Questions Are All You Need to Train a Dense Passage Retriever

Devendra Singh Sachan[1,2], **Mike Lewis**[3], **Dani Yogatama**[4], **Luke Zettlemoyer**[3,5], **Joelle Pineau**[1,2,3], **Manzil Zaheer**[4]
[1]McGill University, Canada; [2]Mila - Quebec AI Institute, Canada; [3]Meta AI, USA; [4]Google DeepMind, USA; [5]University of Washington, USA
sachande@mila.quebec, {dyogatama, manzilzaheer}@google.com
{mikelewis,lsz,jpineau}@meta.com

Abstract

We introduce ART, a new corpus-level autoencoding approach for training dense retrieval models that does not require any labeled training data. Dense retrieval is a central challenge for open-domain tasks, such as Open QA, where state-of-the-art methods typically require large supervised datasets with custom hard-negative mining and denoising of positive examples. ART, in contrast, only requires access to unpaired inputs and outputs (e.g., questions and potential answer passages). It uses a new passage-retrieval autoencoding scheme, where (1) an input question is used to retrieve a set of evidence passages, and (2) the passages are then used to compute the probability of reconstructing the original question. Training for retrieval based on question reconstruction enables effective unsupervised learning of both passage and question encoders, which can be later incorporated into complete Open QA systems without any further finetuning. Extensive experiments demonstrate that ART obtains state-of-the-art results on multiple QA retrieval benchmarks with only generic initialization from a pre-trained language model, removing the need for labeled data and task-specific losses.[1]

1 Introduction

Dense passage retrieval methods (Karpukhin et al., 2020; Xiong et al., 2021), initialized with encoders such as BERT (Devlin et al., 2019) and trained using supervised contrastive losses (Oord et al., 2018), have surpassed the performance achieved by previously popular keyword-based approaches like BM25 (Robertson and Zaragoza, 2009). Such retrievers are core components in models for open-domain tasks, such as Open QA, where state-of-the-art methods typically require large supervised datasets with custom hard-negative mining and denoising of positive examples. In this paper, we introduce the first unsupervised method, based on a new corpus-level autoencoding approach, that can match or surpass strong supervised performance levels with no labeled training data or task-specific losses.

We propose ART, *Autoencoding-based Retriever Training*, which only assumes access to sets of unpaired questions and passages. Given an input question, ART first retrieves a small set of possible evidence passages. It then *reconstructs the original question* by attending to these passages (see Figure 1 for an overview). The key idea in ART is to consider the retrieved passages as a noisy representation of the original question and question reconstruction probability as a way of denoising that provides *soft-labels* for how likely each passage is to have been the correct result.

To bootstrap the training of a strong model, it is important to both have a strong initial retrieval model and to be able to compute reliable initial estimates of question reconstruction probability when conditioned on a (retrieved) passage. Although passage representations from BERT-style models are known to be reasonable retrieval baselines, it is less clear how to do zero-shot question generation. We use a generative pre-trained language model (PLM) and prompt it with the passage as input to generate the question tokens using teacher-forcing. As finetuning of the question-generation PLM is not needed, only the retrieval model, ART can use large PLMs and obtain accurate soft-label estimates of which passages are likely to be the highest quality.

The retriever is trained to penalize the divergence of a passage likelihood from its soft-label score. For example, if the question is ''Where is the bowling hall of fame located?'' as shown in Figure 1, then the training process

[1]Our code and model checkpoints are available at: https://github.com/DevSinghSachan/art.

Transactions of the Association for Computational Linguistics, vol. 11, pp. 600–616, 2023. https://doi.org/10.1162/tacl_a_00564
Action Editor: Jimmy Lin. Submission batch: 8/2022; Revision batch: 11/2022; Published 6/2023.

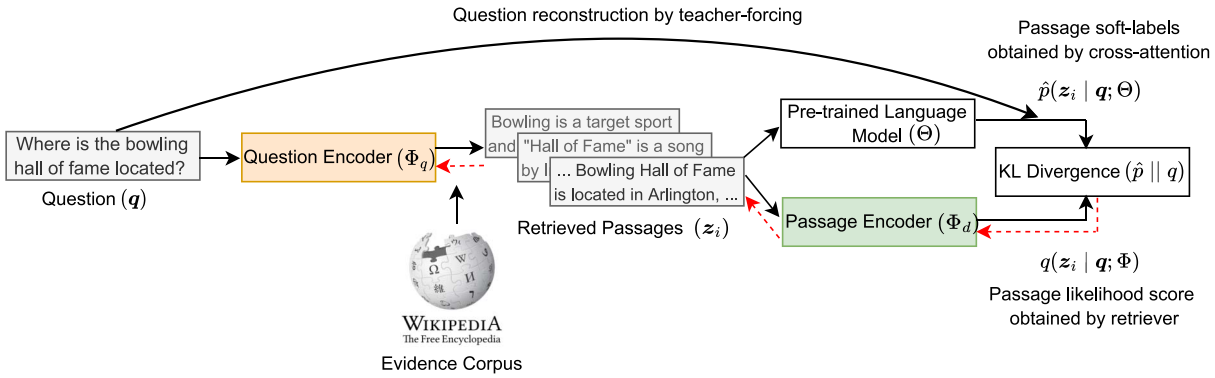

Figure 1: ART maximizes the retrieved passage likelihood computed from the dense retriever by considering the language model question reconstruction score conditioned on the passage as a *soft-label*. Colored blocks indicate trainable parameters. Red arrows show gradient flow during backpropagation.

will boost the retrieval likelihood of the passage "`Bowling Hall of Fame is located in Arlington,`" as it is relevant and would lead to a higher question reconstruction likelihood, while the likelihood of the passage "`Hall of Fame is a song by ...`" would be penalized as it is irrelevant. In this manner, the training process encourages correct retrieval results and vice-versa, leading to an iterative improvement in passage retrieval.

Comprehensive experiments on five benchmark QA datasets demonstrate the usefulness of our proposed training approach. By simply using questions from the training set, ART outperforms models like DPR by an average of 5 points absolute in top-20 and 4 points absolute in top-100 accuracy. We also train using all the questions contained in the Natural Questions (NQ) dataset (Kwiatkowski et al., 2019) and find that even with a mix of answerable and unanswerable questions, ART achieves strong generalization on out-of-distribution datasets due to relying on PLM. Our analysis further reveals that ART is highly sample-efficient, outperforming BM25 and DPR with just 100 and 1000 questions, respectively, on the NQ-Open dataset, and that scaling up to larger retriever models consistently improves performance.

2 Method

2.1 Problem Definition

We focus on open-domain retrieval, where, given a question $\boldsymbol{q}$, the task is to select a small set of matching passages (i.e., 20 or 100) from a large collection of evidence passages $\mathcal{D} = \{\boldsymbol{d}_1, \ldots, \boldsymbol{d}_m\}$. Our goal is to train a retriever in a *zero-shot manner*, that is, without using question-passage pairs, such that it retrieves relevant passages to answer the question. Our proposed approach consists of two core modeling components (§2.2, §2.3) and a novel training method (§2.4).

2.2 Dual Encoder Retriever

For the retriever, we use the dual-encoder model (Bromley et al., 1994) which consists of two encoders, where

- one encoder computes the question embedding $f_q(\boldsymbol{q}; \Phi_q) : \mathcal{X} \mapsto \mathbb{R}^d$, and
- the other encoder computes the passage embedding $f_d(\boldsymbol{d}; \Phi_d) : \mathcal{X} \mapsto \mathbb{R}^d$.

Here, $\mathcal{X} = \mathbb{V}^n$ denotes the universal set of text sequences, $\mathbb{V}$ denotes the vocabulary consisting of discrete tokens, and $\mathbb{R}^d$ denotes the (latent) embedding space. We assume that both the question and passage embeddings lie in the same latent space. The *retrieval score* for a question-passage pair $(\boldsymbol{q}, \boldsymbol{d})$ is then defined as the inner product between their respective embeddings,

$$\text{score}(\boldsymbol{q}, \boldsymbol{d}_i; \Phi) = f_q(\boldsymbol{q}; \Phi_q) \cdot f_d(\boldsymbol{d}_i; \Phi_d), \ \forall \boldsymbol{d}_i \in \mathcal{D}, \tag{1}$$

where $\Phi = [\Phi_q, \Phi_d]$ denotes the retriever parameters. We select the top-K passages with maximum inner product scores and denote them as $\mathcal{Z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_K\}$.[2]

[2]As the selection operation requires performing inner-product with millions of passage embeddings, this can be efficiently performed on accelerators such as GPUs.

We use the transformer network (Vaswani et al., 2017) with BERT tokenization (Devlin et al., 2019) to model both the encoders. To obtain the question or passage embedding, we do a forward pass through the transformer and select the last layer hidden state corresponding to the `[CLS]` token. As the input passage representation, we use both the passage title and text separated by `[SEP]` token.

2.3 Zero-Shot Cross-Attention Scorer

We obtain an estimate of the *relevance score* for a question-(retrieved) passage pair $(\boldsymbol{q}, \boldsymbol{z})$ by using a PLM. In order to do this in a zero-shot manner, we use a large generative PLM to compute the likelihood score of a passage conditioned on the question $p(\boldsymbol{z} \mid \boldsymbol{q})$.

The quantity $p(\boldsymbol{z} \mid \boldsymbol{q})$ can be better approximated by the autoregressive generation of question tokens conditioned on the passage and teacher-forcing (Sachan et al., 2022). More formally, this can be written as

$$\begin{aligned}\log p(\boldsymbol{z} \mid \boldsymbol{q}; \Theta) & \\ &= \log p(\boldsymbol{q} \mid \boldsymbol{z}; \Theta) + \log p(\boldsymbol{z}) + c \quad \text{(2a)} \\ &\propto \frac{1}{|\boldsymbol{q}|} \sum_t \log p(q_t \mid \boldsymbol{q}_{<t}, \boldsymbol{z}; \Theta), \quad \text{(2b)}\end{aligned}$$

where Θ denotes the parameters of the PLM, c is a constant independent of the passage $\boldsymbol{z}$, and $|\boldsymbol{q}|$ denotes the number of question tokens. Here, Eq. 2a follows from a simple application of Bayes' rule to $p(\boldsymbol{z} \mid \boldsymbol{q})$ and assuming that the passage prior $p(\boldsymbol{z})$ in Eq. 2b is uniform for all $\boldsymbol{z} \in \mathcal{Z}$.

We hypothesize that calculating the relevance score using Eq. 2b would be accurate because it requires performing deep cross-attention involving all the question and passage tokens. In a large PLM, the cross-attention step is highly expressive, and in combination with teacher-forcing, requires the model to explain every token in the question resulting in a better estimation.

As the input passage representation, we concatenate the passage title and its text. In order to prompt the PLM for question generation, we follow Sachan et al. (2022) and append a simple natural language instruction ''*Please write a question based on this passage.*'' to the passage text.

2.4 Training Algorithm

For training the model, our only assumption is that a collection of questions $(\mathcal{T})$ and evidence passages $(\mathcal{D})$ are provided as input. During training, the weights of the retriever are updated while the PLM is not finetuned, i.e., it is used in inference mode. Our training algorithm consists of five core steps. The first four steps are performed at every training iteration while the last step is performed every few hundred iterations. Figure 1 presents an illustration of our approach.

Step 1: Top-K Passage Retrieval For fast retrieval, we pre-compute the evidence passage embedding using the initial retriever parameters $(\hat{\Phi}_d)$. Given a question $\boldsymbol{q}$, we compute its embedding using the current question encoder parameters (Φ_q) and then retrieve the top-K passages $(\mathcal{Z})$ according to Eq. 1. We then embed these top-K passages using the current passage encoder parameters (Φ_d) and compute *fresh* retriever scores as

$$\text{score}(\boldsymbol{q}, \boldsymbol{z}_i) = f_q(\boldsymbol{q}; \Phi_q) \cdot f_d(\boldsymbol{z}_i; \Phi_d), \; \forall \boldsymbol{z}_i \in \mathcal{Z}.$$

Step 2: Retriever Likelihood Calculation Computing the exact likelihood of the passage conditioned on the question requires normalizing over all the evidence passages

$$p(\boldsymbol{z}_i \mid \boldsymbol{q}, \mathcal{D}; \Phi) = \frac{\exp(\text{score}(\boldsymbol{q}, \boldsymbol{z}_i)/\tau)}{\sum_{j=1}^{m} \exp(\text{score}(\boldsymbol{q}, \boldsymbol{d}_j)/\tau)},$$

where τ is a temperature hyperparameter. Computing this term is intractable, as this would require re-embedding all the evidence passages using Φ_d. Hence, we define a new distribution to approximate the likelihood of $\boldsymbol{z}_i$ as

$$q(\boldsymbol{z}_i \mid \boldsymbol{q}, \mathcal{Z}; \Phi) = \frac{\exp(\text{score}(\boldsymbol{q}, \boldsymbol{z}_i)/\tau)}{\sum_{j=1}^{K} \exp(\text{score}(\boldsymbol{q}, \boldsymbol{z}_j)/\tau)}, \quad (3)$$

which we also refer to as the *student distribution*. We assume that passages beyond the top-K contribute a very small probability mass, so we only sum over all the retrieved passages $\mathcal{Z}$ in the denominator. While this approximation leads to a biased estimate of retrieved passage likelihood, it works well in practice. Computing Eq. 3 is tractable as it requires embedding and backpropagating through a much smaller set of passages.

Step 3: PLM Relevance Score Estimation We compute the relevance score $\log p(\boldsymbol{z}_i \mid \boldsymbol{q})$ of all the passages in $\mathcal{Z}$ using a large PLM (Θ). This requires scoring the question tokens using teacher-forcing conditioned on a passage as described in §2.3. We then define a *teacher distribution* by applying softmax to the relevance scores

$$\hat{p}(\boldsymbol{z}_i \mid \boldsymbol{q}, \mathcal{Z}) = \frac{\exp(\log p(\boldsymbol{z}_i \mid \boldsymbol{q}; \Theta))}{\sum_{j=1}^{K} \exp\left(\log p(\boldsymbol{z}_j \mid \boldsymbol{q}; \Theta)\right)}.$$

Step 4: Loss Calculation and Optimization We train the retriever (Φ) by minimizing the KL divergence loss between the teacher distribution (obtained by PLM) and the student distribution (computed by retriever).

$$\mathcal{L}(\Phi) = \frac{1}{|\mathcal{T}|} \sum_{\boldsymbol{q} \in \mathcal{T}} \mathbb{KL}(\hat{p}(\boldsymbol{z}_i \mid \boldsymbol{q}, \mathcal{Z}) \mid\mid q(\boldsymbol{z}_i \mid \boldsymbol{q}, \mathcal{Z}; \Phi))$$

Intuitively, optimizing the KL divergence pushes the passage likelihood scores of the retriever to match the passage relevance scores from PLM by considering the relevance scores as soft-labels.

Step 5: Updating Evidence Embeddings During training, we update the parameters of both the question encoder (Φ_q) and passage encoder (Φ_d). Due to this, the pre-computed evidence embeddings that was computed using initial retriever parameters ($\hat{\Phi}_d$) becomes stale, which may affect top-K passage retrieval. To prevent staleness, we re-compute the evidence passage embeddings using current passage encoder parameters (Φ_d) after every 500 training steps.

2.5 ART as an Autoencoder

Since our encoder takes as input question $\boldsymbol{q}$ and the PLM scores (or reconstructs) the same question when computing the relevance score, we can consider our training algorithm as an autoencoder with a retrieved passage as the latent variable.

In the generative process, we start with an observed variable $\mathcal{D}$ (the collection of evidence passages), which is the support set for our latent variable. Given an input $\boldsymbol{q}$, we generate an index i and retrieve the passage $\boldsymbol{z}_i$. This index generation and retrieval process is modeled by our dual encoder architecture. Given $\boldsymbol{z}_i$, we decode it back into the question using our PLM.

Recall that our decoder (the PLM) is frozen and its parameters are not updated. However, the signal from the decoder output is used to train parameters of the dual encoder such that the log-likelihood of reconstructing the question $\boldsymbol{q}$ is maximized. In practice, this improves the dual encoder to select the best passage for a given question, since the only way to maximize the objective is by choosing the most relevant $\boldsymbol{z}_i$ given the input $\boldsymbol{q}$.

Dataset	Train Questions	Dev	Test
Question-Answering Datasets			
WebQ	3,417	361	2,032
NQ-Open	79,168	8,757	3,610
SQuAD-Open	78,713	8,886	10,570
TriviaQA	78,785	8,837	11,313
EQ	–	22,068	22,075
All Questions Datasets			
NQ-Full	307,373	–	–
MS MARCO	502,939	–	–

Table 1: Dataset statistics. During the training process, ART only uses the questions while evaluation is performed over the canonical development and test sets.

3 Experimental Setup

In this section, we describe the datasets, evaluation protocol, implementation details, and baseline methods for our passage retrieval experiments.

3.1 Datasets and Evaluation

Evidence Passages The evidence corpus includes the preprocessed English Wikipedia dump from December 2018 (Karpukhin et al., 2020). Following convention, we split an article into non-overlapping segments containing 100 words each, resulting in over 21 million passages. The same evidence is used for both training and evaluation.

Question-Answering Datasets Following previous work, we use the open-retrieval version of Natural Questions (NQ-Open; Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), SQuAD-1.0 (SQuAD-Open; Rajpurkar et al., 2016), WebQuestions (WebQ; Berant et al., 2013), and EntityQuestions (EQ; Sciavolino et al., 2021) datasets. Table 1 lists their training, development, and test set sizes.

All Questions Datasets For our transfer learning experiments, we use all the questions from

Natural Questions (henceforth referred to as NQ-Full) and MS MARCO passage ranking (Bajaj et al., 2016) datasets. Table 1 lists the number of questions. The questions in NQ-Full are information-seeking, as they were asked by real users. Its size is four times that of NQ-Open. NQ-Full consists of questions having just long-form of answers such as paragraphs, all the questions in NQ-Open (which have both long-form and short-form answers), questions having yes/no answers, and questions that do not contain the answer or are unanswerable. For MS MARCO, we use its provided passage collection (around 8.8 million passages in total) as the evidence corpus.

Evaluation To evaluate retriever performance, we report the conventional top-K accuracy metric. It is the fraction of questions for which at least one passage among the top-K retrieved passages contains a span of words that matches human-annotated answer(s) to the question.

3.2 Implementation Details

Model Sizes We use BERT base configuration (Devlin et al., 2019) for the retriever, which consists of 12 layers, 12 attention heads, and 768 embedding dimensions, leading to around 220M trainable parameters. For the teacher PLM, we use two configurations: (i) T5-XL configuration (Raffel et al., 2020) consisting of 24 layers, 32 attention heads, and 2048 embedding dimensions, leading to 3B parameters, and (ii) a larger T5-XXL configuration consisting of 11B parameters.

Model Initialization We initialize the retriever with unsupervised masked salient spans (MSS) pre-training (Sachan et al., 2021a) as it provides an improved zero-shot retrieval over BERT pre-training.[3] We initialize the cross-attention (or teacher) PLM with the T5-lm-adapted (Lester et al., 2021) or instruction-tuned T0 (Sanh et al., 2022) language models, which have been shown to be effective zero-shot re-rankers for information retrieval tasks (Sachan et al., 2022).

Compute Hardware We perform training on instances containing 8 or 16 A100 GPUs, each containing 40 GB RAM.

Passage Retrieval To perform fast top-K passage retrieval at every training step, we pre-compute the embeddings of all the evidence passages. Computing embeddings of 21M passages takes roughly 10 minutes on 16 GPUs. The total size of these embeddings is around 30 GB (768-dimensional vectors in FP16 format). For scalable retrieval, we shard these embeddings across all the GPUs and perform exact maximum inner product search using distributed matrix multiplication.

Training Details When training with T0 (3B) PLM, for all the datasets except WebQ, we perform training for 10 epochs using Adam with a batch size of 64, 32 retrieved passages, dropout value of 0.1, and peak learning rate of 2×10^{-5} with warmup and linear scheduling. Due to the smaller size of WebQ, we train for 20 epochs with a batch size of 16. When training with the T5-lm-adapted (11B) PLM, we use a smaller batch size of 32 with 16 retrieved passages. We save the retriever checkpoint every 500 steps and perform model selection by evaluating it on the development set. We use mixed precision training to train the retriever and perform inference over the PLM using bfloat16 format. We set the value of the temperature hyperparameter (τ) using cross-validation.

3.3 Baselines

We compare ART to both unsupervised and supervised models. Unsupervised models train a single retriever using unlabeled text corpus from the Internet while supervised models train a separate retriever for each dataset. We report the performance numbers from the original papers when the results are available or run their open-source implementations in case the results are not available.

Unsupervised Models These include the popular BM25 algorithm (Robertson and Zaragoza, 2009) that is based on the sparse bag-of-words representation of text. Dense models typically use Wikipedia paragraphs to create (pseudo-) query and context pairs to perform contrastive training of the retriever. These differ in how the negative examples are obtained during contrastive training: They can be from the same batch (ICT; Lee et al., 2019; Sachan et al., 2021a), or contexts passages from previous batches (Contriever; Izacard et al., 2022), or by using other passages in the same article (Spider; Ram et al., 2022). Context

[3]We use the open-source MSS retriever checkpoint from `https://github.com/DevSinghSachan/emdr2`.

passages can also be sampled from articles connected via hyperlinks (HLP; Zhou et al., 2022).

Supervised Models These consist of approaches that use questions and positive passages to perform contrastive training of the retriever. To obtain improved performance an additional set of hard-negative passages is often used (DPR; Karpukhin et al., 2020), iterative mining of negative contexts is done using model weights (ANCE; Xiong et al., 2021), or the retriever is first initialized with ICT or MSS pre-training followed by DPR-style finetuning (ICT-DPR / MSS-DPR; Sachan et al., 2021a). The pre-trained retriever can be further trained by ANCE-style mining of hard-negative passages to further improve accuracy (coCondenser; Gao and Callan, 2022). Previous methods have also explored finetuning the cross-encoder PLM jointly with the retriever such that cross-encoder provides more accurate training signals to improve retrieval accuracy. Among them include the approaches of end-to-end training of PLM and retriever, which infuses supervision from the annotated answers to a question (EMDR^2; Sachan et al., 2021b), and a multi-stage mixed objective distillation approach to jointly train the re-ranker (Nogueira and Cho, 2019) and retriever (RocketQAv2; Ren et al., 2021). A combination of adversarial and distillation-based training of re-ranker and retriever has been shown to obtain state-of-the-art performance (AR2; Zhang et al., 2022).

4 Experiments and Results

4.1 Zero-Shot Passage Retrieval

For the passage retrieval task, we report results on SQuAD-Open, TriviaQA, NQ-Open, and WebQ and train ART under two settings. In the first setting, we train a separate retriever for each dataset using questions from their training set. In the second setting, to examine the robustness of ART training to different question types, we train a single retriever by combining the questions from all the four datasets, which we refer to as ART-Multi. For both these settings, we train ART using T5-lm-adapted (11B) and T0 (3B) cross-attention PLM scorers. As our training process does not require annotated passages for a question, we refer to this as *zero-shot passage retrieval*.

Table 2 presents the top-20 and top-100 retrieval accuracy in these settings alongside recent baselines that train a similarly sized retriever (110M). All the variants of ART achieve substantially better performance than previous unsupervised approaches. For example, ART trained with T0 (3B) outperforms the recent Spider and Contriever models by an average of 9 points on top-20 and 6 points on top-100 accuracy. When comparing with supervised models, despite using just questions, ART outperforms strong baselines like DPR and ANCE and is on par or slightly better than pre-trained retrievers like MSS-DPR. In addition, ART-Multi obtains comparable performance to its single dataset version, a considerable advantage in practical applications as a single retriever can be deployed rather than training a custom retriever for each use case.

ART's performance also comes close to the state-of-the-art supervised models like AR2 and EMDR^2, especially on the top-100 accuracy but lags behind in the top-20 accuracy. In addition to obtaining reasonable performance and not requiring aligned passages for training, ART's training process is much simpler than AR2. It also does not require cross-encoder finetuning and is thus faster to train. As generative language models continue to become more accurate (Chowdhery et al., 2022), we hypothesize that the performance gap between state-of-the-art supervised models and ART would further narrow down.

Our results showcase that both the PLM scorers, T5-lm-adapt (11B) and T0 (3B), achieve strong results on the QA retrieval tasks, with T0 achieving higher performance gains. This illustrates that the relevance score estimates of candidate passages obtained in the zero-shot cross-attention step are accurate enough to provide strong supervision for retriever training. We believe that this is a direct consequence of the knowledge stored in the PLM weights. While T5-lm-adapt's knowledge is obtained by training on unsupervised text corpora, T0 was further finetuned using instruction-prompted datasets of tasks such as summarization, QA, text classification, etc.[4] Hence, in addition to learning from instructions, the performance gains from T0 can be attributed to the knowledge infused in its

[4]However, we note that T0 was not finetuned on the question generation task and not trained on any of the datasets we have used in this work. We refer the reader to the original paper for more training details.

Retriever	Cross-Attention Language Model	SQuAD-Open		TriviaQA		NQ-Open		WebQ	
		Top-20	Top-100	Top-20	Top-100	Top-20	Top-100	Top-20	Top-100
Unsupervised Approaches (trained using Wikipedia / Internet data)									
BERT		5.2	13.5	7.2	17.8	9.4	20.3	3.7	12.8
ICT		45.1	65.2	57.5	73.6	50.6	66.8	43.4	65.7
MSS	T5* (220M)	51.3	68.4	68.2	79.4	59.8	74.9	49.2	68.4
BM25		71.1	81.8	76.4	83.2	62.9	78.3	62.4	75.5
Contriever		63.4	78.2	74.2	83.2	67.8	82.1	74.9	80.1
Spider		61.0	76.0	75.8	83.5	68.3	81.2	65.9	79.7
cpt-text S†		–	–	75.1	81.7	65.5	77.2	–	–
HLP		–	–	76.9	84.0	70.2	82.0	66.9	80.8
Supervised Approaches (trained using question-passage aligned data)									
DPR		63.2	77.2	79.4	85.0	78.4	85.4	73.2	81.4
DPR-Multi‡		51.6	67.6	78.8	84.7	79.4	86.0	75.0	82.9
ANCE		–	–	80.3	85.3	81.9	87.5	–	–
ICT-DPR		–	–	81.7	86.3	81.8	88.0	72.5	82.3
MSS-DPR◇		73.1	84.5	81.8	86.6	82.1	87.8	76.9	84.6
coCondenser		–	–	83.2	87.3	84.3	89.0	–	–
RocketQAv2	ERNIE* (110M)	–	–	–	–	83.7	89.0	–	–
EMDR²○	T5* (220M)	–	–	83.4	87.3	85.3	89.7	**79.1**	**85.2**
AR2	ERNIE* (330M)	–	–	**84.4**	**87.9**	**86.0**	**90.1**	–	–
Our Approach (trained using questions and Wikipedia text)									
ART	T5-lm-adapt (11B)	74.2	84.3	82.5	86.6	80.2	88.4	74.4	82.7
ART-Multi	T5-lm-adapt (11B)	72.8	83.2	82.2	86.6	81.5	88.5	74.8	83.7
ART	T0 (3B)	75.3	85.0	82.9	87.1	81.6	89.0	75.7	84.3
ART-Multi	T0 (3B)	74.7	84.5	82.9	87.0	82.0	88.9	76.6	**85.0**

Table 2: Top-20 and top-100 retrieval accuracy on the test set of datasets. For more details regarding the unsupervised and supervised models, please see §3.3 in the text. Best supervised results are highlighted in bold while best results from the our proposed model (ART) are underlined. ART substantially outperforms previous unsupervised models and comes close to or matches the performance of supervised models by just using questions during training. * indicates that the cross-attention PLM is finetuned. † denotes that 'cpt-text S' model (Neelakantan et al., 2022) contains around 300M parameters. ‡ denotes that DPR-Multi was not trained on SQuAD-Open. ◇ indicates that the results on SQuAD-Open and WebQ are obtained by finetuning the open-source MSS checkpoint. ○ indicates that EMDR^2 results are obtained using their open-source checkpoints.

weights by (indirect) supervision from these manually curated datasets. Instruction-based finetuning is helpful in the case of smaller datasets like WebQ and especially in improving the performance on lower values of top-K accuracy (such as top-20).[5]

Overall, our results suggest that an *accurate and robust passage retrieval can be achieved by training with questions alone*. This presents a considerably more favorable setting than the current approaches which require obtaining positive and hard-negative passages for such questions. Due to its better performance, we use the T0 (3B) PLM for subsequent experiments unless stated otherwise.

4.2 Sample Efficiency

To measure the sample efficiency of ART, we train the model by randomly selecting a varying number of questions from NQ-Open training questions and compute the top-K accuracy on its development set. These results are presented in Figure 2 and we also include the results of BM25

[5] §4.5 includes more detailed comparisons of different PLMs as cross-encoders.

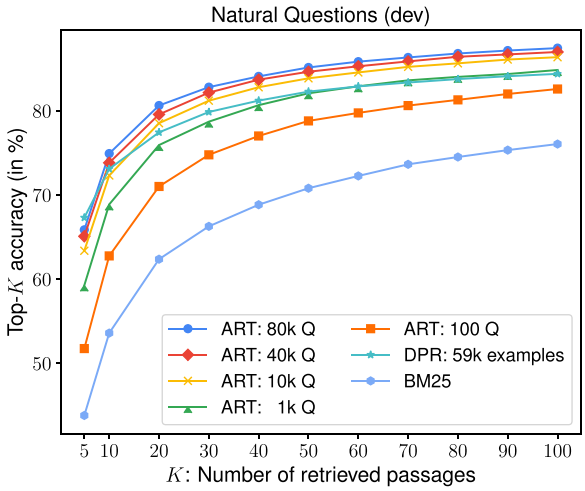

Figure 2: Top-K accuracy as the number of training questions (denoted as 'Q' in the legend) is varied. When trained with 100 questions, ART outperforms BM25 and when trained with 1k questions, it matches DPR's performance for top-$K > 50$ passages, illustrating that ART is highly sample efficient.

and DPR for comparison. We see that performance increases with the increase in questions until about 10k questions, after which the gains become less pronounced.

When trained with just 100 questions, ART significantly outperforms BM25 and when trained with 1k questions, it matches DPR performance levels for top-$\{50, \ldots, 100\}$ accuracy. *This demonstrates that* ART *in addition to using just questions is also much more data efficient than* DPR, as it requires almost ten times fewer questions to reach a similar performance.

4.3 Zero-Shot Out-of-Distribution Transfer

In the previous experiments, both the training and test sets contained questions that were sampled from the same underlying distribution, a setting that we refer to as *in-distribution training*. However, obtaining in-domain questions for training is not always feasible in practice. Instead, a model trained on an existing collection of questions must be evaluated on new datasets, a setting that we refer to as *out-of-distribution* (OOD) transfer.

We train ART using NQ-Open and NQ-Full questions and then evaluate its performance on SQuAD-Open, TriviaQA, WebQ, and EQ datasets. It is desirable to train on answerable questions such as the ones included in NQ-Open, but this is not always possible, as real user questions are often imprecisely worded or ambiguous. Due to this, training on NQ-Full can be considered as a practical testbed for evaluating true OOD generalization as a majority of the questions (51%) were marked as unanswerable from Wikipedia by human annotators.[6]

Table 3 presents OOD generalization results on the four QA datasets including the results of DPR and Spider models trained on NQ-Open.[7] ART trained on NQ-Open always performs significantly better than both DPR and Spider, showing that it is better at generalization than supervised models. When trained using NQ-Full, ART performance further improves by 3 and 0.5–1 points on EQ and other datasets, respectively, over NQ-Open. This highlights that in addition to questions annotated as having short answers, *questions annotated with long answers also provide meaningful supervisory signals* and *unanswerable questions do not necessarily degrade performance*.

We also train ART using MS MARCO questions and perform OOD evaluation. Due to the larger size of MS MARCO and a smaller number of evidence passages, we use a batch size of 512 and retrieve 8 passages for training. Quite surprisingly, it obtains much better performance than previous approaches including BM25 on EQ (more than 10 points gain on top-20 over training ART on NQ-Open). We suspect that this may be due to the similar nature of questions in MS MARCO and EQ. Further finetuning the pre-trained MS MARCO model on NQ-Full significantly improves performance on WebQ.

4.4 Scaling Model Size

We examine if scaling up the retriever parameters can offer further performance improvements. To this end, we train a retriever of BERT-large configuration (24 layers, 16 attention heads, 1024 embedding dimensions) containing around 650M parameters on NQ-Open and TriviaQA. Results are presented in Table 4 for both the development

[6]The reasons for question unanswerability can be partly attributed to imprecise Wikipedia article retrieval during the annotation process, ambiguity in information-seeking questions, information required to answer not being localized to a single paragraph, etc. (Kwiatkowski et al., 2019).

[7]We also include BM25 results for reference but do not directly compare with them because there is a high lexical overlap between question and passage tokens in the SQuAD-Open and EQ datasets which renders dense retrievers at a disadvantage over BM25, especially in the transfer setting.

Retriever	Training Dataset	SQuAD-Open		TriviaQA		WebQ		EQ	
		Top-20	Top-100	Top-20	Top-100	Top-20	Top-100	Top-20	Top-100
Training on answerable questions									
BM25	–	**71.1**	**81.8**	76.4	83.2	62.4	75.5	71.2	79.8
DPR†	NQ-Open	48.9	65.2	69.0	78.7	68.8	78.3	49.7	63.2
EMDR^2	NQ-Open	66.8	79.0	79.7	85.3	74.2	83.2	62.7	75.1
Spider†	NQ-Open	57.7	72.8	77.2	83.7	74.2	82.5	61.9	74.1
ART	NQ-Open	68.0	80.2	79.8	85.1	73.4	83.1	64.3	75.5
ART	MS MARCO	68.4	80.4	78.0	84.1	74.8	83.2	**75.3**	**81.9**
Training on a mix of answerable and unanswerable questions									
ART	NQ-Full	69.4	**81.1**	80.3	**85.7**	74.3	83.9	67.8	78.3
ART	MS MARCO + NQ-Full	69.6	**81.1**	**80.7**	**85.7**	**75.3**	**84.5**	69.2	79.1

Table 3: Top-20 and top-100 retrieval accuracy when evaluating zero-shot out-of-distribution (OOD) generalization of models on the test set of datasets. † denotes that these results are from Ram et al. (2022). ART generalizes better than supervised models on OOD evaluation even when trained on all the questions of the Natural Questions dataset, which contains a mix of answerable and unanswerable questions.

	Retriever	NQ-Open		TriviaQA	
		Top-20	Top-100	Top-20	Top-100
Dev	ICT	44.2	61.0	58.8	74.4
	DPR	79.1	85.5	81.1	85.9
	ICT-DPR	81.4	87.4	82.8	86.9
	EMDR^2	83.1	88.0	83.7	87.4
	ART-base	80.6	87.4	83.6	87.4
	ART-large	**81.0**	**87.8**	**83.7**	**87.5**
Test	ICT	49.3	66.1	58.5	74.1
	DPR	81.0	87.2	81.4	86.0
	ICT-DPR	82.6	88.3	82.9	87.1
	EMDR^2	85.3	89.7	83.4	87.3
	ART-base	81.6	**89.0**	82.9	87.1
	ART-large	**82.1**	88.8	**83.6**	**87.6**

Table 4: Top-20 and top-100 accuracy when training large configuration retriever, which contains around 650M parameters. EMDR^2 (base configuration) (Sachan et al., 2021b) contains 440M parameters. Best supervised results are underlined while the best unsupervised results are highlighted in bold.

and test sets. We also include the results of other relevant baselines containing a similar number of trainable parameters.

By scaling up the retriever size, we see small but consistent improvements in retrieval accuracy across both the datasets. Especially on TriviaQA, ART matches or exceeds the performance of previous best models. On NQ-Open, it comes close to the performance of EMDR^2 (Sachan et al., 2021b), a supervised model trained using thousands of question-answer pairs.

We also attempted to use larger teacher PLMs such as T0 (11B). However, our initial experiments did not lead to any further improvements over the T0 (3B) PLM. We conjecture that this might be either specific to these QA datasets or that we need to increase the capacity of the teacher PLM even more to observe improvements. We leave an in-depth analysis of using larger teacher PLMs as part of the future work.

4.5 Analysis

Sensitivity to Retriever Initialization To examine how the convergence of ART training is affected by the initial retriever parameters, we initialize the retriever with (1) BERT weights, (2) ICT weights (as trained in Sachan et al., 2021a), and (3) MSS weights, and train using NQ-Open questions. Figure 3 displays the top-20 performance on the NQ development set as the training progresses. It reveals that ART training is not sensitive to the initial retriever parameters as all three initialization schemes converge to similar results. However, the convergence properties might be different under low-resource settings, an exploration of which we leave for future work.

Effect of the Number of Retrieved Passages Table 5 quantifies the effect of the number of retrieved passages used during training on performance. A smaller number of retrieved passages

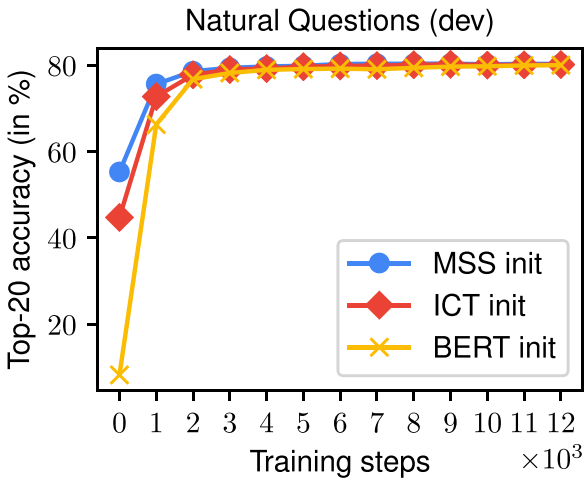

Figure 3: Effect of retriever initialization on ART training. The plot reveals that the training process is not sensitive to initial retriever parameters.

Retrieved Passages	Top-1	Top-5	Top-20	Top-100
32	36.7	65.8	80.6	87.4
2	+2.4	+0.4	−0.9	−0.6
4	+1.9	+0.9	−0.6	−0.6
8	+0.8	+0.8	−0.5	−0.1
16	+0.9	+0.9	−0.3	−0.1
64	−0.5	−0.7	−0.3	0
128	−2.3	−1.7	−0.8	−0.2

Table 5: Effect of using a different number of retrieved passages during ART training as evaluated on the NQ-Open development set. For each case, we list the absolute gain or loss in top-K accuracy when compared to the setting utilizing 32 retrieved passages.

such as 2 or 4 leads to a somewhat better top-$\{1, 5\}$ accuracy, at the expense of a drop in top-$\{20, 100\}$ accuracy. Retrieving 32 passages offers a reasonable middle ground and beyond that, the top-K retrieval performance tends to drop.

A Closer Inspection of ART with Supervised Models In order to have a better understanding of the tradeoff between supervised models and ART, we examine their top-1 and top-5 accuracy in addition to the commonly reported top-20 and top-100 scores. Table 6 presents these results for ART (large) along with supervised models of DPR (large) and EMDR^2. Supervised models achieve much better performance for top-$K \in \{1, \ldots, 5\}$ passages, i.e., these models are more

Retriever	Top-1	Top-5	Top-20	Top-100
NQ-Open (dev)				
DPR	**50.1**	**69.6**	79.1	85.5
EMDR^2	**55.3**	**74.9**	83.1	88.0
ART	37.6	66.8	81.0	87.8
TriviaQA (dev)				
DPR	59.6	74.4	81.1	85.9
EMDR^2	**63.7**	78.0	83.7	87.4
ART	58.3	77.5	83.7	87.5

Table 6: Analysis reveals that ART (large) can even match the performance of end-to-end trained models like EMDR^2 when retrieving a larger number of passages. However, DPR (large) and EMDR^2 still outperform ART when retrieving a small number of passages such as top-$K \in \{1, \ldots, 5\}$ (highlighted in bold).

precise. This is likely because DPR is trained with hard-negative passages and EMDR^2 finetunes PLM using answers resulting in an accurate relevance feedback to the retriever. When considering top-$K \in \{20, \ldots, 100\}$ passages, ART comes close or matches the performance of EMDR^2. As top-performing models for knowledge-intensive tasks such as open-domain QA rely on a larger set of retrieved passages, such as top-$K = 100$ (Sachan et al., 2022), this justifies the argument to adopt zero-shot ART over supervised retrievers.

Why Training using Passage Retrieval? To assess the importance of passages in $\mathcal{Z}$ during the training process, we train the retriever under different settings by varying the passage types. Specifically, we train with a mix of positive, hard-negative, and uniformly sampled passages. We also perform in-batch training by defining $\mathcal{Z}$ to be the union of positive and hard-negative passages for all the questions in a batch. Results in Table 7 illustrate that when $\mathcal{Z}$ consists of uniformly sampled passages, it leads to poor performance. Including a (gold) positive passage in $\mathcal{Z}$ leads to good performance improvements. Results further improve with the inclusion of a hard-negative passage in $\mathcal{Z}$. However, in-batch training leads to a slight drop in performance. As the gold passages are not always available, our method of selecting the top passages from evidence at every training step can be seen as an

P	N	U	IB	Top-1	Top-5	Top-20	Top-100
0	0	32	✗	6.0	16.6	30.8	46.7
1	0	31	✗	31.8	58.9	74.8	84.4
1	1	30	✗	33.7	61.0	76.0	85.5
1	1	0	✓	32.6	59.5	75.1	84.9
Top-32 passages				**36.7**	**65.8**	**80.6**	**87.4**

Table 7: Effect of passage types on ART training when evaluated on the NQ-Open development set. P denotes a positive passage, N denotes a hard-negative passage (mined using BM25), U denotes that the passages are randomly sampled from the evidence, and IB denotes in-batch training.

approximation to using the gold passages. With this, ART obtains even better results than the previous settings, an improvement by 4 points absolute in the top-20 accuracy.

Impact of Language Model Training Strategy We examine which PLMs can provide good cross-attention scores during training. We compare across PLMs trained using three different objectives: (i) generative denoising of masked spans (T5 series; Raffel et al., 2020), (ii) further pre-training using autoregressive language modeling objective (T5-lm-adapt series; Lester et al., 2021), and (iii) finetuning T5-lm-adapt models on unrelated tasks using instructions (T0 series; Sanh et al., 2022). Our results in Table 8 highlight that PLM training methodology and model size can have a large effect on retrieval performance. T5 base model leads to low scores possibly because pre-training using predicting masked spans is not ideal for question reconstruction. However, the accuracy improves with an increase in model size. T5-lm-adapt models are more stable and lead to improved performance with the best result achieved by the 11B model. Instruction finetuned T0 models outperforms the T5-lm-adapt models. However, scaling up the size of T0 to 11B parameters does not result in meaningful improvements.

Ad-Hoc Retrieval Tasks While the previous experiments were conducted on QA datasets, here we examine the robustness of the ART model trained using questions to different ad-hoc retrieval tasks. For this analysis, we evaluate the performance of ART on the BEIR benchmark (Thakur et al., 2021). It is a heterogeneous collection of many retrieval datasets, with each dataset consisting of test set queries, evidence documents, and gold document annotations. BEIR spans multiple domains and diverse retrieval tasks presenting a strong challenge suite, especially to the dense retrievers. We train ART using MS MARCO questions and report its nDCG@10 and Recall@100 scores on each dataset. For comparison, we include the results of three baselines: BM25, Contriever, and DPR trained using NQ-Open. Our results presented in Table 9 show strong generalization performance of ART as it outperforms DPR and Contriever results. ART also achieves at par results with the strong BM25 baseline outperforming BM25 on 8 out of the 15 datasets (according to nDCG@10 scores).

	NQ-Open (dev)			
Language Model (Θ)	Top-1	Top-5	Top-20	Top-100
Models trained using Denoising Masked Spans				
T5-base (250M)	12.8	30.9	47.8	63.0
T5-xl (3B)	25.0	53.9	74.4	85.3
T5-xxl (11B)	29.5	59.8	77.8	86.3
Models trained using Language Modeling Objective				
T5-lm-adapt (250M)	29.4	56.6	74.4	84.7
T5-lm-adapt (800M)	30.9	59.1	76.5	85.9
T5-lm-adapt (3B)	31.8	61.0	77.9	86.5
T5-lm-adapt (11B)	32.7	62.6	78.6	87.0
Model trained using Natural Language Instructions				
T0 (3B)	**36.7**	**65.8**	**80.6**	**87.4**
T0 (11B)	34.3	64.5	79.8	87.2

Table 8: Comparison of different PLMs when used as cross-attention scorers during training (§2.3). T0 (3B) PLM achieves the highest accuracy among the compared PLMs showcasing that training language models using instruction-tuning provides accurate relevance scores.

5 Related Work

Our work is based on training a dense retriever using PLMs, which we have covered in previous sections. Here, we instead focus on other related approaches.

A popular method to train the dual-encoder retriever is to optimize contrastive loss using in-batch negatives (Gillick et al., 2019) and hard-negatives (Karpukhin et al., 2020; Xiong et al., 2021). Alternatives to using hard-negatives such as sampling from cached evidence embeddings

Dataset	#Q	#E	nDCG@10				Recall@100			
			DPR†	BM25†	Contriever	ART	DPR†	BM25†	Contriever	ART
Scifact	300	5K	31.8	**66.5**	64.9	55.2	72.7	90.8	**92.6**	88.0
Scidocs	1000	25K	7.7	**15.8**	14.9	14.4	21.9	35.6	**36.0**	32.4
Nfcorpus	323	3.5K	18.9	**32.5**	31.7	29.9	20.8	25.0	**29.0**	26.6
FIQA-2018	648	57K	11.2	23.6	24.5	**26.5**	34.2	53.9	**56.2**	55.4
Trec-covid	50	0.2M	33.2	**65.5**	27.4	50.3	21.2	**49.8**	17.2	36.9
Touche-2020	49	0.4M	13.1	**36.8**	19.3	16.2	30.1	**53.8**	22.5	44.7
NQ	3452	2.7M	**47.4**	32.9	25.4	40.5	88.0	76.0	77.1	**88.7**
MS-Marco	6980	8.8M	17.7	22.8	20.6	**32.6**	55.2	65.8	67.2	**81.7**
HotpotQA	7405	5.2M	39.1	60.3	48.1	**61.0**	59.1	**74.0**	70.4	73.9
ArguAna	1406	8.7K	17.5	31.5	**37.9**	32.2	75.1	94.2	90.1	**95.3**
CQADupStack	13145	0.5M	15.3	29.9	28.4	**33.5**	40.3	60.6	61.4	**62.6**
Quora	10000	0.5M	24.8	78.9	83.5	**84.2**	47.0	97.3	98.7	**98.8**
DBpedia	400	4.6M	26.3	31.3	29.2	**36.3**	34.9	39.8	45.3	**47.2**
Fever	6666	5.4M	56.2	**75.3**	68.2	72.4	84.0	93.1	**93.6**	93.1
Climate-Fever	1535	5.4M	14.8	21.3	15.5	**21.4**	39.0	43.6	44.1	**47.1**
Average Score			25.0	41.6	36.0	40.4	48.2	63.6	60.1	64.8

Table 9: Results on the BEIR benchmark. #Q and #E denotes the size of the test set and evidence, respectively. Best scores for each dataset are highlighted in bold. ART is trained using MS MARCO questions. DPR is trained using NQ-Open. † denotes that these results are from Thakur et al. (2021).

have also shown to work well in practice (Lindgren et al., 2021). Multi-vector encoders for questions and passages are more accurate than dual-encoders, (Luan et al., 2021; Khattab and Zaharia, 2020; Humeau et al., 2020), although at the cost of an increased latency and storage requirements.

PLMs have been shown to improve passage rankings as they can perform cross-attention between the question and the retrieved passages (Lin et al., 2021). Supervised approaches to re-rank either finetune PLMs using question-passage pairs (Nogueira et al., 2020) or finetune PLMs to generate question conditioned on the passage (Nogueira dos Santos et al., 2020) while unsupervised re-rankers are based on zero-shot question scoring (Sachan et al., 2022). The re-ranking process is slow due to the cross-attention step and is bottlenecked by the accuracy of first-stage retrievers. To address these limitations, cross-attention distillation approaches from the PLM to retriever have been proposed (Qu et al., 2021). Such distillation can be performed either in a single end-to-end training step (Guu et al., 2020; Sachan et al., 2021b) or in a multi-stage process (Khattab et al., 2021; Izacard and Grave, 2021).

An alternative approach to using PLMs is to generate data that can aid retrieval. The data can be either the title or an answer that provides more information about the question (Mao et al., 2021). Generating new questions to augment the training data has also been shown to improve performance (Ma et al., 2021; Bonifacio et al., 2022; Dai et al., 2022). In comparison, we do not generate new questions but train the retriever using existing questions and PLM feedback. Data augmentation is likely complementary, and can further improve accuracy.

6 Conclusions and Future Work

We introduced ART, a novel approach to train a dense passage retriever using only questions. ART does not require question-passage pairs or hard-negative examples for training and yet achieves state-of-the-art results. The key to making ART work is to optimize the retriever to select relevant passages such that conditioning on them, the question generation likelihood computed using a large pre-trained language model iteratively improves. Despite requiring much less supervision, ART substantially outperforms DPR when evaluated on

multiple QA datasets and also generalizes better on out-of-distribution questions.

ART presents several directions for future work. It would be interesting to apply this approach in low-resource retrieval including multilingual (Clark et al., 2020) and cross-lingual question answering (Asai et al., 2021). Our training framework can also be extended to train cross-modality retrievers such as for image or code search (Li et al., 2022; Neelakantan et al., 2022) using textual queries. Finally, other directions worth exploring would be to make use of labeled data when available such as by finetuning PLM on passage-question aligned data and to train multi-vector retrievers (Luan et al., 2021) with ART.

Acknowledgments

We are grateful to the TACL action editor and the three anonymous reviewers for providing us valuable feedback and useful suggestions that helped to improve this work. We would also like to thank Elena Gribovskaya from DeepMind for providing us valuable comments to improve the paper.

References

Akari Asai, Jungo Kasai, Jonathan Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021. XOR QA: Cross-lingual open-retrieval question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. https://doi.org/10.18653/v1/2021.naacl-main.46

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*. https://doi.org/10.48550/arXiv.1611.09268

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*.

Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. 2022. Inpars: Unsupervised dataset generation for information retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. https://doi.org/10.1145/3477495.3531863

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1994. Signature verification using a ‘‘siamese’’ time delay neural network. In *Advances in Neural Information Processing Systems*. https://doi.org/10.1142/9789812797926_0003

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Benton C. Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Oliveira Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathleen S. Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling language modeling with pathways. *ArXiv*, abs/2204.02311. https://doi.org/10.48550/arXiv.2204.02311

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470. https://doi.org/10.1162/tacl_a_00317

Zhuyun Dai, Arun Tejasvi Chaganty, Vincent Y. Zhao, Aida Amini, Qazi Mamunur Rashid, Mike Green, and Kelvin Guu. 2022. Dialog inpainting: Turning documents into dialogs. In *Proceedings of the 39th International Conference on Machine Learning*.

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)*. https://doi.org/10.18653/v1/N19-1423

Luyu Gao and Jamie Callan. 2022. Unsupervised corpus aware language model pre-training for dense passage retrieval. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. https://doi.org/10.18653/v1/2022.acl-long.203

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)*. https://doi.org/10.18653/v1/K19-1049

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*.

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. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Gautier Izacard and Edouard Grave. 2021. Distilling knowledge from reader to retriever for question answering. In *International Conference on Learning Representations*.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. https://doi.org/10.18653/v1/P17-1147

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, 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)*. https://doi.org/10.18653/v1/2020.emnlp-main.550

Omar Khattab, Christopher Potts, and Matei Zaharia. 2021. Relevance-guided supervision for openqa with colbert. *Transactions of the Association for Computational Linguistics*, 9:929–944. https://doi.org/10.1162/tacl_a_00405

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. https://doi.org/10.1145/3397271.3401075

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association of Computational Linguistics*, 7:453–466. https://doi.org/10.1162/tacl_a_00276

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*. https://doi.org/10.18653/v1/P19-1612

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/2021.emnlp-main.243

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de Masson d'Autume, Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey Cherepanov, James Molloy, Daniel J. Mankowitz, Esme Sutherland Robson, Pushmeet Kohli, Nando de Freitas, Koray Kavukcuoglu, and Oriol Vinyals. 2022. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097. https://doi.org/10.1126/science.abq1158, PubMed: 36480631

Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2021. Pretrained transformers for text ranking: BERT and beyond. *Synthesis Lectures on Human Language Technologies*, 14(4):1–325. https://doi.org/10.2200/S01123ED1V01Y202108HLT053

Erik Lindgren, Sashank J. Reddi, Ruiqi Guo, and Sanjiv Kumar. 2021. Efficient training of retrieval models using negative cache. In *Advances in Neural Information Processing Systems*.

Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2021. Sparse, dense, and attentional representations for text retrieval. *Transactions of the Association for Computational Linguistics*, 9:329–345. https://doi.org/10.1162/tacl_a_00369

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*. https://doi.org/10.18653/v1/2021.eacl-main.92

Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2021. Generation-augmented retrieval for open-domain question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. https://doi.org/10.18653/v1/2021.acl-long.316

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 pre-training. *arXiv preprint arXiv:2201.10005*. https://doi.org/10.48550/arXiv.2201.10005

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*. https://doi.org/10.48550/arXiv.1901.04085

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pretrained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. https://doi.org/10.18653/v1/2020.findings-emnlp.63

Cicero Nogueira dos Santos, Xiaofei Ma, Ramesh Nallapati, Zhiheng Huang, and Bing Xiang. 2020. Beyond [CLS] through ranking by generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. https://doi.org/10.18653/v1/2020.emnlp-main.134

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*. https://doi.org/10.48550/arXiv.1807.03748

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*. https://doi.org/10.18653/v1/2021.naacl-main.466

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.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/D16-1264

Ori Ram, Gal Shachaf, Omer Levy, Jonathan Berant, and Amir Globerson. 2022. Learning to retrieve passages without supervision. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. https://doi.org/10.18653/v1/2022.naacl-main.193

Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, QiaoQiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021. 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*. https://doi.org/10.18653/v1/2021.emnlp-main.224

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*. https://doi.org/10.1561/1500000019

Devendra Singh Sachan, Mike Lewis, Mandar Joshi, Armen Aghajanyan, Wen-tau Yih, Joelle Pineau, and Luke Zettlemoyer. 2022. Improving passage retrieval with zero-shot question generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.48550/arXiv.2204.07496

Devendra Singh Sachan, Mostofa Patwary, Mohammad Shoeybi, Neel Kant, Wei Ping, William L. Hamilton, and Bryan Catanzaro. 2021a. End-to-end training of neural retrievers for open-domain question answering. In *Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*. https://doi.org/10.18653/v1/2021.acl-long.519

Devendra Singh Sachan, Siva Reddy, William L. Hamilton, Chris Dyer, and Dani Yogatama. 2021b. End-to-end training of multi-document reader and retriever for open-domain question answering. In *Advances in Neural Information Processing Systems*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. 2021. Simple entity-centric questions challenge dense retrievers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/2021.emnlp-main.496

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)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

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*.

Hang Zhang, Yeyun Gong, Yelong Shen, Jiancheng Lv, Nan Duan, and Weizhu Chen. 2022. Adversarial retriever-ranker for dense text retrieval. In *International Conference on Learning Representations*.

Jiawei Zhou, Xiaoguang Li, Lifeng Shang, Lan Luo, Ke Zhan, Enrui Hu, Xinyu Zhang, Hao Jiang, Zhao Cao, Fan Yu, Xin Jiang, Qun Liu, and Lei Chen. 2022. Hyperlink-induced pre-training for passage retrieval in open-domain question answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. https://doi.org/10.18653/v1/2022.acl-long.493