A Complementing Information

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

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

B Training and In-domain Evaluation

We use the MS MARCO Passage Ranking dataset [42], 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 [43, 15, 48]. We use the official MS MARCO development set for our in-domain evaluation which has been widely used in prior research [43, 15, 48]. It has 6,980 queries. Most of the queries have only 1 document judged relevant; the labels are binary.

C 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:

C.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 [26]. We consider a scientific query as *input* and retrieve bio-medical documents as *output*.

TREC-COVID [63] is an ad-hoc search challenge based on the CORD-19 dataset containing scientific articles related to the COVID-19 pandemic [67]. 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 [59] 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 $\mathbf{T}$.

C.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 [32] 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 [74] 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 [41] 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.

C.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 [57] 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).

C.4 News Retrieval

TREC-NEWS [56] 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 [62] 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 $\mathbf{Q}$ and the complete TREC disks 4 and 5 documents as our corpus $\mathbf{T}$.

C.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 [65] 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 $\mathbf{T}$. We consider the arguments present in the original test split as our queries $\mathbf{Q}$.

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 [64] as corpus $\mathbf{T}$. We include the shared Touché-2020 task data as our test queries $\mathbf{Q}$. 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.

C.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 [23] is a popular dataset for research in community question-answering (cQA). The corpus $\mathbf{T}$ 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 $\mathbf{Q}$, 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 [25]. We achieve 522,931 unique queries as our corpus $\mathbf{T}$ and 5,000 dev and 10,000 test queries $\mathbf{Q}$ respectively.

C.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 [19] is an established entity retrieval dataset. It contains a set of heterogeneous entity-bearing queries $\mathbf{Q}$ containing named entities, IR style keywords, and natural language queries. The task involves retrieving entities from the English part of DBpedia corpus $\mathbf{T}$ from October 2015. We randomly sample out 67 queries from the test split as our dev set.

C.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 $\mathbf{T}$ 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 $\mathbf{Q}$ with 5 relevant papers and 25 (randomly selected) uncited papers for each query.

C.9 Fact Checking

Fact checking verifies a claim against a big collection of evidence [58]. 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 [58] 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 [13] 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 [66] 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 $\mathbf{T}$.

D 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".

E Weighted Jaccard Similarity

The weighted Jaccard similarity $J(S, T)$ [24] is intuitively calculated as the unique word overlap for all words present in both the datasets. More formally, the normalized frequency for an 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.

F 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:

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

where Q is the set of queries, $A_i^{\star}$ is the set of relevant documents for the ith 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:

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

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

G 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 [60] vs. Margin-MSE [22] 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 [60] 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-firstp
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**.

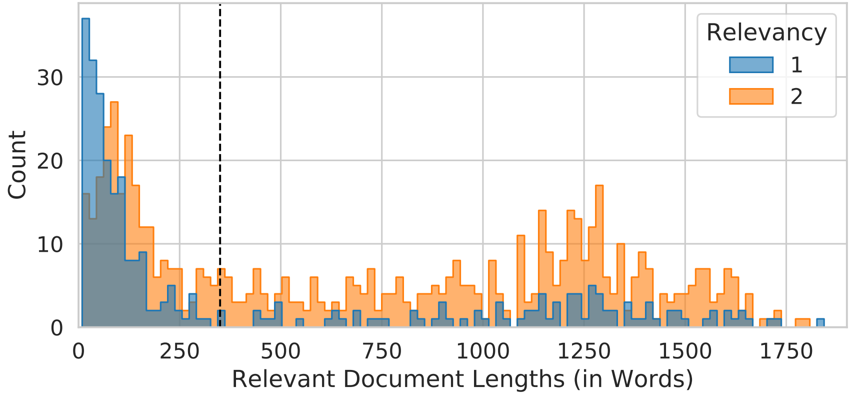


Figure 5: Annotated original relevant document lengths (in words) for Touché-2020 [6]. Majority of the relevant documents (score = 2) on average in the original dataset are longer. Many shorter documents are annotated as less relevant (score = 1).

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`

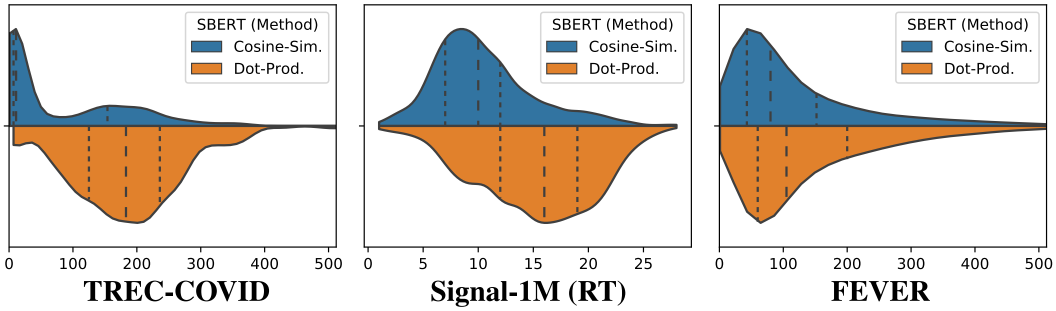
Table 7: Corpus Name and Link used for datasets in **BEIR**.

Dataset	**Query**	**Relevant-Document**
MS MARCO	what fruit is native to australia	<*Paragraph*> 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	<*Title*> Origin of Novel Coronavirus (COVID-19): A Computational Biology Study using Artificial Intelligence <*Paragraph*> Origin of the COVID-19 virus has been intensely debated in the community...
BioASQ	What is the effect of HMGB2 loss on CTCF clustering	<*Title*> HMGB2 Loss upon Senescence Entry Disrupts Genomic Organization and Induces CTCF Clustering across Cell Types. <*Paragraph*> 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	<*Title*> Titanium Dioxide Nanoparticles in Food and Personal Care Products <*Paragraph*> 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?	<*Title*> Tobacco advertising <*Paragraph*> 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)?	<*Title*> Stokely Webster <*Paragraph*> 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?	<*Paragraph*> 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	<*Paragraph*> All people with #HIV should get anti-retroviral drugs: @WHO, by @kkelland via @Reuters_Health #AIDS #TasP
TREC-NEWS	Websites where children are prostituted are immune from prosecution. But why?	<*Title*> Senate launches bill to remove immunity for websites hosting illegal content, spurred by Backpage.com <*Paragraph*> 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.?	<*Paragraph*> 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?	<*Title*> America should support blanket amnesty for illegal immigrants. <*Paragraph*> 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	<*Title*> Combing head and tail in a single call via pipe <*Paragraph*> 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 ./lotsofoutput | headtail...
Quora	How long does it take to methamphetamine out of your blood?	<*Paragraph*> How long does it take the body to get rid of methamphetamine?
DBPedia	Paul Auster novels	<*Title*> The New York Trilogy <*Paragraph*> 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	<*Title*> Application of CFD in building performance simulation for the outdoor environment: an overview <*Paragraph*> 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	<*Title*> DodgeBall: A True Underdog Story <*Paragraph*> 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.	<*Title*> Sea level rise <*Paragraph*> 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**‡	0.865‡	0.658‡
TREC-COVID	0.498*	0.347*	0.409*	**0.541***	0.212*	0.457*	0.387*	0.456*	0.464*	0.498*
BioASQ	**0.714**	0.699	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**	0.254	0.250
NQ	0.760	0.636	0.787	0.832	0.880‡	0.836	0.903	0.862	**0.912**	0.760
HotpotQA	0.740	0.731	0.651	0.709	0.591	0.578	0.728	0.673	**0.748**	0.740
FiQA-2018	0.539	0.489	0.446	0.598	0.342	0.581	0.593	**0.618**	0.603	0.539
Signal-1M (RT)	**0.370**	0.299	0.270	0.351	0.162	0.239	0.304	0.281	0.283	**0.370**
TREC-NEWS	0.422	0.316	0.262	**0.439**	0.215	0.398	0.418	0.412	0.367	0.422
Robust04	**0.375**	0.271	0.215	0.357	0.211	0.274	0.331	0.298	0.310	**0.375**
ArguAna	0.942	0.932	0.893	0.972	0.751	0.937	0.942	**0.978**	0.914	0.942
Touché-2020	0.538	0.406	0.381	**0.557**	0.301	0.458	0.431	0.451	0.439	0.538
CQADupStack	0.606	0.545	0.521	0.638	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	0.988	**0.989**	0.973
DBPedia	0.398	0.372	0.411	0.365	0.349	0.319	**0.499**	0.431	0.461	0.398
SCIDOCS	0.356	0.314	0.297	**0.360**	0.219	0.269	0.335	0.332	0.344	0.356
FEVER	0.931	0.735	0.843	0.916	0.840	0.900	**0.937**	0.928	0.934	0.931
Climate-FEVER	0.436	0.232	0.227	0.427	0.390	0.445	**0.534**	0.450	0.444	0.436
SciFact	0.908	0.893	0.863	**0.914**	0.727	0.816	0.891	0.893	0.878	0.908

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 F).



TREC-COVID		Signal-1M (RT)		FEVER	
Cosine-Sim.	**Dot-Prod.**	**Cosine-Sim.**	**Dot-Prod.**	**Cosine-Sim.**	**Dot-Prod.**
0.482	**0.635**	**0.261**	0.243	0.670	**0.685**

Table 10: Violin plots [20] 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 G.