

# Moving Beyond Downstream Task Accuracy for Information Retrieval Benchmarking \*

Keshav Santhanam<sup>1†</sup> Jon Saad-Falcon<sup>1†</sup> Martin Franz<sup>2</sup> Omar Khattab<sup>1</sup> Avirup Sil<sup>2</sup>  
Radu Florian<sup>2</sup> Md Arafat Sultan<sup>2</sup> Salim Roukos<sup>2</sup> Matei Zaharia<sup>1</sup> Christopher Potts<sup>1</sup>

<sup>1</sup>Stanford University

<sup>2</sup>IBM Research AI

## Abstract

Neural information retrieval (IR) systems have progressed rapidly in recent years, in large part due to the release of publicly available benchmarking tasks. Unfortunately, some dimensions of this progress are illusory: the majority of the popular IR benchmarks today focus exclusively on downstream task accuracy and thus conceal the costs incurred by systems that trade away efficiency for quality. Latency, hardware cost, and other efficiency considerations are paramount to the deployment of IR systems in user-facing settings. We propose that IR benchmarks structure their evaluation methodology to include not only metrics of accuracy, but also efficiency considerations such as a query latency and the corresponding cost budget for a reproducible hardware setting. For the popular IR benchmarks MS MARCO and XOR-TyDi, we show how the best choice of IR system varies according to how these efficiency considerations are chosen and weighed. We hope that future benchmarks will adopt these guidelines toward more holistic IR evaluation.

## 1 Introduction

Benchmark datasets have helped to drive rapid progress in neural information retrieval (IR). When the MS MARCO (Nguyen et al., 2016) Passage Ranking leaderboard began in 2018, the best performing systems had MRR@10 scores around 0.20; the latest entries have since increased accuracy past 0.44. Similarly, the XOR TyDi multilingual question answering (QA) dataset (Asai et al., 2020) was released in 2021 and has seen improvements in recall scores from 0.45 to well past 0.70.

The leaderboards for these datasets are defined by a particular set of accuracy-based metrics, and progress on these metrics can easily become synonymous in people’s minds with progress in gen-

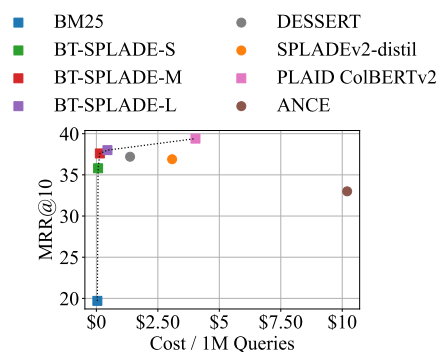


Figure 1: Selected MS MARCO Passage Ranking submissions assessed on both cost and accuracy, with the Pareto frontier marked by a dotted line. The trade-offs evident here are common in real-world applications of IR technologies. These submissions do not represent “optimal” implementations of each respective approach, but rather reflect existing reported implementations and hardware configurations in the literature. Including cost and other efficiency considerations on our leaderboards would lead to more thorough exploration of possible system designs and, in turn, to more meaningful progress.

eral. However, IR and QA systems deployed in production environments must not only deliver high accuracy but also operate within strict resource requirements, including tight bounds on per-query latency, constraints on disk and RAM capacity, and fixed cost budgets for hardware. Within the boundaries of these constraints, the optimal solution for a downstream task may no longer be the system which simply achieves the highest task accuracy.

Figure 1 shows how significant these trade-offs can be. The figure tracks a selection of MS MARCO Passage Ranking submissions, with cost on the x-axis and accuracy (MRR@10) on the y-axis. At one extreme, the BM25 model costs just US\$0.04 per million queries,<sup>1</sup> but it is far behind the other models in accuracy. For very similar

\* Data and code are provided as PrimeQA extensions: <https://github.com/primeqa/primeqa>.

<sup>†</sup> Equal contribution.

<sup>1</sup> Estimated by mapping the minimum necessary hardware to an AWS instance and taking the per-hour on-demand rental cost; see Table 2 for details.

	Hardware			Performance		
	GPU	CPU	RAM (GiB)	MRR@10	Query Latency (ms)	Index Size (GiB)
BM25 (Mackenzie et al., 2021)	0	32	512	18.7	8	1
BM25 (Lassance and Clinchant, 2022)	0	64	-	19.7	4	1
SPLADEv2-distil (Mackenzie et al., 2021)	0	32	512	36.9	220	4
SPLADEv2-distil (Lassance and Clinchant, 2022)	0	64	-	36.8	691	4
BT-SPLADE-S (Lassance and Clinchant, 2022)	0	64	-	35.8	7	1
BT-SPLADE-M (Lassance and Clinchant, 2022)	0	64	-	37.6	13	2
BT-SPLADE-L (Lassance and Clinchant, 2022)	0	64	-	38.0	32	4
ANCE (Xiong et al., 2020)	1	48	650	33.0	12	-
RocketQAv2 (Ren et al., 2021)	-	-	-	37.0	-	-
coCondenser (Gao and Callan, 2021)	-	-	-	38.2	-	-
CoT-MAE (Wu et al., 2022)	-	-	-	39.4	-	-
ColBERTv1 (Khattab and Zaharia, 2020)	4	56	469	36.1	54	154
PLAID ColBERTv2 (Santhanam et al., 2022a)	4	56	503	39.4	32	22
PLAID ColBERTv2 (Santhanam et al., 2022a)	4	56	503	39.4	12	22
DESSERT (Engels et al., 2022)	0	24	235	37.2	16	-

Table 1: Post-hoc leaderboard of MS MARCO v1 dev performance using results reported in corresponding papers. For hardware specifications, we show the precise resources given as the running environment in the paper, even if not all resources were available to the model or the resources were over-provisioned for the particular task. Table 2 provides our estimates of minimum hardware requirements for a subset of these systems. Note that the first PLAID ColBERTv2 result listed was run on a server which includes 4 GPUs but no GPU was actually used for measurement, thereby resulting in a larger latency than the second listed result which does measure GPU execution.

costs to BM25, one can use BT-SPLADE-S and achieve much better performance. On the other hand, the SPLADE-v2-distil model outperforms BT-SPLADE-S by about 1 point, but at a substantially higher cost. Unfortunately, these trade-offs would not be reflected on the MS MARCO leaderboard. Similarly, the top two systems of the XOR TyDi leaderboard as of October 2022 were separated by only 0.1 points in Recall@5000 tokens, but the gap in resource efficiency between these two approaches is entirely unclear.

In this work, we contribute to the growing literature advocating for multidimensional leaderboards that can inform different values and goals (Coleman et al., 2017; Mattson et al., 2020a,b; Baidu Research, 2016; Ma et al., 2021; Liu et al., 2021a; Liang et al., 2022). Our proposal is that researchers should report orthogonal dimensions of performance such as query latency and overall cost, in addition to accuracy-based metrics. Our argument has two main parts.

In part 1 (§2), we create a post-hoc MS MARCO leaderboard from published papers (Table 1). This reveals that systems with similar accuracy often differ substantially along other dimensions, and also that techniques for improving latency and reducing memory and hardware costs are currently

being explored only very sporadically. However, a few of the contributions (Santhanam et al., 2022a; Lassance and Clinchant, 2022; Engels et al., 2022; Li et al., 2022) exemplify the kind of thorough investigation of accuracy and efficiency that we are advocating for, and we believe that improved multidimensional leaderboards could spur further innovation in these areas.

In part 2 (§3), we systematically explore four prominent systems: BM25, Dense Passage Retriever (DPR; Karpukhin et al. 2020), BT-SPLADE-L (Formal et al., 2021; Lassance and Clinchant, 2022), and PLAID ColBERTv2 (Khattab and Zaharia, 2020; Santhanam et al., 2022a,b). These experiments begin to provide a fuller picture of the overall performance of these systems.

We close by discussing practical considerations relating to the multidimensional leaderboards that the field requires. Here, we argue that the *Dynascore* metric developed by Ma et al. (2021) is a promising basis for leaderboards that aim to (1) measure systems along multiple dimensions and (2) provide a single full ranking of systems. Dynascores allow the leaderboard creator to weight different assessment dimensions (e.g., to make cost more important than latency). These weightings transparently reflect a particular set of values, and

	GPU	CPU	RAM	Instance	Cost
BM25	0	1	4	m6g.med	\$0.04
SPLADEv2-distil	0	1	8	r6g.med	\$3.08
BT-SPLADE-S	0	1	8	m6g.med	\$0.07
BT-SPLADE-M	0	1	8	m6g.med	\$0.14
BT-SPLADE-L	0	1	8	r6g.med	\$0.45
ANCE	1	8	64	p3.2x1	\$10.20
ColBERTv1	1	16	256	p3.8x1	\$183.60
PLAID ColBERTv2	0	8	32	r6a.2x1	\$4.03
PLAID ColBERTv2	1	8	64	p3.2x1	\$10.20
DESSERT	0	8	32	m6g.2x1	\$1.37

Table 2: Estimated minimum viable AWS instance type necessary to run each model. RAM is in GiB; Cost is per 1M queries.

we show that they give rise to leaderboards that are likely to incentivize different research questions and system development choices than current leaderboards do.

## 2 A Post-hoc Leaderboard

While existing IR benchmarks facilitate progress on accuracy metrics, the lack of a unified methodology for measuring latency, memory usage, and hardware cost makes it challenging to understand the trade-offs between systems. To illustrate this challenge, we constructed a post-hoc leaderboard for the MS MARCO Passage Ranking benchmark (Table 1). We include the MRR@10 values reported in prior work and, when available, copy the average per-query latency, index size, and hardware configurations reported in the respective papers.<sup>2</sup> We highlight the following key takeaways.

### 2.1 Hardware Provisioning

The hardware configurations in Table 1 are the specific compute environments listed in the corresponding papers rather than the minimum viable hardware necessary to achieve the reported latency. In Table 2, we have sought to specify the minimal configuration that would be needed to run each system. (This may result in an overly optimistic assessment of latency; see §3). The hardware differences between Table 1 and Table 2 reveal that researchers are often using vastly over-provisioned hardware for their experiments. Our proposed leaderboards would create a pressure to be more deliberative about the costs of hardware used when reporting efficiency metrics.

<sup>2</sup>We plan to expand our analysis to include the recently released CITADEL model (Li et al., 2022), first uploaded to arXiv on 11/18/22)

### 2.2 Variation in Methodology

Table 1 shows that both the quality metrics and the hardware used for evaluation across different models vary significantly. Many papers exclusively report accuracy, which precludes any quantitative understanding of efficiency implications (Ren et al., 2021; Gao and Callan, 2021; Wu et al., 2022). For papers that do report efficiency-oriented metrics, the evaluation environment and methodology are often different; for example, the results from Mackenzie et al. 2021 and Lassance and Clinchant 2022 are measured on a single CPU thread whereas Khattab and Zaharia 2020 and Santhanam et al. 2022a leverage multiple CPU threads for intra-query parallelism, and even a GPU for certain settings. We also observe performance variability even for the same model, with Mackenzie et al. 2021 (220 ms) and Lassance and Clinchant 2022 (691 ms) reporting SPLADEv2 latency numbers which are  $3\times$  apart. Similarly, the BM25 latencies reported by these papers differ by a factor of  $2\times$ .

### 2.3 Multidimensional Evaluation Criteria

The optimal model choice for MS MARCO is heavily dependent on how we weight the different evaluation metrics. Based purely on accuracy, CoT-MAE and PLAID ColBERTv2 are the top-performers in Table 1, with an MRR@10 score of 39.4 for both. However, we do not have all the information we need to compare them along other dimensions. On the other hand, BM25 is the fastest model, with a per-query latency of only 4 ms as measured by Lassance and Clinchant (2022), and its space footprint is also small. The trade-off is that it has the lowest accuracy in the cohort. Compared to BM25, one of the highly optimized BT-SPLADE models may be a better choice. Figure 1 begins to suggest how we might reason about these often opposing pressures.

## 3 Experiments with Representative Retrievers

As Table 1 makes clear, the existing literature does not include systematic, multidimensional comparisons of models. In this section, we report on experiments that allow us to make these comparisons. We focus on four models:

**BM25 (Robertson et al., 1995)** A sparse, term-based IR model. BM25 remains a strong baseline in many IR contexts and is notable for its low latency and low costs. We assess a basic implementation.

	Hardware				Performance	
	GPU	CPU	RAM	Instance	Latency	Cost
BM25	0	1	4	m6gd.med	11	\$0.14
BM25					10	\$0.48
DPR					146	\$6.78
ColBERTv2-S	0	1	32	x2gd.lrg	206	\$9.58
ColBERTv2-M					321	\$14.90
ColBERTv2-L					459	\$21.30
BT-SPLADE-L					46	\$2.15
BM25	0	16	4	c7g.4xl	9	\$1.48
BM25					9	\$1.43
DPR					19	\$2.97
ColBERTv2	0	16	32	c7g.4xl	51	\$8.19
ColBERTv2-M					63	\$10.09
ColBERTv2-L					86	\$13.88
BT-SPLADE-L					33	\$5.38
BM25	1	1	4	p3.2xl	11	\$9.09
BM25					10	\$8.46
DPR					19	\$15.73
ColBERTv2-S	1	1	32	p3.2xl	36	\$30.46
ColBERTv2-M					52	\$44.54
ColBERTv2-L					99	\$83.97
BT-SPLADE-L					42	\$35.86
BM25	1	16	4	p3.8xl	9	\$30.51
BM25					9	\$29.94
DPR					18	\$61.06
ColBERTv2-S	1	16	32	p3.8xl	27	\$90.41
ColBERTv2-M					36	\$123.35
ColBERTv2-L					55	\$187.24
BT-SPLADE-L					33	\$112.87

(a) MS MARCO efficiency results.

	Hardware				Performance	
	GPU	CPU	RAM	Instance	Latency	Cost
BM25					37	\$3.45
DPR					208	\$19.29
ColBERTv2-S	0	1	64	x2gd.xlrg	343	\$31.84
ColBERTv2-M					771	\$71.56
ColBERTv2-L					1107	\$102.74
BT-SPLADE-L					70	\$6.49
BM25					36	\$6.11
DPR					84	\$14.38
ColBERTv2-S	0	16	64	m6g.4xlrg	83	\$14.17
ColBERTv2-M					110	\$18.83
ColBERTv2-L					165	\$28.26
BT-SPLADE-L					43	\$7.41
BM25					36	\$123.69
DPR					26	\$89.91
ColBERTv2-S	1	1	64	p3.8xl	57	\$194.81
ColBERTv2-M					74	\$251.76
ColBERTv2-L					121	\$411.62
BT-SPLADE-L					63	\$213.17
BM25					35	\$118.12
DPR					28	\$95.23
ColBERTv2-S	1	16	64	p3.8xl	46	\$155.10
ColBERTv2-M					65	\$219.84
ColBERTv2-L					106	\$359.65
BT-SPLADE-L					43	\$147.53

(b) XOR-TyDi efficiency results.

	MS MARCO		XOR-TyDi	
	MRR@10	Success@10	MRR@10	Success@10
BM25	18.7	38.6	26.3	44.5
DPR	31.7	52.1	16.9	32.4
ColBERTv2-S	39.4	68.8	41.8	57.5
ColBERTv2-M	39.7	69.6	45.4	63.0
ColBERTv2-L	39.7	69.7	47.4	66.0
BT-SPLADE-L	38.0	66.3	43.5	65.4

(c) Accuracy.

Table 3: Experimental results. Latency is average per-query latency (ms), and Cost is per 1M queries.

More sophisticated versions may achieve better accuracy (Berger and Lafferty, 1999; Boytsov, 2020), though often with trade-offs along other dimensions (Lin et al., 2016). For evidence that simple BM25 models often perform best in their class, see Thakur et al. 2021.

**DPR (Karpukhin et al., 2020)** A dense single-vector neural IR model. DPR separately encodes queries and documents into vectors and scores them using fast dot-product-based comparisons.

**BT-SPLADE-L (Lassance and Clinchant, 2022)** SPLADE (Formal et al., 2021) is a sparse neural model. The BT-SPLADE variants are highly optimized versions of this model designed to achieve low latency and reduce the overall computational demands of the original model. To the best of our knowledge, only the Large configuration, BT-SPLADE-L, is publicly available.

**PLAID ColBERTv2 (Santhanam et al., 2022a)** The ColBERT retrieval model (Khattab and Zaharia, 2020) encodes queries and documents into sequences of output states, one per input token, and scoring is done based on the maximum similarity values obtained for each query token. ColBERTv2 (Santhanam et al., 2022b) improves supervision and reduces the space footprint of the index, and the PLAID engine focuses on achieving low latency. The parameter  $k$  to the model dictates the initial candidate passages that are scored by the model. Larger  $k$  thus leads to higher latency but generally more accurate search. In our initial experiments, we noticed that higher  $k$  led to better out-of-domain performance, and thus we evaluated the recommended settings from Santhanam et al. (2022a), namely,  $k \in \{10, 100, 1000\}$ . To distinguish these configurations from the number



of passages evaluated by the MRR or Success metric (also referred to as  $k$ ), we refer to these configurations as the ‘-S’, ‘-M’, and ‘-L’ variants of ColBERTv2, respectively.

We chose these models as representatives of key IR model archetypes: lexical models (BM25), dense single-vector models (DPR), sparse neural models (SPLADE), and late-interaction models (ColBERT). The three ColBERT variants provide a glimpse of how model configuration choices can interact with our metrics.

We use two retrieval datasets: MS MARCO (Nguyen et al., 2016) and XOR-TyDi (Asai et al., 2020). All neural models in our analysis are trained on MS MARCO data. We evaluate on XOR-TyDi without further fine-tuning to test out-of-domain evaluation (see Appendix A for more details).

Our goal is to understand how the relative performance of these models changes depending on the available resources and evaluation criteria. Our approach differs from the post-hoc leaderboard detailed in §2 in two key ways: (1) we fix the underlying hardware platform across all models, and (2) we evaluate each model across a broad range of hardware configurations (AWS instance types), ensuring that we capture an extensive space of compute environments. Furthermore, in addition to quality, we also report the average per-query latency and the corresponding cost of running 1 million queries given the latency and the choice of instance type. This approach therefore enables a more principled and holistic comparison between the models.

We use the open-source PrimeQA framework,<sup>3</sup> which provides a uniform interface to implementations of BM25, DPR, and PLAID ColBERTv2. For SPLADE, we use the open-source implementation maintained by the paper authors.<sup>4</sup> For each model we retrieve the top 10 most relevant passages. We report the average latency of running a fixed sample of 1000 queries from each dataset as measured across 5 trials. See Appendix A for more details about the evaluation environments and model configurations.

Table 3 summarizes our experiments. Tables 3a and 3b report efficiency numbers, with costs estimated according to the same hardware pricing used for Table 2. Table 3c gives accuracy results (MRR@10 and Success@10).

Overall, BM25 is the least expensive model

<sup>3</sup><https://github.com/primeqa/primeqa>

<sup>4</sup><https://github.com/naver/splade>

when selecting the minimum viable instance type: only BM25 is able to run with 4 GB memory. However, its accuracy scores are low enough to essentially remove it from contention.

On both datasets, we find that BT-SPLADE-L and the PLAID ColBERTv2 variants are the most accurate models by considerable margins. On MS MARCO, all the ColBERTv2 variants outperform BT-SPLADE-L in MRR@10 and Success@10 respectively, while BT-SPLADE-L offers faster and cheaper scenarios than ColBERTv2 for applications that permit a moderate loss in quality.

In the out-of-domain XOR-TyDi evaluation, BT-SPLADE-L outperforms the ColBERTv2-S variant, which sets  $k = 10$  (the least computationally-intensive configuration). We hypothesize this loss in quality is an artifact of the approximations employed by the default configuration. Hence, we also test the more computationally-intensive configurations mentioned above: ColBERTv2-M ( $k = 100$ ) and ColBERTv2-L ( $k = 1000$ ). These tests reveals that ColBERTv2-L solidly outperforms BT-SPLADE-L in MRR@10 and Success@10, while allowing BT-SPLADE-L to expand its edge in latency and cost.

Interestingly, despite per-instance costs being higher for certain instances, selecting the more expensive instance can actually reduce cost depending on the model. For example, the c7g.4xlarge instance is  $3.5\times$  more expensive than x2gd.large, but ColBERTv2-S runs  $4\times$  faster with 16 CPU threads and therefore is cheaper to execute on the c7g.4xlarge. These findings further reveal the rich space of trade-offs when it comes to model configurations, efficiency, and accuracy.

## 4 Discussion and Recommendations

In this section, we highlight several considerations for future IR leaderboards and offer recommendations for key design decisions.

### 4.1 Evaluation Platform

A critical design goal for IR leaderboards should be to encourage transparent, reproducible submissions. However, as we see in Table 1, many existing submissions are performed using custom—and likely private—hardware configurations and are therefore difficult to replicate.

Instead, we strongly recommend all submissions be tied to a particular public cloud instance type.<sup>5</sup>

<sup>5</sup>In principle, any public cloud provider (e.g., AWS EC2,

In particular, leaderboards should require that the specific evaluation environment associated with each submission (at inference time) can be easily reproduced. This encourages submissions to find realistic and transparent ways to use public cloud resources that minimize the cost of their submissions in practice, subject to their own goals for latency and quality. We note that our inclusion of “cost” subsumes many individual tradeoffs that systems may consider, like the amount of RAM (or, in principle, storage) required by the index and model, or the number of CPUs, GPUs, or TPUs.

In principle, leaderboards could report the constituent resources instead of reporting a specific reproducible hardware platform. For example, a leaderboard could simply report the number of CPU threads and GPUs per submission. This offers the benefit of decoupling submissions from the offerings available on public cloud providers. However, this approach fails to account for the ever-growing space of hardware resources or their variable (and changing) pricing. For instance, it is likely unrealistic to expect leaderboard builders to quantify the difference in cost between a V100 and a more recent A100 GPU—or newer generations, like H100, let alone FPGAs or other heterogeneous choices. We argue that allowing submissions to select their own public cloud instance (including its capabilities and pricing) reflects a realistic, market-driven, up-to-date strategy for estimating dollar costs. In practice, the leaderboard creators need to set a policy for dealing with changing prices over time. They may, for instance, opt to use the latest pricing at all times. This may lead to shifts in the leaderboard rankings over time, reflecting the changing tradeoffs between cost and the other dimensions evaluated.

## 4.2 Scoring

Efficiency-aware IR leaderboards have several options for scoring and ranking submissions. We enumerate three such strategies here:

1. Fix a latency or cost threshold (for example) and rank eligible systems by accuracy. Many different thresholds could be chosen to facilitate competition in different resource regimes (e.g., mobile phones vs. data centers).
2. Fix an accuracy threshold and rank eligible systems by latency or cost (or other aspects).

Google Cloud, or Azure) is acceptable as long as they offer a transparent way to estimate costs.

The accuracy threshold could be set to the state-of-the-art result from prior years.

3. Weight the different assessment dimensions and distill them into a single score, possibly after filtering systems based on thresholds on accuracy, latency, and/or cost.

Of these approaches, the third is the most flexible and is the only one that can provide a complete ranking of systems. The *Dynascores* of Ma et al. (2021) seem particularly well-suited to IR leaderboards, since they allow the leaderboard creator to assign weights to each of the dimensions included in the assessment, reflecting the relative importance assigned to each. The Dynascore itself is a utility-theoretic aggregation of all the measurements and yields a ranking of the systems under consideration.

Following Ma et al., we define Dynascores as follows. For a set of models  $\mathcal{M} = \{\mathcal{M}_1, \dots, \mathcal{M}_N\}$  and assessment metrics  $\mu = \{\mu_1, \dots, \mu_k\}$ , the Dynascore for a model  $\mathcal{M}_i \in \mathcal{M}$  is defined as

$$\sum_{j=1}^k \mathbf{w}_{\mu_j} \frac{\mu_j(\mathcal{M}_i)}{\text{AMRS}(\mathcal{M}, \text{acc}, \mu_j)} \quad (1)$$

where  $\mathbf{w}_{\mu_j}$  is the weight assigned to  $\mu_j$  (we ensure that the sum of all the weights is equal to 1), and  $\text{acc}$  is an appropriate notion of accuracy (e.g.,  $\text{MRR}@10$ ). The AMRS (average marginal rate of substitution) is defined as

$$\frac{1}{N} \sum_i^N \left| \frac{\mu(\mathcal{M}_i) - \mu(\mathcal{M}_{i+1})}{\text{acc}(\mathcal{M}_i) - \text{acc}(\mathcal{M}_{i+1})} \right| \quad (2)$$

for models  $\mathcal{M}_1, \dots, \mathcal{M}_N$  organized from worst to best performing according to  $\text{acc}$ . In our experiments, we use the negative of Cost and Latency, so that all the metrics are oriented in such a way that larger values are better. If a model cannot be run for a given hardware configuration, it is excluded.

For a default weighting, Ma et al. suggest assigning half of the weight to the performance metric and spreading the other half evenly over the other metrics. For our experiments, this leads to

$$\{\text{MRR}@10: 0.5, \text{Cost}: 0.25, \text{Latency}: 0.25\}$$

In Table 4, we show what the MS MARCO and XOR-TyDi leaderboards would look like if they were driven by this Dynascore weighting. In both leaderboards, ColBERTv2 variants are the winning systems. This is very decisive for XOR-TyDi. For MS MARCO, ColBERTv2 and SPLADE are much closer overall.

	System	Hardware	Dynascore
1	CoBERTv2-M	16 CPU, 32 GB memory	19.127
2	CoBERTv2-S	16 CPU, 32 GB memory	19.118
3	CoBERTv2-L	16 CPU, 32 GB memory	18.857
4	CoBERTv2-S	1 GPU, 1 CPU, 32 GB memory	18.698
5	BT-SPLADE-L	16 CPU, 32 GB memory	18.637
6	BT-SPLADE-L	1 CPU, 32 GB memory	18.616
7	CoBERTv2-M	1 GPU, 1 CPU, 32 GB memory	18.385
8	CoBERTv2-S	1 CPU, 32 GB memory	17.912
9	BT-SPLADE-L	1 GPU, 1 CPU, 32 GB memory	17.839
10	CoBERTv2-S	1 GPU, 16 CPU, 32 GB memory	17.331
11	CoBERTv2-L	1 GPU, 1 CPU, 32 GB memory	17.080
12	CoBERTv2-M	1 GPU, 32 GB memory	17.060
13	CoBERTv2-M	1 GPU, 16 CPU, 32 GB memory	16.619
14	BT-SPLADE-L	1 GPU, 16 CPU, 32 GB memory	16.062
15	CoBERTv2-L	1 CPU, 32 GB memory	15.858
16	DPR	16 CPU, 32 GB memory	15.635
17	DPR	1 GPU, 1 CPU, 32 GB memory	15.330
18	CoBERTv2-L	1 GPU, 16 CPU, 32 GB memory	14.940
19	DPR	1 CPU, 32 GB memory	14.583
20	DPR	1 GPU, 16 CPU, 32 GB memory	14.252
21	BM25	1 CPU, 4 GB memory	9.263
22	BM25	1 CPU, 32 GB memory	9.263
23	BM25	16 CPU, 32 GB memory	9.248
24	BM25	16 CPU, 4 GB memory	9.246
25	BM25	1 GPU, 1 CPU, 32 GB memory	9.072
26	BM25	1 GPU, 1 CPU, 4 GB memory	9.049
27	BM25	1 GPU, 16 CPU, 32 GB memory	8.565
28	BM25	1 GPU, 16 CPU, 4 GB memory	8.551

(a) MS MARCO.

	System	Hardware	Dynascore
1	CoBERTv2-L	16 CPU, 64 GB memory	21.241
2	BT-SPLADE-L	16 CPU, 64 GB memory	21.119
3	CoBERTv2-M	16 CPU, 64 GB memory	21.063
4	BT-SPLADE-L	1 CPU, 64 GB memory	20.753
5	CoBERTv2-M	1 GPU, 16 CPU, 64 GB memory	20.255
6	BT-SPLADE-L	1 GPU, 16 CPU, 64 GB memory	20.123
7	CoBERTv2-M	1 GPU, 1 CPU, 64 GB memory	19.904
8	CoBERTv2-L	1 GPU, 16 CPU, 64 GB memory	19.700
9	CoBERTv2-S	16 CPU, 64 GB memory	19.649
10	BT-SPLADE-L	1 GPU, 1 CPU, 64 GB memory	19.380
11	CoBERTv2-S	1 GPU, 16 CPU, 64 GB memory	19.157
12	CoBERTv2-L	1 GPU, 1 CPU, 64 GB memory	19.123
13	CoBERTv2-S	1 GPU, 1 CPU, 64 GB memory	18.723
14	CoBERTv2-S	1 CPU, 64 GB memory	15.934
15	BM25	1 CPU, 64 GB memory	12.635
16	BM25	16 CPU, 64 GB memory	12.630
17	BM25	1 GPU, 16 CPU, 64 GB memory	11.847
18	BM25	1 GPU, 1 CPU, 64 GB memory	11.794
19	CoBERTv2-M	1 CPU, 64 GB memory	11.563
20	CoBERTv2-L	1 CPU, 64 GB memory	7.708
21	DPR	1 GPU, 1 CPU, 64 GB memory	7.452
22	DPR	1 GPU, 16 CPU, 64 GB memory	7.386
23	DPR	16 CPU, 64 GB memory	7.188
24	DPR	1 CPU, 64 GB memory	5.442

(b) XOR-TyDi.

Table 4: Dynascores for the default weighting scheme {Accuracy: 0.5, Cost: 0.25, Latency: 0.25}.

However, this weighting scheme is not the only reasonable choice one could make. Appendix B presents a range of different leaderboards capturing different relative values. Here, we mention a few highlights. First, if accuracy is very important (e.g., MRR@10: 0.9), then all the CoBERTv2 systems dominate all the others. Second, if we are very cost sensitive, then we could use a weighting {MRR@10: 0.4, Cost: 0.4, Latency: 0.2}. In this setting, CoBERTv2-S rises to the top of the leaderboard for MS MARCO and BT-SPLADE-L is more of a contender. Third, on the other hand, if money is no object, we could use a weighting like {MRR@10: 0.75, Cost: 0.01, Latency: 0.24}. This setting justifies using a GPU with COIBERTv2, whereas most other settings do not justify the expense of a GPU for this system. In contrast, a GPU is never justified for BT-SPLADE-L.

To get a holistic picture of how different weightings affect these leaderboards, we conducted a systematic exploration of different weighting vectors. Figure 2a summarizes these findings in terms of the winning system for each setting. The plots depict Latency on the x-axis and Accuracy on the y-axis. The three weights always sum to 1 (Dynascores are normalized), so the Cost value is determined by the other two, as  $1.0 - \text{Accuracy} - \text{Latency}$ .

The overall picture is clear. For MS MARCO, a CoBERTv2-M or CoBERTv2-S system is generally the best choice overall assuming Accuracy is the most important value, and CoBERTv2-L is never a winner. In contrast, a BT-SPLADE-L system is generally the best choice where Cost and Latency are much more important than Accuracy. DPR is a winner only where Accuracy is relatively unimportant, and BM25 is a winner only where Accuracy is assigned essentially zero importance. For the out-of-domain XOR-TyDi test, the picture is somewhat different: now CoBERTv2-L is the dominant system, followed by BT-SPLADE-L.

### 4.3 Metrics

Here we briefly explore various metrics and their potential role in leaderboard design, beginning with the two that we focused on in our experiments:

**Latency** Latency measures the time for a single query to be executed and a result to be returned to the user. Some existing work has measured latency on a single CPU thread to isolate the system performance from potential noise (Mackenzie et al., 2021; Lassance and Clinchant, 2022). While this approach ensures a level playing field for different systems, it fails to reward systems which do benefit from accelerated computation (e.g., on GPUs) or

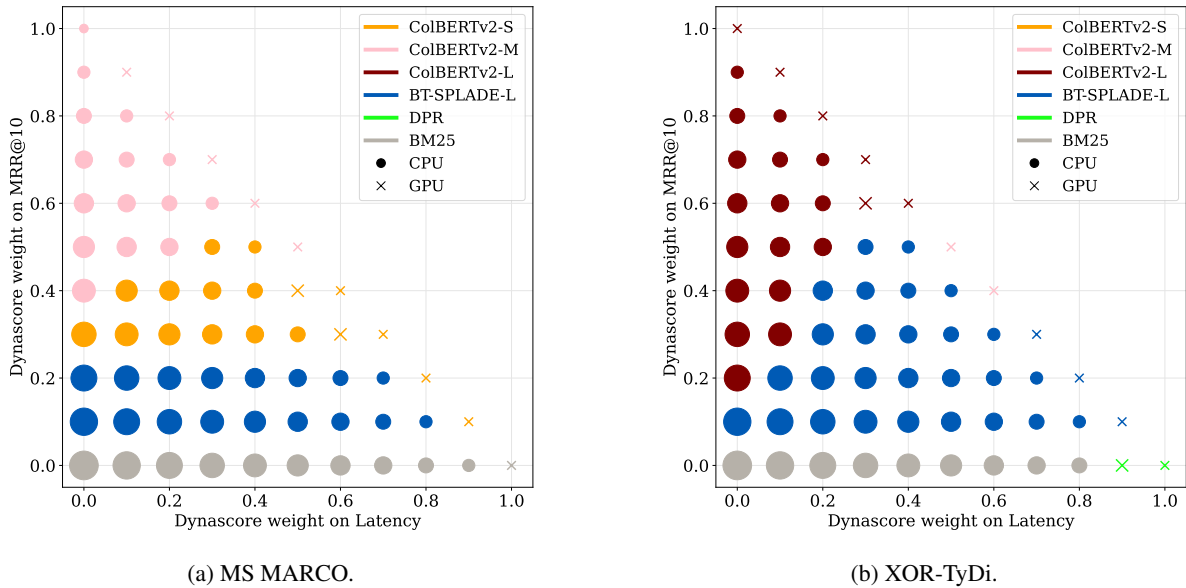


Figure 2: Exploration of Dynascore weighting schemes. Marker sizes are proportional to Cost weights (large dots represent more-cost-sensitive weightings and thus the most expensive systems are along the diagonal).

*intra-query* parallelism such as DPR and PLAID ColBERTv2. Therefore, for leaderboards with raw latency as a primary objective, we recommend allowing flexibility in the evaluation hardware to enable the fastest possible submissions. Such flexibility is then subsumed in the dollar cost below.

**Dollar cost** Measuring the financial overhead of deploying IR systems is key for production settings. One way to measure cost is to select a particular public cloud instance type and simply multiply the instance rental rate by the time to execute some fixed number of queries, as in Table 2.

**Throughput** Throughput measures the total number of queries which can be executed over a fixed time period. Maximizing throughput could entail compromising the average per-query latency in favor of completing a larger volume of queries concurrently. It is important that leaderboards explicitly define the methodology for measuring latency and/or throughput in practice (e.g., in terms of average time to complete one query at a time or average time to complete a batch of 16 queries).

**FLOPs** The number of floating point operations (FLOPs) executed by a particular model gives a hardware-agnostic metric for assessing computational complexity. While this metric is meaningful in the context of compute-bound operations such as language modeling (Liu et al., 2021b), IR systems are often comprised of heterogeneous pipelines where the bottleneck operation may instead be bandwidth-bound (Santhanam et al., 2022a). There-

fore we discourage FLOPs as a metric to compete on for IR leaderboards.

**Memory usage** IR systems often pre-compute large indexes and load them into memory (Johnson et al., 2019; Khattab and Zaharia, 2020), meaning memory usage is an important consideration for determining the minimal hardware necessary to run a given system. In particular, we recommend leaderboard submissions report the index size at minimum as well as the dynamic peak memory usage if possible. The reporting of the dollar cost of each system (i.e., which accounts for the total RAM made available for each system) allows us to quantify the effect of this dimension in practice.

## 5 Related Work

Many benchmarks holistically evaluate the accuracy of IR systems on dimensions such as out-of-domain robustness (Thakur et al., 2021; Santhanam et al., 2022b) and multilingual capabilities (Zhang et al., 2021, 2022). While these benchmarks are key for measuring retrieval effectiveness, they do not incorporate analysis of resource efficiency or cost. The MLPerf benchmark does include such analysis but is focused on vision and NLP tasks rather than retrieval (Mattson et al., 2020a). Several retrieval papers offer exemplar efficiency studies (Mackenzie et al., 2021; Santhanam et al., 2022a; Engels et al., 2022; Li et al., 2022); we advocate in this work for more widespread adoption as well as standardization around the evaluation procedure.



## 6 Conclusion

We argued that current benchmarks for information retrieval should adopt multidimensional leaderboards that assess systems based on latency and cost as well as standard accuracy-style metrics. Such leaderboards would likely have the effect of spurring innovation, and lead to more thorough experimentation and more detailed reporting of results in the literature. As a proof of concept, we conducted experiments with four representative IR systems, measuring latency, cost, and accuracy, and showed that this reveals important differences between these systems that are hidden if only accuracy is reported. Finally, we tentatively proposed Dynascoring as a simple, flexible method for creating multidimensional leaderboards in this space.

## 7 Limitations

We identify two sources of limitations in our work: the range of metrics we consider, and the range of models we explore in our experiments.

Our paper advocates for multidimensional leaderboards. In the interest of concision, we focused on cost and latency as well as system quality. These choices reflect a particular set of values when it comes to developing retrieval models. In §4.3, we briefly consider a wider range of metrics and highlight some of the values they encode. Even this list is not exhaustive, however. In general, we hope that our work leads to more discussion of the values that should be captured in the leaderboards in this space, and so we do not intend our choices to limit exploration here.

For our post-hoc leaderboard (Table 1), we surveyed the literature to find representative systems. We cannot claim that we have exhaustively listed all systems, and any omissions should count as limitations of our work. In particular, we note that we did not consider any re-ranking models, which would consume the top- $k$  results from any of the retrievers we test and produce a re-arranged list. Such models would only add weight to our argument of diverse cost-quality tradeoffs, as re-ranking systems must determine which retriever to re-rank, how many passages to re-rank per query (i.e., setting  $k$ ), and what hardware to use for re-ranking models, which are typically especially accelerator-intensive (i.e., require GPUs or TPUs).

For our experimental comparisons, we chose four models that we take to be representative of broad approaches in this area. However, different

choices from within the space of all possibilities might have led to different conclusions. In addition, our experimental protocols may interact with our model choices in important ways. For example, the literature on SPLADE suggests that it may be able to fit its index on machines with 8 or 16 GB of RAM, but our experiments used 32 GB of RAM.

Our hope is merely that our results help encourage the development of leaderboards that offer numerous, fine-grained comparisons from many members of the scientific community, and that these leaderboards come to reflect different values for scoring and ranking such systems as well.

## Acknowledgements

This work was partially supported by IBM as a founding member of the Stanford Institute for Human-Centered Artificial Intelligence (HAI). This research was supported in part by affiliate members and other supporters of the Stanford DAWN project—Ant Financial, Facebook, Google, and VMware—as well as Toyota Research Institute, Cisco, SAP, and the NSF under CAREER grant CNS-1651570. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. Toyota Research Institute ("TRI") provided funds to assist the authors with their research but this article solely reflects the opinions and conclusions of its authors and not TRI or any other Toyota entity. Omar Khattab is supported by the Apple Scholars in AI/ML fellowship.

## References

- Akari Asai, Jungo Kasai, Jonathan H Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2020. [XOR QA: Cross-lingual Open-Retrieval Question Answering](#). *arXiv preprint arXiv:2010.11856*.
- Baidu Research. 2016. [DeepBench: Benchmarking deep learning operations on different hardware](#). Electronic resource.
- Adam Berger and John Lafferty. 1999. Information retrieval as statistical translation. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 222–229.
- Leonid Boytsov. 2020. [Traditional IR rivals neural models on the MS MARCO document ranking leaderboard](#). *arXiv preprint arXiv:2012.08020*.

- Cody Coleman, Deepak Narayanan, Daniel Kang, Tian Zhao, Jian Zhang, Luigi Nardi, Peter Bailis, Kunle Olukotun, Chris Ré, and Matei Zaharia. 2017. Dawn-bench: An end-to-end deep learning benchmark and competition. *Training*, 100(101):102.
- Joshua Engels, Benjamin Coleman, Vihan Lakshman, and Anshumali Shrivastava. 2022. **DESSERT: An Efficient Algorithm for Vector Set Search with Vector Set Queries**. *arXiv preprint arXiv:2210.15748*.
- Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. Splade: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2288–2292.
- Luyu Gao and Jamie Callan. 2021. **Unsupervised Corpus aware Language Model Pre-training for Dense Passage Retrieval**. *arXiv preprint arXiv:2108.05540*.
- Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-Scale Similarity Search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. **Dense passage retrieval for open-domain question answering**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.
- Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.
- Carlos Lassance and Stéphane Clinchant. 2022. An Efficiency Study for SPLADE Models. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2220–2226.
- Minghan Li, Sheng-Chieh Lin, Barlas Oguz, Asish Ghoshal, Jimmy Lin, Yashar Mehdad, Wen-tau Yih, and Xilun Chen. 2022. **CITADEL: Conditional Token Interaction via Dynamic Lexical Routing for Efficient and Effective Multi-Vector Retrieval**. *arXiv preprint arXiv:2211.10411*.
- Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, E. Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel J. Orr, Lucia Zheng, Mert Yuksekogul, Mirac Suzgun, Nathan S. Kim, Neel Guha, Niladri S. Chatterji, O. Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas F. Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2022. **Holistic evaluation of language models**. *arXiv preprint arXiv:2211.09110*.
- Jimmy Lin, Matt Crane, Andrew Trotman, Jamie Callan, Ishan Chattopadhyaya, John Foley, Grant Ingersoll, Craig Macdonald, and Sebastiano Vigna. 2016. Toward reproducible baselines: The open-source IR reproducibility challenge. In *European Conference on Information Retrieval*, pages 408–420. Springer.
- Pengfei Liu, Jinlan Fu, Yang Xiao, Weizhe Yuan, Shuaichen Chang, Junqi Dai, Yixin Liu, Zihuiwen Ye, and Graham Neubig. 2021a. **ExplainsBoard: An explainable leaderboard for NLP**. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 280–289, Online. Association for Computational Linguistics.
- Xiangyang Liu, Tianxiang Sun, Junliang He, Lingling Wu, Xinyu Zhang, Hao Jiang, Zhao Cao, Xuanjing Huang, and Xipeng Qiu. 2021b. **Towards efficient NLP: A standard evaluation and a strong baseline**. *arXiv preprint arXiv:2110.07038*.
- Zhiyi Ma, Kawin Ethayarajh, Tristan Thrush, Somya Jain, Ledell Wu, Robin Jia, Christopher Potts, Adina Williams, and Douwe Kiela. 2021. **Dynaboard: An evaluation-as-a-service platform for holistic next-generation benchmarking**. In *Advances in Neural Information Processing Systems*, volume 34, pages 10351–10367.
- Joel Mackenzie, Andrew Trotman, and Jimmy Lin. 2021. **Wacky weights in learned sparse representations and the revenge of score-at-a-time query evaluation**. *arXiv preprint arXiv:2110.11540*.
- Peter Mattson, Christine Cheng, Gregory Diamos, Cody Coleman, Paulius Micikevicius, David Patterson, Hanlin Tang, Gu-Yeon Wei, Peter Bailis, Victor Bittorf, David Brooks, Dehao Chen, Debo Dutta, Udit Gupta, Kim Hazelwood, Andy Hock, Xinyuan Huang, Daniel Kang, David Kanter, Naveen Kumar, Jeffery Liao, Deepak Narayanan, Tayo Oguntebi, Gennady Pekhimenko, Lillian Pentecost, Vijay Janapa Reddi, Taylor Robie, Tom St John, Carole-Jean Wu, Lingjie Xu, Cliff Young, and Matei Zaharia. 2020a. **MLperf training benchmark**. In *Proceedings of Machine Learning and Systems*, volume 2, pages 336–349.
- Peter Mattson, Vijay Janapa Reddi, Christine Cheng, Cody Coleman, Greg Diamos, David Kanter, Paulius Micikevicius, David Patterson, Guenther Schmuelling, Hanlin Tang, Gu-Yeon Wei, and Carole-Jean Wu. 2020b. **MLPerf: An industry standard benchmark suite for machine learning performance**. *IEEE Micro*, 40(2):8–16.

- Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated MACHine Reading COMprehension dataset. In *CoCo@ NIPS*.
- Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, Qiaoqiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021. [RocketQv2: A Joint Training Method for Dense Passage Retrieval and Passage Re-ranking](#). *arXiv preprint arXiv:2110.07367*.
- Stephen E. Robertson, Steve Walker, Susan Jones, Micheline M. Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at TREC-3. *NIST Special Publication Sp*, 109:109.
- Keshav Santhanam, Omar Khattab, Christopher Potts, and Matei Zaharia. 2022a. [PLAID: An efficient engine for late interaction retrieval](#). In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, page 1747–1756, New York, NY, USA. Association for Computing Machinery.
- Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022b. [ColBERTv2: Effective and efficient retrieval via lightweight late interaction](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3715–3734, Seattle, United States. Association for Computational Linguistics.
- Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. [Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models](#). In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.
- Xing Wu, Guangyuan Ma, Meng Lin, Zijia Lin, Zhongyuan Wang, and Songlin Hu. 2022. [Contextual Mask Auto-Encoder for dense passage retrieval](#). *arXiv preprint arXiv:2208.07670*.
- Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. [Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval](#). *arXiv preprint arXiv:2007.00808*.
- Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. 2021. [Mr. tydi: A multi-lingual benchmark for dense retrieval](#). *arXiv preprint arXiv:2108.08787*.
- Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2022. [Making a miracle: Multilingual information retrieval across a continuum of languages](#). *arXiv preprint arXiv:2210.09984*.

## Supplementary Materials

### A Experiment Details

This section provides additional detail for the experiments presented in §3.

**Datasets** We use the MS MARCO Passage Ranking task unmodified. We use the data from the XOR-Retrieve task (part of XOR-TyDi benchmark), but pre-translate all queries to English. All systems use the same set of pre-translated queries. Table 5 lists the number of training and dev examples for each dataset. We refer to the original papers for details on filters for personally identifiable information and offensive content, as well as domain coverage. We believe we have used all datasets in accordance with their licensing terms.<sup>6</sup>

Dataset	Training	Dev
MS MARCO	808731	6980
XOR-TyDi	15250	2113

Table 5: Example Counts for Training and Dev Sets in MS MARCO and XOR-TyDi

**Software** We use commit [7b5aa6c](#) of PrimeQA and commit [d96f5f1](#) of SPLADE. We use the provided pip environment files provided by PrimeQA (shared across BM25, DPR, and PLAID ColBERTv2) and SPLADE. The only modification we made to the respective environments was upgrading the PyTorch version in both cases to 1.13. We use Python version 3.9.13 for all experiments. We believe we have used all software in accordance with their licensing terms.<sup>7</sup>

**Hyperparameters** Table 6 lists the maximum query and passage lengths used for each neural model:

Model	$ Q $	$ D $
DPR	32	128
PLAID ColBERTv2	32	300
BT-SPLADE-Large	256	256

Table 6: Maximum query and passage lengths used for each neural model as measured in number of tokens.

**Methodology** We run 10 warm-up iterations for each system to mitigate noise from the initial ramp-up phase. We used Docker containers to ensure precise resource allocations across CPU threads, GPUs, and memory. Our experiments are conducted on AWS instances. The times to instantiate the instance and load model environments are not included in latency calculations.

**Model Pre-training and Finetuning** The BM25 model used in our experiments was not pretrained or finetuned for either MSMARCO or XOR-TyDi. Our DPR model used the *facebook/dpr-question\_encoder-multiset-base* and *facebook/dpr-ctx\_encoder-multiset-base* pretrained models and finetunes them on the MSMARCO training set; for XOR-TyDi, our DPR model is not finetuned beyond the original configuration. For BT-SPLADE-Large, we use the *naver/efficient-splade-VI-BT-large-doc* and *naver/efficient-splade-VI-BT-large-query* pretrained models and finetune them on the MSMARCO training set; for XOR-TyDi, we do not finetune them. For PLAID, we use the original model given in [Santhanam et al. \(2022a\)](#) and finetune it using the MSMARCO training set; for XOR-TyDi, we do not finetune the model. When finetuning for MS MARCO or XOR-TyDi, we finetuned for three epochs.

### B Additional Dynascore-Based Leaderboards

<sup>6</sup>MS MARCO is available without license for non-commercial use. XOR-TyDi is distributed under the CC BY-SA 4.0 license.

<sup>7</sup>PrimeQA is distributed under the Apache 2.0 license. ColBERTv2 is distributed under the MIT license.



	System	Hardware	Dynascore
1	ColBERTv2-M	16 CPU, 32 GB memory	19.127
2	ColBERTv2-S	16 CPU, 32 GB memory	19.118
3	ColBERTv2-L	16 CPU, 32 GB memory	18.857
4	ColBERTv2-S	1 GPU, 1 CPU, 32 GB memory	18.698
5	BT-SPLADE-L	16 CPU, 32 GB memory	18.637
6	BT-SPLADE-L	1 CPU, 32 GB memory	18.616
7	ColBERTv2-M	1 GPU, 1 CPU, 32 GB memory	18.385
8	ColBERTv2-S	1 CPU, 32 GB memory	17.912
9	BT-SPLADE-L	1 GPU, 1 CPU, 32 GB memory	17.839
10	ColBERTv2-S	1 GPU, 16 CPU, 32 GB memory	17.331
11	ColBERTv2-L	1 GPU, 1 CPU, 32 GB memory	17.080
12	ColBERTv2-M	1 CPU, 32 GB memory	17.060
13	ColBERTv2-M	1 GPU, 16 CPU, 32 GB memory	16.619
14	BT-SPLADE-L	1 GPU, 16 CPU, 32 GB memory	16.062
15	ColBERTv2-L	1 CPU, 32 GB memory	15.858
16	DPR	16 CPU, 32 GB memory	15.635
17	DPR	1 GPU, 1 CPU, 32 GB memory	15.330
18	ColBERTv2-L	1 GPU, 16 CPU, 32 GB memory	14.940
19	DPR	1 CPU, 32 GB memory	14.583
20	DPR	1 GPU, 16 CPU, 32 GB memory	14.252
21	BM25	1 CPU, 4 GB memory	9.263
22	BM25	1 CPU, 32 GB memory	9.263
23	BM25	16 CPU, 32 GB memory	9.248
24	BM25	16 CPU, 4 GB memory	9.246
25	BM25	1 GPU, 1 CPU, 32 GB memory	9.072
26	BM25	1 GPU, 1 CPU, 4 GB memory	9.049
27	BM25	1 GPU, 16 CPU, 32 GB memory	8.565
28	BM25	1 GPU, 16 CPU, 4 GB memory	8.551

(a) Default weighting per Ma et al. 2021: {Accuracy: 0.5, Cost: 0.25, Latency: 0.25}.

	System	Hardware	Dynascore
1	ColBERTv2-M	1 GPU, 16 CPU, 32 GB memory	29.388
2	ColBERTv2-M	1 GPU, 1 CPU, 32 GB memory	29.347
3	ColBERTv2-M	16 CPU, 32 GB memory	29.300
4	ColBERTv2-S	1 GPU, 16 CPU, 32 GB memory	29.267
5	ColBERTv2-S	1 GPU, 1 CPU, 32 GB memory	29.259
6	ColBERTv2-L	1 GPU, 16 CPU, 32 GB memory	29.181
7	ColBERTv2-S	16 CPU, 32 GB memory	29.172
8	ColBERTv2-L	16 CPU, 32 GB memory	29.122
9	ColBERTv2-L	1 GPU, 1 CPU, 32 GB memory	28.961
10	BT-SPLADE-L	16 CPU, 32 GB memory	28.278
11	BT-SPLADE-L	1 CPU, 32 GB memory	28.186
12	BT-SPLADE-L	1 GPU, 1 CPU, 32 GB memory	28.183
13	BT-SPLADE-L	1 GPU, 16 CPU, 32 GB memory	28.175
14	ColBERTv2-S	1 CPU, 32 GB memory	28.045
15	ColBERTv2-M	1 CPU, 32 GB memory	27.422
16	ColBERTv2-L	1 CPU, 32 GB memory	26.407
17	DPR	16 CPU, 32 GB memory	23.634
18	DPR	1 GPU, 1 CPU, 32 GB memory	23.622
19	DPR	1 GPU, 16 CPU, 32 GB memory	23.586
20	DPR	1 CPU, 32 GB memory	22.708
21	BM25	16 CPU, 32 GB memory	13.958
22	BM25	16 CPU, 4 GB memory	13.958
23	BM25	1 CPU, 32 GB memory	13.952
24	BM25	1 CPU, 4 GB memory	13.945
25	BM25	1 GPU, 1 CPU, 32 GB memory	13.944
26	BM25	1 GPU, 1 CPU, 4 GB memory	13.936
27	BM25	1 GPU, 16 CPU, 32 GB memory	13.931
28	BM25	1 GPU, 16 CPU, 4 GB memory	13.930

(c) Cost is not a concern, and low latency is key: {Accuracy: 0.75, Cost: 0.01, Latency: 0.24}.

	System	Hardware	Dynascore
1	ColBERTv2-M	16 CPU, 32 GB memory	35.577
2	ColBERTv2-L	16 CPU, 32 GB memory	35.515
3	ColBERTv2-M	1 GPU, 1 CPU, 32 GB memory	35.429
4	ColBERTv2-S	16 CPU, 32 GB memory	35.344
5	ColBERTv2-S	1 GPU, 1 CPU, 32 GB memory	35.260
6	ColBERTv2-M	1 CPU, 32 GB memory	35.164
7	ColBERTv2-L	1 GPU, 1 CPU, 32 GB memory	35.160
8	ColBERTv2-S	1 CPU, 32 GB memory	35.102
9	ColBERTv2-M	1 GPU, 16 CPU, 32 GB memory	35.076
10	ColBERTv2-S	1 GPU, 16 CPU, 32 GB memory	34.986
11	ColBERTv2-L	1 CPU, 32 GB memory	34.916
12	ColBERTv2-L	1 GPU, 16 CPU, 32 GB memory	34.732
13	BT-SPLADE-L	16 CPU, 32 GB memory	34.151
14	BT-SPLADE-L	1 CPU, 32 GB memory	34.147
15	BT-SPLADE-L	1 GPU, 1 CPU, 32 GB memory	33.992
16	BT-SPLADE-L	1 GPU, 16 CPU, 32 GB memory	33.636
17	DPR	16 CPU, 32 GB memory	28.487
18	DPR	1 GPU, 1 CPU, 32 GB memory	28.426
19	DPR	1 CPU, 32 GB memory	28.277
20	DPR	1 GPU, 16 CPU, 32 GB memory	28.210
21	BM25	1 CPU, 4 GB memory	16.813
22	BM25	1 CPU, 32 GB memory	16.813
23	BM25	16 CPU, 32 GB memory	16.810
24	BM25	16 CPU, 4 GB memory	16.809
25	BM25	1 GPU, 1 CPU, 32 GB memory	16.774
26	BM25	1 GPU, 1 CPU, 4 GB memory	16.770
27	BM25	1 GPU, 16 CPU, 32 GB memory	16.673
28	BM25	1 GPU, 16 CPU, 4 GB memory	16.670

(b) Heavy emphasis on quality: {Accuracy: 0.9, Cost: 0.05, Latency: 0.05}.

	System	Hardware	Dynascore
1	ColBERTv2-S	16 CPU, 32 GB memory	15.138
2	ColBERTv2-M	16 CPU, 32 GB memory	15.108
3	BT-SPLADE-L	1 CPU, 32 GB memory	14.851
4	ColBERTv2-L	16 CPU, 32 GB memory	14.820
5	BT-SPLADE-L	16 CPU, 32 GB memory	14.806
6	ColBERTv2-S	1 GPU, 1 CPU, 32 GB memory	14.375
7	ColBERTv2-S	1 CPU, 32 GB memory	14.146
8	ColBERTv2-M	1 GPU, 1 CPU, 32 GB memory	13.855
9	BT-SPLADE-L	1 GPU, 1 CPU, 32 GB memory	13.584
10	ColBERTv2-M	1 CPU, 32 GB memory	13.363
11	DPR	16 CPU, 32 GB memory	12.451
12	ColBERTv2-L	1 CPU, 32 GB memory	12.278
13	ColBERTv2-S	1 GPU, 16 CPU, 32 GB memory	12.132
14	ColBERTv2-L	1 GPU, 1 CPU, 32 GB memory	12.055
15	DPR	1 GPU, 1 CPU, 32 GB memory	11.962
16	DPR	1 CPU, 32 GB memory	11.537
17	ColBERTv2-M	1 GPU, 16 CPU, 32 GB memory	10.932
18	BT-SPLADE-L	1 GPU, 16 CPU, 32 GB memory	10.687
19	DPR	1 GPU, 16 CPU, 32 GB memory	10.231
20	ColBERTv2-L	1 GPU, 16 CPU, 32 GB memory	8.365
21	BM25	1 CPU, 4 GB memory	7.408
22	BM25	1 CPU, 32 GB memory	7.401
23	BM25	16 CPU, 32 GB memory	7.371
24	BM25	16 CPU, 4 GB memory	7.369
25	BM25	1 GPU, 1 CPU, 32 GB memory	7.095
26	BM25	1 GPU, 1 CPU, 4 GB memory	7.065
27	BM25	1 GPU, 16 CPU, 32 GB memory	6.278
28	BM25	1 GPU, 16 CPU, 4 GB memory	6.256

(d) Cost is a very significant concern: {Accuracy: 0.4, Cost: 0.4, Latency: 0.2}.

Table 7: Dynascores for MS MARCO, for different weightings of the metrics.

	System	Hardware	Dynascore
1	ColBERTv2-L	16 CPU, 64 GB memory	21.241
2	BT-SPLADE-L	16 CPU, 64 GB memory	21.119
3	ColBERTv2-M	16 CPU, 64 GB memory	21.063
4	BT-SPLADE-L	1 CPU, 64 GB memory	20.753
5	ColBERTv2-M	1 GPU, 16 CPU, 64 GB memory	20.255
6	BT-SPLADE-L	1 GPU, 16 CPU, 64 GB memory	20.123
7	ColBERTv2-M	1 GPU, 1 CPU, 64 GB memory	19.904
8	ColBERTv2-L	1 GPU, 16 CPU, 64 GB memory	19.700
9	ColBERTv2-S	16 CPU, 64 GB memory	19.649
10	BT-SPLADE-L	1 GPU, 1 CPU, 64 GB memory	19.380
11	ColBERTv2-S	1 GPU, 16 CPU, 64 GB memory	19.157
12	ColBERTv2-L	1 GPU, 1 CPU, 64 GB memory	19.123
13	ColBERTv2-S	1 GPU, 1 CPU, 64 GB memory	18.723
14	ColBERTv2-S	1 CPU, 64 GB memory	15.934
15	BM25	1 CPU, 64 GB memory	12.635
16	BM25	16 CPU, 64 GB memory	12.630
17	BM25	1 GPU, 16 CPU, 64 GB memory	11.847
18	BM25	1 GPU, 1 CPU, 64 GB memory	11.794
19	ColBERTv2-M	1 CPU, 64 GB memory	11.563
20	ColBERTv2-L	1 CPU, 64 GB memory	7.708
21	DPR	1 GPU, 1 CPU, 64 GB memory	7.452
22	DPR	1 GPU, 16 CPU, 64 GB memory	7.386
23	DPR	16 CPU, 64 GB memory	7.188
24	DPR	1 CPU, 64 GB memory	5.442

(a) Default weighting per [Ma et al. 2021](#):  
{Accuracy: 0.5, Cost: 0.25, Latency: 0.25}.

	System	Hardware	Dynascore
1	ColBERTv2-L	1 GPU, 16 CPU, 64 GB memory	34.073
2	ColBERTv2-L	1 GPU, 1 CPU, 64 GB memory	33.859
3	ColBERTv2-L	16 CPU, 64 GB memory	33.385
4	ColBERTv2-M	1 GPU, 16 CPU, 64 GB memory	33.149
5	ColBERTv2-M	1 GPU, 1 CPU, 64 GB memory	33.020
6	ColBERTv2-M	16 CPU, 64 GB memory	32.609
7	BT-SPLADE-L	16 CPU, 64 GB memory	32.076
8	BT-SPLADE-L	1 GPU, 16 CPU, 64 GB memory	32.036
9	BT-SPLADE-L	1 GPU, 1 CPU, 64 GB memory	31.752
10	BT-SPLADE-L	1 CPU, 64 GB memory	31.718
11	ColBERTv2-S	1 GPU, 16 CPU, 64 GB memory	30.689
12	ColBERTv2-S	1 GPU, 1 CPU, 64 GB memory	30.532
13	ColBERTv2-S	16 CPU, 64 GB memory	30.239
14	ColBERTv2-S	1 CPU, 64 GB memory	26.788
15	ColBERTv2-M	1 CPU, 64 GB memory	23.835
16	ColBERTv2-L	1 CPU, 64 GB memory	20.881
17	BM25	16 CPU, 64 GB memory	19.276
18	BM25	1 CPU, 64 GB memory	19.264
19	BM25	1 GPU, 16 CPU, 64 GB memory	19.258
20	BM25	1 GPU, 1 CPU, 64 GB memory	19.243
21	DPR	1 GPU, 1 CPU, 64 GB memory	12.305
22	DPR	1 GPU, 16 CPU, 64 GB memory	12.277
23	DPR	16 CPU, 64 GB memory	11.558
24	DPR	1 CPU, 64 GB memory	9.913

(c) Cost is not a concern, and low latency is key:  
{Accuracy: 0.75, Cost: 0.01, Latency: 0.24}.

	System	Hardware	Dynascore
1	ColBERTv2-L	16 CPU, 64 GB memory	42.200
2	ColBERTv2-L	1 GPU, 16 CPU, 64 GB memory	41.892
3	ColBERTv2-L	1 GPU, 1 CPU, 64 GB memory	41.777
4	ColBERTv2-M	16 CPU, 64 GB memory	40.557
5	ColBERTv2-M	1 GPU, 16 CPU, 64 GB memory	40.395
6	ColBERTv2-M	1 GPU, 1 CPU, 64 GB memory	40.325
7	ColBERTv2-L	1 CPU, 64 GB memory	39.494
8	BT-SPLADE-L	16 CPU, 64 GB memory	39.048
9	BT-SPLADE-L	1 CPU, 64 GB memory	38.975
10	BT-SPLADE-L	1 GPU, 16 CPU, 64 GB memory	38.849
11	BT-SPLADE-L	1 GPU, 1 CPU, 64 GB memory	38.700
12	ColBERTv2-M	1 CPU, 64 GB memory	38.657
13	ColBERTv2-S	16 CPU, 64 GB memory	37.362
14	ColBERTv2-S	1 GPU, 16 CPU, 64 GB memory	37.263
15	ColBERTv2-S	1 GPU, 1 CPU, 64 GB memory	37.177
16	ColBERTv2-S	1 CPU, 64 GB memory	36.619
17	BM25	1 CPU, 64 GB memory	23.599
18	BM25	16 CPU, 64 GB memory	23.598
19	BM25	1 GPU, 16 CPU, 64 GB memory	23.441
20	BM25	1 GPU, 1 CPU, 64 GB memory	23.431
21	DPR	1 GPU, 1 CPU, 64 GB memory	15.010
22	DPR	1 GPU, 16 CPU, 64 GB memory	14.997
23	DPR	16 CPU, 64 GB memory	14.958
24	DPR	1 CPU, 64 GB memory	14.608

(b) Heavy emphasis on quality:  
{Accuracy: 0.9, Cost: 0.05, Latency: 0.05}.

	System	Hardware	Dynascore
1	BT-SPLADE-L	16 CPU, 64 GB memory	16.853
2	ColBERTv2-L	16 CPU, 64 GB memory	16.832
3	ColBERTv2-M	16 CPU, 64 GB memory	16.743
4	BT-SPLADE-L	1 CPU, 64 GB memory	16.565
5	ColBERTv2-S	16 CPU, 64 GB memory	15.638
6	BT-SPLADE-L	1 GPU, 16 CPU, 64 GB memory	15.259
7	ColBERTv2-M	1 GPU, 16 CPU, 64 GB memory	14.954
8	ColBERTv2-M	1 GPU, 1 CPU, 64 GB memory	14.491
9	ColBERTv2-S	1 GPU, 16 CPU, 64 GB memory	14.444
10	BT-SPLADE-L	1 GPU, 1 CPU, 64 GB memory	14.291
11	ColBERTv2-S	1 GPU, 1 CPU, 64 GB memory	13.871
12	ColBERTv2-L	1 GPU, 16 CPU, 64 GB memory	13.714
13	ColBERTv2-L	1 GPU, 1 CPU, 64 GB memory	12.958
14	ColBERTv2-S	1 CPU, 64 GB memory	12.566
15	BM25	1 CPU, 64 GB memory	10.088
16	BM25	16 CPU, 64 GB memory	10.069
17	ColBERTv2-M	1 CPU, 64 GB memory	8.843
18	BM25	1 GPU, 16 CPU, 64 GB memory	8.806
19	BM25	1 GPU, 1 CPU, 64 GB memory	8.731
20	DPR	16 CPU, 64 GB memory	5.669
21	ColBERTv2-L	1 CPU, 64 GB memory	5.582
22	DPR	1 GPU, 1 CPU, 64 GB memory	5.450
23	DPR	1 GPU, 16 CPU, 64 GB memory	5.368
24	DPR	1 CPU, 64 GB memory	4.244

(d) Cost is a very significant concern:  
{Accuracy: 0.4, Cost: 0.4, Latency: 0.2}.

Table 8: Dynascores for XOR-TyDi, for different weightings of the metrics.

## ACL 2023 Responsible NLP Checklist

---

### A For every submission:

- A1. Did you describe the limitations of your work?  
*Section 6*
- A2. Did you discuss any potential risks of your work?  
*Sections 4.3 and 6, both of which acknowledge that particular measurement choices come with their own risks*
- A3. Do the abstract and introduction summarize the paper's main claims?  
*Abstract and Section 1*
- A4. Have you used AI writing assistants when working on this paper?  
*Left blank.*

### B Did you use or create scientific artifacts?

*Section 3, Appendix A*

- B1. Did you cite the creators of artifacts you used?  
*Sections 2 and 3*
- B2. Did you discuss the license or terms for use and / or distribution of any artifacts?  
*Appendix A*
- B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?  
*Appendix A*
- B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?  
*Not applicable. We relied entirely on publicly available and widely used datasets, and we refer readers to the original sources for these details.*
- B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?  
*Not applicable. We relied entirely on publicly available and widely used datasets, and we refer readers to the original sources for these details.*
- B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.  
*Appendix A*

---

*The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.*

**C  Did you run computational experiments?**

*Section 3, Appendix A*

- C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?

*Section 3, Appendix A (These issue are a key focus of our paper)*

- C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?

*Section 3, Appendix A (These issue are a key focus of our paper)*

- C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?

*Section 3, Appendix A (These issue are a key focus of our paper)*

- C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?

*Appendix A*

**D  Did you use human annotators (e.g., crowdworkers) or research with human participants?**

*Left blank.*

- D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?

*Not applicable. Left blank.*

- D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?

*Not applicable. Left blank.*

- D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?

*Not applicable. Left blank.*

- D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?

*Not applicable. Left blank.*

- D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?

*Not applicable. Left blank.*