DL-DPR: Document-level Dense Passage Retrieval for Efficient Question Answering

Anonymous ACL submission

Abstract

In the continuously evolving field of Natural Language Processing (NLP), we introduce a nuanced problem: Document-Level Dense Passage Retrieval (DL-DPR). The specialized task of extracting relevant passages from within individual, often complex, documents has not been adequately addressed, with prevalent dense retrieval methods primarily tailored for broader, corpus-level contexts. This identified gap, where the intricacies and specificities of single-document analysis are often overlooked, motivates our research. We propose a novel approach, embedding a contrastive finetuning method coupled with the augmentation of datasets through queries generated by Large Language Models (LLMs). This fusion of techniques is meticulously designed to finetune dense retrieval methods for the unique challenges presented by DL-DPR. Our approach, when subjected to rigorous evaluation on multiple benchmark datasets and metrics like top-k retrieval accuracy and MRR@10, exhibits a marked enhancement in performance. The findings not only validate our method but also underscore the untapped potentials of refining and adapting existing dense retrieval technologies for specialized tasks. This study, thus, serves as both an introduction and a significant contribution to this intricate sub-domain of NLP, promising enhanced precision and efficiency in information extraction from detailed and lengthy documents.

1 Introduction

Passage retrieval, a cornerstone in fields ranging from ad-hoc information retrieval to retrieval-augmented generation (Lewis et al., 2020), open-domain question answering (Karpukhin et al., 2020), and fact verification (Thorne et al., 2018), is undergoing a paradigm shift. While the traditional sparse retrieval techniques like BM25 (Robertson and Zaragoza, 2009) have stood the test of time, the emergence of large-scale pre-trained language models (Radford and Narasimhan, 2018; Devlin et al., 2019; Liu et al., 2019; He et al., 2020; Beltagy et al., 2020) has catalyzed a transition towards neural dense retrieval methods (Karpukhin et al., 2020; Xiong et al., 2020). These methods, adept at projecting both queries and passages into a low-dimensional vector space, calculate relevance through dot product or cosine similarity, marking an evolution in the passage retrieval landscape.

However, an overlooked yet significant domain is the application of dense retrieval methods at the document level, a realm distinct from the traditional corpus-wide application. In this study, we introduce and delve into Document-Level Dense Passage Retrieval (DL-DPR), characterized by the extraction of contextually relevant passages from specific, individual documents. This nuanced task is distinguished by its focus and precision, addressing the need for targeted information retrieval within a given document's confines - a scenario commonly encountered in legal, medical, academic, and business settings.

The motivation for this research is rooted in the observed limitations of current dense retrieval methods when applied to the document-level context. While adept at corpus-wide tasks, these models exhibit suboptimal performance in the face of the unique challenges posed by DL-DPR. The complexity and context-specific nuances of individual documents necessitate a tailored approach, driving our exploration into optimizing dense retrieval methods for this specific application.

We bridge this identified gap with a two-pronged strategy. First, we conduct a comprehensive evaluation of existing dense retrieval methods within the DL-DPR context, unveiling their strengths and areas for improvement. This baseline analysis serves as a foundation for our second initiative - the introduction of a novel contrastive fine-tuning method. Recognizing the data constraints inherent in current datasets, characterized by the limited availability of passage-level queries, we leverage Large Language Models (LLMs) for query generation, enriching the datasets to facilitate an effective model optimization process. Our code base is available at <redacted> [1].

2 Related Work

The discipline of information retrieval (IR) aspires to locate pertinent information in response to a given ad-hoc query, serving as the backbone of contemporary search engines. (Kobayashi and Takeda, 2000; Manning, 2009; Chowdhury, 2010) In recent times, the focus in IR has started to transition from traditional BM25-based retrieval methods using an inverted index to more innovative dense retrieval techniques (Hofstätter et al., 2022). While BM25 retrieval is characterized by its efficiency and interpretability, it struggles to bridge the lexical mismatch between queries and passages. Efforts have been made to ameliorate this issue via approaches like document expansion (Tao et al., 2006) and query expansion (Carpineto and Romano, 2012; Azad and Deepak, 2019). In stark contrast, dense retrieval techniques, such as DSSM (Huang et al., 2013), C-DSSM (Shen et al., 2014), and DPR (Karpukhin et al., 2020), opt to map queries and passages into a shared low-dimensional vector space, promoting semantic matching. The employment of a bi-encoder architecture (Humeau et al., 2020), drawing from pre-trained language models, has become prevalent for first-stage retrieval in knowledge-intensive endeavors (Karpukhin et al., 2020; Wang et al., 2022), spanning from open-domain question answering to fact verification tasks. The search for close matches can be performed efficiently using approximate nearest neighbor (ANN) algorithms (Aumüller et al., 2020), such as HNSW (Malkov and Yashunin, 2018).

The application of contrastive objectives in training dense retrieval models has emerged as an influential approach, with the potential to augment retrieval effectiveness by influencing representation learning. By creating an embedding space where similar instances are drawn closer and dissimilar instances are distanced, the effectiveness of retrieval tasks can be enhanced. This technique has found notable success in Dense Passage Retrieval studies (Karpukhin et al., 2020) where models are trained to maximize the similarity between relevant queries and passages, while minimizing the similarity with irrelevant ones. Further applications include SimCSE(Gao et al., 2021) where contrastive learning was employed for learning sentence embeddings, and ANCE (Xiong et al., 2020) that leverages approximate nearest neighbor negative contrastive learning for dense text retrieval. Contriever (Izacard et al., 2021) explores the limits of contrastive learning as a way to train unsupervised dense retrievers and shows that it leads to strong performance in various retrieval settings, while the CLIP model (Radford et al., 2021) incorporated a contrastive objective to learn to comprehend and generate images and text simultaneously. SimLM (Wang et al., 2022) also utilizes a contrastive objective for self-supervised pre-training method for dense passage retrieval.

Data augmentation, a critical strategy in machine learning (Shorten and Khoshgoftaar, 2019; Feng et al., 2021), has been extensively employed in dense retrieval, improving model performance by expanding and diversifying the training data. Initially, studies like Dense Passage Retrieval (DPR) (Karpukhin et al., 2020) and work by Xiong et al. relied on traditional techniques such as negative sampling and in-batch negatives, creating negative instances from irrelevant passages or passages from other queries within the same batch. However, more recently, the potential of Large Language Models (LLMs) has been harnessed for data augmentation in information retrieval tasks, as evident in several significant studies. For example, InPars (Bonifacio et al., 2022), Doc2Query (Gospodinov et al., 2023), and Promptagator (Dai et al., 2022) have leveraged LLMs to generate new queries and expand documents. Additional works such as UDEG (Jeong et al., 2021) and Query2doc (Wang et al., 2023) further demonstrated the applicability of LLMs for query and document expansion. These advancements underscore the utility of LLMs in enriching datasets for retrieval tasks, thus broadening the scope of data augmentation techniques beyond the creation of negative examples.

3 Methodology

3.1 Problem Definition

In Document-Level Dense Passage Retrieval (DL-DPR), our objective is to identify the top-k passages from a given document D_i that are most relevant to a given query q. We formally define it as

[1] For review purposes, our code base is available as part of supplementary material.

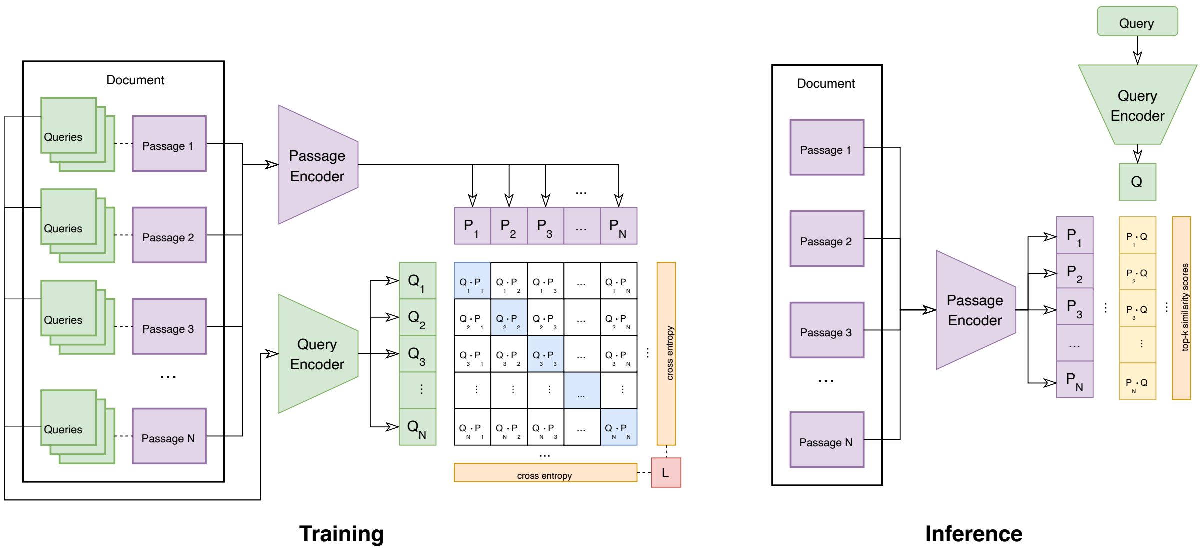


Figure 1: Overview of Document-level Dense Passage Retrieval (DL-DPR). (Left) Training step using the contrastive objective on a single document. (Right) Inference step to retrieve top-k relevant passages from document.

follows:

$$\mathbf{P}^* = \mathrm{Top}_k\{\mathrm{sim}(q,p) \,|\, p \in P_i\}$$

where $k << |D|$ (the number of passages in document), each $p \in P_i$ is a passage in document D_i and $\mathrm{sim}(q,p)$ calculates the similarity score between the query q and passage p, defined as:

$$\mathrm{sim}(q,p) = E_Q(q)^\top E_P(p).$$

Here, E_Q and E_P denote the query and passage encoders that map their inputs into a shared d-dimensional space.

3.2 Training Objective

We consider a collection of training documents $D = \{D_1, D_2, \ldots, D_N\}$, each associated with a set of passages P_i and queries Q_i. A crucial aspect of our methodology is that each training batch is optimized using query-passage pairs from a single document. This constraint is pivotal for tailoring the model to document-level contexts. From a given document D_i, we select m distinct query-passage pairs $(q_1, p_1), \ldots, (q_m, p_m)$.

We initiate the training with pre-trained encoders E_Q and E_P that are already adept at open-domain passage retrieval. The objective then is to fine-tune these base models, namely the query and passage encoders, to specialize in document-level retrieval. The embeddings for queries and passages are computed as:

$$Q_e = [E_Q(q_1), E_Q(q_2), \ldots, E_Q(q_m)],$$
$$P_e = [E_P(p_1), E_P(p_2), \ldots, E_P(p_m)].$$

The similarity matrix $\mathbf{S}$ is computed as the scaled dot-product of these embeddings, adjusted by an exponential temperature parameter t:

$$\mathbf{S} = Q_e \cdot P_e^\top \times e^t.$$

Cross-entropy losses for the queries and passages are then calculated:

$$\mathcal{L}_q = -\sum_{i=1}^{m} y_i \log(f(\mathbf{S}_i)),$$

$$\mathcal{L}_p = -\sum_{i=1}^{m} y_i \log(f(\mathbf{S}_i^\top)),$$

with each element being defined as follows:

- $\mathbf{S}_i$ and $\mathbf{S}_i^\top$ are the i-th row of the similarity matrix and its transpose respectively,
- y_i represents the true label i, encoded as one-hot vector,
- $f(\cdot)$ is the softmax function, transforming the similarity scores into probabilities.

The symmetric contrastive loss $\mathcal{L}$ is then calculated as the average of $\mathcal{L}_q$ and $\mathcal{L}_p$:

$$\mathcal{L} = \frac{\mathcal{L}_q + \mathcal{L}_p}{2}.$$

This objective refines the model to efficiently discern the relevancy of passages in response to a given query, tailored specifically for individual document contexts, thereby boosting its document-level retrieval performance. Figure 1 depicts a schematic representation of the training and inference setups for document-level dense passage retrieval (DL-DPR).

3.3 Question-Generation-based Data-Augmentation

The paradigm of our proposed methodology emphasizes the retrieval of positive question-passage pairs from a singular document at each step. The Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2018) fits well with this requirement due to its rich distribution of question-associated passages at the document level. However, certain large-scale datasets such as Natural Questions (NQ) (Kwiatkowski et al., 2019), NewsQA (Trischler et al., 2017), despite their extensive content, pose a challenge with their sparse provision of question-passage pairs at the document level. In these datasets, an entire document might be linked with only one or two questions, leaving a substantial number of passages without any corresponding questions. Our contrastive objective thrives on a wider set of pairs at the document level to enhance performance. This sparsity of question-passage pairs in these datasets inhibits the optimal operation of our method.

Inspired by recent works that have effectively employed Large Language Models (LLMs) for data augmentation, we decided to incorporate the same approach for our use-case. Specifically, we utilize Flan-T5 (Chung et al., 2022), a publicly available instruction-tuned LLM, to enrich our datasets. We prompt Flan-T5 to generate questions for passages that lack associated questions in the dataset. This method significantly alleviates data sparsity and enriches the documents with a larger set of question-passage pairs, thereby enhancing the performance of our model.

3.4 Fine-tuning Process

Our approach begins by utilizing the pre-trained models specifically designed for dense passage retrieval tasks. These models are trained on large-scale datasets like MS MARCO (Bajaj et al., 2016), Natural Questions (NQ) (Kwiatkowski et al., 2019) and CCNet (Wenzek et al., 2020). Subsequently, these models undergo a fine-tuning process using our unique contrastive objective on the enriched training data, which has been augmented using the aforementioned question-generation process. The fine-tuning phase consists of optimizing the parameters of the model to minimize the discrepancy between embeddings of relevant question-passage pairs using the symmetric cross-entropy loss.

4 Experiments

In this section, we describe our experimental setup, including the datasets we used, the baseline models we compared with, and the evaluation metrics. Furthermore, we delve into the specifics of our fine-tuning process and the implementation details of our proposed method.

4.1 Datasets

In our research, we utilize a diverse set of datasets including widely used English language benchmarks SQuAD (Rajpurkar et al., 2018), Natural Questions (NQ) (Kwiatkowski et al., 2019), and NewsQA (Trischler et al., 2017), along with several non-English datasets SQuAD-es (Carrino et al., 2019), SQuAD-bn (Tasmiah Tahsin Mayeesha and Rahman, 2021), FQuAD (Martin et al., 2020), KorQuAD (Lim et al., 2019), and ARCD (Mozannar et al., 2019) catering to Spanish, Bengali, French, Korean, and Arabic languages respectively. Our method is applied to fine-tune dense retrieval models on these datasets for the document-level passage retrieval task, offering a thorough evaluation across different languages and contexts. Each dataset, with its linguistic and contextual nuances, provides a distinct set of challenges, enabling a comprehensive appraisal of our method's versatility.

SQuAD is a reading comprehension dataset on a set of Wikipedia articles, with each article comprising several paragraphs. Every paragraph is paired with a set of questions that pertain specifically to the information contained within that paragraph. Unlike SQuAD, NQ does not explicitly provide paragraphs for each document. Instead, we use non-overlapping long answer candidates as paragraphs for each document. It is noteworthy that only a few questions are linked to the entire document, leaving many paragraphs without associated questions. In NewsQA, we are not provided with explicit paragraphs for each news story. To construct a comparable structure, we implement a rule-based heuristic to merge several sentences into paragraphs of less than 128 words. However, similar to NQ, not all paragraphs are paired with questions in this dataset. Note that we use the development set as a test set for SQuAD and NQ as their test sets are not available. For NQ, we use a subset of 9173 documents for data-augmentation and training. Table 1 shows the number of paragraphs and documents in the training and test sets for all the datasets.

Dataset	Train		Test	
	Docs	Paras	Docs	Paras
SQuAD	442	19,035	35	1,204
NewsQA	11,469	66,042	634	3,697
NQ	9,173	303,579	3,486	125,601
SQuAD-es	442	18,896	48	2,067
SQuAD-bn	241	10,289	61	2,633
FQuAD	117	4,921	18	768
KorQuAD	1,420	9,681	140	964
ARCD	77	231	78	234

Table 1: Summary of datasets and their train-test splits.

Our proposed method operates by extracting positive question-passage pairs from a single document per training batch. Therefore, if most paragraphs lack associated questions, it would significantly reduce the batch-size, making it unsuitable for our fine-tuning method. To address this, we employ data augmentation on the training data through question generation. For data augmentation, we utilized passages that do not have linked questions within the same datasets. We generated questions for these passages using the Flan-T5 LLM (Chung et al., 2022), expanding our training set significantly. Note that we do not employ data augmentation for SQuAD and other non-English datasets, as most of the paragraphs in these datasets have multiple related questions available. Table 2 shows the number of questions before and after data augmentation in our training datasets.

Dataset	Original	Augmented
NewsQA	68,009	338,047
NQ	10,000	1,099,373

Table 2: Total number of questions in training datasets before and after data augmentation.

4.2 Baseline Models

For our study, we select an array of retrieval models as our baselines, each embodying different strategies and techniques prevalent in the field of dense passage retrieval. We select BM25 (Robertson and Zaragoza, 2009) as our sparse retrieval baseline, given its wide acceptance and use in the field of information retrieval. In the dense retrieval domain, we initially consider DPR (Karpukhin et al., 2020), which laid the foundation for dense passage retrieval. To provide a more comprehensive evaluation, we include recently developed models like RocketQA (Qu et al., 2021), PAIR (Ren et al., 2021a), and RocketQA v2 (Ren et al., 2021b), each introducing unique approaches to dense passage retrieval. Moreover, we consider state-of-the-art models Contriever (Izacard et al., 2021) and SimLM (Wang et al., 2022), which represent the most recent advances in this field. Notably, we focus on evaluating the retrieval component of these DPR models, not the re-ranking component. We assess the performance of these diverse retrieval models specifically in the context of our document-level passage retrieval task. For the fine-tuning experiments, we select DPR, SimLM, and Contriever models. Subsequently, by fine-tuning these baselines with our method, we aim to offer a comprehensive evaluation of our approach's performance and its potential improvements to the field.

4.3 Evaluation Metrics

We use standard evaluation metrics to evaluate the performance of our model, namely Top-k Retrieval Accuracy (Karpukhin et al., 2020) and Mean Reciprocal Rank (MRR). In our datasets, only one relevant passage corresponds to a given question. Thus, we do not use metrics like normalized discounted cumulative gain (NDCG) and mean average precision (MAP) which are generally used to evaluate retrieval systems where multiple passages may be relevant to a single question. For all datasets, we compute Top-1, Top-3, and Top-5 retrieval accuracies and MRR@10.

4.4 Implementation Details

For our experiments, we utilize PyTorch (Paszke et al., 2017) deep learning framework, the Hugging Face Transformers library (Wolf et al., 2020), and the Sentence Transformers library (Reimers and Gurevych, 2019). We use the pre-trained models available in these libraries as the starting point for our fine-tuning process. For each model, we run our experiments on a single NVIDIA Tesla V100 GPU with 32GB memory. We conduct a grid-search to identify optimal hyperparameters for our method, including learning rate, weight decay, and batch size. We use Adam (Kingma and Ba, 2017) optimizer for training. To prevent overfitting, we apply early-stopping during the fine-tuning process.

For our data augmentation process, we use Flan-T5 LLM (Chung et al., 2022) with a maximum sequence length of 256 tokens to generate questions. To prevent the generation of irrelevant or nonsensical questions, we carefully crafted the prompt, providing Flan-T5 with adequate context and setting clear expectations for the output with appropriate stopping conditions. To expedite the data-augmentation process and reduce the inference cost, we perform batch inference for each paragraph to generate multiple questions in a single response in order to avoid making multiple requests for the same paragraph. We experimented with varied prompts to find the best prompt for our use-case. For our final experiments we used the following prompt to augment data: *Generate a question which can be answered using given context: <paragraph>*. We set the inference-time parameters for the Flan-T5 LLM as follows: temperature as 1, repetition penalty as 1.05 and number of beams as 9.

4.5 Evaluation Setup

The evaluation process is designed to measure the efficacy of retrieval model in determining the most relevant passage for each question within a given document. Here is the step-by-step procedure:

1. **Embedding Computation:** We start by computing embeddings for all passages and questions within each document. These embeddings are generated using our trained model, encapsulating the contextual intricacies of each text piece.

2. **Similarity Scoring:** Next, for each question, we measure its similarity scores with all the passages within the same document. This process involves calculating the dot product or cosine similarity between the question and passage embeddings, reflecting how semantically close they are.

3. **Passage Ranking:** Using the computed similarity scores, we rank all the passages within the document for each question. The passage with the highest similarity score for a given question receives the highest rank.

4. **Performance Metrics:** After the ranking process, we compute aforementioned evaluation metrics based on these rankings to measure the model's ability in identifying the correct passage in response to each question.

By following this setup, we are able to evaluate the model's performance in associating relevant passages to the corresponding questions within a document. The goal is to ensure that the correct passage is ranked as high as possible for each question, thereby demonstrating the model's effectiveness. We leave the end-to-end evaluation of question-answering accuracy as future work.

5 Results and Discussion

5.1 Results

First, we benchmark out-of-the-box performance of various pre-trained dense retrieval models applied to our document-level passage retrieval task. The detailed results are provided in Table 3. The top-performing models from DPR, SimLM and Contriever methods are utilized as a baseline for evaluating the impact of our proposed fine-tuning method.

Tables 4, 5, 6, 7 present the results of our method on SQuAD, NewsQA, NQ and other non-English datasets respectively. For each method, the first row shows the retrieval performance of a pre-trained model before fine-tuning and the second row shows the retrieval performance after performing fine-tuning with our method on a given dataset. This allows us to isolate and assess the contribution of our proposed method in improving the model's efficacy. For the non-English datasets, we utilize the multi-lingual dense retriever - the Contriever model pre-trained on CC-net (29 languages) and MS-MARCO datasets.

Figure 2 offers insight into an ablation study we conducted, focusing on the configuration of our fine-tuning method. In our approach, pairs in a given training batch must originate from a single document. We conducted this study to highlight the importance of this constraint. Specifically, we compare results of our method with an alternative scenario - Mixed-Batch, where we loosen this constraint by allowing pairs to come from different documents. The results of this study clearly illustrate the superiority of our method, as it outperforms the alternative configuration.

We also conduct additional experiments to study the impact of document length on the performance of our approach. The detailed report of the results is provided in Appendix A.

Model	SQuAD			NewsQA			NQ			
	Top-1	Top-3	MRR	Top-1	Top-3	MRR	Top-1	Top-3	Top-5	MRR
BM25	68.7	83.1	76.8	57.8	83.9	73.3	19.9	40.1	51.8	33.7
DPR										
Single-NQ	47.7	70.6	61.1	40.1	73.4	59.6	39.7	64.6	75.7	54.8
Multi	51.2	72.1	63.6	49.7	82.3	67.4	40.0	64.6	75.5	55.1
RocketQA										
MS MARCO	67.0	84.4	76.6	61.8	88.6	76.0	41.6	67.3	77.9	56.8
NQ	64.6	83.0	75.0	62.7	89.8	76.9	41.1	69.4	81.0	57.8
PAIR										
MS MARCO	66.6	83.4	76.1	61.4	88.0	75.7	40.5	67.0	77.3	56.1
NQ	61.0	79.7	71.8	58.3	87.1	73.5	42.0	68.8	79.5	57.9
RocketQA v2										
MS MARCO	64.0	82.1	74.1	61.0	88.6	75.5	41.1	66.9	76.5	56.3
NQ	57.9	78.8	69.7	57.9	86.6	73.1	45.3	70.5	79.8	59.9
Contriever										
CC-net & Wiki *pt*	68.1	86.1	78.2	53.1	85.2	70.1	18.0	42.2	55.3	34.0
+ MS MARCO ft	75.2	90.7	83.5	65.2	91.0	78.7	39.0	64.7	75.6	54.6
CC-net 29 lang. *pt*	60.5	81.5	72.4	46.6	81.2	65.3	17.9	40.3	53.2	33.1
+ MS MARCO *ft*	73.4	89.7	82.2	63.6	89.9	77.3	33.3	61.2	72.8	50.1
SimLM										
MS MARCO *pt*	51.2	71.8	63.5	44.4	81.2	64.3	4.8	17.2	29.6	15.9
+ finetune & distill	68.0	84.4	77.2	63.6	88.7	77.0	41.0	66.0	76.5	56.2
Wiki *pt*	47.1	66.9	59.2	45.9	79.9	64.4	9.6	27.0	39.8	22.6

Table 3: Benchmarking various dense retrieval models for document-level passage retrieval task on SQuAD, NewsQA and NQ datasets. Top-k retrieval accuracy and MRR@10 metrics are shown for each model. Note that *pt* and *ft* stand for *pre-training* and *fine-tuning* respectively.

Model	Top-1	Top-3	Top-5	MRR@10
DPR	51.2	72.1	80.2	63.6
+ DL-DPR	55.2	75.6	83.3	67.1
SimLM	68.0	84.4	89.5	77.2
+ DL-DPR	71.1	86.9	91.8	79.9
Contriever	75.2	90.7	94.5	83.5
+ DL-DPR	81.3	94.2	96.8	88.0

Table 4: DL-DPR fine-tuning results on SQuAD.

Model	Top-1	Top-3	Top-5	MRR@10
DPR	40.7	65.4	76.2	55.8
+ DL-DPR	43.3	69.0	79.9	58.6
SimLM	42.1	67.0	77.4	57.1
+ DL-DPR	43.4	69.6	79.9	58.6
Contriever	39.0	64.7	75.6	54.6
+ DL-DPR	40.8	67.8	78.1	56.7

Table 6: DL-DPR fine-tuning results on NQ.

Model	Top-1	Top-3	Top-5	MRR@10
DPR	49.7	82.3	93.0	67.4
+ DL-DPR	59.0	86.7	95.1	73.9
SimLM	63.6	88.7	95.5	77.0
+ DL-DPR	67.3	90.1	96.5	79.5
Contriever	65.2	91.0	97.1	78.7
+ DL-DPR	71.6	93.3	97.7	82.7

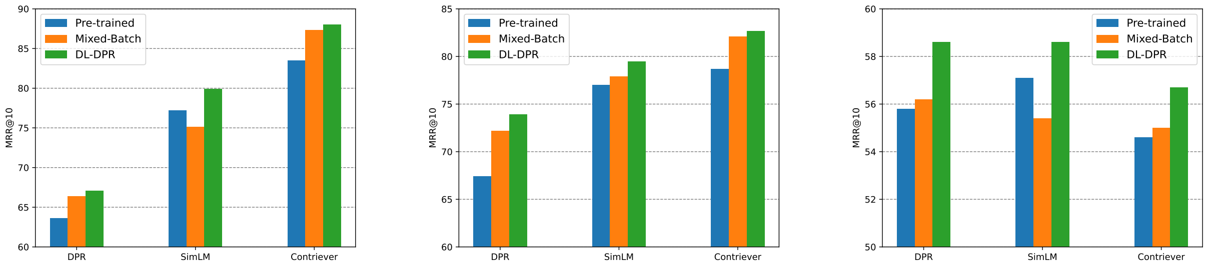
Table 5: DL-DPR fine-tuning results on NewsQA.

Dataset	Top-1	Top-3	Top-5	MRR@10
SQuAD-es	70.1	87.1	91.7	79.4
	75.2	89.8	93.8	83.3
SQuAD-bn	56.7	74.9	81.5	67.3
	61.7	79.5	85.5	71.8
FQuAD	56.0	75.5	82.0	67.1
	63.2	81.0	87.0	73.4
KorQuAD	77.9	91.5	95.0	85.3
	82.9	94.1	96.4	88.9
ARCD	75.4	-	-	86.4
	80.1	-	-	89.1

Table 7: Results on non-English datasets with multilingual Contriever before/after DL-DPR fine-tuning.

5.2 Discussion

Our experimental results underline the significant contribution of our fine-tuning method to the effectiveness of dense retrieval models in document-level passage retrieval tasks. The ablation study elucidates the critical importance of extracting positive question-passage pairs from a single document per training batch. This configuration manifests in a substantial improvement in the model's performance when compared to allowing pairs to be drawn from different documents. This finding im-



(a) SQuAD (b) NewsQA (c) NQ

Figure 2: Ablation study on the effect of batch composition in the fine-tuning process. Each sub-figure presents the performance comparison (MRR@10) of pre-trained models, Mixed-Batch, and DL-DPR on different datasets.

plies that preserving the document-level context is crucial for the model to better understand and infer the relevance of the passages to the questions.

Note that the NQ dataset contains on an average longer documents than other datasets and the number of documents in the test set of NQ is much bigger compared to any other dataset in our study, which makes it more challenging compared to SQuAD and NewsQA.

When comparing our results with the baseline models, it is evident that our method leads to an appreciable enhancement in evaluation metrics across all the datasets. Even with the state-of-the-art models like Contriever and SimLM, our method fine-tunes them to achieve superior performance. This signifies the potential of our fine-tuning approach to serve as a novel strategy in the ongoing evolution of dense passage retrieval techniques.

On the non-English datasets too, we observe a consistent and considerable improvement in all metrics after fine-tuning with our method. This demonstrates the soundness of our approach across different languages.

Our experiments to study the impact of document length (Appendix A) on the document level retrieval suggest that the increase in the number of passages leads to gradual decrease in the performance. This is understandable as the increase in the number of passages increases the search space as well. The key point to note is that, our fine-tuning approach consistently surpasses the baseline, underscoring its effectiveness across a diverse range of document lengths.

5.3 Future Work

Looking forward, we envision abundant opportunities for enhancing our methodology and expanding its applications. Firstly, we aim to refine the question generation process by probing more sophisticated techniques that could yield better quality questions, thus amplifying the efficacy of our data augmentation. This might involve delving into advanced fine-tuning techniques of language models or harnessing novel developments in controllable text generation.

Secondly, to scale up our fine-tuning process for larger datasets and better computational efficiency, we propose exploring strategies that are adept at identifying and ranking multiple relevant passages for a given question. This would more accurately reflect real-world information retrieval scenarios.

Finally, the relevance and impact of our approach extend to domain-specific tasks. Particularly in areas such as legal, academic, or medical fields, where document-level retrieval can significantly aid in information extraction and comprehension. This potential for domain-specific applicability emphasizes the robustness and versatility of our approach, thereby inspiring us to continually push its boundaries in future explorations.

6 Conclusion

This study introduces the problem of document-level dense passage retrieval (DL-DPR) and proposes a novel fine-tuning approach, leveraging a contrastive objective with a constraint to limit query-passage pairs in a batch from the same document. Our method addresses the challenge of sparse query-passage pairs in large-scale datasets by employing LLM for question-generation-based data augmentation, thereby enriching the training set. Through comprehensive experiments, our approach consistently surpasses the efficacy of traditional methods across various datasets and metrics. The promise of this method opens up potential future avenues for its application across a broader range of information retrieval tasks, establishing the state-of-the-art in this domain.

Limitations

Despite the encouraging results of our approach, it is not without its limitations. First and foremost, our method relies heavily on the quality of questions generated by the LLM during data augmentation. While it typically generates questions that are coherent and contextually sensible, there could be instances where the questions lack relevance to the corresponding passage or fail to accurately reflect its content. Moreover, the process of generating questions using an LLM can be time-consuming and computationally costly, which could pose challenges for large-scale applications. Second, our current implementation assumes a single relevant passage per question. The future works can investigate ways to adapt it for the real-world scenarios where multiple passages within a document may provide valuable context or insights in response to a given question. These limitations offer avenues for potential future work to further enhance the applicability and effectiveness of our method.

References

Martin Aumüller, Erik Bernhardsson, and Alexander Faithfull. 2020. Ann-benchmarks: A benchmarking tool for approximate nearest neighbor algorithms. *Information Systems*, 87:101374.

Hiteshwar Kumar Azad and Akshay Deepak. 2019. Query expansion techniques for information retrieval: a survey. *Information Processing & Management*, 56(5):1698–1735.

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

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.

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

Claudio Carpineto and Giovanni Romano. 2012. A survey of automatic query expansion in information retrieval. *ACM Comput. Surv.*, 44(1).

Casimiro Pio Carrino, Marta R. Costa-jussà, and José A. R. Fonollosa. 2019. Automatic spanish translation of the squad dataset for multilingual question answering.

Gobinda G Chowdhury. 2010. *Introduction to modern information retrieval*. Facet publishing.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Zhuyun Dai, Vincent Y Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B Hall, and Ming-Wei Chang. 2022. Promptagator: Few-shot dense retrieval from 8 examples. *arXiv preprint arXiv:2209.11755*.

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

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mitko Gospodinov, Sean MacAvaney, and Craig Macdonald. 2023. Doc2query: When less is more. *arXiv preprint arXiv:2301.03266*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention.

Sebastian Hofstätter, Nick Craswell, Bhaskar Mitra, Hamed Zamani, and Allan Hanbury. 2022. Are we there yet? a decision framework for replacing term based retrieval with dense retrieval systems.

Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. ACM International Conference on Information and Knowledge Management (CIKM).

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *International Conference on Learning Representations*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning.

Soyeong Jeong, Jinheon Baek, ChaeHun Park, and Jong Park. 2021. Unsupervised document expansion for information retrieval with stochastic text generation. In *Proceedings of the Second Workshop on Scholarly Document Processing*, pages 7–17, Online. Association for Computational Linguistics.

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.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Mei Kobayashi and Koichi Takeda. 2000. Information retrieval on the web. *ACM computing surveys (CSUR)*, 32(2):144–173.

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.

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.

Seungyoung Lim, Myungji Kim, and Jooyoul Lee. 2019. Korquad1.0: Korean QA dataset for machine reading comprehension. *CoRR*, abs/1909.07005.

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

Yu A Malkov and Dmitry A Yashunin. 2018. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):824–836.

Christopher D Manning. 2009. *An introduction to information retrieval*. Cambridge university press.

d'Hoffschmidt Martin, Vidal Maxime, Belblidia Wacim, and Brendlé Tom. 2020. FQuAD: French Question Answering Dataset. *arXiv e-prints*, page arXiv:2002.06071.

Hussein Mozannar, Elie Maamary, Karl El Hajal, and Hazem Hajj. 2019. Neural Arabic question answering. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 108–118, Florence, Italy. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Ruiyang Ren, Shangwen Lv, Yingqi Qu, Jing Liu, Wayne Xin Zhao, QiaoQiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021a. PAIR: Leveraging passage-centric similarity relation for improving dense passage retrieval. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2173–2183, Online. Association for Computational Linguistics.

Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, QiaoQiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021b. RocketQAv2: A joint training method for dense passage retrieval and passage re-ranking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2825–2835, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

S. Robertson and H. Zaragoza. 2009. *The Probabilistic Relevance Framework: BM25 and Beyond*. Foundations and trends in information retrieval. Now Publishers.

Yelong Shen, Xiaodong He, Jianfeng Gao, Li Deng, and Grégoire Mesnil. 2014. Learning semantic representations using convolutional neural networks for web search. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, page 373–374, New York, NY, USA. Association for Computing Machinery.

Connor Shorten and Taghi M Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48.

Tao Tao, Xuanhui Wang, Qiaozhu Mei, and ChengXiang Zhai. 2006. Language model information retrieval with document expansion. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 407–414, New York City, USA. Association for Computational Linguistics.

Abdullah Md Sarwar Tasmiah Tahsin Mayeesha and Rashedur M. Rahman. 2021. Deep learning based question answering system in bengali. *Journal of Information and Telecommunication*, 5(2):145–178.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018. The fact extraction and VERification (FEVER) shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Simlm: Pre-training with representation bottleneck for dense passage retrieval.

Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query expansion with large language models. *arXiv preprint arXiv:2303.07678*.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval.

A Impact of Document Length on Performance

We conduct supplementary experiments to study the influence of document length on the efficacy of our approach. The documents within the datasets were categorized based on the number of passages they contained as shown in Tables 8, 9, 10. We evaluated the performance metrics, namely MRR@10 and Top-1 retrieval accuracy, across these different categories to gain insights into the relationship between document length and retrieval accuracy. These results are provided in Tables 11, 12, 13, 14, 15, 16. We notice that, as the number of passages increase, the search space increases which leads to gradual decrease in metric value. Nevertheless, it is evident that our fine-tuning approach consistently surpasses the baseline, underscoring its effectiveness across a diverse range of document lengths.

#passages	count
20-24	10
25-29	5
30-34	2
35-39	7
40-44	5
45-49	6

Table 8: Distribution of documents according to passage count in SQuAD test set.

#passages	Count
1-50	2,741
51-100	536
101-150	138
151-200	50
201-250	15
250+	6

Table 9: Distribution of documents according to passage count in NQ test set.

#passages	Count
1-5	338
6-10	233
11-15	58
16-20	5

Table 10: Distribution of documents according to passage count in NewsQA test set.

Model	20-24	25-29	30-34	35-39	40-44	45-49
DPR	68.4	67.8	61.1	64.3	59.1	61.2
+ DL-DPR	72.4	69.5	61.5	48.5	63.1	65.2
SimLM	80.6	77.6	76.4	77.7	74.8	75.8
+ DL-DPR	83.7	80.6	78.3	80.5	76.8	80.0
Contriever	85.8	85.7	81.9	84.2	80.9	82.1
+ DL-DPR	90.1	88.9	86.3	87.1	84.7	85.4

Table 11: SQuAD - #passages vs MRR@10

Model	20-24	25-29	30-34	35-39	40-44	45-49
DPR	55.6	56.6	46.7	51.7	47.3	48.7
+ DL-DPR	60.6	58.4	48.3	57.0	51.8	52.6
SimLM	71.9	68.9	65.8	68.8	65.4	66.0
+ DL-DPR	75.7	71.7	67.9	72.3	67.2	68.6
Contriever	77.7	78.2	72.5	76.7	72.0	73.2
+ DL-DPR	84.1	82.7	78.3	80.3	77.3	77.7

Table 12: SQuAD - #passages vs Top-1 Accuracy

Model	1-50	51-100	101-150	151-200	201-250	251+
DPR	58.4	44.8	50.3	43.9	48.8	37.5
+ DL-DPR	62.0	44.7	46.7	47.7	48.0	50.0
SimLM	59.7	46.2	54.8	42.2	44.4	18.5
+ DL-DPR	60.5	48.0	52.4	40.4	43.3	22.2
Contriever	58.2	45.2	46.9	41.5	37.5	35.4
+ DL-DPR	59.5	45.9	50.6	46.6	43.5	35.4

Table 13: NQ - #passages vs MRR@10

Model	1-50	51-100	101-150	151-200	201-250	251+
DPR	42.8	31.7	36.2	30.0	40.0	33.3
+ DL-DPR	46.3	29.1	33.3	36.0	40.0	50.0
SimLM	44.2	32.8	43.5	30.0	33.3	16.7
+ DL-DPR	44.5	34.3	37.7	28.0	26.7	16.7
Contriever	42.2	31.1	32.6	28.0	26.7	33.3
+ DL-DPR	42.9	32.8	37.7	32.0	33.3	33.3

Table 14: NQ - #passages vs Top-1 Accuracy

Model	1-5	6-10	11-15	16-20
DPR	74.4	55.1	50.7	33.3
+ DL-DPR	80.5	62.6	56.9	43.2
SimLM	83.4	66.4	60.3	54.9
+ DL-DPR	85.7	69.3	65.5	51.8
Contriever	84.4	69.0	64.7	58.6
+ DL-DPR	87.3	73.9	68.6	72.1

Table 15: NewsQA - #passages vs MRR@10

Model	1-5	6-10	11-15	16-20
DPR	56.7	35.0	30.6	11.3
+ DL-DPR	66.5	44.5	40.5	28.0
SimLM	70.8	50.0	45.3	35.3
+ DL-DPR	74.7	53.7	51.6	26.0
Contriever	72.3	52.1	46.9	38.7
+ DL-DPR	77.3	59.0	54.0	58.7

Table 16: NewsQA - #passages vs Top-1 Accuracy