mGTE: Generalized Long-Context Text Representation and Reranking Models for Multilingual Text Retrieval

Xin Zhang[1,2], **Yanzhao Zhang**[1], **Dingkun Long**[1], **Wen Xie**[1], **Ziqi Dai**[1], **Jialong Tang**[1], **Huan Lin**[1]
Baosong Yang[1], **Pengjun Xie**[1], **Fei Huang**[1], **Meishan Zhang**, **Wenjie Li**[2], **Min Zhang**

[1]Alibaba Group, [2]Department of Computing, The Hong Kong Polytechnic University

{linzhang.zx,zhangyanzhao.zyz,dingkun.ldk}@alibaba-inc.com

https://hf.co/Alibaba-NLP/gte-multilingual-base

Abstract

We present systematic efforts in building long-context multilingual text representation model (TRM) and reranker from scratch for text retrieval. We first introduce a text encoder (base size) enhanced with RoPE and unpadding, pre-trained in a native 8192-token context (longer than 512 of previous multilingual encoders). Then we construct a hybrid TRM and a cross-encoder reranker by contrastive learning. Evaluations show that our text encoder outperforms the same-sized previous state-of-the-art XLM-R. Meanwhile, our TRM and reranker match the performance of large-sized state-of-the-art BGE-M3 models and achieve better results on long-context retrieval benchmarks. Further analysis demonstrate that our proposed models exhibit higher efficiency during both training and inference. We believe their efficiency and effectiveness could benefit various researches and industrial applications.

1 Introduction

Text retrieval aims to find relevant passages or documents from a large corpus given a query (Manning, 2008). It is often implemented as a multi-stage process, consisting of two main components: a *retriever* and a *reranker* (Gao et al., 2021a; Zhang et al., 2022; Zhao et al., 2024). The retriever identifies a set of candidate documents that are potentially relevant to the query based on the similarity between their sparse (lexical term weights) or/and dense representations from a text representation model (TRM). While the reranker reorders these retrieved candidates to refine the results based on the relevance score generated by a more precise yet computationally demanding model that processes both the query and a candidate document together.

Recent advances in large language models (LLMs) and retrieval augmented generation (RAG) (Gao et al., 2023) systems have led to an unprecedented surge in demand for versatile, plug-and-play TRMs and rerankers. These new applications
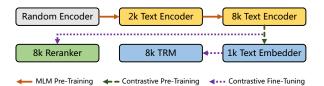

Figure 1: Training pipeline. We first build an 8k long-context multilingual encoder. Then based on it, we train text representation and reranking models for retrieval.

heavily involve processing long and multilingual texts, which could not be addressed by conventional encoder-based models and urgently require upgraded ones. To this end, some resort to enhancing existing multilingual encoders, *e.g.,* XLM-R (Conneau et al., 2020), with extended context window up to 8192 (Chen et al., 2024). Others turn to use multilingual LLMs which already have the required capabilities (Zhang et al., 2023a), but their models might be computationally expensive for self-hosted search services.

In the English community, it has been proven that training long-context encoders from scratch is promising for text retrieval (Günther et al., 2023; Nussbaum et al., 2024). In this work, we continue this journey, presenting systematic efforts in building the long-context multilingual text encoder, TRM, and reranker. We suggest a holistic pipeline (Figure 1) as well as several techniques in modeling and training for multilingual long-context retrieval.

Concretely, we first introduce a text encoder enhanced with RoPE (Su et al., 2024) and unpadding (Portes et al., 2023), pre-trained by masked language modeling (MLM) (Devlin et al., 2019) via a two-stage curriculum for the native 8,192 tokens context. Based on our encoder, we propose a hybrid TRM capable of generating both elastic dense (Kusupati et al., 2022) and sparse vectors for efficient first-stage retrieval, as well as a cross-encoder reranker. We construct them via contrastive learning (Wang et al., 2022; Li et al., 2023) with large-
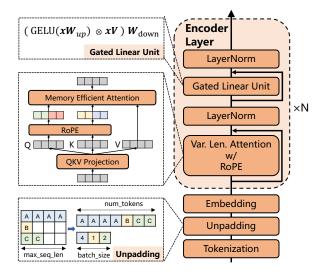
Figure 2: Our text encoder architecture.

scale meticulously curated datasets, providing robust off-the-shelf retrieval models.

We conduct extensive experiments to verify our method. For the text encoder, we evaluate on two natural language understanding (NLU) benchmarks, *i.e.,* XTREME-R (Ruder et al., 2021) and GLUE (Wang et al., 2018), and show that our encoder outperforms the same-sized previous state-of-the-art XLM-R. For the TRM and reranker, we evaluate on multiple retrieval benchmarks with multilingual and long-context settings, *e.g.,* MIRACL (Zhang et al., 2023b) and MLDR (Chen et al., 2024), where our models match the performance of state-of-the-art BGE-M3 (Chen et al., 2024) and achieve better long-context performance by a smaller size. We will open-source our models and code to facilitate further research and applications.

2 Method

2.1 Text Encoder

To construct powerful long-context multilingual text encoder models, we implement several enhancements to BERT (Devlin et al., 2019) architecture and train it from scratch using the vocabulary of XLM-R[1] (Conneau et al., 2020) series.

Specifically, we replace the absolute positional embeddings with Rotary Position Embedding (RoPE, Su et al., 2024), and upgrade the FFN to gated linear unit (GLU) (Shazeer, 2020). To ensure compatibility with libraries like FlashAttention (Dao, 2023), we remove the dropout applied to attention scores. In addition, we pad the token embedding size to be a multiple of 64, which could

---

[1] https://hf.co/FacebookAI/xlm-roberta-base

speedup the model throughput (Portes et al., 2023).

Unpadding Mode Inspired by Portes et al. (2023), we unpad the input batch to reduce redundant computations associated with padding tokens (Figure 2). We use xFormers (Lefaudeux et al., 2022) to implement the variable length attention. It dispatch the attention forward and backward to different kernels[2] based on the numerical precision, attention head size and device type. We unpad the MLM labels as well to reduce the computation cost of predicting non-masked tokens.

Data We assemble our multilingual pre-training data from a combination of the following sources: C4 (Raffel et al., 2020), Skypile (Wei et al., 2023) (2021-2023 subsets), mC4 (Xue et al., 2021), CulturaX (Nguyen et al., 2023), Wikipedia (Foundation) and books (proprietary). We filter them and curate a dataset covering 75 Languages. Appendix Table 7 presents the statistics of our dataset.

Training Curriculum We pre-train the model via masked language modeling (MLM) (Devlin et al., 2019)[3]. The MLM probability is set to 30% (Portes et al., 2023). Following Conneau and Lample (2019) and Conneau et al. (2020), the data from different languages is sampled by a multinomial distribution with probabilities $\{q_i\}_{i=1...N}$, where

$$q_i = \frac{p_i^\alpha}{\sum_{j=1}^N p_j^\alpha} \quad \text{with} \quad p_i = \frac{n_i}{\sum_{j=1}^N n_j}, \quad (1)$$

and n_i is the number of texts in language i. We set $\alpha = 0.5$. This sampling strategy could increase texts from low-resource languages. To train the native 8192-context model more efficiently, we adopt a phased training curriculum (Xiong et al., 2024):

- MLM-2048: we chunk the input into 2048 tokens and set RoPE base to $10,000$.
- MLM-8192: we chunk the input into 8192 tokens and set RoPE base to $160,000$.

Training Setup Following Portes et al. (2023), we use the learning rate decoupled[4] AdamW (Loshchilov and Hutter, 2018) with weight decay $1e-5$. We disable gradient clipping (set to 0) (Liu et al., 2019). All models are trained on A100 GPU servers by BF16 PyTorch native automatic mixed

---

[2] We adopt the memory-efficient attention (Rabe and Staats, 2021) in this work.

[3] We remove the next sentence prediction objective of BERT following (Liu et al., 2019).

[4] However, Xie et al. (2024) state that the decoupled weight decay is not ideal. We recommend to keep the default setting.
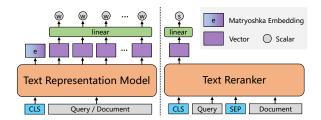
Figure 3: Our TRM and reranker.

precision via `transformers` (Wolf et al., 2020). We list the detailed hyper-parameters of each training stage in Appendix A.2 and Table 8. We denote the resulting models as `mGTE-MLM-2048/8192`.

2.2 Text Representation Model

Based on our encoder, we construct the TRM for the first-stage text retrieval in two steps: contrastive pre-training and fine-tuning (Wang et al., 2022; Li et al., 2023). Both steps share the same InfoNCE (Oord et al., 2018) learning objective:

$$\mathcal{L} = -\log \frac{\exp(s(q, d^+)/\tau)}{\sum_{i=1}^{N} \exp(s(q, d^i)/\tau)}, \quad (2)$$

where τ, q, and d denote the temperature parameter, query and document. The positive d^+ is the relevant document to q, and other irrelevant documents are negatives. These negatives can be either hard-negatives or in-batch negatives (documents of other instances in the same batch). $s(q, d)$ is the relevance score of q and d, measured by the dot product or cosine similarity between their respective representations.

Contrastive Pre-Training   We take the encoder output hidden state of the [CLS] token as the dense representation (*i.e.,* embedding) and compute the relevance score by cosine similarity. Our pre-training data (Appendix Table 9) comprise naturally occurring text pairs (*e.g.,* question-answer pairs from Quora and StackExchange, title-content pairs of CommonCrawl), translation pairs (Team et al., 2024), and crosslingual instruction tuning data (Muennighoff et al., 2022). We train the model with a batch size of $16,384$ and a learning rate of $5e-4$ for $240k$ steps. Each batch is sampled from a single data source by the same distribution of Eq.1. The queries (*resp.* documents) are truncated to the max tokens of 512 (*resp.* 1024). We reverse scale the RoPE base from $160,000$ to $20,000$ to fit the 1024 context length and acquire the long-context retrieval ability (denotes revNTK, ablation in §3.4). We set τ of InfoNCE to 0.01 and only use

in-batch negatives. More details refer to Appendix B.3. We denote this contrastive pre-trained model as `mGTE-CPT`, which is actually an unsupervised embedding model.

Matryoshka Embedding   Many of recently released models and APIs offer elastic embeddings by Matryoshka representation learning (MRL) (Kusupati et al., 2022), providing competitive sub-vectors of embeddings to save index storage and speedup search. Let $e \in \mathbb{R}^H$ denotes an embedding and $e_{:d}$ is the sliced sub-vector from dimension 0 to $d < H$. MRL[5] optimizes the weighted sum of multiple losses from different d dimensional sub-vectors, *i.e.,* compute InfoNCE by $s_d(e_{:d}^q, e_{:d}^d)$. We add this objective to our TRM fine-tuning stage.

Sparse Representation   Chen et al. (2024) show that neural sparse representations (term/token weights predicted by TRM) could greatly improve the long-context retrieval performance. We follow this design, computing the term weight w_t of each token of the input by $w_t = \text{ReLU}(\boldsymbol{W}\boldsymbol{h}_t)$, where $\boldsymbol{h}_t$ is the encoder hidden state of token t with dimension size H and $\boldsymbol{W} \in \mathbb{R}^{H \times 1}$ is randomly initialized. If a token appears multiple times in the text, we keep the max weight. The relevance score is computed by the joint importance of the co-occurring terms (denoted as $q \cap d$) within the query and document pair: $s_{\text{sparse}}(q, d) = \sum_{t \in q \cap d}(w_t^q \cdot w_t^d)$. This is then used to derive the InfoNCE loss for training.

Contrastive Fine-Tuning   Now we construct the TRM by multi-task learning of matryoshka embedding and sparse representation:

$$\mathcal{L}_{\text{TRM}} = \lambda \mathcal{L}_{\text{sparse}} + \sum_{d \in D} w_d \mathcal{L}_{:d}, \quad (3)$$

where $D = \{32k \mid k \in \mathbb{N}, k \geq 1, 32k \leq H\}$ is MRL dimension set, w_d is the weight of dimension d, and λ is the weight of sparse representation loss. We fine-tune our contrastive pre-trained embedding model on diverse high-quality datasets with hard-negatives (*e.g.,* MS MARCO (Nguyen et al., 2016), MIRACL (Zhang et al., 2023b), listed in Table 11). We adopt a dynamic batching strategy (Chen et al., 2024) to fine-tune 8192-context data. The batch sampling strategy is the same as the pre-training stage. The τ of MRL and sparse is set to 0.05 and 0.01 respectively. Other details refer to Appendix B.3. We denote this fine-tuned model as `mGTE-TRM`.

---

[5]Here we mean the MRL-E in Kusupati et al. (2022).

Model	Avg.	Pair Class.	M.C.	Structure Prediction		Question Answering			Cross-lingual Retrieval		
		XNLI	XCOPA	UDPOS	WikiANN	XQuAD	MLQA	TyDiQA-GoldP	Mewsli-X	LAReQA	Tatoeba
#Languages (Total 50)		15	11	38	47	11	7	9	38	11	38
Metrics		Acc.	Acc.	F1	F1	F1/EM	F1/EM	F1/EM	mAP@20	mAP@20	Acc.
mBERT-base	59.43	66.63	55.49	71.80	**62.34**	66.23 / 51.03	57.37 / 42.44	**55.01 / 38.05**	44.65	**75.26**	39.49
XLM-R-base	62.02	**74.50**	50.45	**73.84**	61.23	72.83 / 58.01	61.54 / 46.45	53.09 / 37.11	42.09	63.43	67.20
mGTE-MLM-2048	**65.24**	73.17	**63.62**	73.25	60.87	**75.33 / 60.00**	64.02 / 48.57	53.58 / 36.68	44.41	72.13	**72.02**
mGTE-MLM-8192	64.44	73.37	61.98	73.14	59.83	74.81 / 59.37	**64.24 / 48.80**	49.85 / 33.27	44.52	71.54	71.10

Table 1: XTREME-R (Ruder et al., 2021) results in the cross-lingual zero-shot transfer (models are trained on English data) setting. M.C. stands for Multiple Choice. The EM scores are not included in the average.

Model	Params	Pos.	Seq. Len.	GLUE Avg.
RoBERTa-base[α]	125M	Abs.	512	86.4
XLM-R-base	279M	Abs.	512	80.44
mGTE-MLM-2048	305M	RoPE	2048	83.42
mGTE-MLM-8192			8192	**83.47**

Table 2: GLUE (Wang et al., 2018) devset averages (w/o WNLI). The detailed scores for each subset are shown in Table 13. [α]Taken from Table 8 of Liu et al. (2019). The rest are from our runs, refer to Appendix C.2.

Model	Seq.	en	zh	fr	pl
BGE-M3-unsupervised[†]	8192	56.48	57.53	57.95	55.98
mGTE-CPT	512*	**60.16**	**58.67**	59.72	**57.66**
	8192	60.04	58.63	**59.74**	57.11
mE5-base	514	59.45	56.21	56.19	55.62
mE5-large	514	**61.50**	58.81	56.07	60.08
BGE-M3 (Dense)[†]	8192	59.84	60.80	58.79	**60.35**
mGTE-TRM (Dense)	8192	61.40	**62.72**	**59.79**	58.22
E5-mistral-7b	32768	66.63	60.81	48.33	-
voyage-multilingual-2	32000	-	-	61.65	-
Cohere-multilingual-v3.0	512	64.01	-	56.02	-
OpenAI-3-large	8191	64.59	-	-	-
OpenAI-3-small	8191	62.26	-	-	-

Table 3: Embedding model performance on MTEB English (Muennighoff et al., 2023), Chinese (Xiao et al., 2023), French (Ciancone et al., 2024) and Polish (Poświata et al., 2024). The scores of other models are retrieved from the MTEB online leaderboard. *To be consistent with the setting in contrastive pre-training, in retrieval tasks, the max sequence length of the document side is set to 1024. [†]Denote our runs.

2.3 Text Reranking Model

We also build a reranker using the cross-encoder architecture. It takes the query and document as input: [CLS] q [SEP] d, and directly predicts their relevance score by the [CLS] output state: $s_{\text{rerank}} = \boldsymbol{W}\boldsymbol{h}_{[\text{CLS}]}$. $\boldsymbol{W} \in \mathbb{R}^{H \times 1}$ is randomly initialized. The model is fine-tuned by InfoNCE in one step[6] based on our text encoder. Unless otherwise specified, we employ identical data and training settings as the TRM fine-tuning (§2.2). The difference lies in our adjustment of the hard-negatives. We describe the detailed settings in Appendix B.4. We denote this model as mGTE-reranker.

3 Evaluation

We separately evaluate our text encoder in §3.1, TRM and reranker in §3.2 and §3.3.

3.1 Natural Language Understanding

We evaluate the encoder on the cross-lingual natural language understanding (NLU) benchmark XTREME-R[7] (Ruder et al., 2021) and the English NLU benchmark GLUE (Wang et al., 2018). Results show that our encoder outperforms the same-sized previous state-of-the-art XLM-R (Conneau et al., 2020) on all benchmarks.

---

[6]We found that the contrastive pre-training of reranker does not improve the performance.

[7]We use XTREME-R (Ruder et al., 2021) instead of XTREME (Hu et al., 2020) since we found the retrieval tasks of XTREME is unstable and difficult to evaluate.

XTREME-R We focus on the *zero-shot cross-lingual transfer* setting where models are fine-tuned on English trainset and tested on multi- and cross-lingual data. The fine-tuning setup is described in Appendix C.1. We run mBERT-base, XLM-R-base, and our encoder, as shown in Table 1. Our 2048 and 8192 encoder models achieve average scores that are higher than those of XLM-R by 3.22 and 2.42 points, respectively.

GLUE We also report the performance on the devset of GLUE benchmark (Wang et al., 2018). The fine-tuning details refer to Appendix C.2. Table 2 presents the average scores (Table 13 provides the full results). Our models consistently outperform XLM-R-base and reasonably lag behind the English RoBERTa-base (Liu et al., 2019).

3.2 Text Embedding

Our contrastive pre-training actually yields a text embedding model. To understand the pre-training

Metric #languages (Total 33)	Params	Seq. Len.	Avg.	MLDR nDCG@10 13	MIRACL nDCG@10 18	MKQA recall@20 25	BEIR nDCG@10 1	LoCo nDCG@10 1
BM25	-	-	47.0	53.6	31.9	28.1	41.7	79.9
mE5-base	279M	514	53.5	30.5	62.3	53.7	48.9	72.2
mE5-large	560M	514	57.7	34.2	65.4	63.5	51.4	74.3
E5-mistral-7b	7111M	32768	62.4	42.6	62.2	62.4	56.9	87.8
OpenAI-3-large	-	8191	-	-	54.9	62.1	55.4	79.4
BGE-M3 Dense			64.3	52.5	67.7	67.8	48.7	84.9
BGE-M3 Sparse	568M	8192	55.1	62.2	53.9	36.3	38.3	84.9
BGE-M3 Dense + Sparse			67.7	64.8	**68.9**	**68.1**	49.4	87.4
mGTE-TRM Dense			66.7	56.6	62.1	65.8	51.1	88.9
mGTE-TRM Sparse	304M	8192	57.2	71.0	55.9	31.6	39.2	88.1
mGTE-TRM Dense + Sparse			**68.9**	**71.3**	64.5	66.0	**51.4**	**91.3**

Table 4: Retrieval results on MIRACL (Zhang et al., 2023b) and MLDR (Chen et al., 2024) (multilingual), MKQA (Longpre et al., 2021) (crosslingual), BEIR (Thakur et al., 2021) and LoCo (Saad-Falcon et al., 2024) (English).
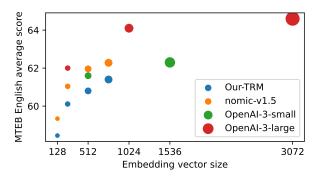

Figure 4: Elastic embedding results on MTEB English.

and fine-tuning of TRM, and to compare with other models, we first run the most popular text embedding benchmark MTEB (Muennighoff et al., 2023) as well as its Chinese, French and Polish versions.

Multilingual MTEB The results in Table 3 also present the scores of LLM-based models and commercial APIs for reference. For contrastive pre-trained models, our model outperforms BGE-M3-unsupervised (Chen et al., 2024) on all four subsets, through our backbone has fewer params than XLM-R-large. Comparing with BGE-M3 and mE5 (Wang et al., 2024b), our final TRM achieves best scores on Chinese and French, and is competitive on English.

Elastic Embedding We compare our TRM (only elastic embeddings) with open-source model and commercial APIs on MTEB English (Figure 4). Our model presents close scores to the same-sized English-only nomic-v1.5, which is promising for a multilingual model. However, it is still behind OpenAI APIs, which is reasonable since they are guessed to be much larger models.

3.3 Text Retrieval

We conduct evaluations to our TRM and reranker on retrieval benchmarks in multilingual (Miracl (Zhang et al., 2023b) and MLDR (Chen et al., 2024)), crosslingual (MKQA (Longpre et al., 2021)) setting, and the commonly used English BEIR (Thakur et al., 2021) and LoCo (Saad-Falcon et al., 2024). Our models are close to the state-of-the-art large models on Miracl, MKQA and BEIR, while achieve better scores on long-context datasets MLDR and LoCo. Details are in Appendix E.

First-Stage Retrieval We compare our TRM to the hybrid model BGE-M3 (Chen et al., 2024), dense models like mE5 (Wang et al., 2024b) and E5-mistral-7b (Wang et al., 2024a), and BM25. As shown in Table 4, our TRM consistently outperforms mE5 and OpenAI APIs, better than BGE-M3 on MLDR, and close to it on the rest parts.

Reranking In Table 5, we evaluate rerankers based on the candidates retrieved by Our-TRM dense model. Our model outperforms the powerful bge-reranker-v2-m3 (Chen et al., 2024) with a smaller size. Moreover, it greatly surpasses the same-sized jina-reranker-v2-multilingual.

3.4 Analysis

Efficiency We compare the efficiency of our TRM with BGE-M3 on dense retrieval in Table 6. To simulate the real-world scenario, the encoding time is the duration of encoding texts without length grouping. Our TRM is up to 14 times faster than BGE-M3 (52s v.s. 744s). The end-to-end unpadding with xFormers is crucial for encoding, which reduces the time by 5 times (52s v.s. 279s).

Metric #languages (Total 33)	Params	Seq. Len.	Avg.	MLDR nDCG@10 13	MIRACL nDCG@10 18	MKQA recall@20 25	BEIR nDCG@10 1
Retrieval (mGTE-TRM Dense)	304M	8192	58.9	56.6	62.1	65.8	50.9
jina-reranker-v2-multilingual	278M	8192	59.4	53.2	65.8	**68.8**	49.7
bge-reranker-v2-m3	568M	8192	65.7	66.8	**72.6**	68.7	54.6
mGTE-reranker	304M	8192	**67.4**	**78.7**	68.5	67.2	**55.4**

Table 5: Results of reranking based on the candidates retrieved by our TRM dense model (refer to Table 4).

Model	Attn.	Unpad.	Encoding Time	Search Latency
BGE-M3	eager	×	1800s	20.35ms
	SDPA-MEA		744s	
mGTE-TRM	eager	×	695s	15.07ms
	SDPA-MEA	×	298s	
	eager	✓	675s	
	SDPA-MEA	✓	279s	
	MEA	✓	52s	

Table 6: Dense retrieval efficiency. Encoding time is running MLDR-hi corpus (3806 texts with average 4456 tokens after truncating to maximum 8192) on one A100 GPU with FP16. Search latency is measured on a faiss index with 8.8M texts. MEA is the memory-efficient attention in xFormers. SDPA-MEA denotes MEA dispatched by scaled dot-product attention of PyTorch.

Scaled Contrastive Pre-Training We utilize the reversed NTK scaling in contrastive pre-training to reduce required text length, where we set the RoPE base to $1/8$ of the original and train the 8k encoder with 1k max length. To evaluate the effectiveness, we run the same training without the reversed NTK, comparing the MLDR scores in Figure 5. With revNTK, models exhibit slightly lower performance on 1k context but achieve more stable 8k performance across different training steps.

4 Related Work

Training long-context TRMs has become a hot topic recently. OpenAI released 8191 context APIs (Neelakantan et al., 2022) have set the target for open-source community. Portes et al. (2023) and Günther et al. (2023) replace position embedding of BERT with Alibi (Press et al., 2022) attention bias and pre-train from scratch, which is shown to be effective in build 8k TRMs. Nussbaum et al. (2024) explore the more powerful RoPE (Su et al., 2024) in BERT pre-training and their 2048-context pre-trained encoder achieve better retrieval performance on English. Zhu et al. (2024) suggest patch E5 (Wang et al., 2022) with RoPE. We also use
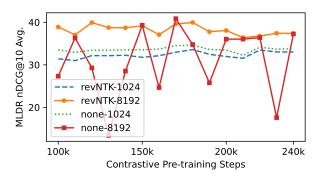
Figure 5: MLDR scores in contrastive pre-training. none keeps the RoPE untouched in pre-training. 1024 and 8192 are the max sequence length in evaluations. revNTK-8912 recovers the 8k context by NTK scaling.

RoPE and provide multi-stage training for native 8192-context text encoder, TRM, and reranker.

Chen et al. (2024) propose long-context multilingual TRM and reranker based on XLM-RoBERTa-large (Conneau et al., 2020) by extending position embedding to 8192 via continue training. We pre-train native 8k multilingual models from scratch for better long-context performance and efficiency.

5 Conclusion

We present the holistic practice of building native 8192-context multilingual retrieval models. We first suggest a text encoder with RoPE and unpadding, which is pre-trained by a two-stage MLM curriculum for 8k context. Evaluations on NLU benchmarks show that our encoder outperforms XLM-RoBERTa in the same size. Based on our encoder, we construct a hybrid TRM and a cross-encoder reranker by contrastive learning. The TRM is pre-trained with reversed RoPE NTK scaling and fine-tuned to generate both Matryoshka embeddings and sparse representations. Results on monolingual and crosslingual retrieval benchmarks show that our TRM and reranker are close to larger ones on regular datasets, and achieve better performance on long-context datasets. This means our models are more efficient for industrial applications.

References

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proc. of the ACL*, pages 4623–4637, Online.

Luiz Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. 2021. mmarco: A multilingual version of the ms marco passage ranking dataset. *arXiv preprint arXiv:2108.13897*.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada.

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 preprint arXiv:2402.03216*.

Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. 2016. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*.

Mathieu Ciancone, Imene Kerboua, Marion Schaeffer, and Wissam Siblini. 2024. Mteb-french: Resources for french sentence embedding evaluation and analysis. *arXiv preprint arXiv:2405.20468*.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

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. In *Proc. of the ACL*, pages 8440–8451, Online.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proc. of the EMNLP*, pages 2475–2485, Brussels, Belgium.

Sławomir Dadas, Michał Perełkiewicz, and Rafał Poświata. 2024. PIRB: A comprehensive benchmark of polish dense and hybrid text retrieval methods.

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. In *The Twelfth International Conference on Learning Representations*.

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational linguistics*, 47(2):255–308.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of the NAACL-HTL*, pages 4171–4186, Minneapolis, Minnesota.

Bill Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Third international workshop on paraphrasing (IWP2005)*.

Facebook. 2019. Tatoeba test set. https://github.com/facebookresearch/LASER/tree/main/data/tatoeba/v1.

Wikimedia Foundation. Wikimedia downloads.

Luyu Gao, Zhuyun Dai, and Jamie Callan. 2021a. Rethink training of bert rerankers in multi-stage retrieval pipeline. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021*, pages 280–286. Springer.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021b. SimCSE: Simple contrastive learning of sentence embeddings. In *Proc. of the EMNLP*, pages 6894–6910, Online and Punta Cana, Dominican Republic.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

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*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Junqin Huang, Zhongjie Hu, Zihao Jing, Mengya Gao, and Yichao Wu. 2024. Piccolo2: General text embedding with multi-task hybrid loss training. *arXiv preprint arXiv:2405.06932*.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proc. of the ACL*, pages 1601–1611, Vancouver, Canada.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, et al. 2022. Matryoshka representation learning. *Advances in Neural Information Processing Systems*, 35:30233–30249.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024a. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*.

Jinhyuk Lee, Zhuyun Dai, Xiaoqi Ren, Blair Chen, Daniel Cer, Jeremy R Cole, Kai Hui, Michael Boratko, Rajvi Kapadia, Wen Ding, et al. 2024b. Gecko: Versatile text embeddings distilled from large language models. *arXiv preprint arXiv:2403.20327*.

Sean Lee, Aamir Shakir, Darius Koenig, and Julius Lipp. 2024c. Open source strikes bread - new fluffy embeddings model.

Benjamin Lefaudeux, Francisco Massa, Diana Liskovich, Wenhan Xiong, Vittorio Caggiano, Sean Naren, Min Xu, Jieru Hu, Marta Tintore, Susan Zhang, Patrick Labatut, and Daniel Haziza. 2022. xformers: A modular and hackable transformer modelling library. https://github.com/facebookresearch/xformers.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proc. of ACL*, pages 7315–7330, Online.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Xiaoran Liu, Hang Yan, Chenxin An, Xipeng Qiu, and Dahua Lin. 2024. Scaling laws of rope-based extrapolation. In *The Twelfth International Conference on Learning Representations*.

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*.

Dingkun Long, Qiong Gao, Kuan Zou, Guangwei Xu, Pengjun Xie, Ruijie Guo, Jian Xu, Guanjun Jiang, Luxi Xing, and Ping Yang. 2022. Multi-cpr: A multi domain chinese dataset for passage retrieval. In *Proc. of the SIGIR*, pages 3046–3056.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. Mkqa: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Christopher D Manning. 2008. *Introduction to information retrieval*. Syngress Publishing,.

Xin Men, Mingyu Xu, Bingning Wang, Qingyu Zhang, Hongyu Lin, Xianpei Han, and Weipeng Chen. 2024. Base of rope bounds context length. *arXiv preprint arXiv:2405.14591*.

Niklas Muennighoff, SU Hongjin, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2024. Generative representational instruction tuning. In *ICLR 2024 Workshop: How Far Are We From AGI*.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proc. of the EACL*, pages 2014–2037, Dubrovnik, Croatia.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, et al. 2022. Text and code embeddings by contrastive pretraining. *arXiv preprint arXiv:2201.10005*.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. *arXiv preprint arXiv:2309.09400*.

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 *Proc. of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016)*, volume 1773 of *CEUR Workshop Proceedings*.

Zach Nussbaum, John X Morris, Brandon Duderstadt, and Andriy Mulyar. 2024. Nomic embed: Training a reproducible long context text embedder. *arXiv preprint arXiv:2402.01613*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proc. of the EMNLP*, pages 2362–2376, Online.

Jacob Portes, Alexander R Trott, Sam Havens, DANIEL KING, Abhinav Venigalla, Moin Nadeem, Nikhil Sardana, Daya Khudia, and Jonathan Frankle. 2023. Mosaicbert: A bidirectional encoder optimized for fast pretraining. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Rafał Poświata, Sławomir Dadas, and Michał Perełkiewicz. 2024. Pl-mteb: Polish massive text embedding benchmark. *arXiv preprint arXiv:2405.10138*.

Ofir Press, Noah Smith, and Mike Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*.

Yifu Qiu, Hongyu Li, Yingqi Qu, Ying Chen, Qiao-Qiao She, Jing Liu, Hua Wu, and Haifeng Wang. 2022. DuReader-retrieval: A large-scale Chinese benchmark for passage retrieval from web search engine. In *Proc. of the EMNLP*, pages 5326–5338, Abu Dhabi, United Arab Emirates.

Markus N Rabe and Charles Staats. 2021. Self-attention does not need o(n2) memory. *arXiv preprint arXiv:2112.05682*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proc. of the ACL*, pages 151–164, Florence, Italy.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proc. of the EMNLP*, pages 2383–2392, Austin, Texas.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI spring symposium series*.

Uma Roy, Noah Constant, Rami Al-Rfou, Aditya Barua, Aaron Phillips, and Yinfei Yang. 2020. LAReQA: Language-agnostic answer retrieval from a multilingual pool. In *Proc. of the EMNLP*, pages 5919–5930, Online.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proc. of the EMNLP*, pages 10215–10245, Online and Punta Cana, Dominican Republic.

Jon Saad-Falcon, Daniel Y Fu, Simran Arora, Neel Guha, and Christopher Re. 2024. Benchmarking and building long-context retrieval models with loco and m2-bert. In *Forty-first International Conference on Machine Learning*.

Noam Shazeer. 2020. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proc. of the EMNLP*, pages 1631–1642, Seattle, Washington, USA.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

NLLB Team et al. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841.

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 *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proc. of the NAACL-HLT*, pages 809–819, New Orleans, Louisiana.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

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

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Tianwen Wei, Liang Zhao, Lichang Zhang, Bo Zhu, Lijie Wang, Haihua Yang, Biye Li, Cheng Cheng, Weiwei Lü, Rui Hu, et al. 2023. Skywork: A more open bilingual foundation model. *arXiv preprint arXiv:2310.19341*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proc. of the NAACL-HLT*, pages 1112–1122, New Orleans, Louisiana.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proc. of the EMNLP: System Demonstrations*, pages 38–45, Online.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighof. 2023. C-pack: Packaged resources to advance general chinese embedding. *arXiv preprint arXiv:2309.07597*.

Xiaohui Xie, Qian Dong, Bingning Wang, Feiyang Lv, Ting Yao, Weinan Gan, Zhijing Wu, Xiangsheng Li, Haitao Li, Yiqun Liu, et al. 2023. T2ranking: A large-scale chinese benchmark for passage ranking. In *Proc. of the SIGIR*, pages 2681–2690.

Zeke Xie, Zhiqiang Xu, Jingzhao Zhang, Issei Sato, and Masashi Sugiyama. 2024. On the overlooked pitfalls of weight decay and how to mitigate them: A gradient-norm perspective. *Advances in Neural Information Processing Systems*, 36.

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, Madian Khabsa, Han Fang, Yashar Mehdad, Sharan Narang, Kshitiz Malik, Angela Fan, Shruti Bhosale, Sergey Edunov, Mike Lewis, Sinong Wang, and Hao Ma. 2024. Effective long-context scaling of foundation models. In *Proc. of the NAACL-HLT*, pages 4643–4663, Mexico City, Mexico.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proc. of the NAACL-HLT*, pages 483–498, Online.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proc. of the EMNLP*, pages 2369–2380, Brussels, Belgium.

Sheng Zhang, Xin Zhang, Hui Wang, Lixiang Guo, and Shanshan Liu. 2018. Multi-scale attentive interaction networks for chinese medical question answer selection. *IEEE Access*, 6:74061–74071.

Xin Zhang, Zehan Li, Yanzhao Zhang, Dingkun Long, Pengjun Xie, Meishan Zhang, and Min Zhang. 2023a. Language models are universal embedders. *arXiv preprint arXiv:2310.08232*.

Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. 2021. Mr. tydi: A multi-lingual benchmark for dense retrieval. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 127–137.

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.

Yanzhao Zhang, Dingkun Long, Guangwei Xu, and Pengjun Xie. 2022. Hlatr: enhance multi-stage text retrieval with hybrid list aware transformer reranking. *arXiv preprint arXiv:2205.10569*.

Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. 2024. Dense text retrieval based on pretrained language models: A survey. *ACM Transactions on Information Systems*, 42(4):1–60.

Dawei Zhu, Liang Wang, Nan Yang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. 2024. Longembed: Extending embedding models for long context retrieval. *arXiv preprint arXiv:2404.12096*.

Appendix

A MLM Pre-Training

In this section, we describe the data and training configurations of the MLM pre-training of our suggested text encoder.

A.1 Data

Our multilingual pre-training data are composed from following sources:

- C4 (Raffel et al., 2020),

- Skypile (Wei et al., 2023) (2021-2023 subsets),
- mC4 (Xue et al., 2021) (excluded English),
- CulturaX (Nguyen et al., 2023),
- Wikipedia (Foundation),
- books (proprietary).

We filter them and curate a dataset with 1,028B tokens (by XLM-R tokenizer), covering 75 languages (Chinese Simplified and Traditional are counted as one). Table 7 presents the statistics of our final dataset.

A.2 Training Details

We pre-train out text encoder with a two-stage curriculum by masked language model (MLM) objective. The first stage model is trained on maximum length 2048 with batch size 8192 for roughly 0.6 epoch (250k steps) on sampled data (by XLM sampling Eq.1). In the second stage, we down sample texts shorter than 2048 and continue train the model for 30k steps with maximum length 8192 and batch size 2048. The RoPE base is set to $10,000$ and $160,000$ for the first and second stage, respectively (Xiong et al., 2024; Liu et al., 2024; Men et al., 2024).

The text encoder is initialized in base size (12 layers of hidden state size 768) by PyTorch default initialization. We train the model by `transformers` library (Wolf et al., 2020) in BF16 precision. Following Portes et al. (2023), we use the learning rate decoupled AdamW optimizer with weight decay 1e-5. The other hyper-parameters are in Table 8. During training, we split texts that exceed the max sequence length into chunks, but we do not modify shorter texts.

The 250k steps of first stage, MLM-2048, took 10.75 days on 32 A100 80G GPUs. The 30k steps of second stage, MLM-8192, took 20.5 hours on 32 A100 80G GPUs. We acknowledge that this is not the optimal setting and recommend further explorations to optimize the pre-training.

A.3 Additional Discussion on RoPE

We chose RoPE (Su et al., 2024) (to replace absolute position embedding) due to its advantageous properties. RoPE offers excellent context extension capabilities, allowing models to be trained on shorter context windows and then run inference on longer ones. Additionally, it implements asymmetric relative distance encoding, meaning $D(i, j) \neq D(j, i)$, which appears to be particularly important for the training of BERT-like encoder-

only models that rely on bidirectional attention. Furthermore, the effectiveness of RoPE has been empirically validated by numerous models, such as RoFormer (Su et al., 2024) and LLaMA (Touvron et al., 2023).

B Contrastive Learning

In this section, we describe the data and training configurations of the contrastive learning of our TRM and reranker.

B.1 Pre-Training Data

Following previous studies, we create large-scale weakly correlated text pairs from diverse sources. The data are primarily consisted of four parts: English pairs (Wang et al., 2022; Li et al., 2023), Chinese pairs (Li et al., 2023; Xiao et al., 2023), multilingual pairs (cc-news[8]), and crosslingual instruction and translation pairs (Muennighoff et al., 2022; Team et al., 2024). We filter the data by removing duplicates and low-quality pairs, resulting in a total of 2,938.8M pairs. Table 9 lists the statistics of our contrastive pre-training data (cc-news is separately presented by languages in Table 10).

B.2 Fine-Tuning Data

We collect publicly available high-quality dataset as our fine-tune data as detailed in Table 11. For English, we utilize seven datasets:
- MS MARCO (Nguyen et al., 2016),
- Natural Questions (NQ) (Kwiatkowski et al., 2019),
- TriviaQA (Joshi et al., 2017),
- HotpotQA (Yang et al., 2018),
- SQuAD (Rajpurkar et al., 2016),
- FEVER (Thorne et al., 2018),
- AllNLI from SimCSE (Gao et al., 2021b).

For Chinese, we compile six datasets:
- DuReader (Qiu et al., 2022),
- mMARCO-zh (Bonifacio et al., 2021),
- T2-Ranking (Xie et al., 2023),
- CmedQAv2 (Zhang et al., 2018),
- SimCLUE[9],
- Multi-CPR (Long et al., 2022).

Additionally, we incorporate three multilingual datasets: Mr.TyDi (Zhang et al., 2021), MIRACL (Zhang et al., 2023b), and MLDR (Chen et al., 2024). We exclusively use the trainset of

---

[8] `commoncrawl.org/blog/news-dataset-available`
[9] `https://github.com/CLUEbenchmark/SimCLUE`

ISO code	Language	Tokens (M)	Size (GiB)	ISO code	Language	Tokens (M)	Size (GiB)
af	Afrikaans	1,489.19	5.30	ky	Kyrgyz	500.40	3.27
ar	Arabic	14,549.36	79.53	lo	Lao	2.43	0.01
az	Azerbaijani	688.72	3.13	lt	Lithuanian	1,824.46	6.38
be	Belarusian	1,090.61	6.17	lv	Latvian	1,823.43	6.38
bg	Bulgarian	1,454.57	8.94	mk	Macedonian	735.46	4.89
bn	Bengali	1,291.58	9.21	ml	Malayalam	778.66	7.27
ca	Catalan	1,294.05	4.65	mn	Mongolian	958.83	5.91
ceb	Cebuano	633.06	2.02	mr	Marathi	861.05	7.48
cs	Czech	1,465.00	5.27	ms	Malay	96.37	0.39
cy	Welsh	582.49	1.84	my	Burmese	902.46	7.26
da	Danish	1,030.30	4.01	ne	Nepali	657.65	6.32
de	German	18,097.31	67.90	nl	Dutch	5,137.98	18.65
el	Greek	874.87	5.09	no	Norwegian	992.51	3.91
en	English	187,110.31	771.79	pa	Punjabi	726.41	4.96
es	Spanish	148,713.06	601.04	pl	Polish	2,949.88	10.42
et	Estonian	1,111.31	4.10	pt	Portuguese	49,594.59	198.64
eu	Basque	787.46	2.99	qu	Quechua	0.07	0.00
fa	Persian	1,203.16	7.22	ro	Romanian	2,215.05	7.98
fi	Finnish	949.88	3.73	ru	Russian	93,966.28	597.92
fr	French	136,785.00	512.28	si	Sinhala	878.65	7.03
gl	Galician	772.47	3.22	sk	Slovak	884.38	3.31
gu	Gujarati	973.27	6.95	sl	Slovenian	1,100.81	4.05
he	Hebrew	1,842.74	8.36	so	Somali	0.82	0.00
hi	Hindi	1,032.67	8.27	sq	Albanian	700.78	2.73
hr	Croatian	480.19	1.54	sr	Serbian	1,139.38	6.84
ht	Haitian	0.03	0.00	sv	Swedish	840.00	3.37
hu	Hungarian	1,341.23	5.10	sw	Swahili	31.58	0.13
hy	Armenian	805.98	4.88	ta	Tamil	926.84	8.54
id	Indonesian	25,564.33	119.84	te	Telugu	857.91	7.01
is	Icelandic	987.89	3.63	th	Thai	12,782.08	119.52
it	Italian	11,068.23	40.50	tl	Filipino	275.16	1.01
ja	Japanese	135,684.28	601.19	tr	Turkish	1,065.05	4.42
jv	Javanese	0.62	0.00	uk	Ukrainian	893.70	5.68
ka	Georgian	834.90	7.25	ur	Urdu	1,051.83	6.19
kk	Kazakh	1,020.27	6.57	vi	Vietnamese	67,850.87	305.51
km	Khmer	746.15	6.54	yo	Yoruba	0.04	0.00
kn	Kannada	919.83	7.15	zh-cn	Chinese (Simplified)	43,727.30	167.23
ko	Korean	22,865.85	91.78	zh-tw	Chinese (Traditional)	73.39	0.26

Table 7: MLM pre-training data, where we have a total of 1,028B tokens (by XLM-RoBERTa tokenizer). The raw texts are stored in 4.47 TiB arrow files. We report the list of 75 languages (Chinese Simplified and Traditional are counted as one) and include the number of tokens and the size of the data (arrow files, in GiB) for each language.

Hyper-param	MLM-2048	MLM-8192
Number of Params	304M	
Number of Layers	12	
Hidden Size	768	
FFN Inner Size	3072	
Number of Attention Heads	12	
Attention Head Size	64	
Dropout	0.1	
Attention Dropout	0	
Learning Rate Decay	Linear	
Adam ϵ	1e-6	
Adam β_1	0.9	
Adam β_2	0.98	
Gradient Clipping	0.0	
Precision	PyTorch BF16 AMP	
Weight Decay	1e-5	
Max Length	2048	8192
Batch Size	8192	2048
Peak Learning Rate	5e-4	5e-5
Warm-up Ratio	0.06	0.06
Max Steps	250000	30000
RoPE base	10000	160000

Table 8: MLM pre-training hyper-parameters.

each dataset and employ our contrastive pre-trained model to mine hard negatives.

B.3 TRM Training Setup

Here we separately describe the training setting of the contrastive pre-training and TRM fine-tuning.

Contrastive Pre-Training In the contrastive pre-training, we train a dense representation model (embedder) which take the [CLS] hidden state as the embedding of the input. We use the same XLM sampling strategy (eq.1) to sample batches from each source of Table 9 or cc-news subset of Table 10, where the texts of one batch only come from one single source, and the batch size is $16, 384$. We train the model by transformers with deepspeed ZeRO (Rajbhandari et al., 2020) stage 1 in FP16 precision for roughly 0.4 epoch (240k steps, took 154 hours on 16 A100 80G GPUs) of our data (3.93B pairs on sampled data by Eq.1). We use the AdamW optimizer with the learning rate 2e-4, linear decay, and warm-up ratio 0.05. The $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1e - 07$. We set gradient clipping to 1.0.

TRM Fine-Tuning In the fine-tuning stage, we further train our embedding model with high-quality datasets as detailed in §B.2. For each query, we incorporate one positive passage and 8 hard negative passages. To enhance long-context retrieval capabilities and maximize training efficiency, we

adopt a dynamic batch size strategy as previous work (Chen et al., 2024). Firstly, we group the training data according to their lengths for each dataset. Different batch sizes are then used for varying lengths during training. Additionally, we divide the entire batch into multiple sub-batches, encoding each sub-batch iteratively with gradient checkpointing (Chen et al., 2016) and then gather them to get the final batch's embeddings. We train the embedding model with 10 epochs with 8 A100 80G GPUs. All other hyper-parameters remain consistent with those used in the contrastive pre-training stage. In Table 12, we list the batch size of different length.

B.4 Reranker Training Setup

We utilize the identical fine-tuning dataset for both the reranker and the TRM. For each query, we introduce 10 negative samples, comprising 6 hard negatives and 4 randomly selected negatives. All training parameters expect batch size are kept consistent with those employed for the TRM. The batch sizes are listed in Table 12.

C NLU Evaluation

We evaluate our text encoder as well as baselines on the multilingual XTREME-R (Ruder et al., 2021) and English GLUE (Wang et al., 2018) benchmarks. We describe the fine-tuning setup and the evaluation details in the following subsections. The evaluation scripts are available in our github repo[10].

C.1 XTREME-R

We only run XTREME-R (Ruder et al., 2021) in the zero-shot cross-lingual transfer learning setting, where models are fine-tuned on English trainset and tested on multi- and cross-lingual data. We compare our encoder with mBERT-base-cased[11] and XLM-RoBERTa-base[12]. All models are fine-tuned in the same setting and hyper-parameters.

The results are already presented in Table 1.

As XTREME-R has no final release, we implement the evaluation code based on the code of XTREME[13]. However, there are some differences in the retrieval evaluation, where our code will deduplicate the retrieval corpus. In addition, we implement the XCOPA in multiple choice,

---

Source	Language	Pairs (M)	Size (GiB)	Source	Language	Pairs (M)	Size (GiB)
agnews	English	1.15	0.30	stackoverflow_title_body	English	18.01	20.49
amazon_qa	English	1.10	0.37	wikihow	English	0.13	0.03
amazon_review_title_body	English	87.86	43.58	wikipedia	English	33.17	19.39
arxiv_title_abstract	English	2.26	2.26	yahoo_body_answer	English	0.68	0.44
baai_mtp_en	English	196.60	178.70	yahoo_qa	English	1.20	0.55
beir_dbpedia	English	4.64	1.59	yahoo_question_body	English	0.66	0.20
beir_debate	English	0.38	0.63	baai_mtp_zh	Chinese	100.13	231.42
beir_pubmed_title_abstract	English	0.13	0.19	baidu_baike	Chinese	34.21	39.05
biorxiv_title_abstract	English	0.20	0.32	baike_qa_train	Chinese	1.43	1.34
clueweb	English	3.94	6.62	commoncrawl_zh	Chinese	28.42	92.79
clueweb_anchor	English	4.51	7.69	gpt3_qa_all	Chinese	4.97	2.39
cnn_dailymail	English	0.31	1.28	gpt3_summarization	Chinese	4.48	1.62
commoncrawl	English	139.94	506.84	medical_quac_wenda_10m	Chinese	10.00	4.55
dpr_reddit	English	199.82	125.71	medical_scholar	Chinese	8.43	7.81
gooaq_qa	English	3.01	0.97	qcl	Chinese	7.40	43.23
hlp_wikipedia	English	19.48	13.55	web_text_zh_train	Chinese	4.12	2.07
medrxiv_title_abstract	English	0.20	0.32	wikipedia	Chinese	4.45	1.07
msmarco	English	2.89	19.56	wodao	Chinese	59.13	190.29
npr	English	0.59	1.03	zh_sft_data_v1	Chinese	0.45	0.43
reddit_title_body	English	124.89	90.36	zh_sft_data_v2	Chinese	2.24	1.37
s2orc_citation_abstract	English	30.58	67.81	zhihu_qa	Chinese	53.42	40.99
s2orc_citation_title	English	51.03	10.84	zhihu_title_body	Chinese	0.94	0.29
s2orc_title_abstract	English	41.77	30.29	xp3x	Crosslingual	351.87	463.85
stackexchange_qa	English	3.00	3.36	translation_eg_NLLB	Crosslingual	940.63	323.06
stackexchange_title_body	English	4.74	4.00				

Table 9: Contrastive pre-training data, where cc-news multilingual data are not included (Table 10). For this Table, we have a total of 2,595.57M pairs (raw texts stored by 2.55 TiB jsonl files).

Lang.	Pairs (M)	Size (GiB)	Lang.	Pairs (M)	Size (GiB)	Lang.	Pairs (M)	Size (GiB)	Lang.	Pairs (M)	Size (GiB)
ar	20.407	32.45	fy	0.044	0.03	lb	0.048	0.05	sk	1.093	1.16
az	0.401	0.23	gl	0.114	0.20	lt	0.321	0.24	sl	1.046	0.93
be	0.039	0.06	gu	0.061	0.06	lv	0.438	0.37	sq	0.282	0.51
bg	3.005	5.03	he	0.397	0.84	mk	0.173	0.44	sr	0.910	1.09
bn	0.463	0.33	hi	14.253	29.90	ml	0.408	0.48	sv	3.361	2.90
ca	0.909	1.30	hr	1.268	1.77	mr	0.278	0.35	sw	0.059	0.07
cs	1.834	2.18	hu	2.668	3.40	my	0.045	0.04	ta	2.125	1.26
da	1.090	1.58	hy	0.125	0.09	nl	6.700	7.41	te	0.355	0.33
de	39.715	57.98	id	6.048	7.46	nn	0.162	0.12	tg	0.038	0.03
el	7.170	14.93	is	0.100	0.05	no	1.978	2.21	th	0.124	0.17
en	0.615	1.47	it	27.827	40.57	or	0.038	0.03	tl	0.055	0.07
es	55.201	86.87	ja	4.139	3.95	pa	0.036	0.04	tr	23.840	26.81
et	0.950	0.85	ka	0.074	0.06	pl	3.530	5.77	uk	5.021	8.42
eu	0.051	0.02	kn	0.192	0.16	pt	12.611	19.28	ur	1.625	0.87
fa	4.839	7.99	ko	8.605	12.48	ro	6.678	9.15	vi	4.375	7.03
fi	1.532	1.93	ky	0.061	0.03	ru	39.451	65.74	MIX*	0.359	0.28
fr	21.242	32.67	la	0.035	0.06	sh	0.220	0.18			

Table 10: The cc-news multilingual pairs (343.26M in total, raw texts stored by 512.8 GiB jsonl files), used in contrastive pre-training together with all data of Table 9. MIX* denotes the mixed pairs of languages that are less than 1GiB (such as af, ceb). We utilize a very large batch size (16,384), and since each batch contains text exclusively from a single source, these low-resource languages might not fill an entire batch. Consequently, we have merged these languages together.

Dataset	Language	Size
MS MARCO, HotpotQA, NQ, NLI, etc.	English	1.4M
DuReader, T^2-Ranking, SimCLUE, etc.	Chinese	2.0M
MIRACL, Mr.TyDi, MLDR	Multilingual	118.9K

Table 11: Specification of training data adopted in Fine-tuning stage.

length	BS(E)	S-BS(R)	BS(E)	S-BS(R)
0-500	768	256	512	256
500-1000	384	128	384	128
1000-2000	256	64	256	64
2000-3000	160	48	160	48
3000-8000	80	16	80	16

Table 12: Batch size (BS) and sub batch size (S-BS) of different length for embedding (E) and reranker (R) model in the fine-tune stage.

which might be different from XTREME-R. In fine-tuning, if not specified, we use the epoch number of 3, learning rate of 2e-5, batch size of 32, and max sequence length of 128 (Hu et al., 2020).

XNLI We fine-tune the model on MNLI[14] (Williams et al., 2018) trainset and then evaluate the checkpoint on XNLI[15] (Conneau et al., 2018).

XCOPA We run this data as the multiple choice task. The model is first trained on SIQA[16] citesapetal-2019-social and then COPA[17] (Roemmele et al., 2011) for 5 epochs on each dataset. The checkpoint of COPA is evaluated on XCOPA[18] (Ponti et al., 2020).

UDPOS We extract pos-tagging data from the UD (De Marneffe et al., 2021) v2.7 and train the model on trainset of English parts by 10 epochs.

WikiANN We fine-tune the model on the trainset of English by 10 epochs and evaluate on selected WikiANN (Rahimi et al., 2019) testsets[19].

XQuAD We fine-tune on the trainset of SQuAD (Rajpurkar et al., 2016) v1.1[20] for 3 epochs with the learning rate 3e-5 and max length 384. Then

we evaluate the checkpoint on XQuAD[21] (Artetxe et al., 2020).

MLQA We directly evaluate the same checkpoint of XQuAD on MLQA[22] (Lewis et al., 2020) with the same setting.

TyDiQA-GoldP We train the model on TyDiQA-GoldP[23] (Clark et al., 2020) trainset in the same setting as XQuAD. Then we evaluate the checkpoint on the testset.

Mewsli-X We generate the data following their github[24]. This is a updated version so that we can not compare with the results in the XTREME-R paper. We train the model on the English wikipedia (mention, entity)-pairs for 2 epochs with the batch size 64 and max length 64. Then we evaluate the checkpoint in the language agnostic retrieval setting, refer to Ruder et al. (2021) for more details.

LAReQA This task is actually conducted on XQuAD-R[25] (Roy et al., 2020). We fine-tune the model on the trainset of SQuAD v1.1 in dual-encoder architecture ([CLS] as the embedding) and retrieval setting for 3 epochs with the batch size 16, max query length 96, and max document length 256. Then we evaluate the checkpoint on XQuAD-R in same setting.

Tatoeba We directly evaluate the checkpoint from LAReQA on Tatoeba[26] (Facebook, 2019) in the same setting.

C.2 GLUE

The GLUE benchmark (Wang et al., 2018) is English transfer learning, *i.e.,* models are trained and tested on the trainset and testset of each dataset (CoLA (Warstadt et al., 2019), SST-2 (Socher et al., 2013), MRPC (Dolan and Brockett, 2005), STS-B (Cer et al., 2017), QQP, MNLI (Williams et al., 2018), QNLI (Rajpurkar et al., 2016), RTE).

We evaluate the GLUE benchmark based on the

---

[14] hf.co/datasets/nyu-mll/glue MNLI subset.
[15] hf.co/datasets/facebook/xnli
[16] hf.co/datasets/allenai/social_i_qa
[17] hf.co/datasets/aps/super_glue copa split.
[18] hf.co/datasets/cambridgeltl/xcopa
[19] hf.co/datasets/unimelb-nlp/wikiann
[20] hf.co/datasets/rajpurkar/squad

[21] hf.co/datasets/google/xquad
[22] hf.co/datasets/facebook/mlqa
[23] hf.co/datasets/juletxara/tydiqa_xtreme
[24] https://github.com/google-research/google-research/blob/master/dense_representations_for_entity_retrieval/mel/mewsli-x.md#getting-started
[25] hf.co/datasets/google-research-datasets/xquad_r
[26] hf.co/datasets/mteb/tatoeba-bitext-mining

scripts[27] and data[28] provided by `transformers`. In fine-tuning of each dataset, we use the epoch number of 3, learning rate of 2e-5, batch size of 32, and max sequence length of 128. For MRPC, STS-B, and RTE, we start from the checkpoint of MNLI following (Liu et al., 2019). The MNLI checkpoint is shared with XNLI of XTREME-R (§C.1).

The detailed results are in Table 13. We also include scores of our English models (`Our-en-*`, pre-trained on C4-en) and baselines (Portes et al., 2023; Günther et al., 2023; Nussbaum et al., 2024).

D Text Embedding Evaluation

We have demonstrated the average scores on MTEB English, Chinese, French and Polish (Table 3). In this section, we delve into the details, presenting results of different tasks on each language. For a fair comparison, we do not include the derived models (developed by secondary training on other public off-the-shelf models) in English and Chinese. In addition to the results obtained from the online leaderboard, our own MTEB evaluations were conducted using version 1.2.0 of `mteb` library.

MTEB-en Table 14 shows the results on English MTEB (Muennighoff et al., 2023). For reference, we include our English embedding models (`Our-en-base/large-embed`, trained by the two-stage contrastive learning on the English part of our data) and top-performing systems from the online leaderboard. We can see that the multilingual models still have a noticeable gap compared to the English models.

MTEB-zh Table 15 presents the C-MTEB (Xiao et al., 2023) (MTEB Chinese subset) results. We include the results of several LLM-based embedding models and APIs. Given that the Chinese community is also keen on optimizing embedding models, the gap between multilingual models and Chinese models is quite noticeable.

MTEB-fr Table 16 demonstrates the F-MTEB (Ciancone et al., 2024) (MTEB French subset) results. Our TRM dense is comparable to the specialized French API `mistral-embed`. However, compared to our our-cpt model, the improvement from fine-tuning is not significant.

MTEB-pl Table 17 lists the Polish MTEB (Poświata et al., 2024) results. Our model does not outperform large-sized BGE and mE5. We speculate this may be due to the limited amount of Polish pairs in the contrastive pre-training, resulting in insufficient training.

E Text Retrieval Evaluation

The retrieval process can be divided into two main stages: recall and reranking. In the recall stage, documents are retrieved using both dense vectors and sparse representations. The final recall score is calculated by weighting the dense retrieval score with a fixed coefficient of 1 and the sparse retrieval score with coefficients ranging from 0.001 to 0.01. Documents not retrieved by either method receive a score of 0. During the ranking stage, the top 100 documents from the recall results are selected as candidates. These candidates are then sorted using our reranker model to produce the final retrieval results.

We present the detail results of MLDR (Chen et al., 2024) (multilingual long-context retrieval, Table 18), MKQA (Longpre et al., 2021) (multilingual, Table 19), MIRACL (Zhang et al., 2023b) (multilingual, Table 20, BEIR (Thakur et al., 2021) (English, Table 21) and LoCo (Saad-Falcon et al., 2024) (English long-context, Table 22).

---

[27]`github.com/huggingface/transformers/tree/main/examples/pytorch/text-classification#glue-tasks`
[28]`hf.co/datasets/nyu-mll/glue`

Model	Params	Pos.	Seq.	Avg.	Single Sentence		Paraphrase and Similarity			Natural Language Inference		
					CoLA	SST-2	MRPC	STS-B	QQP	MNLI	QNLI	RTE
RoBERTa-base[α]	125M	Abs.	512	**86.4**	63.6	94.8	90.2	91.2	91.9	87.6	92.8	78.7
MosaicBERT-base-128[β]	137M	Alibi	128	85.4	58.2	93.5	89.0	90.3	92.0	85.6	91.4	83.0
MosaicBERT-base-2048[γ]	137M	Alibi	2048	85	54	93	87	90	92	86	92	82
JinaBERT-base[δ]	137M	Alibi	512	82.6	51.4	94.5	88.4	89.5	80.7	85.7	92.2	78.7
nomic-bert-2048[γ]	137M	RoPE	2048	84	50	93	88	90	92	86	92	82
GTEv1.5-en-base-2048	137M	RoPE	2048	85.15	54.46	93.81	93.21	90.00	88.61	86.73	91.67	82.67
GTEv1.5-en-base-8192	137M	RoPE	8192	85.61	57.02	93.35	92.14	90.21	88.78	86.69	91.85	84.84
XLM-R-base	279M	Abs.	512	80.44	30.74	92.43	92.74	89.16	87.74	84.54	90.37	75.81
mGTE-MLM-2048	305M	RoPE	2048	83.42	49.65	92.66	91.17	89.95	88.41	85.40	91.38	78.70
mGTE-MLM-8192	305M	RoPE	8192	**83.47**	48.41	92.32	90.94	89.77	88.50	85.58	91.34	80.87
RoBERTa-large[α]	355M	Abs.	512	**88.9**	68.0	96.4	90.9	92.4	92.2	90.2	94.7	86.6
MosaicBERT-large-128[β]	434M	Alibi	128	86.1	59.7	93.7	88.2	90.9	92.0	86.9	93.0	84.5
JinaBERT-large[δ]	435M	Alibi	512	83.7	59.6	95.0	88.5	88.2	80.9	86.6	92.5	78.5
GTEv1.5-en-large-512	434M	RoPE	512	88.16	64.80	94.50	92.09	91.50	89.23	89.12	93.78	90.25
GTEv1.5-en-large-2048	434M	RoPE	2048	87.02	60.09	94.61	92.14	91.47	89.12	89.02	92.31	87.36
GTEv1.5-en-large-8192	434M	RoPE	8192	87.58	60.39	95.07	93.45	91.37	89.19	89.20	93.90	88.09

Table 13: GLUE (Wang et al., 2018) devset scores (w/o WNLI). [α]Taken from Table 8 of Liu et al. (2019). [β]Taken from Table S3 of Portes et al. (2023). [γ]Taken from Table 2 of Nussbaum et al. (2024). [δ]Taken from Table 2 of Günther et al. (2023). The rest of the numbers are from our runs, refer to §C.2 for details.

MTEB English #Datasets (→)	Param.	Dim.	Seq.	Avg. 56	Class. 12	Clust. 11	PairC. 3	Rerank. 4	Retr. 15	STS 10	Summ. 1
gte-Qwen2-7b-instruct (Li et al., 2023)	7613M	3584	131072	70.24	86.58	56.92	85.79	61.42	60.25	83.04	31.35
neural-embedding-v1	-	-	-	69.94	87.91	54.32	87.68	61.49	58.12	85.24	30.87
NV-Embed-v1 (Lee et al., 2024a)	7851M	4096	32768	69.32	87.35	52.8	86.91	60.54	59.36	82.84	31.2
voyage-large-2-instruct	-	1024	16000	68.28	81.49	53.35	89.24	60.09	58.28	84.58	30.84
gte-Qwen2-1.5B-instruct (Li et al., 2023)	1776M	1536	131072	67.16	82.47	48.75	87.51	59.98	58.29	82.73	31.17
google-gecko (Lee et al., 2024b)	1200M	768	2048	66.31	81.17	47.48	87.61	58.9	55.7	85.07	32.63
GritLM-7B (Muennighoff et al., 2024)	7242M	4096	32768	66.76	79.46	50.61	87.16	60.49	57.41	83.35	30.37
E5-mistral-7b (Wang et al., 2024a)	7111M	4096	32768	66.63	78.47	50.26	88.34	60.21	56.89	84.63	31.4
text-embedding-3-large	-	3072	8191	64.59	75.45	49.01	85.72	59.16	55.44	81.73	29.92
mxbai-embed-large-v1 (Lee et al., 2024c)	335M	1024	512	64.68	75.64	46.71	87.2	60.11	54.39	85	32.71
nomic-embed-text-v1 (Nussbaum et al., 2024)	137M	768	8192	62.39	74.12	43.91	85.15	55.69	52.81	82.06	30.08
gte-en-large-v1.5	434M	1024	8192	65.39	77.75	47.96	84.53	58.5	57.91	81.43	30.91
gte-en-base-v1.5	137M	768	8192	64.11	77.17	46.82	85.33	57.66	54.09	81.97	31.17
mE5-base (Wang et al., 2024b)	278M	768	514	59.45	73.02	37.89	83.57	54.84	48.88	80.26	30.11
mE5-large (Wang et al., 2024b)	560M	1024	514	61.5	74.81	41.06	84.75	55.86	51.43	81.56	29.69
BGE-m3 (dense)[†] (Chen et al., 2024)	568M	1024	8192	59.84	74.08	37.27	84.50	55.28	48.82	81.37	31.55
mGTE-TRM (dense)	305M	768	8192	61.40	70.89	44.31	84.23	57.47	51.08	82.11	30.58
BGE-m3-unsupervised[†] (Chen et al., 2024)	560M	1024	8192	56.48	69.28	38.52	80.92	54.03	42.26	78.30	32.11
mGTE-CPT	305M	768	512*	60.16	72.89	45.05	84.60	58.41	44.93	80.77	29.94
			8192	60.04	72.70	45.35	84.63	58.36	44.46	80.59	30.77

Table 14: Results on MTEB English subset (Muennighoff et al., 2023). We compare models from the online leaderboard, where derived models (developed by secondary training on other public off-the-shelf models) are not listed. [†]Denote our runs. *To be consistent with the setting in contrastive pre-training, in retrieval tasks, the max sequence length of the document side is set to 1024.

C-MTEB #Datasets (→)	Param.	Dim.	Seq.	Avg. 35	Class. 9	Clust. 4	PairC. 2	Rerank. 4	Retr. 8	STS 8
gte-Qwen2-7b-instruct (Li et al., 2023)	7613M	3584	131072	**72.05**	75.09	66.06 8	7.48	68.92	76.03	65.33
piccolo-large-zh-v2 (Huang et al., 2024)	-	-	-	70.95	74.59	62.17	90.24	70	74.36	63.5
OpenSearch-text-hybrid	-	1792	512	68.71	71.74	53.75	88.1	68.27	74.41	62.46
Baichuan-text-embedding	-	1024	512	68.34	72.84	56.88	82.32	69.67	73.12	60.07
gte-Qwen2-1.5B-instruct (Li et al., 2023)	1776M	1536	131072	67.65	71.12	54.61	86.91	68.21	71.86	60.96
E5-mistral-7b (Wang et al., 2024a)	7111M	4096	32768	60.81	70.17	52.3	72.19	61.86	61.75	50.22
mE5-base (Wang et al., 2024b)	278M	768	514	56.21	65.35	40.68	67.07	54.35	61.63	46.49
mE5-large (Wang et al., 2024b)	560M	1024	514	58.81	67.34	48.23	69.89	56	63.66	48.29
BGE-m3 (dense)† (Chen et al., 2024)	568M	1024	8192	60.80	66.95	45.75	73.98	62.88	65.43	52.43
mGTE-TRM (dense)	305M	768	8192	62.72	64.27	47.48	78.34	68.17	71.95	52.73
BGE-m3-unsupervised† (Chen et al., 2024)	560M	1024	8192	57.53	65.04	47.10	64.09	58.14	61.45	48.42
mGTE-CPT	305M	768	512*	58.67	64.64	50.21	63.95	63.77	64.23	46.74
			8192	58.63	64.38	49.84	63.99	64.13	64.30	46.77

Table 15: Results on C-MTEB (Xiao et al., 2023) (MTEB Chinese). We compare models from the online leaderboard, where derived models (developed by secondary training on other public off-the-shelf models) are not listed. †Denote our runs. *To be consistent with the setting in contrastive pre-training, in retrieval tasks, the max sequence length of the document side is set to 1024.

F-MTEB #Datasets (→)	Param.	Dim.	Seq.	Avg. 26	Class. 6	Clust. 7	PairC. 2	Rerank. 2	Retr. 5	STS 3	Summ. 1
gte-Qwen2-7b-instruct (Li et al., 2023)	7613M	3584	131072	**68.25**	81.76	55.56	90.43	78.7	55.65	82.31	31.45
gte-Qwen2-1.5B-instruct (Li et al., 2023)	1776M	1536	131072	66.6	78.02	55.01	86.88	83.76	52.56	81.26	30.5
voyage-multilingual-2	-	1024	32000	61.65	68.56	46.57	78.66	82.59	54.56	80.13	29.96
voyage-law-2	-	1024	16000	60.58	68.45	44.23	77.3	82.06	52.98	80.29	30.34
mistral-embed	-	1024	-	59.41	68.61	44.74	77.32	80.46	46.81	79.56	31.47
E5-mistral-7b (Wang et al., 2024a)	7111M	4096	32768	48.33	57.72	41.16	76.08	62.2	23.44	65.36	32.22
mE5-base (Wang et al., 2024b)	278M	768	514	56.19	66.8	42.66	74.82	71.76	41.19	77.22	30.76
mE5-large (Wang et al., 2024b)	560M	1024	514	56.07	68.39	38.7	76.19	72.14	42.17	79.37	30.92
BGE-m3 (dense)† (Chen et al., 2024)	568M	1024	8192	58.79	71.57	36.54	79.78	77.36	51.13	80.78	31.05
mGTE-TRM (dense)	305M	768	8192	59.79	68.72	41.66	79.47	76.47	52.97	81.36	29.74
BGE-m3-unsupervised† (Chen et al., 2024)	560M	1024	8192	57.95	69.87	38.43	78.51	75.42	50.05	77.18	28.80
mGTE-CPT	305M	768	512*	59.72	70.79	41.15	80.29	76.19	53.44	76.87	29.04
			8192	59.74	70.69	41.07	79.56	77.10	53.55	77.24	28.74

Table 16: Results on F-MTEB (Ciancone et al., 2024) (MTEB French). We compare top-performing models from the online leaderboard. †Denote our runs. *To be consistent with the setting in contrastive pre-training, in retrieval tasks, the max sequence length of the document side is set to 1024.

MTEB Polish #Datasets (→)	Param.	Dim.	Seq.	Avg. 26	Class. 7	Clust. 1	PairClass. 4	Retr. 11	STS 3
gte-Qwen2-7b-instruct (Li et al., 2023)	7613M	3584	131072	**67.86**	77.84	51.36	88.48	54.69	70.86
gte-Qwen2-1.5B-instruct (Li et al., 2023)	1776M	1536	131072	64.04	72.29	44.59	84.87	51.88	68.12
mmlw-roberta-large (Dadas et al., 2024)	435M	1024	514	63.23	66.39	31.16	89.13	52.71	70.59
mmlw-e5-large (Dadas et al., 2024)	560M	1024	514	61.17	61.07	30.62	85.9	52.63	69.98
mmlw-roberta-base (Dadas et al., 2024)	124M	768	514	61.05	62.92	33.08	88.14	49.92	70.7
mE5-base (Wang et al., 2024b)	278M	768	514	55.62	59.01	24.97	82.15	44.01	65.13
mE5-large (Wang et al., 2024b)	560M	1024	514	60.08	63.82	33.88	85.5	48.98	66.91
BGE-m3 (dense)† (Chen et al., 2024)	568M	1024	8192	60.35	65.15	25.21	86.46	48.51	69.44
mGTE-TRM (dense)	305M	768	8192	58.22	60.15	33.67	85.45	46.40	68.92
BGE-m3-unsupervised† (Chen et al., 2024)	560M	1024	8192	55.98	60.30	40.17	79.01	43.26	67.05
mGTE-CPT	305M	768	512*	57.66	62.72	38.04	79.70	45.55	67.39
			8192	57.11	61.55	38.15	79.53	45.29	66.53

Table 17: Results on MTEB Polish subset (Poświata et al., 2024) We compare top-performing models from the online leaderboard. †Denote our runs. *To be consistent with the setting in contrastive pre-training, in retrieval tasks, the max sequence length of the document side is set to 1024.

	Max Length	Avg	ar	de	en	es	fr	hi	it	ja	ko	pt	ru	th	zh
BM25	8192	53.6	45.1	52.6	57.0	78.0	75.7	43.7	70.9	36.2	25.7	82.6	61.3	33.6	34.6
mE5$_{large}$	512	34.2	33.0	26.9	33.0	51.1	49.5	21.0	43.1	29.9	27.1	58.7	42.4	15.9	13.2
mE5$_{base}$	512	30.5	29.6	26.3	29.2	45.2	46.7	19.0	40.9	24.9	20.9	50.8	37.8	12.2	12.8
E5$_{mistral-7b}$	8192	42.6	29.6	40.6	43.3	70.2	60.5	23.2	55.3	41.6	32.7	69.5	52.4	18.2	16.8
BGE-m3-Dense	8192	52.5	47.6	46.1	48.9	74.8	73.8	40.7	62.7	50.9	42.9	74.4	59.5	33.6	26.0
BGE-m3-Sparse	8192	62.2	58.7	53.0	62.1	87.4	82.7	49.6	74.7	53.9	47.9	85.2	72.9	40.3	40.5
BGE-m3-Dense+Sparse	8192	64.8	63.0	56.4	64.2	88.7	84.2	52.3	75.8	58.5	53.1	86.0	75.6	42.9	42.0
mGTE-TRM Dense	8192	56.6	55.0	54.9	51.0	81.2	76.2	45.2	66.7	52.1	46.7	79.1	64.2	35.3	27.4
mGTE-TRM Sparse	8192	71.0	74.3	66.2	66.4	93.6	88.4	61.0	82.2	66.2	64.2	89.9	82.0	47.4	41.8
mGTE-TRM Dense+Sparse	8192	71.3	74.6	66.6	66.5	93.6	88.6	61.6	83.0	66.7	64.6	89.8	82.1	47.7	41.4
+ mGTE-reranker	8192	73.8	76.6	70.4	69.3	96.4	89.6	67.8	81.9	68.1	71.1	90.2	86.1	46.7	44.8

Table 18: Evaluation of multilingual long-doc retrieval on the MLDR (Chen et al., 2024) testset (measured by nDCG@10).

	Baselines							M3-Embedding					**mGTE-TRM**			**mGTE-reranker**
	BM25	mDPR	mContriever	mE5$_{large}$	mE5$_{base}$	E5$_{mistral-7b}$	OpenAI-3	Dense	Sparse	Multi-vec	D+S	All	Dense	Sparse	D+S	ReRank
ar	13.4	33.8	43.8	59.7	44.3	47.6	55.1	61.9	19.5	62.6	61.9	**63.0**	55.9	17.5	56.0	58.2
da	36.2	55.7	63.3	71.7	63.6	**72.3**	67.6	71.2	45.1	71.7	71.3	72.0	69.8	37.9	69.7	71.0
de	23.3	53.2	60.2	**71.2**	62.3	70.8	67.6	69.8	33.2	69.6	70.2	70.4	68.9	27.0	69.1	70.1
es	29.8	55.4	62.3	70.8	63.8	**71.6**	68.0	69.8	40.3	70.3	70.2	70.7	69.6	35.1	70.0	71.0
fi	33.2	42.8	58.7	67.7	53.0	63.6	65.5	67.8	41.2	68.3	68.4	**68.9**	64.2	35.3	64.5	64.9
fr	30.3	56.5	62.6	69.5	61.2	**72.7**	68.2	69.6	43.2	70.1	70.1	70.8	68.9	36.9	70.4	71.0
he	16.1	34.0	50.5	61.4	37.4	32.4	46.3	63.4	24.5	64.4	63.5	**64.6**	55.4	22.0	55.4	56.5
hu	26.1	46.1	57.1	68.0	55.9	**68.3**	64.0	67.1	34.5	67.3	67.7	67.9	64.6	28.8	65.0	66.1
it	31.5	53.8	62.0	71.2	61.6	**71.3**	67.6	69.7	41.5	69.9	69.9	70.3	69.0	36.2	69.2	70.1
ja	14.5	46.3	50.7	63.1	51.7	57.6	64.2	67.0	23.3	67.8	67.1	**67.9**	65.3	19.5	65.2	67.2
km	20.7	20.6	18.7	18.3	28.2	23.3	25.7	58.5	24.4	59.2	58.9	**59.5**	53.6	21.9	53.8	54.7
ko	18.3	36.8	44.9	58.9	40.4	49.4	53.9	61.9	24.3	63.2	62.1	**63.3**	55.9	21.4	56.1	58.9
ms	42.3	53.8	63.7	70.2	62.4	71.1	66.1	71.6	52.5	72.1	71.8	**72.3**	69.9	47.8	70.2	70.9
nl	42.5	56.9	63.9	**73.0**	65.0	74.5	68.8	71.3	52.9	71.8	71.7	72.3	70.7	47.4	70.9	71.5
no	38.5	55.2	63.0	71.1	62.0	70.8	67.0	70.7	47.0	71.4	71.1	**71.6**	69.1	39.7	69.2	70.2
pl	28.7	50.4	60.9	70.5	57.2	**71.5**	66.1	69.4	36.4	70.0	69.9	70.4	68.4	31.4	68.3	69.6
pt	31.8	52.5	61.0	66.8	58.7	**71.6**	67.7	69.3	40.2	70.0	69.8	70.6	69.6	34.9	69.6	70.7
ru	21.8	49.8	57.9	**70.6**	58.7	68.7	65.1	69.4	29.2	70.0	69.4	70.0	69.5	25.8	69.5	69.6
sv	41.1	54.9	62.7	72.0	61.3	**73.3**	67.8	70.5	49.8	71.3	71.5	71.5	69.5	43.3	69.9	70.6
th	28.4	40.9	54.4	69.7	59.7	57.1	55.2	69.6	34.7	70.5	69.8	**70.8**	65.0	30.6	65.2	66.9
tr	33.5	45.5	59.9	67.3	59.2	65.5	64.9	68.2	40.9	69.0	69.1	**69.6**	67.7	36.0	67.7	69.0
vi	33.6	51.3	59.9	68.7	60.0	62.3	63.5	69.6	42.2	70.5	70.2	**70.9**	69.4	37.6	69.3	70.3
zh_cn	19.4	50.1	55.9	44.3	38.3	61.2	62.7	66.4	26.9	66.7	66.6	**67.3**	68.2	23.2	68.4	69.5
zh_hk	23.9	50.2	55.5	46.4	38.3	55.9	61.4	65.8	31.2	66.4	65.9	**66.7**	63.7	27.8	63.8	65.8
zh_tw	22.5	50.6	55.2	45.9	39.0	56.5	61.6	64.8	29.8	65.3	64.9	**65.6**	63.8	26.6	63.9	65.7
Avg	28.1	47.9	56.3	63.5	53.7	62.4	62.1	67.8	36.3	68.4	68.1	**68.8**	65.8	31.6	66.0	67.2

Table 19: Recall@20 on MKQA (Longpre et al., 2021) dataset for cross-lingual retrieval in all 25 languages. The All of M3-Embedding denotes the hybrid retrieval result of dense, sparse, and multi-vec scores.

Model	Avg	ar	bn	en	es	fa	fi	fr	hi	id	ja	ko	ru	sw	te	th	zh	de	yo
BM25	31.9	39.5	48.2	26.7	7.7	28.7	45.8	11.5	35.0	29.7	31.2	37.1	25.6	35.1	38.3	49.1	17.5	12.0	56.1
mE5$_{large}$	65.4	76.0	75.9	52.9	52.9	59.0	77.8	54.5	62.0	52.9	70.6	66.5	67.4	74.9	84.6	80.2	56.0	56.4	56.5
mE5$_{base}$	60.13	71.6	70.2	51.2	51.5	57.4	74.4	49.7	58.4	51.1	64.7	62.2	61.5	71.1	75.2	75.2	51.5	43.4	42.3
E5$_{mistral-7b}$	62.2	73.3	70.3	57.3	52.2	52.1	74.7	55.2	52.1	52.7	66.8	61.8	67.7	68.4	73.9	74.0	54.0	54.0	58.8
OpenAI-3	54.9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
BGE-M3-Dense	67.8	78.4	80.0	56.9	55.5	57.7	78.6	57.8	59.3	56.0	72.8	69.9	70.1	78.6	86.2	82.6	61.7	56.8	60.7
BGE-M3-Sparse	53.9	67.1	68.7	43.7	38.8	45.2	65.3	35.5	48.2	48.9	56.3	61.5	44.5	57.9	79.0	70.9	36.3	32.2	70.0
BGE-M3-Multi-vec	69.0	79.6	81.1	59.4	57.2	58.8	80.1	59.0	61.4	58.2	74.5	71.2	71.2	79.0	87.9	83.0	62.7	57.9	60.4
BGE-M3-Dense+Sparse	68.9	79.6	80.7	58.8	57.5	59.2	79.7	57.6	62.8	58.3	73.9	71.3	69.8	78.5	87.2	83.1	62.5	57.6	**61.8**
BGE-M3 All	**70.0**	80.2	81.5	59.8	59.2	60.3	80.4	60.7	63.2	59.1	75.2	72.2	71.7	79.6	88.2	83.8	63.9	59.8	61.5
mGTE-TRM Dense	62.1	71.4	72.7	54.1	51.4	51.2	73.5	53.9	51.6	50.3	65.8	62.7	63.2	69.9	83.0	74.0	60.8	49.7	58.3
mGTE-TRM Sparse	55.9	66.5	70.4	35.6	46.2	40.0	47.6	66.5	39.8	48.9	47.9	59.3	64.3	47.1	59.4	83.0	70.5	73.7	39.9
mGTE-TRM Dense+Sparse	63.5	73.4	75.1	49.9	57.6	62.7	52.0	74.7	53.5	56.4	52.8	67.1	66.7	63.5	69.5	85.2	75.8	58.4	58.8
+ mGTE-reranker	68.5	77.1	63.1	78.6	56.3	72.4	80.3	79.6	58.6	59.1	74.6	75.5	59.4	56.3	56.5	62.2	72.2	86.3	65.1

Table 20: Multi-lingual retrieval performance on the MIRACL (Zhang et al., 2023b) dev set (measured by nDCG@10).

BEIR	Avg.	Argu-Ana	Cli-mate-Fever	CQA-Dup-Stack	DB-Pedia	Fever	FiQA	Hotpot-QA	MS MAR-CO	NF-Corpus	NQ	Quora	Sci-docs	Sci-fact	Touche-2020	Trec-Covid
gte-Qwen2-7B-instruct	60.25	64.27	45.88	46.43	52.42	95.11	62.03	73.08	45.98	40.6	67	90.09	28.91	79.06	30.57	82.26
NV-Embed-v1	59.36	68.2	34.72	50.51	48.29	87.77	63.1	79.92	46.49	38.04	71.22	89.21	20.19	78.43	28.38	85.88
gte-Qwen2-1.5B-instruct	58.29	69.72	42.91	44.76	48.69	91.57	54.7	68.95	43.36	39.34	64	89.64	24.98	78.44	27.89	85.38
voyage-large-2-instruct	58.28	64.06	32.65	46.6	46.03	91.47	59.76	70.86	40.6	40.32	65.92	87.4	24.32	79.99	39.16	85.07
neural-embedding-v1	58.12	67.21	32.3	49.11	48.05	89.46	58.94	78.87	42	42.6	68.36	89.02	27.69	78.82	24.06	75.33
GritLM-7B	57.41	63.24	30.91	49.42	46.6	82.74	59.95	79.4	41.96	40.89	70.3	89.47	24.41	79.17	27.93	74.8
e5-mistral-7b-instruct	56.89	61.88	38.35	42.97	48.89	87.84	56.59	75.72	43.06	38.62	63.53	89.61	16.3	76.41	26.39	87.25
google-gecko	55.7	62.18	33.21	48.89	47.12	86.96	59.24	71.33	32.58	40.33	61.28	88.18	20.34	75.42	25.86	82.62
text-embedding-3-large	55.44	58.05	30.27	47.54	44.76	87.94	55	71.58	40.24	42.07	61.27	89.05	23.11	77.77	23.35	79.56
gte-en-large-v1.5	57.91	72.11	48.36	42.16	46.3	93.81	63.23	68.18	42.93	36.95	56.08	89.67	26.35	82.43	22.55	77.49
gte-en-base-v1.5	54.09	63.49	40.36	39.52	39.9	94.81	48.65	67.75	42.62	35.88	52.96	88.42	21.92	76.77	25.22	73.13
BM25	41.7	31.5	21.3	29.9	31.3	75.3	23.6	60.3	22.8	32.5	32.9	78.9	15.8	66.5	36.7	65.6
mE5-large	51.43	54.38	25.73	39.68	41.29	82.81	43.8	71.23	43.7	33.99	64.06	88.18	17.47	70.41	23.39	71.33
mE5-base	48.88	44.23	23.86	38.52	40.36	79.44	38.17	68.56	42.27	32.46	60.02	87.65	17.16	69.35	21.35	69.76
BGE-M3 Dense†	48.34	53.95	29.52	39.09	39.80	81.38	41.30	69.44	38.32	31.43	60.60	88.57	16.39	64.36	22.63	55.59
BGE-M3 Sparse†	38.30	25.08	24.69	27.51	23.21	88.36	26.79	68.45	19.59	27.5	17.98	73.82	8.89	64.37	30.26	48.00
BGE-M3 Dense+Sparse†	49.41	53.88	30.21	39.10	39.89	81.24	40.25	70.11	37.62	32.53	59.58	88.62	15.59	65.74	31.12	55.67
mGTE-TRM Dense	51.07	58.36	34.83	38.12	40.11	92.07	44.99	63.03	39.92	36.66	58.10	88.02	18.26	73.42	22.76	57.4
mGTE-TRM Sparse	39.24	40.06	24.17	25.11	20.0	88.32	28.58	64.68	19.39	28.34	19.71	76.84	10.92	67.72	21.52	53.33
mGTE-TRM Dense+Sparse	51.43	58.48	34.89	38.36	39.72	93.14	44.98	65.01	39.99	36.67	56.90	89.05	18.26	73.45	24.09	58.46
+ mGTE-reranker	55.42	58.53	44.93	38.37	45.62	93.9	44.38	74.51	44.99	36.29	65.21	81.67	18.42	75.59	31.29	77.75
BGE-M3-unsupervised†	42.26	59.07	23.05	38.10	31.16	59.15	36.57	53.39	27.79	30.67	39.69	86.38	15.08	61.26	17.62	54.90
mGTE-CPT-512,1024	44.93	52.99	17.93	45.01	37.63	34.13	48.38	54.39	31.76	39.01	48.48	86.82	22.95	72.46	18.56	63.46
mGTE-CPT-8192	44.46	55.14	15.85	44.73	38.74	27.42	47.45	55.93	31.79	38.62	49.27	86.81	22.72	73.08	17.08	62.27

Table 21: BEIR benchmark (Thakur et al., 2021) nDCG@10 scores. We include top models from MTEB Retrieval English leaderboard. †Denote our runs.

Model	Param.	Dim.	Seq	Avg.	Tau Scr.	Tau Gov.	Tau QMS.	QASP. Tit. Art.	QASP. Abs. Art.
Jina_{base-v2} (Günther et al., 2023)	137M	768	8192	85.5	93.3	98.6	40.8	95.1	99.3
nomic-embed-text-v1 (Nussbaum et al., 2024)	137M	768	8192	85.5	90.9	97.8	44.2	94.9	99.9
text-embedding-3-small	-	1536	8192	82.4	92.2	97.7	27.4	95.9	98.9
text-embedding-3-large	-	3072	8192	79.4	88.0	93.6	25.5	93.2	96.8
mGTE-en-base-embed	137M	768	8192	87.4	91.8	98.6	49.9	97.1	99.8
mGTE-en-large-embed	434M	1024	8192	86.7	92.6	98.7	44.5	97.8	99.8
mE5_{base} (Wang et al., 2024b)	279M	768	512	72.2	68.9	87.6	30.5	85.1	88.9
mE5_{large} (Wang et al., 2024b)	279M	1024	512	74.3	70.4	89.5	37.6	89.5	85.4
E5_{mistral} (Wang et al., 2024a)	7B	4096	4096	87.8	95.9	98.3	46.8	98.4	99.8
BGE-M3-Dense† (Chen et al., 2024)	568M	1024	8192	84.9	93.8	97.4	41.9	93.2	98.3
BGE-M3-Sparse† (Chen et al., 2024)	568M	1024	8192	84.9	95.5	97.9	46.7	85.7	98.9
BGE-M3-Dense+Sparse† (Chen et al., 2024)	568M	1024	8192	87.4	97.7	98.2	47.7	93.6	99.7
mGTE-TRM Dense	434M	768	8192	88.9	95.1	97.7	58.5	94.6	98.7
mGTE-TRM Sparse	434M	768	8192	88.1	97.6	97.9	60.1	85.5	99.2
mGTE-TRM Dense+Sparse	434M	768	8192	91.3	98.2	98.3	66.5	94.6	98.7

Table 22: The nCDG@10 scores on the LoCo benchmark (Saad-Falcon et al., 2024). †Denote our runs.