms	3GB
(7)	ANCE	768	20ms	275ms	3GB
(8)	SPARTA	2000	–	20ms	12GB
(9)	DeepCT	–	–	25ms	0.4GB
(10)	DPR	768	19ms	230ms	3GB

Table 3: Estimated average retrieval latency and index sizes for a single query in DBPedia [21]. Ranked from best to worst on zero-shot BEIR. Lower the latency or memory is desired.

6 Impact of Annotation Selection Bias

Creating a perfectly unbiased evaluation dataset for retrieval is inherently complex and is subject to multiple biases induced by the: (i) annotation guidelines, (ii) annotation setup, and by the (iii) human annotators. Further, it is impossible to manually annotate the relevance for all (query, document)-pairs. Instead, existing retrieval methods are used to get a pool of candidate documents which are then marked for their relevance. All other unseen documents are assumed to be irrelevant. This is a source for *selection bias* [39]: A new retrieval system might retrieve vastly different results than the system used for the annotation. These hits are automatically assumed to be irrelevant.

Many BEIR datasets are found to be subject to a lexical bias, i.e. a lexical based retrieval system like TF-IDF or BM25 has been used to retrieve the candidates for annotation. For example, in BioASQ, candidates have been retrieved for annotation via term-matching with boosting tags [61]. Creation of Signal-1M (RT) involved retrieving tweets for a query with 7 out of these 8 techniques relying upon

Model (→)	BM25	DeepCT	SPARTA	docT5query	DPR	ANCE	TAS-B	ColBERT	BM25+CE
Hole@10 (in %)	6.4%	19.4%	12.4%	2.8%	30.6%	14.4%	31.8%	12.4%	1.6%
nDCG@10 performances before and after manual annotation on TREC-COVID [65]									
Original (w/ holes)	0.656	0.406	0.538	<u>0.713</u>	0.332	0.654	0.481	0.677	0.757
Annotated (w/o holes)	0.668	0.472	0.624	0.714	0.445	<u>0.735</u>	0.555	<u>0.735</u>	0.760

Table 4: Hole@10 analysis on TREC-COVID. Annotated scores show improvement in performances after removing holes@10 (documents in top-10 hits unseen by annotators) across each model.

lexical term-matching signals [59]. Such a lexical bias disfavours approaches that don’t rely on lexical matching, like dense retrieval methods, as retrieved hits without lexical overlap are automatically assumed to be irrelevant, even though the hits might be relevant for a query.

In order to study the impact of this particular type of bias, we conducted a study on the recent TREC-COVID dataset. TREC-COVID used a pooling method [38, 40] to reduce the impact of the aforementioned bias: The annotation set was constructed by using the search results from the various systems participating in the challenge. Table 4 shows the Hole@10 rate [73] for the tested systems, i.e., how many top-10 hits is each system retrieving that have not been seen by annotators.

The results reveal large differences between approaches: Lexical approaches like BM25 and docT5query have a rather low Hole@10 value of 6.4% and 2.8%, indicating that the annotation pool contained the top-hits from lexical retrieval systems. In contrast, dense retrieval systems like ANCE and TAS-B have a much higher Hole@10 of 14.4% and 31.8%, indicating that a large fraction of hits found by these systems have not been judged by annotators. Next, we manually added for all systems, the missing annotation (or holes) following the original annotation guidelines. During annotation, we were unaware of the system who retrieved the missing annotation to avoid a preference bias. In total, we annotated 980 query-document pairs in TREC-COVID. We then re-computed nDCG@10 for all systems with this additional annotations.

As shown in Table 4, we observe that lexical approaches improves only slightly, e.g. for docT5query just from 0.713 to 0.714 after adding the missing relevance judgements. In contrast, for the dense retrieval system ANCE, the performance improves from 0.654 (slightly below BM25) to 0.735, which is 6.7 points above the BM25 performance. Similar improvements are noticed in ColBERT (5.8 points). Even though many systems contributed to the TREC-COVID annotation pool, the annotation pool is still biased towards lexical approaches.

7 Conclusions and Future Work

In this work, we presented BEIR: a heterogeneous benchmark for information retrieval. We provided a broader selection of target tasks ranging from narrow expert domains to open domain datasets. We included nine different retrieval tasks spanning 18 diverse datasets.

By open-sourcing BEIR, with a standardized data format and easy-to-adapt code examples for many different retrieval strategies, we take an important steps towards a unified benchmark to evaluate the zero-shot capabilities of retrieval systems. It hopefully steers innovation towards more robust retrieval systems and to new insights which retrieval architectures perform well across tasks and domains.

We studied the effectiveness of ten different retrieval models and demonstrate, that in-domain performance cannot predict how well an approach will generalize in a zero-shot setup. Many approaches that outperform BM25 on an in-domain evaluation, perform poorly on the BEIR datasets. Cross-attentional re-ranking, late-interaction ColBERT, and the document expansion technique docT5query performed overall well across the evaluated tasks.

Our study on annotation selection bias highlights the challenge of evaluating new models on existing datasets: Even though TREC-COVID is based on the predictions from many systems, contributed by a diverse set of teams, we found largely different *Hole@10* rates for the tested systems, negatively affecting non-lexical approaches. Better datasets, that use diverse pooling strategies, are needed for a fair evaluation of retrieval approaches. By integrate a large number of diverse retrieval systems into BEIR, creating such diverse pools becomes significantly simplified.

References

- [1] Amin Ahmad, Noah Constant, Yinfei Yang, and Daniel Cer. 2019. **ReQA: An Evaluation for End-to-End Answer Retrieval Models**. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 137–146, Hong Kong, China. Association for Computational Linguistics. 18
- [2] Akari Asai, Jungo Kasai, Jonathan Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021. **XOR QA: Cross-lingual Open-Retrieval Question Answering**. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 547–564, Online. Association for Computational Linguistics. 17
- [3] Petr Baudiš and Jan Šedivý. 2015. **Modeling of the question answering task in the yodaqa system**. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 222–228. Springer. 6
- [4] Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. **Semantic Parsing on Freebase from Question-Answer Pairs**. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, Seattle, Washington, USA. Association for Computational Linguistics. 6
- [5] Adam Berger, Rich Caruana, David Cohn, Dayne Freitag, and Vibhu Mittal. 2000. **Bridging the lexical chasm: statistical approaches to answer-finding**. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 192–199. 1, 3
- [6] Alexander Bondarenko, Maik Fröbe, Meriem Beloucif, Lukas Gienapp, Yamen Ajjour, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. 2020. **Overview of Touché 2020: Argument Retrieval**. In *Working Notes Papers of the CLEF 2020 Evaluation Labs*, volume 2696 of *CEUR Workshop Proceedings*. 4, 19, 22
- [7] Vera Boteva, Demian Gholipour, Artem Sokolov, and Stefan Riezler. 2016. **A full-text learning to rank dataset for medical information retrieval**. In *Proceedings of the 38th European Conference on Information Retrieval (ECIR 2016)*, pages 716–722. 4, 18
- [8] 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. 1, 18
- [9] 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. 4, 19
- [10] Davind Corney, Dyaa Albakour, Miguel Martinez, and Samir Moussa. 2016. **What do a Million News Articles Look like?** In *Proceedings of the First International Workshop on Recent Trends in News Information Retrieval co-located with 38th European Conference on Information Retrieval (ECIR 2016)*, pages 42–47. 18
- [11] Zhuyun Dai and Jamie Callan. 2020. **Context-Aware Term Weighting For First Stage Passage Retrieval**. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’20*, page 1533–1536, New York, NY, USA. Association for Computing Machinery. 3, 6
- [12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. **BERT: Pre-training of deep bidirectional transformers for language understanding**. *arXiv preprint arXiv:1810.04805*. 1

- [13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics. 3
- [14] Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. **CLIMATE-FEVER: A Dataset for Verification of Real-World Climate Claims**. 4, 20
- [15] Yingqi Qu Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2020. **RocketQA: An Optimized Training Approach to Dense Passage Retrieval for Open-Domain Question Answering**. 1, 17
- [16] Anlei Dong, Ruiqiang Zhang, Pranam Kolari, Jing Bai, Fernando Diaz, Yi Chang, Zhaohui Zheng, and Hongyuan Zha. 2010. **Time is of the Essence: Improving Recency Ranking Using Twitter Data**. In *Proceedings of the 19th International Conference on World Wide Web, WWW '10*, page 331–340, New York, NY, USA. Association for Computing Machinery. 17
- [17] Luyu Gao, Zhuyun Dai, Tongfei Chen, Zhen Fan, Benjamin Van Durme, and Jamie Callan. 2020. **Complementing Lexical Retrieval with Semantic Residual Embedding**. 3, 17
- [18] Daniel Gillick, Alessandro Presta, and Gaurav Singh Tomar. 2018. **End-to-End Retrieval in Continuous Space**. 3
- [19] Mandy Guo, Yinfei Yang, Daniel Cer, Qinlan Shen, and Noah Constant. 2020. **MultiReQA: A Cross-Domain Evaluation for Retrieval Question Answering Models**. 2, 3
- [20] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. 2008. Exploring Network Structure, Dynamics, and Function using NetworkX. In *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA. 5
- [21] 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, SIGIR '17*, pages 1265–1268. ACM. 4, 8, 19
- [22] Jerry L. Hintze and Ray D. Nelson. 1998. **Violin Plots: A Box Plot-Density Trace Synergism**. *The American Statistician*, 52(2):181–184. 8, 24
- [23] Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently Teaching an Effective Dense Retriever with Balanced Topic Aware Sampling. In *Proc. of SIGIR*. 6
- [24] Sebastian Hofstätter, Sophia Althammer, Michael Schröder, Mete Sertkan, and Allan Hanbury. 2021. **Improving Efficient Neural Ranking Models with Cross-Architecture Knowledge Distillation**. 6, 21
- [25] 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. 4, 19
- [26] Sergey Ioffe. 2010. **Improved consistent sampling, weighted minhash and l1 sketching**. In *2010 IEEE International Conference on Data Mining*, pages 246–255. IEEE. 4, 20
- [27] Ming Ji, Yizhou Sun, Marina Danilevsky, Jiawei Han, and Jing Gao. 2010. Graph Regularized Transductive Classification on Heterogeneous Information Networks. In *Machine Learning and Knowledge Discovery in Databases*, pages 570–586, Berlin, Heidelberg. Springer Berlin Heidelberg. 19
- [28] Jing Jiang and ChengXiang Zhai. 2007. An empirical study of tokenization strategies for biomedical information retrieval. *Information Retrieval*, 10(4-5):341–363. 18

- [29] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with GPUs. *arXiv preprint arXiv:1702.08734*. 6
- [30] Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. **TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension**. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics. 6
- [31] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. **Dense Passage Retrieval for Open-Domain Question Answering**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics. 1, 3, 4, 6
- [32] Omar Khattab and Matei Zaharia. 2020. **ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT**. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '20*, page 39–48, New York, NY, USA. Association for Computing Machinery. 3, 4, 6, 8
- [33] Jon M. Kleinberg. 1999. **Authoritative Sources in a Hyperlinked Environment**. *J. ACM*, 46(5):604–632. 17
- [34] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: a Benchmark for Question Answering Research. *Transactions of the Association of Computational Linguistics*. 1, 4, 6, 18
- [35] Davis Liang, Peng Xu, Siamak Shakeri, Cicero Nogueira dos Santos, Ramesh Nallapati, Zhiheng Huang, and Bing Xiang. 2020. **Embedding-based Zero-shot Retrieval through Query Generation**. 3
- [36] Jimmy Lin, Matt Crane, Andrew Trotman, Jamie Callan, Ishan Chattopadhyaya, John Foley, Grant Ingersoll, Craig Macdonald, and Sebastiano Vigna. 2016. Toward reproducible baselines: The open-source IR reproducibility challenge. In *European Conference on Information Retrieval*, pages 408–420. Springer. 5
- [37] Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2020. **Pretrained Transformers for Text Ranking: BERT and Beyond**. 1, 3
- [38] Aldo Lipani. 2016. **Fairness in Information Retrieval**. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '16*, page 1171, New York, NY, USA. Association for Computing Machinery. 9
- [39] Aldo Lipani. 2019. *On Biases in Information retrieval models and evaluation*. Ph.D. thesis, Technische Universität Wien. 8
- [40] Aldo Lipani, Mihai Lupu, and Allan Hanbury. 2016. The Curious Incidence of Bias Corrections in the Pool. In *European Conference on Information Retrieval*, pages 267–279. Springer. 9
- [41] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. **RoBERTa: A Robustly Optimized BERT Pretraining Approach**. 6
- [42] Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2021. **Sparse, Dense, and Attentional Representations for Text Retrieval**. 3
- [43] Ji Ma, Ivan Korotkov, Yinfei Yang, Keith Hall, and Ryan McDonald. 2021. **Zero-shot Neural Passage Retrieval via Domain-targeted Synthetic Question Generation**. 3
- [44] 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, WWW '18*, page 1941–1942, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee. 4, 18

- [45] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated MACHine Reading COMprehension Dataset. *choice*, 2640:660. 1, 3, 4, 6, 17
- [46] Rodrigo Nogueira and Kyunghyun Cho. 2020. **Passage Re-ranking with BERT**. *arXiv preprint arXiv:1901.04085*. 1, 3, 17
- [47] Rodrigo Nogueira, Jimmy Lin, and AI Epistemic. 2019. **From doc2query to docTTTTTquery**. *Online preprint*. 6
- [48] Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. **Document Expansion by Query Prediction**. 3
- [49] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. **The PageRank Citation Ranking: Bringing Order to the Web**. Technical Report 1999-66, Stanford InfoLab. Previous number = SIDL-WP-1999-0120. 17
- [50] Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2020. **KILT: a Benchmark for Knowledge Intensive Language Tasks**. 2
- [51] Filip Radlinski, Madhu Kurup, and Thorsten Joachims. 2008. **How Does Clickthrough Data Reflect Retrieval Quality?** In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, page 43–52, New York, NY, USA. Association for Computing Machinery. 17
- [52] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. **Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer**. *Journal of Machine Learning Research*, 21(140):1–67. 6
- [53] Nils Reimers and Iryna Gurevych. 2019. **Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks**. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. 3, 4
- [54] Nils Reimers and Iryna Gurevych. 2020. **The Curse of Dense Low-Dimensional Information Retrieval for Large Index Sizes**. *arXiv preprint arXiv:2012.14210*. 2
- [55] Stephen Robertson and Hugo Zaragoza. 2009. **The Probabilistic Relevance Framework: BM25 and Beyond**. *Foundations and Trends in Information Retrieval*, 3(4):333–389. 1, 3, 5
- [56] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. **DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter**. 6
- [57] Minjoon Seo, Jinhyuk Lee, Tom Kwiatkowski, Ankur Parikh, Ali Farhadi, and Hannaneh Hajishirzi. 2019. **Real-Time Open-Domain Question Answering with Dense-Sparse Phrase Index**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4430–4441, Florence, Italy. Association for Computational Linguistics. 3
- [58] Ian Soboroff, Shudong Huang, and Donna Harman. 2019. **TREC 2019 News Track Overview**. In *TREC*. 4, 18
- [59] Axel Suarez, Dyaa Albakour, David Corney, Miguel Martinez, and Jose Esquivel. 2018. **A Data Collection for Evaluating the Retrieval of Related Tweets to News Articles**. In *40th European Conference on Information Retrieval Research (ECIR 2018)*, Grenoble, France, March, 2018., pages 780–786. 4, 9, 18
- [60] 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, New Orleans, Louisiana. Association for Computational Linguistics. 1, 4, 19, 20

- [61] George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):138. 4, 8, 18
- [62] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. [Representation Learning with Contrastive Predictive Coding](#). 21
- [63] Christophe Van Gysel and Maarten de Rijke. 2018. Pytrec_eval: An Extremely Fast Python Interface to trec_eval. In *SIGIR*. ACM. 5
- [64] Ellen Voorhees. 2005. [Overview of the TREC 2004 Robust Retrieval Track](#). 4, 19
- [65] 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](#). *SIGIR Forum*, 54(1). 2, 4, 9, 18
- [66] Henning Wachsmuth, Martin Potthast, Khalid Al-Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. 2017. [Building an Argument Search Engine for the Web](#). In *4th Workshop on Argument Mining (ArgMining 2017) at EMNLP*, pages 49–59. Association for Computational Linguistics. 19
- [67] 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. Association for Computational Linguistics. 4, 19
- [68] 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. 4, 20
- [69] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex Wade, Kuansan Wang, Nancy Xin Ru Wang, Chris Wilhelm, Boya Xie, Douglas Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. [CORD-19: The COVID-19 Open Research Dataset](#). 18
- [70] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. [MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 5776–5788. Curran Associates, Inc. 6
- [71] Yining Wang, Liwei Wang, Yuanzhi Li, Di He, Wei Chen, and Tie-Yan Liu. 2013. A theoretical analysis of NDCG ranking measures. In *Proceedings of the 26th annual conference on learning theory (COLT 2013)*, volume 8, page 6. 5
- [72] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. [Transformers: State-of-the-Art Natural Language Processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics. 4
- [73] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. [Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval](#). 6, 9
- [74] Peilin Yang, Hui Fang, and Jimmy Lin. 2017. [Anserini: Enabling the Use of Lucene for Information Retrieval Research](#). In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’17, page 1253–1256, New York, NY, USA. Association for Computing Machinery. 4

- [75] Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, et al. 2020. Multilingual Universal Sentence Encoder for Semantic Retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94. 4
- [76] 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, Brussels, Belgium. Association for Computational Linguistics. 4, 18
- [77] Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. 2021. **Mr. TyDi: A Multi-lingual Benchmark for Dense Retrieval**. 17
- [78] Yun Zhang, David Lo, Xin Xia, and Jian-Ling Sun. 2015. Multi-factor duplicate question detection in stack overflow. *Journal of Computer Science and Technology*, 30(5):981–997. 1
- [79] Tiancheng Zhao, Xiaopeng Lu, and Kyusong Lee. 2021. **SPARTA: Efficient Open-Domain Question Answering via Sparse Transformer Matching Retrieval**. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 565–575, Online. Association for Computational Linguistics. 3, 6

Checklist

1. For all authors...
 - (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes]
 - (b) Did you describe the limitations of your work? [Yes] See Appendix B.
 - (c) Did you discuss any potential negative societal impacts of your work? [No]
 - (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
 - (a) Did you state the full set of assumptions of all theoretical results? [N/A]
 - (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
 - (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] URL mentioned in Abstract.
 - (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] All results can be reproduced by the code in our repository.
 - (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] We evaluate existing available pre-trained models that often come without suitable training code. Hence, in many cases, re-training the model is not feasible.
 - (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] We include the type of GPU and CPU resources we used, but not the total amount of compute that was used.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
 - (a) If your work uses existing assets, did you cite the creators? [Yes] Original papers are cited (if available), Table 5 contains the original website links for the used datasets.
 - (b) Did you mention the license of the assets? [Yes] See Appendix E.
 - (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] No supplemental material attached to this submission. Further supplemental material can be found in our repository mentioned in the URL.
 - (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [N/A] Used datasets provide a specific dataset license, which we follow.
 - (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] We re-use existing datasets, which most are freely available. Most datasets are from less sensitive sources, like Wikipedia or scientific publications, where don’t expect personally identifiable information. Checking for offensive content in more than 50 million documents is difficult and removing it would alter the underlying dataset.
5. If you used crowdsourcing or conducted research with human subjects...
 - (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] We ourselves performed annotation on the TREC-COVID dataset, where we followed the instructions from the original task website.
 - (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
 - (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] Annotations were done by the authors of the paper.

A Complementing Information

We provide the following additional sections in detail and information that complement discussions in the main paper:

- Limitations of the BEIR benchmark in Appendix [B](#).
- Training and in-domain evaluation task details in Appendix [C](#).
- Description of all zero-shot tasks and datasets used in BEIR in Appendix [D](#).
- Details of dataset licenses in Appendix [E](#).
- Overview of the weighted jaccard similarity metric in Appendix [F](#).
- Overview of the capped recall at k metric in Appendix [G](#).
- Length preference for dense retrieval system in Appendix [H](#).

B Limitations of the BEIR Benchmark

Even though we cover a wide range of tasks and domains in BEIR, no benchmark is perfect and has its limitations. Making those explicit is a critical point in understanding the results on the benchmark and, for future work, to improve up-on the benchmark.

1. Multilingual Tasks: Although we aim for a diverse retrieval evaluation benchmark, due to the limited availability of multilingual retrieval datasets, all datasets covered in the BEIR benchmark are currently English. It is worthwhile to add more multilingual datasets [\[2, 77\]](#) (in consideration of the selection criteria) as a next step for the benchmark. Future work could include multi- and cross-lingual tasks and models.

2. Long Document Retrieval: Most of our tasks have average document lengths up-to a few hundred words roughly equivalent to a few paragraphs. Including tasks that require the retrieval of longer documents would be highly relevant. However, as transformer-based approaches often have a length limit of 512 word pieces, a fundamental different setup would be required to compare approaches.

3. Multi-factor Search: Until now, we focused on pure textual search in BEIR. In many real-world applications, further signals are used to estimate the relevancy of documents, such as PageRank [\[49\]](#), recency [\[16\]](#), authority score [\[33\]](#) or user-interactions such as click-through rates [\[51\]](#). The integration of such signals in the tested approaches is often not straight-forward and is an interesting direction for research.

4. Multi-field Retrieval: Retrieval can often be performed over multiple fields. For example, for scientific publication we have the title, the abstract, the document body, the authors list, and the journal name. So far we focused only on datasets that have one or two fields.

5. Task-specific Models: In our benchmark, we focus on evaluating models that are able to generalize well for a broad range of retrieval tasks. Naturally in real-world, for some few tasks or domains, specialized models are available which can easily outperform generic models as they focus and perform well on a single task, lets say on question-answering. Such task-specific models do not necessarily need to generalize across all diverse tasks.

C Training and In-domain Evaluation

We use the MS MARCO Passage Ranking dataset [\[45\]](#), which contains 8.8M Passages and an official training set of 532,761 query-passage pairs for fine-tuning for a majority of retrievers. The dataset contains queries from Bing search logs with one text passage from various web sources annotated as relevant. We find the dataset useful for training, in terms of covering a wide variety of topics and providing the highest number of training pairs. It has been extensively explored and used for fine-tuning dense retrievers in recent works [\[46, 17, 15\]](#). We use the official MS MARCO development set for our in-domain evaluation which has been widely used in prior research [\[46, 17, 15\]](#). It has 6,980 queries. Most of the queries have only 1 document judged relevant; the labels are binary.

D Zero-shot Evaluation Tasks

Following the selection criteria mentioned in Section [3](#), we include 18 evaluation datasets that span across 9 heterogeneous tasks. Each dataset mentioned below contains a document corpus denoted

by \mathbf{T} and test queries for evaluation denoted by \mathbf{Q} . We additionally provide dataset website links in Table 5 and intuitive examples in Table 8. We now describe each task and dataset included in the BEIR benchmark below:

D.1 Bio-Medical Information Retrieval

Bio-medical information retrieval is the task of searching relevant scientific documents such as research papers or blogs for a given scientific query in the biomedical domain [28]. We consider a scientific query as *input* and retrieve bio-medical documents as *output*.

TREC-COVID [65] is an ad-hoc search challenge based on the CORD-19 dataset containing scientific articles related to the COVID-19 pandemic [69]. We include the July 16, 2020 version of CORD-19 dataset as corpus \mathbf{T} and use the final cumulative judgements with query descriptions from the original task as queries \mathbf{Q} .

NFCorpus [7] contains natural language queries harvested from NutritionFacts (NF). We use the original splits provided alongside all content sources from NF (videos, blogs, and Q&A posts) as queries \mathbf{Q} and annotated medical documents from PubMed as corpus \mathbf{T} .

BioASQ [61] Task 8b is a biomedical semantic question answering challenge. We use the original train and test splits provided in Task 8b as queries \mathbf{Q} and collect around 15M articles from PubMed provided in Task 8a as our corpus \mathbf{T} .

D.2 Open-domain Question Answering (QA)

Retrieval in open domain question answering [8] is the task of retrieving the correct answer for a question, without a predefined location for the answer. In open-domain tasks, model must retrieve over an entire knowledge source (such as Wikipedia). We consider the question as *input* and the passage containing the answer as *output*.

Natural Questions [34] contains Google search queries and documents with paragraphs and answer spans within Wikipedia articles. We did not use the NQ version from ReQA [1] as it focused on queries having a short answer. As a result, we parsed the HTML of the original NQ dataset and include more complex development queries that often require a longer passage as answer compared to ReQA. We filtered out queries without an answer, or having a table as an answer, or with conflicting Wikipedia pages. We retain 2,681,468 passages as our corpus \mathbf{T} and 3452 test queries \mathbf{Q} from the original dataset.

HotpotQA [76] contains multi-hop like questions which require reasoning over multiple paragraphs to find the correct answer. We include the original full-wiki task setting: utilizing processed Wikipedia passages as corpus \mathbf{T} . We held out randomly sampled 5447 queries from training as our dev split. We use the original (paper) task’s development split as our test split \mathbf{Q} .

FiQA-2018 [44] Task 2 consists of opinion-based question-answering. We include financial data by crawling StackExchange posts under the Investment topic from 2009-2017 as our corpus \mathbf{T} . We randomly sample out 500 and 648 queries \mathbf{Q} from the original training split as dev and test splits.

D.3 Tweet Retrieval

Twitter is a popular micro-blogging website on which people post real-time messages (i.e. tweets) about their opinions on a variety of topics and discuss current issues. We consider a news headline as *input* and retrieve relevant tweets as *output*.

Signal-1M Related Tweets [59] task retrieves relevant tweets for a given news article title. The Related Tweets task provides news articles from the Signal-1M dataset [10] which we use as queries \mathbf{Q} . We construct our twitter corpus \mathbf{T} by manually scraping tweets from the provided tweet-ids in the relevancy judgements using Python package: Tweepy (<https://www.tweepy.org>).

D.4 News Retrieval

TREC-NEWS [58] 2019 track involves background linking: Given a news headline, we retrieve relevant news articles that provide important context or background information. We include the original shared task query description (single sentence) as our test queries \mathbf{Q} and the TREC Washington Post as our corpus \mathbf{T} . For simplicity, we convert the original exponential gain relevant judgements to linear labels.

Robust04 [64] provides a robust dataset focusing on evaluating on poorly performing topics. We include the original shared task query description (single sentence) as our test queries **Q** and the complete TREC disks 4 and 5 documents as our corpus **T**.

D.5 Argument Retrieval

Argument retrieval is the task of ranking argumentative texts in a collection of focused arguments (*output*) in order of their relevance to a textual query (*input*) on different topics.

ArguAna Counterargs Corpus [67] involves the task of retrieval of the best counterargument to an argument. We include pairs of arguments and counterarguments scraped from the online debate portal as corpus **T**. We consider the arguments present in the original test split as our queries **Q**.

Touché-2020 [6] Task 1 is a conversational argument retrieval task. We use the conclusion as title and premise for arguments present in args.me [66] as corpus **T**. We include the shared Touché-2020 task data as our test queries **Q**. The original relevance judgements (qrels) file also included negative judgements (-2) for non-arguments present within the corpus, but for simplicity we substitute them as zero.

D.6 Duplicate Question Retrieval

Duplicate question retrieval is the task of identifying duplicate questions asked in community question answering (cQA) forums. A given query is the *input* and the duplicate questions are the *output*.

CQADupStack [25] is a popular dataset for research in community question-answering (cQA). The corpus **T** comprises of queries from 12 different StackExchange subforums: Android, English, Gaming, Gis, Mathematica, Physics, Programmers, Stats, Tex, Unix, Webmasters and Wordpress. We utilize the original test split for our queries **Q**, and the task involves retrieving duplicate query (title + body) for an input query title. We evaluate each StackExchange subforum separately and report the overall mean scores for all tasks in BEIR.

Quora Duplicate Questions dataset identifies whether two questions are duplicates. Quora originally released containing 404,290 question pairs. We add transitive closures to the original dataset. Further, we split it into train, dev, and test sets with a ratio of about 85%, 5% and 10% of the original pairs. We remove all overlaps between the splits and ensure that a question in one split of the dataset does not appear in any other split to mitigate the transductive classification problem [27]. We achieve 522,931 unique queries as our corpus **T** and 5,000 dev and 10,000 test queries **Q** respectively.

D.7 Entity Retrieval

Entity retrieval involves retrieving unique Wikipedia pages to entities mentioned in the query. This is crucial for tasks involving Entity Linking (EL). The entity-bearing query is the *input* and the entity abstract and title are retrieved as *output*.

DBPedia-Entity-v2 [21] is an established entity retrieval dataset. It contains a set of heterogeneous entity-bearing queries **Q** containing named entities, IR style keywords, and natural language queries. The task involves retrieving entities from the English part of DBpedia corpus **T** from October 2015. We randomly sample out 67 queries from the test split as our dev set.

D.8 Citation Prediction

Citations are a key signal of relatedness between scientific papers [9]. In this task, the model attempts to retrieve cited papers (*output*) for a given paper title as *input*.

SCIDOCS [9] contains a corpus **T** of 30K held-out pool of scientific papers. We consider the direct-citations (1 out of 7 tasks mentioned in the original paper) as the best suited task for retrieval evaluation in BEIR. The task includes 1k papers as queries **Q** with 5 relevant papers and 25 (randomly selected) uncited papers for each query.

D.9 Fact Checking

Fact checking verifies a claim against a big collection of evidence [60]. The task requires knowledge about the claim and reasoning over multiple documents. We consider a sentence-level claim as *input* and the relevant document passage verifying the claim as *output*.

FEVER [60] The Fact Extraction and VERification dataset is collected to facilitate the automatic fact checking. We utilize the original paper splits as queries \mathbf{Q} and retrieve evidences from the pre-processed Wikipedia Abstracts (June 2017 dump) as our corpus \mathbf{T} .

Climate-FEVER [14] is a dataset for verification of real-world climate claims. We include the original dataset claims as queries \mathbf{Q} and retrieve evidences from the same FEVER Wiki corpus \mathbf{T} . We manually included few Wikipedia articles (25) missing from our corpus, but present within our relevance judgements.

SciFact [68] verifies scientific claims using evidence from the research literature containing scientific paper abstracts. We use the original publicly available dev split from the task containing 300 queries as our test queries \mathbf{Q} , and include all documents from the original dataset as our corpus \mathbf{T} .

E Dataset Licenses

The authors of 4 out of the 19 datasets in the BEIR benchmark (NFCorpus, FiQA-2018, Quora, Climate-Fever) do not report the dataset license in the paper or a repository; We overview the rest:

- MSMARCO: Provided under “MIT License” for non-commercial research purposes.
- FEVER, NQ, DBPedia, Signal-1M: All provided under CC BY-SA 3.0 license.
- TREC-NEWS, Robust04, BioASQ: Data collection archives are under **Copyright**.
- ArguAna, Touché-2020: Provided under CC BY 4.0 license.
- CQADupStack: Provided under Apache License 2.0 license.
- SciFact: Provided under the CC BY-NC 2.0 license.
- SCIDOCS: Provided under the GNU General Public License v3.0 license.
- HotpotQA: Provided under the CC BY-SA 4.0 license.
- TREC-COVID: Provided under the “Dataset License Agreement”.

F Weighted Jaccard Similarity

The weighted Jaccard similarity $J(S, T)$ [26] is intuitively calculated as the unique word overlap for all words present in both the datasets. More formally, the normalized frequency for a unique word k in a dataset is calculated as the frequency of word k divided over the sum of frequencies of all words in the dataset.

S_k is the normalized frequency of word k in the source dataset S and T_k for the target dataset T respectively. The weighted Jaccard similarity between S and T is defined as:

$$J(S, T) = \frac{\sum_k \min(S_k, T_k)}{\sum_k \max(S_k, T_k)}$$

where the sum is over all unique words k present in datasets S and T .

G Capped Recall@k Score

Recall at k is calculated as the fraction of the relevant documents that are successfully retrieved within the top k extracted documents. More formally, the $R@k$ score is calculated as:

$$R@k = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{|\max_k(A_i) \cap A_i^*|}{|A_i^*|}$$

where Q is the set of queries, A_i^* is the set of relevant documents for the i th query, and A_i is a scored list of documents provided by the model, from which top k are extracted.

However measuring recall can be counterintuitive, if a high number of relevant documents ($> k$) are present within a dataset. For example, consider a hypothetical dataset with 500 relevant documents for a query. Retrieving all relevant documents would produce a maximum $R@100$ score = 0.2, which

is quite low and unintuitive. To avoid this we cap the recall score ($R_{cap}@k$) at k for datasets if the number of relevant documents for a query greater than k . It is defined as:

$$R_{cap}@k = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{|\max_k(A_i) \cap A_i^*|}{\min(k, |A_i^*|)}$$

where the only difference lies within the denominator where we compute the minimum of k and $|A_i^*|$, instead of $|A_i^*|$ present in the original recall.

H Document Length Preference for Dense Retrieval System

As we show in Figure 4, TAS-B prefers retrieval of shorter documents, and in comparison, ANCE retrieves longer documents. The difference is especially extreme for the TREC-COVID dataset: TAS-B retrieves lots of top hit documents containing only a title and an empty abstract, while ANCE retrieves top hit documents with a non-empty abstract.

Identifying the source for this contrasting behaviour is difficult, as TAS-B and ANCE use different models (DistilBERT vs. RoBERTa-base), a different loss function (InfoNCE [62] vs. Margin-MSE [24] with in-batch negatives), and different hard negative mining strategies. Hence, we decided to harmonize the training setup and to alter the training by just one aspect: The similarity function.

Dense models require a similarity function to retrieve relevant documents for a given query within an embedding space. This similarity function is also used during training dense models with the InfoNCE [62] loss:

$$\mathcal{L}_q = -\log \frac{\exp(\tau \cdot \text{sim}(q, d_+))}{\sum_{i=0}^n \exp(\tau \cdot \text{sim}(q, d_i))}$$

using n in-batch negatives for each query q and a scaling factor τ . where d_+ denotes the relevant (positive) document for query q . Commonly used similarity functions ($\text{sim}(q, d)$) are cosine-similarity or dot-product.

We trained two distilbert-base-uncased models with an identical training setup on MS MARCO (identical training parameters) and only changed the similarity function from cosine-similarity to dot-product. As shown in Table 10, we observe significant performance differences for some BEIR datasets. For TREC-COVID, the dot-product model achieves the biggest improvement with 15.3 points, while for a majority on other datasets, it performs worse than the cosine-similarity model.

We observe that these (nearly) identical models retrieve documents with vastly different lengths as shown in the violin plots in Table 10. For all datasets, we find the cosine-similarity model to prefer shorter documents over longer ones. This is especially severe for TREC-COVID: a large fraction of the scientific papers (approx. 42k out of 171k) consist only of publication titles without an abstract. The cosine-similarity model prefers retrieving these documents. In contrast, the dot-product model primarily retrieves longer documents, i.e., publications with an abstract. Cosine-similarity uses vectors of unit length, thereby having no notion of the encoded text length. In contrast, for dot-product, longer documents result in vectors with higher magnitudes which can yield higher similarity scores for a query.

Further, as we observe in Figure 5, relevance judgement scores are not uniformly distributed over document lengths: for some datasets, longer documents are annotated with higher relevancy scores, while in others, shorter documents are. This can be either due to the annotation process, e.g., the candidate selection method prefers short or long documents, or due to the task itself, where shorter or longer documents could be more relevant to the user information need. Hence, it can be more advantageous to train a model with either cosine-similarity or dot-product depending upon the nature and needs of the specific task.

Dataset	Website (Link)
MS MARCO	https://microsoft.github.io/msmarco/
TREC-COVID	https://ir.nist.gov/covidSubmit/index.html
NFCorpus	https://www.cl.uni-heidelberg.de/statnlpgroup/nfcorpus/
BioASQ	http://bioasq.org
NQ	https://ai.google.com/research/NaturalQuestions
HotpotQA	https://hotpotqa.github.io
FiQA-2018	https://sites.google.com/view/fiqa/
Signal-1M (RT)	https://research.signal-ai.com/datasets/signal1m-tweetir.html
TREC-NEWS	https://trec.nist.gov/data/news2019.html
Robust04	https://trec.nist.gov/data/t13_robust.html
ArguAna	http://argumentation.bplaced.net/arguana/data
Touchè-2020	https://webis.de/events/touche-20/shared-task-1.html
CQADupStack	http://nlp.cis.unimelb.edu.au/resources/cqadupstack/
Quora	https://www.quora.com/q/quoradata/First-Quora-Dataset-Release-Question-Pairs
DBPedia-Entity	https://github.com/iai-group/DBpedia-Entity/
SCIDocs	https://allenai.org/data/scidocs
FEVER	http://fever.ai
Climate-FEVER	http://climatefever.ai
SciFact	https://github.com/allenai/scifact

Table 5: Original dataset website (link) for all datasets present in **BEIR**.

Model	Public Model Checkpoints (Link)
BM25 (Anserini)	https://github.com/castorini/anserini
DeepCT	http://boston.lti.cs.cmu.edu/appendices/arXiv2019-DeepCT-Zhuyun-Dai/
SPARTA	https://huggingface.co/BeIR/sparta-msmarco-distilbert-base-v1
DocT5query	https://huggingface.co/BeIR/query-gen-msmarco-t5-base-v1
DPR (Query)	https://huggingface.co/sentence-transformers/facebook-dpr-question_encoder-multiset-base
DPR (Context)	https://huggingface.co/sentence-transformers/facebook-dpr-ctx_encoder-multiset-base
ANCE	https://huggingface.co/sentence-transformers/msmarco-roberta-base-ance-firsttp
TAS-B	https://huggingface.co/sentence-transformers/msmarco-distilbert-base-tas-b
ColBERT	https://public.ukp.informatik.tu-darmstadt.de/thakur/BEIR/models/ColBERT/msmarco.psg.l2.zip
MiniLM-L6 (CE)	https://huggingface.co/cross-encoder/ms-marco-MiniLM-L-6-v2

Table 6: Publicly available model links used for evaluation in **BEIR**.

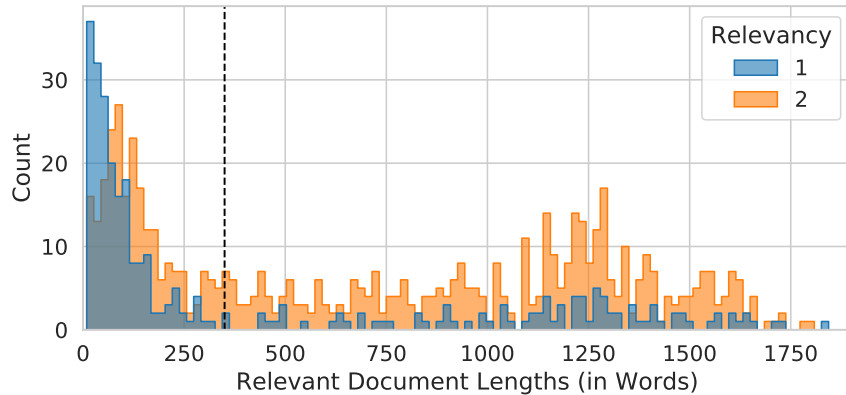


Figure 5: Annotated original relevant document lengths (in words) for Touchè-2020 [6]. Majority of the relevant documents (score = 2) on average in the original dataset are longer. Many shorter documents are annotated as less relevant (score = 1).

Corpus	Website (Link)
CORD-19	https://www.semanticscholar.org/cord19
NutritionFacts	https://nutritionfacts.org
PubMed	https://pubmed.ncbi.nlm.nih.gov
Signal-1M	https://research.signal-ai.com/datasets/signal1m.html
TREC Washington Post	https://ir.nist.gov/wapo/
TREC disks 4 and 5	https://trec.nist.gov/data/cd45/
Args.me	https://zenodo.org/record/4139439/
DBPedia (2015-10)	http://downloads.dbpedia.org/wiki-archive/Downloads2015-10.html
TREC-COVID (Annotated)	https://public.ukp.informatik.tu-darmstadt.de/thakur/BEIR/datasets/trec-covid-beir.zip

Table 7: Corpus Name and Link used for datasets in **BEIR**.

Dataset	Query	Relevant-Documents
MS MARCO	what fruit is native to australia	<i><Paragraph></i> Passiflora herbertiana. A rare passion fruit native to Australia. Fruits are green-skinned, white fleshed, with an unknown edible rating. Some sources list the fruit as edible, sweet and tasty, while others list the fruits as being bitter and inedible. assiflora herbertiana. A rare passion fruit native to Australia...
TREC-COVID	what is the origin of COVID-19	<i><Title></i> Origin of Novel Coronavirus (COVID-19): A Computational Biology Study using Artificial Intelligence <i><Paragraph></i> Origin of the COVID-19 virus has been intensely debated in the community...
BioASQ	What is the effect of HMGB2 loss on CTCF clustering	<i><Title></i> HMGB2 Loss upon Senescence Entry Disrupts Genomic Organization and Induces CTCF Clustering across Cell Types. <i><Paragraph></i> Processes like cellular senescence are characterized by complex events giving rise to heterogeneous cell populations. However, the early molecular events driving this cascade remain elusive...
NFCorpus	Titanium Dioxide & Inflammatory Bowel Disease	<i><Title></i> Titanium Dioxide Nanoparticles in Food and Personal Care Products <i><Paragraph></i> Titanium dioxide is a common additive in many food, personal care, and other consumer products used by people, which after use can enter the sewage system, and subsequently enter the environment as treated effluent discharged to surface waters or biosolids applied to agricultural land, or incinerated wastes...
NQ	when did they stop cigarette advertising on television?	<i><Title></i> Tobacco advertising <i><Paragraph></i> The first calls to restrict advertising came in 1962 from the Royal College of Physicians, who highlighted the health problems and recommended stricter laws...
HotpotQA	Stockely Webster has paintings hanging in what home (that serves as the residence for the Mayor of New York)?	<i><Title></i> Stokely Webster <i><Paragraph></i> Stokely Webster (1912 – 2001) was best known as an American impressionist painter who studied in Paris. His paintings can be found in the permanent collections of many museums, including the Metropolitan Museum of Art in New York, the National Museum...
FiQA-2018	What is the PEG ratio? How is the PEG ratio calculated? How is the PEG ratio useful for stock investing?	<i><Paragraph></i> PEG is Price/Earnings to Growth. It is calculated as Price/Earnings/Annual EPS Growth. It represents how good a stock is to buy, factoring in growth of earnings, which P/E does not. Obviously when PEG is lower, a stock is more undervalued, which means that it is a better buy, and more likely...
Signal-1M (RT)	Genvoya, a Gentler Anti-HIV Cocktail, Okayed by EU Regulators	<i><Paragraph></i> All people with #HIV should get anti-retroviral drugs: @WHO, by @kkelland via @Reuters_Health #AIDS #IasP
TREC-NEWS	Websites where children are prostituted are immune from prosecution. But why?	<i><Title></i> Senate launches bill to remove immunity for websites hosting illegal content, spurred by Backpage.com <i><Paragraph></i> The legislation, along with a similar bill in the House, sets the stage for a battle between Congress and some of the Internet's most powerful players, including Google and various free-speech advocates, who believe that Congress shouldn't regulate Web content or try to force websites to police themselves more rigorously...
Robust04	What were the causes for the Islamic Revolution relative to relations with the U.S.?	<i><Paragraph></i> BFN [Editorial: "Sow the Wind and Reap the Whirlwind"] Yesterday marked the 14th anniversary of severing of diplomatic relations between the Islamic Republic and the United States of America. Several occasions arose in the last decade and a half for improving Irano-American relations...
Touché-2020	Should the government allow illegal immigrants to become citizens?	<i><Title></i> America should support blanket amnesty for illegal immigrants. <i><Paragraph></i> Undocumented workers do not receive full Social Security benefits because they are not United States citizens " nor should they be until they seek citizenship legally. Illegal immigrants are legally obligated to pay taxes...
CQADupStack	Command to display first few and last few lines of a file	<i><Title></i> Combing head and tail in a single call via pipe <i><Paragraph></i> On a regular basis, I am piping the output of some program to either 'head' or 'tail'. Now, suppose that I want to see the first AND last 10 lines of piped output, such that I could do something like ./lotsoutput headtail...
Quora	How long does it take to methamphetamine out of your blood?	<i><Paragraph></i> How long does it take the body to get rid of methamphetamine?
DBPedia	Paul Auster novels	<i><Title></i> The New York Trilogy <i><Paragraph></i> The New York Trilogy is a series of novels by Paul Auster. Originally published sequentially as City of Glass (1985), Ghosts (1986) and The Locked Room (1986), it has since been collected into a single volume.
SCIDOCS	CFD Analysis of Convective Heat Transfer Coefficient on External Surfaces of Buildings	<i><Title></i> Application of CFD in building performance simulation for the outdoor environment: an overview <i><Paragraph></i> This paper provides an overview of the application of CFD in building performance simulation for the outdoor environment, focused on four topics...
FEVER	DodgeBall: A True Underdog Story is an American movie from 2004	<i><Title></i> DodgeBall: A True Underdog Story <i><Paragraph></i> DodgeBall: A True Underdog Story is a 2004 American sports comedy film written and directed by Rawson Marshall Thurber and starring Vince Vaughn and Ben Stiller. The film follows friends who enter a dodgeball tournament...
Climate-FEVER	Sea level rise is now increasing faster than predicted due to unexpectedly rapid ice melting.	<i><Title></i> Sea level rise <i><Paragraph></i> A sea level rise is an increase in the volume of water in the world's oceans, resulting in an increase in global mean sea level. The rise is usually attributed to global climate change by thermal expansion of the water in the oceans and by melting of ice sheets and glaciers...

Table 8: Examples of queries and relevant documents for all datasets included in **BEIR**. (*<Title>*) and (*<Paragraph>*) are used to distinguish the title separately from the paragraph within a document in the table above. These tokens were not passed to the respective models.

Model (→)	Lexical	Sparse			Dense				Late-Interaction	Re-ranking
Dataset (↓)	BM25	DeepCT	SPARTA	docT5query	DPR	ANCE	TAS-B	GenQ	ColBERT	BM25+CE
MS MARCO	0.658	0.752 [‡]	0.793 [‡]	0.819 [‡]	0.552	0.852 [‡]	0.884[‡]	0.884[‡]	<u>0.865[‡]</u>	0.658 [‡]
TREC-COVID	<u>0.498*</u>	0.347*	0.409*	0.541*	0.212*	0.457*	0.387*	0.456*	0.464*	<u>0.498*</u>
BioASQ	0.714	<u>0.699</u>	0.351	0.646	0.256	0.463	0.579	0.627	0.645	0.714
NFCorpus	0.250	0.235	0.243	0.253	0.208	0.232	0.280	0.280	<u>0.254</u>	0.250
NQ	0.760	0.636	0.787	0.832	0.880 [‡]	0.836	<u>0.903</u>	0.862	0.912	0.760
HotpotQA	<u>0.740</u>	0.731	0.651	0.709	0.591	0.578	0.728	0.673	0.748	<u>0.740</u>
FiQA-2018	0.539	0.489	0.446	0.598	0.342	0.581	0.593	0.618	<u>0.603</u>	0.539
Signal-1M (RT)	0.370	0.299	0.270	<u>0.351</u>	0.162	0.239	0.304	0.281	0.283	0.370
TREC-NEWS	<u>0.422</u>	0.316	0.262	0.439	0.215	0.398	0.418	0.412	0.367	<u>0.422</u>
Robust04	0.375	0.271	0.215	<u>0.357</u>	0.211	0.274	0.331	0.298	0.310	0.375
ArguAna	0.942	0.932	0.893	<u>0.972</u>	0.751	0.937	0.942	0.978	0.914	0.942
Touché-2020	<u>0.538</u>	0.406	0.381	0.557	0.301	0.458	0.431	0.451	0.439	<u>0.538</u>
CQADupStack	0.606	0.545	0.521	<u>0.638</u>	0.403	0.579	0.622	0.654	0.624	0.606
Quora	0.973	0.954	0.896	0.982	0.470	0.987	0.986	<u>0.988</u>	0.989	0.973
DBPedia	0.398	0.372	0.411	0.365	0.349	0.319	0.499	0.431	<u>0.461</u>	0.398
SCIDOCS	<u>0.356</u>	0.314	0.297	0.360	0.219	0.269	0.335	0.332	0.344	<u>0.356</u>
FEVER	0.931	0.735	0.843	0.916	0.840	0.900	0.937	0.928	<u>0.934</u>	0.931
Climate-FEVER	0.436	0.232	0.227	0.427	0.390	0.445	0.534	<u>0.450</u>	0.444	0.436
SciFact	<u>0.908</u>	0.893	0.863	0.914	0.727	0.816	0.891	0.893	0.878	<u>0.908</u>

Table 9: In-domain and zero-shot retrieval performance on BEIR datasets. Scores denote **Recall@100**. The best retrieval performance on a given dataset is marked in **bold**, and the second best performance is underlined. [‡] indicates in-domain retrieval performance. * shows the capped Recall@100 score (Appendix G).

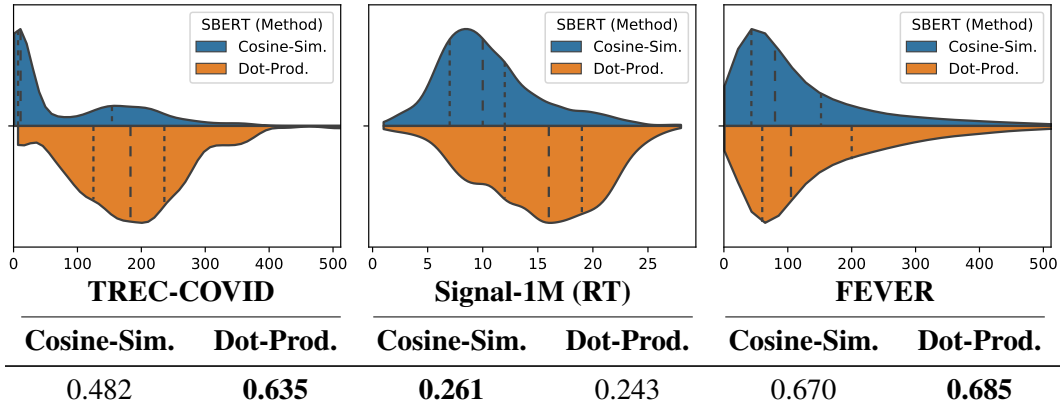


Table 10: Violin plots [22] of document lengths for the top-10 retrieved hits and nDCG@10 scores using a distilbert-base-uncased model trained with either cosine similarity (blue, top) or dot product (orange, bottom) as described in Appendix H.