OVERVIEW OF THE TREC 2020 DEEP LEARNING TRACK

Nick Craswell¹, Bhaskar Mitra^{1,2}, Emine Yilmaz², and Daniel Campos³

¹Microsoft AI & Research, {nickcr, bmitra}@microsoft.com
²University College London, {bhaskar.mitra.15,emine.yilmaz}@ucl.ac.uk
³University of Illinois Urbana-Champaign, {dcampos3}@illinois.edu

ABSTRACT

This is the second year of the TREC Deep Learning Track, with the goal of studying ad hoc ranking in the large training data regime. We again have a document retrieval task and a passage retrieval task, each with hundreds of thousands of human-labeled training queries. We evaluate using singleshot TREC-style evaluation, to give us a picture of which ranking methods work best when large data is available, with much more comprehensive relevance labeling on the small number of test queries. This year we have further evidence that rankers with BERT-style pretraining outperform other rankers in the large data regime.

1 Introduction

Deep learning methods, where a computational model learns an intricate representation of a large-scale dataset, yielded dramatic performance improvements in speech recognition and computer vision [LeCun et al., 2015]. When we have seen such improvements, a common factor is the availability of large-scale training data [Deng et al., 2009, Bellemare et al., 2013]. For ad hoc ranking in information retrieval, which is a core problem in the field, we did not initially see dramatic improvements in performance from deep learning methods. This led to questions about whether deep learning methods were helping at all [Yang et al., 2019a]. If large training data sets are a factor, one explanation for this could be that the training sets were too small.

The TREC Deep Learning Track, and associated MS MARCO leaderboards [Bajaj et al., 2016], have introduced human-labeled training sets that were previously unavailable. The main goal is to study information retrieval in the *large training data* regime, to see which retrieval methods work best.

The two tasks, document retrieval and passage retrieval, each have hundreds of thousands of human-labeled training queries. The training labels are sparse, with often only one positive example per query. Unlike the MS MARCO leaderboards, which evaluate using the same kind of sparse labels, the evaluation at TREC uses much more comprehensive relevance labeling. Each year of TREC evaluation evaluates on a new set of test queries, where participants submit before the test labels have even been generated, so the TREC results are the gold standard for avoiding multiple testing and overfitting. However, the comprehensive relevance labeling also generates a reusable test collections, allowing reuse of the dataset in future studies, although people should be careful to avoid overfitting and overiteration.

The main goals of the Deep Learning Track in 2020 have been: 1) To provide large reusable training datasets with associated large scale click dataset for training deep learning and traditional ranking methods in a large training data regime, 2) To construct reusable test collections for evaluating quality of deep learning and traditional ranking methods, 3) To perform a rigorous blind single-shot evaluation, where test labels don't even exist until after all runs are submitted, to compare different ranking methods, and 4) To study this in both a traditional TREC setup with end-to-end retrieval and in a re-ranking setup that matches how some models may be deployed in practice.

2 Task description

The track has two tasks: Document retrieval and passage retrieval. Participants were allowed to submit up to three runs per task, although this was not strictly enforced. Submissions to both tasks used the same set of 200 test queries.

In the pooling and judging process, NIST chose a subset of the queries for judging, based on budget constraints and with the goal of finding a sufficiently comprehensive set of relevance judgments to make the test collection reusable. This led to a judged test set of 45 queries for document retrieval and 54 queries for passage retrieval. The document queries are not a subset of the passage queries.

When submitting each run, participants indicated what external data, pretrained models and other resources were used, as well as information on what style of model was used. Below we provide more detailed information about the document retrieval and passage retrieval tasks, as well as the datasets provided as part of these tasks.

2.1 Document retrieval task

The first task focuses on document retrieval, with two subtasks: (i) Full retrieval and (ii) top-100 reranking.

In the full retrieval subtask, the runs are expected to rank documents based on their relevance to the query, where documents can be retrieved from the full document collection provided. This subtask models the end-to-end retrieval scenario.

In the reranking subtask, participants were provided with an initial ranking of 100 documents, giving all participants the same starting point. This is a common scenario in many real-world retrieval systems that employ a telescoping architecture [Matveeva et al., 2006, Wang et al., 2011]. The reranking subtask allows participants to focus on learning an effective relevance estimator, without the need for implementing an end-to-end retrieval system. It also makes the reranking runs more comparable, because they all rerank the same set of 100 candidates.

The initial top-100 rankings were retrieved using Indri [Strohman et al., 2005] on the full corpus with Krovetz stemming and stopwords eliminated.

Judgments are on a four-point scale:

- [3] **Perfectly relevant:** Document is dedicated to the query, it is worthy of being a top result in a search engine.
- [2] Highly relevant: The content of this document provides substantial information on the query.
- [1] **Relevant:** Document provides some information relevant to the query, which may be minimal.
- [0] Irrelevant: Document does not provide any useful information about the query.

For metrics that binarize the judgment scale, we map document judgment levels 3,2,1 to relevant and map document judgment level 0 to irrelevant.

2.2 Passage retrieval task

Similar to the document retrieval task, the passage retrieval task includes (i) a full retrieval and (ii) a top-1000 reranking tasks.

In the full retrieval subtask, given a query, the participants were expected to retrieve a ranked list of passages from the full collection based on their estimated likelihood of containing an answer to the question. Participants could submit up to 1000 passages per query for this end-to-end retrieval task.

In the top-1000 reranking subtask, 1000 passages per query were provided to participants, giving all participants the same starting point. The sets of 1000 were generated based on BM25 retrieval with no stemming as applied to the full collection. Participants were expected to rerank the 1000 passages based on their estimated likelihood of containing an answer to the query. In this subtask, we can compare different reranking methods based on the same initial set of 1000 candidates, with the same rationale as described for the document reranking subtask.

Judgments are on a four-point scale:

- [3] **Perfectly relevant:** The passage is dedicated to the query and contains the exact answer.
- [2] **Highly relevant:** The passage has some answer for the query, but the answer may be a bit unclear, or hidden amongst extraneous information.
- [1] **Related:** The passage seems related to the query but does not answer it.
- [0] **Irrelevant:** The passage has nothing to do with the query.

For metrics that binarize the judgment scale, we map passage judgment levels 3,2 to relevant and map document judgment levels 1,0 to irrelevant.

Data	Document task Number of records	Passage task Number of records
Corpus	3,213,835	8,841,823
Train queries Train qrels	367,013 384,597	502,939 532,761
Dev queries Dev qrels	$5,193 \\ 5,478$	$6,980 \\ 7,437$
2019 TREC queries 2019 TREC qrels	$\begin{array}{c} 200 \rightarrow 43 \\ 16,258 \end{array}$	$\begin{array}{c} 200 \rightarrow 43 \\ 9,260 \end{array}$
2020 TREC queries 2020 TREC qrels	$\begin{array}{c} 200 \rightarrow 45 \\ 9,098 \end{array}$	$\begin{array}{c} 200 \rightarrow 54 \\ 11,386 \end{array}$

Table 1: Summary of statistics on TREC 2020 Deep Learning Track datasets.

Table 2: Summary of ORCAS data. Each record in the main file (orcas.tsv) indicates a click between a query (Q) and a URL (U), also listing a query ID (QID) and the corresponding TREC document ID (DID). The run file is the top-100 using Indri query likelihood, for use as negative samples during training.

Filename	Number of records	Data in each record
orcas.tsv	18.8M	QID Q DID U
orcas-doctrain-qrels.tsv	18.8M	QID DID
orcas-doctrain-queries.tsv	10.4M	QID Q
orcas-doctrain-top100	983M	QID DID score

3 Datasets

Both tasks have large training sets based on human relevance assessments, derived from MS MARCO. These are sparse, with no negative labels and often only one positive label per query, analogous to some real-world training data such as click logs.

In the case of passage retrieval, the positive label indicates that the passage contains an answer to a query. In the case of document retrieval, we transferred the passage-level label to the corresponding source document that contained the passage. We do this under the assumption that a document with a relevant passage is a relevant document, although we note that our document snapshot was generated at a different time from the passage dataset, so there can be some mismatch. Despite this, machine learning models trained with these labels seem to benefit from using the labels, when evaluated using NIST's non-sparse, non-transferred labels. This suggests the transferred document labels are meaningful for our TREC task.

This year for the document retrieval task, we also release a large scale click dataset, The ORCAS data, constructed from the logs of a major search engine [Craswell et al., 2020]. The data could be used in a variety of ways, for example as additional training data (almost 50 times larger than the main training set) or as a document field in addition to title, URL and body text fields available in the original training data.

For each task there is a corresponding MS MARCO leaderboard, using the same corpus and sparse training data, but using sparse data for evaluation as well, instead of the NIST test sets. We analyze the agreement between the two types of test in Section 4.

Table 1 and Table 2 provide descriptive statistics for the dataset derived from MS MARCO and the ORCAS dataset, respectively. More details about the datasets—including directions for download—is available on the TREC 2020 Deep Learning Track website¹. Interested readers are also encouraged to refer to [Bajaj et al., 2016] for details on the original MS MARCO dataset.

¹https://microsoft.github.io/TREC-2020-Deep-Learning

Table 3: Summary of statistics of runs for the two retrieval tasks at the TREC 2020 Deep Learning Track.

	Document retrieval	Passage retrieval
Number of groups	14	14
Number of total runs	64	59
Number of runs w/ category: nnlm	27	43
Number of runs w/ category: nn	11	2
Number of runs w/ category: trad	26	14
Number of runs w/ category: rerank	19	18
Number of runs w/ category: fullrank	45	41



Figure 1: NDCG@10 results, broken down by run type. Runs of type "nnlm", meaning they use language models such as BERT, performed best on both tasks. Other neural network models "nn" and non-neural models "trad" had relatively lower performance this year. More iterations of evaluation and analysis would be needed to determine if this is a general result, but it is a strong start for the argument that deep learning methods may take over from traditional methods in IR applications.

4 Results and analysis

Submitted runs The TREC 2020 Deep Learning Track had 25 participating groups, with a total of 123 runs submitted across both tasks.

Based run submission surveys, we manually classify each run into one of three categories:

- **nnlm:** if the run employs large scale pre-trained neural language models, such as BERT [Devlin et al., 2018] or XLNet [Yang et al., 2019b]
- **nn:** if the run employs some form of neural network based approach—*e.g.*, Duet [Mitra et al., 2017, Mitra and Craswell, 2019] or using word embeddings [Joulin et al., 2016]—but does not fall into the "nnlm" category
- **trad:** if the run exclusively uses traditional IR methods like BM25 [Robertson et al., 2009] and RM3 [Abdul-Jaleel et al., 2004].

We placed 70 (57%) runs in the "nnlm" category, 13 (10%) in the "nn" category, and the remaining 40 (33%) in the "trad" category. In 2019, 33 (44%) runs were in the "nnlm" category, 20 (27%) in the "nn" category, and the remaining 22 (29%) in the "trad" category. While there was a significant increase in the total number of runs submitted compared to last year, we observed a significant reduction in the fraction of runs in the "nn" category.

We further categorize runs based on subtask:

- **rerank:** if the run reranks the provided top-k candidates, or
- fullrank: if the run employs their own phase 1 retrieval system.

We find that only 37 (30%) submissions fall under the "rerank" category—while the remaining 86 (70%) are "full-rank". Table 3 breaks down the submissions by category and task.

Overall results Our main metric in both tasks is Normalized Discounted Cumulative Gain (NDCG)—specifically, NDCG@10, since it makes use of our 4-level judgments and focuses on the first results that users will see. To get a picture of the ranking quality outside the top-10 we also report Average Precision (AP), although this binarizes the judgments. For comparison to the MS MARCO leaderboard, which often only has one relevant judgment per query, we report the Reciprocal Rank (RR) of the first relevant document on the NIST judgments, and also using the sparse leaderboard judgments.

Some of our evaluation is concerned with the quality of the top-k results, where k = 100 for the document task and k = 1000 for the passage task. We want to consider the quality of the top-k set without considering how they are ranked, so we can see whether improving the set-based quality is correlated with an improvement in NDCG@10. Although we could use Recall@k as a metric here, it binarizes the judgments, so we instead use Normalized Cumulative Gain (NCG@k) [Rosset et al., 2018]. NCG is not supported in trec_eval. For trec_eval metrics that are correlated, see Recall@k and NDCG@k.

The overall results are presented in Table 4 for document retrieval and Table 5 for passage retrieval. These tables include multiple metrics and run categories, which we now use in our analysis.

Neural vs. traditional methods. The first question we investigated as part of the track is which ranking methods work best in the large-data regime. We summarize NDCG@10 results by run type in Figure 1.

For document retrieval runs (Figure 1a) the best "trad" run is outperformed by "nn" and "nnlm" runs by several percentage points, with "nnlm" also having an advantage over "nn". We saw a similar pattern in our 2019 results. This year we encouraged submission of a variety of "trad" runs from different participating groups, to give "trad" more chances to outperform other run types. The best performing run of each category is indicated, with the best "nnlm" and "nn" models outperforming the best "trad" model by 23% and 11% respectively.

For passage retrieval runs (Figure 1b) the gap between the best "nnlm" and "nn" runs and the best "trad" run is larger, at 42% and 17% respectively. One explanation for this could be that vocabulary mismatch between queries and relevant results is greater in short text, so neural methods that can overcome such mismatch have a relatively greater advantage in passage retrieval. Another explanation could be that there is already a public leaderboard, albeit without test labels from NIST, for the passage task. (We did not launch the document ranking leaderboard until after our 2020 TREC submission deadline.) In passage ranking, some TREC participants may have submitted neural models multiple times to the public leaderboard, so are relatively more experienced working with the passage dataset than the document dataset.

In query-level win-loss analysis for the document retrieval task (Figure 2) the best "nnlm" model outperforms the best "trad" run on 38 out of the 45 test queries (*i.e.*, 84%). Passage retrieval shows a similar pattern in Figure 3. Similar to last year's data, neither task has a large class of queries where the "nnlm" model performs worse.

End-to-end retrieval vs. reranking. Our datasets include top-k candidate result lists, with 100 candidates per query for document retrieval and 1000 candidates per query for passage retrieval. Runs that simply rerank the provided candidates are "rerank" runs, whereas runs that perform end-to-end retrieval against the corpus, with millions of potential results, are "fullrank" runs. We would expect that a "fullrank" run should be able to find a greater number of relevant candidates than we provided, achieving higher NCG@k. A multi-stage "fullrank" run should also be able to optimize the stages jointly, such that early stages produce candidates that later stages are good at handling.

According to Figure 4, "fullrank" did not achieve much better NDCG@10 performance than "rerank" runs. In fact, for the passage retrieval task, the top two runs are of type "rerank". While it was possible for "fullrank" to achieve better NCG@k, it was also possible to make NCG@k worse, and achieving significantly higher NCG@k does not seem necessary to achieve good NDCG@10.

Specifically, for the document retrieval task, the best "fullrank" run achieves 5% higher NDCG@10 over the best "rerank' run; whereas for the passage retrieval task, the best "fullrank" run performs slightly worse (0.3% lower NDCG@10) compared to the best "rerank' run.

Similar to our observations from Deep Learning Track 2019, we are not yet seeing a strong advantage of "fullrank" over "rerank". However, we hope that as the body of literature on neural methods for phase 1 retrieval (*e.g.*, [Boytsov et al., 2016, Zamani et al., 2018, Mitra et al., 2019, Nogueira et al., 2019]) grows, we would see a larger number of runs with deep learning as an ingredient for phase 1 in future editions of this TREC track.

Effect of ORCAS data Based on the descriptions provided, ORCAS data seems to have been used by six of the runs (ndrm3-orc-full, ndrm3-orc-re, uogTrBaseL17, uogTrBaseQL170, uogTr310R, relemb_mlm_0_2). Most runs seem to be make use of the ORCAS data as a field, with some runs using the data as an additional training dataset as well.

Table 4: Document retrieval runs. RR (MS) is based on MS MARCO labels. All other metrics are based on NIST labels. Rows are sorted by NDCG@10.

cl.d2q.m3.duo b2oloo fullraak nnlm 0.4471 0.9476 0.6934 0.7789 0.5427 d_m3.duo b2oloo fullraak nnlm 0.4547 0.9476 0.6934 0.7789 0.5270 GLP_nnl ICIP rerank nnlm 0.8487 0.9476 0.6623 0.6233 0.6233 0.6233 0.6233 0.6233 0.6233 0.6233 0.6233 0.6233 0.4330 ICIP_nna ICIP rerank nnlm 0.4947 0.9667 0.6623 0.6233 0.4130 Poheta-larg BITEM rerank nnlm 0.412 0.0231 0.6232 0.6283 0.4230 Profin-re- MSA1 rerank nn 0.4451 0.9241 0.6217 0.6624 0.4060 ndm3-re- MSA1 rerank nn 0.4258 0.9333 0.6161 0.6233 0.4120 ndm3-re- MSA1 rerank nnlm 0.3255 0.9119 0.6331 0.6260 <	run	group	subtask	neural	RR (MS)	RR	NDCG@10	NCG@100	AP
	d d2q duo	h2oloo	fullrank	nnlm	0.4451	0.9476	0.6934	0.7718	0.5422
	d_d2q_rm3_duo	h2oloo	fullrank	nnlm	0.4541	0.9476	0.6900	0.7769	0.5427
	d_rm3_duo	h2oloo	fullrank	nnlm	0.4547	0.9476	0.6794	0.7498	0.5270
	ICIP_run1	ICIP	rerank	nnlm	0.3898	0.9630	0.6623	0.6283	0.4333
$ fr_doc_roberta BITEM fullrank nnlm 0.3943 0.9365 0.6440 0.6406 0.4425 roberta-large BITEM rerark nnlm 0.3782 0.9185 0.6295 0.6283 0.4206 roberta-large BITEM rerark nnlm 0.3782 0.9185 0.6295 0.6283 0.4308 ndm3-orc-full MSAI fullrank nnlm 0.4102 0.929 0.6278 0.6604 0.4308 ndm3-orc-full MSAI fullrank nn 0.4450 0.9444 0.6249 0.6764 0.4280 ndm3-orc-full MSAI rerark nn 0.4451 0.9241 0.6217 0.6283 0.4194 ndm3-full MSAI fullrank nn 0.4213 0.9333 0.6162 0.6628 0.4192 ndm3-re MSAI rerark nn 0.4421 0.9333 0.6161 0.6283 0.4122 ndm1-re MSAI rerark nn 0.4427 0.9333 0.6161 0.6283 0.4305 ndm3-rel MSAI rerark nn 0.4427 0.9333 0.6161 0.6283 0.4305 ndm1-re MSAI rerark nn 0.4427 0.9333 0.6161 0.6283 0.4305 ndm1-re MSAI rerark nnlm 0.3235 0.9119 0.6031 0.6283 0.3936 ndm1-full MSAI fullrank nnlm 0.3303 0.9000 0.6017 0.6283 0.3363 0.9300 ndm1-full MSAI fullrank nnlm 0.3384 0.9366 0.5907 0.6669 0.4259 1.02372 0.9284 0.929 0.5949 0.6283 0.3368 0.9368 0.3588 0.6752 0.4230 TUW-TKL-2k TU_Viena rerark nnlm 0.3384 0.9296 0.5882 0.6353 0.3406 0.4242 ndp1-re reark nnlm 0.3321 0.8727 0.8889 0.5846 0.6383 0.3740 rug-TrQ-BMP U-GTF fullrank nnlm 0.3321 0.872 0.571 0.6633 0.3740 ndg7 0.9188 0.571 0.6633 0.3740 ndg7 0.9188 0.5846 0.6383 0.3749 hg/R-DH-TS-F QU fullrank nnlm 0.3321 0.872 0.571 0.6633 0.3740 ndg7 0.9182 0.571 0.6633 0.3740 ndg7 0.9182 0.571 0.6633 0.3742 ndg7 0.5810 0.6537 0.6609 0.4752 ndg-TrQ-BMP U-GTF fullrank nnlm 0.321 0.8852 0.571 0.6638 0.3740 hg/T-DH-TS-F QU fullrank nnlm 0.321 0.8852 0.571 0.6638 0.3740 hg/T-QH-TS-F QU fullrank nnlm 0.321 0.8852 0.571 0.6638 0.3740 hg/T-QH-TS-F QU fullrank nnlm 0.324 0.2859 0.6539 0.6539 0.6399 1.06283 0.3740 hg/T-DH-TS-F QU fullrank nnlm 0.324 0.8850 0.5731 0.6638 0.3542 ndg-TA 0.5851 0.6523 0.0629 0.6399 0.0392 hg/T-DH-TS-F QU fullrank nnlm 0.3251 0.8750 0.6531 0.3638 0.3740 hg/T-DH-TS-F QU fullrank nnlm 0.324 0.8852 0.571 0.6642 0.3368 hg/T-TS-F QU fullrank nnlm 0.324 0.8850 0.571 0.6642 0.3368 ndf-TS 0.6420 0.3368 ndf-TS 0.6420 0.3368 ndf-TS 0.6420 0.3368 ndf-TS 0.6420 0.3368 ndf-$	ICIP_run3	ICIP	rerank	nnlm	0.4479	0.9667	0.6528	0.6283	0.4360
	fr_doc_roberta	BITEM	fullrank	nnlm	0.3943	0.9365	0.6404	0.6806	0.4423
roberta-large BITEM rerank nnlm 0.3782 0.9185 0.6295 0.6238 0.4199 ndrm3-ore-full MSAI fullrank nnlm 0.4369 0.9444 0.6249 0.6764 0.4280 ndrm3-ore-full MSAI rerank nn 0.4451 0.9241 0.6217 0.6283 0.4194 ndrm3-re-full MSAI rerank nn 0.4451 0.9241 0.6217 0.6283 0.4194 ndrm3-re-full MSAI rerank nn 0.4421 0.9333 0.6162 0.6628 0.4069 ndrm3-re MSAI rerank nn 0.4427 0.9333 0.6161 0.6283 0.4122 ndrm1-re MSAI rerank nn 0.4427 0.9333 0.6161 0.6283 0.4122 ndrm1-re MSAI rerank nnlm 0.3228 0.8833 0.6151 0.6283 0.4305 nupiL_run rupi rerank nnlm 0.3228 0.8833 0.6151 0.6283 0.4305 nupiL_run rupi rerank nnlm 0.3235 0.9119 0.6011 0.6283 0.3368 nupiL_runi rupi rerank nnlm 0.3234 0.9233 0.5901 0.6283 0.3368 nupiL_runi rupi rerank nnlm 0.3304 0.9000 0.6017 0.6283 0.3946 nupiL_runi rupi rerank nnlm 0.3394 0.9256 0.5855 0.6725 0.4230 nupiL_runi rupi rerank nnlm 0.3384 0.8916 0.5907 0.6669 0.4259 d_d24_l.pm25 naserini fullrank nnlm 0.3184 0.8916 0.5907 0.6669 0.4259 d_d24_l.pm25 naserini fullrank nnlm 0.3374 0.9000 0.5885 0.6752 0.4230 uob_runid UoB rerank nnlm 0.3521 0.9720 0.5846 0.6283 0.3848 biglR-DH-T5-R QU retrank nnlm 0.3524 0.9100 0.5830 0.6283 0.3848 tuob_runid UoB rerank nnlm 0.3524 0.9100 0.5830 0.6283 0.3976 uob_runid UoGTr fullrank nnlm 0.3724 0.8829 0.5749 0.6283 0.3786 TUW-TKL-4k TU_Vienna rerank nnlm 0.3470 0.8889 0.5577 0.6609 0.4177 bi_bcai_multid b_bcai fullrank rad 0.3022 0.5744 0.6669 0.4177 bi_bcai_multid b_bcai fullrank rad 0.3124 0.8858 0.5557 0.6628 0.3303 uogTr120 UoGTr fullrank rad 0.3270 0.8926 0.5435 0.66283 0.3786 terrie-sycbc bcai fullrank rad 0.3270 0.8926 0.5436 0.6283 0.3786 terrie-sycbc bcai fullrank rad 0.3224 0.8250 0.6438 0.3578 0.6738 0.3731 uogTr120 UoGTr fullrank rad 0.3270 0.8946 0.5557 0.6642 0.3906 biglR-DH-T5-F QU fullrank rad 0.3270 0.8648 0.5557 0.6642 0.3906 biglR-DH-T5-F QU fullrank rad 0.3082 0.8648 0.5557 0.6642 0.3906 biglR-DH-T5-F QU fullrank rad 0.3082 0.8648 0.5557 0.6649 0.3774 bi_bcai_multid b_reai fullrank rad 0.3140 0.8371 0.8451 0.6433 0.3764 bi_glca_hd=1.528 0.6673 0.8669 0.34	ICIP_run2	ICIP	rerank	nnlm	0.4081	0.9407	0.6322	0.6283	0.4206
beal_serb_docv bear tulirank nnim 0.4102 0.92.99 0.6278 0.66278 0.4280 ndrm3-orc-te MSAI rerank nn 0.4451 0.9241 0.6217 0.6233 0.4190 ndrm3-re MSAI rerank nn 0.4451 0.9241 0.6217 0.6233 0.44280 ndrm3-re MSAI rerank nn 0.4258 0.9333 0.6162 0.6626 0.4669 ndrm3-re MSAI rerank nn 0.4258 0.9333 0.6162 0.6628 0.4150 npii_run2 mpii rerank nn 0.4258 0.8833 0.6135 0.6623 0.4120 npii_run1 npii rerank nnlm 0.3228 0.8833 0.6135 0.6623 0.4250 npii_run1 npii rerank nnlm 0.3228 0.9000 0.6617 0.6623 0.4325 nub_rnid2 UoB rerank nnlm 0.3234 0.9259 0.5949 0.6233 0.4365 ub_rnid3 UoB rerank nnlm 0.3248 0.9259 0.5949 0.6233 0.4365 ub_rnid3 UoB rerank nnlm 0.3244 0.9259 0.5949 0.6233 0.4365 ub_rnid3 UoB rerank nnlm 0.3284 0.9259 0.5852 0.6623 0.4325 TUW-TKL-2k TU_Vienna rerank nnlm 0.3388 0.9366 0.5885 0.6752 0.4230 TUW-TKL-2k TU_Vienna rerank nnlm 0.3531 0.9000 0.5880 0.6523 0.3484 ub_rnid3 UoB rerank nnlm 0.3531 0.9276 0.5889 0.5846 0.6283 0.3842 ub_rnid2 UoB rerank nnlm 0.3531 0.9276 0.5889 0.5832 0.6633 0.3946 ub_rnid2 UoB rerank nnlm 0.3531 0.8722 0.5781 0.6633 0.3946 bigR-DH-TS-F QU rerank nnlm 0.3124 0.8852 0.5781 0.6633 0.3746 bigR-DH-TS-F QU rerank nnlm 0.3124 0.8852 0.5781 0.6633 0.3746 bigR-DH-TS-F QU rerank nnlm 0.3124 0.8852 0.5781 0.6633 0.3746 bigR-DH-TS-F QU rerank nnlm 0.3124 0.8852 0.5781 0.6638 0.3749 bigR-DH-TS-F QU rerank nnlm 0.3124 0.8892 0.5570 0.6640 0.4177 bigR-DH-TS-F QU rerank nnlm 0.3124 0.8892 0.5570 0.6640 0.3734 bigR-DH-TS-F QU rerank nnlm 0.3124 0.8892 0.5570 0.6643 0.3749 bigR-DH-TS-F QU rerank nnlm 0.3124 0.8892 0.5520 0.6629 0.3904 bigR-DH-TS-F QU rerank nnlm 0.3124 0.8892 0.5520 0.6629 0.3904 bigR-DH-TS-F QU rerank nnlm 0.3144 0.8898 0.5520 0.6623 0.3740 bigR-DH-TS-F QU rerank nnlm 0.3144 0.8894 0.5520 0.6633 0.3764 bigR-DH-TS-F QU rerank nnlm 0.3144 0.8894 0.5520 0.6631 0.3784 bigR-DH-TS-F QU rerank nnlm 0.3178	roberta-large	BITEM	rerank	nnlm	0.3782	0.9185	0.6295	0.6283	0.4199
hdm-3-cre-teili MSAI tuirank nn 0.4509 0.9444 0.6249 0.6764 0.4280 0.4194	bcai_bertb_docv	bcai	fullrank	nnlm	0.4102	0.9259	0.6278	0.6604	0.4308
$ \begin{array}{cccccccccccccccccccccccccccccccccccc$	ndrm3-orc-full	MSAI	Tullrank	nn	0.4369	0.9444	0.6249	0.6764	0.4280
$ \begin{array}{cccccccccccccccccccccccccccccccccccc$	ndrm3-orc-re	MSAI	fullronk	nn	0.4451	0.9241	0.6217	0.6283	0.4194
$ \begin{array}{cccccccccccccccccccccccccccccccccccc$	ndrm3 re	MSAI	roronk	nn	0.4213	0.9333	0.0102	0.0020	0.4009
$\begin{array}{cccccccccccccccccccccccccccccccccccc$	ndrm1_re	MSAI	rerank	nn	0.4238	0.9333	0.6161	0.6283	0.4122
$\begin{tabular}{transform} \begin{tabular}{l $	mpii run?	mpii	rerank	nnlm	0.3228	0.8833	0.6135	0.6283	0.4205
mpii runi mpii rerank mnim 0.3503 0.9000 0.6017 0.6283 0.4030 ndrm1-full MSA1 fullrank nn 0.4350 0.9333 0.5991 0.6283 0.3888 bigR-DTH-T5-F QU fullrank nnlm 0.3184 0.8916 0.5907 0.6669 0.4259 d_42q_Dm25 anserini fullrank nnlm 0.3383 0.9266 0.5885 0.6752 0.4230 0.3848 bigR-DH-T3-R QU rerank nnlm 0.3521 0.8722 0.5781 0.6283 0.3876 uob_runid1 UoB rerank nnlm 0.3521 0.8722 0.5781 0.6283 0.3749 bigR-DH-T5-F QU fullrank trad 0.2622 0.9195 0.5734 0.6669 0.4177 bcai_classic bcai fullrank trad 0.3124 0.8679 0.6298 0.33749 bigR-DH-T5-F QU fullrank trad 0.3241 0.8679	higIR-DTH-T5-R	OU	rerank	nnlm	0.3235	0.9119	0.6031	0.6283	0.3936
ndm1-full M	mpii run1	mpii	rerank	nnlm	0.3503	0.9000	0.6017	0.6283	0.4030
uob_runid3UoBreranknnlm0.32940.92590.59490.62830.3948bigIR-DTH-TS-FQUfullranknnlm0.31840.89160.59070.66690.4259TUW-TKL-2kTU_Viennareranknnlm0.33380.92960.58850.67520.4230TUW-TKL-2kTU_Viennareranknnlm0.33340.91000.58360.67830.3842uob_runid2UoBreranknnlm0.33240.91000.58300.62830.3842uob_runid1UoBreranknnlm0.35210.87220.57910.60340.3752uob_runid1UoBreranknnlm0.31240.88520.57810.62830.3766TUW-TKL-4kTU_Viennareranknnlm0.30210.87220.59710.60080.3762uob_runid1UoBreranknnlm0.30210.88520.57340.66690.4177bl_cai_anulfd6bl_bcaifullranktrad0.30220.86480.55570.64200.3906longformer_1USGTfullranktrad0.31220.88290.55470.64200.3906longformer_1USGTfullranknnlm0.32120.82290.54760.5486rererexC2bl_rmitfullranknnlm0.32270.84210.3603logfLCDTTS-RQUreranknnlm0.22370.84110.54350.66230.3373uogTTA20UoGTr<	ndrm1-full	MSAI	fullrank	nn	0.4350	0.9333	0.5991	0.6280	0.3858
$ bigIR-DTH-TS-F QU fullrank nnlm 0.3184 0.8916 0.5907 0.6669 0.4259 0 d.22, hn25 anscrini fullrank nnlm 0.3383 0.9369 0.5885 0.6752 0.4230 0 bigIR-DH-TS-R QU rerank nnlm 0.2367 0.8889 0.5846 0.6283 0.3840 uob_runi2U UoB rerank nnlm 0.2377 0.8889 0.5846 0.6283 0.3842 uob_runi2U UoB rerank nnlm 0.3521 0.8722 0.5791 0.6634 0.3786 UD-runi1U UoB rerank nnlm 0.3521 0.8722 0.5791 0.6634 0.3786 UD-runi1U UoB rerank nnlm 0.3121 0.8722 0.5791 0.6623 0.3786 UD-runi1U UoB rerank nnlm 0.3124 0.8852 0.5781 0.6283 0.3786 UD-runi1U UoB rerank nnlm 0.3124 0.8852 0.5781 0.6283 0.3786 UD-runi1U UoB rerank nnlm 0.3124 0.8852 0.5781 0.6283 0.3786 UD-runi1U UoB rerank nnlm 0.3124 0.8852 0.5751 0.6623 0.3789 bigIR-DH-T5-F QU fullrank trad 0.2622 0.9195 0.5629 0.6299 0.3829 indri-sdmf RMIT fullrank trad 0.2302 0.9195 0.5529 0.6298 0.3376 terrier-expC2 b_rmi1 fullrank trad 0.3081 0.8868 0.5557 0.6420 0.3906 longformer_1 USI rerank nnlm 0.3614 0.8889 0.5520 0.6283 0.3503 uogTr310R UoGTr fullrank trad 0.3122 0.8259 0.5475 0.6423 0.3363 uogTr310R UoGTr fullrank trad 0.3122 0.8259 0.5475 0.6423 0.3363 uogTr310R UoGTr fullrank trad 0.3122 0.8259 0.5475 0.6423 0.3363 uogTr310R UoGTr fullrank trad 0.3122 0.8259 0.5475 0.6423 0.33672 uogTr72U UoGTr fullrank trad 0.2374 0.8716 0.5431 0.6979 0.4087 rmit_indri-fdm bl_rmi1 fullrank trad 0.2779 0.8481 0.5416 0.6812 0.3805 bigIR-DT-T5-R QU fullrank trad 0.2779 0.8411 0.5453 0.6331 0.3773 bigIR-DT-T5-F QU fullrank trad 0.2702 0.8376 0.5431 0.6670 0.3619 bi_bcai_nodell bl_bcai fullrank trad 0.2716 0.8316 0.4228 indri-bm25 bl_rmi1 fullrank trad 0.2304 0.8358 0.5378 0.6633 0.3774 bl_bcai_pcai_pcai_fullrank trad 0.2763 0.8164 0.5342 0.66761 0.3066 0.3764 terrier-isflaf bl_rmi1 fullrank trad 0.2716 0.8358 0.5378 0.6630 0.3774 bl_bcai_pcai_pcai_fullrank trad 0.2720 0.8470 0.5328 0.6631 0.3791 dl_m252m3 anserini fullrank trad 0.3303 0.8267 0.5231 0.6663 0.3784 terrier-isflase bl_rmi1 fullrank trad 0.2716 0.8410 0.3734 0.6650 0.3784 terrier-isflase bl_rmi1 fullrank trad 0.3310 0.8204 0.5321 0.6663 0.3$	uob_runid3	UoB	rerank	nnlm	0.3294	0.9259	0.5949	0.6283	0.3948
	bigIR-DTH-T5-F	QU	fullrank	nnlm	0.3184	0.8916	0.5907	0.6669	0.4259
$\begin{array}{cccccccccccccccccccccccccccccccccccc$	d_d2q_bm25	anserini	fullrank	nnlm	0.3338	0.9369	0.5885	0.6752	0.4230
bigR-DH-T5-R QU rerank nnlm 0.2877 0.8889 0.5846 0.6283 0.3842 uob_runid1 UoB rerank nnlm 0.3521 0.8722 0.5791 0.6034 0.3752 uob_runid1 UoB rerank nnlm 0.3124 0.8852 0.5781 0.6283 0.3749 twork1 Vuema rerank nnlm 0.4097 0.9185 0.5749 0.6283 0.3749 bigRD-DH-T5-F QU fullrank nnlm 0.2622 0.9195 0.5629 0.6229 0.3829 indri-sdmf RMTT fullrank trad 0.3431 0.8766 0.5557 0.6640 0.3573 uogTr310R UoGTr fullrank trad 0.3217 0.8229 0.5476 0.5496 0.3468 retrier-expC2 bi_rmit <fullrank< td=""> nnlm 0.3223 0.8407 0.5453 0.5334 0.3602 bigR-DT-T5-R QU rerak nnlm 0.3274 0.8481 0.5416<</fullrank<>	TUW-TKL-2k	TU_Vienna	rerank	nn	0.3683	0.9296	0.5852	0.6283	0.3810
$ uob_runid2 UoB rerank nnlm 0.3534 0.9100 0.5380 0.6283 0.3976 0.6034 0.3752 $	bigIR-DH-T5-R	QU	rerank	nnlm	0.2877	0.8889	0.5846	0.6283	0.3842
uog1rQCBMP UoG1r fullrank nnlm 0.3521 0.8722 0.5791 0.6034 0.3732 TUW-TKL-4k TU_Vienna rerank nn 0.4097 0.9185 0.5749 0.6283 0.3786 biglR-DH-T5-F QU fullrank rnd 0.2622 0.9195 0.5629 0.6299 0.3829 indri-sdmf RMIT fullrank trad 0.2622 0.9195 0.5629 0.6290 0.3892 longformer_1 USI rerank nnlm 0.3614 0.8889 0.5520 0.6283 0.3596 longformer_1 USI rerank nnlm 0.3212 0.8825 0.5476 0.5496 0.3468 terrier-expC2 b_1mit fullrank rnd 0.2933 0.9407 0.5455 0.6283 0.3373 uogT720 UoGTr fullrank nnlm 0.2212 0.8259 0.5476 0.5431 0.6999 0.4387 midif-DFPE RMIT fullrank nnlm 0.2	uob_runid2	UoB	rerank	nnlm	0.3534	0.9100	0.5830	0.6283	0.3976
$ \begin{array}{cccccccccccccccccccccccccccccccccccc$	uogTrQCBMP	UoGTr	fullrank	nnlm	0.3521	0.8722	0.5791	0.6034	0.3752
$ \begin{array}{cccccccccccccccccccccccccccccccccccc$	uob_runid l	UoB	rerank	nnlm	0.3124	0.8852	0.5781	0.6283	0.3786
$\begin{array}{cccccccccccccccccccccccccccccccccccc$	IUW-IKL-4K	IU_vienna	feiller	nn	0.4097	0.9185	0.5749	0.6283	0.3749
$ \begin{array}{llllllllllllllllllllllllllllllllllll$	bl basi multfld	QU bl. basi	fullronk	nnim trad	0.2704	0.8902	0.5734	0.0009	0.41//
Indersolini Iorriti Infinition Indiant Indiant <thindiant< th=""> Indiant Indiant</thindiant<>	indri-sdmf	BMIT	fullrank	trad	0.2022	0.9195	0.5629	0.0299	0.3829
$\begin{array}{c} \begin{tabular}{lllllll} blue limit in the large blue blue limit in the large blue blue blue blue blue blue blue blu$	beai classic	beai	fullrank	trad	0.3082	0.8648	0.5557	0.6420	0.3906
$ \begin{array}{cccccccccccccccccccccccccccccccccccc$	longformer 1	USI	rerank	nnlm	0.3614	0.8889	0.5520	0.6283	0.3503
$\begin{array}{cccccccccccccccccccccccccccccccccccc$	uogTr31oR	UoGTr	fullrank	nnlm	0.3257	0.8926	0.5476	0.5496	0.3468
$\begin{array}{cccccccccccccccccccccccccccccccccccc$	rterrier-expC2	bl_rmit	fullrank	trad	0.3122	0.8259	0.5475	0.6442	0.3805
uogTrT20UoGTrfullranknnlm 0.3787 0.8711 0.5453 0.5354 0.3692 RMIT_DFReeRMITfullranktrad 0.2984 0.8756 0.5431 0.6979 0.4087 rmit_indri-fdmbl_rmitfullranktrad 0.2794 0.8481 0.5416 0.6812 0.3859 q_d2q_bm25rm3anserinifullranknnlm 0.2314 0.8147 0.5407 0.6831 0.4228 rindri-bm25bl_rmitfullranktrad 0.3302 0.8572 0.5394 0.6503 0.3773 bjBcai_model1bl_bcaifullranktrad 0.2349 0.9060 0.5390 0.6669 0.3619 bl_bcai_proxbl_bcaifullranktrad 0.22763 0.8164 0.5364 0.6405 0.3766 terrier-jsklsbl_rmitfullranktrad 0.2702 0.8470 0.5328 0.6733 0.3780 rterrier-tifdfbl_rmitfullranktrad 0.2867 0.8611 0.5283 0.6061 0.3466 RMIT_DPHRMITfullranktrad 0.2687 0.8611 0.5283 0.6061 0.3466 RMIT_DPHRMITfullranktrad 0.2645 0.8541 0.5248 0.6632 0.4006 BTrun2BT.UAfullranktrad 0.2645 0.8541 0.5248 0.6632 0.4006 BTru14fullranktrad 0.2303 0.8267 0.5226 0.6634 0.3846 <	bigIR-DT-T5-R	QŪ	rerank	nnlm	0.2293	0.9407	0.5455	0.6283	0.3373
RMIT_DFRee RMIT fullrank trad 0.2984 0.8756 0.5431 0.6979 0.4087 rmit_indri-fdm b_rmit fullrank trad 0.2779 0.8481 0.5416 0.6812 0.3859 d_d2q_bm25rm3 anserini fullrank trad 0.302 0.8572 0.5394 0.6503 0.3773 bigIR-DT-T5-F QU fullrank trad 0.2201 0.8578 0.5390 0.6669 0.3619 bl_bcai_model1 b_bcai fullrank trad 0.2010 0.8578 0.5390 0.3766 terrier-jskls b_rmit fullrank trad 0.2702 0.8470 0.5328 0.6733 0.3780 rterrier-fdf b_rmit fullrank trad 0.2702 0.8470 0.5328 0.6733 0.3780 rterrier-fdf b_rmit fullrank trad 0.2117 0.8241 0.5171 0.6410 0.3734 BT-run2 BT.UA fullrank trad 0.3171 <td< td=""><td>uogTrT20</td><td>UoGTr</td><td>fullrank</td><td>nnlm</td><td>0.3787</td><td>0.8711</td><td>0.5453</td><td>0.5354</td><td>0.3692</td></td<>	uogTrT20	UoGTr	fullrank	nnlm	0.3787	0.8711	0.5453	0.5354	0.3692
$\begin{array}{cccccccccccccccccccccccccccccccccccc$	RMIT_DFRee	RMIT	fullrank	trad	0.2984	0.8756	0.5431	0.6979	0.4087
$ \begin{array}{cccccccccccccccccccccccccccccccccccc$	rmit_indri-fdm	bl_rmit	fullrank	trad	0.2779	0.8481	0.5416	0.6812	0.3859
$\begin{array}{cccccccccccccccccccccccccccccccccccc$	d_d2q_bm25rm3	anserini	fullrank	nnlm	0.2314	0.8147	0.5407	0.6831	0.4228
bigR-D1-15-F QU fullrank nnim 0.2349 0.9060 0.3390 0.6669 0.3619 bl_bcai_prox bl_bcai fullrank trad 0.2763 0.8164 0.5364 0.6405 0.3766 terrier-jskls bl_rmit fullrank trad 0.2763 0.8164 0.5342 0.6761 0.4008 rmit_indri-sdm bl_rmit fullrank trad 0.2702 0.8470 0.5328 0.6733 0.3780 rterrier-tfidf bl_rmit fullrank trad 0.2869 0.8241 0.5317 0.6410 0.3734 BIT-run2 BIT.UA fullrank trad 0.3117 0.8278 0.5280 0.6531 0.3879 d_bm25 anserini fullrank trad 0.2814 0.8521 0.5271 0.6453 0.3791 d_bm25 anserini fullrank trad 0.3033 0.8267 0.5226 0.6634 0.3884 rterrier-tfidf2 bl_rmit fullrank trad	rindri-bm25	bl_rmit	fullrank	trad	0.3302	0.8572	0.5394	0.6503	0.3773
bi_bcai_model1 bi_bcai fullrank frad 0.2763 0.8164 0.5378 0.6390 0.3774 bi_bcai_prox bi_bcai fullrank trad 0.2763 0.8164 0.5342 0.6761 0.4008 rmit_indri-sdm bi_rmit fullrank trad 0.2702 0.8204 0.5342 0.6713 0.3766 rerrier-isdf bi_rmit fullrank trad 0.2269 0.8241 0.5317 0.6410 0.3734 BIT-run2 BIT.UA fullrank trad 0.2267 0.8611 0.5283 0.6061 0.3466 RMIT_DPH RMIT fullrank trad 0.2117 0.8278 0.5280 0.6531 0.3879 d_bm25 anserini fullrank trad 0.2645 0.8541 0.5238 0.6632 0.4006 BIT-run1 BIT.UA fullrank trad 0.3033 0.8267 0.5226 0.6634 0.3884 rterrier-dph bi_rmit fullrank trad 0.3010 0.8407 0.5219 0.6287 0.3607 uogTrBaseQL17o <td>bigIR-DI-IS-F</td> <td>QU</td> <td>fullrank</td> <td>nnlm</td> <td>0.2349</td> <td>0.9060</td> <td>0.5390</td> <td>0.6669</td> <td>0.3619</td>	bigIR-DI-IS-F	QU	fullrank	nnlm	0.2349	0.9060	0.5390	0.6669	0.3619
bi_ccal_prox bi_ccal rultrank trad 0.2735 0.8164 0.5304 0.6403 0.3766 terrier-jskls bi_rmit fullrank trad 0.3190 0.8204 0.5342 0.6761 0.4008 rmit_indri-sdm bi_rmit fullrank trad 0.2702 0.8470 0.5328 0.6733 0.3780 rterrier-tfdf bi_rmit fullrank trad 0.2869 0.8241 0.5317 0.6410 0.3734 BIT-run2 BIT.UA fullrank trad 0.3117 0.8278 0.5280 0.6531 0.3879 d_bm25 anserini fullrank trad 0.2814 0.8521 0.5271 0.6453 0.3791 d_bm25rm3 anserini fullrank trad 0.2645 0.8541 0.5248 0.6632 0.4006 BIT-run1 BIT.UA fullrank trad 0.3033 0.8267 0.5226 0.6634 0.3884 rterrier-dph bi_rmit fullrank trad 0.3010 0.8407 0.5219 0.6228 0.3529 uogTrBaseQL1	bl_bcai_model1	bl_bcai	fullrank	trad	0.2901	0.8358	0.5378	0.6390	0.3774
Iterrier-iskisbi_rmitfullranktrad0.27020.82040.53220.67010.4008rmit_indri-sdmbi_rmitfullranktrad0.27020.84700.53280.67330.3780rterrier-tfildbi_rmitfullranktrad0.28690.82410.53170.64100.3734BIT-un2BIT.UAfullranknn0.26870.86110.52830.60610.3466RMIT_DPHRMITfullranktrad0.21170.82780.52800.65310.3879d_bm25anserinifullranktrad0.26450.85410.52480.66320.4006BIT-run1BIT.UAfullranktrad0.20450.88410.52480.66320.4006BIT-run1BIT.UAfullranktrad0.30330.82670.52260.66340.3846rterrier-dphbi_rmitfullranktrad0.30100.84070.52190.62870.3607uogTrBaseQL170bi_uogTrfullranktrad0.32430.82760.52030.60280.3529uogTrBaseL170bi_uogTrfullranktrad0.32430.82960.51100.66500.3784BIT-run3BIT.UAfullranktrad0.32430.82960.51100.66500.3784BIT-run3BIT.UAfullranktrad0.32430.82960.50630.60720.3267uogTrBaseDPHQbi_uogTrfullranktrad0.32110.7930<	bi_bcai_prox	bl_bcai	fullronk	trad	0.2703	0.8104	0.5304	0.0403	0.5700
Internet fullInternetInternetInternetInternetInternetInternetInternetrterrier-tfidfbl_rmitfullranktrad0.28690.82410.53170.64100.3734BIT-un2BIT.UAfullranktrad0.26870.86110.52830.60610.3466RMIT_DPHRMITfullranktrad0.21170.82780.52800.65310.3879d_bm25anserinifullranktrad0.26450.85410.52480.66320.4006BIT-run1BIT.UAfullranktrad0.20450.83890.52390.60610.3466rterrier-dphbl_rmitfullranktrad0.30330.82670.52260.66340.3884rterrier-dphbl_rmitfullranktrad0.42030.82760.52030.60280.3529uogTrBaseQL170bl_uogTrfullranktrad0.32330.82760.52030.60280.3529uogTrBaseL170bl_uogTrfullranktrad0.32430.82960.51100.66500.3784BIT-run3BIT.UAfullranktrad0.32430.82960.51100.66500.3784BIT-run3BIT.UAfullranktrad0.32430.82960.51100.66500.3784BIT-run3BIT.UAfullranktrad0.32430.82960.50630.60720.3267uogTrBaseDPHQbl_uogTrfullranktrad0.32110.7930	rmit indri-sdm	bl_rmit	fullrank	trad	0.3190	0.8204	0.5342	0.6733	0.4008
Initial BIT. UA fulliank <	rterrier-tfidf	bl_rmit	fullrank	trad	0.2702	0.8741	0.5317	0.6410	0.3734
RMIT_DPH RMIT fullrank trad 0.3117 0.8278 0.5280 0.6331 0.3879 d_bm25 anserini fullrank trad 0.2814 0.8521 0.5280 0.6531 0.3879 d_bm25m3 anserini fullrank trad 0.2645 0.8541 0.5280 0.6632 0.4006 BIT-run1 BIT.UA fullrank trad 0.3045 0.8389 0.5229 0.6061 0.3466 rterrier-dph bl_rmit fullrank trad 0.3033 0.8267 0.5226 0.6634 0.3884 rterrier-dph bl_rmit fullrank trad 0.4233 0.8267 0.5203 0.60287 0.3607 uogTrBaseQL170 bl_uogTr fullrank trad 0.4233 0.8276 0.5203 0.6028 0.3529 uogTrBaseL170 bl_uogTr fullrank trad 0.3243 0.8296 0.5110 0.6650 0.3784 BIT-run3 BIT.UA fullrank trad 0.3	BIT-run2	BITUA	fullrank	nn	0.2687	0.8611	0.5283	0.6061	0.3466
d_bm25anserinifullranktrad0.28140.85210.52710.64530.3791d_bm25rm3anserinifullranktrad0.26450.85410.52480.66320.4006BIT-run1BIT.UAfullranknn0.30450.83890.52390.60610.3466rterrier-dphbl_rmitfullranktrad0.30330.82670.52260.66340.3884rterrier-tfidf2bl_rmitfullranktrad0.42030.82670.52030.60280.3607uogTrBaseQL170bl_uogTrfullranktrad0.42330.82760.52030.60280.3529uogTrBaseL170bl_uogTrfullranktrad0.38700.79800.51200.55010.3248rterrier-dph_sdbl_rmitfullranktrad0.32430.82960.51100.66500.3784BIT-run3BIT.UAfullranktrad0.34590.80520.50630.60720.3267uogTrBaseDPHQbl_uogTrfullranktrad0.33210.79300.49980.60300.3436uogTrBaseD16bl_uogTrfullranktrad0.32210.79300.49980.60300.3428uogTrBaseDPHbl_uogTrfullranktrad0.31790.84150.48710.54950.3248uogTrBaseDPHbl_uogTrfullranktrad0.31790.84150.48710.54900.3248uogTrBaseDPHbl_uogTrfullranktrad0.	RMIT DPH	RMIT	fullrank	trad	0.3117	0.8278	0.5280	0.6531	0.3879
$ \begin{array}{cccccccccccccccccccccccccccccccccccc$	d $bm25$	anserini	fullrank	trad	0.2814	0.8521	0.5271	0.6453	0.3791
BIT-run1 BIT.UA fullrank nn 0.3045 0.8389 0.5239 0.6061 0.3466 rterrier-dph bl_rmit fullrank trad 0.3033 0.8267 0.5226 0.6634 0.3884 rterrier-dph bl_rmit fullrank trad 0.3010 0.8407 0.5219 0.6287 0.3607 uogTrBaseQL170 bl_uogTr fullrank trad 0.4233 0.8276 0.5203 0.6028 0.3529 uogTrBaseL170 bl_uogTr fullrank trad 0.3243 0.8276 0.5100 0.5501 0.3248 rterrier-dph_sd bl_rmit fullrank trad 0.3243 0.8296 0.5110 0.6650 0.3784 BIT-run3 BIT.UA fullrank trad 0.3459 0.8052 0.5063 0.6072 0.3267 uogTrBaseDHQ bl_uogTr fullrank trad 0.3321 0.7930 0.4998 0.6030 0.3436 uogTrBaseDFHQ bl_uogTr fullrank trad <td>d_bm25rm3</td> <td>anserini</td> <td>fullrank</td> <td>trad</td> <td>0.2645</td> <td>0.8541</td> <td>0.5248</td> <td>0.6632</td> <td>0.4006</td>	d_bm25rm3	anserini	fullrank	trad	0.2645	0.8541	0.5248	0.6632	0.4006
rterrier-dphbl_rmitfullranktrad0.30330.82670.52260.66340.3884rterrier-tfidf2bl_rmitfullranktrad0.30100.84070.52190.62870.3607uogTrBaseQL170bl_uogTrfullranktrad0.42330.82760.52030.60280.3529uogTrBaseL170bl_uogTrfullranktrad0.38700.79800.51200.55010.3248rterrier-dph_sdbl_rmitfullranktrad0.32430.82960.51100.66500.3784BIT-run3BIT.UAfullranknn0.26960.82960.50630.60720.3267uogTrBaseDPHQbl_uogTrfullranktrad0.33210.79300.49980.60300.3436uogTrBaseL16bl_uogTrfullranktrad0.31290.49640.54950.3248uogTrBaseDPHbl_uogTrfullranktrad0.31790.84150.48710.54900.3070nlm-bm25-prf-2NLMfullranktrad0.27320.80990.47050.52180.2912nlm-bm25-prf-1NLMfullranktrad0.23900.80860.46750.49580.2720mpii_run3mpiireranknnlm0.14990.63880.32860.62830.2587	BIT-run1	BIT.UA	fullrank	nn	0.3045	0.8389	0.5239	0.6061	0.3466
rterrier-tfidf2bl_rmitfullranktrad0.30100.84070.52190.62870.3607uogTrBaseQL17obl_uogTrfullranktrad0.42330.82760.52030.60280.3529uogTrBaseL17obl_uogTrfullranktrad0.38700.79800.51200.55010.3248rterrier-dph_sdbl_rmitfullranktrad0.32430.82960.51100.66500.3784BIT-run3BIT.UAfullranktrad0.32430.82960.50630.60720.3267uogTrBaseDPHQbl_uogTrfullranktrad0.33210.79300.49980.60300.3461uogTrBaseL16bl_uogTrfullranktrad0.31210.79300.49980.60300.3436uogTrBaseDPHbl_uogTrfullranktrad0.31790.84150.48710.54950.3248uogTrBaseDPHbl_uogTrfullranktrad0.31790.84150.48710.54900.3070nlm-bm25-prf-2NLMfullranktrad0.27320.80990.47050.52180.2912nlm-bm25-prf-1NLMfullranktrad0.23900.80860.46750.49580.2720mpii_run3mpiireranknnlm0.14990.63880.32860.62830.2587	rterrier-dph	bl_rmit	fullrank	trad	0.3033	0.8267	0.5226	0.6634	0.3884
uogTrBaseQL17o bl_uogTr fullrank trad 0.4233 0.8276 0.5203 0.6028 0.3529 uogTrBaseL17o bl_uogTr fullrank trad 0.3870 0.7980 0.5120 0.5501 0.3248 rterrier-dph_sd bl_rmit fullrank trad 0.3243 0.8296 0.5110 0.6650 0.3784 BIT-run3 BIT.UA fullrank trad 0.3243 0.8296 0.5063 0.6072 0.3267 uogTrBaseDPHQ bl_uogTr fullrank trad 0.3459 0.8052 0.5052 0.6041 0.3461 uogTrBaseQL16 bl_uogTr fullrank trad 0.3321 0.7930 0.4998 0.6030 0.3436 uogTrBaseDPH bl_uogTr fullrank trad 0.3121 0.7930 0.4998 0.6030 0.3436 uogTrBaseDPH bl_uogTr fullrank trad 0.3179 0.8415 0.4871 0.5495 0.3248 uogTrBaseDPH bl_uogTr fullrank	rterrier-tfidf2	bl_rmit	fullrank	trad	0.3010	0.8407	0.5219	0.6287	0.3607
uogTrBaseL17o bl_uogTr fullrank trad 0.3870 0.7980 0.5120 0.5501 0.3248 rterrier-dph_sd bl_rmit fullrank trad 0.3243 0.8296 0.5110 0.6650 0.3784 BIT-run3 BIT.UA fullrank trad 0.3243 0.8296 0.5063 0.6072 0.3267 uogTrBaseDPHQ bl_uogTr fullrank trad 0.3459 0.8052 0.5052 0.6041 0.3461 uogTrBaseQL16 bl_uogTr fullrank trad 0.3210 0.7930 0.4998 0.6030 0.3436 uogTrBaseDPH bl_uogTr fullrank trad 0.3210 0.7930 0.4998 0.6030 0.3436 uogTrBaseDPH bl_uogTr fullrank trad 0.3219 0.4964 0.5495 0.3248 uogTrBaseDPH bl_uogTr fullrank trad 0.3179 0.8415 0.4871 0.5490 0.3070 nlm-bm25-prf-2 NLM fullrank trad 0.2	uogTrBaseQL17o	bl_uogTr	fullrank	trad	0.4233	0.8276	0.5203	0.6028	0.3529
rterrier-dph_sdbl_rmitfullranktrad0.32430.82960.51100.66500.3784BIT-run3BIT.UAfullranknn0.26960.82960.50630.60720.3267uogTrBaseDPHQbl_uogTrfullranktrad0.34590.80520.50520.60410.3461uogTrBaseQL16bl_uogTrfullranktrad0.33210.79300.49980.60300.3436uogTrBaseL16bl_uogTrfullranktrad0.31790.84150.48710.54950.3248uogTrBaseDPHbl_uogTrfullranktrad0.31790.84150.48710.54900.3070nlm-bm25-prf-2NLMfullranktrad0.23900.80860.46750.49580.2720mpii_run3mpiireranknnlm0.14990.63880.32860.62830.2587	uogTrBaseL17o	bl_uogTr	fullrank	trad	0.3870	0.7980	0.5120	0.5501	0.3248
B11-run3 B11.UA fullrank nn 0.2696 0.8296 0.5063 0.6072 0.3267 uogTrBaseDPHQ bl_uogTr fullrank trad 0.3459 0.8052 0.5052 0.6041 0.3461 uogTrBaseQL16 bl_uogTr fullrank trad 0.3321 0.7930 0.4998 0.6030 0.3436 uogTrBaseL16 bl_uogTr fullrank trad 0.3179 0.8219 0.4964 0.5495 0.3248 uogTrBaseDPH bl_uogTr fullrank trad 0.3179 0.8415 0.4861 0.5490 0.3070 nlm-bm25-prf-2 NLM fullrank trad 0.2732 0.8099 0.4705 0.5218 0.2912 nlm-bm25-prf-1 NLM fullrank trad 0.2390 0.8086 0.4675 0.4958 0.2720 mpii_run3 mpii reank nnlm 0.1499 0.6388 0.3286 0.6283 0.2587	rterrier-dph_sd	bl_rmit	fullrank	trad	0.3243	0.8296	0.5110	0.6650	0.3784
uog1rBaseDr1Q bl_uog1r rullrank trad 0.3459 0.8052 0.5052 0.6041 0.3461 uogTrBaseQL16 bl_uogTr fullrank trad 0.3321 0.7930 0.4998 0.6030 0.3436 uogTrBaseL16 bl_uogTr fullrank trad 0.3022 0.8219 0.4964 0.5495 0.3248 uogTrBaseDPH bl_uogTr fullrank trad 0.3179 0.8415 0.4871 0.5490 0.3070 nlm-bm25-prf-2 NLM fullrank trad 0.2732 0.8099 0.4705 0.5218 0.2912 nlm-bm25-prf-1 NLM fullrank trad 0.2390 0.8086 0.4675 0.4958 0.2720 mpii_run3 mpii reank nnlm 0.1499 0.6388 0.3286 0.6283 0.2587	BIT-run3	BII.UA	fullrank	nn	0.2696	0.8296	0.5063	0.6072	0.3267
uog ITBaseQL10 buog IT Iulirank irad 0.321 0.7950 0.4998 0.6030 0.3436 uog TrBaseL16 bl_uogTr fullrank trad 0.3062 0.8219 0.4964 0.5495 0.3248 uog TrBaseDPH bl_uogTr fullrank trad 0.3179 0.8415 0.4871 0.5490 0.3070 nlm-bm25-prf-2 NLM fullrank trad 0.2732 0.8099 0.4705 0.5218 0.2912 nlm-bm25-prf-1 NLM fullrank trad 0.2390 0.8086 0.4675 0.4958 0.2720 mpii_run3 mpii rerank nnlm 0.1499 0.6388 0.3286 0.6283 0.2587	uog IrBaseDPHQ	bl_uog1r	fullrank	trad	0.3459	0.8052	0.5052	0.6041	0.3461
uog11base170 b1_uog11 fullmank frad 0.3002 0.8219 0.4904 0.3495 0.3248 uog17BaseDPH b1_uog1r fullrank trad 0.3179 0.8415 0.4871 0.5490 0.3070 nlm-bm25-prf-2 NLM fullrank trad 0.2732 0.8099 0.4705 0.5218 0.2912 nlm-bm25-prf-1 NLM fullrank trad 0.2390 0.8086 0.4675 0.4958 0.2720 mpii_run3 mpii rerank nnlm 0.1499 0.6388 0.3286 0.6283 0.2587	uog IrBaseQL16	ol_uog1r	fullmont	trad	0.3321	0.7930	0.4998	0.6030	0.3430
augrinuscont olagon fulliank	uog 11 DaseL 10	bl uogTr	fullronk	u au trad	0.3002	0.0219	0.4904	0.3493	0.3248
Initiani	nlm_hm25_prf_2	NI M	fullrank	u au trad	0.3179	0.0415	0.4071	0.5490	0.3070
mpii_run3 mpii rerank nnlm 0.1499 0.6388 0.3286 0.6283 0.2587	nlm-bm25-prf-1	NLM	fullrank	trad	0.2390	0.8086	0.4675	0.4958	0.2720
	mpii_run3	mpii	rerank	nnlm	0.1499	0.6388	0.3286	0.6283	0.2587

	run	group	subtask	neural	RR (MS)	RR	NDCG@10	NCG@1000	AP
$ \begin{array}{cccccccccccccccccccccccccccccccccccc$	pash_r3	PASH	rerank	nnlm	0.3678	0.9147	0.8031	0.7056	0.5445
	pash_r2	PASH	rerank	nnlm	0.3677	0.9023	0.8011	0.7056	0.5420
	pash_f3	PASH	fullrank	nnlm	0.3506	0.8885	0.8005	0.7255	0.5504
	pash_f1	PASH	fullrank	nnlm	0.3598	0.8699	0.7956	0.7209	0.5455
$ \begin{array}{c} p.d2q.m32.duo \\ p.d2q.$	pash_f2	PASH	fullrank	nnlm	0.3603	0.8931	0.7941	0.7132	0.5389
	p_d2q_bm25_duo	h2oloo	fullrank	nnlm	0.3838	0.8798	0.7837	0.8035	0.5609
	p_d2q_rm3_duo	h2oloo	fullrank	nnlm	0.3795	0.8798	0.7821	0.8446	0.5643
$ \begin{array}{cccccc} CoRT-electra HSRM-LAVIS fullrank nnlm 0.4039 0.8703 0.7566 0.752 0.5399 \\ RMIT-Baur RMIT fullrank nnlm 0.3990 0.8447 0.7536 0.7682 0.5121 \\ pash,rl PASH rerank nnlm 0.3691 0.8440 0.7488 0.8211 0.5245 \\ NLE_pr3 NLE fullrank nnlm 0.3651 0.8440 0.7488 0.8211 0.5245 \\ pinganNLP2 pinganNLP rerank nnlm 0.3653 0.8506 0.7352 0.7056 0.4881 \\ pinganNLP1 pinganNLP rerank nnlm 0.3653 0.8593 0.7343 0.7056 0.4898 \\ NLE_pr2 NLE fullrank nnlm 0.3654 0.8593 0.7343 0.7056 0.4899 \\ pigR-BER-R QU rerank nnlm 0.3654 0.8551 0.7325 0.6938 0.5117 \\ NLE_pr1 NLE fullrank nnlm 0.3630 0.8502 0.7121 0.7055 0.4899 \\ pigR-BER-R QU rerank nnlm 0.3540 0.8562 0.7121 0.7056 0.4895 \\ fr. pass,roberta BITEM fullrank nnlm 0.3540 0.8638 0.7173 0.8093 0.5004 \\ rr-pass-roberta BITEM fullrank nnlm 0.3540 0.8638 0.7173 0.8093 0.5004 \\ rr-pass-roberta BITEM rerank nnlm 0.3701 0.8635 0.7169 0.7066 0.4823 \\ beal, bertl. pass bcai fullrank nnlm 0.3570 0.86453 0.7169 0.7066 0.4784 2 \\ rr-ass-roberta BITEM rerank nnlm 0.3570 0.8608 0.7138 0.7056 0.4845 2 \\ pigR-5T-FF QU fullrank nnlm 0.3574 0.8668 0.7138 0.7056 0.4784 2 \\ rr-ass-roberta BITEM rerank nnlm 0.3500 0.8504 0.7190 0.7066 0.4784 2 \\ rr-ass-roberta BITEM rerank nnlm 0.3574 0.8668 0.7138 0.7056 0.4784 2 \\ rr-ass-roberta BITEM rerank nnlm 0.3574 0.8668 0.7138 0.7056 0.4784 2 \\ rr-ass-roberta BITEM rerank nnlm 0.3574 0.8668 0.7138 0.7056 0.4374 2 \\ rr-ass-roberta BITEM rerank nnlm 0.3500 0.8507 0.7113 0.7447 0.4866 0.4784 2 \\ rr-ass-roberta BITEM rerank nnlm 0.3560 0.8507 0.7138 0.7056 0.4341 2 \\ rerank nnlm 0.3540 0.8573 0.7161 0.7056 0.4341 2 \\ rr-ass-st-3 NLM fullrank nnlm 0.3540 0.8578 0.7073 0.8393 0.5001 \\ nlm-ens-bs-3 NLM fullrank nnlm 0.3420 0.8579 0.7034 0.4350 \\ nlm-ens-bs-3 NLM fullrank nnlm 0.3450 0.777 0.66610 0.7056 0.4341 \\ retemb.nlm.0_2 2 \\ rerank nnlm 0.3420 0.8570 0.6610 0.7056 0.4350 \\ nlm-sertm-PIL NLM fullrank nnlm 0.3480 0.7770 0.6619 0.7056 0.4319 \\ p.242_bm25m 3 anserini fullrank nnlm 0.2420 0.8570 0.6641 0.6975 0.4750 0.4179 \\ p.242_bm25m 3 anserini fullrank nnlm 0$	p_bm25rm3_duo	h2oloo	fullrank	nnlm	0.3814	0.8759	0.7583	0.7939	0.5355
$ \begin{array}{c c c c c c c c c c c c c c c c c c c $	CoRT-electra	HSRM-LAVIS	fullrank	nnlm	0.4039	0.8703	0.7566	0.8072	0.5399
	RMIT-Bart	RMIT	fullrank	nnlm	0.3990	0.8447	0.7536	0.7682	0.5121
$ \begin{array}{llllllllllllllllllllllllllllllllllll$	pash_r1	PASH	rerank	nnlm	0.3622	0.8675	0.7463	0.7056	0.4969
	NLE_pr3	NLE	fullrank	nnlm	0.3691	0.8440	0.7458	0.8211	0.5245
$ \begin{array}{cccccccccccccccccccccccccccccccccccc$	pinganNLP2	pinganNLP	rerank	nnlm	0.3579	0.8602	0.7368	0.7056	0.4881
$\begin{split} \hline pinganNLP1 & pinganNLP & rerank nnlm 0.3553 0.8593 0.7343 0.7056 0.44896 \\ NLE_pr1 & NLE & fullrank nnlm 0.3658 0.8454 0.7341 0.6938 0.5017 \\ NLE_pr1 & NLE & fullrank nnlm 0.3634 0.8551 0.7325 0.6938 0.5050 \\ 1 & nvidia_ai_apps & rerank nnlm 0.3709 0.8691 0.7325 0.6938 0.5005 \\ pigIR-BERT-R & QU & rerank nnlm 0.4040 0.8562 0.7201 0.7056 0.4899 \\ pigIR-DCTT5-F & QU & fullrank nnlm 0.3580 0.8769 0.7192 0.7982 0.4990 \\ bigIR-DCT5-F & QU & fullrank nnlm 0.3701 0.8635 0.7115 0.8093 0.5004 \\ rr-pass-roberta & BITEM & rerank nnlm 0.3711 0.8635 0.71151 0.7990 0.4641 \\ bigIR-T5-R & QU & retrank nnlm 0.3574 0.8668 0.7113 0.7447 0.4686 \\ 2 & nvidia_ai_apps fullrank nnlm 0.3574 0.8668 0.7138 0.7056 0.4784 \\ 2 & nvidia_ai_apps fullrank nnlm 0.3574 0.8668 0.7138 0.7056 0.4784 \\ 1nlm-en-Sbt-2 & NLM & fullrank nnlm 0.3420 0.8579 0.7013 0.7447 0.4866 \\ nlm-en-Sbt-2 & NLM & fullrank nnlm 0.3420 0.8579 0.7013 0.8493 0.5001 \\ nlm-en-Sbt-2 & NLM & fullrank nnlm 0.3420 0.8579 0.7034 0.8393 0.5011 \\ bigIR-T5xp-T5-F & QU & fullrank nnlm 0.3699 0.7785 0.6623 0.7056 0.4350 \\ nlm-en-Sbt-3 & NLM & fullrank nnlm 0.3699 0.7785 0.6613 0.7056 0.4531 \\ nlm-en-Sbt-3 & NLM & fullrank nnlm 0.3699 0.7785 0.6613 0.7056 0.4350 \\ nlm-bert-rr & NLM & fullrank nnlm 0.3699 0.7785 0.6618 0.07056 0.4139 \\ nlm-en-Sbt-3 & NLM & fullrank nnlm 0.3445 0.8603 0.6648 0.6927 0.4265 \\ nlm-bert-r & NLM & fullrank nnlm 0.3445 0.8603 0.6618 0.6927 0.4265 \\ nlm-bert-S & NLM & fullrank nnlm 0.3445 0.8603 0.6618 0.6927 0.4265 \\ nlm-bert-r & NLM & fullrank nnlm 0.2366 0.7677 0.6662 0.7056 0.4139 \\ nlm-aschst-3 & NLM & fullrank nnlm 0.2369 0.7854 0.6187 0.8035 0.7076 0.4164 \\ TUW-TK-2Layer & U_Vienna & rerank nnlm 0.2375 0.7564 0.6618 0.6927 0.3260 \\ nlm-bert-fall & Lamsterdam & rerank nnlm 0.2364 0.7386 0.6187 0.8035 0.7776 0.2992 \\ bert_6 & UAmsterdam & rerank nnlm 0.2412 0.8112 0.5926 0.6002 0.3380 \\ bei_ac_iamscrini & fullrank trad 0.1636 0.6439 0.7056 0.4169 0.7056 0.3760 \\ berta_i-DM2 & HSRM-LAVIS & fullrank trad 0.1648 0.6438 0.4985 0.7649 0.3135 \\ berta_clas$	pinganNLP3	pinganNLP	rerank	nnlm	0.3653	0.8586	0.7352	0.7056	0.4918
$ \begin{array}{c c c c c c c c c c c c c c c c c c c $	pinganNLP1	pinganNLP	rerank	nnlm	0.3553	0.8593	0.7343	0.7056	0.4896
	NLE_pr2	NLE	fullrank	nnlm	0.3658	0.8454	0.7341	0.6938	0.5117
$ l = nvidia_ai_apps retank nnlm 0.3709 0.8691 0.7271 0.7056 0.4889 \\ bigIR-BERT-R QU retark nnlm 0.4040 0.8562 0.7201 0.7056 0.4849 \\ fr_pass_roberta BITEM fullrank nnlm 0.3540 0.8638 0.7192 0.7982 0.4990 \\ bigIR-DCT-T5-F QU fullrank nnlm 0.3540 0.8635 0.7169 0.7056 0.4842 \\ beai_bert_pass bcai fullrank nnlm 0.3715 0.8453 0.7151 0.7090 0.4641 \\ bigIR-TS-R QU retrank nnlm 0.3716 0.8635 0.7113 0.7447 0.4866 \\ bigIR-TS-BERT-F QU fullrank nnlm 0.3560 0.8507 0.7113 0.7447 0.4866 \\ bigIR-TS-BERT-F QU fullrank nnlm 0.3560 0.8507 0.7113 0.7447 0.4866 \\ bigIR-TS-beta_N V_LM fullrank nnlm 0.3560 0.8507 0.7113 0.7447 0.4866 \\ bigIR-TS-beta_N V_LM fullrank nnlm 0.3542 0.8579 0.7034 0.8393 0.5001 \\ nlm-ens-bst-3 NLM fullrank nnlm 0.3542 0.8579 0.7034 0.8393 0.5001 \\ nlm-ens-bst-3 NLM fullrank nnlm 0.3542 0.8579 0.7034 0.8393 0.5001 \\ nlm-ens-bst-3 NLM fullrank nnlm 0.3542 0.8579 0.7034 0.8393 0.5001 \\ nlm-ens-bst-3 NLM fullrank nnlm 0.3595 0.8491 0.6662 0.7056 0.44341 \\ relemb_nlm_0_2 UAmsterdam retank nnlm 0.3285 0.7677 0.6662 0.7056 0.44350 \\ nlm-frun-betr NLM fullrank nnlm 0.2357 0.7634 0.6610 0.7056 0.41350 \\ nlm-frun-betr NLM fullrank nnlm 0.2375 0.7634 0.6619 0.7056 0.4164 \\ TUW-TK-Sparse TU_Vienna retank nn 0.3188 0.7970 0.66110 0.7056 0.4164 \\ p_d2a_bm25 anserini fullrank nnlm 0.2434 0.7386 0.6147 0.8035 0.4074 \\ p_d2a_bm25 maserini fullrank nnlm 0.2434 0.7386 0.6149 0.7056 0.3760 \\ CoRT-standalone HSRM-LAVIS fullrank nnlm 0.2201 0.8372 0.5667 0.7430 0.3380 \\ bea_class_pass bcai fullrank trad 0.1653 0.6277 0.5992 0.7430 0.3391 \\ b_bca_mdl_v t b_bcai fullrank trad 0.1654 0.6277 0.5992 0.7302 0.3611 \\ CoRT-standalone HSRM-LAVIS fullrank trad 0.1634 0.7986 0.6149 0.7056 0.3760 \\ b_bca_andl_v t b_bcai fullrank trad 0.1654 0.6498 0.5003 0.7778 0.2989 \\ terrier-BM25 b_l_mit fullrank trad 0.1634 0.6488 0.7049 0.3135 \\ b_bca_andl_v t b_bcai fullrank trad 0.1634 0.6498 0.7722 0.22870 \\ p_bm25m3 anserini fullrank trad 0.1634 0.6498 0.7748 0.2380 \\ b_bca_andl_v t b_bcai fullrank trad 0.1634 0.6498 0.7749 0.3137 \\ b_b_b_$	NLE_pr1	NLE	fullrank	nnlm	0.3634	0.8551	0.7325	0.6938	0.5050
bigIR-BERT-R QU rerank nnlm 0.4040 0.8562 0.7201 0.7056 0.74845 fr_pass.roberta BITEM fullrank nnlm 0.3540 0.8638 0.7173 0.8093 0.5004 rrspass-roberta BITEM rerank nnlm 0.3711 0.8635 0.7151 0.7090 0.4641 bigIR-TS-R QU rerank nnlm 0.3574 0.8668 0.7113 0.7474 0.4866 bigIR-TS-BERT-F QU fullrank nnlm 0.3540 0.8579 0.7034 0.8393 0.5101 nlm-ens-bst-2 NLM fullrank nnlm 0.3420 0.8579 0.7034 0.8393 0.5001 nlm-ens-bst-3 NLM fullrank nnlm 0.3422 0.8203 0.6631 0.7196 0.4526 nlm-bert-rr NLM fullrank nnlm 0.3455 0.7677 0.66612 0.7056 0.4350 nlm-bert-rr NLM fullrank nnlm 0.2348	1	nvidia_ai_apps	rerank	nnlm	0.3709	0.8691	0.7271	0.7056	0.4899
$ fr_pass_roberta BITEM fullrank nnlm 0.3580 0.8769 0.7192 0.7982 0.4990 bigIR-DCTF3-F QU fullrank nnlm 0.3701 0.8635 0.7169 0.7056 0.4823 bcai_bertl_pass bcai fullrank nnlm 0.3711 0.8453 0.7151 0.7990 0.4641 bigIR-T5-R QU rerank nnlm 0.3774 0.8668 0.7138 0.7065 0.4784 2 mvidia_ai_apps fullrank nnlm 0.3574 0.8668 0.7138 0.7065 0.4784 2 mvidia_ai_apps fullrank nnlm 0.3570 0.8668 0.7138 0.7067 0.4784 bigIR-T5-BERT-F QU fullrank nnlm 0.3420 0.8579 0.7073 0.8393 0.5101 bigIR-T5xp-T5-F QU fullrank nnlm 0.3420 0.8579 0.7034 0.8393 0.5101 nlm-ens-bst-2 NLM fullrank nnlm 0.3420 0.8579 0.7034 0.8393 0.5101 nlm-ens-bst-3 NLM fullrank nnlm 0.3420 0.8579 0.7034 0.8393 0.5001 nlm-ens-bst-3 NLM fullrank nnlm 0.3420 0.8579 0.7054 0.4526 nlm-bert-r NLM fullrank nnlm 0.3452 0.8030 0.6694 0.7190 0.4598 nlm-prtn-bert NLM fullrank nnlm 0.2455 0.7677 0.6662 0.7056 0.4331 relemb_mlm_0_2 UAmsterdam rerank nn lm 0.3488 0.7970 0.6610 0.7056 0.4330 nlm-prtn-bert NLM fullrank nnlm 0.3445 0.8603 0.6648 0.6927 0.4265 TUW-TK-Sparse TU_Vienna rerank nn 0.3078 0.7670 0.6610 0.7056 0.41179 pd2q_bm25 anserini fullrank nnlm 0.2445 0.7366 0.61187 0.8035 0.4074 p.d2q_bm25 anserini fullrank nnlm 0.2440 0.7386 0.6149 0.7056 0.37660 CoRT-bm25 HSRM-LAVIS fullrank nnlm 0.2210 0.8372 0.5992 0.8072 0.3611 CoRT-standalone HSRM-LAVIS fullrank nnlm 0.2412 0.8112 0.5060 0.7430 0.3374 bl_bcai_md1_vt bl_bcai fullrank trad 0.1563 0.6277 0.7360 0.6149 0.7356 0.3308 bl_bcai_dl1_vt bl_bcai fullrank trad 0.1564 0.6379 0.5667 0.7430 0.33374 bl_bcai_adl1_vs bl_bcai fullrank trad 0.1454 0.5094 0.4935 0.8175 0.3199 indri-fdm bl_rmit fullrank trad 0.1563 0.6277 0.4560 0.7430 0.3374 bl_bcai_dl1_vs bl_bcai fullrank trad 0.1454 0.5094 0.4935 0.8175 0.3199 p.m25 maserini fullrank trad 0.1454 0.5094 0.4935 0.8175 0.3199 p.m25 maserini fullrank trad 0.1563 0.6277 0.7356 0.7430 0.3374 $	bigIR-BERT-R	QU	rerank	nnlm	0.4040	0.8562	0.7201	0.7056	0.4845
bigIR-DCT-T5-F QU fullrank nnlm 0.3540 0.8638 0.7173 0.8093 0.5004	fr_pass_roberta	BITEM	fullrank	nnlm	0.3580	0.8769	0.7192	0.7982	0.4990
$ \begin{array}{cccccccccccccccccccccccccccccccccccc$	bigIR-DCT-T5-F	QU	fullrank	nnlm	0.3540	0.8638	0.7173	0.8093	0.5004
$ beai_bertl_pass beai fullrank nnlm 0.3715 0.8453 0.7151 0.7990 0.4641 $	rr-pass-roberta	BITEM	rerank	nnlm	0.3701	0.8635	0.7169	0.7056	0.4823
bigIR-T5-R QU rerank nnlm 0.3574 0.8668 0.7138 0.7056 0.4784 0.4866 0.5507 0.7113 0.7147 0.4866 0.4784 0.3560 0.8507 0.7113 0.7147 0.4866 0.4866 0.5118 0.5207 0.7034 0.8393 0.5101 0.5118 0.5207 0.7034 0.8393 0.5101 0.5118 0.5207 0.7034 0.8393 0.5001 0.5118 0.5207 0.7034 0.8393 0.5001 0.5320 0.5512 0.520 0.520 0.523 0.6934 0.7190 0.4598 0.4598 0.5320 0.6934 0.7190 0.4598 0.4598 0.542 0.555 0.6721 0.7056 0.4591 0.3450 0.7577 0.6662 0.7056 0.4541 0.4598 0.7677 0.6662 0.7056 0.4351 0.7057 0.4350 0.7551 0.7057 0.4350 0.7551 0.7056 0.4351 0.7057 0.4350 0.7654 0.6539 0.7056 0.4164 0.4351 0.7057 0.7654 0.6539 0.7056 0.4164 0.4704 0.5352 0.7654 0.6539 0.7056 0.4169 0.3075 0.7654 0.6539 0.7056 0.4179 0.242 0.5757 0.7326 0.6187 0.8351 0.4074 0.2848 0.7424 0.6172 0.8391 0.4295 0.4764 0.4792 0.4265 0.4724 0.6172 0.8391 0.4295 0.4764 0.7386 0.6149 0.7056 0.3760 0.6767 0.5667 0.7330 0.3760 0.577 0.5667 0.7430 0.3380 0.527 0.3266 0.6020 0.3308 0.5267 0.3386 0.6149 0.7056 0.3760 0.5776 0.5667 0.7430 0.3380 0.527 0.5922 0.8072 0.3611 0.3240 0.7386 0.6149 0.7056 0.3760 0.5776 0.5667 0.7430 0.3380 0.527 0.5092 0.7430 0.3380 0.527 0.5607 0.7430 0.3380 0.527 0.5092 0.7430 0.3380 0.528 0.503 0.7778 0.2989 0.5011 0.0068 0.4985 0.7749 0.2484 0.1576 0.566 0.4912 0.7741 0.2961 0.4745 0.5585 0.4796 0.7428 0.2856 0.5025 0.5866 0.4912 0.7741 0.2961 0.4758 0.5655 0.4796 0.7428 0.2856 0.5025 0.4796	bcai_bertl_pass	bcai	fullrank	nnlm	0.3715	0.8453	0.7151	0.7990	0.4641
$\begin{array}{cccccccccccccccccccccccccccccccccccc$	bigIR-T5-R	QU	rerank	nnlm	0.3574	0.8668	0.7138	0.7056	0.4784
$ bigIR-T5-BERF-F QU fullrank nnlm 0.3916 0.8478 0.7073 0.8393 0.5101 \\ bigIR-T5xp-T5-F QU fullrank nnlm 0.3420 0.8579 0.7034 0.8393 0.5001 \\ nlm-ens-bst-2 NLM fullrank nnlm 0.3542 0.8203 0.6934 0.7190 0.4598 \\ nlm-ens-bst-3 NLM fullrank nnlm 0.3195 0.8491 0.6803 0.7594 0.4526 \\ nlm-bert-rr NLM fullrank nnlm 0.3699 0.7785 0.6721 0.7056 0.4341 \\ relemb_mlm_0_2 UAmsterdam rerank nnlm 0.2856 0.7677 0.6662 0.7056 0.4351 \\ nlm-prfun-bert NLM fullrank nnlm 0.3455 0.7677 0.66610 0.7056 0.44361 \\ ruW-TK-Sparse TU_Vienna rerank nn 0.3188 0.7970 0.6610 0.7056 0.4164 \\ TUW-TK-2Layer TU_Vienna rerank nn 0.3075 0.7654 0.6539 0.7056 0.4179 \\ p_d2q_bm25rm 3 anserini fullrank nnlm 0.2757 0.7326 0.6187 0.8035 0.4074 \\ p_d2q_bm25rm 3 anserini fullrank nnlm 0.3240 0.7386 0.6149 0.7056 0.3760 \\ CoRT-bm25 HSRM-LAVIS fullrank nnlm 0.2201 0.8372 0.5992 0.8072 0.3611 \\ CoRT-standalone HSRM-LAVIS fullrank nnlm 0.2412 0.8112 0.5926 0.6002 0.3308 \\ bl_bcai_mdl_v s b_bcai fullrank trad 0.1798 0.6498 0.5003 0.7743 0.3374 \\ bl_bcai_mdl_v s b_bcai fullrank trad 0.1636 0.6277 0.5092 0.7430 0.3374 \\ bl_bcai_mdl_v s b_bcai fullrank trad 0.1636 0.6498 0.5003 0.7778 0.2989 \\ terrier-InL2 bl_rmit fullrank trad 0.1636 0.6498 0.5003 0.7778 0.2989 \\ terrier-InL2 bl_rmit fullrank trad 0.1636 0.6498 0.7572 0.3021 \\ DLH_d_5_L25 RMTT fullrank trad 0.1636 0.6498 0.7572 0.3021 \\ DLH_d_5_L25 RMTT fullrank trad 0.1636 0.6498 0.7772 0.3021 \\ DLH_d_5_L25 RMTT fullrank trad 0.1646 0.6436 0.4985 0.7649 0.3135 \\ terrier-BM25 bl_rmit fullrank trad 0.1630 0.6585 0.4796 0.7428 0.2856 \\ bm25 bert_token UAmsterdam fullrank trad 0.1640 0.6585 0.4796 0.7428 0.2856 \\ bm25 maserini fullrank trad 0.1640 0.6239 0.4822 0.7726 0.2870 \\ p_bm25rm3 anserini fullrank trad 0.1786 0.6585 0.4796 0.7428 0.2856 \\ bm25 bert_token UAmsterdam fullrank trad 0.1786 0.6585 0.4796 0.7428 0.2856 \\ bm25 bert_bcen Cerank nnlm 0.0202 0.2785 0.2767 0.7056 0.2112 $	2	nvidia_ai_apps	fullrank	nnlm	0.3560	0.8507	0.7113	0.7447	0.4866
$ bigIR-T5xp-T5-F QU fullrank nnlm 0.3420 0.8579 0.7034 0.8393 0.5001 nlm-ens-bst-2 NLM fullrank nnlm 0.3542 0.8203 0.6934 0.7190 0.4598 nlm-ens-bst-3 NLM fullrank nnlm 0.3195 0.8491 0.6803 0.7594 0.4526 nlm-bert-rr NLM rerank nnlm 0.3699 0.7785 0.6721 0.7056 0.4350 nlm-pfun-bert NLM fullrank nnlm 0.2856 0.7677 0.6662 0.7056 0.4350 nlm-pfun-bert NLM fullrank nnlm 0.3445 0.8603 0.6648 0.6927 0.4265 TUW-TK-Sparse TU_Vienna rerank nn 0.3188 0.7970 0.6610 0.7056 0.4164 TUW-TK-2Layer TU_Vienna rerank nn 0.3075 0.7654 0.6539 0.7056 0.4179 p_d2q_bm25 anserini fullrank nnlm 0.2757 0.7326 0.6187 0.8035 0.4074 p_d2q_bm25 anserini fullrank nnlm 0.2248 0.7424 0.6172 0.8391 0.4295 bert_6 UAmsterdam rerank nn 0.3240 0.7386 0.6149 0.7056 0.3760 CoRT-shm25 HSRM-LAVIS fullrank nnlm 0.2210 0.8372 0.5992 0.8072 0.3611 CoRT-standalone HSRM-LAVIS fullrank nnlm 0.2412 0.8112 0.5926 0.6002 0.3308 bl_bcai_mdll_vt bl_bcai fullrank trad 0.1563 0.6277 0.5067 0.7430 0.3380 beat_class_pass bcai fullrank trad 0.1563 0.6277 0.5092 0.7430 0.3380 bcai_class_pass bcai fullrank trad 0.1631 0.6186 0.4988 0.7572 0.3021 DLFai fullrank trad 0.1631 0.6186 0.4988 0.7572 0.3021 DLH_d_5_t_25 RMIT fullrank trad 0.1631 0.6186 0.4988 0.7572 0.3021 DLH_d_5_t_25 RMIT fullrank trad 0.1630 0.6277 0.5092 0.77430 0.33094 indri-fdm bl_rmit fullrank trad 0.1631 0.6186 0.4988 0.7572 0.3021 DLH_d_5_t_25 RMIT fullrank trad 0.1630 0.6239 0.4822 0.7726 0.2870 p_bm25rm3 anserini fullrank trad 0.1630 0.6239 0.4822 0.7726 0.2870 p_bm25rm3 anserini fullrank trad 0.1760 0.5667 0.7438 0.3135 bl_rmit fullrank trad 0.1600 0.6239 0.4822 0.7726 0.2870 p_bm25 m3 anserini fullrank trad 0.1760 0.5667 0.4671 0.7353 0.3278 TF_IDF_d_2_t_02 RMIT fullrank trad 0.1760 0.5409 0.4486 0.7169 0.2506 terrier-DPH bl_rmit fullrank trad 0.1760 0.5409 0.4686 0.7169 0.2606 terrier-DPH bl_rmit fullrank tra$	bigIR-T5-BERT-F	QU	fullrank	nnlm	0.3916	0.8478	0.7073	0.8393	0.5101
$ \begin{array}{llllllllllllllllllllllllllllllllllll$	bigIR-T5xp-T5-F	QU	fullrank	nnlm	0.3420	0.8579	0.7034	0.8393	0.5001
$\begin{array}{llllllllllllllllllllllllllllllllllll$	nlm-ens-bst-2	NLM	fullrank	nnlm	0.3542	0.8203	0.6934	0.7190	0.4598
$\begin{array}{llllllllllllllllllllllllllllllllllll$	nlm-ens-bst-3	NLM	fullrank	nnlm	0.3195	0.8491	0.6803	0.7594	0.4526
	nlm-bert-rr	NLM	rerank	nnlm	0.3699	0.7785	0.6721	0.7056	0.4341
$\begin{array}{llllllllllllllllllllllllllllllllllll$	relemb_mim_0_2	UAmsterdam	rerank	nnim	0.2856	0.7677	0.6662	0.7056	0.4350
$\begin{array}{cccccccccccccccccccccccccccccccccccc$	nim-priun-bert	NLM TU Vienne	Tuttrank	nnim	0.3445	0.8003	0.0048	0.0927	0.4205
$\begin{array}{cccccccccccccccccccccccccccccccccccc$	TUW TK 2L aver	TU_Vienna	rerank	1111 mm	0.3188	0.7970	0.0010	0.7056	0.4104
$\begin{array}{cccccccccccccccccccccccccccccccccccc$	n d2a hm25	10_vienna	fullropk	nnlm	0.3073	0.7034	0.0339	0.7030	0.4179
$\begin{array}{cccccccccccccccccccccccccccccccccccc$	$p_d2q_bm25rm^2$	anserini	fullronk	nnlm	0.2737	0.7520	0.0187	0.8055	0.4074
$\begin{array}{cccccccccccccccccccccccccccccccccccc$	p_u2q_om2om5			nnlm	0.2848	0.7424	0.0172	0.8391	0.4293
CoRT-standaloneHSRM-LAVISfullrankmillin 0.2201 0.372 0.392 0.3072 0.3072 0.3072 CoRT-standaloneHSRM-LAVISfullranknnlm 0.2412 0.8112 0.5926 0.6002 0.3308 bl_bcai_mdll_vtbl_bcaifullranktrad 0.1854 0.7037 0.5667 0.7430 0.3374 bl_bcai_indll_vsbl_bcaifullranktrad 0.1999 0.7115 0.5600 0.7430 0.3374 bl_bcai_indll_vsbl_bcaifullranktrad 0.1563 0.6277 0.5092 0.7430 0.3094 indri-fdmbl_mitfullranktrad 0.1563 0.6277 0.5092 0.7430 0.3094 indri-fdmbl_mitfullranktrad 0.1648 0.6498 0.5003 0.7778 0.2989 terrier-BM25bl_mitfullranktrad 0.1631 0.6186 0.4985 0.7649 0.3125 indri-imdsbl_mitfullranktrad 0.1454 0.5094 0.4935 0.8175 0.3119 indri-sdmbl_mitfullranktrad 0.1623 0.4822 0.7726 0.2870 p_bm25anserinifullranktrad 0.1786 0.6585 0.4796 0.7428 0.2856 bm25_bert_tokenUAmsterdamfullranktrad 0.1576 0.4661 0.7533 0.2778 TF_IDF_d_2_t50RMITfullranktrad 0.1391 0.5317 0.468	CoPT hm25		fullropk	nnlm	0.3240	0.7380	0.0149	0.7030	0.3700
CorrestandationeFISRM-LAVISfulliarikfull	CoRT-011123	HSKM-LAVIS	fullronk	nnlm	0.2201	0.0572	0.3992	0.6072	0.3011
$\begin{array}{cccccccccccccccccccccccccccccccccccc$	bl bosi mdl1 vt	horni hori	fullronk	trad	0.2412	0.0112	0.5920	0.0002	0.3308
$\begin{array}{cccccccccccccccccccccccccccccccccccc$	begi class pass	bi_bcai	fullrank	trad	0.1854	0.7037	0.5600	0.7430	0.3380
$\begin{array}{cccccccccccccccccccccccccccccccccccc$	bl begi mdl1 vs	bl beai	fullrank	trad	0.1999	0.7115	0.5000	0.7430	0.3374
Indirival b_{1} mitfullianktrad 0.1736 0.3476 0.3035 0.7745 0.2987 terrier-InL2bl_rmitfullianktrad 0.1864 0.6436 0.4985 0.7649 0.3135 terrier-BM25bl_rmitfullianktrad 0.1631 0.6186 0.4985 0.7572 0.3021 DLH_d_5_t_25RMITfullianktrad 0.1454 0.5094 0.4935 0.8175 0.3199 indri-Imdsbl_rmitfullianktrad 0.1250 0.5866 0.4912 0.7741 0.2961 indri-sdmbl_rmitfullianktrad 0.1600 0.6239 0.4822 0.7726 0.2870 p_bm25m3anserinifullianktrad 0.1495 0.6360 0.4821 0.7939 0.3019 p_bm25anserinifullianktrad 0.1786 0.6585 0.4796 0.7428 0.2856 bm25_bert_tokenUAmsterdamfullianktrad 0.1766 0.6409 0.4686 0.7169 0.2606 terrier-DPHbl_rmitfullianktrad 0.1391 0.5317 0.4580 0.7722 0.2923 small_1kreSearch2vecreranknnlm 0.0222 0.2767 0.7056 0.20112 med_1kreSearch2vecreranknnlm 0.0208 0.2740 0.2661 0.7056 0.2072 DoRA_Large_1kreSearch2vecfullianknnlm 0.0000 0.1287 0.0484 $0.$	indri_fdm	bl_bcai	fullrank	trad	0.1505	0.0277	0.5092	0.7450	0.2089
$\begin{array}{cccccccccccccccccccccccccccccccccccc$	terrier_InI 2	bl_mit	fullrank	trad	0.1798	0.0498	0.5005	0.7778	0.3135
$\begin{array}{cccccccccccccccccccccccccccccccccccc$	terrier_BM25	bl_rmit	fullrank	trad	0.1631	0.6186	0.4980	0.7572	0.3133
$\begin{array}{cccccccccccccccccccccccccccccccccccc$	DIH d 5 t 25	RMIT	fullrank	trad	0.1051	0.0100	0.4935	0.8175	0.3199
$ \begin{array}{cccccccccccccccccccccccccccccccccccc$	indri-lmds	bl rmit	fullrank	trad	0.1250	0.5866	0.4912	0 7741	0.2961
$ \begin{array}{cccccccccccccccccccccccccccccccccccc$	indri-sdm	bl rmit	fullrank	trad	0.1600	0.6239	0.4822	0.7726	0.2870
$ \begin{array}{cccccccccccccccccccccccccccccccccccc$	p bm25rm3	anserini	fullrank	trad	0.1495	0.6360	0.4821	0.7939	0.3019
$ \begin{array}{cccccccccccccccccccccccccccccccccccc$	p_bm25	anserini	fullrank	trad	0.1786	0.6585	0.4796	0.7428	0.2856
$\begin{array}{cccccccccccccccccccccccccccccccccccc$	bm25 bert token	UAmsterdam	fullrank	trad	0.1576	0.6409	0.4686	0.7169	0.2606
TF_IDF_d_2_t_50 RMIT fullrank trad 0.1391 0.5317 0.4580 0.7722 0.2923 small_1k reSearch2vec rerank nnlm 0.0232 0.2785 0.2767 0.7056 0.2112 med_1k reSearch2vec rerank nnlm 0.0222 0.2720 0.2708 0.7056 0.2081 DoRA_Large_1k reSearch2vec rerank nnlm 0.0208 0.2740 0.2661 0.7056 0.2072 DoRA_Small reSearch2vec fullrank nnlm 0.0000 0.1287 0.0484 0.0147 0.0088 DoRA_Med reSearch2vec fullrank nnlm 0.0000 0.1075 0.0431 0.0147 0.0087 DoRA_Large reSearch2vec fullrank nnlm 0.0000 0.1111 0.0414 0.0146 0.0079	terrier-DPH	bl rmit	fullrank	trad	0.1420	0.5667	0.4671	0.7353	0.2758
small_1k reSearch2vec rerank nnlm 0.0232 0.2785 0.2767 0.7056 0.2112 med_1k reSearch2vec rerank nnlm 0.0232 0.2785 0.2767 0.7056 0.2112 DoRA_Large_1k reSearch2vec rerank nnlm 0.0222 0.2740 0.2661 0.7056 0.2072 DoRA_Small reSearch2vec fullrank nnlm 0.0000 0.1287 0.0484 0.0147 0.0088 DoRA_Med reSearch2vec fullrank nnlm 0.0000 0.1075 0.0431 0.0147 0.0087 DoRA_Large reSearch2vec fullrank nnlm 0.0000 0.1111 0.0414 0.0146 0.0079	TF IDF d 2 t 50	RMIT	fullrank	trad	0.1391	0.5317	0.4580	0.7722	0.2923
med_lk reSearch2vec rerank nnlm 0.0222 0.2720 0.2708 0.7056 0.2081 DoRA_Large_1k reSearch2vec rerank nnlm 0.0208 0.2740 0.2661 0.7056 0.2072 DoRA_Small reSearch2vec fullrank nnlm 0.0000 0.1287 0.0484 0.0147 0.0088 DoRA_Med reSearch2vec fullrank nnlm 0.0000 0.1075 0.0431 0.0147 0.0087 DoRA_Large reSearch2vec fullrank nnlm 0.0000 0.1111 0.0414 0.0146 0.0079	small 1k	reSearch2vec	rerank	nnlm	0.0232	0.2785	0.2767	0.7056	0.2112
DoRA_Large_1k reSearch2vec rerank nnlm 0.0208 0.2740 0.2661 0.7056 0.2072 DoRA_Small reSearch2vec fullrank nnlm 0.0000 0.1287 0.0484 0.0147 0.0088 DoRA_Med reSearch2vec fullrank nnlm 0.0000 0.1075 0.0431 0.0147 0.0087 DoRA_Large reSearch2vec fullrank nnlm 0.0000 0.1111 0.0414 0.0146 0.0079	med_1k	reSearch2vec	rerank	nnlm	0.0222	0.2720	0.2708	0.7056	0.2081
DoRA_Small reSearch2vec fullrank nnlm 0.0000 0.1287 0.0484 0.0147 0.0088 DoRA_Med reSearch2vec fullrank nnlm 0.0000 0.1075 0.0431 0.0147 0.0087 DoRA_Large reSearch2vec fullrank nnlm 0.0000 0.1111 0.0414 0.0146 0.0079	DoRA_Large 1k	reSearch2vec	rerank	nnlm	0.0208	0.2740	0.2661	0.7056	0.2072
DoRA_Med reSearch2vec fullrank nnlm 0.0000 0.1075 0.0431 0.0147 0.0087 DoRA_Large reSearch2vec fullrank nnlm 0.0000 0.1111 0.0414 0.0146 0.0079	DoRA_Small	reSearch2vec	fullrank	nnlm	0.0000	0.1287	0.0484	0.0147	0.0088
DoRA_Large reSearch2vec fullrank nnlm 0.0000 0.1111 0.0414 0.0146 0.0079	DoRA_Med	reSearch2vec	fullrank	nnlm	0.0000	0.1075	0.0431	0.0147	0.0087
	DoRA_Large	reSearch2vec	fullrank	nnlm	0.0000	0.1111	0.0414	0.0146	0.0079

Table 5: Passage retrieval runs. RR (MS) is based on MS MARCO labels. All other metrics are based on NIST labels.



Figure 2: Comparison of the best "nnlm" and "trad" runs on individual test queries for the document retrieval task. Queries are sorted by difference in mean performance between "nnlm" and "trad" runs. Queries on which "nnlm" wins with large margin are at the top.



Figure 3: Comparison of the best "nnlm" and "trad" runs on individual test queries for the passage retrieval task. Queries are sorted by difference in mean performance between "nnlm" and "trad" runs. Queries on which "nnlm" wins with large margin are at the top.



(c) NCG@100 for runs on the document retrieval task

(d) NCG@1000 for runs on the passage retrieval task

Figure 4: Analyzing the impact of "fullrank" vs. "rerank" settings on retrieval performance. Figure (a) and (b) show the performance of different runs on the document and passage retrieval tasks, respectively. Figure (c) and (d) plot the NCG@100 and NCG@1000 metrics for the same runs for the two tasks, respectively. The runs are ordered by their NDCG@10 performance along the x-axis in all four plots. We observe, that the best run under the "fullrank" setting outperforms the same under the "rerank" setting for both document and passage retrieval tasks—although the gaps are relatively smaller compared to those in Figure 1. If we compare Figure (a) with (c) and Figure (b) with (d), we do not observe any evidence that the NCG metric is a good predictor of NDCG@10 performance.

Most runs used the ORCAS data for the document retrieval task, with relemb_mlm_0_2 being the only run using the ORCAS data for the passage retrieval task.

This year it was not necessary to use ORCAS data to achieve the highest NDCG@10. However, when we compare the performance of the runs that use the ORCAS dataset with those that do not use the dataset within the same group, we observe that usage of the ORCAS dataset always led to an improved performance in terms of NDCG@10, with maximum increase being around 0.0513 in terms of NDCG@10. This suggests that the ORCAS dataset is providing additional information that is not available in the training data. This could also imply that even though the training dataset provided as part of the track is very large, deep models are still in need of more training data.

NIST labels *vs.* **Sparse MS MARCO labels.** Our baseline human labels from MS MARCO often have one known positive result per query. We use these labels for training, but they are also available for test queries. Although our official evaluation uses NDCG@10 with NIST labels, we now compare this with reciprocal rank (RR) using MS MARCO labels. Our goal is to understand how changing the labeling scheme and metric affects the overall results of the track, but if there is any disagreement we believe the NDCG results are more valid, since they evaluate the ranking more comprehensively and a ranker that can only perform well on labels with exactly the same distribution as the training set is not robust enough for use in real-world applications, where real users will have opinions that are not necessarily identical to the preferences encoded in sparse training labels.

Figure 5 shows the agreement between the results using MS MARCO and NIST labels for the document retrieval and passage retrieval tasks. While the agreement between the evaluation setup based on MS MARCO and TREC seems

Table 6: Leaderboard metrics breakdown. The Kendall agreement (τ) of NDCG@10 and RR (MS) varies across task and run type. Agreement on the best neural network runs is high, but agreement on the best document trad runs is very low. We do not list the agreement for passage nn runs since there are only two runs.



Figure 5: Leaderboard metrics agreement analysis. For document runs, the agreement between the leaderboard metric RR (MS) and the main TREC metric NDCG@10 is lower this year. The Kendall correlation is $\tau = 0.46$, compared to $\tau = 0.69$ in 2019. For the passage task, we see $\tau = 0.69$ in 2020, compared to $\tau = 0.68$ in 2019.

reasonable for both tasks, agreements for the document ranking task seems to be lower (Kendall correlation of 0.46) than agreements for the passage task (Kendall correlation of 0.69). This value is also lower than the correlation we observed for the document retrieval task for last year.

In Table 6 we show how the agreement between the two evaluation setups varies across task and run type. Agreement on which are the best neural network runs is high, but correlation for document trad runs is close to zero.

One explanation for this low correlation could be use of the ORCAS dataset. ORCAS was mainly used in the document retrieval task, and could bring search results more in line with Bing's results, since Bing's results are what may be clicked. Since MS MARCO sparse labels were also generated based on top results from Bing, we would expect to see some correlation between ORCAS runs and MS MARCO labels (and Bing results). By contrast, NIST judges had no information about what results were retrieved or clicked in Bing, so may have somewhat less correlation with Bing's results and users.

In Figure 6 we compare the results from the two evaluation setups when the runs are split based on the usage of the ORCAS dataset. Our results suggest that runs that use the ORCAS dataset did perform somewhat better based on the MS MARCO evaluation setup. While the similarities between the ORCAS dataset and the MS MARCO labels seem to be one reason for the mismatch between the two evaluation results, it is not enough to fully explain the 0.03 correlation in Table6. Removing the ORCAS "trad" runs only increases the correlation to 0.13. In the future we plan to further analyze the possible reasons for this poor correlation, which could also be related to 1) the different metrics used in the two evaluation setups (RR vs. NDCG@10), 2) the different sensitivity of the datasets due to the different number of queries and number of documents labelled per query), or 3) difference in relevance labels provided by NIST assessors vs. labels derived from clicks.



Figure 6: This year it was not necessary to use ORCAS data to achieve the highest NDCG@10. ORCAS runs did somewhat better on the leaderboard metric RR (MS), which uses different labels from the other metrics. This may indicate an alignment between the Bing user clicks in ORCAS with the labeled MS MARCO results, which were also generated by Bing.

5 Conclusion

The TREC 2020 Deep Learning Track has provided two large training datasets, for a document retrieval task and a passage retrieval task, generating two ad hoc test collections with good reusability. The main document and passage training datasets in 2020 were the same as those in 2019. In addition, as part of the 2020 track, we have also released a large click dataset, the ORCAS dataset, which was generated using the logs of the Bing search engine.

For both tasks, in the presence of large training data, this year's non-neural network runs were outperformed by neural network runs. While usage of the ORCAS dataset seems to help improve the performance of the systems, it was not necessary to use ORCAS data to achieve the highest NDCG@10.

We compared reranking approaches to end-to-end retrieval approaches, and in this year's track there was not a huge difference, with some runs performing well in both regimes. This is another result that would be interesting to track in future years, since we would expect that end-to-end retrieval should perform better if it can recall documents that are unavailable in a reranking subtask.

This year the number of runs submitted for both tasks have increased compared to last year. In particular, number of non-neural runs have increased. Hence, test collections generated as part of this year's track may be more reusable compared to last year since these test collections may be fairer towards evaluating the quality of unseen non-neural runs. We note that the number of "nn" runs also seems to be smaller this year. We will continue to encourage a variety of approaches in submission, to avoid converging too quickly on one type of run, and to diversify the judging pools.

Similar to last year, in this year's track we have two types of evaluation label for each task. Our official labels are more comprehensive, covering a large number of results per query, and labeled on a four point scale at NIST. We compare this to the MS MARCO labels, which usually only have one positive result per query. While there was a strong correlation between the evaluation results obtained using the two datasets for the passage retrieval task, the correlation for the document retrieval task was lower. Part of this low correlation seems to be related to the usage of the ORCAS dataset (which is generated using similar dataset as the one used to generate the MS MARCO labels) by some runs, and evaluation results based on MS MARCO data favoring these runs. However, our results suggest that while the ORCAS dataset could be one reason for the low correlation, there might be other reasons causing this reduced correlation, which we plan to explore as future work.

References

Nasreen Abdul-Jaleel, James Allan, W Bruce Croft, Fernando Diaz, Leah Larkey, Xiaoyan Li, Mark D Smucker, and Courtney Wade. Umass at trec 2004: Novelty and hard. 2004.

- Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. Ms marco: A human generated machine reading comprehension dataset. arXiv preprint arXiv:1611.09268, 2016.
- Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013.
- Leonid Boytsov, David Novak, Yury Malkov, and Eric Nyberg. Off the beaten path: Let's replace term-based retrieval with k-nn search. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 1099–1108. ACM, 2016.
- Nick Craswell, Daniel Campos, Bhaskar Mitra, Emine Yilmaz, and Bodo Billerbeck. Orcas: 18 million clicked query-document pairs for analyzing search. *arXiv preprint arXiv:2006.05324*, 2020.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In Proc. CVPR, pages 248–255. Ieee, 2009.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759, 2016.
- Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. Nature, 521(7553):436-444, 2015.
- Irina Matveeva, Chris Burges, Timo Burkard, Andy Laucius, and Leon Wong. High accuracy retrieval with multiple nested ranker. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 437–444. ACM, 2006.
- Bhaskar Mitra and Nick Craswell. An updated duet model for passage re-ranking. *arXiv preprint arXiv:1903.07666*, 2019.
- Bhaskar Mitra, Fernando Diaz, and Nick Craswell. Learning to match using local and distributed representations of text for web search. In *Proc. WWW*, pages 1291–1299, 2017.
- Bhaskar Mitra, Corby Rosset, David Hawking, Nick Craswell, Fernando Diaz, and Emine Yilmaz. Incorporating query term independence assumption for efficient retrieval and ranking using deep neural networks. *arXiv preprint arXiv:1907.03693*, 2019.
- Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. Document expansion by query prediction. *arXiv* preprint arXiv:1904.08375, 2019.
- Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends*(**R**) *in Information Retrieval*, 3(4):333–389, 2009.
- Corby Rosset, Damien Jose, Gargi Ghosh, Bhaskar Mitra, and Saurabh Tiwary. Optimizing query evaluations using reinforcement learning for web search. In *Proc. SIGIR*, pages 1193–1196. ACM, 2018.
- Trevor Strohman, Donald Metzler, Howard Turtle, and W Bruce Croft. Indri: A language model-based search engine for complex queries. In *Proceedings of the International Conference on Intelligent Analysis*, volume 2, pages 2–6. Citeseer, 2005.
- Lidan Wang, Jimmy Lin, and Donald Metzler. A cascade ranking model for efficient ranked retrieval. In *Proceedings* of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, pages 105–114. ACM, 2011.
- Wei Yang, Kuang Lu, Peilin Yang, and Jimmy Lin. Critically examining the "neural hype": Weak baselines and the additivity of effectiveness gains from neural ranking models. In *Proc. SIGIR*, pages 1129–1132. ACM, 2019a.
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*, 2019b.
- Hamed Zamani, Mostafa Dehghani, W Bruce Croft, Erik Learned-Miller, and Jaap Kamps. From neural re-ranking to neural ranking: Learning a sparse representation for inverted indexing. In *Proc. CIKM*, pages 497–506. ACM, 2018.