Jina-ColBERT-v2: A General-Purpose Multilingual Late Interaction Retriever

Rohan Jha[1*] and **Bo Wang**[2] and **Michael Günther**[2] and **Georgios Mastrapas**[2] and **Saba Sturua**[2] and **Isabelle Mohr**[2] and **Andreas Koukounas**[2] and **Mohammad Kalim Akram**[2] and **Nan Wang**[2] and **Han Xiao**[2]
[1]The University of Texas at Austin, Austin, Texas, USA
[2]Jina AI GmbH, Prinzessinnenstr. 19-20, 10969 Berlin, Germany
research@jina.ai

Abstract

Multi-vector dense models, such as ColBERT, have proven highly effective in information retrieval. ColBERT's late interaction scoring approximates the joint query-document attention seen in cross-encoders while maintaining inference efficiency closer to traditional dense retrieval models, thanks to its bi-encoder architecture and recent optimizations in indexing and search. In this work we propose a number of incremental improvements to the ColBERT model architecture and training pipeline, using methods shown to work in the more mature single-vector embedding model training paradigm, particularly those that apply to heterogeneous multilingual data or boost efficiency with little tradeoff. Our new model, Jina-ColBERT-v2, demonstrates strong performance across a range of English and multilingual retrieval tasks.

1 Introduction

Neural retrieval has gained popularity in recent years following the arrival of capable pre-trained language models (PLMs) (Devlin et al., 2019; Liu et al., 2019; Clark et al., 2020). Two types of approaches have been employed to apply PLMs to retrieval. Sparse neural retrieval systems, such as SPLADE (Formal et al., 2021), represent texts as weighted bags of words that are interpreted as sparse high-dimensional vectors for maximum inner product search (MIPS). Dense retrievers similarly encode queries and documents as *dense* vectors, capturing relevance signals through spatial relationships extending beyond exact term matching.

Most dense retrievers encode a query or document as a single vector, commonly the result of mean-pooling or the [CLS]-embedding over the transformer's final layer token embeddings. In contrast, recent multi-vector retrievers like ColBERT (Khattab and Zaharia, 2020) generalize this embedding process to maintain an embedding for each token, computing relevance scores as a function of the similarities of query and document tokens instead. To make the ColBERT usable in practice, the output dimensionality is restricted to be much smaller than the single-vector models. This approach has the benefit of remaining compatible with much of the vector similarity infrastructure that makes single-vector methods efficient, but requires more space to store even a smaller embedding per token and compute at inference time to aggregate token interactions into a single score. This late interaction over token embeddings achieves greater in-domain performance and tends to be more robust out-of-domain than single-vector similarity. While ColBERTv2 is trained only on English MSMARCO triplets (Bajaj et al., 2016) and has a monolingual BERT backbone, making it incapable of multilingual retrieval, some previous works extend the model to multilingual retrieval.

ColBERT-XM (Louis et al., 2024) does this by using parameter extensions for each additional language, and (Lawrie et al., 2023) trains solely on machine-translated English MSMARCO data to get effective heterogeneous multilingual performance. These approaches, however, come with trade-offs in terms of model usability and training data diversity. Other multilingual multi-vector models like BGE-M3 (Chen et al., 2024) produce extremely large token representations that limit their practical utility for first-stage retrieval.

In this work, we propose Jina-ColBERT-v2, which introduces an improved training recipe for ColBERT models with the following features:

Training with diverse weakly-supervised data: We additionally pretrain our modified PLM with rotary position embedding and train on large-scale unlabeled text pairs from various corpus with a weakly-supervised single-vector contrastive objective. A second-stage of ColBERT finetuning with labeled triplet data and supervised distillation is used to further boost its performance.

General multilingual performance: We train with data from a variety of high- and low-resource languages using both labeled and unlabeled data, including human- and machine-translated training data, and show that this improves even out-of-domain multilingual performance.

Inference-agnostic efficiency: We introduce

[*] Work done while at Jina AI.

multiple sizes of linear projection heads, jointly trained using the non-weight tying variant of Matryoshka Representation Loss (Kusupati et al., 2022), enabling the selection of token embedding size at inference time with minimal performance degradation. We demonstrate that reducing the embedding dimensionality in half from 128 to 64 yields only a minor performance tradeoff. Additionally, our flash-attention optimized backbone, Jina-XLM-RoBERTa provides further free performance improvement during inference.

Our experimental results show competitive retrieval performance across both English and multilingual benchmarks. We also present controlled experiments demonstrating the benefits, or lack thereof, of the training modifications we consider in developing our training recipe.

2 Related Work

In this section, we discuss related work in single- and multi-vector retrieval, as well as the non-English late-interaction retrievers from which our training recipe draws inspiration.

2.1 Single-Vector Retrieval

Single-vector encoder models have demonstrated their potential as general-purpose embedding models across a number of downstream tasks (Muennighoff et al., 2023). When used in a bi-encoder retrieval model, they asymmetrically encode queries and documents as separate dense vectors, and measure their pairwise relevance as the cosine similarity between the vectors. Owing to their strong in-domain performance and straightforward inference scheme, there has been a growing focus on improving their training. Studies demonstrate that large-scale unsupervised pair training utilizing in-batch negatives, followed by a small-scale triplet finetuning stage, significantly improves performance compared to a dense retriever trained solely on triplet data (Li et al., 2023; Günther et al., 2023). Other works have incorporated asymmetric task-specific instructions for queries and documents to further enhance performance (Wang et al., 2024) and demonstrated the efficacy of using synthetically generated training data, including using diverse task instructions and machine translations, to further improve model representations. (Wang et al., 2023; Lee et al., 2024)

2.2 Multi-Vector Retrieval

Multi-vector retrievers like ColBERT also employ a bi-encoder structure, but queries and passages are represented by a collection of smaller token embeddings rather than one large vector. As such, ColBERTv2's training uses many of the same techniques as state-of-the-art single-vector models: cross-encoder distillation, multiple negatives per query, and self-mined hard negatives. Recent models have continued to improve on this training recipe, particularly for multilingual or non-English training. BGE-M3 (Chen et al., 2024) adopts the two-stage pairs-to-triplets training pipeline, and does self-knowledge distillation, treating the combination of its sparse, dense, and multi-vector scores as the teacher score.

2.3 Multilingual Retrieval

Owing to the quality of English-based pre-trained models (BERT) and annotated data (MSMARCO), many advances in neural retrieval have been applied first to the monolingual English setting (Karpukhin et al., 2020; Xiong et al., 2020; Khattab and Zaharia, 2020). Researchers, however, have also made advances in non-English capabilities.

On the modeling front, multilingual PLMs like mBERT (Devlin et al., 2019) and later XLM-RoBERTa (Conneau et al., 2020) have expanded pre-training to include text in up to 100 languages, including in cross-language contexts. For multilingual retrieval data, there are two approaches: natural and translated. Datasets like Mr-Tydi and MIRACL (Zhang et al., 2021, 2023b) are built from human-generated and annotated queries, whereas mMARCO (Bonifacio et al., 2022) is a collection of machine-translated copies of MSMARCO which inherit their judgments from the original dataset. The former method tends to be of higher quality and lacks the subtle distributional/idiomatic errors, dubbed "translationese", that the latter sometimes exhibits. Naturally, however, human generation costs more per example.

Recent multi-vector work has also proposed further modifications along the dimensions of architecture and data. ColBERT-XM (Louis et al., 2024) addresses the so-called *curse of multilinguality* (Conneau et al., 2020), the performance degradation of models pre-trained on too many tasks, with shared- and per-language parameters that allow for more robust zero-shot language transfer and post-hoc language extension. On the data approach, ColBERT-X (Nair et al., 2022; Lawrie et al., 2023; Yang et al., 2024) uses language-mixed batches of machine-translated English data, and BGE-M3 (Chen et al., 2024) curates unsupervised and high-quality supervised corpora of diverse multilingual training data.

3 Training Overview

`Jina-ColBERT-v2`'s training paradigm has three parts:

1. **Modified Encoder Architecture**: We use a modified encoder backbone, derived from XLM-RoBERTa with improvements made to its architecture and pre-training regime. We further

extend ColBERT's linear projection head by jointly training a collection of different-size heads for embedding size reduction.

2. **Pair Training**: To learn from the semantic structure of large quantities of diverse data in many languages, we first train our encoder model on weakly supervised text pairs from a variety of embedding datasets.
3. **Triplet Training**: Our model is further finetuned using retrieval examples in many languages with both positives and hard negatives, supervised by a highly-capable multilingual cross-encoder.

The following sections describe our experiments on these three components of training `Jina-ColBERT-v2`.

4 Architecture

4.1 Backbone Improvements

Following many prior single- and multi-vector multilingual training efforts, we adopted XLM-RoBERTa as our backbone model due to its strong performance across various downstream tasks (Nair et al., 2022; Louis et al., 2024; Chen et al., 2024). To improve the efficiency, we enhance the XLM-RoBERTa architecture with flash attention (Dao, 2024).

We replace the absolute positional embeddings with rotary positional embeddings (RoPE, Su et al. (2023)), which are empirically understood to be better. They also have the advantage of supporting context lengths far longer than 512 tokens, although we do not explicitly focus on long-context in this work. To warm up its new positional embeddings, we continued pre-training the modified backbone with the same masked language modeling objective for 160,000 steps on the RefinedWeb dataset (Penedo et al., 2023), a modern, high-quality corpus, under the masked language modeling objective. During this pre-training phase, we set the maximum sequence length to 8,192 tokens with a rotary base of 10,000 and employed whole-word-masking (Devlin et al., 2019), masking out 30% of the tokens. We call this modified language model Jina-XLM-RoBERTa.

4.2 Multiple Linear Heads

To reduce index sizes, ColBERT includes a linear head that projects its token embeddings from the hidden dimension of its language model down to a lower dimension ($768 \rightarrow 128$). As a notable exception, BGE-M3's multi-vector retrieval does not take this step, keeping its token embeddings at a full 1024 dimensions.

We jointly train six linear heads with dimensions $d \in \{64, 96, 128, 256, 512, 768\}$ using Matryoshka Representation Loss (MRL, Kusupati et al. (2022)). This allows users to choose greater or lesser space efficiency, with an associated performance trade-off. Figure 1 quantifies this tradeoff, showing the strong performance preservation of our reduced-dimension linear heads. Halving the token dimension ($128 \rightarrow 64$) only causes its nDCG@10 to drop by 0.01 (1.59%). We unfortunately find that MRL's weight-tying efficient variant (MRL-E), where losses are computed on *truncations* of the same token vector does not preserve performance well, which we hypothesize is a consequence of the already-low projected dimension of the original ColBERT formulation.

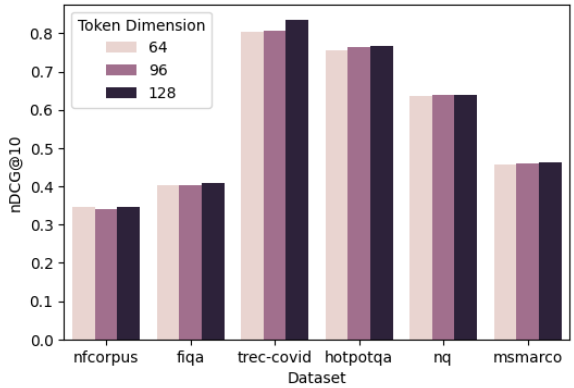

Figure 1: nDCG@10 scores for BEIR datasets using 64-, 96-, and 128-dimension linear projection heads for token embeddings.

5 Pair Training

To leverage an abundance of text pairs with varying richness of semantic structure, we draw inspiration from common practices in single-vector embedding model training and begin with training on these text pairs, focusing on optimizing the embedding model's performance on general semantic similarity and relatedness tasks. This weakly-supervised stage is in contrast to previous ColBERT works, which typically start directly from a PLM like BERT with triplet training on 32-way or 64-way retrieval triplets consisting of a query, a positive passage, and multiple mined negatives.

5.1 Data Composition

Our pair training data consists of a broad range of weakly supervised datasets harvested from the web. We adjusted sampling rates across different languages and domains based on intuition, resulting in a set of 450 million weakly supervised, semantically related sentence pairs, question-answer pairs, and query-document pairs. Of these 450 million pairs, 50.0% are in English. Our non-English pair-wise datasets contain a diverse collection of 29 major languages, including 3.0% code data, with 4.3% representing cross-lingual data.

5.2 Contrastive Loss

We utilize the same *single-vector* pair-training loss function as described in (Günther et al., 2023). Due to the often symmetric nature of our text pairs, the loss is calculated in both directions. During the pair training stage, we set the temperature $\tau = 0.02$ and used a peak learning rate of 5×10^{-5} with a warm-up period of 1,000 steps. The model was trained using the Adam optimizer for 100,000 steps with a global batch size of 16,384.

6 Triplet Training

6.1 Data Composition

Our triplet dataset consists of 1) high-quality, human-annotated research datasets such as MSMARCO, DuReader, and MIRACL (Bajaj et al., 2016; He et al., 2018; Zhang et al., 2023b) with diversely mined negatives 2) high-quality datasets like MSMARCO and NQ translated from English into Chinese, French, German, Japanese, Russian and Spanish, following our previous work (Mohr et al., 2024) and 3) synthetically generated datasets to address common failure modes of dense vector models such as negation and to cover niche domains like legal IR.

The triplet dataset covers 14 widely used languages, with a strong emphasis on Arabic, Chinese, English, French, German, Japanese, Russian, and Spanish. We sample the datasets to create a language distribution similar to that used in pair training. English accounts for 45.9% of the triplets, with 52.1% roughly evenly split between the mentioned high-resource non-English languages and a small 2.0% share for lower-resource languages.

Notably, owing to the limitations of our various sources of data, we train on triplets with only 7 negatives per example, in contrast to the 32- or 64-way triplets of ColBERTv2.

6.2 Supervision Loss

Following ColBERTv2, we finetune our pair-trained checkpoint on samples with hard negatives using a KL divergence loss function to distill soft labels from the teacher model. For the teacher model, we use `jina-reranker-v2-base-multilingual`[1], a highly capable multilingual cross encoder.

This stage trains for 100,000 steps with a batch size of 32 and a cosine decay learning rate schedule with 5% warm-up that peaks at 1×10^{-5}. We use pure BFLOAT-16 precision, and apply magnitude-based gradient clipping with a threshold of 1 for stability.

7 Results

We evaluate `Jina-ColBERT-v2` on four widely used benchmarks, BEIR, LoTTE, and MIRACL and mMARCO. For general English performance, we use the same subset of 14 retrieval and text-similarity tasks from the BEIR benchmark as in Santhanam et al. (2022). Additionally, we assess performance on the LoTTE benchmark, which focuses on long-tail queries, and the MIRACL and mMARCO benchmarks (Zhang et al., 2023b; Bonifacio et al., 2022), which assess non-English retrieval performance. We report nDCG@10 for the BEIR and MIRACL collections, MRR@10 for mMARCO, and Success@5 for LoTTE. Scores are reported on the test split for BEIR, development split for MIRACL and mMARCO, and search test split for LoTTE. We use the same maximum query/document lengths as reported in Santhanam et al. (2022), and use the default (32/300) for MIRACL and mMARCO.

Table 1 shows `Jina-ColBERT-v2`'s strong English performance compared to ColBERTv2, while still trailing the monolingual answerai-colbert-small-v1. Notably, however, we perform well below ColBERTv2 on ArguAna (**ar**), which we might attribute to either its unusual task: *counterargument retrieval* being at odds with our retrieval-heavy triplet training data distribution, or as an indication of the limitation of our stronger augmentation attention (discussed in Section 8.4) when applied to much longer (300 token) queries. Similarly for LoTTE, we see in Table 2 an improvement over ColBERTv2.

Table 3 compares `Jina-ColBERT-v2` to BM25, mDPR, and BGE-M3. While we handily outperform BM25 and zero-shot mDPR (Zhang et al., 2023b) as expected, our model is slightly outperformed by the finetuned mDPR (Zhang et al., 2023a). For context, each mDPR-FT is only tuned on one language, rather than many like ours which may suffer to some extent from the *curse of multilinguality*.

Finally, comparing against ColBERT-XM's zero-shot evaluation on mMARCO in Table 4, we see a strong improvement across the board, including on languages whose mMARCO training set does not occur in our pair or triplet training data (dt, hi, id, it, pt, vi).

8 Ablation Studies

In this section we present short ablation studies on modifications to three various aspects of ColBERT modeling and training.

8.1 Efficient Evaluation

Due to the compute and time costs of indexing corpora containing tens of millions of documents, evaluating every model checkpoint and ablation on every task is not feasible. Therefore, we follow recent works (Clavié, 2024; Merrick et al., 2024) by comparing models' quality on smaller sampled-corpus versions of HotpotQA, NQ, MS MARCO, and MIRACL (Chinese, French, German, Japanese, Spanish). These sampled corpora are constructed by combining the

[1] https://huggingface.co/jinaai/jina-reranker-v2-base-multilingual

BEIR	avg	nf	fi	tc	ar	qu	sd	sf	to	db	fe	cf	hp	nq
BM25	44.0	32.5	23.6	65.6	31.5	78.9	15.8	66.5	**36.7**	31.3	75.3	21.3	60.3	32.9
ColBERTv2	49.6	33.7	35.4	72.6	46.5	85.5	15.4	68.9	26.0	45.2	78.5	17.6	67.5	52.4
answerai-v1	**55.7**	**37.3**	**41.2**	**84.6**	**50.1**	87.7	18.4	**74.8**	25.7	45.6	**91.0**	**33.1**	76.1	59.1
Ours	53.1	34.6	40.8	83.4	36.6	**88.7**	**18.6**	67.8	27.4	**47.1**	80.5	23.9	**76.6**	**64.0**

Table 1: Comparison of nDCG@10 scores between BM25, ColBERTv2, answer-colbert-small and Jina-ColBERT-v1 and `Jina-ColBERT-v2` on the BEIR test set. **nf** for NFCorpus, **fi** for FIQA (Fact In Question Answering), **tc** for TREC-COVID (Text Retrieval Conference COVID), **ar** for Arguana, **qu** for Quora, **sd** for SciDocs, **sf** for SciFact, **to** for Webis-Touche, **db** for DBpedia-Entity, **fe** for FEVER (Fact Extraction and Verification), **cf** for Climate-FEVER, **hp** for HotpotQA, and **nq** for Natural Questions

LoTTE	avg	Life.	Rec.	Wri.	Sci.	Tech.
BM25	67.8	80.2	68.5	74.7	53.6	61.9
ColBERTv2	72.0	84.7	72.3	80.1	56.7	66.1
Ours	**76.4**	**87.0**	**77.6**	**83.8**	**60.5**	**73.0**

Table 2: Comparison of Success@5 of various models across different LoTTE search query subsets.

top 250 BM25-retrieved[2] passages with all judged passages. We observe good agreement between the sampled-corpus evaluation scores and the full-fidelity ones when used to make binary or ranking-based model comparisons, but we leave a more rigorous analysis of this observation to future work. We only use the sampled corpora for ablation studies. For the final model, we evaluate on the full version of every dataset.

8.2 Task Instructions

Inspired by the use of instruction prefixes in single-vector works like Su et al. (2022), we experimented with adding task-specific natural language instructions for retrieval (RET), and question answering (QA), and semantic text similarity (STS). However, results in Table 5 show a generally negative effect across most BEIR datasets. We hypothesize that this is because instructions are not well-suited for late interaction models, which operate at the token level. Any embedding conditioning that the instructions might provide likely becomes less effective when aggregated at the token similarity level. Furthermore, these instructions occupy valuable space within the system's fixed token capacity.

8.3 Score Normalization

Recently, Clavié (2024) applied min-max normalization to both the student and teacher scores before computing the KL loss. This adjustment brings the score distributions of the ColBERT model and its CE teacher into closer alignment, as the original score distribution for ColBERT theoretically ranges from zero to the number of query tokens, and is model-dependent for the teacher CE. Our experiment presented in Table 6, however, shows this method to have inconclusive benefit to nDCG@10 on the BEIR and MIRACL datasets when applied to our model. We consider this result to be understandable given Clavié (2024)'s very small observed effect.

8.4 Query Augmentation Attention

An important feature of ColBERT's implementation is its query augmentation mechanism. By padding queries with `[MASK]` tokens to a uniform length, ColBERT uses BERT's masked language modeling ability to produce additional soft term embeddings which interact with document token embeddings during MaxSim scoring. However, prior ColBERT models do not modify the attention mask to allow query tokens to attend to the mask tokens, which some hypothesize might harm generalization by making this augmentation feature too integral to the embedding process. Our controlled triplet training experiment in Table 7, however, demonstrates a positive effect across a variety of tasks, with particular benefit to non-English tasks in MIRACL. We therefore allow this attention in our training and inference.

9 Conclusion

This work presents `Jina-ColBERT-v2`, a capable multilingual ColBERT model that is the result of improvements to its architecture and training process. We implement modifications to the model architecture that yield efficiency gains with effectively no downside, and subsequently train it on a heterogeneous mix of data of varying tasks, languages, and supervision structures in order to bolster its performance as a general purpose retriever. Our ablation experiments demonstrate the sensitivity of ColBERT to modifications to its representations.

We hope that our work will support future multilingual ColBERT development, and prompt further exploration into the properties and optimal configuration of its query augmentation mechanism. We are also encouraged by the many inference-only optimization works on ColBERT representations, and

[2] We use the standard pre-built Lucene indices in Pyserini (Lin et al., 2021) for MIRACL found at `https://github.com/castorini/pyserini`, and use BM25s (Lù, 2024) for BEIR.

MIRACL	avg	ar	bn	de	es	en	fa	fi	fr	hi	id	ja	ko	ru	sw	te	th	yo	zh
BM25	38.5	48.1	50.8	22.6	31.9	35.1	33.3	55.1	18.3	45.8	44.9	36.9	41.9	33.4	38.3	49.4	48.4	40.6	18.0
mDPR-ZS	41.8	49.9	44.3	49.0	47.8	39.4	48.0	47.2	43.5	38.3	27.2	43.9	41.9	40.7	29.9	35.6	35.8	39.6	51.2
mDPR-FT	**62.7**	72.5	68.4	-	48.8	56.5	**59.3**	71.4	**58.9**	51.6	49.6	**64.2**	59.0	59.7	**68.5**	**80.4**	69.5	-	**65.0**
Ours	62.3	**75.3**	**75.0**	**50.4**	**53.8**	**57.0**	56.3	**74.0**	54.1	**60.0**	**54.7**	63.2	**67.1**	**64.3**	49.9	74.2	**77.2**	**62.3**	52.3

Table 3: Comparison of nDCG@10 scores for BM25, mDPR-ZeroShot (ZS), mDPR-FineTuned (FT), and `Jina-ColBERT-v2` models on the MIRACL dev set across various languages.

mMARCO	avg	ar	de	nl	es	fr	hi	id	it	ja	pt	ru	vi	zh
BM-25	13.9	11.1	13.6	14.0	15.8	15.5	13.4	14.9	15.3	14.1	15.2	12.4	13.6	11.6
ColBERT-XM	25.4	19.5	27.0	27.5	28.5	26.9	23.8	26.3	26.5	24.1	27.6	25.1	22.6	24.6
Ours	**31.3**	**27.2**	**33.1**	**33.0**	**34.1**	**33.5**	**30.9**	**31.9**	**33.7**	**27.6**	**33.7**	**29.8**	**28.7**	**30.2**

Table 4: Comparison of mRR@10 scores between BM25, ColBERT-XM and `Jina-ColBERT-v2` models on the mMARCO dev set across various languages.

	RET								QA			STS		
	nf	tc	sf	to	db	fe	cf	ms*	fq	hp*	nq*	ar	qu	sd
Mark.	32.4	59.3	**67.9**	**19.3**	**35.3**	**67.1**	**18.3**	**34.4**	**37.5**	**25.9**	40.8	**37.5**	**86.1**	**18.4**
Inst.	**32.9**	**63.2**	67.5	18.8	33.9	64.4	16.7	34.0	37.1	24.9	**42.9**	34.2	86.0	17.9

Table 5: nDCG@10 scores on BEIR datasets, grouped by task type (retrieval, question answering, and semantic text similarity) when using natural language instructions versus query/document marker tokens (default). Datasets marked with a * use the BM25-sampled corpus technique discussed in Section 8.1.

	BEIR				MIRACL				
	tc	hp	nq	ms	de	es	fr	ja	zh
Baseline	78.7	**36.6**	**58.0**	**45.4**	57.3	**40.6**	50.7	**63.4**	**63.2**
+ Score Norm.	**80.1**	36.4	56.6	45.1	**57.7**	39.3	**51.3**	61.8	62.5

Table 6: nDCG@10 scores with and without score normalization on a retrieval-oriented subset of BEIR and MIRACL tasks. Results are performed on the BM25-sampled versions of all datasets presented except TREC-COVID (**tc**).

	BEIR				MIRACL				
	tc	hp	nq	ms	de	es	fr	ja	zh
Baseline	77.2	70.4	54.6	37.6	33.3	40.3	35.9	54.9	34.4
+ `[MASK]` attn.	**80.2**	**71.5**	**58.8**	**44.3**	**45.6**	**49.8**	**44.8**	**58.8**	**52.9**

Table 7: nDCG@10 scores with and without query augmentation `[MASK]` token attention on a retrieval-oriented subset of BEIR and MIRACL tasks. Results report full-fidelity scores.

suggest further effort be invested in tying these methods more closely with the models training objective.

10 Acknowledgement

We thank Qi Liu and Jiaxin Mao from Renmin University of China for the contributions to `Jina-ColBERT-v1` and offer the insights about MRL over MRL-E for ColBERT models.

References

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2016. Ms marco: A human generated machine reading comprehension dataset.

Luiz Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. 2022. mMARCO: A Multilingual Version of the MS MARCO Passage Ranking Dataset. ArXiv:2108.13897 [cs].

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu

Lian, and Zheng Liu. 2024. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. ArXiv:2402.03216 [cs].

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. ArXiv:2003.10555 [cs].

Benjamin Clavié. 2024. JaColBERTv2.5: Optimising Multi-Vector Retrievers to Create State-of-the-Art Japanese Retrievers with Constrained Resources. ArXiv:2407.20750 [cs].

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. ArXiv:1911.02116 [cs].

Tri Dao. 2024. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv:1810.04805 [cs].

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking. ArXiv:2107.05720 [cs].

Michael Günther, Jackmin Ong, Isabelle Mohr, Alaeddine Abdessalem, Tanguy Abel, Mohammad Kalim Akram, Susana Guzman, Georgios Mastrapas, Saba Sturua, Bo Wang, et al. 2023. Jina embeddings 2: 8192-token general-purpose text embeddings for long documents. *arXiv preprint arXiv:2310.19923*.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, Xuan Liu, Tian Wu, and Haifeng Wang. 2018. Dureader: a chinese machine reading comprehension dataset from real-world applications.

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. ArXiv:2004.12832 [cs].

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. 2022. Matryoshka representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 30233–30249. Curran Associates, Inc.

Dawn Lawrie, Eugene Yang, Douglas W. Oard, and James Mayfield. 2023. Neural Approaches to Multilingual Information Retrieval. ArXiv:2209.01335 [cs].

Jinhyuk Lee, Zhuyun Dai, Xiaoqi Ren, Blair Chen, Daniel Cer, Jeremy R. Cole, Kai Hui, Michael Boratko, Rajvi Kapadia, Wen Ding, Yi Luan, Sai Meher Karthik Duddu, Gustavo Hernandez Abrego, Weiqiang Shi, Nithi Gupta, Aditya Kusupati, Prateek Jain, Siddhartha Reddy Jonnalagadda, Ming-Wei Chang, and Iftekhar Naim. 2024. Gecko: Versatile Text Embeddings Distilled from Large Language Models. ArXiv:2403.20327 [cs] version: 1.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, pages 2356–2362.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Antoine Louis, Vageesh Saxena, Gijs van Dijck, and Gerasimos Spanakis. 2024. ColBERT-XM: A Modular Multi-Vector Representation Model for Zero-Shot Multilingual Information Retrieval. ArXiv:2402.15059 [cs].

Xing Han Lù. 2024. Bm25s: Orders of magnitude faster lexical search via eager sparse scoring.

Luke Merrick, Danmei Xu, Gaurav Nuti, and Daniel Campos. 2024. Arctic-embed: Scalable, efficient, and accurate text embedding models.

Isabelle Mohr, Markus Krimmel, Saba Sturua, Mohammad Kalim Akram, Andreas Koukounas, Michael Günther, Georgios Mastrapas, Vinit Ravishankar, Joan Fontanals Martínez, Feng Wang, et al. 2024. Multi-task contrastive learning for 8192-token bilingual text embeddings. *arXiv preprint arXiv:2402.17016*.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Suraj Nair, Eugene Yang, Dawn Lawrie, Kevin Duh, Paul McNamee, Kenton Murray, James Mayfield, and Douglas W. Oard. 2022. Transfer learning approaches for building cross-language dense retrieval models. In *Proceedings of the 44th European Conference on Information Retrieval (ECIR)*.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction. ArXiv:2112.01488 [cs].

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A Smith, Luke Zettlemoyer, and Tao Yu. 2022. One embedder, any task: Instruction-finetuned text embeddings. *arXiv preprint arXiv:2212.09741*.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2023. Roformer: Enhanced transformer with rotary position embedding.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2023. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

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:2007.00808 [cs].

Eugene Yang, Dawn Lawrie, and James Mayfield. 2024. Distillation for Multilingual Information Retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2368–2373. ArXiv:2405.00977 [cs].

Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. 2021. Mr. TyDi: A multi-lingual benchmark for dense retrieval. *arXiv:2108.08787*.

Xinyu Zhang, Kelechi Ogueji, Xueguang Ma, and Jimmy Lin. 2023a. Toward best practices for training multilingual dense retrieval models. *ACM Trans. Inf. Syst.*, 42(2).

Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023b. Miracl: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131.