DRAMA: Diverse Augmentation from Large Language Models to Smaller Dense Retrievers

Xueguang Ma[2,*,†] **Xi Victoria Lin**[1] **Barlas Oguz**[1] **Jimmy Lin**[2]
Wen-tau Yih[1] **Xilun Chen**[1,*]

[1]FAIR at Meta [2]University of Waterloo

x93ma@uwaterloo.ca, xilun@meta.com

Abstract

Large language models (LLMs) have demonstrated strong effectiveness and robustness when fine-tuned as dense retrievers. However, their large parameter size presents significant computational challenges at inference time. While smaller retrievers offer better efficiency, they often fail to generalize effectively with limited supervised fine-tuning data. In this work, we introduce DRAMA, a training framework that leverages LLMs to train smaller generalizable dense retrievers. In particular, we adopt pruned LLMs as the backbone and train on diverse LLM-augmented data in a single-stage contrastive learning setup. Experiments show that DRAMA offers better multilingual and long-context capabilities than traditional encoder-based retrievers, and achieves strong performance across multiple tasks and languages.[1]

1 Introduction

Recent advancements in large language models (LLMs) have demonstrated their effectiveness and robustness in text retrieval tasks (Muennighoff et al., 2024; Sun et al., 2023; Li et al., 2025; BehnamGhader et al., 2024; Lee et al., 2025). Directly fine-tuning advanced billion-parameter LLMs with available annotated data can generalize significantly better than fine-tuning a pre-LLM-era smaller model with only a few hundred million parameters (Ma et al., 2024; Luo et al., 2024). However, the large parameter size of LLMs brings non-negligible inference-time compute costs, such as encoding large-scale corpora and increased query latency. For example, using Llama3.1_{8B} (Grattafiori et al., 2024) as the backbone increases the inference cost nearly 40× compared to a dense retriever based on BERT.

In this work, we holistically explore how to effectively leverage LLMs to create smaller retrievers, in terms of both *data* and *model backbone*, to develop generalizable yet efficient dense retrievers with fewer than 1B parameters.

Although several works have discussed using LLMs for retrieval data augmentation, such as directly generating training triplet (Wang et al., 2024b) or using LLM to mine positive and negative documents from a real corpus (Lee et al., 2024), the effectiveness of these methods has not been rigorously compared. We comprehensively study the effectiveness of multiple LLM data augmentation methods with a controlled setup: using the same models and corpora across different data creation methods and only relying on open-sourced models and open-access data. Specifically, we utilize an LLM retriever based on Llama3.1_{8B} and an instruction-tuned LLM based on Llama3.3_{70B}-Instruct to generate augmentation data. This includes computationally cheap approaches such as generating cropped sentences as queries and using an LLM retriever to mine positive and negative documents over a corpus, as well as computationally expensive methods that further utilize Instruct-LLMs to generate queries and provide relevance judgment. We investigate the effectiveness of various combinations of these diverse LLM augmentations, providing high-quality augmented training data for English and multilingual retrieval.

Existing smaller dense retriever models are mostly based on encoder-only architectures (Wang et al., 2023; Chen et al., 2024; Warner et al., 2024). We instead propose to leverage LLMs as the backbone for dense retrievers by pruning the decoder-only LLM into a smaller size to initialize the text encoder. Specifically, we further prune Llama3.2_{1B} (which is pruned from Llama3.1_{8B}) into 0.1B (on par with BERT-base) and 0.3B (on par with XLM-RoBERTa-Large), while preserving multilingual and long-context capability. We demonstrate that

[*]Equal contribution. [†]Work done while at Meta.
[1]Code and checkpoints will be available at https://github.com/facebookresearch/dpr-scale/tree/main/drama.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 30170–30186

by simply turning on the bi-directional attention during retriever training, the pruned decoder-only models perform well as retrievers. This offers a more flexible way to create smaller dense retrievers with arbitrary sizes while still leveraging pretrained LLM weights, making smaller retrievers compatible with current and future LLM advancements.

Combining LLM-based data augmentation and backbones, we introduce a single-stage training framework: DRAMA (smaller Dense Retriever from diverse LLM AugMentAtion). Our smaller retriever models achieve high performance on BEIR (Thakur et al., 2021), MIRACL (Zhang et al., 2023), and multiple multilingual retrieval tasks on MTEB (Muennighoff et al., 2022), outperforming traditional encoder-based retrievers that rely on multi-stage contrastive learning. This demonstrates that our training framework produces models that excel across diverse English retrieval tasks and exhibit robust multilingual performance, showing the potential for unified smaller retrievers that perform effectively across tasks and languages.

In summary, our contributions are as follows:

- We investigate diverse methods for leveraging LLMs to generate data augmentation for training smaller models, analyzing their individual and combined effectiveness.
- We prune LLMs to derive smaller decoder-only language models as backbones for retrievers, demonstrating their advantages in generalizability, multilingual effectiveness, and long-context extrapolation.
- Our training framework produces a series of multilingual and generalizable smaller retrievers, highlighting the benefits of aligning smaller retriever training with ongoing advancements in LLMs.

2 Related Work

2.1 Robust Dense Retrieval

Dense Passage Retrieval (Karpukhin et al., 2020) utilizes a pre-trained language model such as BERT (Devlin et al., 2019), to encode text into dense vectors and conduct passage retrieval as a nearest neighbor search. This approach has shown strong in-domain effectiveness compared to traditional lexical retrievers such as BM25 (Robertson and Zaragoza, 2009). However, dense retrievers have been found to struggle with generalization when applied to out-of-domain retrieval tasks (Thakur et al., 2021). To address this issue, various works have aimed to improve the generalization of dense retrievers through continuous pre-training tailored for retrieval tasks.

Works such as Condenser (Gao and Callan, 2021), RetroMAE (Xiao et al., 2022), and SimLM (Wang et al., 2023) have enhanced the dense representation of BERT via customized architectures during language modeling. Other works, including Contriever (Izacard et al., 2022), GTE (Li et al., 2023), E5 (Wang et al., 2024a) have further adapted two-stage contrastive learning. These models are first trained with unsupervised or weakly supervised large-scale contrastive learning, followed by supervised contrastive learning with available relevance-judged data (Nussbaum et al., 2024; Yu et al., 2024). CDE (Morris and Rush, 2025) further proposes a two-stage model architecture that integrates corpus-level information into document embeddings.

2.2 LLM for Text Ranking

On the other hand, recent large language models have shown strong potential in relevance modeling for text ranking. Fine-tuning LLM as dense retriever models have shown significantly stronger effectiveness across various tasks and languages compared to smaller ones (Wang et al., 2024b; Muennighoff et al., 2024; Springer et al., 2024; Li et al., 2025). For example, RepLlama (Ma et al., 2024), which uses straightforward supervised fine-tuning based on the Llama2-7B model, outperforms previous smaller retriever models that were based on multi-stage continuous pre-training, with a lower training cost. This demonstrates the data efficiency and naturally strong generalization of LLM-based retrievers (Luo et al., 2024).

Moreover, instruction-following LLMs have also shown better performance when directly prompted as rerankers (Ma et al., 2023; Sun et al., 2023). Reflecting the excel relevance understanding of large language models for retrieval. In this work, we aim to leverage the characteristics of LLM-based ranking methods that are data-efficient and generalizable, shifting their high inference time costs into training time costs as data augmentation.

2.3 Data Augmentation for Retriever

InPars (Bonifacio et al., 2022) and Promptagator (Dai et al., 2023) generate synthetic queries that align with given documents sampled from the task corpus, creating training data for retrieval corpora with limited human queries and judgments.

DRAGON (Lin et al., 2023) enhances the robustness of dense retrievers by employing sentence cropping as pseudo-queries and generating augmented data based on retrieval results from multiple retrievers (e.g., sparse, multi-vector models).

With the emergence of LLMs, Mistral-E5 (Wang et al., 2024b) directly prompts an LLM to generate synthetic query-positive-negative triplets, using them as augmentation data to train a 7B LLM retriever across diverse text embedding tasks. Gecko (Lee et al., 2024) takes a different approach by leveraging real documents: it generates synthetic queries from sampled real documents, retrieves top candidate passages, and uses an LLM to rerank them in pointwise way. While these methods introduce various strategies for data augmentation in retrievers, they have not been systematically compared within a single framework where LLMs and corpora are controlled for fair comparison. We explore various types of LLM-based data augmentation and evaluate their individual and combined effectiveness.

2.4 Multilingual Retriever

Multilingual capabilities are crucial for effective retrieval systems. While numerous multilingual retrievers have been developed (Izacard et al., 2022; Wang et al., 2024c; Zhang et al., 2024; Chen et al., 2024), they often face a trade-off between achieving strong performance in multilingual retrieval across various languages and preserves good English generalization performance on English retrieval. While concurrent work ArcticEmbV2 (Yu et al., 2024) also aims for effectiveness in both English and multilingual, they follow the previous training paradigm that firstly pretrain the model with contrastive learning over weakly supervised data pairs and then followed by supervised fine-tuning. In our work, we address this challenge from a different view, by conducting data augmentation from LLM and using pruned LLM as the backbone of smaller retriever.

3 Methods

3.1 Data Augmentation for Contrastive Dense Retriever Training

Given a query q, a positive document D^+ relevant to the query, and a set of hard negative documents $\{D_{\text{HN}}\}$ that are similar to the positive document but are not highly relevant to the query, a dense retriever model is trained using the InfoNCE loss (van den Oord et al., 2019) as follows:

$$\mathcal{L}(q, D^+, \{D_{\text{N}}\}) = -\log p(D = D^+ \mid q)$$
$$= -\log \frac{\exp(\text{Sim}(q, D^+)/\tau)}{\sum_{D_i \in \{D^+\} \cup \{D_{\text{N}}\}} \exp(\text{Sim}(q, D_i)/\tau)},$$

where $\{D_{\text{N}}\}$ is the union of the hard negative documents $\{D_{\text{HN}}\}$ for each query and in-batch negative documents, which are positive or hard negatives from other queries in the same training batch. The similarity $\text{Sim}(Q, D)$ is commonly computed as the cosine similarity between the embedding vectors of the query and document.

Data augmentation for dense retrieval focuses on creating triplets of queries q, positive documents D_{P}, and hard negative documents $\{D_{\text{HN}}\}$. In this work, we make the following assumptions regarding available resources for data augmentation:

- **Initial Supervised Data** (D_{sft}): A commonly accessible general-domain retrieval dataset.
- **Large Retrieval Model** (LLM_{Ret}): An LLM-based retrieval model, fine-tuned on D_{sft}.
- **Instruction-following LLM** (LLM_{Inst}): An LLM with strong instruction-following capability that can generate synthetic data reflecting its relevance preferences.
- **Large Corpus** (C): A diverse or multilingual document corpus that serves as the basis for synthetic query generation and relevance assessment.

With the above assumption, we explored various ways of utilizing LLM to conduct data augmentation for smaller retrievers, ranging from lower to higher computational costs for data creation.

3.1.1 Data Augmentation via Llama-3.1$_{\text{8B}}$ Retriever

Given an LLM-based retriever model, one of the simplest approaches to data augmentation, without relying on even larger LLMs, is to enable the smaller retriever to learn from the relevance preferences of the 8B embedding model LLM_{emb}. Inspired by methods such as SPAR (Chen et al., 2022) and DRAGON (Lin et al., 2023), we begin with the corpus C. For each document in C, we perform random sentence cropping to extract a smaller segment, which is treated as pseudo-query q. These pseudo-queries, along with the full corpus, are encoded using the 8B retriever model. Retrieval is

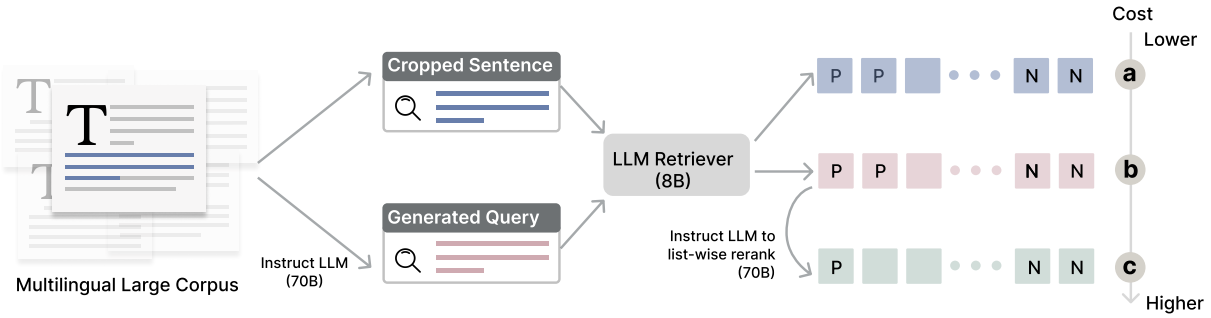

Figure 1: Methods to create data augmentation for smaller retriever with LLMs: (a) Using cropped sentences as queries, selecting the top-ranked documents from top-k retrieval as positives and the remaining as hard negatives. (b) Replacing cropped sentences with synthetic queries generated by prompting instruction-following LLM. (c) Refining retrieval results from the LLM retriever using an instruction-following LLM as a listwise reranker.

then conducted for each pseudo-query q to identify the top-k candidate documents. Among these candidates, the top $[1, m]$ documents are regarded as positive D^+, while the top $[k-n, k]$ documents are designated as hard negatives D_{HN}. The process is illustrated in Figure 1.a. In this work, we set $k = 50, m = 10, n = 20$. This process generates an augmented dataset of size $|C|$, as each document contributes to the creation of pseudo-queries and their associated retrieval results. The computational cost of this method is limited to encoding the $|C|$ documents and pseudo-queries, followed by the retrieval of top-k candidates for each query.

3.1.2 Synthetic Queries from Llama-3.3$_{\text{70B}}$-Instruct

The availability of instruction-following LLMs, such as Llama-3.3$_{\text{70B}}$-Instruct, enables the generation of synthetic queries that are more similar to real queries compared to those from random sentence cropping. For each document in the corpus C, we prompt the LLM to generate a synthetic query q. Similar to the above process, these LLM-generated queries are fed into the 8B LLM$_{\text{Ret}}$ to perform retrieval. Based on the retrieval results, we can identify positive documents and hard negative documents for the synthetic queries as illustrated in Figure 1.b. While this approach retains the costs of encoding and retrieval, it introduces an additional computational cost associated with generating synthetic queries using the instruction-following LLM.

3.1.3 LLM Ranking Preference from Llama-3.3$_{\text{70B}}$-Instruct

Instead of relying solely on the relevance preferences of the 8B embedding model, which are influenced by its fine-tuning on supervised data D_{sft}, the instruction-following LLM such as Llama-3.3$_{\text{70B}}$-Instruct can be further leveraged to refine relevance judgments. Specifically, we prompt the LLM to perform listwise reranking of the top-k candidates retrieved for each synthetic query, as illustrated in Figure 1.c. In this process, the LLM provides its relevance judgments by reranking the candidates. The top-1 candidate after reranking is treated as the positive document D^+, while the top $[k-n, k]$ candidates from the reranked list are designated as hard negatives D_{HN}. In our experiments, we set $k = 20, n = 10$. This listwise reranking approach aligns more closely with how humans select the most relevant one among multiple candidates. This data augmentation incurs additional computational costs of prompting the instruction-following LLM for reranking k candidates for every query.

In practice, having the data augmentation from LLM listwise rerank can further improve the LLM$_{\text{Ret}}$ by combining the augmented data with the initial supervised data D_{sft}. We sampled LLM listwise rerank augmented data as the same amount of D_{sft} to re-train the LLM$_{\text{Ret}}$. The effectiveness of this operation is further analyzed in Section. 6.1.

3.1.4 Triplet Generation from Llama-3.3$_{\text{70B}}$-Instruct

Another approach to leverage the LLM's relevance preferences for data augmentation is to directly prompt the LLM to generate triplets consisting of a query, a positive document, and a hard negative document. This approach does not rely on a pre-existing corpus to provide seed documents. Following Mistral-E5 (Wang et al., 2024b), but adhering to our controlled data augmentation framework (i.e., creating the same amount of augmentation data with the same LLM), we first prompt the LLM to brainstorm $|C|$ retrieval tasks. Each task includes a retrieval scenario t, a query q, and its context. Based on the task and query, the LLM is

then prompted to generate a corresponding positive document and a hard negative document. While this method appears promising in theory, our experiments revealed that purely synthetic triplet data generated in this manner does not substantially improve the training of smaller retriever models. Detailed analyses can be found in Section 6.1.

3.2 Pruning

Previous pre-LLM-era retriever models predominantly utilized encoder-only architectures, such as BERT-base for English retrieval and XLM-RoBERTa-Large for multilingual retrieval. In this work, in addition to leveraging LLMs for data augmentation, we investigate whether recent decoder-only LLMs can provide better backbones for smaller retriever models. We perform structured pruning on an LLM to obtain models with non-embedding parameter sizes of 0.1B and 0.3B, making them comparable to BERT-base and XLM-RoBERTa-Large, respectively. Specifically, we initialize the pruning process with Llama3.2$_{\text{1B}}$, itself a pruned version of Llama3.1$_{\text{8B}}$. Following the methodology from ShearedLlama (Xia et al., 2024), the pruning process is performed in two stages.

In the pruning stage, a parameter mask is learned to selectively prune the model, formulated as a constrained optimization problem. Pruning masks z are applied to model hard concrete distributions (Louizos et al., 2018). Constraints are enforced via Lagrange multipliers to ensure the resulting model adheres to the target architecture. For example, the loss function applied to a layer for a target number of attention heads $H_{\mathcal{T}}$ is defined as (Xia et al., 2024):

$$\tilde{\mathcal{L}}^{\text{head}}(\lambda, \phi, z) = \lambda^{\text{head}} \cdot (\sum_i z_i^{\text{head}} - H_{\mathcal{T}}) + \phi^{\text{head}} \cdot (\sum_i z_i^{\text{head}} - H_{\mathcal{T}})^2$$

Similarly, the full pruning loss integrates constraints on other structural components, including the layer mask z^{layer}, hidden dimension mask z^{hidden}, attention head mask z^{head}, and intermediate dimension mask z^{int}. These are combined with the standard language modeling objective as:

$$\mathcal{L}_{\text{prune}}(\theta, z, \lambda, \phi) = \mathcal{L}(\theta, z) + \sum_{j=1}^{L_S} \tilde{\mathcal{L}}_j^{\text{head}} + \sum_{j=1}^{L_S} \tilde{\mathcal{L}}_j^{\text{int}} + \tilde{\mathcal{L}}^{\text{layer}} + \tilde{\mathcal{L}}^{\text{hidden}}$$

This is followed by a continuous pretraining stage, during which the pruned model is further trained only on the language modeling objective to recover its performance.

Pruning from an LLM offers several potential advantages compared to training traditional pretrained language models. First, it allows us to leverage the latest advancements in LLMs, which are trained on large-scale, high-quality datasets and exhibit strong generalization and multilingual capabilities. Secondly, it supports longer contexts than earlier models, allowing for improved handling of retrieval scenarios requiring extended input sequences. Thirdly, the pruning process provides the flexibility to tailor model sizes based on specific deployment needs.

4 Experiment Setup

4.1 Fine-tuning Data

Controlling the supervised fine-tuning data is critical for ensuring a fair comparison across methods when studying the generalizability of retrieval models. BEIR (Thakur et al., 2021) was originally designed for zero-shot evaluation, encouraging the use of MS MARCO Passage Retrieval as the sole fine-tuning dataset. However, many recent retrievers incorporate supervised data from the evaluation tasks, making the evaluation not entirely zero-shot. To balance fairness in assessing model generalization while maintaining adequate baselines for comparison, we follow the fine-tuning data setup of E5 (Wang et al., 2024a). This setup includes general-domain retrieval datasets but not include fine-tuning data for domain-specific retrieval tasks such as financial QA or scientific document retrieval. For our experiments, we use the open-source replication of the E5 fine-tuning data (Li et al., 2025).

4.2 Data Augmentation

For the LLM retriever model LLM$_{\text{ret}}$, we initialize it with Llama3.1$_{\text{8B}}$ and first fine-tune it following the training recipe of RepLlama (Ma et al., 2024) for one epoch on the MS MARCO Passage Ranking training set (Bajaj et al., 2018). We then further fine-tune it on the aforementioned E5 fine-tuning data to obtain an LLM retriever focusing on English retrieval. We train another multilingual LLM retriever by continuous fine-tuning of the MS MARCO-trained LLM retriever using only the MIRACL (Zhang et al., 2023) training data.

This allows us to better study generalization in the multilingual retrieval setting.

For the large corpus C used in English data augmentation, we sample 25M documents from a diverse open web-crawled dataset. For multilingual augmentation, we use a combination of multilingual Wikipedia and a multilingual web-crawled corpus covering 19 non-English languages, with each corpus containing 25M documents. In both cases, we segment documents into text chunks of up to 256 tokens.

4.3 Pruning

We prune Llama3.2$_{1B}$ into 0.1B and 0.3B models, where the first pruning stage cost about 0.5 billion tokens and the second continuous pertaining stage cost 26 billion tokens. The data covering English and 19 non-English languages from web-crawled corpora. The pruned models support a maximum context length of 8,192 tokens. Detailed model configurations can be found in Appendix A.3.

4.4 Training

The full training data for the smaller retriever models consists of: (1) LLM augmented data based on cropped sentences. (2) 25M LLM retriever augmented data based on generated queries. (3) 25M Inst-LLM listwise reranker augmented data based on generated queries. These three types of data augmentation are applied to all sources, including English web-crawl corpora, multilingual Wikipedia, and multilingual web-crawl corpora (denoted as enWeb, mWiki, and mWeb respectively). The sampling ratio of augmented data across these three sources is 2:1:1. Additionally, the augmented data is mixed with the E5 supervised fine-tuning data, which contains approximately 2M instances. See Appendix A.1 for a detailed ablation study on different sampling ratios, and Appendix A.4 for additional training details.

We train the model with each query paired with one positive document and seven hard negative documents for the 0.1B and 0.3B models and three hard negative documents for the 1B model. We adopt the Matryoshka Representation Learning (MRL) during training to enable flexible dimensionality choice (Kusupati et al., 2022). See Appendix A.2 for the effectiveness of DRAMA with different dimensionality configuration.

4.5 Evaluation

Our main evaluations are conducted on BEIR (Thakur et al., 2021) and MIRACL (Zhang et al., 2023), to assess the generalization of dense retrievers and multilingual retrieval capability. To further analyze the generalization of multilingual retrievers, we also evaluate on retrieval subsets of MTEB-FR (Ciancone et al., 2024), MTEB-ZH (Xiao et al., 2024) and MTEB-DE. To assess the effectiveness of long-context retrieval, which benefits from pruning an LLM, we evaluate on MLDR (Chen et al., 2024), a benchmark for long-context multilingual retrieval across 13 languages. We also include evaluations on synthetic long-context retrieval benchmarks, NeedleRetrieval and PasskeyRetrieval, from the LongEmbed study (Zhu et al., 2024). We use nDCG@10 as the metrics for all evaluations.

4.6 Baselines

We select representative baselines with similar retrieval task training data settings, as described in Sec. 4.1. The major baselines include Contriever (Izacard et al., 2022), DRAGON (Lin et al., 2023), E5 (Wang et al., 2024a), BGE (Xiao et al., 2024), mE5 (Wang et al., 2024c), BGE-M3 (Chen et al., 2024), mGTE (Zhang et al., 2024), ArcticEmbV2 (Yu et al., 2024), Gecko (Lee et al., 2024), and MistralE5 (Wang et al., 2024b).

5 Results

5.1 Generalization of Smaller Retrievers

Table 1 shows the performance of our DRAMA variants on both English and multilingual retrieval tasks. The results indicate that DRAMA is a strong and generalizable retriever at different model sizes. For example, DRAMA$_{0.1B}$ achieves an nDCG@10 of 56.9 on BEIR, on par with ArcticEmb-v2-M, and outperforms other English-only and multilingual retrievers. When scaling up to DRAMA$_{0.3B}$, the score increases to 58.0, outperforming ArcticEmb-v2-L by 0.8 points and matching Gecko, which is a much larger 1B-parameter model. Beyond English retrieval, DRAMA exhibits strong multilingual capabilities. On MIRACL, all DRAMA variants (from 0.1B to 1B) outperform previous best models like M3-BGE-Dense, while also maintaining strong English retrieval performance. This suggests that DRAMA works well across different languages without losing effectiveness in English.

Method	Non-Emb. Param.	Repre. Dim.	Contra. Pretrain.	Data Aug.	Multi. Lang.	English	Multilingual			
						BEIR (13)	MIRACL (18)	MTEB-FR (5)	MTEB-ZH (8)	MTEB-DE (4)
BM25	-	-	×	×	✓	43.7	38.5	-	-	-
Contriever	86M	768	✓	×	×	47.5	-	-	-	-
DRAGON	86M	768	×	✓	×	50.2	-	-	-	-
E5-v2-base	86M	768	✓	×	×	51.9	-	-	-	-
bge-base-en-v1.5	86M	768	✓	×	×	55.0	-	-	-	-
mE5-base	86M	768	✓	×	✓	50.2	60.1	45.4	61.6	49.2
mGTE-Dense	113M	768	✓	×	✓	54.3	62.1	50.6	**72.0**	49.1
ArcticEmb-v2-M	113M	768	✓	×	✓	56.9	59.2	53.7	55.7	55.0
DRAMA$_{0.1B}$	113M	768	×	✓	✓	56.9	70.4	52.1	61.7	55.1
E5-large-v2	303M	1024	✓	×	×	52.1	-	-	-	-
bge-large-en-v1.5	303M	1024	✓	×	×	56.1	-	-	-	-
mE5-large	303M	1024	✓	×	✓	52.9	65.4	47.7	63.7	50.4
mE5-Inst	303M	1024	✓	✓	✓	54.1	66.0	49.9	64.2	52.5
M3-BGE-Dense	303M	1024	✓	×	✓	50.0	69.2	48.6	65.6	50.4
ArcticEmb-v2-L	303M	1024	✓	×	✓	57.2	64.9	54.5	63.6	**55.9**
DRAMA$_{0.3B}$	265M	1024	×	✓	✓	58.0	71.4	54.8	63.0	55.6
Gecko	1B	768	✓	✓	✓	58.0	56.2	-	-	-
DRAMA$_{1B}$	1B	2048	×	✓	✓	**59.1**	**71.7**	**57.6**	63.7	56.2
DRAMA$_{1B}$ (768d)	1B	768	×	✓	✓	58.5	70.9	56.5	62.8	55.8
MistralE5	7B	4096	×	✓	✓	59.0	62.2	-	-	-

Table 1: Effectiveness of DRAMA compared to baseline methods (measured in nDCG@10). For each method, we indicate the number of non-embedding parameters, the text embedding dimensionality, whether contrastive pretraining is needed, whether data augmentation is applied during supervised fine-tuning, and whether the retriever supports multilingual retrieval. The notation (x) after a dataset name indicates the average value across x subsets within the dataset. Detailed results for each subset are provided in the Appendix A.5. We highlight the highest score for each dataset in bold and the highest score within each parameter level with an underscore. The notation (768d) indicates that we use the first 768 dimensions of representations from DRAMA$_{1B}$, as our model is trained with MRL.

As discussed by Lin et al., 2023, there is often a trade-off between in-domain retrieval performance and generalization capability. DRAMA achieves very high in-domain multilingual effectiveness: for example, DRAMA$_{0.3B}$ is 5.5 points higher than ArcticEmb-v2-L on MIRACL (which has training data included in D_{sft}). However, it also maintains robust generalization performance in multilingual settings such as MTEB-FR. On MTEB-ZH, DRAMA$_{0.3B}$ performs slightly lower than ArcticEmb-v2, but the difference is within 1 point. Overall, these results suggest DRAMA is generalizable across retrieval tasks and languages.

Method	Param.	L-CPT.	L-FT.	Max Len	MLDR Avg
BM25	-	×	×	∞	53.6
mE5-large	303M	×	×	512	34.2
M3-BGE-Dense	303M	✓	×	8192	45.0
ArcticEmb-v2-M	113M	×	×	8192	34.0
DRAMA$_{0.1B}$	113M	×	×	8192	47.1
DRAMA$_{0.3B}$	265M	×	×	8192	48.8
DRAMA$_{1B}$	1B	×	×	128k	**54.8**
M3-BGE-Dense	303M	✓	✓	8192	52.5
mGTE-Dense	113M	✓	✓	8192	56.6
DRAMA$_{0.1B}$-MLDR	113M	×	✓	8192	60.2
DRAMA$_{0.3B}$-MLDR	265M	×	✓	8192	58.9
DRAMA$_{1B}$-MLDR	1B	×	✓	128k	**62.3**

Table 2: Effectiveness of DRAMA on the multilingual long-context retrieval task. L-CPT: Model has seen long-context data during contrastive pretraining. L-FT: Model has seen long-context data during supervised fine-tuning. Max Len: Maximum input length supported.

5.2 Pruned LLM as Retriever Backbone

Pruning a state-of-the-art LLM to create smaller retriever backbones offers two key advantages. First, it helps preserve multilingual capability. Most existing retrievers at the 0.1B parameter scale use `bert-base-uncased` as their backbone. While these models achieve strong performance in English retrieval, they do not support multilingual retrieval. By pruning an LLM instead, we achieve strong English retrieval effectiveness while retaining its multilinguality with only a small amount of multilingual web data (less than 10B tokens). Second, as recent LLMs are designed to handle long contexts, pruning an LLM as the retriever backbone allows better long-context retrieval capabilities. Table 2 shows that even though DRAMA's fine-tuning data does not include MLDR training data, and DRAMA is not trained with text beyond 256 tokens, it still performs well in length extrapolation. For example, DRAMA$_{0.1B}$ achieves an nDCG@10 of 46.8 on MLDR, despite never being trained on long-context retrieval data. Comparing DRAMA$_{0.1B}$ to M3-BGE-Dense, which was trained with long-context data during contrastive pretraining but not fine-tuned on MLDR, DRAMA

Model	Size	C-Max	LM-Max	512	4096	8192	32768
E5-v2-base	0.1B	512	512	88.0	13.8	7.3	6.2
E5-v2-base-RoPE	0.1B	512	512	92.6	11.3	6.3	6.3
DRAMA$_{0.1B}$	0.1B	256	8192	93.8	48.5	32.9	8.2
BGE-M3-Dense	0.3B	8192	8192	84.4	49.9	26.2	12.6
DRAMA$_{0.3B}$	0.3B	256	8192	92.3	48.0	25.6	14.7
DRAMA$_{1B}$	1B	256	128k	**95.2**	**77.8**	**53.8**	**19.0**
MistralE5	7B	512	4096	93.4	68.7	30.4	6.6

Table 3: Effectiveness of DRAMA on synthetic long-context retrieval task NeedleRetrieval (nDCG@10) across different max input lengths. C-Max denotes the maximum input length during contrastive learning, and LM-Max denotes the maximum input length during language modeling. Full results of NeedleRetrieval and PasskeyRetrieval are provided in Appendix A.5.

outperforms it by 2.1 points. This demonstrates the advantage of using a pruned LLM, which inherently supports longer contexts.

It is also important to note that BM25, a traditional lexical retrieval method, performs well in long-context retrieval. However, after further fine-tuning DRAMA on MLDR training data, it surpasses BM25 and other methods that have MLDR in training data. This result shows the potential of further adapting DRAMA to long-context multilingual retrieval tasks.

We also extend our evaluation to synthetic long-context retrieval tasks to further analyze effectiveness across different input lengths. Table 3 presents results on NeedleRetrieval (PasskeyRetrieval follows a similar trend and is included in Appendix A.5). When comparing our pruned model to existing E5-v2-base and its variants that apply RoPE positional embeddings (Zhu et al., 2024), DRAMA$_{0.1B}$ (with the same parameter size) demonstrates a clear advantage across all input lengths. Compared to BGE-M3-Dense, which is trained with contrastive learning up to a maximum length of 8192, DRAMA0.3B outperforms it at input lengths of 512 and 32768. At 4096 and 8192 tokens, DRAMA0.3B slightly lags behind, suggesting that training DRAMA with even longer context lengths could further enhance its performance. DRAMA1B achieves the highest scores across all input lengths. We attribute this to its longer maximum input length inherited from its original language modeling, showing the potential of increasing context length during the pruning stage.

6 Analysis and Ablation Study

6.1 Effectiveness of Data Augmentation

Figure 2 illustrates the effectiveness of different data augmentation combinations. We observe that directly fine-tuning the model without data augmentation results in poor generalization performance. Incorporating any form of LLM-based data augmentation significantly improves BEIR performance, with one exception: directly prompting Llama3.3$_{70B}$-Instruct to generate synthetic triplets (queries, positive documents, and negative documents) does not yield meaningful improvements. This suggests that training a smaller retriever model benefits more from using real-world documents.

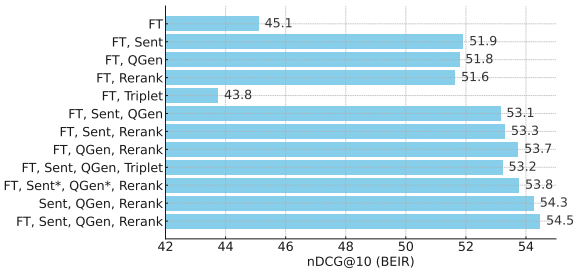

Figure 2: Effectiveness of different data augmentation combinations FT: SFT data, Sent: augmentation data based on cropped sentence query, QGen: augmentation data based on LLM generated query, Rerank: augmentation data based on LLM rerank. Triplet: augmentation data based on LLM triplet generation.). The model is trained based on 0.1B backbone, using only the English data augmentation and with 1 hard negative per query.

Moreover, combining multiple types of data augmentation further enhances effectiveness beyond using any single augmentation method alone. The highest performance is achieved when all three types of data augmentation are combined. Notably, when all augmentation strategies are applied together, the importance of fine-tuning data is diminishing, showing the effectiveness of our data augmentation approach. The data point noted by [FT, Sent*, QGen*, Rerank] shows the performance of using LLM$_{\text{Ret}}$ without further improvement from LLM listwise rerank augmentation. Its lower effectiveness compared to the final combination underscores that incorporating LLM-based rerank augmentation enhances the performance of LLM$_{\text{Ret}}$ and further improving the effectiveness of the smaller retriever model.

In practice, the computational cost estimation of creating Sent, QGen and Rerank data augmentation is in the order of 0.1k, 1k and 10k GPU hours respectively, where cost is dominated by the expensive LLM ranking for full data augmentation. On the other hand, we show that having Sent and QGen achieves a reasonable performance compared to using all 3 augmentations (53.1 vs. 54.5 on BEIR as shown in Figure 2). As a result, the computational cost of synthetic data generation can be lowered

Backbone	Param.	BEIR
BERT	0.1B	53.50
ModernBERT	0.1B	54.22
Llama3.2$_{1B \rightarrow 0.1B}$	0.1B	54.47
XLM-RoBERTa-Large	0.3B	54.74
Llama3.2$_{1B \rightarrow 0.3B}$	0.3B	56.14

Table 4: Effectiveness of using pruned Llama3.2 as smaller retriever backbone compares to pre-LLM-era or recent encoder-only backbone. The models are trained using only the English data augmentation and with 1 hard negative per query.

Model Size	Attention	Pooling	BEIR
0.1B	Bi-direction	Mean	54.47
	Bi-direction	EOS	54.37
	Uni-direction	Mean	53.88
	Uni-direction	EOS	53.58
0.3B	Bi-direction	Mean	56.14
	Bi-direction	EOS	55.85
	Uni-direction	Mean	55.18
	Uni-direction	EOS	54.79

Table 5: Impact of different attention and pooling mechanisms for the smaller retriever. The model is trained using only the English data augmentation and with 1 hard negative per query.

by an order of magnitude by omitting the LLM Ranking data, at the expense of losing around 2-3% of retrieval effectiveness.

6.2 Effectiveness of Model Backbone

In Table 4, we compare the effectiveness of using a pruned Llama model as the retriever backbone against small encoder models. At the 0.1B scale, the pruned model outperforms BERT by approximately 1 point on average. Similarly, at the 0.3B scale, the pruned model surpasses XLM-RoBERTa-Large by about 1.5 points. This demonstrates the effectiveness of using pruned-decoder-only LLM as a retriever backbone for text encoding tasks. Additionally, the 0.1B pruned model performs slightly better than ModernBERT, a recently developed encoder-only model. However, unlike ModernBERT, our approach retains multilingual support and leverages existing LLM pretraining, dropping the need to train the backbone from scratch.

6.3 Attention and Pooling Mechanism

In Table 5, we analyze how the attention mechanism and pooling strategy affect retrieval performance when training the pruned model as a text encoder. It shows that bi-directional attention outperforms uni-directional attention. While mean pooling yields higher scores than last-token pooling, the impact of the attention mechanism is greater than that of the pooling strategy. Even with massive augmented training data, uni-directional attention remains a limiting factor. However, simply enabling bi-directional attention allows the small decoder-only model to function more effectively.

Backbone	MIRACL-de	MIRACL-yo	MTEB-pl
1B $\rightarrow$ 0.1B	45.48	68.77	32.38
1B $\rightarrow$ 0.3B	55.83	83.85	36.85
1B	58.20	76.20	51.08

Table 6: Cross-lingual generalization performance of models trained with English data augmentation, evaluated on zero-shot languages. DE and YO are seen during the pruning stage, while PL is unseen. For MTEB-pl, results are averaged over 11 retrieval tasks.

6.4 Cross-lingual Generalization

In Table 6, we analyze how our model generalizes to zero-shot languages. The models are trained using English data augmentation and evaluated on languages that were not explicitly included in the fine-tuning stage. First, we examine German (de), a higher-resource language. The results show a clear trend where zero-shot effectiveness improves as the model size increases, suggesting that scaling up enhances cross-lingual generalization. For Yoruba (yo), an interesting pattern emerges: the 0.3B pruned model outperforms the larger 1B model. This may be due to the fact that the 1B model was not well-trained in Yoruba. The pruning stage of our approach includes yo data, leading to stronger performance in this language. In contrast, Polish (pl), which was not covered in either the fine-tuning or pruning stages, shows a noticeable performance gap compared to the 1B model. This shows the importance of including a language during pruning, as exposure at this stage significantly benefits zero-shot retrieval effectiveness.

7 Conclusion

We introduce DRAMA, a training framework that leverages large language models to train smaller, generalizable dense retrievers by cohesively integrating LLM pruning and diverse LLM data augmentation. DRAMA achieves strong performance across English and multilingual retrieval tasks, enabling the training of smaller retrievers to improve together with advancements in LLMs.

Limitations

While DRAMA achieves strong retrieval effectiveness across English and multilingual tasks, several areas remain open for further investigation.

Firstly, the scope of language support. As observed in Section 6.4, including a language during the pruning stage is crucial for enabling the smaller model to generalize well to that language. While the 0.1B and 0.3B variants of DRAMA covers 20 languages, expanding this coverage could improve performance for low-resource languages that lack sufficient contrastive learning data. A more comprehensive pruning strategy, incorporating additional languages, would likely enhance zero-shot multilingual retrieval.

Another limitation lies in the amount of supervised fine-tuning data. To maintain a fair evaluation of generalization, we followed the E5 fine-tuning setup, which does not include domain-specific retrieval tasks such as financial and medical. However, incorporating a broader range of supervised datasets could further improve retrieval performance across diverse domains.

Additionally, DRAMA is trained with up to 256 context length. Although it demonstrates strong extrapolation potential in long-context retrieval, it is worth more exploration on how to better integrate the long-context training data into the data augmentation efficiently. One possible approach is to organize training batches based on context length (Chen et al., 2024).

Besides, DRAMA follows a single-stage training approach, where the model is directly fine-tuned from a pruned LLM. While this simplifies the pipeline and produces strong generalization, it remains an open question whether combining with multi-stage pertaining (Yu et al., 2024) or recently proposed multi-stage distillation (Zhang et al., 2025) will help further improve the effectiveness of DRAMA.

Finally, DRAMA focused on retrieval tasks. Many recent models additionally optimize for broader text embedding tasks such as clustering and classification as well as instruction following or reasoning-intensive retrieval (Lee et al., 2024; Wang et al., 2024c; Su et al., 2023; Asai et al., 2023; Weller et al., 2025; Shao et al., 2025). We leave further integrate supervised fine-tuning data and LLM data augmentation for these tasks into DRAMA training framework as future work.

Ethics Statement

This work complies with the ACL Ethics Policy. We declare that there are no ethical issues in this paper, to the best of our knowledge.

Acknowledgments

We sincerely thank Sheng-Chieh Lin, Rulin Shao, John X. Morris, Luyu Gao, and Minghan Li for the insightful discussions throughout this work. We also extend our appreciation to the anonymous reviewers for their invaluable suggestions.

References

Akari Asai, Timo Schick, Patrick Lewis, Xilun Chen, Gautier Izacard, Sebastian Riedel, Hannaneh Hajishirzi, and Wen-tau Yih. 2023. Task-aware retrieval with instructions. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3650–3675, Toronto, Canada. Association for Computational Linguistics.

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. 2018. MS MARCO: A human generated machine reading comprehension dataset. *arXiv:1611.09268*.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. LLM2Vec: Large language models are secretly powerful text encoders. In *First Conference on Language Modeling*.

Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. 2022. InPars: Data augmentation for information retrieval using large language models. *arXiv:2202.05144*.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Xilun Chen, Kushal Lakhotia, Barlas Oguz, Anchit Gupta, Patrick Lewis, Stan Peshterliev, Yashar Mehdad, Sonal Gupta, and Wen-tau Yih. 2022. Salient phrase aware dense retrieval: Can a dense retriever imitate a sparse one? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 250–262, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Zhuyun Dai, Vincent Y. Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B. Hall, and Ming-Wei Chang. 2023. Promptagator: Few-shot dense retrieval from 8 examples. In *ICLR*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Luyu Gao and Jamie Callan. 2021. Condenser: a pre-training architecture for dense retrieval. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 981–993, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat

Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The Llama 3 herd of models. *arXiv:2407.21783*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *arXiv:2112.09118*.

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.

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.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025. NV-Embed: Improved techniques for training LLMs as generalist embedding models. In *The Thirteenth International Conference on Learning Representations*.

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*.

Chaofan Li, Minghao Qin, Shitao Xiao, Jianlyu Chen, Kun Luo, Defu Lian, Yingxia Shao, and Zheng Liu. 2025. Making text embedders few-shot learners. In *The Thirteenth International Conference on Learning Representations*.

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

Sheng-Chieh Lin, Akari Asai, Minghan Li, Barlas Oguz, Jimmy Lin, Yashar Mehdad, Wen-tau Yih, and Xilun Chen. 2023. How to train your dragon: Diverse augmentation towards generalizable dense retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6385–6400, Singapore. Association for Computational Linguistics.

Christos Louizos, Max Welling, and Diederik P. Kingma. 2018. Learning sparse neural networks through l_0 regularization. *arXiv:1712.01312*.

Kun Luo, Minghao Qin, Zheng Liu, Shitao Xiao, Jun Zhao, and Kang Liu. 2024. Large language models as foundations for next-gen dense retrieval: A comprehensive empirical assessment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1354–1365, Miami, Florida, USA. Association for Computational Linguistics.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2024. Fine-tuning llama for multi-stage text retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2421–2425, New York, NY, USA. Association for Computing Machinery.

Xueguang Ma, Xinyu Zhang, Ronak Pradeep, and Jimmy Lin. 2023. Zero-shot listwise document reranking with a large language model. *arXiv:2305.02156*.

John Xavier Morris and Alexander M Rush. 2025. Contextual document embeddings. In *The Thirteenth International Conference on Learning Representations*.

Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2024. Generative representational instruction tuning. *arXiv:2402.09906*.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. MTEB: Massive text embedding benchmark. *arXiv:2210.07316*.

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

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Rulin Shao, Rui Qiao, Varsha Kishore, Niklas Muennighoff, Xi Victoria Lin, Daniela Rus, Bryan Kian Hsiang Low, Sewon Min, Wen tau Yih, Pang Wei Koh, and Luke Zettlemoyer. 2025. ReasonIR: Training retrievers for reasoning tasks. *arXiv:2504.20595*.

Jacob Mitchell Springer, Suhas Kotha, Daniel Fried, Graham Neubig, and Aditi Raghunathan. 2024. Repetition improves language model embeddings. *arXiv:2402.15449*.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2023. One embedder, any task: Instruction-finetuned text embeddings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1102–1121, Toronto, Canada. Association for Computational Linguistics.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14918–14937, Singapore. Association for Computational Linguistics.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding. *arXiv:1807.03748*.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2023. SimLM: Pre-training with representation bottleneck for dense passage retrieval. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2244–2258, Toronto, Canada. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2024a. Text embeddings by weakly-supervised contrastive pre-training. *arXiv:2212.03533*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024b. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916, Bangkok, Thailand. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024c. Multilingual E5 text embeddings: A technical report. *arXiv:2402.05672*.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv:2412.13663*.

Orion Weller, Benjamin Chang, Sean MacAvaney, Kyle Lo, Arman Cohan, Benjamin Van Durme, Dawn

Lawrie, and Luca Soldaini. 2025. FollowIR: Evaluating and teaching information retrieval models to follow instructions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11926–11942, Albuquerque, New Mexico. Association for Computational Linguistics.

Mengzhou Xia, Tianyu Gao, Zhiyuan Zeng, and Danqi Chen. 2024. Sheared LLaMA: Accelerating language model pre-training via structured pruning. In *The Twelfth International Conference on Learning Representations*.

Shitao Xiao, Zheng Liu, Yingxia Shao, and Zhao Cao. 2022. RetroMAE: Pre-training retrieval-oriented language models via masked auto-encoder. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 538–548, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-Pack: Packed resources for general chinese embeddings. *arXiv:2309.07597*.

Puxuan Yu, Luke Merrick, Gaurav Nuti, and Daniel Campos. 2024. Arctic-Embed 2.0: Multilingual retrieval without compromise. *arXiv:2412.04506*.

Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. 2025. Jasper and Stella: distillation of SOTA embedding models. *arXiv:2412.19048*.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval. *arXiv:2407.19669*.

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

Dawei Zhu, Liang Wang, Nan Yang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. 2024. LongEmbed: Extending embedding models for long context retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 802–816, Miami, Florida, USA. Association for Computational Linguistics.

A Appendix

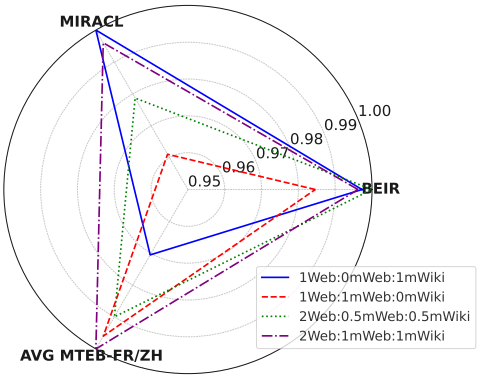

Figure 3: Effectiveness of different mixture ratios of English and multi-lingual data augmentation ratio for the data source of Web, mWeb and mWiki. The model is trained based on 0.1B backbone with 1 hard negative per query.

A.1 Multilingual Data Balance

Figure 3 illustrates how different mixtures of data sources affect effectiveness across English retrieval, in-domain multilingual retrieval, and multilingual generalization. We observe that excluding mWeb negatively impacts multilingual generalization, likely due to overfitting on the Wikipedia corpus. Conversely, excluding mWiki leads to a drop in in-domain multilingual retrieval effectiveness. However, mixing both mWiki and mWeb enables strong performance across both in-domain effectiveness and multilingual generalization. Additionally, we find that maintaining a 1:1 balance between English and multilingual data yields better overall performance than doubling the proportion of English data. While increasing the English proportion slightly improves BEIR effectiveness, it significantly weakens multilingual retrieval performance. Overall, using a 1:1 ratio of English to multilingual data and incorporating augmentation data from both Wikipedia and web-crawled multilingual sources achieves the best trade-off, covering the largest area in the radar chart and ensuring robust performance across retrieval tasks.

A.2 Matryoshka Representation Learning

In Figure 4, we compare the effectiveness of DRAMA variants across different dense representation dimensionalities. For dimensions larger than 256, the trend of model size scaling is clear—larger model achieves higher effectiveness. Additionally, text representations largely retain their effective-

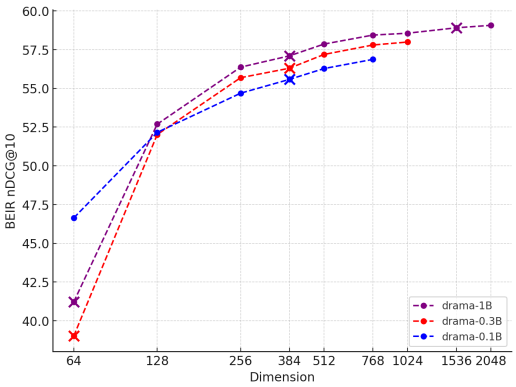

Figure 4: Effectiveness of DRAMA across different text representation dimensions. Points marked with × indicate dimensionalities that were not explicitly optimized in the MRL process.

ness compared to using the full-dimensionality representation. However, at 128 dimensions, the scaling trend is not guaranteed. At 64 dimensions, the 0.1B model outperforms both the 0.3B and 1B models, likely because 64 dimensions were not a target setting during MRL training for the larger models. In contrast, for dimensions 384 and 1536, despite also not being target dimensions for MRL, the effectiveness is well preserved. This observation raises the importance of considering the range of target dimensionalities during MRL training to ensure effectiveness at test time.

A.3 Pruned Model Configuration

Llama3.2-1B has a hidden size of 2048, an intermediate size of 8192, 32 attention heads, and 16 hidden layers. **DRAMA$_{0.3B}$** reduces these dimensions to a hidden size of 1024, intermediate size of 4096, 16 attention heads, while maintaining 16 hidden layers. **DRAMA$_{0.1B}$** further scales down to a hidden size of 768, intermediate size of 3072, 12 attention heads, and 12 hidden layers.

A.4 Detailed Training Setup

Model License: Our LLM retriever is trained based on Llama3.1-8B follows Llama 3.1 Community License Agreement. Data augmentation based on Inst-LLM is based on Llama3.3-70B-Instruct follows Llama 3.3 Community License Agreement. Our backbone model is pruned based on Llama3.2-1B, following Llama 3.2 Community License Agreement.

Languages: For pruning and data augmentation, our web crawl text corpora cover the following 20 languages: English, Arabic, Bengali, Spanish, Persian, Finnish, French, Hindi, Indonesian, Japanese, Korean, Russian, Swahili, Telugu, Thai, Chinese, German, Yoruba, Italian, Portuguese.

Training: The models are trained using the `dpr-scale`[2] codebase on 32 A100 GPUs over approximately two days. The training configurations for different model sizes are as follows:

DRAMA$_{0.1B}$: Batch size of 2048, with each query paired with seven hard negatives. DRAMA$_{0.3B}$: Batch size of 1024, with each query paired with seven hard negatives. DRAMA$_{1B}$: Batch size of 256, with each query paired with three hard negatives. All three variants are trained for 200,000 steps.

A.5 Detailed Evaluation Results

We use the `tevatron`[3] codebase to evaluate BEIR and MIRACL. For retrieval tasks in MTEB-FR/ZH/DE, we utilize the `mteb` codebase. For BEIR and MIRACL, we set the maximum context length as 512 for both query and document following previous works. For baselines, we adopt BEIR and MIRACL scores directly from the original works. In MLDR, we reference baseline results from the mGTE work for mE5, BGE-M3, and mGTE. For Arctic-Embedding, we conduct the MLDR evaluation ourselves. While some MTEB scores are reported in previous works, we observe version changes in certain datasets within MTEB-FR. To ensure consistency, we re-evaluate MTEB-FR/ZH/DE baselines ourselves. We set the maximum context length as 1024 following (Zhang et al., 2024).

The full evaluation results are presented in Table 7, Table 8, Table 9, Table 10, Table 11, Table 12, Table 13, and Table 14.

[2] https://github.com/facebookresearch/dpr-scale
[3] https://github.com/texttron/tevatron

Method	Param.	CPT	Multi.	Avg	BEIR (nDCG@10) TREC-COVID	NF Corpus	Sci Fact	SCI DOCS	FiQA	Argu Ana	Touche-2020	DB Pedia	Climate-FEVER	FEVER	NQ	Hotpot QA	Quora
BM25	-	×	✓	43.7	59.5	32.2	67.9	14.9	23.6	39.7	44.2	31.8	16.5	65.1	30.5	63.3	78.9
Contriever	86M	✓	×	47.5	59.6	32.8	67.7	16.5	32.9	44.6	23.0	41.3	23.7	75.8	49.8	63.8	86.5
DRAGON	86M	×	×	50.2	75.9	33.9	67.9	15.9	35.6	46.9	26.3	41.7	22.7	78.1	53.7	66.2	87.5
E5-v2-base	86M	✓	×	51.9	69.6	35.4	71.9	18.7	39.9	44.5	26.4	42.2	26.6	85.0	58.2	69.2	86.6
bge-base-en-v1.5	86M	✓	×	55.0	78.1	37.4	74.0	21.7	40.6	63.6	25.7	40.8	31.2	86.3	54.1	72.6	88.9
mE5-base	86M	✓	✓	50.2	69.7	32.5	69.3	17.2	38.2	44.2	21.4	40.4	23.9	79.4	60.0	68.6	87.6
mGTE-Dense	113M	✓	✓	54.3	57.4	36.7	73.4	18.3	63.0	58.4	22.8	40.1	34.8	92.1	58.1	63.0	88.0
ArcticEmb-v2-M	113M	✓	✓	56.9	80.3	35.9	71.8	20.3	44.0	58.0	29.8	43.9	38.3	91.6	64.6	72.4	88.7
DRAMA$_{0.1B}$	113M	×	✓	56.9	83.3	36.9	75.7	19.1	44.2	54.8	29.1	44.8	38.0	89.4	60.8	74.9	88.3
E5-large-v2	303M	✓	×	52.1	66.5	37.1	72.2	20.5	41.1	46.4	20.7	44.0	22.2	82.8	63.4	73.1	86.8
bge-large-en-v1.5	303M	✓	×	56.1	74.8	38.1	74.6	22.6	45.0	63.5	24.8	44.1	36.6	87.2	55.0	74.1	89.1
mE5-large	303M	✓	✓	52.9	71.3	34.0	70.4	17.5	43.8	54.4	23.4	41.3	25.7	82.8	64.1	71.2	88.2
mE5-Inst	303M	✓	✓	54.1	82.0	35.5	71.9	18.7	47.7	58.4	27.2	38.4	29.9	78.0	57.8	69.3	89.1
M3-BGE-Dense	303M	✓	✓	50.0	55.6	31.4	64.4	16.4	41.3	54.0	22.6	39.8	24.2	81.4	60.6	69.4	88.6
ArcticEmb-v2-L	303M	✓	✓	57.2	83.9	35.3	70.6	20.2	45.5	59.2	29.5	43.4	43.5	91.9	63.7	68.2	89.0
DRAMA$_{0.3B}$	265M	×	✓	58.0	83.8	37.9	76.1	19.7	46.9	54.1	28.1	47.7	41.9	89.5	64.1	75.6	88.4
Gecko	1B	✓	✓	58.0	82.6	40.3	75.4	20.4	59.2	62.2	25.9	47.1	33.2	87.0	61.3	71.3	88.2
DRAMA$_{1B}$	1B	×	✓	59.1	85.8	37.6	77.9	20.7	50.6	53.5	29.6	50.0	38.7	89.9	67.3	77.4	88.7
DRAMA$_{1B}$(768d)	1B	×	✓	58.4	85.2	37.1	77.5	20.7	50.2	53.1	29.0	49.2	37.9	89.5	66.5	75.5	88.5
MistralE5	7B	×	✓	59.0	87.2	38.6	76.4	16.3	56.6	61.9	26.4	48.9	38.4	87.8	63.5	75.7	89.6

Table 7: Full BEIR evaluation of DRAMA.

Method	Param.	Avg	MIRACL (nDCG@10) ar	bn	en	es	fa	fi	fr	hi	id	ja	ko	ru	sw	te	th	zh	de	yo
BM25	-	38.5	48.1	50.8	35.1	31.9	33.3	55.1	18.3	45.8	44.9	36.9	41.9	33.4	38.3	49.4	48.4	18.0	22.6	40.6
mE5-base	86M	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
mGTE-Dense	113M	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
ArcticEmb-v2-M	113M	59.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
DRAMA$_{0.1B}$	113M	70.4	80.5	74.5	56.3	61.4	62.8	78.9	62.2	61.9	58.0	74.2	70.5	72.3	77.1	81.5	80.4	64.8	62.3	88.5
mE5-large	303M	60.1	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
mE5-Inst	303M	66.0	76.8	73.9	51.5	53.7	59.4	77.3	53.7	60.3	52.1	69.0	65.3	67.9	72.5	83.4	78.6	56.2	55.5	81.5
M3-BGE-Dense	303M	69.2	78.4	80.0	56.9	56.1	60.9	78.6	58.3	59.5	56.1	72.8	69.9	70.1	78.7	86.2	82.6	62.7	56.7	81.8
ArcticEmb-v2-L	303M	64.9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
DRAMA$_{0.3B}$	265M	71.4	81.4	77.2	58.5	62.4	63.7	79.9	62.4	64.8	58.3	75.6	70.0	73.6	78.1	81.8	81.4	65.1	63.4	87.2
Gecko	1B	56.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
DRAMA$_{1B}$	1B	71.7	81.1	76.6	58.4	62.2	64.5	80.9	62.8	65.7	58.7	76.4	69.3	74.6	77.6	80.6	81.8	68.2	63.9	88.1
MistralE5	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

Table 8: Full MIRACL evaluation of DRAMA.

Method	Param.	L-CPT.	L-FT.	Max Len	Avg	MLDR (nDCG@10) ar	de	en	es	fr	hi	it	ja	ko	pt	ru	th	zh
BM25	-	×	×	∞	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	303M	×	×	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
M3-BGE-Dense	303M	✓	×	512	45.0	37.9	43.3	41.2	67.7	64.6	32.0	55.8	43.4	33.1	67.8	52.8	27.2	18.2
ArcticEmb-v2-M	113M	×	×	8192	34.0	15.9	35.4	32.4	67.0	63.9	15.2	56.8	10.0	17.7	66.1	42.9	11.2	7.4
DRAMA$_{0.1B}$	113M	×	×	8192	47.1	39.6	46.7	40.6	73.4	72.9	27.0	57.9	44.5	36.2	69.5	55.3	30.0	19.0
DRAMA$_{0.3B}$	265M	×	×	8192	48.8	42.9	49.8	44.1	75.1	73.2	30.1	62.3	42.4	34.4	71.2	58.4	32.7	17.7
DRAMA$_{1B}$	1B	×	×	128k	54.8	49.0	53.2	51.1	79.2	76.9	36.7	68.4	53.3	43.3	77.5	63.0	37.1	24.0
M3-BGE-Dense	303M	✓	✓	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
mGTE-Dense	113M	✓	✓	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
DRAMA$_{0.1B}$-MLDR	113M	×	✓	8192	60.2	60.6	55.3	56.6	84.0	81.3	43.6	72.2	55.9	48.7	82.3	73.8	38.8	29.1
DRAMA$_{0.3B}$-MLDR	265M	×	✓	8192	58.9	58.2	53.1	57.0	83.1	81.0	39.9	71.0	54.9	47.5	80.8	71.8	39.2	28.7
DRAMA$_{1B}$-MLDR	1B	×	✓	128k	62.3	59.9	58.2	62.1	84.6	81.6	49.2	77.6	57.9	52.7	84.3	70.8	43.7	32.9

Table 9: Full MLDR evaluation of DRAMA.

Method	Param.	Avg	MTEB-FR-Retrieval (nDCG@10) AlloprofRetrieval	BSARDRetrieval	MintakaRetrieval	SyntecRetrieval	XPQARetrieval
mE5-base	86M	45.4	34.4	18.8	31.0	82.9	59.6
mGTE-Dense	113M	50.6	49.4	19.1	34.7	82.6	67.4
ArcticEmb-v2-M	113M	53.7	54.6	18.4	31.4	89.8	74.4
DRAMA$_{0.1B}$	113M	52.1	51.9	24.7	26.7	85.5	71.5
mE5-large	303M	47.7	39.3	21.4	34.2	82.4	61.3
mE5-Inst	303M	49.9	51.4	24.3	30.3	86.2	57.4
M3-BGE-Dense	303M	48.6	48.3	16.6	22.9	84.5	70.9
ArcticEmb-v2-L	303M	54.5	53.9	21.9	30.7	88.5	77.3
DRAMA$_{0.3B}$	265M	54.8	55.8	26.6	28.8	89.9	72.8
DRAMA$_{1B}$	1B	57.6	55.9	29.9	37.5	91.6	72.9

Table 10: Full MTEB-FR-Retrieval evaluation of DRAMA.

Method	Param.	MTEB-ZH-Retrieval (nDCG@10) Avg	Cmedqa	Covid	Du	Ecom	Medical	MMarco	T2	Video
mE5-base	86M	61.6	27.2	73.5	81.7	54.2	48.4	76.0	70.8	61.3
mGTE-Dense	113M	72.0	43.8	81.0	87.5	64.8	61.9	79.4	84.7	72.8
ArcticEmb-v2-M	113M	55.7	19.7	72.2	68.4	48.6	38.3	71.2	71.3	56.1
DRAMA$_{0.1B}$	113M	61.7	21.2	78.4	74.9	57.9	42.4	76.2	76.4	66.0
mE5-large	303M	63.7	28.7	75.6	85.3	54.7	51.5	79.2	76.1	58.2
mE5-Inst	303M	64.2	33.9	76.1	85.2	53.7	56.2	78.6	82.9	47.2
M3-BGE-Dense	303M	65.6	33.8	78.3	84.0	58.5	54.2	77.3	81.5	57.0
ArcticEmb-v2-L	303M	63.6	27.8	78.8	78.4	56.4	51.1	78.4	79.7	58.6
DRAMA$_{0.3B}$	265M	63.0	21.2	78.4	74.9	57.9	42.4	76.2	76.4	66.0
DRAMA$_{1B}$	1B	63.7	23.6	76.1	77.8	60.1	45.8	79.4	79.0	67.8

Table 11: Full MTEB-ZH-Retrieval evaluation of DRAMA.

Method	Param.	MTEB-DE-Retrieval (nDCG@10) Avg	GerDaLIR	GermanDPR	GermanQuAD-Retrieval	XMarket
mE5-base	86M	49.2	6.9	79.6	93.9	16.3
mGTE-Dense	113M	49.1	9.4	80.0	91.1	16.0
ArcticEmb-v2-M	113M	55.0	16.1	81.8	94.4	27.6
DRAMA$_{0.1B}$	113M	55.1	15.4	82.8	95.9	26.2
mE5-large	303M	50.4	6.5	82.9	94.6	17.5
mE5-Inst	303M	52.5	10.7	79.4	94.5	25.3
M3-BGE-Dense	303M	50.4	10.9	82.5	95.1	13.1
ArcticEmb-v2-L	303M	55.9	17.5	83.7	95.2	27.0
DRAMA$_{0.3B}$	265M	55.6	15.7	82.6	96.4	27.7
DRAMA$_{1B}$	1B	56.2	15.3	84.4	97.1	28.0

Table 12: Full MTEB-DE-Retrieval evaluation of DRAMA.

Model	Size	C-Max	LM-Max	256	512	1024	2048	4096	8192	16384	32768
E5-v2-base	0.1B	512	512	91.8	88.0	40.8	21.4	13.8	7.3	2.7	6.2
E5-v2-base-RoPE	0.1B	512	512	96.3	92.6	44.0	20.0	11.3	6.3	3.8	6.3
DRAMA$_{0.1B}$	0.1B	256	8192	95.2	93.8	92.8	84.2	48.5	32.9	24.7	8.2
BGE-M3-Dense	0.3B	8192	8192	94.2	84.4	80.0	53.6	49.9	26.2	30.4	12.6
DRAMA$_{0.3B}$	0.3B	256	8192	95.4	92.3	88.2	71.7	48.0	25.6	15.4	14.7
DRAMA$_{1B}$	1B	256	128k	96.6	95.2	88.8	88.8	77.8	53.8	24.4	19.0
MistralE5	7B	512	4096	97.4	93.4	86.9	74.8	68.7	30.4	21.3	6.6

Table 13: Full NeedleRetrieval evaluation of DRAMA.

Model	Size	C-Max	LM-Max	256	512	1024	2048	4096	8192	16384	32768
E5-v2-base	0.1B	512	512	100.0	98.3	62.0	22.0	14.0	6.0	6.5	6.4
E5-v2-base-RoPE	0.1B	512	512	100.0	100.0	62.0	22.0	14.0	6.0	4.7	6.6
DRAMA$_{0.1B}$	0.1B	256	8192	100.0	100.0	100.0	100.0	65.5	34.4	14.6	12.6
BGE-M3-Dense	0.3B	8192	8192	100.0	100.0	99.3	85.9	43.8	39.8	17.9	17.4
DRAMA$_{0.3B}$	0.3B	256	8192	100.0	100.0	100.0	98.8	51.1	30.2	17.8	7.9
DRAMA$_{1B}$	1B	512	128k	100.0	100.0	100.0	100.0	100.0	96.2	48.0	16.0
MistralE5	7B	512	4096	98.3	97.8	96.3	99.3	100.0	50.0	32.0	8.0

Table 14: Full PasskeyRetrieval evaluation of DRAMA.