BMRETRIEVER: Tuning Large Language Models as Better Biomedical Text Retrievers

Ran Xu[♡*], **Wenqi Shi**[♠*], **Yue Yu**[♠*], **Yuchen Zhuang**[♠], **Yanqiao Zhu**[♢]
May D. Wang[♠], **Joyce C. Ho**[♡], **Chao Zhang**[♠], **Carl Yang**[♡]
[♡] Emory University [♠] Georgia Tech [♢] UCLA
{ran.xu,joyce.c.ho,j.carlyang}@emory.edu, yzhu@cs.ucla.edu
{wqshi,yueyu,yczhuang,maywang,chaozhang}@gatech.edu

Abstract

Developing effective biomedical retrieval models is important for excelling at knowledge-intensive biomedical tasks but still challenging due to the lack of sufficient publicly annotated biomedical data and computational resources. We present BMRETRIEVER, a series of dense retrievers for enhancing biomedical retrieval via unsupervised pre-training on large biomedical corpora, followed by instruction fine-tuning on a combination of labeled datasets and synthetic pairs. Experiments on 5 biomedical tasks across 11 datasets verify BMRETRIEVER's efficacy on various biomedical applications. BMRETRIEVER also exhibits strong parameter efficiency, with the 410M variant outperforming baselines up to 11.7 times larger, and the 2B variant matching the performance of models with over 5B parameters. The training data and model checkpoints are released at https://huggingface.co/BMRetriever to ensure transparency, reproducibility, and application to new domains.

1 Introduction

In the field of biomedicine, the ability to effectively retrieve knowledge from external corpora is crucial for large language models (LLMs) to excel at biomedical NLP tasks (Lewis et al., 2020; Zhang et al., 2024; Xiong et al., 2024). By tapping into up-to-date domain knowledge, retrieval-augmented LLMs have demonstrated promising results in various biomedical downstream applications, including knowledge discovery (Frisoni et al., 2022), question answering (Wang et al., 2023; Yu et al., 2024), and clinical decision-making (Naik et al., 2022; Shi et al., 2023; Xu et al., 2024).

Several works have designed specialized retrieval models for biomedical domains (Mohan et al., 2017; Liu et al., 2021; Jin et al., 2023; Luo et al., 2022a; Singh et al., 2023; Zhang et al., 2023).

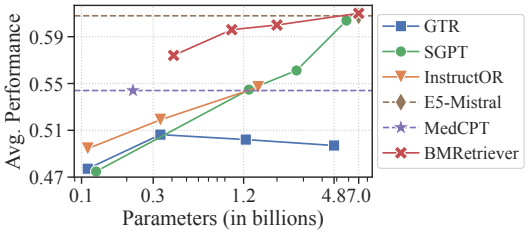


Figure 1: The average performance of BMRETRIEVER on 5 popular biomedical search tasks compared to baselines with different parameters. X-axis in log scale.

However, these models are typically built upon BERT-series models, which have limited representative power. Besides, they often rely on proprietary data (e.g., private search logs or patient records), making it challenging to scale them up to accommodate larger models effectively due to privacy concerns. While recent studies in the general domain have improved neural retrieval models via scaling up model size (Ni et al., 2022; Muennighoff, 2022; Wang et al., 2024) and training data (Izacard et al., 2022; Wang et al., 2022b; Lin et al., 2023), adapting such models to the biomedical domain may lead to suboptimal performance due to the distribution shift issue (Thakur et al., 2021). Developing large-scale retrieval models dedicated to the biomedical domain without requiring massive proprietary datasets remains crucial yet challenging.

In this work, we propose BMRETRIEVER, a series of dense text retrievers at various scales using LLMs as backbones to improve biomedical retrieval performance. Firstly, we inject biomedical knowledge into BMRETRIEVER by unsupervised contrastive pre-training on a *large-scale* unlabeled biomedical corpora, which comprises an extensive and diverse collection of data, with rich biomedical background knowledge invaluable for domain-specific understanding (Lála et al., 2023; Xiong et al., 2024). Besides, unlabeled corpora are readily accessible, overcoming the bottleneck of scarce annotated data that often plagues specialized domains. Pre-training on them allows us to adapt our

[*] Equal contribution.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 22234–22254
November 12-16, 2024

models to the biomedical domain, equipping them with necessary linguistic patterns and terminology.

To further boost the embedding quality and align the retriever with downstream applications, we conduct instruction fine-tuning with *high-quality* labeled datasets. Specifically, we gather various *public human-annotated* biomedical retrieval tasks, such as medical question-answering (QA) and dialogue pairs, and create instructions for each to improve BMRETRIEVER with task-specific understanding. Given the relatively small sample size and limited task types in public biomedical datasets, we further leverage the powerful GPT models to generate additional synthetic retrieval tasks under various scenarios with query and passage pairs to augment training samples and diversify instructions. This allows the model to acquire a comprehensive understanding of biomedical retrieval tasks and facilitates its generalization across various downstream tasks and input formats.

We conduct extensive experiments across *five* tasks on *eleven* biomedical datasets to demonstrate the strong performance of BMRETRIEVER. As shown in Figure 1, BMRETRIEVER outperforms existing dense retrievers with orders of magnitude more parameters: with 410M parameters, it surpasses the performance of GTR-4.8B (Ni et al., 2022) and SGPT-2.7B (Muennighoff, 2022), which have 7× more parameters. At the 7B scale, BMRETRIEVER outperforms the recently proposed E5-Mistral (Wang et al., 2024), which uses extra-large batch-size and nonpublic data mixture. In addition, BMRETRIEVER presents a lightweight yet high-performing domain adaptation solution, with its 1B variant achieving more than 98% performance of E5-Mistral using only 14.3% of parameters. Our contribution can be summarized as follows:

- We develop a family of BMRETRIEVER models ranging from 410M to 7B parameters, achieving efficient scaling via a two-stage framework to improve biomedical text retrieval performance.
- We assess BMRETRIEVER's efficacy with an extensive evaluation against 18 baselines on 5 tasks across 11 biomedical datasets. Results demonstrate BMRETRIEVER's parameter efficiency yet strong domain adaptation capabilities, achievable within academic computational budgets.
- BMRETRIEVER ensures transparency, reproducibility, and potential generalization to additional domain-specific adaptations by providing a detailed training recipe with public datasets and accessible model checkpoints.

Parameters	410M	1B	2B	7B
Backbone	Pythia (2023)	Pythia (2023)	Gemma (2024)	BioMistral (2024)
Model Layers	24	16	18	32
Embedding Dim.	1024	2048	2048	4096

Table 1: An overview of BMRETRIEVER.

2 Related Work

Earlier research explores various approaches for learning representations suitable for text retrieval (Deerwester et al., 1990; Huang et al., 2013). More recently, several studies introduce dual-encoder architectures based on BERT for dense retrieval (Karpukhin et al., 2020; Xiong et al., 2021; Qu et al., 2021; Izacard et al., 2022). With the advent of LLMs with billions of parameters, several studies attempt to scale up model size (Ni et al., 2022; Neelakantan et al., 2022), often fine-tuned on multi-task instruction data (Asai et al., 2023; Su et al., 2023; Wang et al., 2024; Lee et al., 2024). However, the benefit of scaling up is more pronounced for general domain datasets where massive annotated data are available.

To design effective retrievers for specialized domains, several works propose continuously pre-train the retrieval model on domain-specific corpora (Yu et al., 2022; Zhang et al., 2023) or fine-tuning the model on proprietary search datasets (Mohan et al., 2017; Jin et al., 2023). On the other hand, synthetic data has also been used to improve the generalization ability of dense retrieval model (Ma et al., 2021; Wang et al., 2022a; Jiang et al., 2023; Wang et al., 2024). Despite these advancements, how to combine public, open data to formulate a dataset curation recipe for adapting LLMs as high-performing biomedical retrievers remains unresolved. Our method efficiently integrates diverse supervision signals for biomedical retrieval model training, which achieves better performance than baselines trained with more data.

3 Method

BMRETRIEVER leverages the pre-trained autoregressive transformer as the backbone, taking advantage of the availability of various model sizes within this model family. This flexibility allows us to scale up the retrieval model. Specifically, we utilize the publicly available autoregressive transformers with 410M, 1B, 2B, and 7B parameters (Biderman et al., 2023; Team et al., 2024; Labrak et al., 2024). Our model details are illustrated in Table 1.

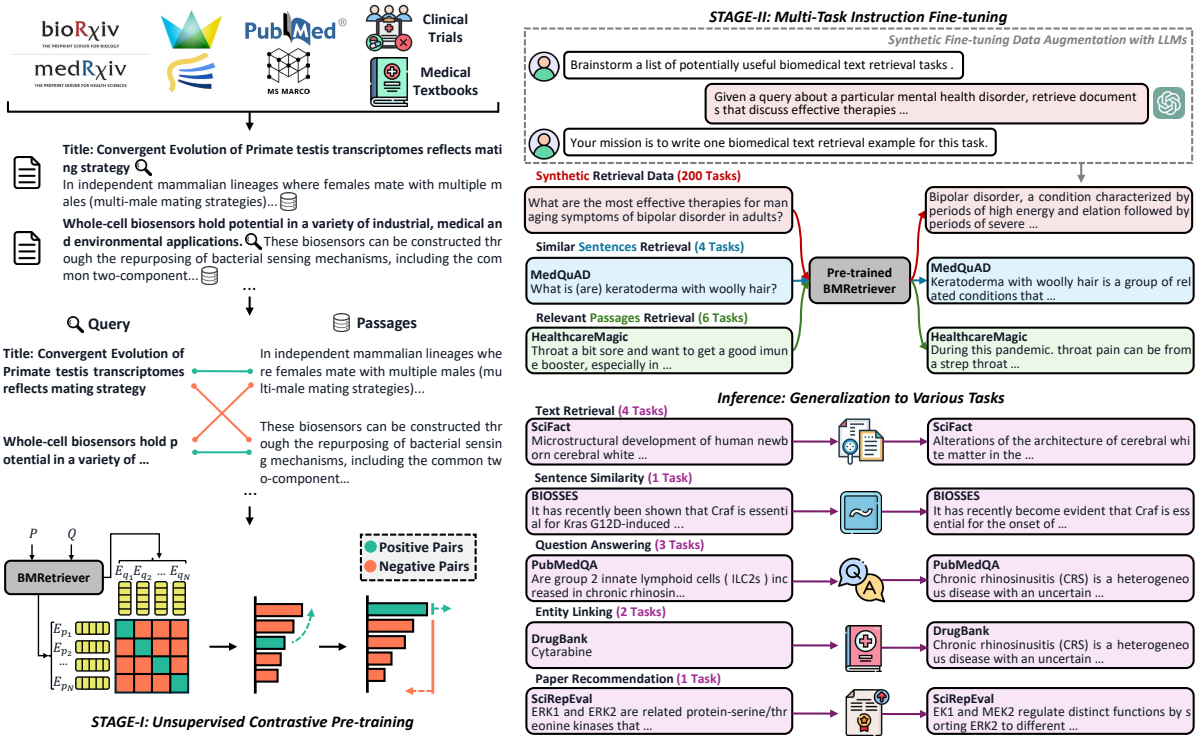


Figure 2: The overview of the two-stage pre-training framework in BMRETRIEVER. Stage I performs unsupervised contrastive pre-training on large-scale biomedical query-passage pairs, while Stage II conducts instruction fine-tuning using diverse labeled data, including synthetic examples generated by LLMs, to adapt BMRETRIEVER to various biomedical downstream tasks.

3.1 Background of Dense Text Retrieval

In dense retrieval (Lee et al., 2019; Karpukhin et al., 2020), the language model $\mathbf{E}$ is used to represent queries and passages in dense embeddings. Denote the query q and passage p with the corresponding task instruction I_q and I_p[1], the embedding is calculated as $\boldsymbol{e}_q = \mathbf{E}(I_q \oplus q)$, $\boldsymbol{e}_p = \mathbf{E}(I_p \oplus p)$. The relevance score $\text{sim}(q, p)$ is calculated with the dot product between query and passage embeddings:

$$\text{sim}(q, p) = \boldsymbol{e}_q^\top \boldsymbol{e}_p. \quad (1)$$

In this work, where autoregressive LLMs are used for $\mathbf{E}$, an <EOS> token is appended to the end of the query and passage. The embedding of the <EOS> token from the final layer of LLM is used as the representation for both queries and passages.

To effectively adapt BMRETRIEVER to the biomedical domain, a two-stage training procedure is proposed (see Figure 2): (1) an unsupervised contrastive pre-training stage (§ 3.2) using *silver* query-passage pairs from extensive biomedical corpora, and (2) a fine-tuning stage (§ 3.3) using *gold* labeled data from various tasks. The details of two stages will be introduced in the following sections.

[1]The instruction format is in Appendix B.

3.2 Unsupervised Contrastive Pre-training

Pre-training Corpus Collection. To provide BMRETRIEVER with an initial understanding of biomedical contexts, we collect a diverse range of publicly available biomedical corpora, including *biomedical publications* (Chen et al., 2021; Xiong et al., 2024; Lo et al., 2020), *medical textbooks* (Jin et al., 2021), as well as *general-domain web corpus* (Bajaj et al., 2016), as detailed in Table 8.

Contrastive Pre-training. We construct positive and negative query-passage pairs from raw unlabeled corpora to facilitate contrastive pre-training of the retrieval model. For positive pairs, we employ two strategies: (1) for corpora with titles, we treat the title as the query and the corresponding abstract as the passage; (2) for untitled corpora, we randomly sample two disjoint passages from documents, using one as the query and the other as the passage (Izacard et al., 2022). To obtain negative pairs, we sample in-batch negatives (Gillick et al., 2019) where the passages from other pairs in the same batch serve as negative examples. With the collected pairs, we employ contrastive learning to distinguish the relevant query-passage pairs from the irrelevant ones. For each mini-batch $\mathcal{B}$, we leverage the InfoNCE loss as the pre-training ob-

jective to rank the positive text pairs $\{(q_i, p_i)\}_{i=1}^{n}$ higher than in-batch negative passages $\{p_{ij}^{-}\}_{j=1}^{N}$:

$$\mathcal{L}_{\text{cpt}} = -\log \frac{e^{\text{sim}(q_i,p_i)/\tau}}{\sum_{j\in\mathcal{B}} e^{\text{sim}(q_i,p_j)/\tau}}. \quad (2)$$

Contrastive pre-training improves the quality of representations by better aligning similar text sequences while ensuring the uniformity of unrelated text sequences, which helps adapt the retrieval model to biomedical domains (Gururangan et al., 2020; Yu et al., 2022; Luo et al., 2022b).

3.3 Supervised Instruction Fine-tuning

To further enhance the model's specialized domain knowledge and align the model with downstream application tasks, we conduct instruction fine-tuning, which integrates a diverse collection of retrieval tasks into the instruction tuning blend. We present a detailed procedure below.

Instruction Fine-tuning Dataset. To incorporate the model with a wide range of biomedical downstream tasks, we leverage a series of biomedical tasks with varying granularity, including both sentence-level *medical natural language inference* (MedNLI) (Shivade, 2017), *medical question pairs* (McCreery et al., 2020), and passage-level biomedical QA tasks, including *MedQuad* (Ben Abacha et al., 2019), *StackExchange* (Team, 2021), and *medical dialogues* (Li et al., 2023b). Besides, we also include several general-domain retrieval datasets, including MS MARCO (Bajaj et al., 2016), NQ (Kwiatkowski et al., 2019), Fever (Thorne et al., 2018), ELI5 (Fan et al., 2019), and NLI (Bowman et al., 2015), to enhance the model's ability for relevance estimation. The instruction format and data conversion details are exhibited in Appendix B.

Synthetic Data Augmentation with LLMs. To supplement the limited task types and relatively small sample sizes in labeled biomedical datasets, we employ a data augmentation approach to generate synthetic query and passage pairs. Two approaches are utilized for this generation process.

We leverage GPT-3.5 (`gpt-3.5-turbo-1106`) for *instance-level* augmentation to enrich (query, passage) pairs resembling standard biomedical information retrieval (IR) formats. Given a passage from PubMed and Meadow used in contrastive pre-training, we prompt GPT-3.5 to generate a relevant query based on the passage context. This allows the model to better capture the relevance within biomedical contexts for effective retrieval.

Beyond relevance signals, task generalization is also crucial for building a general retriever, as user intent and input formats vary while public data captures only a fraction of tasks. To address this, we perform *task-level* augmentation, which involves prompting GPT-4 (`gpt-4-turbo-1106`) to conceptualize a diverse list of potential scenarios for biomedical retrieval tasks (Wang et al., 2024). Subsequently, we prompt GPT-4 again to generate examples for each scenario, including a query, a relevant (positive) passage, and a challenging irrelevant (hard negative) passage. This approach allows us to enhance the diversity of instructions.

Hard Negative Mining and Data Filter. In both labeled instruction fine-tuning datasets and data-label synthetic datasets, positive pairs are available, while negative examples are missing. To obtain the negatives, we randomly select 1 passage from the top 100 passages retrieved by `E5-base` (Wang et al., 2022b) when using the given query to search the entire corpus of the corresponding dataset. As the generated synthetic data can be noisy, consistency filtering is adopted to filter low-quality pairs (Alberti et al., 2019; Dai et al., 2023), where for each synthetic (query q, passage p) pair, we use the `E5-base` to predict the most relevant passages for q. We only retain q when p occurs among the top three retrieved passages.

Fine-tuning Objectives. After constructing positive and negative text pairings $\{(q_i, p_i^{+}, p_i^{-})\}_{i=1}^{M}$ where p_i^{+} and p_i^{-} stands for the positive passage and the hard negative, respectively, we employ the InfoNCE loss function for each minibatch $\mathcal{B}$ as

$$\mathcal{L}_{\text{ft}} = \frac{e^{\text{sim}(q_i,p_i^{+})/\tau}}{\sum_{j\in\mathcal{B}} e^{\text{sim}(q_i,p_j^{+})/\tau} + e^{\text{sim}(q_i,p_j^{-})/\tau}}, \quad (3)$$

where both in-batch negatives and hard negatives are utilized to further improve model training.

4 Experimental Results

4.1 Experimental Setups

Tasks and Datasets. We conduct experiments on eleven datasets across five biomedical retrieval-oriented tasks, including (1) IR, (2) sentence similarity (STS), (3) QA, (4) entity linking, and (5) paper recommendation. There is *no overlap* between the training and test pairs. Task and dataset details are available in Appendix B.

Baselines. We compare to sparse retrieval models *BM25* (Robertson et al., 2009) and *open-source*

Task	Scale	# PT Pairs	# FT Pairs	Standard IR				Sent. Sim.	Avg. Retr.	Avg. All
Model				NFCorpus	SciFact	SciDocs	Trec-COVID	BIOSSES		
Sparse Retrieval										
BM25 (Robertson et al., 2009)	—	—	—	0.325	0.665	0.158	0.656	—	0.451	—
Base Size (< 1B)										
Contriever (Izacard et al., 2022)	110M	1B	500K	0.328	0.677	0.165	0.596	0.833	0.442	0.520
Dragon (Lin et al., 2023)	110M	—	28.5M	0.339	0.679	0.159	0.759	0.819	0.484	0.551
SPECTER 2.0 (Singh et al., 2023)	110M	3.3M	—	0.228	0.671	—	0.584	—	—	—
SciMult (Zhang et al., 2023)	110M	5.5M	—	0.308	0.707	—	0.712	—	—	—
COCO-DR (Yu et al., 2022)	110M	15M	500K	0.355	0.709	0.160	0.789	0.829	0.503	0.567
SGPT-125M (Muennighoff, 2022)	125M	unknown	500K	0.228	0.569	0.122	0.703	0.752	0.406	0.475
MedCPT (Jin et al., 2023)	220M	—	255M	0.340	0.724	0.123	0.697	0.837	0.471	0.544
GTR-L (Ni et al., 2022)	335M	2B	662K	0.329	0.639	0.158	0.557	0.849	0.421	0.506
InstructOR-L (Su et al., 2023)	335M	—	1.24M	0.341	0.643	0.186	0.581	0.844	0.438	0.519
E5-Large-v2† (Wang et al., 2022b)	335M	270M	1M	0.371	0.726	0.201	0.665	0.836	0.491	0.560
BGE-Large*‡ (Chen et al., 2024)	335M	1.2B	1.62M	0.345	0.723	0.222	0.753	0.804	**0.511**	0.569
BMRETRIEVER-410M	410M	10M	1.4M	0.321	0.711	0.167	0.831	0.840	0.508	**0.574**
Large Size (1B - 5B)										
InstructOR-XL (Su et al., 2023)	1.5B	—	1.24M	0.360	0.646	0.174	0.713	0.842	0.473	0.547
GTR-XL (Ni et al., 2022)	1.2B	2B	662K	0.343	0.635	0.159	0.584	0.789	0.430	0.502
GTR-XXL (Ni et al., 2022)	4.8B	2B	662K	0.342	0.662	0.161	0.501	0.819	0.417	0.497
SGPT-1.3B (Muennighoff, 2022)	1.3B	unknown	500K	0.320	0.682	0.162	0.730	0.830	0.473	0.545
SGPT-2.7B (Muennighoff, 2022)	2.7B	unknown	500K	0.339	0.701	0.166	0.752	0.848	0.489	0.561
BMRETRIEVER-1B	1B	10M	1.4M	0.344	0.760	0.180	0.840	0.858	0.531	0.596
BMRETRIEVER-2B	2B	10M	1.4M	0.351	0.760	0.199	0.863	0.828	**0.543**	**0.600**
XL Size (> 5B)										
SGPT-5.8B (Muennighoff, 2022)	5.8B	unknown	500K	0.362	0.747	0.199	0.849	0.863	0.539	0.604
LLaRA (Li et al., 2023a)	7B	21M	500K	0.372	0.757	0.172	0.853	—	0.539	—
RepLLaMA (Ma et al., 2023)	7B	—	500K	0.378	0.756	0.181	0.847	—	0.541	—
LLM2Vec* (BehnamGhader et al., 2024)	7B	1.2M	1.5M	0.393	0.788	0.225	0.776	0.852	0.545	0.606
E5-Mistral* (Wang et al., 2024)	7B	—	1.8M	0.386	0.764	0.162	0.872	0.855	0.546	0.608
CPT-text-XL (Neelakantan et al., 2022)	175B	unknown	unknown	0.407	0.754	—	0.649	—	—	—
BMRETRIEVER-7B	7B	10M	1.4M	0.364	0.778	0.201	0.861	0.847	**0.551**	**0.610**

Table 2: Main experiments on biomedical text representation tasks in various scales. **Bold** and underline indicate the best and second best results on average performance over the four retrieval tasks, and over all five tasks. $*$ denotes concurrent works (for reference only). † uses reranker distillation. ‡ employs hybrid retrieval. We highlight the biomedical or scientific domain-specific retrieval models. Notations are consistent across tables. "PT", "FT", and "Sent. Sim." denote "Pre-training", "Fine-tuning", and "Sentence Similarity", respectively.

dense retrieval models with varying model sizes: *Contriever* (Izacard et al., 2022), *Dragon* (Lin et al., 2023), *SciMult* (Zhang et al., 2023), *SPECTER 2.0* (Singh et al., 2023), *COCO-DR* (Yu et al., 2022), *SGPT* (Muennighoff, 2022), *MedCPT* (Jin et al., 2023), *GTR* (Ni et al., 2022), *InstructOR* (Su et al., 2023), *E5-Large-v2* (Wang et al., 2022b), *BGE-Large* (Chen et al., 2024), *LLaRA* (Li et al., 2023a), *RepLLaMA* (Ma et al., 2023), *LLM2Vec* (BehnamGhader et al., 2024), *E5-Mistral* (Wang et al., 2024), and *CPT-text* (Neelakantan et al., 2022). The details of baselines and parameter sizes are in Appendix C.

Implementation Details. The backbones used for BMRETRIEVER are available in Table 1. The learning rates are set to $5e-5$ for the 410M and 1B variants, $4e-5$ for the 2B variant, and $2e-5$ for the 7B variant during pre-training; $5e-5$ for the 410M and 1B variants, $2e-5$ for the 2B variant, and $1e-5$ for the 7B variant during fine-tuning. The global batch size is set to 256 for the 410M and 1B variants, 128 for the 2B variant, and 64 for 7B variants. To optimize GPU memory consumption, we train our models with LoRA ($r = 16$, $\alpha = 32$) (Hu et al., 2022), brain floating point (bfloat16) quantization, and DeepSpeed gradient checkpointing (Rasley et al., 2020). The training is performed on 4 NVIDIA H100 GPUs for 2 epochs during pre-training and 1 epoch during fine-tuning, using a maximum sequence length of 512 tokens. We use the AdamW optimizer (Loshchilov and Hutter, 2019) with a linear learning rate warm-up for the first 100 steps. For contrastive learning, we set $\tau = 1$ without any further tuning.

Evaluation. We use nDCG@10 to measure standard IR performance and Spearman correlation for STS based on *cosine similarity*. To evaluate the retrieval performance of QA, we report Recall@{5,20} and nDCG@20. For entity linking, we report mean reciprocal rank (MRR)@5 and Recall@{1,5}. For paper recommendation, we follow Singh et al. (2023) and report mean average precision (MAP) and nDCG.

4.2 Results on Text Representation Tasks

Table 2 presents a comprehensive evaluation of the embedding quality on four standard biomedical IR tasks and an additional task focused on

Task	Question Answering									Entity Linking						Paper Rec.	
Model	BioASQ			PubMedQA			iCliniq			DrugBank			MeSH			RELISH	
	R@5	R@20	nDCG@20	R@5	R@20	nDCG@20	R@5	R@20	nDCG@20	R@1	R@5	MRR@5	R@1	R@5	MRR@5	MAP	nDCG
Base Size (< 1B)																	
Dragon (2023)	36.2	**54.6**	49.1	71.8	74.0	72.0	50.6	65.2	47.4	81.0	87.6	83.3	28.2	47.0	34.8	72.6	80.6
MedCPT (2023)	34.7	54.4	45.2	66.3	71.1	60.4	26.8	42.0	24.9	75.1	88.0	80.6	27.7	54.2	37.4	83.6	89.7
E5-Large-v2† (2022b)	36.8	54.0	50.4	71.6	74.2	72.2	57.6	72.0	55.8	**81.8**	86.5	81.5	**32.8**	**55.0**	**41.3**	84.9	91.0
BMRETRIEVER-410M	**39.9**	54.2	**53.1**	**73.8**	**74.6**	**72.4**	**60.6**	**72.8**	**56.6**	81.4	**88.2**	**83.7**	31.5	53.8	39.8	**85.2**	**91.2**
Large Size (1B - 5B)																	
InstructOR-XL (2023)	29.9	43.2	41.8	70.5	74.0	69.1	64.9	78.1	58.3	75.3	84.2	80.3	33.6	56.2	45.7	84.5	90.6
SGPT-2.7B (2022)	33.9	47.4	47.3	68.3	73.7	63.2	45.0	52.2	41.2	71.9	77.0	62.9	20.2	39.7	28.5	84.9	90.8
BMRETRIEVER-1B	40.4	55.8	53.4	73.6	74.4	72.7	61.1	73.7	56.8	**84.7**	89.1	**86.5**	35.5	60.3	48.8	85.2	91.3
BMRETRIEVER-2B	**42.5**	**56.5**	**55.7**	**74.0**	**74.6**	**73.1**	**70.0**	**81.2**	**65.7**	82.6	**90.2**	85.8	**45.6**	**71.3**	**59.5**	**85.4**	**91.5**
XL Size (> 5B)																	
E5-Mistral* (2024)	39.6	55.4	52.7	72.6	74.2	70.0	56.7	72.2	51.8	78.5	92.2	84.0	47.9	76.2	**61.3**	85.2	90.8
BMRETRIEVER-7B	**43.7**	**60.2**	**57.4**	**74.2**	**74.6**	**73.8**	**68.4**	**79.7**	**63.7**	**84.7**	**92.8**	**88.0**	**49.8**	**76.5**	61.1	**86.7**	**92.2**

Table 3: Experiments on retrieval-oriented biomedical NLP applications compared with strongest and fair baselines.

biomedical sentence similarity. Across different scales, BMRETRIEVER outperforms the majority of baseline methods, achieving either the highest or second-highest performance in terms of average scores on the four IR tasks, as well as on the combined set of all five tasks. It even outperforms E5-Large-v2 (Wang et al., 2022b) with additional supervision signals and matches BGE-Large's hybrid retrieval approach combining dense, lexical, and multi-vector retrieval (Chen et al., 2024). Here we focus on scaling up biomedical retrieval models with mixed data types, leaving the combination of BMRETRIEVER with other more complex and larger scale language systems for future work.

A notable aspect of BMRETRIEVER is its efficiency and lightweight nature. Its 410M, 1B, and 2B variants achieve 94.1%, 97.7%, and 98.4% performance using only 5.9%, 14.3%, and 28.6% of 7B variant's parameters, respectively. Moreover, BMRETRIEVER-410M outperforms all the baselines in large size (1B-5B) with up to 11.7× more parameters, and BMRETRIEVER-2B matches performance with baselines in XL size (> 5B). Remarkably, BMRETRIEVER also provides a reasonable training setup within an academic budget, requiring only 10M pre-training data and 1.5M fine-tuning data, which is significantly less than the data usage in most baselines, such as GTR (Ni et al., 2022) and MedCPT (Jin et al., 2023). Yet, BMRETRIEVER still outperforms these data-intensive methods.

4.3 Results on Retrieval-Oriented Biomedical Applications

Table 3 evaluates BMRETRIEVER's performance on biomedical downstream applications. The results demonstrate BMRETRIEVER's efficacy over most baselines across different tasks and datasets, justifying the adaptability of our learned representations to various retrieval-oriented applications.

Task	Size	Standard IR					Sent. Sim.	Avg.	Avg.
Model		NFC.	Sci-Fact	Sci-Docs	Trec-COVID	BIOSSES		Retr.	All
Contriever (2022)	110M	0.328	0.677	0.165	0.274	0.781	0.347	0.434	
COCO-DR (2022)	110M	0.243	0.724	0.150	0.483	0.801	0.400	0.480	
QExt (2022)	110M	0.303	0.644	0.147	0.535	—	0.407	—	
E5-Large-v2 (2022b)	335M	0.337	0.723	0.218	0.618	0.822	0.474	0.543	
LLM2Vec* (2024)	7B	0.271	0.687	0.153	0.557	0.832	0.417	0.500	
BMRETRIEVER	410M	0.306	0.677	0.180	0.802	0.834	0.491	0.560	
BMRETRIEVER	1B	0.330	0.744	0.187	0.800	0.833	0.515	0.579	
BMRETRIEVER	2B	0.342	0.738	0.198	0.848	0.847	0.531	0.593	
BMRETRIEVER	7B	0.355	0.750	0.208	0.833	0.861	**0.537**	**0.601**	

Table 4: The performance of unsupervised dense retrieval models on biomedical representation tasks. Directly using the backbone model of BMRETRIEVER (before contrastive pre-training) leads to performance < 0.03 for all datasets, thus we do not report them.

Furthermore, our proposed BMRETRIEVER exhibits strong generalization capabilities across diverse tasks and input formats, including retrieving long context from short questions (BioASQ, PubMedQA), retrieving long answers from patient questions (iCliniq), retrieving definitions from entity names (DrugBank, MeSH), and retrieving relevant abstracts given an abstract (RELISH). Notably, BMRETRIEVER performs well on unseen tasks, such as entity linking and paper recommendation, verifying its ability to generalize to new tasks unseen in the instruction fine-tuning stage.

4.4 Unsupervised Retrieval Performance

To highlight the effectiveness of our contrastive pre-training approach, we evaluate the performance of unsupervised dense retrieval models that only use unlabeled corpora for pre-training and synthetic data for finetuning. As shown in Table 4, our model outperforms existing unsupervised baselines and even surpasses many fully supervised models reported in Table 2. The strong unsupervised results have important implications for real-world

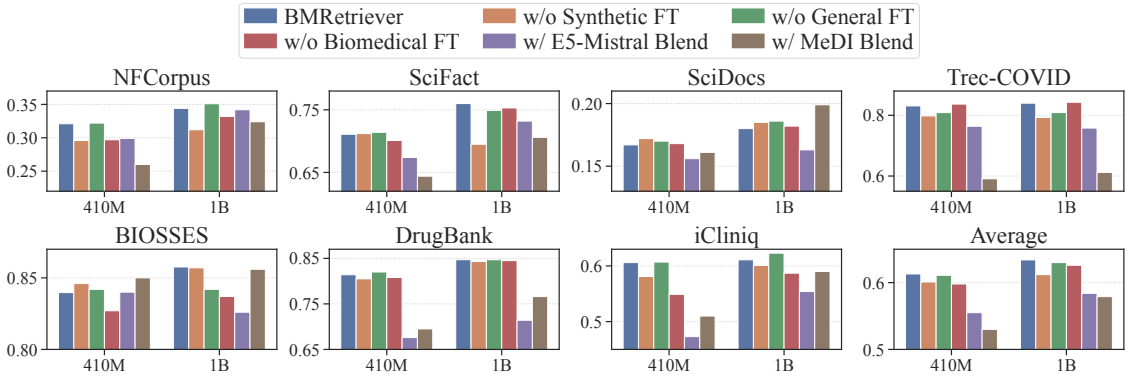
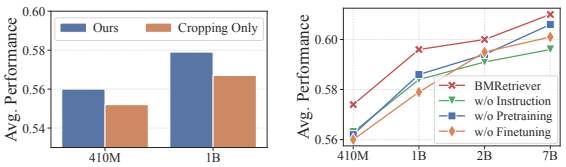


Figure 3: Effect of different fine-tuning data on various datasets. "FT" denotes "Fine-tuning".



(a) Effect of CL strategies (b) Ablation Studies

Figure 4: Additional results over five tasks in the main experiments. "CL" stands for "Contrastive Learning".

Stage (↓)	Volume (→)	10%	50%	100%
Pre-training	BMRETRIEVER-410M	0.540	0.554	0.560
	BMRETRIEVER-1B	0.564	0.575	0.579
Fine-tuning	BMRETRIEVER-410M	0.562	0.571	0.574
	BMRETRIEVER-1B	0.590	0.595	0.596

Table 5: Effect of data volume in pre-training and fine-tuning. Pre-training results do not involve subsequent fine-tuning. Fine-tuning results are based on the pre-training checkpoints with full pre-training data.

biomedical applications, where curating large labeled datasets is often prohibitively expensive and time-consuming. Our approach presents an attractive alternative, enabling the development of high-quality retrieval models in a data-efficient manner.

We further investigate the performance of employing cropping alone as the contrastive pre-training strategy, which entails randomly selecting two passages from the corpus as a positive query-passage pair (Izacard et al., 2022). The results presented in Table 4(a) demonstrate that utilizing cropping as the sole contrastive learning objective yields suboptimal performance.

4.5 Studies on Instruction Fine-tuning

Figure 3 illustrates the impact of different fine-tuning data sources on model performance across various datasets[2]. Among all the utilized data types, *synthetic data* contributes the most significant performance gain, which can be attributed to its larger volume compared to biomedical data and its coverage of a more diverse range of task types. It is particularly beneficial for NFCorpus, SciFact, and Trec-COVID, as these datasets follow the standard IR format of short queries and long passages, aligning with the format of the synthetic data. Furthermore, synthetic data proves advantageous for the iCliniq dataset, as it potentially includes various retrieval scenarios, such as dialog data. *General domain fine-tuning data*, consisting of short queries and long passages, generally enhances relevance estimation capabilities, benefiting standard biomedical IR tasks like Trec-COVID. However, it may slightly diminish performance on datasets like iCliniq, which contain conversational patient-provider data that deviates from traditional IR formats, potentially introducing distribution shifts. *Biomedical fine-tuning data*, on the other hand, demonstrates broad usefulness, particularly for BIOSSES and iCliniq tasks, likely due to its inclusion of sentence similarity and dialog data aligning with downstream task formats. *Public fine-tuning data* from *E5-Mistral* (Wang et al., 2024) and *MEDI*, used by InstructOR (Su et al., 2023), achieves relatively poor performance, possibly due to their focus on the general domain with limited biomedical knowledge. Additionally, the lack of publicly available synthetic generated data from E5-Mistral (Wang et al., 2024) may contribute to the suboptimal performance observed.

4.6 Additional Studies

Ablation Studies. We inspect different components of BMRETRIEVER in Figure 4(b), including instruction, pre-training, and fine-tuning[3]. The re-

[2]Removing biomedical data retains the synthetic data.

[3]Removing instruction solely eliminates the retrieval guidance, while retaining both pre-training and fine-tuning.

Case Study 1	
Dataset / Task	iCliniq / Given a patient query, retrieve its corresponding answer from the community website.
Query	My grandmother is having dizziness, vomiting and headache. These problems occur constantly with a gap of five to six months.
BMRETRIEVER	From the symptoms that you mentioned it seems your grand mother has migraine which reccurs when she is either exposed to the allergen, which triggers her migraine, or it can also manifest as a seasonal change. How long has she has been having this problem? During her migraine episodes, how long does it last? I will recommend the following medicines :- 1) Tablet Crocin pain relief one tablet twice daily after food. 2) Tablet Ondem (Ondansetron) 4 mg, if there is severe nausea two tablets can be taken at once. Make sure her BP (blood pressure) is under control.
BMRETRIEVER w/o Instruction	Please provide details regarding your headache and abuse. Also, provide details about your current symptoms and like how they started and how they progress over time. What are the medications you are taking currently and for how long? Once I receive this information, I will be able to help you.
MedCPT	Since the movement is worsening the symptoms, it must be an inner ear problem. There could be additionally middle ear block or eustachian tube block as the pressure change worsening it. However, since you have not mentioned the duration, I am mentioning all possible differentials. Recurrent vertigo lasting under one minute is usually benign paroxysmal positional vertigo. A single episode of vertigo lasting several minutes to hours may be due to migraine or transient ischemia of the labyrinth or brainstem. · · ·
E5-Mistral	You are suffering from Episodes of headache with nausea, which are not present every day. also having left eye pain and twisting with tingling, imbalance while getting up, tingling in throat and over genitals, coldness over forehead ... I would like to suggest to do some routine blood investigation tests and this symptoms seems like vitamin B12 deficiency or hypothyroidism. Are you Vegetarian?
Case Study 2	
Dataset / Task	SciFact / Given a biomedical claim, retrieve biomedical literature to support or refute the claim.
Query	Forkhead 0 (fox0) transcription factors are involved in apoptosis.
BMRETRIEVER	Cell Stem Cell Review FoxO Transcription Factors and Stem Cell Homeostasis: Insights from the Hematopoietic System The forkhead O (FoxO) family of transcription factors participates in diverse physiologic processes, including induction of cell-cycle arrest, stress resistance, differentiation, apoptosis, and metabolism. Several recent studies indicate that FoxO-dependent signaling is required for long-term regenerative potential of the hematopoietic stem cell (HSC) compartment through regulation of HSC response to physiologic oxidative stress, quiescence, and survival. · · ·
MedCPT	Forkhead box transcription factor, class O (FOXO) is a mammalian homologue of DAF-16, which is known to regulate the lifespan of Caenorhabditis elegans and includes subfamilies of forkhead transcription factors such as AFX, FKHRL1, and FKHR. FKHR is phosphorylated on three sites (Thr-24, Ser-256, and Ser-319) in a phosphatidylinositol 3-kinase (PI3K)/Akt-dependent manner, thereby inhibiting death signals. We here documented dephosphorylation of FKHR following transient forebrain ischemia with its concomitant translocation into the nucleus in neurons in gerbil and mouse brains. The activation of FKHR preceded delayed neuronal death in the vulnerable hippocampal regions following ischemic brain injury. · · ·
E5-Mistral & BMRETRIEVER w/o Instruction	Novel Foxo1-dependent transcriptional programs control Treg cell function Regulatory T (Treg) cells, characterized by expression of the transcription factor forkhead box P3 (Foxp3), maintain immune homeostasis by suppressing self-destructive immune responses. Foxp3 operates as a late-acting differentiation factor controlling Treg cell homeostasis and function, whereas the early Treg-cell-lineage commitment is regulated by the Akt kinase and the forkhead box O (Foxo) family of transcription factors. However, whether Foxo proteins act beyond the Treg-cell-commitment stage to control Treg cell homeostasis and function remains largely unexplored. Here we show that Foxo1 is a pivotal regulator of Treg cell function. · · ·

Table 6: A case study with two examples illustrating the quality of retrieved passages from BMRETRIEVER compared with baseline models. Blue text denotes keywords present in the original query, while green and red represent relevant and irrelevant keywords, respectively, in the retrieved passages. "· · ·" at the end indicates that the remaining portion of the passage is omitted due to space constraints.

sults indicate that removing any component would hurt the performance. We also observe that pre-training is particularly beneficial for smaller models, as larger models may already possess sufficient capacity to capture domain knowledge.

Effect of Data Volume. Table 5 evaluates the effect of data volume during pre-training and fine-tuning. The results demonstrate the remarkable efficiency of BMRETRIEVER, achieving comparable performance even when trained on substantially less data. Notably, using only 10% of the data, the 1B variant of BMRETRIEVER outperforms all baselines in either the pre-training or fine-tuning stage, while the 410M variant also achieves better performance than most baselines in fine-tuning.

4.7 Case Study

We present two case studies in Table 6 illustrating the quality of retrieved passages from BMRETRIEVER compared to strong baselines. The first example, from the iCliniq dataset, considers a patient query and retrieves the corresponding answer from a community website. In the given example, BMRETRIEVER retrieves a passage directly addressing symptoms like *headaches* and *nausea*, recommending medication aligning with the condition. In contrast, the retrieved passage from MedCPT focuses on *inner ear problems* and *vertigo*, not covering the vomiting or the specific periodicity of the episodes described in the query. The passage from E5-Mistral talks about symptoms not mentioned by the patient, such as *left eye pain* and *tingling*. Besides, we also present the result from BMRETRIEVER without using instructions, which is also imprecise since it mentions *abuse*, a topic not relevant to the query.

The second example involves retrieving biomedical literature to support or refute a claim about *apoptosi*. The passage retrieved by BMRETRIEVER specifically mentions that the *FoxO* family of transcription factors participates in *apoptosis*. Although the passage retrieved by MedCPT discusses the role of FoxO transcription factors in *cell death*, it is specific to neuronal cells under *ischemic conditions*, rather than general apoptosis. Furthermore, both E5-Mistral and BMRETRIEVER

without instructions retrieve an irrelevant passage about the role of FoxO1 in regulating *regulatory T cells*, unrelated to the claim. We further illustrate the cosine similarity distributions of relevant and irrelevant (query, passage) pairs in Appendix E.

5 Conclusion

We present BMRETRIEVER, a series of dense retrieval models designed for knowledge-intensive biomedical NLP tasks with various scales. BMRETRIEVER is pre-trained on a large-scale biomedical corpus and further instruction fine-tuned on diverse, high-quality biomedical tasks. Through extensive experimentation, we have demonstrated that BMRETRIEVER exhibits state-of-the-art performance across a range of biomedical applications. Furthermore, BMRETRIEVER demonstrates impressive parameter efficiency, with its smaller variants achieving 94-98% of the performance of the 7B model using only 6-29% as many parameters, while the 410M version surpasses larger baselines (1B-5B) up to 11.7 times larger. We hope BMRETRIEVER can be incorporated into a broad suite of biomedical tasks to advance biomedical NLP research.

Acknowledgement

We thank the anonymous reviewers and area chairs for valuable feedbacks. This research was also partially supported by the National Science Foundation under Award Number 2319449 and Award Number 2312502, the National Institute Of Diabetes And Digestive And Kidney Diseases of the National Institutes of Health under Award Number K25DK135913, the Emory Global Diabetes Center of the Woodruff Sciences Center, Emory University. JH was supported by NSF grants IIS-1838200 and IIS-2145411. YY and CZ was supported in part by NSF IIS-2008334 and CAREER IIS-2144338.

Limitation

Efficiency. One specific caveat for scaling up model size is the increment in the latency overhead. We have reported both the *passage indexing* speed and *retrieval latency* in Appendix F, which indicates that our model does not incur much additional time when compared to models with similar size (e.g., BMRETRIEVER-2B v.s. InstructOR-1.5B). One important future work is to explore how to reduce the inference latency and lower the storage cost for text embeddings produced by LLMs.

Cost Estimation. Generating synthetic data using GPT models incurs additional costs. In our work, the total API cost of BMRETRIEVER is less than $500[4], which remains affordable within an academic budget. This cost is significantly lower than recent works (Wang et al., 2024), which have an estimated cost of more than $6000.

Ethics Consideration

Misinformation. One specific issue for LLM-generated biomedical text is the potential for misinformation and hallucination (Pal et al., 2023). It is important to note that for the *generated queries*, the majority are short sentences or phrases without presenting any scientific facts. Regarding the generated (query, passage) pairs, to ensure that our generated synthetic text does not introduce misinformation or hallucination, we randomly selected 200 examples and asked medical students to evaluate the factuality of the generated text. The evaluation results did not reveal misinformation or hallucination in the randomly selected examples.

Data Contamination. A potential issue is test set contamination (Sainz et al., 2023), where some test examples overlap with the training data. This can be especially problematic for text generated by LLMs, as they are often pre-trained on massive corpora spanning various domains. To address this concern, we follow Wang et al. (2024) to conduct a string match-based analysis between the test set and our training set, where we *do not* observe any overlap between the train and test queries. While some of the corpora (e.g., PubMed) are also utlized in the test tasks, this is a standard practice even in zero-shot or few-shot evaluation of retrieval models (Ma et al., 2021; Wang et al., 2022a; Yu et al., 2022), and it is not considered as contamination.

References

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. Synthetic QA corpora generation with roundtrip consistency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6168–6173, Florence, Italy. Association for Computational Linguistics.

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*

[4]As of April 2024.

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, et al. 2016. MS MARCO: A human generated machine reading comprehension dataset. *arXiv preprint 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*.

Asma Ben Abacha, Chaitanya Shivade, and Dina Demner-Fushman. 2019. Overview of the MEDIQA 2019 shared task on textual inference, question entailment and question answering. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 370–379, Florence, Italy. Association for Computational Linguistics.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 2397–2430. PMLR.

Vera Boteva, Demian Gholipour, Artem Sokolov, and Stefan Riezler. 2016. A full-text learning to rank dataset for medical information retrieval. In *European Conference on Information Retrieval*, pages 716–722. Springer.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Peter Brown, Aik-Choon Tan, Mohamed A El-Esawi, Thomas Liehr, Oliver Blanck, Douglas P Gladue, Gabriel MF Almeida, Tomislav Cernava, Carlos O Sorzano, Andy WK Yeung, et al. 2019. Large expert-curated database for benchmarking document similarity detection in biomedical literature search. *Database*, 2019.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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*.

Qingyu Chen, Alexis Allot, and Zhiyong Lu. 2021. Litcovid: an open database of covid-19 literature. *Nucleic acids research*, 49(D1):D1534–D1540.

Shu Chen, Zeqian Ju, Xiangyu Dong, Hongchao Fang, Sicheng Wang, Yue Yang, Jiaqi Zeng, Ruisi Zhang, Ruoyu Zhang, Meng Zhou, Penghui Zhu, and Pengtao Xie. 2020. Meddialog: A large-scale medical dialogue dataset. *CoRR*, abs/2004.03329.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

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

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Nicolas Fiorini, Robert Leaman, David J Lipman, and Zhiyong Lu. 2018. How user intelligence is improving pubmed. *Nature biotechnology*, 36(10):937–945.

Giacomo Frisoni, Miki Mizutani, Gianluca Moro, and Lorenzo Valgimigli. 2022. BioReader: a retrieval-enhanced text-to-text transformer for biomedical literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5770–5793, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Daniel Gillick, Sayali Kulkarni, Larry Lansing, Alessandro Presta, Jason Baldridge, Eugene Ie, and Diego Garcia-Olano. 2019. Learning dense representations for entity retrieval. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 528–537, Hong Kong, China. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey,

and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry P. Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *22nd ACM International Conference on Information and Knowledge Management, CIKM'13, San Francisco, CA, USA, October 27 - November 1, 2013*, pages 2333–2338. ACM.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Fan Jiang, Tom Drummond, and Trevor Cohn. 2023. Noisy self-training with synthetic queries for dense retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11991–12008, Singapore. Association for Computational Linguistics.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651.

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.

Daniel Khashabi, Amos Ng, Tushar Khot, Ashish Sabharwal, Hannaneh Hajishirzi, and Chris Callison-Burch. 2021. GooAQ: Open question answering with diverse answer types. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 421–433, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. BioMistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5848–5864, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Jakub Lála, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G Rodriques, and Andrew D White. 2023. Paperqa: Retrieval-augmented generative agent for scientific research. *arXiv preprint arXiv:2312.07559*.

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

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Chaofan Li, Zheng Liu, Shitao Xiao, and Yingxia Shao. 2023a. Making large language models a better foundation for dense retrieval. *arXiv preprint arXiv:2312.15503*.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023b. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).

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.

Carolyn E Lipscomb. 2000. Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238, Online. Association for Computational Linguistics.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Man Luo, Arindam Mitra, Tejas Gokhale, and Chitta Baral. 2022a. Improving biomedical information retrieval with neural retrievers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11038–11046.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022b. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409.

Ji Ma, Ivan Korotkov, Yinfei Yang, Keith Hall, and Ryan McDonald. 2021. Zero-shot neural passage retrieval via domain-targeted synthetic question generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1075–1088, Online. Association for Computational Linguistics.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2023. Fine-tuning llama for multi-stage text retrieval. *arXiv preprint arXiv:2310.08319*.

Clara H McCreery, Namit Katariya, Anitha Kannan, Manish Chablani, and Xavier Amatriain. 2020. Effective transfer learning for identifying similar questions: matching user questions to covid-19 faqs. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3458–3465.

Rui Meng, Ye Liu, Semih Yavuz, Divyansh Agarwal, Lifu Tu, Ning Yu, Jianguo Zhang, Meghana Bhat, and Yingbo Zhou. 2022. Augtriever: Unsupervised dense retrieval by scalable data augmentation. *arXiv preprint arXiv:2212.08841*.

Sunil Mohan, Nicolas Fiorini, Sun Kim, and Zhiyong Lu. 2017. Deep learning for biomedical information retrieval: Learning textual relevance from click logs. In *BioNLP 2017*, pages 222–231, Vancouver, Canada,. Association for Computational Linguistics.

Niklas Muennighoff. 2022. Sgpt: Gpt sentence embeddings for semantic search. *arXiv preprint arXiv:2202.08904*.

Aakanksha Naik, Sravanthi Parasa, Sergey Feldman, Lucy Lu Wang, and Tom Hope. 2022. Literature-augmented clinical outcome prediction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 438–453, Seattle, United States. Association for Computational Linguistics.

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*.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, and Yinfei Yang. 2022. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9844–9855, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. Med-HALT: Medical domain hallucination test for large language models. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 314–334, Singapore. Association for Computational Linguistics.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2021. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5835–5847, Online. Association for Computational Linguistics.

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.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

Wenqi Shi, Yuchen Zhuang, Yuanda Zhu, Henry Iwinski, Michael Wattenbarger, and May Dongmei Wang. 2023. Retrieval-augmented large language models for adolescent idiopathic scoliosis patients in shared decision-making. In *Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 1–10.

Chaitanya Shivade. 2017. Mednli — a natural language inference dataset for the clinical domain.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2023. SciRepEval: A multi-format benchmark for scientific document representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5548–5566, Singapore. Association for Computational Linguistics.

Gizem Soğancıoğlu, Hakime Öztürk, and Arzucan Özgür. 2017. Biosses: a semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics*, 33(14):i49–i58.

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.

Flax Sentence Embeddings Team. 2021. Stack exchange question pairs.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

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 *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):1–28.

Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R. Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2021. TREC-COVID: Constructing a pandemic information retrieval test collection. *SIGIR Forum*, 54(1).

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Kexin Wang, Nandan Thakur, Nils Reimers, and Iryna Gurevych. 2022a. GPL: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2345–2360, Seattle, United States. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022b. 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. 2024. 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.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Nancy Xin Ru Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The COVID-19 open

research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

Yubo Wang, Xueguang Ma, and Wenhu Chen. 2023. Augmenting black-box llms with medical textbooks for clinical question answering. *arXiv preprint arXiv:2309.02233*.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, et al. 2018. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic acids research*, 46(D1):D1074–D1082.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6233–6251, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*.

Ran Xu, Wenqi Shi, Yue Yu, Yuchen Zhuang, Bowen Jin, May Dongmei Wang, Joyce Ho, and Carl Yang. 2024. RAM-EHR: Retrieval augmentation meets clinical predictions on electronic health records. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 754–765, Bangkok, Thailand. Association for Computational Linguistics.

Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. 2024. RankRAG: Unifying context ranking with retrieval-augmented generation in LLMs. In *Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Yue Yu, Chenyan Xiong, Si Sun, Chao Zhang, and Arnold Overwijk. 2022. COCO-DR: Combating distribution shift in zero-shot dense retrieval with contrastive and distributionally robust learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1462–1479, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tianjun Zhang, Shishir G Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E. Gonzalez. 2024. RAFT: Adapting language model to domain specific RAG. In *First Conference on Language Modeling*.

Yu Zhang, Hao Cheng, Zhihong Shen, Xiaodong Liu, Ye-Yi Wang, and Jianfeng Gao. 2023. Pre-training multi-task contrastive learning models for scientific literature understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12259–12275, Singapore. Association for Computational Linguistics.

A Additional Synthetic Data Augmentation Details

A.1 Prompt format to Generate Query from Passage

Listing 1: Prompt Format for synthetic query generation.

```
Given the passage in [dataset],
please generate a query that is
relevant to the provided passage.
```

[dataset]: The dataset from which the provided passage is selected.

A.2 Prompt Format to Generate Task and Pairs

Listing 2: Prompt format for synthetic retrieval task generation.

```
Brainstorm a list of potentially
useful biomedical text retrieval
tasks.

Here are a few examples for your
reference:
1. Provided a scientific claim as
 query, retrieve documents that
help verify or refute the claim.
2. Search for documents that
answers a FAQ-style query on
children's nutrition.

Please adhere to the following
guidelines:
1. Specify what the query is, and
 what the desired documents are.
2. Each retrieval task should
cover a wide range of queries,
and should not be too specific.
3. Focus on biomedical related
topics.

Your output should always be a
python list of strings only, with
 about 20 elements, and each
element corresponds to a distinct
 retrieval task in one sentence.
Do not explain yourself or output
 anything else. Be creative!
```

Listing 3: Prompt format for synthetic retrieval examples generation.

```
You have been assigned a
biomedical retrieval task: [task]

Your mission is to write one
biomedical text retrieval example
 for this task in JSON format.
The JSON object must contain the
following keys:
1. "user_query": a string, a
random user search query
specified by the retrieval task.
2. "positive_document": a string,
 a relevant document for the user
 query.
3. "hard_negative_document": a
string, a hard negative document
that only appears relevant to the
 query.

Please adhere to the following
guidelines:
1. The "user_query" should be
[query_type], [query_length],
[clarity], and diverse in topic.
2. All documents should be at
least [num_words] words long.
3. Both the query and documents
should be in English.
4. Both the query and documents
require [difficulty] level
education to understand.

Your output must always be a JSON
 object only, do not explain
yourself or output anything else.
 Be creative!
```

[task]: The task names generated from the previous step.

[query_type]: Randomly sampled from ["extremely long-tail", "long-tail", "common"].

[query_length]: Randomly sampled from ["less than 5 words", "5-10 words", "at least 10 words"]

[clarity]: Randomly sampled from ["clear", "understandable with some effort", "ambiguous"]

[num_words]: Randomly sampled from ["50 words", "50-100 words", "200 words", "300 words", "400 words"]

[difficulty]: Randomly sampled from ["high school", "college", "PhD"]

A.3 Case Study

We present a list of generated retrieval scenarios as examples:

- "Search for articles discussing the latest advancements in neurology."
- "Retrieval of articles discussing the symptoms and treatments of rare diseases given a query on rare diseases."
- "Find documents that discuss the impact of lifestyle changes on a specific medical condition."
- "Locate documents that provide information on the epidemiology of a certain disease in a specific region."
- ...

Table 7 presents two illustrative examples where GPT-4 generates corresponding queries, positive passages, and negative passages for each synthetic retrieval task. The complete set of task names is provided in the supplementary materials.

B Task and Dataset Information

B.1 Pre-training Corpus

We publicly release the training recipe used in both the pre-training and fine-tuning stages to ensure transparency, reproducibility, and potential applicability to new domains. To equip BMRETRIEVER with a strong foundation in biomedical contexts, we compile a diverse corpus of biomedical data sources. Table 8 summarizes the unlabeled corpora used for contrastive pre-training of our model, including their sizes and public availability. For pre-training on BMRETRIEVER-7b, we only use 1M passages due to the efficiency issue.

For queries and passages, the instruction used in the contrastive pre-training stage is "`Given a query, retrieve passages that are relevant to the query. Query: {}`", "`Represent this passage. Passage: {}`".

B.2 Fine-tuning Task and Dataset

Real Datasets. Table 9 displays the datasets used for instruction fine-tuning besides synthetic augmentation, which include a diverse range of tasks at both the sentence and passage levels across biomedical and general domains. Biomedical datasets cover biomedical QA (Team 2021, Ben Abacha et al. 2019), sentence similarity (Shivade 2017, McCreery et al. 2020), and dialogue (Li et al. 2023b). General domain datasets tackle long-form QA (Fan et al. 2019), web search (Bajaj et al., 2016), open-domain QA (Khashabi et al. 2021, Kwiatkowski et al. 2019), fact verification (Thorne et al. 2018), NLI (Bowman et al. 2015), and web search (Bajaj et al. 2016). For MS Marco[5] and NQ dataset[6], we use the ground-truth annotations as well as the provided hard negative to form the fine-tuning data.

For non-retrieval tasks, we convert them into a retrieval format as follows:

- For standard QA datasets, we directly use the question as the query and the gold evidence passages as the ground-truth passages.
- For NLI and sentence similarity tasks, we treat sentence pairs identified as "entail" or "similar" as positive examples, while those labeled as "contradict" or "non-similar" serve as *hard negatives*.
- For medical dialogue datasets, we consider the answer to the user query as the ground-truth passage for retrieval.

The query instructions are listed in the corresponding tables, while for passages, we use the same instruction format as the template used in the pre-training stage: "`Represent this passage. Passage: {}`".

Synthetic Datasets. We leverage LLM-generated synthetic data to augment the training set. For the query generation scenario, we generate 500K synthetic queries. After round-trip filtering, we retain approximately 420K (query, passage) pairs. The instructions used for generating synthetic queries are:

- For the PubMed corpus: "`Given a question, retrieve Pubmed passages that answer the question.`"
- For the Meadow corpus on COVID-19: "`Given a query on COVID-19, retrieve COVID-19 related articles that answer the query.`"

[5] https://msmarco.z22.web.core.windows.net/msmarcoranking/triples.train.small.tar.gz

[6] https://dl.fbaipublicfiles.com/dpr/data/retriever/biencoder-nq-adv-hn-train.json.gz

Synthetic Data Augmentation Example 1	
Synthetic Retrieval Task	Retrieve documents that discuss the potential health risks associated with obesity
Query	Potential health hazards related to obesity
Positive Passage	The escalating obesity epidemic is associated with a range of serious health risks and complications. Obesity is a chronic disease characterized by excessive accumulation and storage of fat in the body, which in turn is associated with a multitude of adverse health effects. These effects include an increased risk of premature death, cardiovascular diseases including hypertension, stroke and coronary artery disease, type 2 diabetes, certain types of cancer, and sleep apnea among others. The psychosocial consequences of obesity are also significant, with reduced quality of life and depression being common. The pathophysiological basis of these complications is complex and involves a myriad of interrelated metabolic, inflammatory, and vascular processes. For instance, the excess adiposity in obesity contributes to systemic inflammation and insulin resistance, which are central to the development of type 2 diabetes and cardiovascular diseases. The increased mechanical load of obesity can also lead to degenerative joint diseases, while the increased metabolic demand can lead to heart failure. Furthermore, the excess adipose tissue can secrete a variety of adipokines and cytokines that can promote carcinogenesis. Thus, the potential health risks associated with obesity are multifold, and the prevention and management of obesity should be a public health priority.
Negative Passage	Obesity is increasingly recognized as a major health problem worldwide. While it is associated with a multitude of health risks, the potential benefits of obesity are less well known. Some studies have suggested that obese individuals may have a lower risk of certain conditions such as osteoporosis, due to their higher bone mineral density. Furthermore, some obese individuals seem to be 'metabolically healthy', with no apparent signs of the metabolic abnormalities typically associated with obesity. These findings, however, do not negate the overall detrimental effects of obesity on health. Rather, they highlight the complexity of obesity as a disease and the need for personalized approaches in its management. It is also important to note that the 'obesity paradox', where obesity seems to confer survival advantages in certain conditions such as heart failure, remains controversial and is not yet fully understood. Overall, while there may be potential benefits associated with obesity, these are likely outweighed by the substantial health risks, and efforts should be focused on preventing and managing obesity to improve health outcomes.
Synthetic Data Augmentation Example 2	
Synthetic Retrieval Task	Search for documents that provide information on the latest treatments for autoimmune diseases
Query	I am looking for scholarly articles or scientific papers that describe the most recent advancements in therapies for autoimmune diseases, such as rheumatoid arthritis, lupus, celiac disease, or multiple sclerosis.
Positive Passage	In recent years, there have been significant advancements in the treatment of autoimmune diseases. One major development is the use of biologics, which are drugs derived from living organisms. Biologics have been successfully used in the treatment of rheumatoid arthritis, lupus, and other autoimmune disorders. They work by targeting specific components of the immune system that cause inflammation and damage. Another promising treatment is stem cell therapy, which has potential in treating diseases such as multiple sclerosis. In this procedure, the patient's immune system is suppressed and then re-established with the patient's own stem cells, essentially 'resetting' the immune system. Moreover, dietary intervention, such as a strict gluten-free diet, has been proven to manage celiac disease effectively. However, these treatments all have their own risks and side effects, and research is ongoing to refine these therapies and develop new ones.
Negative Passage	Autoimmune disorders are a group of diseases where the body's immune system attacks its own cells. There are many types of autoimmune diseases, including Rheumatoid Arthritis, Lupus, Celiac Disease, and Multiple Sclerosis. Each of these diseases has different symptoms, causes, and requires different treatments. Some common symptoms of autoimmune diseases are fatigue, joint pain, and swelling, skin problems, and abdominal pain. The causes of these diseases are not fully understood, but they are thought to be a combination of genetic and environmental factors. There is currently no cure for autoimmune diseases, but treatments can help manage the symptoms. Treatments include medication, physical therapy, and in some cases surgery. In the case of celiac disease, a strict gluten-free diet is necessary. It is important to work with a healthcare provider to develop a treatment plan that is tailored to the individual's needs.

Table 7: Synthetic retrieval tasks and examples generated by GPT-4.

We generate 20,000 synthetic tasks and query-passage pairs using GPT-4. Table 7 presents some examples of synthetic retrieval tasks and query-passage pairs.

B.3 Evaluation Task and Dataset

We conduct a comprehensive evaluation of BMRETRIEVER on eleven datasets (Table 10) across five biomedical tasks, including:

Information Retrieval. For passage retrieval tasks in biomedicine, we select four datasets from the BEIR benchmark (Thakur et al., 2021), each focusing on biomedical or scientific-related IR tasks involving complex, terminology-rich documents: (1) **NFCorpus** (Boteva et al., 2016) contains 323 queries related to nutrition facts for medical IR, sourced from 3.6K PubMed documents; (2) **SciFact** (Wadden et al., 2020) includes 300 queries, aiming to retrieve evidence-containing abstracts

Dataset	Size	Line
PubMed (2024)	8M*	https://huggingface.co/datasets/MedRAG/pubmed
arXiv, MedRxiv, BioRxiv	577K	https://huggingface.co/datasets/mteb/raw_arxiv
Meadow (2020)	460k	https://huggingface.co/datasets/medalpaca/medical_meadow_cord19
Textbooks (2021)	50K	https://huggingface.co/datasets/MedRAG/textbooks
StatPearls (2024)	54K	https://huggingface.co/datasets/MedRAG/statpearls
LitCovid (2021)	70K	https://huggingface.co/datasets/KushT/LitCovid_BioCreative
S2ORC (2020)	600K	https://github.com/allenai/s2orc
MS Marco (2016)	1.2M	https://huggingface.co/datasets/Tevatron/msmarco-passage-corpus

Table 8: Biomedical corpora collection for unsupervised contrastive pre-training. *: We randomly select 8M corpus from the full collections.

from 5K scientific papers for fact-checking; (3) **SciDocs** (Cohan et al., 2020) consists of 25K scientific papers for citation prediction with 1K queries containing article titles; (4) **TREC-COVID** (Voorhees et al., 2021) includes 50 queries, with an average of 493.5 relevant documents per query, specifically curated for biomedical IR related to COVID-19.

Sentence Similarity. For sentence retrieval tasks, we evaluate retrieval models on (5) **BIOSSES** (Soğancıoğlu et al., 2017), which comprises 100 sentence pairs extracted from PubMed articles. The similarity of each sentence pair is annotated using a 5-point scale, ranging from 0 (no relation) to 4 (equivalent).

Question-and-Answering. Besides passage and sentence retrieval tasks, we further evaluate the effectiveness of retrieval models on several retrieval-oriented downstream tasks, including biomedical QA. (6) **BioASQ** (Tsatsaronis et al., 2015) and (7) **PubMedQA** (Jin et al., 2019) are large-scale biomedical multi-choice QA datasets derived from PubMed articles. (8) **iCliniq** (Chen et al., 2020) contains medical QA pairs from the public health forum derived from conversations between clinicians and patients.

Entity Linking. For additional retrieval-oriented downstream applications, we conduct two biomedical entity-linking experiments: (9) **DrugBank** (Wishart et al., 2018) for drug entity matching, and (10) **MeSH** (Lipscomb, 2000) for biomedical concept linking.

Paper Recommendation. We evaluate the performance of retrieval models on a paper recommendation task using the (11) **RELISH** dataset (Singh et al., 2023; Brown et al., 2019). It assigns similarity scores ranging from 0 (not similar) to 2 (similar) for locating relevant literature from more than 180K PubMed abstracts.

C Baseline Information

We consider both sparse and dense retrieval models to provide a comprehensive evaluation of retrieval models in biomedical applications.

C.1 Baselines for Retrieval Tasks in Main Experiments

Sparse Retrieval Models. Sparse retrieval models rely on lexical matching between query and document terms to calculate similarity scores.

- **BM25** (Robertson et al., 2009) is the most commonly used sparse retrieval model, employing a scoring function that calculates the similarity between two high-dimensional sparse vectors based on token matching and weighting.

Dense Retrieval Models. Dense retrieval models utilize dense vector representations to capture semantic similarity between queries and documents. In our experiments, we consider dense retrieval models at various scales for a comprehensive evaluation: (1) **Base Size** (<1B parameters), (2) **Large Size** (1B-5B), and (3) **XL Size** (>5B).

- **Contriever** (Izacard et al., 2022) is a dense retrieval model (110M) pre-trained via contrastive learning on documents sampled from Wikipedia and CC-Net (Wenzek et al., 2020) corpora.
- **Dragon** (Lin et al., 2023) is a BERT-base-sized dense retrieval model (110M) that undergoes progressive training using a data augmentation approach, incorporating diverse queries and sources of supervision.
- **SPECTER 2.0** (Singh et al., 2023) is a scientific document representation model (110M) pretrained using multi-format representation learning.

Dataset	Size	Task	Link	Instruction Format
BioMedical Domain				
StackExchange (2021)	43K	QA	https://huggingface.co/datasets/flax-sentence-embeddings/stackexchange_titlebody_best_voted_answer_jsonl	Given a biological query from the stackexchange, retrieve replies most relevant to the query
MedNLI (2017)	4.6K	Sentence Similarity	https://physionet.org/content/mednli/1.0.0/	Given a sentence, retrieve sentences with the same meaning
MQP (2020)	3K	Sentence Similarity	https://huggingface.co/datasets/medical_questions_pairs	Given a sentence, retrieve sentences with the same meaning
MedQuad (2019)	47K	QA	https://huggingface.co/datasets/lavita/MedQuAD	Given a question, retrieve relevant documents that answer the question
HealthcareMagic (2023b)	30K	Dialogue	https://huggingface.co/datasets/medical_dialog	Given a question with context from online medical forums, retrieve responses that best answer the question
General Domain				
ELI5 (2019)	20K*	Longform QA	https://huggingface.co/datasets/eli5	Given a question, retrieve the highest voted answers on Reddit forum
GooAQ (2021)	100K*	QA	https://huggingface.co/datasets/gooaq	Given a question, retrieve relevant passages that answer the question
MS Marco (2016)	500K	Web Search	https://huggingface.co/datasets/ms_marco	Given a web search query, retrieve relevant passages that answer the query
NQ (2019)	58K	QA	https://github.com/facebookresearch/DPR/blob/main/dpr/data/download_data.py	Given a question, retrieve Wikipedia passages that answer the question
FEVER (2018)	10K*	Fact Verification	https://huggingface.co/datasets/BeIR/fever	Given a claim, retrieve documents that support or refute the claim
NLI (2015)	150K*	Natural Language Inference	https://github.com/princeton-nlp/SimCSE/blob/main/data/download_nli.sh	Given a premise, retrieve hypotheses that are entailed by the premise

Table 9: Labeled data collection for instruction fine-tuning with a diverse range of tasks, including both sentence-level NLI and passage-level QA. *: Only a subset of the original dataset is sampled.

- **SciMult** (Zhang et al., 2023) is a retrieval model (110M) that employs a multi-task contrastive learning framework with task-aware specialization and instruction tuning to enhance performance on scientific literature retrieval tasks.
- **COCO-DR** (Yu et al., 2022) is a dense retrieval model (110M) pre-trained using continuous contrastive learning and implicit distributionally robust optimization on domain-specific corpora, enabling adaptation to various downstream tasks.
- **QExt** (Meng et al., 2022) is a data augmentation method that trains dense retrieval models by selecting salient spans from the original document, and generating pseudo queries using transferred language models.
- **SGPT** (Muennighoff, 2022) is a dense retrieval model that employs position-weighted mean pooling and fine-tunes only bias tensors to learn effective representations for semantic search.
- **MedCPT** (Jin et al., 2023) is a biomedical embedding model (220M) specifically designed for biomedical literature retrieval, leveraging contrastive pre-training on medical corpora consisting of 255M user clicks from PubMed search logs (Fiorini et al., 2018).
- **GTR** (Ni et al., 2022) is a generalizable dense retriever that initializes its dual encoders from T5 (Raffel et al., 2020). We conduct a comprehensive comparison with GTR at varying scales, including GTR-Large (335M), GTR-XL (1.2B), and GTR-XXL (4.8B).
- **InstructOR** (Su et al., 2023) is a multitask embedder that generates task- and domain-aware embeddings for a given text input and its corresponding task instructions, without requiring any additional training. We evaluate InstructOR at both base (335M) and large (1.5B) scales.
- **E5-Large-v2** (Wang et al., 2022b) adopts a com-

Dataset	Task	# Queries	# Documents	Link	Instruction Format
NFCorpus (2016)	Biomedical Search	323	3.6K	https://huggingface.co/datasets/BeIR/nfcorpus	Given a question, retrieve relevant documents that best answer the question
SciFact (2020)	Fact Verification	300	5K	https://huggingface.co/datasets/BeIR/scifact	Given a scientific claim, retrieve documents that support or refute the claim
SciDocs (2020)	Citation Prediction	1,000	25K	https://huggingface.co/datasets/BeIR/scidocs	Given a scientific paper title, retrieve paper abstracts that are cited by the given paper
Trec-COVID (2021)	Biomedical Search	50	171K	https://huggingface.co/datasets/BeIR/trec-covid	Given a query on COVID-19, retrieve documents that answer the query
BIOSSES (2017)	Biomedical Sentence Similarity	100	—	https://huggingface.co/datasets/biosses	Given a sentence, retrieve sentences with the same meaning
BioASQ (2015)	Biomedical QA	500	500K	http://participants-area.bioasq.org/datasets/	Given a question, retrieve Pubmed passages that answer the question
PubMedQA (2019)	Biomedical QA	500	211K	https://huggingface.co/datasets/qiaojin/PubMedQA	Given a question, retrieve Pubmed passages that answer the question
iCliniq (2020)	Biomedical CQA	7.3K	7.3K	https://huggingface.co/datasets/medical_dialog	Given a question with context from online medical forums, retrieve responses that best answer the question
DrugBank (2018)	Biomedical Entity Linking	4.1K	4.1K	https://go.drugbank.com/	Given a drug, retrieve passages for its definition
MeSH (2000)	Biomedical Entity Linking	29.6K	29.6K	https://www.nlm.nih.gov/databases/download/mesh.html	Given a concept, retrieve passages for its definition
RELISH (2023; 2019)	Biomedical Paper Recommendation	3.2K	191.2K	https://huggingface.co/datasets/allenai/scirepeval/viewer/relish	Given an article, retrieve Pubmed articles that are relevant to this article

Table 10: Evaluation datasets for biomedical text representation tasks and retrieval-oriented downstream applications.

plex multi-stage training paradigm that first pretrains on large-scale weakly-supervised text pairs and then fine-tunes on several labeled datasets.

- **BGE-Large** (Chen et al., 2024) is a dense retrieval model (335M) that uses graph-based embedding techniques and a multi-stage training paradigm similar to E5 (Wang et al., 2022b).
- **LLaRA** (Li et al., 2023a) is a post-hoc adaptation of LLMs for dense retrieval (7B) that uses LLM-generated text embeddings to reconstruct input sentence tokens and predict next sentence tokens.
- **RepLLaMA** (Ma et al., 2023) is a dense retriever (7B) that fine-tunes the LLaMA model for effective representation learning in passage and document retrieval using MS MARCO (Bajaj et al., 2016).
- **LLM2Vec** (BehnamGhader et al., 2024) is an unsupervised approach that transforms LLMs into text encoders by enabling bidirectional attention via masked next token prediction and adopts unsupervised contrastive learning for sequence representation learning.
- **E5-Mistral** (Wang et al., 2024) is an enhanced version of the E5 (Wang et al., 2022b) that incorporates synthetic data generated by LLMs for a diverse range of text embedding tasks. We consider E5-Mistral (7B) as a concurrent work and report its performance for reference only.
- **CPT-text** (Neelakantan et al., 2022) is a dense retrieval model pre-trained on web-scale data. We only consider its performance as a reference rather than a fair comparison due to its large size, as it is initialized from GPT-3 (Brown et al., 2020) with 175B parameters.

C.2 Baselines for Retrieval-Oriented Downstream Applications

In experiments for retrieval-oriented downstream applications, we only compare BMRETRIEVER

to the strongest, most relevant, and fair baselines, including: (1) **Base Size** (<1B): Dragon (Lin et al., 2023), MedCPT (Jin et al., 2023), and E5-Large-v2 (Wang et al., 2022b); (2) **Large Size** (1B-5B): InstructOR (Su et al., 2023) and SGPT-2.7B (Muennighoff, 2022); and (3) **XL Size** (>5B): E5-Mistral (Wang et al., 2024).

D Cosine Similarity v.s. Dot Product

We explore different objectives for embedding similarity, namely dot product and cosine similarity. From the experimental results in Figure 5, we empirically observe that the dot product could achieve a better empirical performance. Thus, we choose to use dot product by default as our similarity metrics.

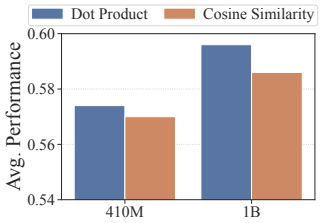
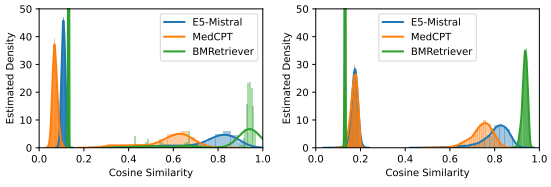


Figure 5: Comparison of performance using dot product and cosine similarity.

E Similarity Score

Figure 6 depicts the distributions of cosine similarity scores for positive and negative embedding pairs across two datasets. The left side displays the similarity distributions for negative examples, while the right side shows the distributions for positive examples. These figures illustrate that BMRETRIEVER exhibits a larger separation between positive and negative examples, showing its enhanced ability to effectively retrieve relevant passages.



(a) SciFact (b) iCliniq

Figure 6: The cosine similarity on positive pair embeddings and negative pair embeddings.

F Efficiency

Table 11 exhibits the document encoding speed and retrieval latency of BMRETRIEVER and baseline dense retrieval models. While BMRETRIEVER introduces additional encoding latency compared to BERT-based retrievers, we do not incorporate significant overhead when compared to baselines of similar model size.

Models	Size	Document Encoding Speed (# docs / s / GPU)	Retrieval Latency (ms)
MedCPT (2023)	220M	1390.1	11.6
InstructOR (2023)	1.5B	181.2	14.6
SGPT (2022)	2.7B	98.5	35.5
E5-Mistral* (2024)	7B	51.8	58.6
BMRETRIEVER	410M	471.2	14.6
BMRETRIEVER	1B	194.0	28.6
BMRETRIEVER	2B	166.2	28.6
BMRETRIEVER	7B	51.8	58.6

Table 11: Time complexity of BMRETRIEVER.