

Corpus Subsampling: Estimating the Effectiveness of Neural Retrieval Models on Large Corpora

Maik Fröbe¹, Andrew Parry², Harrison Scells³, Shuai Wang⁴, Shengyao Zhuang^{4,5}, Guido Zuccon⁴, Martin Potthast⁶, and Matthias Hagen¹

¹ Friedrich-Schiller-Universität Jena

² University of Glasgow

³ University of Kassel and hessian.AI

⁴ University of Queensland

⁵ CSIRO

⁶ University of Kassel, hessian.AI, and ScaDS.AI

Abstract. Due to their low efficiency, neural retrieval models are usually evaluated on small corpora (e.g. MS MARCO or BEIR subsets) or in re-ranking scenarios using a more efficient first-stage retriever. To estimate their effectiveness on larger corpora independently of a first-stage retriever, we propose a new corpus subsampling strategy based on the top- k results of the pooled systems that contributed to the relevance judgments of a corpus. Our experiments on nine TREC tasks covering different corpus sizes show that using the top-1,000 or even only the top-100 pools provides a reliable effectiveness estimate for neural models. This reduces the required experimental resources for large corpora by a factor of up to 1,000 and enables a “green” IR evaluation.

Keywords: Evaluation · Neural Retrieval · Green IR Evaluation

1 Introduction

Over time, the document collections used in information retrieval research have become larger to make laboratory experiments more realistic. For example, the ClueWeb09 and the ClueWeb12⁷ each comprise around one billion documents, while the ClueWeb22 [44] comprises ten billion documents. To evaluate a given set of systems on large collections, a considerable number of judgments is required (e.g., 20 per topic for MTC [1]), whereas the validity of an evaluation is typically low with only a few judgments. Few (binary) judgments per topic impair the discriminatory power of evaluation measures like nDCG [30], which discriminate systems better the more (graded) judgments are available [41]. The ClueWeb09/12 were frequently used in TREC-like evaluations, so a large number of graded relevance judgments are available for many topics. Evaluations on collections with many judgments are more likely valid when they build on pools of different systems [66], but also failed validity tests have been reported [67].

⁷ <https://lemurproject.org/clueweb09/> and <https://lemurproject.org/clueweb12/>

Table 1: Our recommended subsamples for reliable evaluation at substantially reduced computational effort by reducing documents not on topic (\notin_J).

Corpus	Tracks	Judgm.	Complete			Subsampled		
			Docs.	\notin_J	Size	Docs.	\notin_J	Size
ClueWeb09	Web 09–12 [13–16]	84 366	1.0 b	99 %	4.0 TB	0.3 m	73 %	0.9 GB
ClueWeb12	Web 13/14 [17, 18]	28 906	0.7 b	99 %	4.5 TB	0.1 m	72 %	0.5 GB
Disks 4/5	Robust04 [64]	311 410	0.5 m	41 %	0.6 GB	0.4 m	31 %	0.5 GB
MS MARCO	DL 19/20 [21, 22]	20 646	8.8 m	99 %	2.9 GB	0.3 m	97 %	42.1 MB

Still, evaluating neural retrievers (e.g., LLMs [36]) on large collections is not really practical: indexing 50% of the ClueWeb09 with an efficient neural retrieval model already takes 71 days on an Nvidia V100 GPU [33]. More feasible are smaller collections like the ones from the BEIR meta-dataset [60] (19 collections with median size of 530,000 documents) but they have only a limited number of relevant documents (median of 3 per topic). At the same time, when evaluating a new retriever in post-hoc experiments (i.e., after a collection has been judged), many of the retrieved results may be unjudged and the common practice of then assuming them to be non-relevant or measuring effectiveness on condensed lists [49] often considerably under- or overestimates effectiveness [25, 50]. As a way out, subsampling (subsets of) the judged documents and combining them with random documents has been proposed [34, 57, 58]. But for large collections, random documents are “easy negatives” and unlikely to be retrieved. This, too, leads to an overestimation of effectiveness as in the case of using condensed lists.

In this paper, we compare different subsampling approaches to systematically study which strategies allow for reliable post-hoc experiments at a reasonable compute budget. But in contrast to previous sampling approaches [34, 57, 58], we use the runs that contributed to the judgment pool. The runs’ top results below a collection’s pooling depth include a diversity of hard negative unjudged documents, minimizing the overestimation of effectiveness. We evaluate eight sampling strategies (Section 3) using leave-one-group-out retrieval (Section 4) on diverse collections (between 500,000 and 1 billion documents) and find that evaluation becomes reliable when subsamples are pooled to a depth of 100 up to 1,000. Table 1 provides an overview of our recommended subsamples that we found in our experiments for the nine TREC tasks on four corpora, yielding dataset sizes of only a few hundred Megabytes. As these subsamples are way easier to index and to process at retrieval time than whole collections, the subsampling-based evaluation fulfills one of the goals of Green IR [53]: minimizing the resources of IR evaluation.⁸

⁸ Our code and data are available at: <https://github.com/webis-de/ECIR-25>

2 Related Work

Reliability of IR Experiments. IR experiments are reliable if observations (e.g., System A > System B) transfer to similar scenarios with a high probability [66]. Two main aspects impact reliability [66]: subjectiveness and incompleteness of relevance judgments. TREC-style test collections are constructed as a community effort. Teams submit runs that are subsequently pooled and query–document pairs from the judgment pool are independently labeled by assessors that are unaware of the originating retrieval system [23, 51]. Relevance judgments are subjective and, therefore, might vary among different assessors [40]. Low inter-assessor agreement can indicate low reliability [51]. To obtain high agreement among assessors, topics usually have a narrative describing what separates relevance from non-relevance and assessors undergo extensive training [28].

Once relevance judgments have been collected, TREC-style test collections are re-used with the assumption that the judgment pool is “essentially complete” [66], i.e., unjudged documents retrieved by a new system are considered not relevant. In cases where this assumption is true (e.g., recent findings on TREC-8 [68]), the system rankings do not change substantially depending on whether the systems contributed to the judgments or not [66]. However, especially if corpora are large or diverse or manual submissions cannot be pooled (e.g., during the time–constrained construction of TREC-COVID [48, 60]), this assumption might not hold. These limitations motivate dedicated evaluation procedures using either specialized measures [6, 42, 65, 74, 75] or predicted relevance labels [2, 3, 7, 8, 37]. Importantly, corpus subsampling strategies might limit which of those procedures can be applied, e.g., for subsamples that resemble re-ranking only (parts of) the judgment pool [57, 58, 34], none of these specialized methods can be applied as no retrievable unjudged documents remain in the corpus.

Overall, the high effort to construct TREC-style test collections and make them reliable, usually involving work by multiple expert assessors for several weeks [54], motivates us to study how the underlying corpora can be subsampled to allow reliable evaluation of expensive modern neural retrieval models, while still ensuring that the subsampled corpora allow evaluation setups that account for the uncertainties of post-hoc experimentation due to unjudged documents.

Corpus Subsampling for Validation. Validation queries are used to stop neural model training when the model converges before over–fitting [77]. As the validation step compares model checkpoints, we include the subsampling strategies employed there in our experiments. During the test phase, dense retrieval models create document embeddings for the corpus for efficient retrieval [76]. Encoding the corpus would be an unrealistic expense for validation [77], especially as often only one relevant document is available per query [76]. Different types of subsampling make validation more efficient [77], either by re-instantiating the training objective on the validation data [31, 72] or by re-ranking candidate documents [47, 29]. While the training objective permits efficient validation [72, 31], it can not be applied during test, as this would require knowledge of document

relevance. Re-ranking documents retrieved by a first-stage retriever [47], like BM25, makes the validation more realistic, as the documents in the comparison are harder to distinguish from the relevant document. Still, this re-ranking can hide problems in neural retrieval models (like missing length normalization [59]), but as this strategy can be realistically used during test, we include it in our experiments. In sum, training and validation often differ greatly from the test phase. Consequently, subsampling strategies that work well during validation may not be reliable when used to evaluate using TREC-style test collections.

Corpus Subsampling for Evaluation. The general rule of thumb that “software is getting slower more rapidly than hardware becomes faster” [70] can also be observed for research-oriented IR systems. Approaches that either store the corpus within the weights of transformers [58], embed document corpora with large language models [36], or expand documents with transformers [46], can rarely be evaluated on large corpora. Enriching MS MARCO v2 with DocT5Query expansions took 2 500 hours and cost 6 000 USD [35], highlighting the importance of sharing such resources, but also showing that it is not feasible to apply it to substantially larger corpora like the ClueWeb. Consequently, there have been experiments on subsampled corpora [57, 58, 34] to evaluate computationally expensive neural models using relevance judgments from large TREC-style collections. LOFT’s subsampling strategy [34] includes all documents that are judged as relevant for a given query and subsequently adds random documents. The subsampling strategy by Tay et al. [58], recently re-used by Lee et al. [34], includes the complete judgment pool and subsequently adds random documents. While both sampling strategies allow to evaluate computationally expensive retrieval approaches, it remains unclear whether the resulting subsamples provide reliable evaluation. Given that both scenarios greatly increase the proportion of relevant documents, they might overestimate retrieval effectiveness or favor some techniques over others, similar to observations on condensed lists [25, 50]. Therefore, we include this family of sampling strategies into our experiments; they also form the basis of our own sampling approach. Rather than sampling random documents, we sample from runs that contributed to the judgment pool.

Green and Efficiency-Oriented IR. Both efficiency and utilization play equal yet separate roles during development and operation of IR systems. Concerns of excessive utilization by large language models were first raised by Strubell et al. [55] by measuring energy and CO₂e usage. From there, Scells et al. [53] and Zuccon et al. [78] investigated energy, water, and CO₂e usage in IR systems. Despite several studies having investigated energy usage of search systems in CPU-bound environments [5, 9–12], few studies have focused on search systems in GPU-based environments. The popularity of leaderboard-driven effectiveness-focused benchmarks may be one reason for this [52]. Despite several studies noting the importance of utilization [24, 27, 32, 45, 56], there is still no generally accepted measurement [26]. However, we are the first to focus on tackling the problem of utilization from a data perspective. Despite the rich history of efficiency IR research [71, 61, 20], we found no studies investigating the evaluation efficiency.

3 Post-Hoc Subsampling From Pooled Retrieval Runs

In TREC-style evaluation campaigns, organizers publish a document corpus D and a set of topics T . Research groups then submit retrieval runs R that retrieve documents $d \in D$ for each topic $t \in T$. A subset of query—document pairs $J \subset T \times D$ form the judgment pool (typically, the top-10 results of each system) that is assessed by experts to form relevance labels $rel(t, d)$ with $(t, d) \in J$.

In this section, we describe prior subsampling strategies and introduce pooling as a new strategy to create reliable subcorpora. As most corpora are already a sample (e.g., a web crawl), we refer to this “second” round of deriving a subcorpus as subsampling that can be applied either before or after relevance assessment.

3.1 Subsampling Before the Relevance Assessments

Cascading re-ranking pipelines can subsample large corpora without needing the relevance assessments of topics. A cascading re-ranking pipeline starts with an initial efficient retrieval model to reduce the corpus to a few hundred or thousands of documents. Progressively more expensive retrieval models, e.g., neural retrieval models, are applied to re-rank the documents [38]. Formally, a retrieval model retrieves the top results for each topic $t \in T$ from the document corpus D . The subsampled corpus D' is the union of all retrieval results.

In practice, BM25 is predominantly used as the first stage retrieval system in such scenarios (e.g., the TREC Deep Learning tracks provided an official BM25 subsample [22, 21]). Subsampling with BM25 has the advantage that it is implemented in a wide range of research retrieval systems [39, 73], can be used efficiently, and has a natural order to adjust the size of the subsample. However, as the judgments and submitted runs are not used, the resulting subcorpus might miss difficult relevant documents. All documents in the subcorpus must have a lexical overlap with the query. The subcorpus thereby imposes a view that semantic overlap without lexical overlap is not existent (especially manual runs in TREC-campaigns aim to identify those documents) favoring a certain type of systems. Both factors might negatively impact reliability of BM25 subcorpora.

3.2 Subsampling After the Relevance Assessments

The judgment pool J and the submitted runs R provide valuable resources for reliable subsampling. We first describe existing subsampling strategies that only incorporate the judgment pool J and then describe our pooling approach that additionally uses the submitted runs R to include more realistic hard negatives.

LOFT Subsampling. The LOFT subsampling approach [34] was used to create small subcorpora that can fit into the input of long context large language models with a context size of up to one million tokens. LOFT subsampling takes all documents that are relevant for some topic and adds random documents from the corpus until a desired size is reached. Formally, if k is the desired size of the subcorpus, $D_{rel} = \{d \mid (t, d) \in J \wedge rel(t, d) > 0\}$ are all relevant documents,

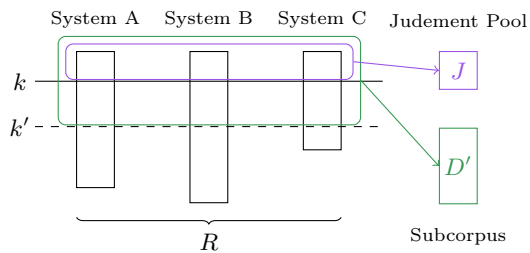


Fig. 1: Corpus subsampling via run pooling. k is the depth of the judgement pool J on pooled runs R . The subcorpus D' has depth k' .

Algorithm 1 LOFT Sampling

Input: T set of topics
 J relevance judgments
 k target corpus size
Output: set of documents C

```

1:  $C \leftarrow \emptyset, i \leftarrow 0$ 
2:  $m \leftarrow \min(\{\forall j \in D_{rel}^T : |j|\})$ 
3: repeat  $i \leftarrow i + 1$ 
4:   if  $i < m$  then
5:     for all topics  $t \in T$  do
6:        $C \leftarrow d_i \in D_{rel}^T$ 
7:   else
8:      $C \leftarrow d \in \{D \setminus D_{rel}\}$ 
9: until  $C = k$ 

```

and $D_{rand} \subset D \setminus D_{rel}$ are random documents with $|D_{rand}| = k - |D_{rel}|$, LOFT sampling yields $D' = D_{rel} \cup D_{rand}$ as a subcorpus. Given that the document length can differ among corpora, the LOFT benchmark has an average k of 885 documents for a subcorpus (minimum: 328; maximum: 3306) to fit documents of different sizes into the fixed size context. We instantiate LOFT sampling twice, with $k = 1000$ and $k = 10000$. As some of our test collections have more than 1000 relevant documents, we include the relevant documents round-robin per query, ensuring that each query has at most a difference of one relevant document. Algorithm 1 details how we use LOFT for corpus subsampling.

A potential problem of LOFT subsampling is that it may be too easy to distinguish the relevant documents from random documents. The subsampling strategy by Tay et al. [58] reduces this effect by taking the complete judgment pool $D_J = \{d \mid (t, d) \in J\}$ as corpus that is then expanded with random documents until the target size. However, as the judgment pools are usually rather small, random documents make the majority of the subcorpus, thereby the tasks still remains substantially more easy than more realistic subsampling strategies.

Corpus Subsampling via Pooling. In our corpus sampling strategy, we perform a second pooling round after the initial pooling for relevance judgments but at a higher depth. This allows us to incorporate a diverse set of hard-negative documents. Previous subsampling approaches have not used the submitted retrieval runs R . We argue that they form an excellent source to construct reliable subcorpora as the submitted runs were produced as a community effort from independent retrieval pipelines and should, therefore, cover diverse notions of non-relevant documents, making the subcorpus more realistic and also more accurately biased towards the topics than random sampling. As we pool to a higher depth than the judgment pool, all judged documents are included, and the subcorpus size can be varied naturally via the depth of the second pooling.

Figure 1 shows corpus subsampling via pooling. The input is a set of topics T and the set of pooled runs R . We assume the judgment pool J is an top- k pool (e.g., $k=10$) of $r \in R$ with the subcorpus pooling size k' being substantially larger than k . Let $r_{k'} \subset D$ denote the set of documents within the top- k' results of $r \in R$. Then, pooling subsampling yields $D' = \bigcup_{r \in R} r_{k'}$ as subsample.

Table 2: Overview of the subsamples (documents and bytes) for all datasets.

Sampling	ClueWeb09		ClueWeb12		MS MARCO		Robust04	
	Docs.	Size	Docs.	Size	Docs.	Size	Docs.	Size
BM25	48091	134.3 MB	48875	241.2 MB	46392	5.9 MB	165165	209.2 MB
LOFT _{1k}	1000	3.9 MB	1000	5.9 MB	1000	0.1 MB	1000	3.3 MB
LOFT _{10k}	10000	28.9 MB	10000	45.8 MB	10000	1.4 MB	10000	22.2 MB
Pool _J	9303	26.5 MB	6971	33.1 MB	15864	2.2 MB	27239	59.4 MB
Pool ₂₅	21469	60.2 MB	15416	71.0 MB	38244	5.3 MB	57579	110.2 MB
Pool ₅₀	40520	111.9 MB	28032	126.5 MB	74388	10.2 MB	98653	170.9 MB
Pool ₁₀₀	77013	211.2 MB	52477	232.0 MB	143083	19.6 MB	160825	253.8 MB
Pool ₁₀₀₀	629140	1.7 GB	435859	1.9 GB	983765	135.9 MB	449371	529.5 MB

4 Evaluation

We compare all subsampling strategies in theoretical and empirical experiments on four corpora with 9 TREC tracks using nDCG@10 as evaluation measure. For theoretical validation, we use simulated experiments on all runs submitted to the 9 tracks. For empirical validation, we run 20 diverse retrieval systems on all subsampled corpora while contrasting energy usage with reliability.

Subsampling Strategies in our Experiments. We experiment with 9 subsampling strategies: (1) the top-1000 BM25 results, (2) two LOFT subsamples at $k=1\,000$ and $k=10\,000$, and five pooling variants ($k \in \{10 = |J|, 25, 50, 100, 1000\}$). Where applicable, we use the full-corpus (i.e., not removing any documents) as baseline.

Corpora and Tracks. We use ClueWeb09, ClueWeb12, MS MARCO passage v1, and Disks 4/5 (between 0.5 m and 1 b documents; cf. Table 1). We use all runs submitted to the 2009–2014 Web tracks (30 to 71 runs), the Robust04 track (110 runs), and the MS MARCO v1 Deep Learning tracks (37 to 59 runs). Following previous setups that use TREC runs [4], we sort all runs by their nDCG@10 score and remove the least effective 25 % to mitigate the effect of potentially erroneous runs. MS MARCO comes with passage-level relevance judgments. Still, document level relevance judgments are frequently used for neural evaluations (e.g., BEIR has pre-neural datasets with document-level relevance judgments like Robust04 and Touché, truncating documents by default [60]).

Subsampled Corpora. Table 2 overviews the subcorpora sizes (averaged across the tracks). Even the largest subcorpus of pooling to 1 000 reduces the corpus size by more than 1 500 for large corpora like the ClueWeb09, and still by 8.9 for smaller corpora like MS MARCO (61 times for a top-100 pool), making it feasible to share and re-use subcorpora. Furthermore, Table 2 already hints how reliable subsampling strategies might be: the size of the judgment pool (Pool_J) shows how many documents are retrieved at top positions, allowing to calculate

the probability that a random document makes it into the top results, which is negligible (e.g., $\frac{9303}{1e+09} = 9.3e - 06$ for the ClueWeb09). Therefore, random sampling might be unlikely to include documents that are retrieved into the top positions that are important for evaluation, contrary to our corpus pooling.

4.1 Theoretical Validation of Corpus Subsampling

We validate the subsampling strategies using simulated incompleteness in three experiments. First, we analyze how nDCG@10 scores differ. Second, we analyze the impact on system rankings when a team that did not participate is compared with systems that participated in judgment and subsampling. Third, we analyze the impact on system rankings when multiple teams did not participate.

Simulated Incompleteness. We group all runs by their submitting team for leave-one-group-out experiments to simulate that a team did not participate in a track [63]. If the corpus is not changed, leave-one-group-out experiments impact only the judgment pool. However, as we subsample the corpus, leave-one-group-out experiments additionally impact the subsampled corpora, as subsampling strategies can not include documents solely retrieved by a left-out group. We simulate judgment incompleteness for top-10 judgment pools, i.e., removing documents from the judgment pool solely contributed by the left-out group (i.e., does not occur in the top-10 results of any system of other groups). All subsampling strategies only have access to the modified judgment pool and the remaining contributing runs. This yields one incomplete subsampled corpus and judgment pool per group, where runs of other groups fully contributed to the subsampled corpus and remain fully judged. In all cases, we use the nDCG@10 on the complete corpus on all judgments as ground-truth. For systems that retrieve on the subsampled corpora, we assume that subsampling does not affect the order of documents (we validate this in Section 4.2), and thereby remove documents that are not in a subsample from a run to simulate retrieval from the subcorpus.

The Impact of Corpus Subsampling on nDCG@10. We compare the nDCG@10 scores of left-out systems with their ground-truth to show the behavior of the subsampling strategies. Two factors contribute to wrong nDCG@10 scores: (1) unjudged documents and (2) missing documents. A document is unjudged if retrieved only by the left-out group, potentially underestimating nDCG@10 [25, 50]. A document is missing if a left-out system would retrieve this document from the full corpus that is not in the subsample. Without post-judgments under the default assumption that unjudged documents are not relevant [66], missing documents would be unjudged and assumed not relevant, thereby overestimating nDCG@10 (a frequent scenario as post-hoc experiments rarely add judgments).

To show the impact of unjudged and missing documents independently, in Table 3 we report the root mean squared error without post-judgments (RMSE), with post-judgments (RMSE_{Judged}), and the average difference of the subsampled versus the actual nDCG@10 score to show which subsampling strategies

Table 3: Leave-one-group-out differences in nDCG@10 on ClueWeb09 (C09), ClueWeb12 (C12), MS MARCO (MSM), and Robust04 (R04). We report the RMSE without and with post-judgments (RMSE vs. RMSE_{Judged}), and the difference of subsampled vs. correct nDCG@10 ($\Delta_{\text{nDCG@10}} \pm \text{std dev.}$).

* indicates Bonferroni corrected significant differences to the full corpus.

Sampl.	RMSE				RMSE _{Judged}				$\Delta_{\text{nDCG@10}}$			
	C09	C12	MSM	R04	C09	C12	MSM	R04	C09	C12	MSM	R04
BM25	.057	.084	.096*	.007	.083*	.052*	.110*	.010	-.013 \pm .06	-.053 \pm .07	.049* \pm .08	-.005 \pm .00
LOFT _{1k}	.372*	.333*	.529*	.042*	.372*	.333*	.529*	.047*	.365* \pm .08	.311* \pm .12	.528* \pm .04	.015* \pm .04
LOFT _{10k}	.381*	.342*	.110*	.263*	.386*	.349*	.110*	.271*	.375* \pm .07	.325* \pm .11	.062* \pm .09	.259* \pm .05
Pool _J	.041	.039*	.009	.014*	.042*	.040*	.009*	.018*	.030 \pm .03	.031* \pm .02	.005 \pm .01	.011* \pm .01
Pool ₂₅	.025*	.040*	.006	.005*	.037*	.029*	.007	.011*	-.011*	-.028*	-.002\pm.01	-.002\pm.00
Pool ₅₀	.032*	.060	.007	.006*	.034*	.019*	.006	.008	-.023* \pm .02	-.050 \pm .03	-.004 \pm .01	-.005* \pm .00
Pool ₁₀₀	.039	.071	.006	.008	.030*	.015*	.006	.007	-.030 \pm .02	-.060 \pm .04	-.004 \pm .00	-.007 \pm .00
Pool ₁₀₀₀	.047	.082	.006	.009	.023	.008*	.005	.007	-.039 \pm .03	-.070 \pm .04	-.005 \pm .00	-.007 \pm .00
Full	.050	.085	.006	.009	.000	.000	.000	.000	-.041 \pm .03	-.072 \pm .04	-.005 \pm .00	-.007 \pm .00

overestimate (positive $\Delta_{\text{nDCG@10}}$) or underestimate (negative $\Delta_{\text{nDCG@10}}$). Evaluations on the full corpus underestimate nDCG@10 scores the most, by 0.005 for MS MARCO up to 0.072 for the ClueWeb12. With post-judgments, nDCG@10 scores are correct (RMSE_{Judged} = 0.0). LOFT sampling substantially overestimates nDCG@10 scores, in the worst case by 0.528 for MS MARCO. Using only the judgment pool (Pool_J) also overestimates the effectiveness, but much less than LOFT as more realistic non-relevant documents are included. BM25 may over- or underestimate the effectiveness, having a much higher standard deviation than pooling. Pooling to a higher depth than the original judgment pool produces the most reliable subcorpora, as nDCG@10 scores are underestimated, but to a lesser extent as processing the complete corpus as fewer unjudged documents are included but still enough to not overestimate. Without post-judgments, pooling to 25 produces reliable scores, whereas the depth-100 and depth-1000 scores are statistically indistinguishable from the full corpus. With post-judgments, pooling beyond 100 only yields negligible improvements, making a top-100 pool a good choice for all scenarios.

Corpus Subsampling for Comparisons Against Participating Systems. We study the scenario where systems of the left-out group are compared with systems that contributed to the judgments and the subsampling. All contributing systems are correctly evaluated, whereas only the left-out systems could introduce changes in the system rankings. To compare system rankings we use τ_{AP} variant of Kendall’s correlation with handling of ties [62] (1: perfect, 0: random, -1: perfect inverse correlation), as τ_{AP} gives more importance to top-ranked systems. We report changes in the system rankings in three scenarios, without post-judgments (τ), with condensed lists that remove all unjudged documents (τ_C), and with post-judgments (τ_{PJ}) in Table 4. Across all scenarios, LOFT yields the most changes in system rankings, often significantly different to the full corpus.

Table 4: Overview of the reliability of system rankings measured as τ_{AP} without post-judgments (τ), with condensed evaluation (τ_C), and with post judgments (τ_{PJ}) for comparing systems of a non-participating group with participating systems. *: Bonferroni corrected significant differences to the full corpus.

Sampl.	ClueWeb09			ClueWeb12			MS MARCO			Robust04		
	τ	τ_C	τ_{PJ}	τ	τ_C	τ_{PJ}	τ	τ_C	τ_{PJ}	τ	τ_C	τ_{PJ}
BM25	.919	.807	.936*	.894	.922	.938	.847*	.827*	.836*	.980	.945	.994
LOFT _{1k}	.793	.690*	.793*	.775	.775*	.775*	.776*	.774*	.776*	.940	.904	.940*
LOFT _{10k}	.799	.693*	.799*	.763	.762*	.765*	.790*	.786*	.789*	.794*	.764*	.794*
Pool _J	.944	.832	.944*	.941	.939	.941*	.983	.980	.983	.978	.948	.978*
Pool ₂₅	.956	.832	.967*	.917	.939	.959	.978	.980	.992	.987	.948	.993*
Pool ₅₀	.940	.832	.974	.902	.939	.976	.976	.980	.993	.981	.948	.997*
Pool ₁₀₀	.934	.832	.980	.898	.939	.987	.974	.980	.995	.979	.948	.999
Pool ₁₀₀₀	.931	.832	.994	.896	.939	.997	.974	.980	.996	.978	.948	1.00
Full	.930	.832	1.00	.895	.939	1.00	.972	.980	1.00	.978	.948	1.00

BM25 subsamples are more reliable, but pooling yields even better results in all cases, statistically not distinguishable from evaluating on the complete corpus for top-100 and top-1000 pools. Still, comparisons against participating systems are rarely used, as research often compares systems across multiple corpora, and the same system rarely is pooled on different corpora. Hence, we next study scenarios where multiple non-participating systems are compared.

Corpus Subsampling for Comparisons of Non-Participating Systems. To simulate a setting where systems of different groups that did not contribute to the judgments and subsampling are compared, we contrast the system ranking obtained from only leave-one-group-out nDCG@10 scores with the ground-truth system ranking in Table 5 (no significance tests are possible in this scenario, as no averages are reported). Every system can have unjudged and missing documents. Consequently, system ranking correlations for the full corpus without post judgments (τ) are rather low for large corpora with few pooled runs (e.g., $\tau = 0.302$ for the ClueWeb12 with 30 runs) whereas reliable for corpora with many pooled runs (e.g., $\tau = 0.884$ for Robust04 with 110 runs). However, with post-judgments (τ_{PJ}), the full corpora produce perfect correlations, whereas the top-100 and top-1000 pooled subcorpora also achieve high ranking correlations; in case of condensed lists (τ_C) even indistinguishable from the full corpora. Consistent with the previous experiments, BM25 subsampling is better than LOFT but less reliable than pooling (especially for the condensed and the post-judgment setting). Altogether, our three experiments show that pooling to a depth of 100 or 1000 is often indistinguishable from processing the full corpora, at only a tiny fraction of the compute.

Table 5: The reliability of system rankings as τ_{AP} without post-judgments (τ), with condensed evaluation (τ_C), and with post judgments (τ_{PJ}) when comparing systems that all did not participate in the judgment and subsampling.





















Sampling	ClueWeb09			ClueWeb12			MS MARCO			Robust04		
	τ	τ_C	τ_{PJ}	τ	τ_C	τ_{PJ}	τ	τ_C	τ_{PJ}	τ	τ_C	τ_{PJ}
BM25	.265	.298	.435	.211	.541	.542	.601	.593	.627	.888	.859	.949
LOFT _{1k}	.278	.281	.278	.329	.329	.329	.294	.294	.294	.416	.392	.416
LOFT _{10k}	.373	.360	.372	.366	.337	.373	.496	.497	.497	.312	.259	.305
Pool _J	.611	.579	.611	.699	.699	.699	.872	.872	.872	.909	.899	.909
Pool ₂₅	.706	.583	.734	.560	.699	.732	.854	.872	.926	.940	.899	.951
Pool ₅₀	.658	.584	.781	.470	.699	.816	.841	.872	.940	.912	.899	.971
Pool ₁₀₀	.618	.585	.827	.376	.699	.889	.835	.872	.952	.888	.899	.991
Pool ₁₀₀₀	.574	.586	.946	.337	.699	.965	.830	.872	.957	.884	.899	1.00
Full	.565	.586	1.00	.302	.699	1.00	.869	.869	1.00	.884	.899	1.00

4.2 Empirical Validation of Corpus Subsampling for Retrieval

While our previous experiments assumed that corpus subsampling does not change the relative order of retrieval results to validate theoretically the effect of incompleteness and subsampling, in practice, removing documents from the collection changes the corpus statistics and thereby potentially the ranking. Therefore, we evaluate how rankings of 20 diverse first-stage retrieval systems change over all subcorpora monitoring their energy consumption to identify which subsampling yield reliable retrieval results at acceptable energy consumption.

Retrieval Systems. We include retrieval systems of three popular paradigms: 10 lexical retrieval models in PyTerrier [39], 7 bi-encoder models in BEIR [60], and 3 late interaction models in PLAID-X [43].⁹ All systems operate in their default configuration. We do not include re-rankers like monoT5 or duoT5 [46] as they solely depend on the first-stage, not on global aspects like corpus statistics.

Experimental Setup. We run all retrieval systems on all subcorpora on the same machine (four A100 GPUs with 96 CPU cores) sequentially measuring their energy consumption with codecarbon [19]. To make the experiments comparable across the diverse corpora, we use the retrieval results of a system for the top-1000 pool as the ground-truth ranking and report the similarities for all smaller subsamples measured as RBO [69] as this allows to compare rankings of different documents. The energy consumption includes indexing and retrieval, i.e., the footprint of the complete experiment.

⁹ Bi-Encoder: ANCE , DistilBERT , MiniLM-L6 , MiniLM-L12 , TAS-B , MultiQA (DistilBERT) , MultiQA (MpNet) ; Late Interaction: ColBERT v1.9 , ColBERT v2.0 , PLAID-X English Large ; Lexical: BM25 , DirichletLM , DFIZ , DLH , DPH , Hiemstra LM , IFB2 , InB2 , PL2 , TF IDF .

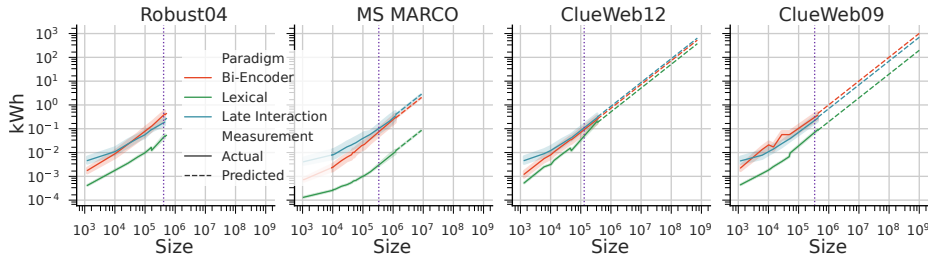


Fig. 2: Energy consumption in kWh for all paradigms and subsamples interpolated to the full corpora. The vertical line indicates the recommended subsample.

Table 6: RBO similarity of Bi-Encoder, Late-Interaction, and Lexical models to themselves when retrieving from a subsample versus the largest Pool₁₀₀₀ subcorpus.

Sampling	ClueWeb09			ClueWeb12			MS MARCO			Robust04		
	Bi-E.	Late	Lex.	Bi-E.	Late	Lex.	Bi-E.	Late	Lex.	Bi-E.	Late	Lex.
BM25	.139	.192	.037	.146	.159	.0837	.89	.749	.858	.921	.758	.963
LOFT _{1k}	.070	.090	.040	.115	.103	.176	.68	.634	.513	.113	.13	.138
LOFT _{10k}	.096	.111	.056	.185	.167	.308	.777	.693	.601	.431	.459	.495
Pool _J	.297	.263	.295	.289	.256	.606	.908	.756	.880	.712	.682	.885
Pool ₂₅	.405	.333	.449	.407	.358	.686	.954	.761	.901	.824	.726	.910
Pool ₅₀	.495	.394	.550	.506	.44	.752	.973	.781	.914	.888	.755	.927
Pool ₁₀₀	.600	.481	.660	.613	.524	.792	.981	.787	.933	.936	.759	.946

Results. Figure 2 shows the energy consumption in kWh for all approaches across the different corpus subsamples. We extrapolate the energy consumption beyond the top-1000 pools to the full corpora and indicate the consumption for our recommended top-100 and top-1000 pooled subsamples. For large corpora, our subsamples reduce the energy consumption by more than 1000, and even for small corpora, energy savings of around a factor of 10 are reached.

Table 6 shows how similar the rankings on the subsampled corpora are towards the top-1000 pool for the three retrieval paradigms. Consistently, the LOFT and BM25 subcorpora do not yield similar rankings, whereas a top-25 pool already shows substantial improvement compared to just re-ranking the judgment pool. Still, using the judgment pool shows reasonable ranking similarities for small corpora, as there the judgment pools cover a substantial part of the complete collection, whereas huge corpora such as the ClueWeb09 and ClueWeb12 show a very low ranking similarity for judged documents only (e.g., RBO of 0.295 for lexical models on ClueWeb09 vs. 0.880 on MS MARCO). The top-100 pool yields the most similar rankings. Overall, our experiments were consistent across all tested scenarios, providing evidence that pooling allows the construction of reliable subcorpora, whereas top-100 pools can be recommended as they already perform as reliably as top-1000 pools.

5 Conclusion and Future Work

We introduced pooling as an approach to produce subcorpora that allows reliable evaluation of expensive retrieval approaches on large corpora. We reduced the computational requirements by up to 1000 times while providing statistically indistinguishable evaluations from processing the complete corpus. The resulting subcorpora allow exploration of many expensive re-ranking cascades more systematically. Even corpora that are already small could be substantially reduced, yielding much faster and more Green experiments. An interesting direction for future work could be to study how corpus subsampling can be incorporated up-front into the design of evaluation campaigns. Retrieval from huge, noisy corpora is difficult; thereby, running a campaign on such huge, ClueWeb-style corpora might be able to construct a realistic and diverse set of hard negatives, while subsequent corpus subsampling can ensure that the post-hoc experiments with the fixed set of queries only need a small computational budget.

Acknowledgements This publication has received funding from the European Union’s Horizon Europe research and innovation programme under grant agreement No 101070014 (OpenWebSearch.EU, <https://doi.org/10.3030/101070014>).

References

- [1] Allan, J., Aslam, J.A., Pavlu, V., Kanoulas, E., Carterette, B.: Million Query track 2008 overview. In: Voorhees, E.M., Buckland, L.P. (eds.) Proceedings of The Seventeenth Text REtrieval Conference, TREC 2008, Gaithersburg, Maryland, USA, November 18–21, 2008. NIST Special Publication, vol. 500-277. National Institute of Standards and Technology (NIST) (2008)
- [2] Aslam, J.A., Pavlu, V., Yilmaz, E.: A statistical method for system evaluation using incomplete judgments. In: Efthimiadis, E.N., Dumais, S.T., Hawking, D., Järvelin, K. (eds.) SIGIR 2006: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA, August 6–11, 2006. pp. 541–548. ACM (2006)
- [3] Aslam, J.A., Yilmaz, E.: Inferring document relevance from incomplete information. In: Silva, M.J., Laender, A.H.F., Baeza-Yates, R.A., McGuinness, D.L., Olstad, B., Olsen, Ø.H., Falcão, A.O. (eds.) Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, CIKM 2007, Lisbon, Portugal, November 6–10, 2007. pp. 633–642. ACM (2007)
- [4] Bernstein, Y., Zobel, J.: Redundant documents and search effectiveness. In: Herzog, O., Schek, H., Fuhr, N., Chowdhury, A., Teiken, W. (eds.) Proceedings of the 2005 ACM CIKM International Conference on Information and Knowledge Management, Bremen, Germany, October 31 – November 5, 2005. pp. 736–743. ACM (2005)

- [5] Blanco, R., Catena, M., Tonello, N.: Exploiting green energy to reduce the operational costs of multi-center web search engines. In: Bourdeau, J., Hendler, J., Nkambou, R., Horrocks, I., Zhao, B.Y. (eds.) Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11–15, 2016. pp. 1237–1247. ACM (2016)
- [6] Buckley, C., Voorhees, E.M.: Retrieval evaluation with incomplete information. In: Sanderson, M., Järvelin, K., Allan, J., Bruza, P. (eds.) SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, July 25–29, 2004. pp. 25–32. ACM (2004)
- [7] Büttcher, S., Clarke, C.L.A., Yeung, P.C.K., Soboroff, I.: Reliable information retrieval evaluation with incomplete and biased judgements. In: Kraaij, W., de Vries, A.P., Clarke, C.L.A., Fuhr, N., Kando, N. (eds.) SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23–27, 2007. pp. 63–70. ACM (2007)
- [8] Carterette, B., Jones, R.: Evaluating search engines by modeling the relationship between relevance and clicks. In: Platt, J.C., Koller, D., Singer, Y., Roweis, S.T. (eds.) Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3–6, 2007. pp. 217–224. Curran Associates, Inc. (2007)
- [9] Catena, M.: Energy efficiency in web search engines. In: Azzopardi, L., Wilson, M.L. (eds.) Sixth BCS-IRSG Symposium on Future Directions in Information Access, FDIA 2015, 31 August – 4 September 2015, Thessaloniki, Greece. Workshops in Computing, BCS (2015)
- [10] Catena, M., Frieder, O., Tonello, N.: Efficient energy management in distributed web search. In: Cuzzocrea, A., Allan, J., Paton, N.W., Srivastava, D., Agrawal, R., Broder, A.Z., Zaki, M.J., Candan, K.S., Labrinidis, A., Schuster, A., Wang, H. (eds.) Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22–26, 2018. pp. 1555–1558. ACM (2018)
- [11] Catena, M., Macdonald, C., Tonello, N.: Load-sensitive CPU power management for web search engines. In: Baeza-Yates, R., Lalmas, M., Moffat, A., Ribeiro-Neto, B.A. (eds.) Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9–13, 2015. pp. 751–754. ACM (2015)
- [12] Catena, M., Tonello, N.: A study on query energy consumption in web search engines. In: Boldi, P., Perego, R., Sebastiani, F. (eds.) Proceedings of the 6th Italian Information Retrieval Workshop, Cagliari, Italy, May 25–26, 2015. CEUR Workshop Proceedings, vol. 1404. CEUR-WS.org (2015)
- [13] Clarke, C.L.A., Craswell, N., Soboroff, I.: Overview of the TREC 2009 Web track. In: Voorhees, E.M., Buckland, L.P. (eds.) Proceedings of The Eighteenth Text REtrieval Conference, TREC 2009, Gaithersburg,

- Maryland, USA, November 17–20, 2009. NIST Special Publication, vol. 500-278. National Institute of Standards and Technology (NIST) (2009)
- [14] Clarke, C.L.A., Craswell, N., Soboroff, I., Cormack, G.V.: Overview of the TREC 2010 Web track. In: Voorhees, E.M., Buckland, L.P. (eds.) Proceedings of The Nineteenth Text REtrieval Conference, TREC 2010, Gaithersburg, Maryland, USA, November 16–19, 2010. NIST Special Publication, vol. 500-294. National Institute of Standards and Technology (NIST) (2010)
- [15] Clarke, C.L.A., Craswell, N., Soboroff, I., Voorhees, E.M.: Overview of the TREC 2011 Web track. In: Voorhees, E.M., Buckland, L.P. (eds.) Proceedings of The Twentieth Text REtrieval Conference, TREC 2011, Gaithersburg, Maryland, USA, November 15–18, 2011. NIST Special Publication, vol. 500-296. National Institute of Standards and Technology (NIST) (2011)
- [16] Clarke, C.L.A., Craswell, N., Voorhees, E.M.: Overview of the TREC 2012 Web track. In: Voorhees, E.M., Buckland, L.P. (eds.) Proceedings of The Twenty-First Text REtrieval Conference, TREC 2012, Gaithersburg, Maryland, USA, November 6–9, 2012. NIST Special Publication, vol. 500-298. National Institute of Standards and Technology (NIST) (2012)
- [17] Collins-Thompson, K., Bennett, P.N., Diaz, F., Clarke, C., Voorhees, E.M.: TREC 2013 Web track overview. In: Voorhees, E.M. (ed.) Proceedings of The Twenty-Second Text REtrieval Conference, TREC 2013, Gaithersburg, Maryland, USA, November 19–22, 2013. NIST Special Publication, vol. 500-302. National Institute of Standards and Technology (NIST) (2013)
- [18] Collins-Thompson, K., Macdonald, C., Bennett, P.N., Diaz, F., Voorhees, E.M.: TREC 2014 Web track overview. In: Voorhees, E.M., Ellis, A. (eds.) Proceedings of The Twenty-Third Text REtrieval Conference, TREC 2014, Gaithersburg, Maryland, USA, November 19–21, 2014. NIST Special Publication, vol. 500-308. National Institute of Standards and Technology (NIST) (2014)
- [19] Courty, B.: mlco2/codecarbon: v2.4.1 (May 2024), <https://doi.org/10.5281/zenodo.11171501>
- [20] Crane, M., Culpepper, J.S., Lin, J., Mackenzie, J.M., Trotman, A.: A comparison of document-at-a-time and score-at-a-time query evaluation. In: de Rijke, M., Shokouhi, M., Tomkins, A., Zhang, M. (eds.) Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, WSDM 2017, Cambridge, United Kingdom, February 6–10, 2017. pp. 201–210. ACM (2017)
- [21] Craswell, N., Mitra, B., Yilmaz, E., Campos, D.: Overview of the TREC 2020 Deep Learning Track. In: Voorhees, E.M., Ellis, A. (eds.) Proceedings of the 29th Text REtrieval Conference, TREC 2020, Virtual Event, Gaithersburg, MD, USA, November 16–20, 2020. NIST Special Publication, vol. 1266. National Institute of Standards and Technology (NIST) (2020)

- [22] Craswell, N., Mitra, B., Yilmaz, E., Campos, D., Voorhees, E.M.: Overview of the TREC 2019 Deep Learning Track. In: Voorhees, E., Ellis, A. (eds.) 28th International Text Retrieval Conference, TREC 2019, Gaithersburg, Maryland, USA. NIST Special Publication, National Institute of Standards and Technology (NIST) (Nov 2019)
- [23] Ferro, N., Maistro, M.: Evaluation of IR systems. In: Alonso, O., Baeza-Yates, R. (eds.) Information Retrieval: Advanced Topics and Techniques, ACM Books, vol. 60, pp. 111–191. ACM (2024)
- [24] Frayling, E., MacAvaney, S., Macdonald, C., Ounis, I.: Effective adhoc retrieval through traversal of a query-document graph. In: Goharian, N., Tonellotto, N., He, Y., Lipani, A., McDonald, G., Macdonald, C., Ounis, I. (eds.) Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24–28, 2024, Proceedings, Part III. Lecture Notes in Computer Science, vol. 14610, pp. 89–104. Springer (2024)
- [25] Fröbe, M., Gienapp, L., Potthast, M., Hagen, M.: Bootstrapped nDCG estimation in the presence of unjudged documents. In: Kamps, J., Goeuriot, L., Crestani, F., Maistro, M., Joho, H., Davis, B., Gurrin, C., Kruschwitz, U., Caputo, A. (eds.) Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part I. Lecture Notes in Computer Science, vol. 13980, pp. 313–329. Springer (2023)
- [26] Fröbe, M., Mackenzie, J., Mitra, B., Nardini, F.M., Potthast, M.: ReNeuIR at SIGIR 2024: The third workshop on reaching efficiency in neural information retrieval. In: Yang, G.H., Wang, H., Han, S., Hauff, C., Zuccon, G., Zhang, Y. (eds.) Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14–18, 2024. pp. 3051–3054. ACM (2024)
- [27] Gienapp, L., Deckers, N., Potthast, M., Scells, H.: Learning effective representations for retrieval using self-distillation with adaptive relevance margins. arXiv 2407.21515 (2024)
- [28] Harman, D.: TREC-style evaluations. In: Agosti, M., Ferro, N., Forner, P., Müller, H., Santucci, G. (eds.) Information Retrieval Meets Information Visualization - PROMISE Winter School 2012, Zinal, Switzerland, January 23-27, 2012, Revised Tutorial Lectures. Lecture Notes in Computer Science, vol. 7757, pp. 97–115. Springer (2012)
- [29] Hofstätter, S., Lin, S., Yang, J., Lin, J., Hanbury, A.: Efficiently teaching an effective dense retriever with balanced topic aware sampling. In: Diaz, F., Shah, C., Suel, T., Castells, P., Jones, R., Sakai, T. (eds.) SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11–15, 2021. pp. 113–122. ACM (2021)
- [30] Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems* **20**(4), 422–446 (2002)

- [31] Karpukhin, V., Oguz, B., Min, S., Lewis, P.S.H., Wu, L., Edunov, S., Chen, D., Yih, W.: Dense passage retrieval for open-domain question answering. In: Webber, B., Cohn, T., He, Y., Liu, Y. (eds.) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16–20, 2020. pp. 6769–6781. ACL (2020)
- [32] Khandel, P., Yates, A., Varbanescu, A.L., de Rijke, M., Pimentel, A.D.: Distillation vs. sampling for efficient training of learning to rank models. In: Oosterhuis, H., Bast, H., Xiong, C. (eds.) Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2024, Washington, DC, USA, 13 July 2024. pp. 51–60. ACM (2024)
- [33] Lawrie, D.J., Kayi, E.S., Yang, E., Mayfield, J., Oard, D.W.: PLAID SHIRTTT for large-scale streaming dense retrieval. In: Yang, G.H., Wang, H., Han, S., Hauff, C., Zuccon, G., Zhang, Y. (eds.) Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14–18, 2024. pp. 2574–2578. ACM (2024)
- [34] Lee, J., Chen, A., Dai, Z., Dua, D., Sachan, D.S., Boratko, M., Luan, Y., Arnold, S.M.R., Perot, V., Dalmia, S., Hu, H., Lin, X., Pasupat, P., Amini, A., Cole, J.R., Riedel, S., Naim, I., Chang, M., Guu, K.: Can long-context language models subsume retrieval, RAG, SQL, and more? arXiv 2406.13121 (2024)
- [35] Ma, X., Pradeep, R., Nogueira, R.F., Lin, J.: Document expansion baselines and learned sparse lexical representations for MS MARCO V1 and V2. In: Amigó, E., Castells, P., Gonzalo, J., Carterette, B., Culpepper, J.S., Kazai, G. (eds.) SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022. pp. 3187–3197. ACM (2022)
- [36] Ma, X., Wang, L., Yang, N., Wei, F., Lin, J.: Fine-tuning LLaMA for multi-stage text retrieval. In: Yang, G.H., Wang, H., Han, S., Hauff, C., Zuccon, G., Zhang, Y. (eds.) Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14–18, 2024. pp. 2421–2425. ACM (2024)
- [37] MacAvaney, S., Soldaini, L.: One-shot labeling for automatic relevance estimation. In: Chen, H., Duh, W.E., Huang, H., Kato, M.P., Mothe, J., Poblete, B. (eds.) Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23–27, 2023. pp. 2230–2235. ACM (2023)
- [38] MacAvaney, S., Tonellotto, N., Macdonald, C.: Adaptive re-ranking with a corpus graph. In: Hasan, M.A., Xiong, L. (eds.) Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022. pp. 1491–1500. ACM (2022)
- [39] Macdonald, C., Tonellotto, N., MacAvaney, S., Ounis, I.: PyTerrier: Declarative experimentation in Python from BM25 to dense retrieval. In:

- Demartini, G., Zuccon, G., Culpepper, J.S., Huang, Z., Tong, H. (eds.) CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 – 5, 2021. pp. 4526–4533. ACM (2021)
- [40] Mizzaro, S.: Relevance: The whole history. *Journal of the American Society for Information Science* **48**(9), 810–832 (1997)
- [41] Moffat, A., Mackenzie, J.: How much freedom does an effectiveness metric really have? *Journal of the Association for Information Science and Technology* **75**(6), 686–703 (2024)
- [42] Moffat, A., Zobel, J.: Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems* **27**(1), 2:1–2:27 (2008)
- [43] Nair, S., Yang, E., Lawrie, D.J., Duh, K., McNamee, P., Murray, K., Mayfield, J., Oard, D.W.: Transfer learning approaches for building cross-language dense retrieval models. In: Hagen, M., Verberne, S., Macdonald, C., Seifert, C., Balog, K., Nørvåg, K., Setty, V. (eds.) *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part I. Lecture Notes in Computer Science*, vol. 13185, pp. 382–396. Springer (2022)
- [44] Overwijk, A., Xiong, C., Callan, J.: ClueWeb22: 10 billion web documents with rich information. In: Amigó, E., Castells, P., Gonzalo, J., Carterette, B., Culpepper, J.S., Kazai, G. (eds.) *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Madrid, Spain, July 11 - 15, 2022. pp. 3360–3362. ACM (2022)
- [45] Pasin, A., Cunha, W., Gonçalves, M.A., Ferro, N.: A quantum annealing instance selection approach for efficient and effective transformer fine-tuning. In: Oosterhuis, H., Bast, H., Xiong, C. (eds.) *Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2024, Washington, DC, USA, 13 July 2024*. pp. 205–214. ACM (2024)
- [46] Pradeep, R., Nogueira, R.F., Lin, J.: The Expando-Mono-Duo design pattern for text ranking with pretrained sequence-to-sequence models. *arXiv 2101.05667* (2021)
- [47] Qu, Y., Ding, Y., Liu, J., Liu, K., Ren, R., Zhao, W.X., Dong, D., Wu, H., Wang, H.: RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. In: Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tür, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., Zhou, Y. (eds.) *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6–11, 2021*. pp. 5835–5847. ACL (2021)
- [48] Roberts, K., Alam, T., Bedrick, S., Demner-Fushman, D., Lo, K., Soboroff, I., Voorhees, E.M., Wang, L.L., Hersh, W.R.: TREC-COVID: Rationale and structure of an information retrieval shared task for

- COVID-19. *Journal of the American Medical Informatics Association* **27**(9), 1431–1436 (2020)
- [49] Sakai, T.: Alternatives to Bpref. In: Kraaij, W., de Vries, A.P., Clarke, C.L.A., Fuhr, N., Kando, N. (eds.) *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Amsterdam, The Netherlands, July 23–27, 2007. pp. 71–78. ACM (2007)
- [50] Sakai, T.: Comparing metrics across TREC and NTCIR: The robustness to system bias. In: Shanahan, J.G., Amer-Yahia, S., Manolescu, I., Zhang, Y., Evans, D.A., Kolcz, A., Choi, K., Chowdhury, A. (eds.) *Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008*, Napa Valley, California, USA, October 26–30, 2008. pp. 581–590. ACM (2008)
- [51] Sakai, T.: How to run an evaluation task – With a primary focus on ad hoc information retrieval. In: Ferro, N., Peters, C. (eds.) *Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF*, The Information Retrieval Series, vol. 41, pp. 71–102. Springer (2019)
- [52] Santhanam, K., Saad-Falcon, J., Franz, M., Khattab, O., Sil, A., Florian, R., Sultan, M.A., Roukos, S., Zaharia, M., Potts, C.: Moving beyond downstream task accuracy for information retrieval benchmarking. In: Rogers, A., Boyd-Graber, J.L., Okazaki, N. (eds.) *Findings of the Association for Computational Linguistics: ACL 2023*, Toronto, Canada, July 9–14, 2023. pp. 11613–11628. ACL (2023)
- [53] Scells, H., Zhuang, S., Zuccon, G.: Reduce, reuse, recycle: Green information retrieval research. In: Amigó, E., Castells, P., Gonzalo, J., Carterette, B., Culpepper, J.S., Kazai, G. (eds.) *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Madrid, Spain, July 11 – 15, 2022. pp. 2825–2837. ACM (2022)
- [54] Soboroff, I.: Don't use LLMs to make relevance judgments. arXiv 2409.15133 (2024)
- [55] Strubell, E., Ganesh, A., McCallum, A.: Energy and policy considerations for deep learning in NLP. In: Korhonen, A., Traum, D.R., Màrquez, L. (eds.) *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019*, Florence, Italy, July 28 – August 2, 2019, Volume 1: Long Papers. pp. 3645–3650. ACL (2019)
- [56] Sun, S., Zhuang, S., Wang, S., Zuccon, G.: An investigation of prompt variations for zero-shot LLM-based rankers. arXiv 2406.14117 (2024)
- [57] Tang, Y., Zhang, R., Guo, J., de Rijke, M., Chen, W., Cheng, X.: Listwise generative retrieval models via a sequential learning process. *ACM Transactions on Information Systems* **42**(5), 133:1–133:31 (2024)
- [58] Tay, Y., Tran, V., Dehghani, M., Ni, J., Bahri, D., Mehta, H., Qin, Z., Hui, K., Zhao, Z., Gupta, J.P., Schuster, T., Cohen, W.W., Metzler, D.: Transformer memory as a differentiable search index. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) *Advances*

- in *Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 – December 9, 2022* (2022)
- [59] Thakur, N., Bonifacio, L., Fröbe, M., Bondarenko, A., Kamaloo, E., Potthast, M., Hagen, M., Lin, J.: Systematic evaluation of neural retrieval models on the Touché 2020 argument retrieval subset of BEIR. In: Yang, G.H., Wang, H., Han, S., Hauff, C., Zuccon, G., Zhang, Y. (eds.) *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14–18, 2024*. pp. 1420–1430. ACM (2024)
- [60] Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., Gurevych, I.: BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In: Vanschoren, J., Yeung, S. (eds.) *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual* (2021)
- [61] Trotman, A.: Compressing inverted files. *Information Retrieval* **6**(1), 5–19 (2003)
- [62] Urbano, J., Marrero, M.: The treatment of ties in AP correlation. In: Kamps, J., Kanoulas, E., de Rijke, M., Fang, H., Yilmaz, E. (eds.) *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2017, Amsterdam, The Netherlands, October 1–4, 2017*. pp. 321–324. ACM (2017)
- [63] Voorhees, E.M.: The philosophy of information retrieval evaluation. In: Peters, C., Braschler, M., Gonzalo, J., Kluck, M. (eds.) *Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Darmstadt, Germany, September 3–4, 2001, Revised Papers. Lecture Notes in Computer Science, vol. 2406*, pp. 355–370. Springer (2001)
- [64] Voorhees, E.M.: Overview of the TREC 2004 Robust track. In: Voorhees, E.M., Buckland, L.P. (eds.) *Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, USA, November 16–19, 2004. NIST Special Publication, vol. 500-261. National Institute of Standards and Technology (NIST)* (2004)
- [65] Voorhees, E.M.: The effect of sampling strategy on inferred measures. In: Geva, S., Trotman, A., Bruza, P., Clarke, C.L.A., Järvelin, K. (eds.) *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast , QLD, Australia - July 06–11, 2014*. pp. 1119–1122. ACM (2014)
- [66] Voorhees, E.M.: The evolution of Cranfield. In: Ferro, N., Peters, C. (eds.) *Information Retrieval Evaluation in a Changing World – Lessons Learned from 20 Years of CLEF, The Information Retrieval Series, vol. 41*, pp. 45–69. Springer (2019)
- [67] Voorhees, E.M., Craswell, N., Lin, J.: Too many relevants: Whither Cranfield test collections? In: Amigó, E., Castells, P., Gonzalo, J., Carterette, B., Culpepper, J.S., Kazai, G. (eds.) *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in*

- Information Retrieval, Madrid, Spain, July 11–15, 2022. pp. 2970–2980. ACM (2022)
- [68] Voorhees, E.M., Soboroff, I., Lin, J.: Can old TREC collections reliably evaluate modern neural retrieval models? arXiv 2201.11086 (2022)
- [69] Webber, W., Moffat, A., Zobel, J.: A similarity measure for indefinite rankings. *ACM Transactions on Information Systems* **28**(4), 20:1–20:38 (2010)
- [70] Wirth, N.: A plea for lean software. *IEEE Computer* **28**(2), 64–68 (1995)
- [71] Witten, I.H., Moffat, A., Bell, T.C.: *Managing Gigabytes: Compressing and Indexing Documents and Images*, Second Edition. Morgan Kaufmann (1999)
- [72] Xiong, L., Xiong, C., Li, Y., Tang, K., Liu, J., Bennett, P.N., Ahmed, J., Overwijk, A.: Approximate nearest neighbor negative contrastive learning for dense text retrieval. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3–7, 2021. OpenReview.net (2021)
- [73] Yang, P., Fang, H., Lin, J.: Anserini: Enabling the use of Lucene for information retrieval research. In: Kando, N., Sakai, T., Joho, H., Li, H., de Vries, A.P., White, R.W. (eds.) *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Shinjuku, Tokyo, Japan, August 7–11, 2017. pp. 1253–1256. ACM (2017)
- [74] Yilmaz, E., Aslam, J.A.: Estimating average precision with incomplete and imperfect judgments. In: Yu, P.S., Tsotras, V.J., Fox, E.A., Liu, B. (eds.) *Proceedings of the 2006 ACM CIKM International Conference on Information and Knowledge Management*, Arlington, Virginia, USA, November 6–11, 2006. pp. 102–111. ACM (2006)
- [75] Yilmaz, E., Kanoulas, E., Aslam, J.A.: A simple and efficient sampling method for estimating AP and NDCG. In: Myaeng, S., Oard, D.W., Sebastiani, F., Chua, T., Leong, M. (eds.) *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2008, Singapore, July 20–24, 2008. pp. 603–610. ACM (2008)
- [76] Zhao, W.X., Liu, J., Ren, R., Wen, J.: Dense text retrieval based on pretrained language models: A survey. *ACM Transactions on Information Systems* **42**(4), 89:1–89:60 (2024)
- [77] Zhuang, S., Zuccon, G.: Asyncval: A toolkit for asynchronously validating dense retriever checkpoints during training. In: Amigó, E., Castells, P., Gonzalo, J., Carterette, B., Culpepper, J.S., Kazai, G. (eds.) *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Madrid, Spain, July 11 - 15, 2022. pp. 3235–3239. ACM (2022)
- [78] Zuccon, G., Scells, H., Zhuang, S.: Beyond CO2 emissions: The overlooked impact of water consumption of information retrieval models. In: Yoshioka, M., Kiseleva, J., Aliannejadi, M. (eds.) *Proceedings of the 2023*

ACM SIGIR International Conference on Theory of Information Retrieval,
ICTIR 2023, Taipei, Taiwan, 23 July 2023. pp. 283-289. ACM (2023)